Universidad Nacional Autónoma
de México

Facultad de Ciencias

Aplicación de distintos métodos de agrupamiento en piezas musicales

T E S I S

que para obtener el título de:

Actuario

P R E S E N T A :

Carlos Fernando Vásquez Guerra

Tutora:
Dra. Ruth Selene Fuentes García

CIUDAD DE MÉXICO 2022

Este trabajo lo quiero dedicar a todas las personas que han formado parte de mi vida en todos estos años, en especial a mis padres: Maria Leticia Guerra Flores y Elpidio Vásquez Vásquez por su apoyo incondicional, su paciencia y su cariño al igual que mis hermanos. A mi asesora la Dra. Ruth Selene Fuentes por su infinita paciencia, a Sarahy Huesca por haberme reconfortado y recorrido con su amor este camino conmigo y a Allison Odette Merino, por ser la mejor amiga en todo este tiempo. ¡Gracias!

Índice general

Capítulo 1

Preliminares

1.1. Resumen

En el presente trabajo se estudian y se da una breve introducción a diferentes métodos de agrupamiento, como lo son el aglomerativo, divisivo, DSBCAN y espectral; así como ciertas estadísticas para evaluar el rendimiento de los algoritmos antes mencionados. Esto con la finalidad de aplicarlos en una muestra de 300 canciones de uso libre, bajo ciertas restricciones, y así obtener resultados importantes sobre una efectiva segregación de este tipo de información. Para determinar si existe alguna evidencia estadística que justifique la búsqueda de grupos en este tipo de información, se hace uso de la estadística de Hopkins. Además, se presentan distintos temas importantes del análisis musical, como la digitalización de señales continuas y la teoría pertinente relacionada a las transformaciones de Fourier, sus variantes, creación del espectrograma, los denominados coeficientes cepstrales de Mel (MFCC) y otras características de resumen auditiva como lo son: el centroide y la dispersión espectral, esto con la finalidad de resumir las canciones y utilizar dicha información transformada en los algoritmos de agrupamiento antes mencionados. Los resultados muestran que el uso del método espectral, considerando los MFCC, crean conglomerados con diversas características auditivas que hacen discernibles a los grupos creados; además, el índice CVNN es adecuado en este tipo de información.

1.2. Introducción

Este proyecto comenzó con la finalidad de estudiar diversos métodos de agrupamiento y analizar su comportamiento en grandes cantidades de información añadiendo un enfoque computacional. A medida que se fueron estudiando dichos métodos, el propósito de este trabajo fue perdiendo claridad por la falta de aplicación directa, ya que como cualquier método de aprendizaje supervisado y no supervisado en el aprendizaje de máquina ("Machine learning"), los resultados de una metodología no generalizan el rendimiento y la calidad de resultados de algún algoritmo. Por esta razón, se realizó un cambio radical en el propósito de este trabajo llevando a temas sumamente interesantes y obteniendo conclusiones útiles sobre algunos métodos de agrupamiento en un tipo de información específica: la música.

Actualmente existen muchos sistemas y aplicaciones que han comercializado con este tipo de información y, por razones meramente monetarias, poco se sabe acerca del funcionamiento de estos. Otro problema al tratar de entender estos sistemas es la obtención

de la información, ya que la mayoría del contenido auditivo popular no se puede obtener sin algún tipo de remuneración económica hacia el autor. Por estas razones y como una oportunidad interesante para aplicar algunos métodos del área del aprendizaje no supervisado, se decidió estudiar si solo con la información proporcionada por la canción, sin algún otra característica externa, que bien podría aportar información relevante para segregar la información, al menos en una muestra tomada de manera aleatoria de un conjunto grande de canciones, algún método de agrupamiento era capaz de proporcionar alguna segregación significativa en la música.

Es importante mencionar que la base de datos creada para este trabajo (véase más detalles en la sección 4.1), no otorgaba más características fiables más que la propia canción en formato mp3, por lo que cada canción utilizada, descritas en el apéndice A, fue normalizada, tratada de manera individual y resumida en diferentes objetos. Así, con solo algunas transformaciones de la información que contienen los archivos de audio, fueron realizados los métodos de agrupamiento considerados.

Respecto a la estructura de este trabajo, en el capítulo 2 se presenta un breve repaso de ciertos métodos de agrupación, como lo son los métodos tradicionales (aglomerativos y divisivos), métodos basados en densidades, en este trabajo solo se considera el algoritmo Density-Based Spatial Clustering of Applications with Noise (DBSCAN), y el método espectral. Además se agrega información sobre distintas métricas, ventajas y desventajas de estos algoritmos y ciertos aspectos importantes a tener en consideración cuando se realizan este tipo de métodos, como lo son diversas métricas de rendimiento y una prueba estadística para determinar si tiene sentido o no realizar este tipo de algoritmos.

En el capítulo 3 se presenta una introducción de diversos temas relacionados con el análisis de señales auditivas. En este apartado se muestra el conocimiento necesario sobre la física de las señales de audio y su digitalización para ser estudiadas en un ordenador; se da un análisis desde la transformada de Fourier, así como sus variantes para el caso discreto y su implementación, hasta la obtención de los espectrogramas y como estos son modificados para plasmar la información que es comprendida de manera psicológica. Al final, se agregan otras formas de resumir información auditiva que son de uso común en un análisis de audio.

Respecto a los últimos capítulos, en el capítulo 4 se muestran los resultados que fueron obtenidos con toda la teoría aplicada a la muestra de canciones que fue considerada para terminar en el capítulo 5 con las conclusiones finales. En caso de querer tener una mejor concepción de los resultados descritos en estos últimos capítulos, se puede consultar algunas de las canciones[1], ya separadas en los grupos que se obtienen con el mejor resultado[2], en el repositorio referenciado al siguiente enlace: https://github.com/CarlosFernandoVG/CancionesTesisCFVG. Aquí mismo se pueden encontrar objetos de tipo Tensor (como son definidos por PyTorch) respectivos a los MFCC de cada una de las canciones alojadas en dicho repositorio.

[1] A la fecha de término de este trabajo, diferentes canciones fueron retiradas de la plataforma, por esto y por motivos de almacenamiento, las canciones almacenadas en el repositorio (separadas por grupos) serán representativas de los grupos (algunos de ellos por la poca cantidad de elementos estarán completos). Todas las canciones de las cuales se presenta su espectrograma en el capítulo 4, se encuentran presentes en el repositorio.

[2] El mejor resultado es el obtenido por el método espectral, considerando los coeficientes cepstrales de frecuencia de Mel y utilizando como hiperparámetros 7 grupos y 144 vecinos.

Capítulo 2

Creación y evaluación de conglomerados

En la búsqueda de conglomerados será necesario conocer y aplicar algún tipo de función que determine que tan cercana o alejada está una observación con respecto a otra o a una partición de los datos a la cual no pertenece dicha observación. Dicha actividad será esencial para la creación de los conglomerados; de hecho, determinar una medida de disimilaridad apropiada es incluso más importante para obtener éxito al proponer conglomerados que la propia elección del algoritmo [20]. Se dará una introducción a este tema en la primera parte de este capítulo.

Una vez establecidas diferentes maneras de calcular la similitud entre observaciones, se procederá al estudio de diferentes técnicas para la obtención de conglomerados en un conjunto de datos. En este caso solo se dará una introducción a los métodos jerárquicos y partitivos, los cuales serán considerados como métodos tradicionales, métodos basados en densidades, se dará especial atención al método DBSCAN, y el método espectral.

Finalmente, se hará una breve introducción de algunas estadísticas para la evaluación de los resultados obtenidos por un método de agrupamiento de manera interna. Antes de eso, se describirá la estadística de Hopkins, la cual nos ayudará, de manera estadística, a determinar si es factible o no utilizar alguno de estos algoritmos sobre nuestra información.

2.1. Medidas de similaridad y disimilaridad

El objetivo de crear conglomerados es obtener una partición finita de grupos entre las observaciones que compartan características de tal manera que dentro de cada grupo se tengan observaciones lo más similares posibles y lo más distintas posibles fuera de los grupos, para dicho fin es necesario tener un conjunto de funciones matemáticas que determinen que tan alejadas o separadas están las observaciones en el espacio donde existan. Estas funciones varían de acuerdo al tipo de dato que se esté tratando y de las características que el investigador o usuario deseé aplicar, las cuales pueden obtenerse por la experiencia o el mismo comportamiento de los datos.

De acuerdo al método de agrupación y las características de este, la función que cuantificará lo cercano que dos observaciones se encuentran en el espacio donde existen se llamará función de **similaridad** y para considerar lo distintas que son dichas observaciones se tomará una función de **disimilaridad**.

***Definición* 2.1.1:** Sean x y y dos elementos de un conjunto $E \neq \emptyset$. Una **función de similaridad** $s : E \times E \to \mathbb{R}$ es aquella que satisface las siguientes propiedades para todo elemento en E:

1. $0 \leq s(x, y) \leq 1$
2. $s(x, x) = 1$. Si $s(x, y) = 1 \Leftrightarrow x = y$
3. $s(x, y) = s(y, x)$.

Es decir, una función de similaridad es aquella que tiene su soporte en $[0, 1]$ y cumple las propiedades de simetría y conmutabilidad. Por otro lado, una función de disimilaridad generalmente es una métrica, la cual queda definida de la siguiente manera:

***Definición* 2.1.2:** Sean x, y y z tres elementos de un conjunto $E \neq \emptyset$. Una **métrica** $d : E \times E \to \mathbb{R}$ es una función que satisface las siguientes propiedades para todo elemento de E:

1. $d(x, y) \geq 0$ y $d(x, y) = 0 \Leftrightarrow x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$.

Véase que el rango de una función de similaridad es limitado comparado al de una función de disimilaridad. Además, cuando esta sea una métrica, se cumplirá la desigualdad del triángulo, lo cual añade una percepción natural de distancia en el plano euclidiano. Para algunas funciones de similaridad es posible demostrar que alguna función definida como $d(\cdot, \cdot) = 1 - s(\cdot, \cdot)$ es una métrica, por lo que es sencillo crear una función de disimilaridad con una función de similaridad que cumpla la desigualdad triangular.

Para algunas técnicas de agrupamiento, se utiliza una matriz donde cada uno de sus elementos representa la distancia entre dos puntos o la similitud entre ellos. Suponga que se tienen n elementos de E. Para el caso de la matriz de distancias $\mathcal{M}$, $d_{i,j}$ es la distancia entre los elementos i y j; dicha matriz es simétrica cuando la función correspondiente es simétrica. La matriz de similaridad es análoga salvo cuando se compara el mismo elemento, en dicho caso se tiene que $s_{i,i} = 1$:

$$\mathcal{M} = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,n} \\ d_{1,2} & 0 & \cdots & d_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n,1} & d_{n,2} & \cdots & 0 \end{pmatrix}.$$

Existen una gran variedad de funciones de disimilaridad y similaridad y son de gran importancia en la construcción de un conglomerado. Un ejemplo de éstas es la distancia de **Minkowski**, la cual generaliza algunas métricas.

***Definición* 2.1.3:** Sean dos vectores en $\mathbb{R}^d$; $X_i = (x_{i1}, x_{i2}, \ldots, x_{id}), X_j = (x_{j1}, x_{j2}, \ldots, x_{jd})$. Se define la **distancia de Minkowski de orden** p entre los anteriores dos puntos como

$$d_p(X_i, X_j) = \left(\sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p \right)^{\frac{1}{p}} = \|X_i - X_j\|_p \text{ con } p \geq 1.$$

La segunda igualdad hace mención de la norma de grado p entre los dos elementos de $\mathbb{R}^d$, ya que la norma es la longitud de un vector y $\mathbb{R}^d$ es un espacio vectorial, lo cual otorga una mejor interpretación a la distancia de Minkowski (la longitud del vector que se obtiene de la diferencia entre los vectores X_i y X_j). Es común encontrarse con esta notación en la extensa bibliografía sobre medidas de disimilaridad; aunque, es importante mencionar que, si bien es cierto que todo espacio normado[1] induce una métrica, una métrica está definida sobre un conjunto que no necesariamente debe ser un espacio vectorial.

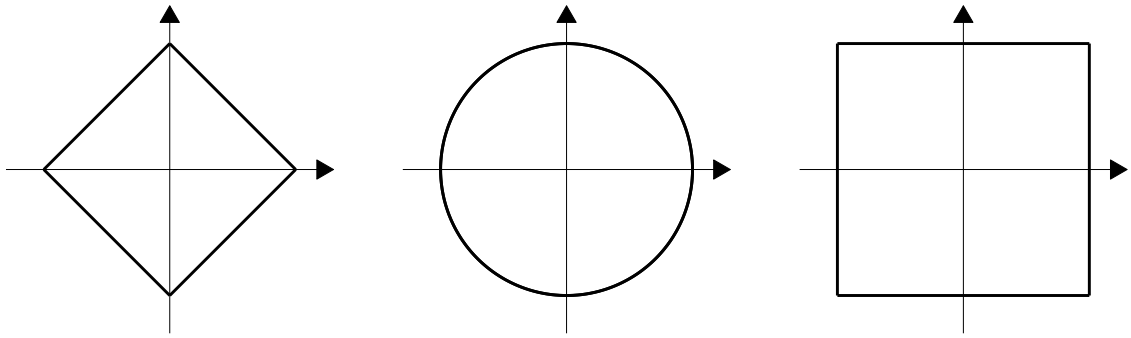


Figura 2.1: Circunferencias tomando como distancia algunas métricas de Minkowski.

La distancia de Minkowski es una de las más populares y más en su caso particular cuando $p = 2$, en tal caso se obtiene la **distancia euclidiana** o la distancia estándar entre dos vectores en $\mathbb{R}^d$, también conocida como métrica pitagórica ya que, con el uso iterativo del teorema de Pitágoras, se puede obtener dicha distancia en $\mathbb{R}^d$. Esta métrica es útil cuando se desea obtener la distancia absoluta entre las observaciones, aunque no considera redundancias que puedan existir entre las variables, lo cual podría ser lo deseado:

$$d_2(X_i, X_j) = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} = \sqrt{(X_i - X_j)(X_i - X_j)'}.$$

Cuando $p = 1$ se obtiene la **distancia Manhattan** o también conocida como **City block distance**. Dichos nombres tienen un contexto histórico con el diseño de las calles de la ciudad de Manhattan, ya que el objetivo fue que el camino más corto en la mayoría de las intersecciones entre calles de dicha ciudad se obtuviera desde cualquier ruta, esto mediante segmentos verticales y horizontales. Por la fórmula se puede interpretar que esta distancia no es más que la suma de las proyecciones de los segmentos de recta entre las coordenadas de los puntos X_i y X_j:

$$d_1(X_i, X_j) = \sum_{k=1}^{d} |x_{ik} - x_{jk}|.$$

[1] Un espacio normado es un espacio vectorial E sobre $\mathbb{K}$ un campo en el cual se puede definir una norma, la cual es una función $\| \ \| : E \to [0, \infty)$ la cual cumple que si $v \in E$; $\|v\| = 0 \Leftrightarrow v = 0$, $\|av\| = |a|\|v\|$ donde $a \in \mathbb{K}$ y que cumple la desigualdad del triángulo.

Finalmente, cuando $p = \infty$ se obtiene la **distancia de Chebyshev** o métrica máxima, la cual no es más que la distancia máxima entre todas las que se pueden obtener entre las coordenadas de dos puntos respetando la dimensión que estas representan:

$$d_\infty(X_i, X_j) = \lim_{p\to\infty} \left(\sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p \right)^{\frac{1}{p}} = \max_{1\leq k\leq d} |x_{i,k} - x_{j,k}|.$$

Una manera clásica de representar cada una de estas métricas es con el conjunto de todos los puntos con los cuales se obtendría que la distancia de Minkowski desde el centro a dichos puntos es la unidad; es decir, las circunferencias que se obtienen considerando la métrica de Minkowski variando el parámetro p, lo cual queda plasmado en la gráfica 2.1. En dicha gráfica se puede ver que a medida que esta incrementa en su parámetro p, disminuye en su valor e independientemente del valor de p, la métrica de Minkowski tiende a dar preferencia a los valores más grandes, en valor absoluto, en la variable que se esté considerando, lo cual puede verse en cada uno de los ejes (variables) de las gráficas.

Una de las prácticas más comunes es otorgar un escalamiento a los datos para que estos no perjudiquen el cálculo de la distancia por tener diferentes unidades de medición; por ejemplo utilizando el rango de las variables cuantitativas, el inverso de la desviación estándar o restando o no la media de dichas variables. Con esto se da una menor importancia a las variables con mayor variabilidad, lamentablemente esto puede ser perjudicial en la creación de agrupamientos. Tómese como ejemplo lo visto en la gráfica 2.3 donde al normalizar las observaciones se tiene la misma subgráfica del panel izquierdo, pero al aumentar los datos, y por lo tanto la variabilidad, como en la gráfica 2.2 (a) y posteriormente normalizar estos, como en (b), se podrían dar malos agrupamientos.

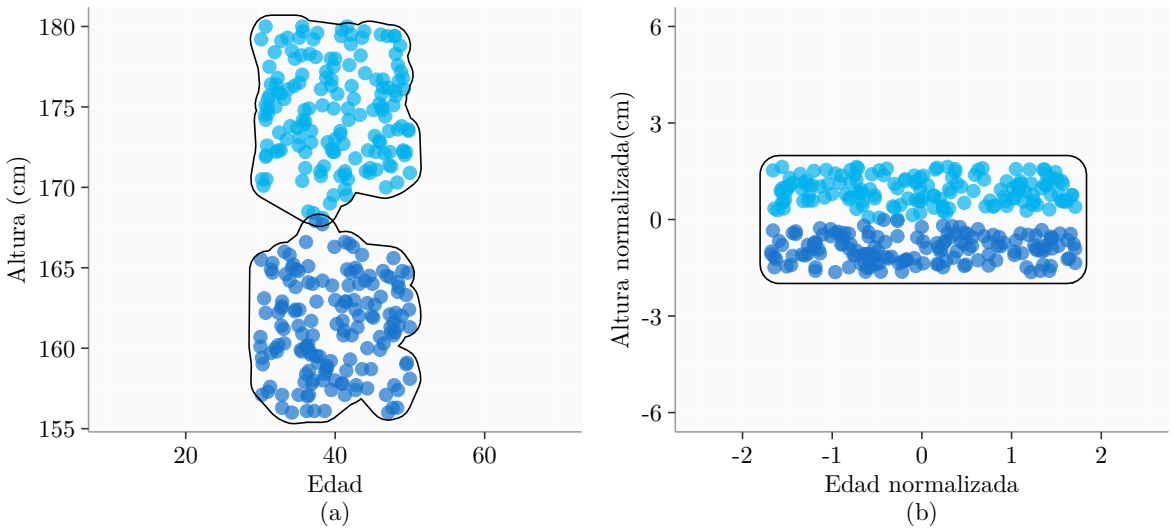


Figura 2.2: Agrupamientos en datos simulados y el efecto de estos en datos normalizados.

Para evitar este problema en la creación de grupos, se puede usar algún método que sea invariante a las escalas cuando no se conocen las unidades adecuadas de las variables al igual que en las distancias [18].

Es común utilizar las métricas anteriores, ya que tienen una buena interpretación, aunque hay que considerar que éstas trabajan adecuadamente con conglomerados compactos y aislados (es decir que entre los grupos las métricas son pequeñas y entre dos miembros

de dos grupos distintos se tienen distancias grandes) lo cual podría no suceder. Al darle una mayor importancia a las observaciones con mayor valor en la variable que se esté comparando, pueden surgir problemas al crear los grupos. Un ejemplo claro de esto puede verse en la figura 2.3 donde, mediante datos simulados, al cambiar la escala de centímetros a pies, en la variable *Altura* se tienen grupos distintos. Esto podría evitarse dando una misma escala a los datos (normalización).

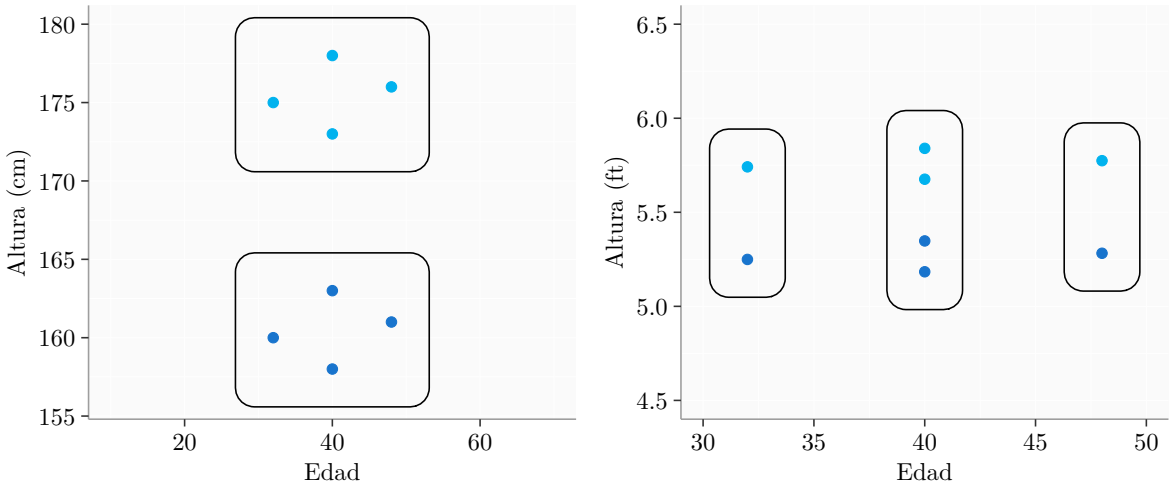


Figura 2.3: Conglomerados en un conjunto de datos simulados de individuos donde se representa su edad y altura. Véase que al cambiar la escala en una variable se obtienen conglomerados diferentes.

Existen otras funciones derivadas de las anteriores que tratan diversas deficiencias; por ejemplo, la distancia Canberra[2] da mayor prioridad a los valores más alejados o a los rangos más grandes que se pueden formar con dichos valores, además, dicha distancia es sensible a datos cercanos a 0. Un problema al utilizar la distancia euclidiana se presenta cuando dos observaciones no presentan información en común en sus variables, lo que ocasionaría que dicha distancia podría ser más pequeña que las dos observaciones cercanas.

La métrica de similaridad **Coseno** da una alternativa al anterior problema, ya que en ésta se considera el ángulo entre los vectores para determinar lo cercano que son sin considerar su tamaño, lo cual haría que en la distancia euclideana se tengan distancias muy grandes. La distancia coseno queda expresada como $1 - s_{cos}(X_i, X_j)$ donde

$$s_{cos}(X_i, X_j) = \frac{X_i \cdot X_j}{\|X_i\|_2 \|X_j\|_2}.$$

El coseno ha sido utilizado en otras distancias como la distancia Chord [3] la cual determina la longitud de la cuerda que se forma por dos observaciones estandarizadas en una hiperesfera de radio uno o en lugar de considerar la cuerda que se forma en dicha hiperesfera

[2]La distancia Canberra queda determinada como $d_{canb}(X_i, X_j) = \sum_{k=1}^{d} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$ donde es necesario que al menos $x_{ik} \neq 0$ o $x_{ik} \neq 0$.

[3]László Orlóci en 1967 [48] propuso esta distancia que tiene siguiente expresión: $d_{chord}(X_i, X_j) = \sqrt{2 - 2\frac{\sum_{k=1}^{d} x_{ik}x_{jk}}{\|X_i\|_2\|X_j\|_2}} = \sqrt{2(1 - cos(\theta))}$. En la segunda igualdad se está considerando que el $cos(\theta)$ entre los vectores que se forman con las observaciones X_i y X_j se puede determinar con el uso del producto punto.

se puede tomar la longitud de arco que se forma entre dichas observaciones estandarizadas, como lo hace la métrica geodésica[4] [34]. Finalmente, la distancia de Bhattacharyya también utiliza el ángulo entre los vectores que representan a las observaciones [33][5].

De acuerdo al tipo de información con el que se esté trabajando, se pueden tener no solo variables numéricas, también categóricas o de manera mixta. En este trabajo no se tiene dicho caso pero se recomienda tener en cuenta este importante factor y se recomienda revisar el **coeficiente Gower**[6] [23].

Varias funciones de disimilaridad se han propuesto en muchos ámbitos; como en la genética, biología, antropología, etc. Y muchas de éstas se derivan como funciones de otra función de similaridad; para ver más de éstas y un análisis más profundo para las funciones de disimilitud y similitud véase [21], [28] y [34]. En nuestro caso, es necesario hablar de la distancia Kullback–Leibler[7].

En el área de la teoría de la información[8], se cuenta con una medida de distancia entre dos distribuciones de probabilidad conocida como entropía relativa[9] que es una medida de la deficiencia de suponer que la distribución es $\mathcal{Q}$ cuando la distribución verdadera es $\mathcal{P}$ [15]; es decir, es una estimación de qué tan bien puede ser representada la distribución $\mathcal{P}$ con la distribución $\mathcal{Q}$. Esta surge del logaritmo esperado de la razón de verosimilitudes.

***Definición* 2.1.4:** La entropía relativa o distancia de Kullback-Leibler entre dos funciones de masa de probabilidad $\mathcal{P}(X)$ y $\mathcal{Q}(X)$ está definida como

$$KL(\mathcal{P}||\mathcal{Q}) = \sum_{x\in X} \mathcal{P}(X) log \frac{\mathcal{P}(X)}{\mathcal{Q}(X)} = \mathbb{E}_{\mathcal{P}} \log \frac{\mathcal{P}(X)}{\mathcal{Q}(X)}.$$

Por convención, se establece que $0 \log \frac{0}{0} = 0$ y, basado en argumentos de continuidad, $0 \log \frac{0}{\mathcal{Q}} = 0$ y $\mathcal{P} \log \frac{\mathcal{P}}{0} = \infty$. Esta distancia siempre es no negativa y es cero sí y solo sí $\mathcal{P} = \mathcal{Q}$. Esta distancia no es una métrica ni una verdadera distancia entre distribuciones

[4] $d_{geo}(X_i, X_j) = arc\ cos\left(1 - \frac{d_{chord}(X_i,X_j)}{2}\right)$.

[5] Originalmente Bhattacharyya propuso esta función como una medida de divergencia entre distribuciones de probabilidad considerando que las dos poblaciones se distribuían de manera multinomial en k clases con p_i y $p\prime_i$ probabilidades respectivamente. Dado lo anterior, su función se interpreta como el coseno del ángulo entre los vectores $(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k})$, $(\sqrt{p\prime_1}, \sqrt{p\prime_2}, \ldots, \sqrt{p\prime_k})$ [3], es decir $cos(\theta) = \sum_{i=1}^{k} \sqrt{p_i p\prime_i}$. Véase que cuando ambas distribuciones son idénticas $cos(\theta) = 1$.

[6] $S_{Gower}(X_i, X_j) = \frac{\sum_{k=1}^{d} S(x_{ik},x_{jk})\delta(x_{ik},x_{jk})}{\sum_{k=1}^{d} \delta(x_{ik},x_{jk})}$, donde la función delta de Kronecker tomará el valor de 1 cuando se pueda hacer una comparación (cuando ninguna de las características sean valores perdidos) y 0 en el caso contrario. Los valores que tomará $S(x_{ik}, x_{jk})$ se obtendrán de acuerdo al tipo de variable que se esté tratando. Para más detalles consúltese [23].

[7] También conocida como divergencia Kullback–Leibler.

[8] Esta fue desarrollada por Claude E. Shannon y Warren Weaver a finales de la década de los años 1940. Esta teoría está estrictamente relacionada con la estadística y probabilidad enfocadas a modelar la transmisión, medida y el procesamiento de la información.

[9] La entropía de una variable aleatoria X puede ser interpretada como el valor esperado de la variable aleatoria $\log \frac{1}{\mathbb{P}(X)}$ y esta, conceptualmente, se puede interpretar como la cantidad de información que contiene una señal.

ya que no satisface la desigualdad del triangulo ni es simétrica, pero es común utilizar la siguiente expresión como una función de disimilaridad

$$D_{KL}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2}\left(KL(\mathcal{P}||\mathcal{Q}) + KL(\mathcal{Q}||\mathcal{P})\right).$$

Por ejemplo, si consideramos que $\mathcal{P}$ y $\mathcal{Q}$ son modelos gaussianos con vectores de medias $\mu_{\mathcal{P}}$, $\mu_{\mathcal{Q}}$ en $\mathbb{R}^d$ y matrices de varianzas y covarianzas $\Sigma_{\mathcal{P}}$, $\Sigma_{\mathcal{Q}}$. Si $|\Sigma_{\mathcal{P}}|$ denota el determinante de $\Sigma_{\mathcal{P}}$ y $Tr(\cdot)$ la traza, la distancia entre $\mathcal{P}$ y $\mathcal{Q}$ queda representada por la siguiente expresión

$$KL(\mathcal{P}||\mathcal{Q}) = \frac{1}{2}\left[\log\frac{|\Sigma_{\mathcal{P}}|}{|\Sigma_{\mathcal{Q}}|} + Tr(\Sigma_{\mathcal{P}}^{-1}\Sigma_{\mathcal{Q}}) + (\mu_{\mathcal{Q}} - \mu_{\mathcal{P}})^T\Sigma_{\mathcal{P}}^{-1}(\mu_{\mathcal{Q}} - \mu_{\mathcal{P}}) - d\right].$$

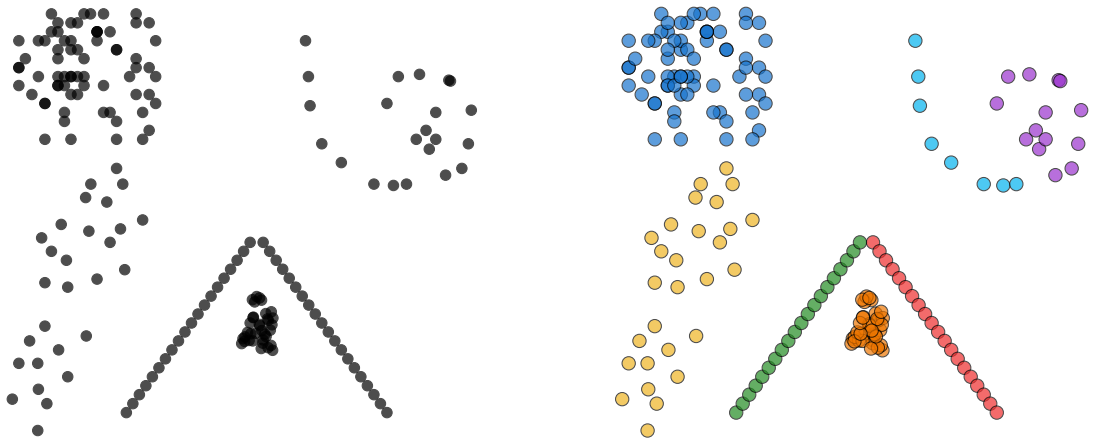
$D_{KL}(\mathcal{P}, \mathcal{Q})$ nos será de utilidad ya que utilizaremos ésta función para determinar la distancia entre canciones; más adelante se explicará como será resumida dicha información.

2.2. Técnicas de Agrupación

2.2.1. Métodos tradicionales de agrupación

Parte del análisis de datos se puede clasificar en la tarea de confirmar información o en un análisis exploratorio, el cual podemos hacer mediante distintas técnicas, ya sea con la evaluación de una estadística de resumen para una prueba de bondad y ajuste sobre algún modelo o encontrando agrupamientos naturales (***clústers***) [27].

Al final de cualquier método que busque patrones sobre los datos, se espera que dentro de cada agrupación, el contenido sea similar entre el mismo y sea, a la vez, distinto de los otros agrupamientos; como se puede hacer evidente en la figura 2.4 donde los datos se encuentran en el espacio de $\mathbb{R}^2$ (a) y los grupos deseados son discernidos con distintos colores (b).



(a) (b)

Figura 2.4: Agrupamiento de datos.

Otro punto importante sobre este tema, es que en muchos de los ámbitos donde se tenga información y se desee descubrir o validar resultados, se busca tener la menor cantidad de supuestos sobre esta información. Las técnicas de agrupación o **clusterización** nos pueden ayudar al encontrar interrelaciones en su estructura, lo cual puede realizarse de manera preliminar.

Los métodos clásicos son basados en el uso de distancias sobre vectores. Hay que notar que, a diferencia del aprendizaje supervisado, la cantidad de grupos puede variar de acuerdo a diferentes criterios que determinen la eficacia del método que se está ejecutando y que este número de grupos siempre está supeditado a los propios datos.

De acuerdo a la fuente que se consulte, todos los métodos de clusterización pueden clasificarse en **Jerárquicos** y **Partitivos** [27], pero en este trabajo se considerará a los anteriores métodos como clásicos por el hecho de utilizar una métrica para segregar los datos y utilizar un algoritmo relativamente sencillo; más adelante se verán ejemplos de otro tipo de clusterización que conllevan una mayor complejidad en su ejecución.

Una de las principales diferencias entre los anteriores algoritmos es que los métodos partitivos dividen un conjunto de datos dentro de particiones aisladas y los métodos jerárquicos en particiones anidadas. Además, en un método partitivo generalmente se requieren una cantidad de parámetros iniciales; como la cantidad de clústeres a encontrar y un conjunto inicial de observaciones (puntos prototipos o centroides) con el cual, iterativamente, comenzará a separar todos los datos en grupos. El ejemplo más común de este tipo de algoritmos es el K-Medias o **K-Means** [2].

Algoritmo 1 K-means Clustering

1: Selecciona K observaciones como centroides iniciales.
2: **repetir**
3: Crear K clústeres asignando a cada observación a su centroide más cercano.
4: Recalcular los centroides de cada clúster.
5: **hasta** que el criterio de convergencia es alcanzado.

Los métodos jerárquicos, a diferencia de los métodos partitivos, tienen la ventaja de no requerir una cantidad fija de conglomerados a crear para iniciar el algoritmo. Estos se dividen en dos: **aglomerativos** y **divisivos**. El primero de estos inicia tomando un solo elemento como miembro de un clúster, por lo que empieza creando n grupos (suponiendo que se cuentan con n observaciones) los cuales, en cada iteración, se irán fusionando hasta formar un solo clúster. El segundo trabajará de manera inversa, primero creará un conglomerado con todos los datos e irá dividiendolo hasta crear un grupo por cada observación.

La figura 2.5 es un ejemplo de un **dendrograma**[10], el cual es una representación gráfica donde se muestran como fueron formados los conglomerados. Esta es una de las ventajas de utilizar un método jerárquico aunque, para decidir la mejor partición de los datos, se debe decidir, mediante distintos criterios, a qué nivel se desea realizar un corte en el dendograma, lo cual representará nuestra configuración final de los grupos para los datos. Por la manera en que se construyen los conglomerados, se dice que los métodos

[10]Un dendograma es un n-árbol en el cual cada nodo interno es asociado con una altura que satisfaga la condición $h(A) \leq h(B) \Leftrightarrow A \subseteq B$ para todos los nodos.

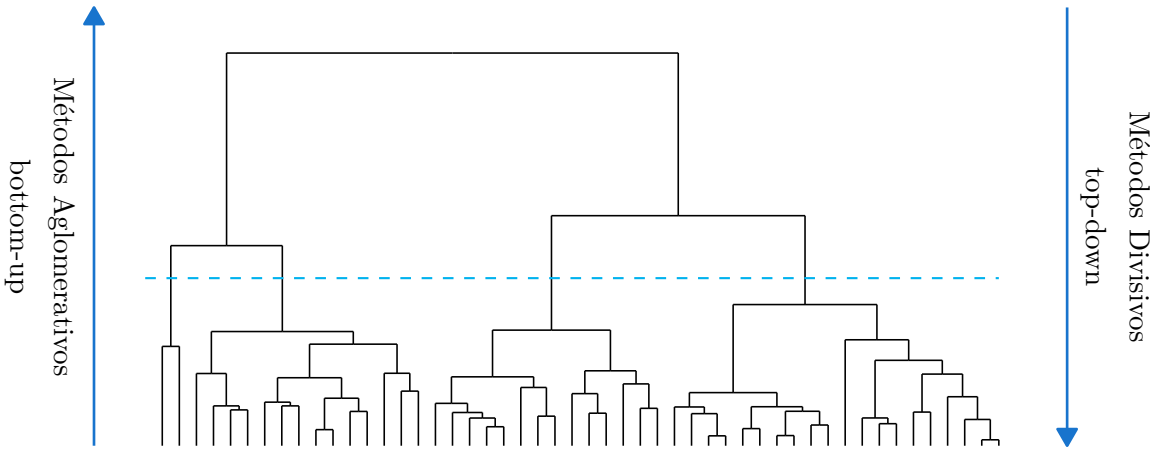


Figura 2.5: Ejemplo de un dendrograma indicando como se realizan los distintos métodos jerárquicos. La línea punteada muestra un posible corte en el dendograma haciendo que se consigan 4 agrupaciones para los datos.

aglomerativos siguen una jerarquía de abajo hacia arriba (*bottom-up*) y los divisivos una jerarquía de arriba hacia abajo (*top-down*).

Los métodos aglomerativos pueden ser clasificados en métodos geométricos y basados en gráficas, ya que diferentes métodos pueden ser representados por una subgráfica o por puntos inter conectados y en los métodos geométricos un grupo puede ser representado por un punto central [21]. En los métodos basados en gráficas se encuentran los métodos *single-link*, *complete-link* y *average*, y en los métodos geométricos se encuentra el método ward, los cuales serán explicados a continuación.

Algoritmo 2 Cluster Jerárquico Aglomerativo

1: Calcular la matriz de disimilaridad entre todos las observaciones.
2: **repetir**
3: Fusionar los grupos como $C_{a\cup b} = C_a \cup C_b$. Establecer la cardinalidad del nuevo grupo como $N_{a\cup b} = N_a + N_b$.
4: Insertar un nuevo renglón y columna que contengan las distancias entre el nuevo grupo $C_{a\cup b}$ y los restantes grupos.
5: **hasta** que solo un grupo maximal permanezca.

2.2.1.1. Métodos de enlace único y completo

El método de enlace único (***single-link***), es también conocido como el método del vecino más cercano o método mínimo ya que este considera lo cercano que son dos grupos de acuerdo a la distancia del vecino más cercano. Para aclarar esto, si $x_1 \in C_1$ y $x_2 \in C_2$, se dirá que x_1 y x_2 son **vecinos**. La **distancia del vecino más cercano**, respecto a $d(\cdot,\cdot)$, queda determinada por la ecuación (2.1). Cuando se sustituye el mínimo por el máximo se tiene la **distancia del vecino más lejano** (d_{fn}), tal como se representa en la gráfica 2.6 para un conjunto de puntos simulados.

$$d_{nn}(C_1, C_2) = \min_{i\in I_{C_1}; j\in I_{C_2}} d(x_i, x_j); \qquad I_C = \{i | x_i \in C\} \tag{2.1}$$

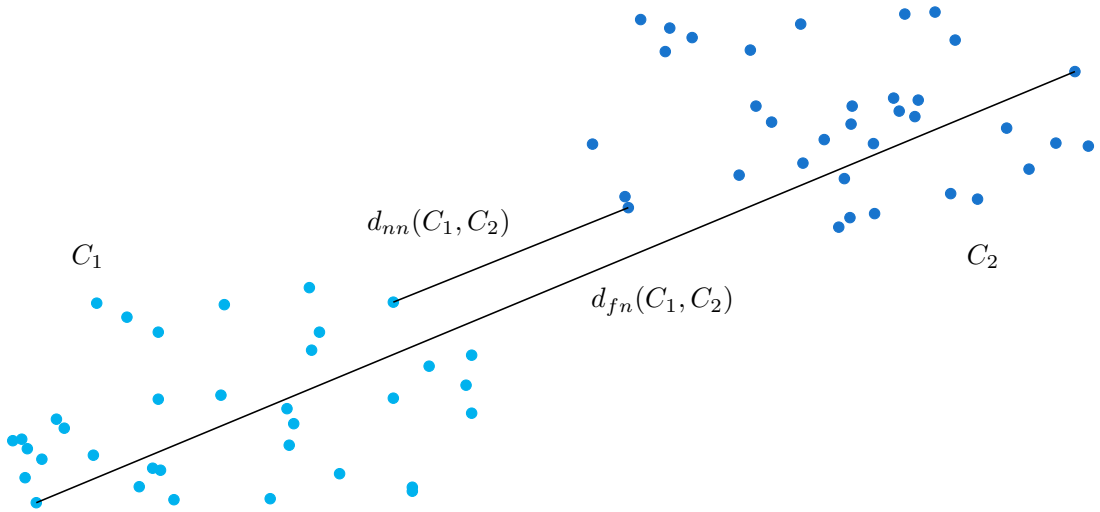


Figura 2.6: Representación del vecino más cercano y más lejano en dos grupos de datos simulados.

Dicho método da una mayor importancia a los grupos más cercanos, pero no considera la estructura de estos; es decir, que tiene un buen tratamiento para los grupos que no tienen una forma elíptica o con una forma alargada, pero no tiene un buen comportamiento cuando se tienen datos atípiticos o ruido en los datos.

El método de enlace completo (***complete-link***) considera el caso contrario, ya que en este se considera la distancia del vecino más lejano para discernir a los grupos. Este considera la estructura de los grupos y generalmente se obtienen grupos con una forma compacta, al igual que en método de enlace único, este método es sensible a datos atípicos [2].

Solo por considerar un ejemplo, véase la gráfica 2.7 donde se muestra un grupo de 5 puntos en $\mathbb{R}^2$ y el respectivo dendrograma utilizando el método single-link con la métrica euclideana. Con base a dichas observaciones, el siguiente grupo de matrices muestran la aplicación del método mencionado:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0 & 7.21 & 5 & 4.12 & 2.23 \\ 7.21 & 0 & 4.12 & 3.60 & 7.81 \\ 5 & 4.12 & 0 & 1.41 & 4.47 \\ 4.12 & 3.60 & 1.41 & 0 & 4.24 \\ 2.23 & 7.81 & 4.47 & 4.24 & 0 \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \\ x_1 \\ x_2 \\ \{x_3, x_4\} \\ x_5 \\ . \end{array}\!\!\begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \{x_3, x_4\} & x_5 \end{array} \\ \begin{bmatrix} 0 & 7.21 & 4.12 & 2.23 \\ 7.21 & 0 & 3.60 & 7.81 \\ 4.12 & 3.60 & 0 & 4.24 \\ 2.23 & 7.81 & 4.24 & 0 \end{bmatrix} \end{array}$$

$$\longrightarrow \begin{array}{c} \\ \{x_1, x_5\} \\ x_2 \\ \{x_3, x_4\} \end{array}\!\!\begin{array}{c} \begin{array}{ccc} \{x_1, x_5\} & x_2 & \{x_3, x_4\} \end{array} \\ \begin{bmatrix} 0 & 7.211103 & 4.123106 \\ 7.211103 & 0 & 3.605551 \\ 4.123106 & 3.605551 & 0 \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \\ \{x_1, x_5\} \\ \{x_2, x_3, x_4\} \end{array}\!\!\begin{array}{c} \begin{array}{cc} \{x_1, x_5\} & \{x_2, x_3, x_4\} \end{array} \\ \begin{bmatrix} 0 & 4.12 \\ 4.12 & 0 \end{bmatrix} \end{array}$$

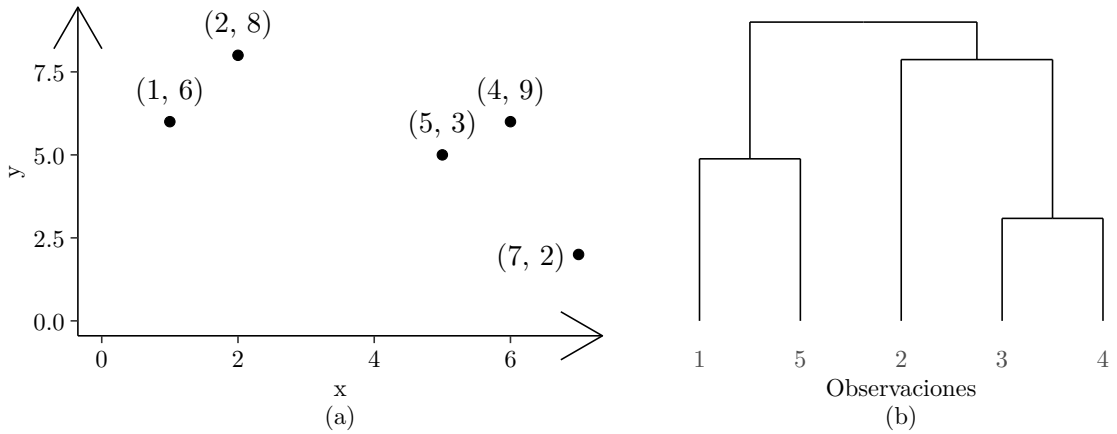


Figura 2.7: Ejemplo de la aplicación de un algoritmo aglomerativo con el método complete-link. En la primera sub gráfica se aprecia la configuración de las observaciones y en la segunda el dendrograma resultante.

2.2.1.2. Métodos del grupo promedio y criterio de Ward

Así como se considera el máximo o el mínimo, se puede tomar el promedio de las distancias entre los distintos vecinos, esto queda expresado en la siguiente distancia suponiendo que $|C_1| = n_1$ y $|C_2| = n_2$.

$$d_{ave}(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d(x_i, x_j)$$

En el método del grupo promedio (***Group Averaged Agglomerative Clustering***), la distancia entre dos grupos se calcula en base a la anterior expresión, con esto evitamos las debilidades que tenían los métodos *single* y *complete*. Una de las desventajas de este método es lo costoso de su cálculo en comparación de los otros métodos, ya que este debe calcular las distancias promedio por pares de todos los elementos en cada iteración.

Otra técnica propuesta por Ward Jr. [53] busca crear las particiones de los datos de tal manera que minice la suma de los errores cuadráticos, es decir, en cada iteración se busca minimizar el aumento en la suma del error cuadratico (SSE) total dentro del clúster [18]. Podemos interpretar lo anterior como una técnica que busca minimizar la pérdida de información asociada en cada fusión y que, gracias al uso de la ecuación (2.2), este método es comúnmente llamado método de "mínima varianza" [21].

$$SSE = \sum_{x \in C} (x - \mu(C))(x - \mu(C))^T; \quad \mu(C) = \frac{1}{|C|} \sum_{x \in C} x. \tag{2.2}$$

Véase que gracias a la ecuación (2.2), donde C es un grupo de datos, el criterio de Ward puede ser interpretado como la distancia euclidiana entre los centros de los conglomerados.

A pesar de las ventajas que tiene realizar un método jerárquico o partitivo, se tienen distintas desventajas al solo optar por este tipo de algoritmos; por ejemplo, el método K-medias tiene un buen comportamiento al encontrar grupos compactos con alguna forma esférica o elíptica, y si no se tiene este tipo estructura, el método puede dar resultados espurios, un ejemplo de esto puede verse en la gráfica 2.8 donde claramente se tienen 3

conglomerados horizontales pero el algoritmo fusionó elementos de diferentes líneas para crear dos grupos de manera errónea.

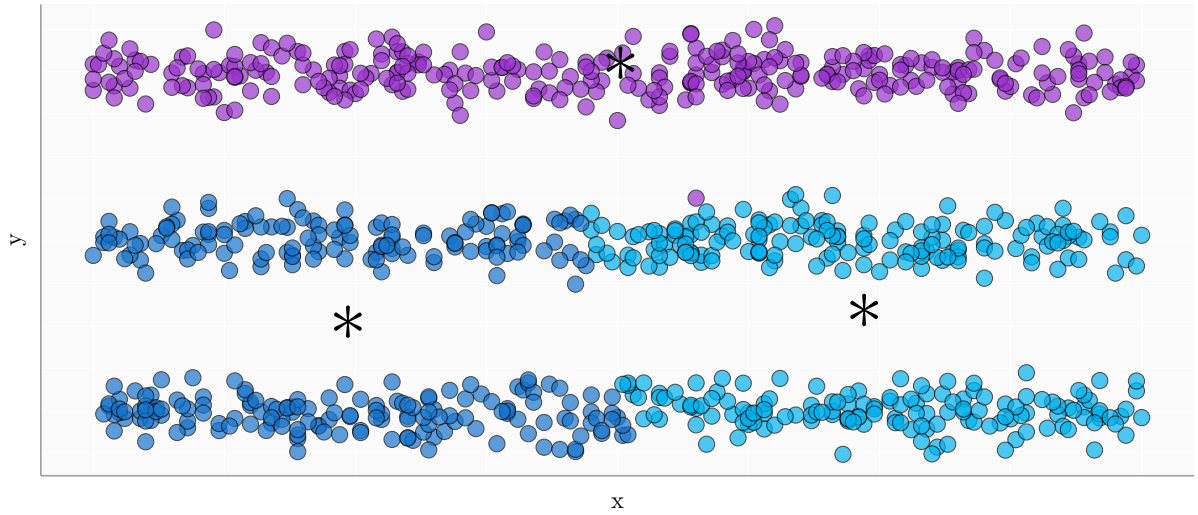


Figura 2.8: Resultado del algoritmo K-medias en un conjunto de puntos simulados. Los elementos en color negro representan los centroides finales que el algoritmo consideró.

Otro inconveniente en este algoritmo es que la mayoría de implementaciones no permiten el uso de distintas métricas[11]. En el caso de los métodos vistos anteriormente, aunque nos proporciona la facilidad de crear una representación gráfica de los resultados, a medida que estos crean grupos más grandes o pequeños, estos ya no se pueden deshacer o modificar, por lo que son poco flexibles con posteriores iteraciones.

2.2.2. Conglomerados basados en densidades

Una alternativa diseñada para la búsqueda de grupos con forma arbitraria es aquella basada en la densidad que contienen ciertas regiones; esta, además, contempla el problema de observaciones extremas y la presencia de ruido. Este paradigma es considerado una metodología no paramétrica, ya que no es necesario tener conocimiento previo sobre el número de grupos; además, el propio algoritmo se encarga de separar el ruido y los *outliers* de los grupos más densos y estos, a su vez, son separados entre si.

Existen diversas propuestas con este paradigma, como lo son los métodos **DBSCAN**, **DENCLUE**, **BRIDGE** y **OPTICS**, siendo el primero de estos el más popular. En este trabajo solo se darán detalles del primer método, pero si se desea obtener un estudio más profundo del resto de métodos, se recomiendan [2], [18] y [21].

[11] El método K-medias puede verse como un problema de optimización que busca minimizar la suma de errores al cuadrado (SSE) del centro con el resto de las observaciones, la cual puede ser interpretada como la distancia euclidiana; esto con el objetivo de segregar la información a un equivalente de un diagrama de Voronoi de los centroides. Existen otro tipo de métodos que se basan solamente en la implementación y que utilizan otras métricas, dejando de lado la esencia del algoritmo y con el riesgo de no tener soluciones que, iterativamente, lleguen a un mínimo local como sí lo hace el método K-means.

2.2.2.1. DBSCAN

El algoritmo **DBSCAN** fue propuesto en 1996 [17]. Este estima la densidad de ciertas regiones mediante un conteo de observaciones en una vecindad de radio fijo ϵ. Para entender dicho algoritmo, será necesario definir ciertos aspectos.

***Definición* 2.2.1:** Sea x un elemento de un conjunto $E \neq \emptyset$. La **ϵ-vecindad de un punto** x queda definida como:

$$N_\epsilon(x) = \{y \in E : d(x, y) \leq \epsilon\},$$

donde $d(\cdot, \cdot)$ es una función de distancia.

Dependiendo de la bibliografía que se consulte, al punto x en la anterior definición será llamado punto interno (***core point***) si la vecindad $N_\epsilon(x)$ contiene al menos N_{min} puntos; es decir, se solicita que la densidad en su vecindad exceda un cierto umbral en su cardinalidad.

Definición* 2.2.2:** Se dice que una observación x es directamente alcanzable por densidad (directly density-reacheable***) desde otra observación y (con respecto a ϵ y N_{min}) si

1. $x \in N_\epsilon(y)$
2. $|N_\epsilon(y)| \geq N_{min}$,

donde $|N_\epsilon(y)|$ es la cardinalidad de $N_\epsilon(y)$ y N_{min} el número de observaciones mínimo.

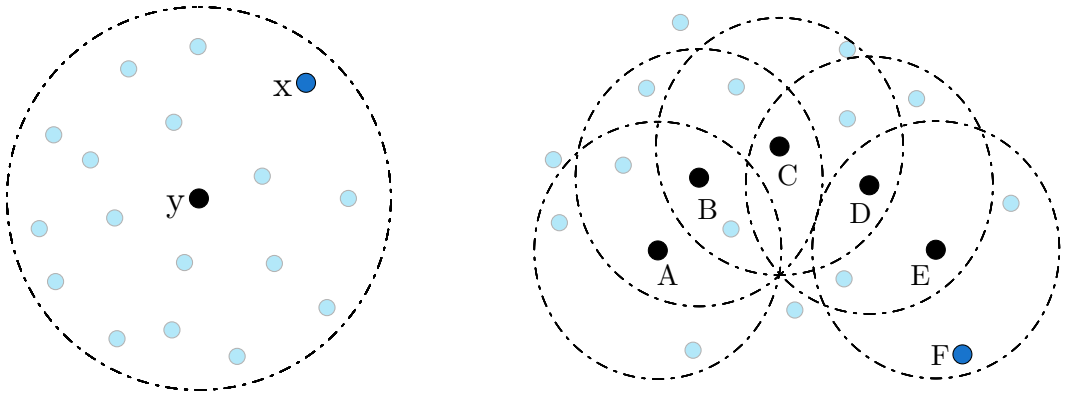
Véase que si x y y son dos puntos internos, tanto x es directamente alcanzable por densidad desde y como este lo es directamente alcanzable por x, por lo que esta relación es simétrica; aunque si y se encontrara en el borde de un grupo, esta relación ya no sería simétrica. Por lo mismo, la siguiente relación, en general, no es simétrica.

Definición* 2.2.3:** Se dice que una observación x es alcanzable por densidad (density-reacheable***) desde otra observación y si existe una secuencia de puntos $x = x_1, x_2, x_3, \ldots, x_i = y$ tal que x_l es directamente alcanzable por densidad desde x_{l+1} para $1, 2, \ldots, i - 1$.

En la figura 2.9 se tienen representados las anteriores definiciones. Por un lado, la sub gráfica (a) representa que x es directamente alcanzable por densidad desde y. Por otra parte, la sub gráfica (b) representa que la observación F es alcanzable por densidad desde A, ya que esta es alcanzable directamente por densidad desde E y este a su vez lo es por D, este por C y así hasta que B lo es por A; aunque, por ejemplo, E no es directamente alcanzable por densidad desde F ni alcanzable por densidad.

Definición* 2.2.4:** Sean x y y dos observaciones. Se dice que ambos son conectados por densidad (density-connected***) con respecto a ϵ y N_{min} si existe otra observación z tal que x y y son alcanzables por densidad desde z con respecto a ϵ y N_{min}.

Esta última relación sí es una relación simétrica. En la figura 2.10 se tiene una representación de dicha definición. Con lo anterior, la concepción de un grupo basado en densidades



(a) (b)

Figura 2.9: Conceptos Directly density-reacheable y Density-reacheable. Los puntos internos son codificados de color negro y aquellas observaciones de azul con etiqueta son puntos que por si solos no tienen una vecindad densa. El resto de elementos gráficos son observaciones.

es sencilla, ya que se buscarán conjuntos de datos maximales de puntos conectados por densidad.

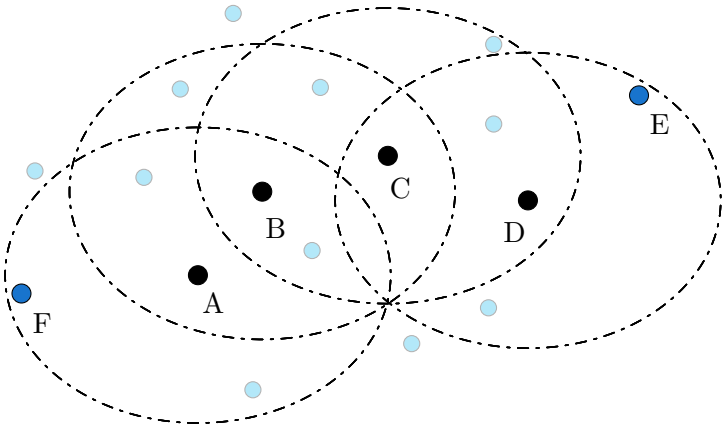


Figura 2.10: Concepto Density-connected. Véase que los puntos E y F no podrían crear vecindades densas como lo hacen los demás puntos internos.

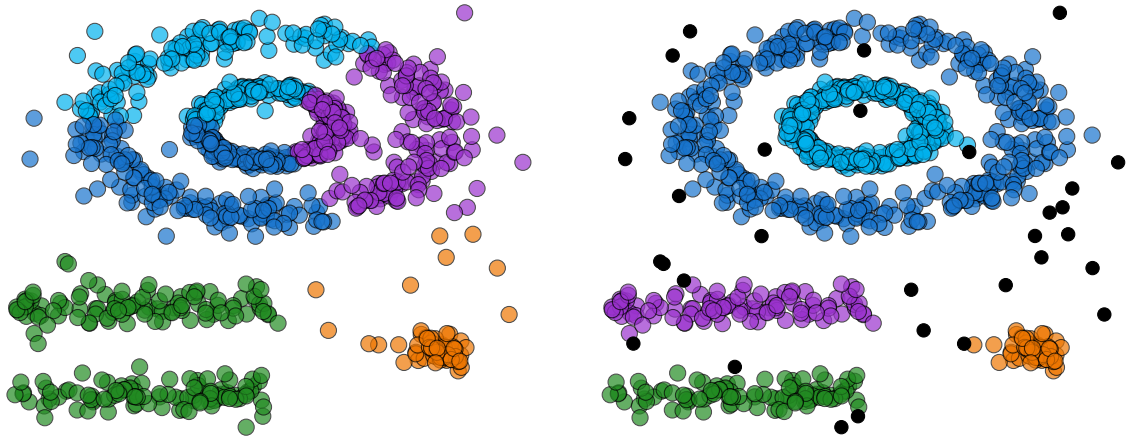
***Definición* 2.2.5:** Sea un conjunto de datos $E \neq \emptyset$. Un conglomerado C con respecto a ϵ y N_{min} es un subconjunto no vacío de E que satisface las siguientes condiciones:

1. $\forall x, y \in E$, si $x \in C$ y y es alcanzable por densidad desde x con respecto a ϵ y N_{min}, entonces $y \in C$ (**maximalidad**).
2. $\forall x, y \in C$, x y y están conectados por densidad con respecto a ϵ y N_{min} (**conectividad**).

Si un punto pertenece a un conglomerado pero su vecindad no es densa, a este punto se le llamará punto borde (***border point***). En la figura 2.10, el punto F pertenece a esta categoría. Si un punto no pertenece a ningún grupo, este será considerado como punto ruido (***noise point***), con lo que se puede dar la siguiente definición:

***Definición* 2.2.6:** Sea $C_1, \ldots, C_k$ conglomerados con respecto a ϵ_i y N^i_{min} del conjunto $E \neq \emptyset$ para $i = 1, \ldots, k$. Se define el **ruido** como el conjunto de puntos de E que no pertenecen a ningún conglomerado; es decir Ruido $= \{p \in E | \forall i : p \notin C_i\}$

Los siguientes dos lemas, como los mismos autores lo establecen [17], son importantes para validar la efectividad del algoritmo DBSCAN.



(a) (b)

Figura 2.11: Comportamiento de los métodos Kmeans (a) y DBSCAN (b) en observaciones no esféricas. En la sub figura (b), las observaciones codificadas en color negro representan lo que el método DBSCAN categorizó como ruido. La información corresponde a la base de datos "multishape"del paquete factoextra desarrollado para el lenguaje de programación R.

***Lema* 2.2.1:** Sea $x \in E$ y $|N_\epsilon(x)| \geq N_{min}$. Entonces el conjunto $W = \{w | w \in E$ y w es alcanzable por densidad desde x con respecto a ϵ y $N_{min}\}$ es un conglomerado con respecto a ϵ y N_{min}.

***Lema* 2.2.2:** Sea C un conglomerado con respecto a ϵ y N_{min} y sea $x \in C$ con $|N_\epsilon(x)| \geq N_{min}$. Entonces C es igual al conjunto $W = \{w | w$ es alcanzable por densidad desde x con respecto a ϵ y $N_{min}\}$.

Gracias al primer lema podemos establecer las etapas principales para la construcción de conglomerados en nuestros datos. Primero, se selecciona un punto arbitrario en el conjunto de datos que cumpla los criterios para ser un punto central y será considerado como una semilla. Después, se recuperan todos las observaciones que sean alcanzadas por densidad desde la anterior semilla, obteniendo así un conglomerado. Con el segundo lema, podemos concluir que un grupo respecto a ϵ y N_{min} queda determinado de forma única por cualquiera de sus puntos internos.

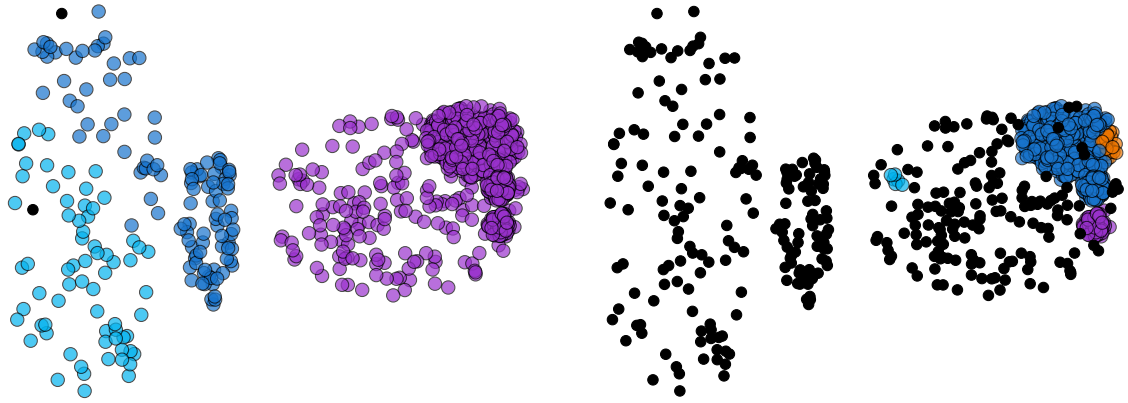
Algoritmo 3 Algoritmo DBSCAN

1: Clasificar todas las observaciones en puntos internos, borde o ruido.
2: Remover el ruido.
3: Conectar con un borde los puntos internos vecinos.
4: Crear conglomerados separados de cada grupo de puntos internos conectados.
5: Asignar a cada punto borde a uno de los conglomerados de sus puntos internos asociados.

Este algoritmo tiene diversas ventajas [55], varias de las cuales fueron previamente mencionadas en la introducción de esta sección:

- No es necesario conocer el número de conglomerados para segregar la información.
- Los conglomerados pueden tener formas arbitrarias.
- La información puede contener observaciones que sean consideradas como ruido o extremas.
- El algoritmo solo requiere dos parámetros: ϵ y N_{min}.
- El algoritmo tiene poca sensibilidad en cuanto al orden en que las observaciones son proporcionadas.

En la figura 2.11 se puede distinguir claramente la eficiencia del algoritmo DBSCAN con estructuras que tienen formas no elípticas en comparación con otro tipo de algoritmo. Aún así, es importante mencionar que el algoritmo tiene una fuerte dependencia a la función de distancia $d(x, y)$ que se esté utilizando. De hecho, la distancia entre dos conglomerados C_1 y C_2 queda determinada por $d(C_1, C_2) = min_{x \in C_1, y \in C_2} d(x, y)$, lo que hace que el método DBSCAN tienda a juntar muchos conglomerados ligeramente conectados [21].



(a) (b)

Figura 2.12: Aplicación del método DBSCAN en un conjunto de puntos simulados con diferentes valores en el parámetro ϵ. En (a), $\epsilon = 0.35$; en (b), $\epsilon = 0.08$. En ambas aplicaciones, $N_{min} = 5$.

Además de que el método es susceptible a manejar erróneamente densidades variables, también es sensible a sus parámetros ϵ, N_{min}. Un ejemplo de esto es apreciable en la figura 2.12 donde con un pequeño cambio en el parámetro ϵ, se pueden obtener resultados indeseables. Los propios autores del método sugieren el uso de una heurística para obtener ambos parámetros de acuerdo a cada conjunto de datos con la finalidad de obtener los conglomerados "más delgados".

Dicha heurística es mediante la construcción de una gráfica llamada ***k-dist graph***. Para esto, defínase a la función $F_k : E \to \mathbb{R}$ como la distancia entre x y su $k-$ésimo vecino, donde $x \in E$. La manera de proceder con esta heurística es ordenar en orden descendente los valores de $F_k(x)$[12] y graficarlos en un plano cartesiano, tal como se muestra en la figura 2.13.

Obtenida la gráfica, se procederá a buscar el primer punto de inflexión z_0 importante donde se pueda visualizar el primer "valle" en la gráfica. De a cuerdo a dicho lugar

[12]Pueden ser en orden ascendente, ya que lo único importante es tener un orden en los valores de $F_k(x)$ y ver el comportamiento de la gráfica.

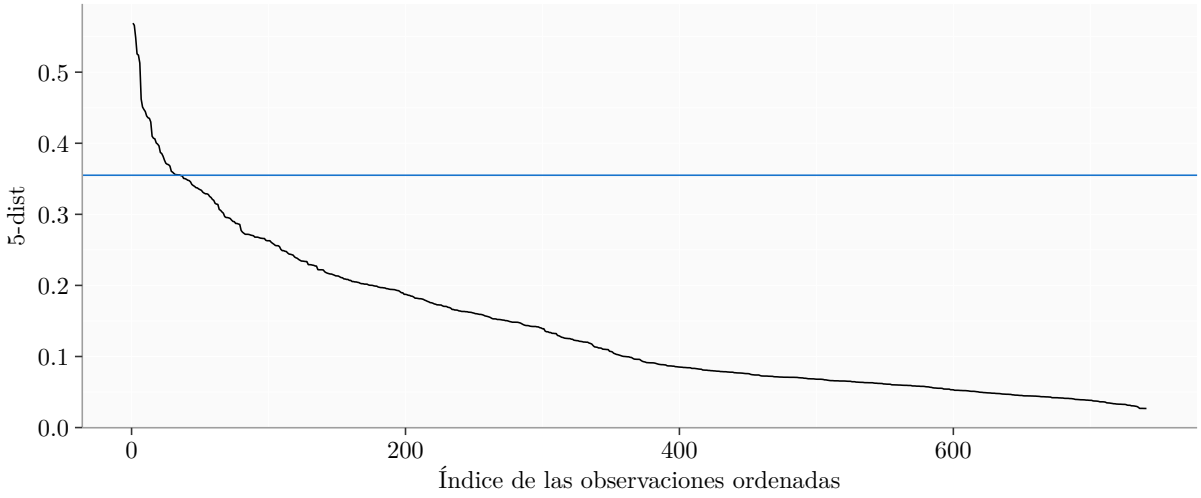


Figura 2.13: Gráfica k-dist con $k = 5$ respecto a los datos de la figura 2.12. Véase que un posible punto de inflexión se encuentra al rededor de $\epsilon = 0.35$, marcado con la línea de color azul, la cual corresponde a la sub gráfica (a) de la antes figura antes mencionada.

geométrico, ϵ es seleccionado con el valor de $F_k(z_0)$ y N_{min} se deberá seleccionar en base al k donde se obtenga el ϵ más pequeño. Se ha observado que para bases de datos con dos dimensiones, para $k > 4$ no difieren de manera significante con la gráfica 4-dist [21] [55], por lo que en dicho tipo de información, solo bastará hacer la búsqueda de ϵ.

2.2.3. Método espectral

Otra de tantas alternativas para obtener conglomerados que, al igual que los métodos basados en densidades, buscan solucionar los problemas derivados de no tener formas elípticas o esféricas, son los **métodos espectrales**, los cuales utilizan teoría de gráficas donde, posteriormente, reducen la dimensionalidad de la información mediante *eigenvectores* y ejecutan un algoritmo de agrupación (por ejemplo, K-medias) en dicho espacio de baja dimensionalidad. Para entender la logística del algoritmo, se debe dar una introducción a diferentes conceptos; pero antes de esto, recordemos un poco acerca de diversos temas relacionados al álgebra lineal.

Considérese una matriz $\mathbf{A} \in \mathcal{M}_{n\times n}(\mathbb{R})$[13]. A λ se le llamará ***eigenvalor***, **valor propio** o valor característico de $\mathbf{A}$ si existe un vector $x \neq 0$, llamado ***eigenvector***, tal que

$$\mathbf{A}v = \lambda v.$$

Véase que la anterior expresión se interpreta de la siguiente manera: el efecto que tiene la matriz $\mathbf{A}$ al aplicarla al vector v es un elongación sobre este, lo que puede significar en un cambio de signo en el vector más no un cambio de dirección. Si se considera la expresión equivalente $(\mathbf{A} - \lambda\mathbf{I})v = 0$, esto forma un sistema lineal de ecuaciones en λ y v [47], por

[13]De manera general, si $T \in \mathcal{L}(V)$, donde $\mathcal{L}(V)$ es conjunto de transformaciones lineales en el K-espacio vectorial V, K un campo; el escalar $\lambda \in K$ es un eigenvalor de T si existe un $v \in V$, llamado *eigenvector*, tal que $v \neq 0$ y $Tv = \lambda v$ [6]. Es importante recordar que existe un isomorfismo entre el espacio de matrices y el de transformaciones lineales, por lo que podemos obtener una matriz asociada a una transformación y viceversa.

lo que existe una solución no trivial ($v \neq 0$) sí y solo sí la matriz $(\mathbf{A} - \lambda\mathbf{I})$ es singular[14], por lo que es común resolver la siguiente expresión para encontrar los correspondientes eigenvalores

$$det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

A lo anterior se le conoce como ecuación característica y a $det(\mathbf{A} - \lambda\mathbf{I})$ como polinomio característico. Al conjunto de todos los escalares λ se le denomina el **espectro** de la matriz, o de manera general de la transformación lineal[15]. Si $\mathbf{A}$ es una matriz simétrica, tal como lo son las matrices de distancias o similaridad, entonces existe una matriz ortogonal[16] $\mathbf{Q}$ tal que $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathbf{T}}$, donde $\mathbf{\Lambda}$ es una matriz diagonal la cual contiene los valores propios de la matriz $\mathbf{A}$, mientras que en la matriz $\mathbf{Q}$ se encuentran los correspondientes vectores ortonormales propios. A esto último se le conoce como **factorización espectral** [47].

Los métodos de agrupación espectral pueden ser resumidos en las siguientes tres etapas [2]:

- Construir una gráfica de similitud para todas las observaciones.
- Las observaciones son proyectadas en un espacio, en el cual los grupos son más "obvios" con el uso de *eigenvectores* de la gráfica Laplaciana. Es decir, se reduce la dimensionalidad de la información.
- Aplicar un algoritmo clásico para la búsqueda de conglomerados, tal como el algoritmo K-medias; y en lugar de crear una partición sobre los datos, el problema es trasladado a un problema de partición de grafos.

Este algoritmo, al igual que en los métodos de agrupación previamente mencionados, se pueden utilizar mediante una matriz de distancias o de similaridad previamente calculada; en este caso, será necesario dicha matriz para la construcción de la gráfica requerida.

***Definición* 2.2.7:** Un **grafo** o una **gráfica** $G = (V, E)$ es un par ordenado conformado por un conjunto $V \neq \emptyset$ de vértices o nodos y un conjunto E de aristas o arcos; estos últimos son pares no ordenados de vértices.

Nuestras observaciones serán representadas por vértices en una gráfica no dirigida y ponderada[17] llamada gráfica de similitud; en esta, las relaciones entre las observaciones, o las aristas, se determinarán mediante ciertos criterios basados en la similitud entre cada dos observaciones. La ponderación o los pesos entre los vértices ayudarán a describir la gráfica mediante una **matriz de adyacencia** $\mathbf{W}$, de tal manera que cuando $\mathbf{W}_{ij}$ es igual a cero, significa que v_i y v_j no están conectados; a la vez, esto nos ayudará a conocer el grado de similitud entre dos observaciones. La construcción de esta puede determinar la

[14] Es decir, que no tenga una matriz, o transformación inversa.

[15] El término *espectro* fue motivado por temas relacionados a la física. Las líneas de energía espectral de átomos, moléculas y núcleos se caracterizan como los valores propios del operador mecánico cuántico gobernante de Schrödinger [47].

[16] Una matriz $\mathbf{A}$ es ortogonal cuando $\mathbf{A} \cdot \mathbf{A}^{\mathbf{T}} = \mathbf{I}$, es decir: $\mathbf{A}^{\mathbf{T}} = \mathbf{A}^{-\mathbf{1}}$.

[17] En una gráfica dirigida, el conjunto de aristas E está conformado por pares ordenados de vértices, por lo que será diferente $\{v_i, v_j\}$ a $\{v_j, v_i\}$, lo cual no sucede en la definición de un grafo. Un grafo es ponderado cuando existe una asociación en sus vértices con un número real.

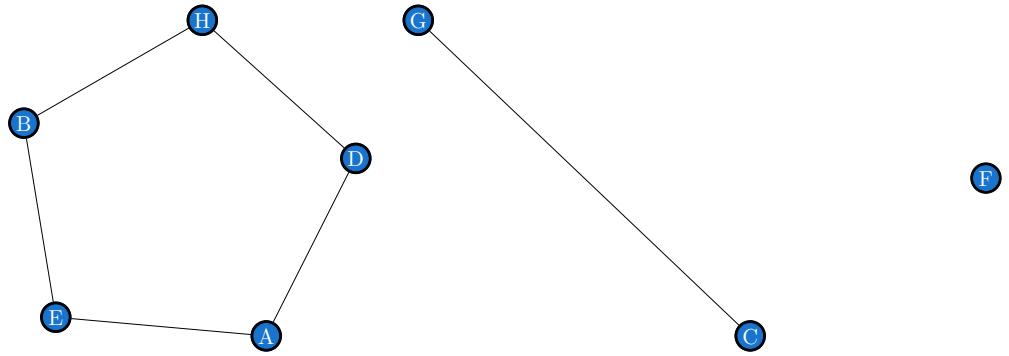


Figura 2.14: Ejemplo de un grafo. En este caso: G está conformado por $V = \{A, B, C, D, E, F, G, H\}$ y $E = \{\{A, D\}, \{A, E\}, \{B, E\}, \{B, H\}, \{C, G\}, \{D, H\}\}$.

efectividad del algoritmo sobre la información proporcionada, por lo que aquí se presentan las opciones más comunes en este tipo de algoritmos:

1. **Gráfico de ϵ-vecino**: Los vértices de esta gráfica estarán conectados cuando $||x_i - x_j||^2 < \epsilon$, por lo que en la matriz $\mathbf{W}$, en las entradas $\mathbf{W}_{ij}$, se tendrá el valor 1 cuando las observaciones estén conectadas. En este método se debe tener mucho cuidado con la elección del valor de ϵ, ya que podría eliminar conexiones importantes.

2. **Gráfico K-NN**: La forma de relacionar los vértices y asignarles sus ponderaciones es determinando si el elemento x_i está entre el k-vecino más cercano de x_j o si este está entre el $k-$vecino más cercano de x_i; cuando sucede esto, $\mathbf{W}_{ij} = 1$ y a la gráfica resultante se la llama gráfico del k-vecino más cercano. Otra manera de utilizar la distancia del vecino más cercado es pidiendo que tanto x_i y x_j se encuentren en la vecindad de cada uno de ellos; la gráfica resultante de este enfoque se le llama gráfica del k-vecino más cercano mutuo.

3. **Gráfico totalmente conectado**: En este caso, todas las observaciones estarán conectadas con una similaridad positiva. Una de las maneras más comunes es el uso del núcleo de función de base radial (*RBF kernel*); así, $\mathbf{W}_{ij} = \exp(-d_{ij}^2/c)$ donde comúnmente $d = ||x_i - x_j||$ y $c > 0$ es un parámetro de escala [2], [20].

Sobre el último caso, la elección del parámetro de escala tiene una gran influencia en la eficacia del algoritmo, un ejemplo de esto se puede visualizar en la figura 2.15 donde, en la parte superior se tienen los resultados de aplicar el método espectral con los correctos parámetros de escala para cada uno de los conjuntos de datos. Por otro lado, la parte inferior muestra malos resultados del método con diferentes valores de c[18].

En [59] se sugiere que el parámetro de escala sea establecido de manera local por el ancho de las vecindades que se crean por cada observación x_i con su k-vecino más cercano. Así, considerando que la distancia entre las observaciones x_i y x_j, puede ser vista como $d(x_i, x_j)/\sigma_i$ y $d(x_j, x_i)/\sigma_j$, $d^2 = d(x_i, x_j)d(x_j, x_j)/\sigma_i\sigma_j = d^2(x_i, x_j)/\sigma_i\sigma_j$; donde σ, así como se sugiere en el mismo artículo, puede ser igual a $d(x_i, x_K)$ donde x_K es el k-ésimo vecino de x_i. Resumiendo, los valores en la matriz de afinidad $\mathbf{W}$ serán calculados de la siguiente manera:

[18]La función original del documento del cual fue obtenido la figura 2.15 es $\exp(-d^2(x_i, x_j)/\sigma^2)$.

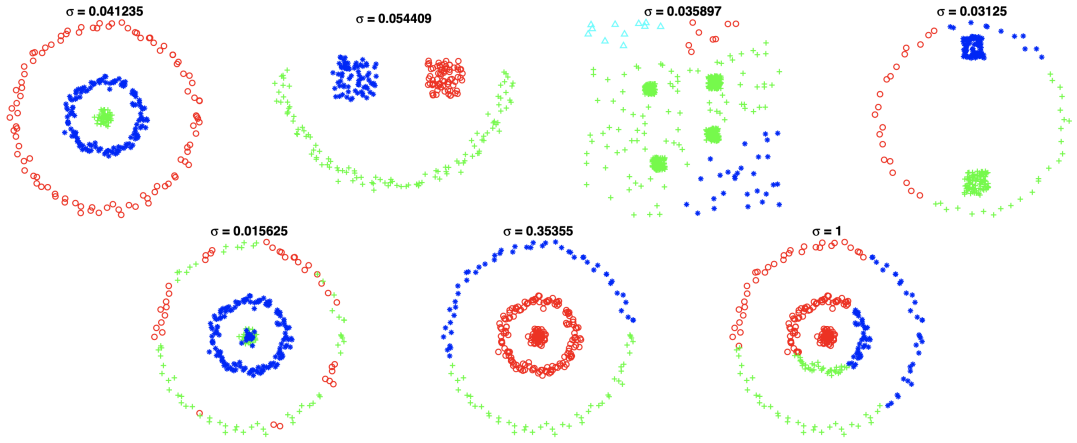


Figura 2.15: Sensibilidad en la creación de grupos mediante el método espectral sobre parámetro de escala en la función kernel RBF en diferentes conjuntos de datos. (Desde Zelnik-manor, L. and Perona, P. Self-tuning spectral clustering. Advances in neural information processing systems, 2005.)

$$\mathbf{W_{ij}} = \exp\left(\frac{-d^2(s_i, s_j)}{\sigma_i \sigma_j}\right).$$

El efecto de la elección del valor de c de manera local puede verse en la figura 2.16, donde en la figura (b) y (c) se puede ver, mediante el ancho de las aristas que conectan a las observaciones, la afinidad o la similitud de los puntos. En particular, en la última de estas, la afinidad que se tiene de manera interna en los dos grupos, la cual se pueden distinguir claramente en la figura (a), es mejor, ya que ayuda a discernir de una manera más clara los grupos en el grafo y, por lo tanto, a particionar la gráfica con los conglomerados adecuados.

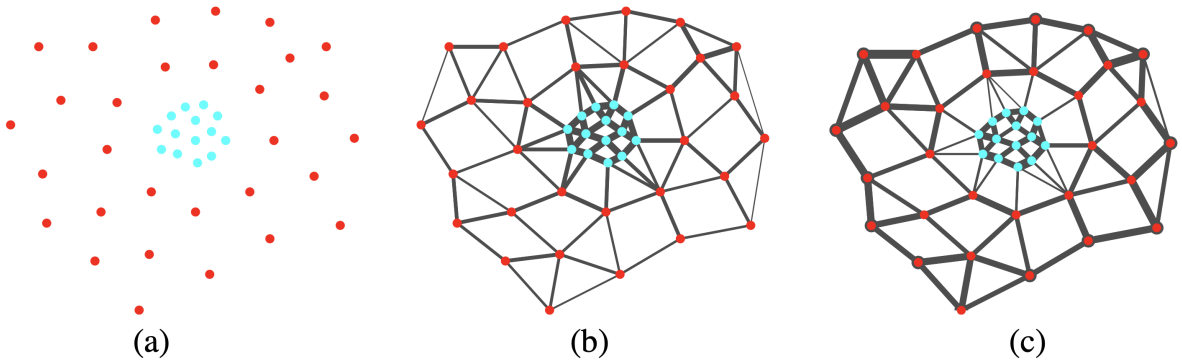


Figura 2.16: Efecto del escalamiento local en la función kernel RBF. (Desde Zelnik-manor, L. and Perona, P. Self-tuning spectral clustering. Advances in neural information processing systems, 2005.)

Obtenida la matriz de adyacencia $\mathbf{W}$ de la gráfica[19], se buscará construir la matriz Laplaciana, la cual está definida en términos de la matriz $\mathbf{W}$ y la matriz de grados, esta contiene los grados de cada vértice en la gráfica. Para cada vértice, su **grado** será calculado

[19]Véase que la matriz $\mathbf{W}$ es simétrica.

como la suma de los pesos incidentes en el[20], es decir, si se tienen n observaciones, para el vértice i, su grado se calcula como: $d_i = \sum_{j=1}^{n} \mathbf{W}_{ij}$. Con esto, la **matriz de grados D** es definida como la matriz diagonal donde $\mathbf{D}_{ii} = d_i$.

La **gráfica Laplaciana**[21] quedará determinada como:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}.$$

El método espectral encontrará los m *eigenvectores* $Z_{n\times m}$ que correspondan a los m eigenvalores más pequeños de $\mathbf{L}$. La matriz $\mathbf{L}$ tiene propiedades interesantes; por ejemplo:

1. Para cualquier vector $f \in \mathbb{R}^n$

$$\begin{aligned} f^T\mathbf{L}f = f^T(\mathbf{D}-\mathbf{W})f &= \sum_{i=1}^{n} f_i^2 d_i - \sum_{i,j=1}^{n} f_i f_j \mathbf{W}_{ij} \\ &= \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}(f_i - f_j)^2. \end{aligned}$$

2. $\mathbf{L}$ es simétrica y positiva definida.
3. El eigenvalor más pequeño de $\mathbf{L}$ es 0, con el eigenvector constante $\mathbb{1}_n$.
4. $\mathbf{L}$ tiene n no negativos eigenvalores reales: $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Las demostraciones de estas propiedades pueden encontrarse en [2] y [55]. Para entender la importancia de utilizar este tipo de gráfica, es importante tener en cuenta las siguientes definiciones:

***Definición* 2.2.8:** Un grafo $G = (V, E)$ se dice **conexo** si, dados cualesquiera dos vértices $u, v \in V$, existe una $\{u, v\}$-trayectoria; es decir, si u y v están conectados.

***Definición* 2.2.9:** Diremos que H es un **componente conexo** de G si H es una subgráfica conexa de G y G es un grafo no conexo.

Como se puede distinguir en la figura 2.17, se tienen claramente 3 grupos, los cuales son las componentes conexas del grafo $G(V, E)$ representada en la figura antes mencionada. Esto sugiere buscar dichas componentes conexas en un grafo que sea creado en base a nuestra información y esta es la relevancia de considerar la gráfica Laplaciana, ya que esta tiene la siguiente propiedad [2]:

***Proposición* 2.2.3:** La multiplicidad k del eigenvalor 0 de $\mathbf{L}$ es igual al número de componentes conexas $A_1, A_2, \ldots, A_k$ en el grafo y el espacio de eigenvectores de eigenvalor igual a 0 es generado por los vectores indicadores $\mathbb{1}_{A_1}, \mathbb{1}_{A_2}, \ldots, \mathbb{1}_{A_k}$ de dichas componentes.

[20] El grado de un vértice es el número de vecinos, o los otros vértices, que están conectados a él, por lo que el grado de un vértice, en una gráfica con n vértices sin pesos, es un número entero entre 0 y $n-1$. En nuestro caso, se tienen pesos en cada arista, por lo que el grado de cada vértice se calcula de la manera mencionada.

[21] En realidad, el término correcto es **gráfica Laplaciana no normalizada**, ya que existen algunas versiones que normalizan a dicha matriz de adyacencia con respecto a los grados de los nodos; por ejemplo: $\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$ y $\mathbf{D}^{-1}\mathbf{L}$ [2].

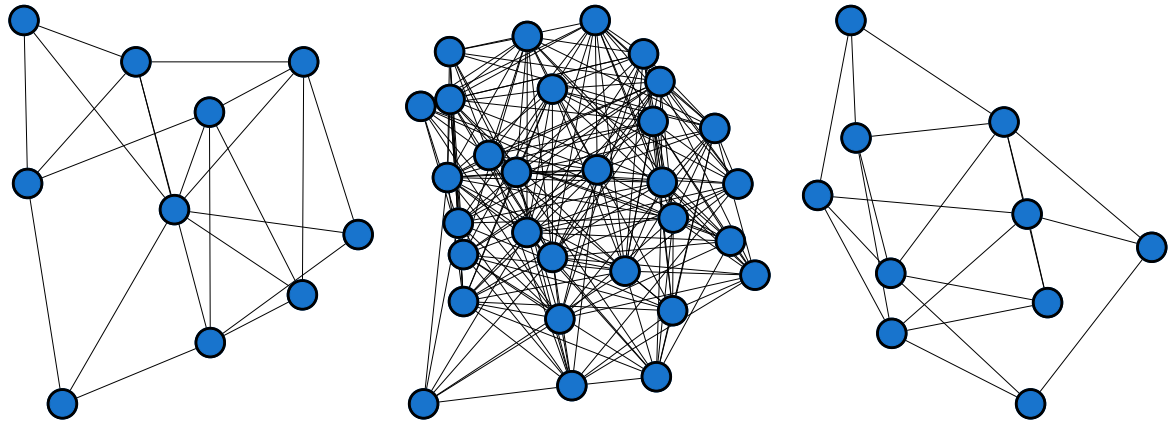


Figura 2.17: Una gráfica no conexa conformada por 50 vértices y 263 aristas en la cual son distinguibles 3 componentes conexas

En la figura 2.18 podemos visualizar el resultado de aplicar el método espectral basándonos en toda la teoría que se ha mencionado, la cual queda sintetizada en el algoritmo 4. En el caso de la figura antes mencionada, se creó un gráfico totalmente conectado utilizando un escalamiento local considerando 7 vecinos. En la subgráfica (b) podemos ver que solo un eigenvalor es igual a cero, pero en la práctica, se pueden tener fuertes o débiles similaridades entre los vértices, por lo que se utilizan aproximaciones a los eigenvalues iguales a cero con los más cercanos a este; para este caso se consideraron los primeros 5 eigenvectores que corresponden a los 5 eigenvalores más pequeños. En (c) se grafican los valores de las entradas del eigenvector más pequeño, visualizando así la proyección que se tienen de los datos, mostrando claramente una segregación.

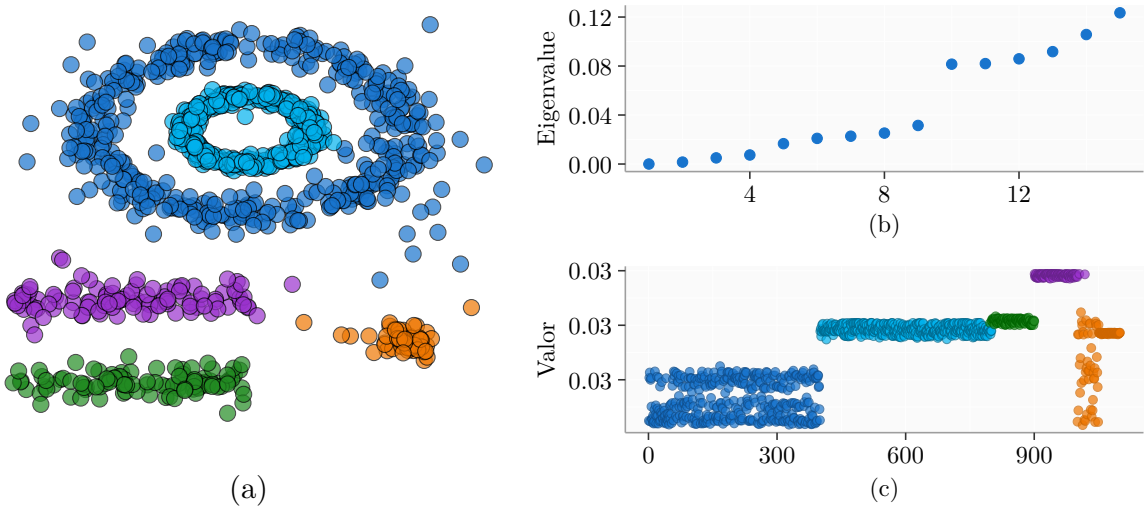


Figura 2.18: Método espectral en la base de datos utilizada en la figura 2.11. En (a) se tiene el resultado final del método. En (b) se graficaron los valores de los eigenvalores ordenados de manera ascendente. Respecto a (c), se visualizan los valores de las entradas del eigenvector correspondiente al eigenvalor más pequeño.

Por el mismo hecho de que pueden existir grupos débilmente separados, se utiliza otra técnica de agrupación, en el caso de los resultados de la figura 2.18 se aplico el método

k-means con los 5 eigenvectores antes mencionados.

Algoritmo 4 Agrupación Espectral

1: Construir una gráfica de similaridad con alguno de los métodos mencionados. Sea $\mathbf{W}$ la matriz de adyacencia y $\mathbf{D}$ la matriz de grados.
2: Calcular la gráfica Laplaciana $\mathbf{L}$.
3: Determinar los mejores k eigenvectores $f_1, f_2, \ldots, f_k$ con base a los k menores eigenvalores de $\mathbf{L}$.
4: Construir la matriz $\mathbf{F} = [f_1, f_2, \ldots, f_k] \in \mathbb{R}^{n \times k}$.
5: Considerar cada renglón de $\mathbf{F}$ como las observaciones a ser agrupadas y realizar un método para obtener conglomerados, como K-means, sobre estos.

2.3. Estadísticas de evaluación para conglomerados

2.3.1. Estadística de Hopkins

Una de las preguntas que se deberían contestar antes de iniciar con el uso de alguna técnica de agrupamiento es: ¿Existe evidencia para crear grupos entre los datos? Es decir, se debe evaluar de alguna forma si las estructuras que contiene la información no son simplemente aleatorias. El problema con los algoritmos que buscarán conglomerados es que, al aplicar cualquier método de agrupamiento este devolverá una partición de los datos aunque no existan grupos en realidad.

Para abarcar dicho problema, es común utilizar la estadística de Hopkins para determinar si es factible realizar algún método de agrupación en las observaciones que se estén estudiando. Esto se hace midiendo la probabilidad de que el conjunto de datos se genere mediante una distribución de datos uniforme. Es decir, se realiza una prueba de aleatoriedad espacial de los datos [24] [29].

La manera más común de estudiar la aleatoriedad espacial es mediante una prueba de Aleatoriedad Espacial Completa (CSR), donde los eventos (las observaciones) se distribuyen de forma independiente, aleatoria y uniforme en el área de estudio. Esto implica que no hay regiones donde los eventos sean más (o menos) probables de ocurrir y que la presencia de un evento dado no modifica la probabilidad de que aparezcan otros eventos cercanos [8]. Gracias a las anteriores condiciones, a un CSR se le puede considerar un sinónimo de un proceso de Poisson homogéneo [39].

La manera de calcular el estadístico de Hopkins utiliza lo anterior, por lo que este es parte fundamental de una prueba de hipótesis donde la hipótesis nula establecería si los datos o los patrones fueron generados por un proceso Poisson con intensidad λ por unidad de volumen [7], en otras palabras:

$$H_0 : \text{El conjunto de datos } D \text{ se distribuyen de manera uniforme}$$
$$\text{vs}$$
$$H_a : \text{El conjunto de datos } D \text{ no se distribuyen de manera uniforme}$$

En el primer caso, no existirán grupos significativos que buscar, mientras que el segundo caso nos daría indicios de grupos importantes. De manera práctica, el estadístico de Hopkins H se calcula de la manera siguiente [29]:

1. Obtener un muestreo de manera uniforme de n observaciones $(p_1, \ldots, p_n \in D)$.
2. Para cada punto p_i, encontrar su vecino más cercano p_j y obtener $x_i = d(p_i, p_j)$.
3. Generar un conjunto de datos simulados (S) desde una distribución uniforme con n observaciones $(q_1, \ldots, q_n)$ que tenga la misma variación que el conjunto de datos D.
4. Para cada punto q_i, encontrar su vecino más cercano q_j en D y obtener $y_i = d(q_i, q_j)$.
5. Calcular el estadístico de Hopkins como la distancia media del vecino más cercano en s dividida por la suma de las distancias medias del vecino más cercano en el conjunto de datos real y simulado. Es decir:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}.$$

Si el conjunto de datos que se está estudiando (D) se distribuyera de manera uniforme, entonces $\sum_{i=1}^{n} x_i$ y $\sum_{i=1}^{n} y_i$ serían muy cercanos y el estadístico $H \approx 0.5$. Sin embargo, si D contiene indicios de tener información sesgada, se esperaría que $\sum_{i=1}^{n} y_i$ fuera considerablemente más pequeño que $\sum_{i=1}^{n} x_i$ y por lo tanto $H \approx 0$ [24]. Es decir, buscaremos H sea cercano a cero para así rechazar la hipótesis nula y concluir que D es significativamente un conjunto de datos en el cual se puedan obtener conglomerados [29]. En la figura 2.19 se puede ver un ejemplo del comportamiento de dicha estadística con diferentes grupos de datos, donde es fácil entender que buscamos un valor de H pequeño.

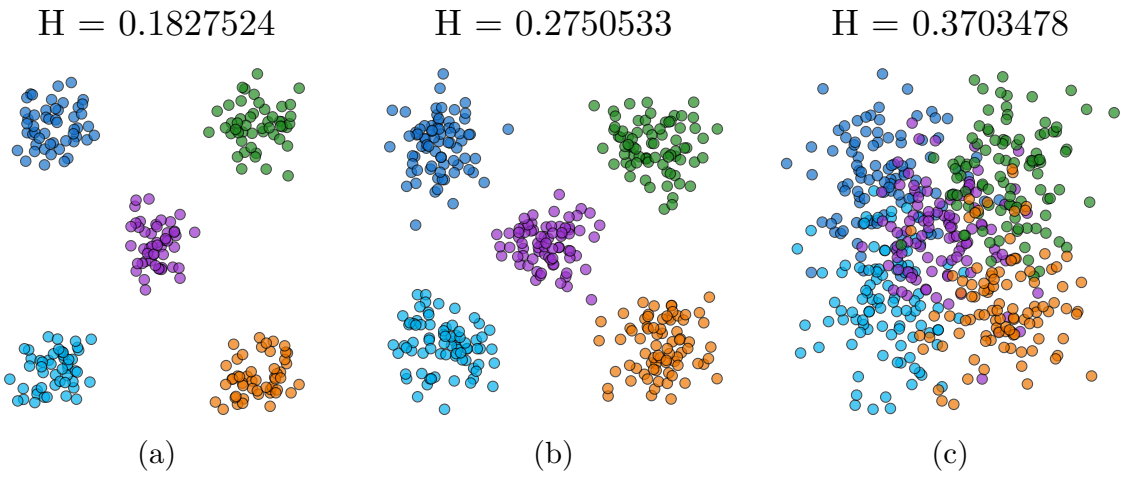


Figura 2.19: Comportamiento de la estadística de Hopkins en diferentes datos simulados. Estos varían por la cantidad de ruido agregado mediante una distribución normal variante en desviación estándar para cada subgráfica. La H es el valor de la estadística para cada conjunto de datos.

Es importante mencionar que, bajo la hipótesis nula, el estadístico H sigue una distribución $Beta(n, n)$ y solo se realiza una prueba de cola izquierda, ya que si los datos fueran espacialmente más aleatorios de lo esperado por casualidad, aún se considerarían no agrupables, por lo que podemos obtener el $p - value$ como $\mathbb{P}(H < q_\alpha(n, n))$ donde q_α es el cuantil de la distribución $Beta(n, n)$ al $\alpha \times 100\,\%$ [1].

2.3.2. Métricas de evaluación

Determinar la efectividad de un método de agrupamiento es un tema importante, ya que puede ser complicado analizar la homogeneidad de los grupos creados, ya sea por la gran cantidad de información que se esté considerando o porque no se tiene una manera práctica de ver en que son discernibles los grupos. De acuerdo al tipo de información que se esté analizando, se pueden dividir las métricas de evaluación en medidas de evaluación internas y externas.

Las métricas de evaluación externas utilizan información que no se usa para la creación de grupos, tal como etiquetas que identifiquen previamente a los datos, de tal manera que pueden ayudar a determinar si una observación fue categorizada correctamente o no. Ejemplos de este tipo hay muchos, tales como la pureza, el estadístico de Rand o hasta una simple tabla de contingencia. Por otro lado, si no se tiene información externa, solo se puede hacer uso de la información que considera el método de agrupamiento y no se tiene manera de validar una categorización a priori. Este es el que se desarrollará en el resto del capítulo ya que este es el tipo de información con la que se trabajó.

El objetivo de un método de agrupación es crear grupos donde las observaciones sean lo más parecidas a ellas dentro de los grupos (cohesión) y lo más separadas con observaciones fuera de sus propios grupos (separación), en esto están basadas las métricas de evaluación interna. A continuación se verán algunas de las métricas más utilizadas haciendo énfasis en sus ventajas y desventajas [2].

2.3.2.1. Índice de la Silueta (S)

El índice de la silueta (*Silhouette index*) es una de las métricas más utilizadas y conocidas en este tema. Ésta valida el rendimiento de los conglomerados creados considerando el cambio que se produce en la forma (la silueta) de los grupos al agregar una observación, esto midiendo la fuerza que tiene dicha observación de pertenecer al grupo al que fue asignado y también observando la poca pertenencia que tiene en otros grupos. Para esto, se definen las siguientes funciones:

$$a(x) = \frac{1}{n_i - 1} \sum_{y \in C_i, y \neq x} d(x, y); \quad b(x) = \min_{j, j \neq i} \left[\frac{1}{n_j} \sum_{y \in C_j} d(x, y) \right].$$

$a(x)$ es la distancia promedio que tiene la observación x en su grupo de pertenencia C_i y $b(x)$ es la distancia promedio hacia los otros puntos de los otros conglomerados. Dado esto, se define el **índice de silueta** para el punto x como

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}.$$

Dicho coeficiente tiene un rango de valores entre -1 y 1, donde -1 indica que el punto x tendría una mayor fuerza de pertenencia en otro grupo y no al que fue asignado, se obtiene un valor de 1 cuando x está en el grupo correcto y 0 cuando es indiferente el grupo donde esté, lo cual tampoco es lo esperado. Considerando todas las observaciones y que se tienen k grupos, se tiene el **índice de silueta global** (ASW):

$$\frac{1}{k}\sum_{k=1}^{k}\left[\frac{1}{n_k}\sum_{x\in C_k} s(x)\right].$$

A pesar de que el índice de silueta global se utiliza para determinar la eficiencia de un algoritmo de agrupamiento, hay que considerar que este mide la fuerza de pertenencia en promedio, por lo que diversas observaciones podrían estar sesgando este índice[22]. Es pertinente visualizar una gráfica donde se muestren los índices que se obtuvieron de manera individual en todos los grupos y así evitar problemas de sesgo, un ejemplo se puede encontrar en la figura 2.20.

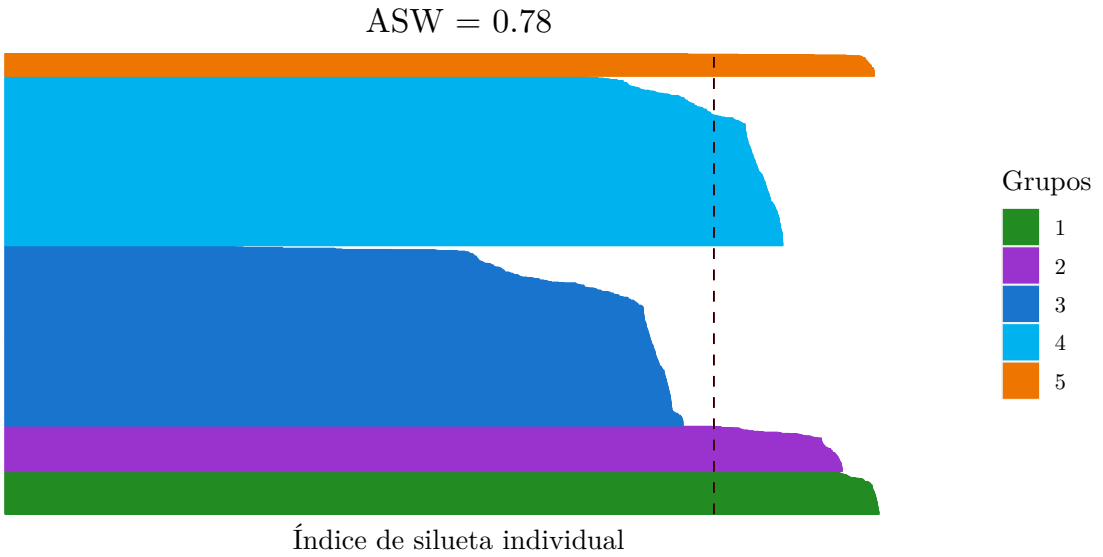


Figura 2.20: Gráfico de silueta correspondiente a los resultados de la figura 2.18. La línea puntuada está ubicada en el índice de silueta global. Se utilizó la métrica euclideana para los cálculos correspondientes.

Es común en este tipo de gráficas localizar el índice de silueta global para determinar si los grupos tienen una buena estructura; esto sucede cuando el grupo, en general, sobre pasa el índice global.

2.3.2.2. Índice Calinski-Harabasz (CH)

El índice de Calinski-Harabasz (CH) [11] evalúa la eficiencia del algoritmo considerando la variación interna de los grupos y la variación entre los grupos [25] [41]. Considerando k grupos y n observaciones, la variación entre los grupos queda determinada por

$$B = \sum_{i=1}^{k} n_i d^2(x_i, \bar{x}),$$

donde $\bar{x}$ es centroide de todo el conjunto de datos y n_i el número de elementos en el grupo i. La variación interna de los grupos se calcular mediante la siguiente expresión:

[22] Esto puede evitarse normalizando las observaciones.

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} d^2(x_i, \bar{x}_j).$$

En este caso, $\bar{x}_i$ es el centroide del grupo j. Con esto, el índice CH queda definido de la siguiente manera:

$$CH = \frac{B/(k-1)}{W(n-k)}.$$

Un índice alto, indicará una mejor agrupación entre los datos aunque se pueden conseguir índices más altos para agrupaciones convexas y esféricas, las cuales no siempre suceden [25].

2.3.2.3. Índice Davies-Bouldin (DB)

El índice de Davies-Bouldin (DB) fue propuesto en 1979 [16], el cual solicita distintas propiedades para medir la separación general entre conglomerados; para esto define la dispersión interna de los k grupos como:

$$S_k = \left(\frac{1}{n_k} \sum_{i=1}^{k} ||x_i - \bar{x}_k||_2^q \right)^{1/q}.$$

La similitud entre los k grupos queda determinada en términos de S_i y S_j, con $i, j \in 1, 2, \ldots, k$, y considerando la distancia entre los centroides M_{ij} de los grupos i y j:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}; \quad M_{ij} = ||\bar{x_i} - \bar{x_j}||_p.$$

R_{ij} es una función no negativa y simétrica, además, $R_{ij} = 0 \Leftrightarrow S_i = 0$ y $S_j = 0$, si la distancia entre los grupos aumenta mientras su dispersión se mantiene constante, la similitud entre los grupos decrece y si la distancia entre los grupos se mantiene constante mientras la dispersión incrementa, la similitud incrementa [25] [16].

Finalmente, para cada conglomerado se calcula su máxima similitud con otros grupos y se promedian todas estas para obtener el índice DB.

$$DB = \frac{1}{k} \sum_{i=1}^{k} D_i; \quad D_i = \max_{i \neq j} R_{ij}.$$

En este caso, se buscará que los grupos tengan poca similitud entre ellos y así obtener un índice bajo. Al igual que en el índice anterior, se pueden tener mejores resultados con agrupaciones convexas y esféricas. Además, la manera de calcular la distancia entre los centroides limita la métrica de la distancia al espacio euclideano.

2.3.2.4. Índice basado en vecinos cercanos ($CVNN$)

El índice basado en vecinos más cercanos (Clustering validation index based on nearest neighbors: $CVNN$) fue propuesto en 2013 [37] con la finalidad de no medir la separación utilizando valores de disimilaridad, sino en considerar la cantidad de los k vecinos más cercanos en cada observación en cada uno de los k conglomerados.

Con este índice se tratará de encontrar el mejor número de agrupaciones dado un método de agrupación por lo que los autores proponen comparar un conjunto de K agrupaciones $\mathcal{K} = \{\mathcal{C}_{K_{min}}, \ldots, \mathcal{C}_{K_{max}}\}$, por lo que no puede ser calculado para un solo grupo aislado. Este índice requiere de dos estadísticas. La estadística de separación en el grupo K queda definida como

$$Sep_k(C_K) = \underset{i=1,\ldots,k}{\text{máx}} \left(\frac{1}{n_i} \sum_{x \in C_i} \frac{q_k(x)}{k} \right),$$

donde $q_k(x)$ es el número de observaciones entre los k vecinos más cercanos de x que no están en el mismo grupo en el grupo C. Por otra parte, la estadística de compacidad, que es solo el promedio de la disimilaridad en los grupos [25], queda definida como

$$Com(C_K) = \frac{\sum_{j=1}^{K} \sum_{x_h \neq x_i \in C_j} d(x_h, x_i)}{\sum_{i=1}^{K} n_i(n_i - 1)}.$$

Se buscará que ambas estadísticas sean pequeñas, ya que si una observación está ubicada en el centro de un grupo y está rodeado por observaciones del mismo grupo, está bien separado de otros grupos y, por lo tanto, contribuye poco a la separación entre grupos. Si un punto está ubicado en el borde de un grupo y está rodeado principalmente por objetos en otros grupos, se conecta estrechamente a otros grupos y, por lo tanto, contribuye mucho a la separación entre conglomerados [2]. Con esto, se define el índice $CVNN$ como

$$CVNN_k(C_K) = \frac{Sep_k(C_K)}{\underset{\mathcal{C} \in \mathcal{K}}{\text{máx}}\, Sep_k(C)} + \frac{Com(C_K)}{\underset{\mathcal{C} \in \mathcal{K}}{\text{máx}}\, Com(C)}.$$

En el documento donde se propone éste, el índice $CVNN$ se comparó con distintos índices de evaluación internos, mostrando que este índice puede tener una menor sensibilidad en presencia de ruido, en grupos densos o en conglomerados con grupos internos, grupos con distintos tamaños y grupos con formas arbitrarias [2] [37].

Capítulo 3

Análisis de audio y de Fourier

En este trabajo se mostrarán diferentes resultados de la aplicación de los anteriores métodos de agrupación en diferentes piezas musicales, para esto será necesario entender qué es el sonido y cómo pueden ser estudiadas dichas señales de audio. De acuerdo al tratamiento que se le desee dar al sonido, las señales de audio pueden ser estudiadas desde sus propiedades físicas más básicas o mediante el uso de transformaciones para la obtención de características más interesantes. Las señales de cualquier sonido son, en general, una mezcla de diferentes sonidos y será importante descomponer dicha señal en diferentes componentes para su análisis.

Debido al comportamiento físico del sonido, es posible crear un espectro de dichas componentes mediante la modelación con transformaciones de Fourier, esta es la razón por la que es importante dar una introducción sobre este tema, ya que será parte fundamental para la obtención de características de una canción, resumir esta misma y así poder compararla con otras piezas musicales. Aquí también se dará una breve introducción sobre las características del sonido, con énfasis en obras musicales y la obtención de diferentes características que se han desarrollado en este ámbito considerando las interpretaciones psicológicas, en especial de los Coeficientes Cepstrales en la Frecuencia de Mel (MFCC).

3.1. Señales de audio

Un **sonido** es una onda longitudinal[1] que se propaga por un medio elástico, como el aire o el agua, y es generado por un objeto en vibración, por ejemplo las cuerdas vocales o las cuerdas de cualquier guitarra, piano, violín, etc. Dichas vibraciones producen desplazamientos y oscilaciones de las partículas del aire haciendo que en este se generen, de manera local, compresiones y refracciones. Dichos cambios de presión viajan en forma de onda desde la fuente hasta el receptor, un micrófono por ejemplo o el oído humano. En el caso del oído, las vibraciones en el aire se transportan hasta el tímpano, el cual es forzado a vibrar de acuerdo a oscilaciones de presión. Después de un procedimiento intermedio en el oído interno, las ondas de sonido son transformadas en pulsos nerviosos, los cuales son enviados e interpretados por el cerebro [45].

Es común considerar a las ondas longitudinales desde el punto de vista de variaciones en la presión en vez de como estas se desplazan. El cambio de presión en el aire (usualmente

[1]Existen dos tipos de ondas: transversales y longitudinales. En las primeras, las partículas vibran de manera perpendicular a la dirección de la propagación de la onda, tal efecto se puede visualizar en un lago al lanzar un objeto en él. En las ondas longitudinales las partículas individuales vibran de manera paralela en dirección a la propagación de la onda.

medido en pascales) puede ser visualizado como en la figura 3.1, donde el comportamiento de las partículas, en promedio, se grafican con el transcurso del tiempo[2]. De la misma figura se pueden visualizar algunos conceptos importantes; por ejemplo, si se alcanzan los puntos más altos y bajos en la presión del aire de manera repetida y de manera alternada, la onda es llamada **periódica**, y el **periodo** queda definido como el tiempo entre dos puntos de presión máximos (o mínimos) sucesivos[3]. En el momento en que se llega a tener una presión de 0, se dice que se tiene la **posición de equilibrio**. La **frecuencia**, medida en hertz (HZ) es el número de ondas que pasan por un punto de equilibrio y es el recíproco del periodo ($f = 1/\lambda$) [44].

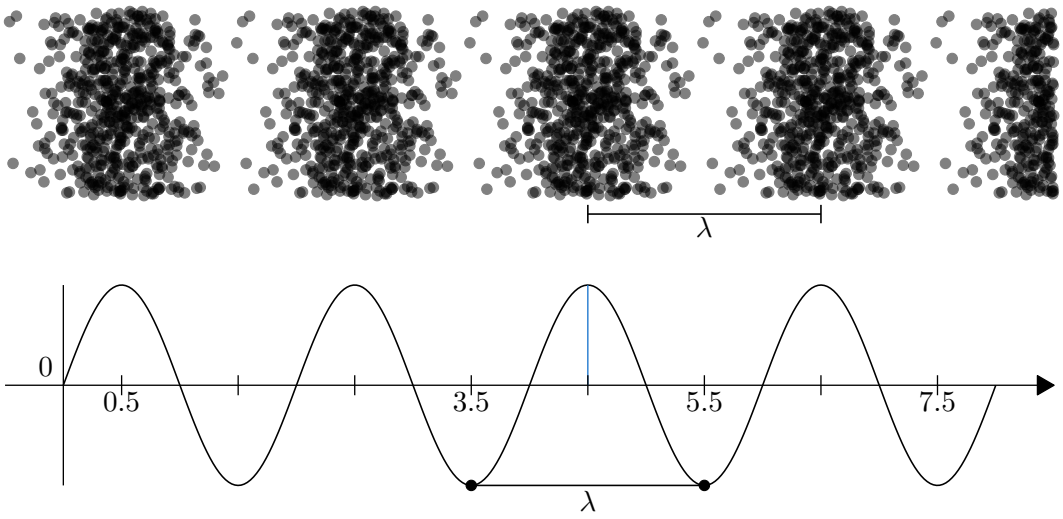


Figura 3.1: Representación gráfica de un sonido periódico con frecuencia de 1/2 Hz. λ representa el periodo (2 segundos en este caso) y la línea azul la amplitud o la distancia máxima en las zonas de compresión.

Ya que el sonido es interpretado por el cerebro, se tienen distintas sensaciones subjetivas en la conciencia de quien escucha, como la intensidad y el tono, las cuales tiene su relación con una cantidad física medible [22]. La **intensidad sonora**, o mejor dicho la **sonoridad** (*loudness*) se relaciona con la intensidad de una onda[4]. El oído humano, en promedio, puede detectar sonidos entre $10^{-12}W/m^2$ y $1W/m^2$; un sonido con una potencia mayor provocaría una sensación de dolor. Dentro de este rango, la sonoridad percibida no es directamente proporcional a la intensidad; por ejemplo, una duplicación de la sonoridad percibida corresponde aproximadamente a un incremento en la intensidad por un factor de 10 [56], por lo que es común medir la intensidad en **decibelios** (dB)

$$dB(I) := 10 \cdot \log_{10}\left(\frac{I}{I_{THO}}\right)$$

donde I_{THO} es el umbral auditivo, es decir $I_{THO} = 10^{-12}W/m^2$.

[2]La sub gráfica inferior de la figura 3.1 es conocida como "*waveform*" del sonido.

[3]En el análisis físico de las ondas, el periodo puede ser entendido como la longitud de onda.

[4]La intensidad de una onda es la cantidad de energía transportada por esta por unidad de tiempo a través de una unidad de área perpendicular al flujo de energía. Esta se mide en unidades de potencia por unidad de área, o watts/metro^2 [22]. Otro punto importante es que la sonoridad no es lo mismo al **volumen**, ya que esté es una percepción subjetiva de la potencia.

La sonoridad también depende de que tan sensible sea el oído, lo cual puede variar con la edad, y de la frecuencia en que este se encuentre, ya que el oído humano es más sensible a frecuencias entre 2000 y 5000 Hz. Desde 1933, se ha determinado, mediante diferentes estudios, como los niveles de presión del sonido son igualmente percibidos en diferentes frecuencias, obteniendo así las llamadas **curvas de Fletcher-Munson**[5], las cuales pueden ser visualizadas en la figura 3.2.

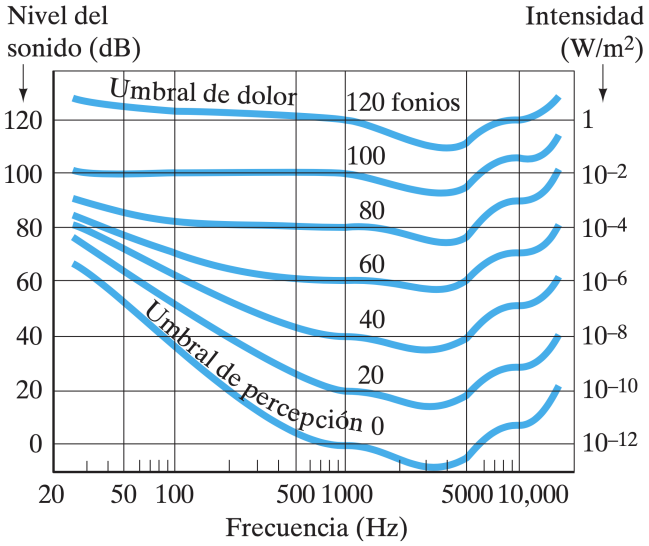


Figura 3.2: Contornos de igual sonoridad. Por ejemplo, La curva correspondiente a 40 fonios representa los sonidos que, en promedio, se perciben con la misma intensidad que un sonido a 1000 Hz con 40 dB, por lo que un sonido a 100 Hz se percibirá igual de intenso si este está a un nivel de 62 dB aproximadamente. (Desde Giancoli, D.C. Física para ciencias e ingeniería con física moderna, 2009.)

En dicha gráfica, cada curva describe la relación entre la presión del sonido en dB y un nivel particular de percepción sonora (los sonidos en toda la curva parecen tener la misma intensidad), variando sobre las frecuencias. Las unidades de percepción sonora o los niveles de intensidad se llaman **fonios** y es numéricamente igual al nivel de sonido en dB a 1000 Hz. Considerando esta escala, que aún es logarítmica, podemos tener información adecuada de la relación no lineal entre la presión sonido y la sensación sonora del humano, lo cual es de suma importancia en el procesamiento de señales [32].

Por otra parte, el **tono** (*pitch*), o la altura, de un sonido es una característica psicológica que los identifica como agudos (como el sonido de un violín) o graves (como el sonido de un tambor) en función de la frecuencia; entre más baja sea la frecuencia, menos agudo será el tono y entre más alta la frecuencia, más agudo este resultará. En el caso del oído humano, este puede percibir sonidos entre 20 Hz y 20,000 Hz, conocido este como **rango audible**.

Un **tono puro** o **sonido armónico**, por definición, corresponde a una onda senoidal $y(t) = A \cdot sen(2\pi ft)$, como la representada en la figura 3.1, donde A representa la amplitud, y f la frecuencia del sonido. Por ejemplo, para lo que se ha catalogado como la nota

[5]El nombre de dichas curvas es debido a los desarrolladores del primer estudio, y aunque ya no son utilizadas dichas curvas, el nombre sigue siendo utilizado para describir a los contornos de igual sonoridad. Las curvas actuales fueron integradas dentro de la estandarización ISO 226:2003 [32].

la central, se tiene una frecuencia de 440 Hz [6]. Así como la intensidad se interpreta en términos logarítmicos, también se hace con el tono, por lo que la distancia percibida entre los tonos A3 (220 Hz) y A4 (440 Hz) es la misma que entre A4 (440 Hz) y A5 (880 Hz). El intervalo entre dos sonidos con el doble o la mitad de la frecuencia es llamado una **octava** [45].

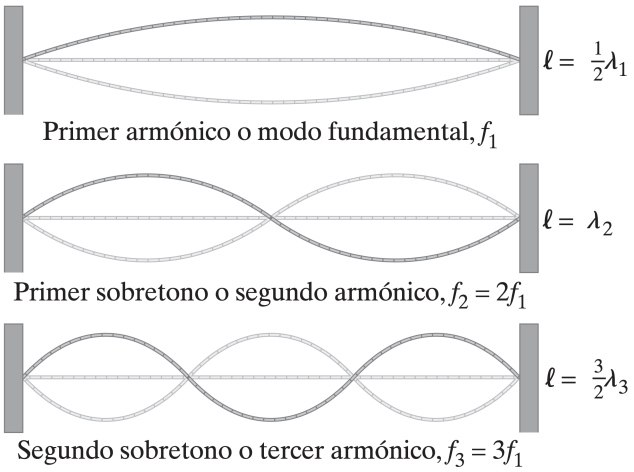


Figura 3.3: Diferentes armónicos de una cuerda en vibración. (Desde Giancoli, D.C. Física para ciencias e ingeniería con física moderna, 2009.)

Considérese el caso de una cuerda estirada. Al sacudir dicha cuerda desde alguno de los extremos, se generará una onda estacionaria obteniendo un tono con la menor frecuencia resonante, a la que se le denomina **frecuencia fundamental**. Si se sacudiera dicha cuerda con el fin de aumentar la frecuencia, se generaría diferentes puntos intermedios que no oscilarían llamados **nodos** como se aprecia en la figura 3.3. En esta misma se puede visualizar que la longitud de onda del modo fundamental de la cuerda es igual al doble de la longitud de la cuerda L; así, las longitudes de onda de cada modo (λ_n) y las frecuencias correspondientes (f_n) quedan determinadas por las siguientes relaciones:

$$\lambda_n = \frac{2L}{n}; \quad f_n = \frac{nv}{2L}; \quad n = 1, 2, \cdots,$$

donde v es la rapidez de la onda sobre la cuerda. Debido a este factor, en un instrumento como la guitarra, se puede modificar el tono por cada cuerda.

Cuando se toca una nota musical en un instrumento, se está realizando una combinación de varios armónicos, esto a consecuencia del llamado **principio de superposición**, el cual establece que cuando dos o más ondas pasan a través de la misma región del espacio al mismo tiempo, el desplazamiento resultando en cualquier punto e instante es la suma vectorial de los desplazamientos individuales. En el caso de las ondas sonoras, el principio de superposición, el cual puede ser visualizado en la figura 3.4, permite crear sonidos complejos a partir de los sonidos más básicos, como los son los múltiplos del armónico fundamental.

Considerando el tono, la intensidad, la **duración** y el **timbre**, se tienen las cuatro cualidades esenciales de cualquier sonido. El último de estos ayuda a diferenciar dos sonidos ejecutados con dos instrumentos diferentes con la misma duración sobre el mismo

[6] Debido a la poca percepción que se tiene a cambios ligeros de frecuencia, un tono es generalmente asociado a un rango de frecuencias.

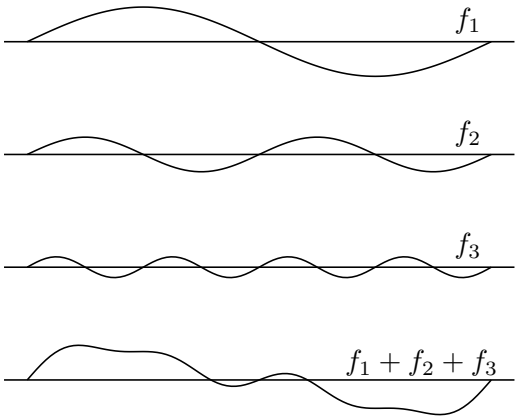


Figura 3.4: Principio de superposición aplicado en los primeros dos armónicos de una onda junto con el armónico fundamental.

tono en la misma intensidad, además de poder otorgar un "color" a diferentes tonos. Esta es una característica multidimensional que considera el comportamiento del sonido en toda su ejecución, como toda la distribución de energía en sus armónicos y la variación de la amplitud de la onda. Para más detalles puede consultarse [44].

3.2. Digitalización del audio

Teóricamente, el sonido es una señal en tiempo continuo, es decir una **señal analógica**, por lo que no es posible analizarla en un sistema digital, así que esta debe ser discretizada en una **señal digital**. Dicho procedimiento recibe el nombre de ADC (*Analog to digital conversion*) y consta de dos etapas: **Muestreo** y **Cuantificación**.

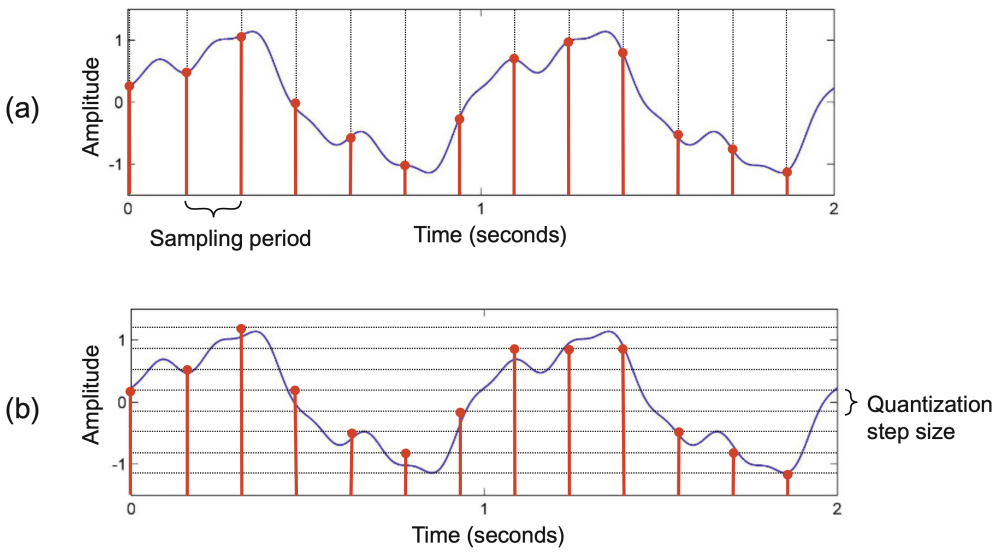


Figura 3.5: Etapas del proceso de digitalización de una señal análoga. (Desde Müller, M. Fundamentals of music processing: Audio, analysis, algorithms, applications. Springer, 2015.)

3.2.1. Muestreo

En esta etapa, la señal análoga es leída o muestreada en intervalos uniformes de tiempo; es decir que se mide la amplitud de la señal a una cierta **tasa de muestreo** $S_r = 1/T$, donde T es el periodo de muestreo. En la figura 3.6 se puede apreciar este proceso.

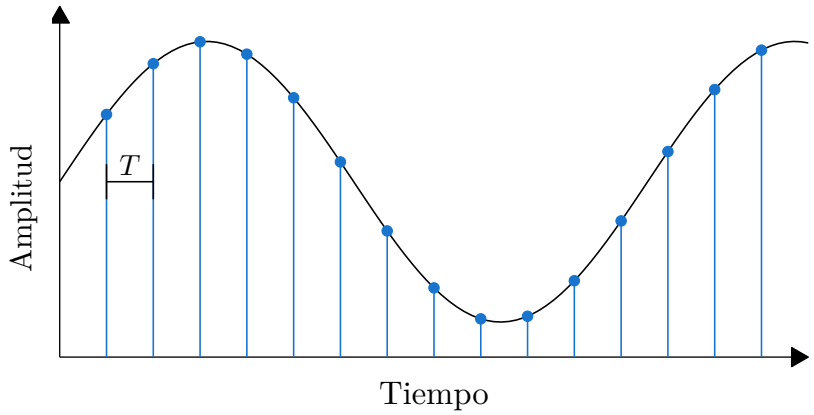


Figura 3.6: Ejemplo de un muestreo de una señal análoga. T representa el periodo de muestreo.

Solo por establecer un ejemplo, en un disco compacto las señales de audio que son almacenadas en él fueron muestreadas a 44,100 Hz. No es coincidencia que este número sea el seleccionado para dicho dispositivo; ya que, al tratar de discretizar una señal continua, se puede establecer el periodo de muestreo tan pequeño como se decida haciendo que se tenga más información y sea posible reconstruir la señal original de la manera más fiel. Esto puede ser un problema ya que se cuenta con una cantidad limitada de almacenamiento en los dispositivos digitales, por lo que se buscará establecer la tasa de muestreo más pequeña posible con la que se pueda reconstruir la señal análoga original.

Cuando la tasa de muestreo es muy pequeña, en la reconstrucción de la señal original, se pueden obtener componentes de la señal indistinguibles, haciendo que se pueda reconstruir una señal similar más no igual a la original. A este efecto se le conoce como ***aliasing*** y puede ser visualizado en la figura 3.7.

Con el fin de evitar el aliasing, se utiliza el **teorema de muestreo de Nyquist-Shannon** [9], [32], [43]; el cual establece que una señal puede ser reconstruida perfectamente si es muestreada al doble de la tasa más alta de frecuencia presente. Así obtenemos la **frecuencia de Nyquist** $f_N = S_r/2$, la cual indica la frecuencia límite de las señales que pueden ser perfectamente reconstruidas a un tasa de muestreo S_r. Para el caso de un disco compacto, $f_N = 44,100Hz/2 = 22,500Hz$ lo cual es cercano al límite del rango audible.

3.2.2. Cuantificación

En la siguiente etapa se buscará remplazar los valores que se obtuvieron en el muestreo, los cuales están codificados como números reales, por un rango discreto de valores, generalmente se utilizan bits; así, por ejemplo, si se consideran 4 bits para codificar las amplitudes, se tendrán $2^4 = 16$ posibles valores, tal como se muestra en la figura 3.8. En tal caso, diremos que se tiene una resolución de 4 bits.

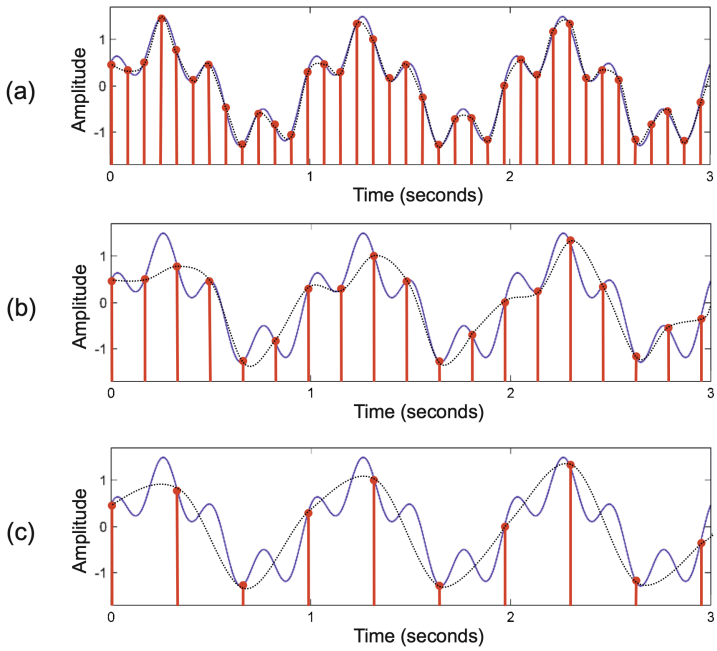


Figura 3.7: Efecto aliasing obtenido de realizar un muestreo a diferentes tasas menores a la ideal. En cada una de las sub gráficas, la línea solida es la señal análoga y la línea punteada es la señal reconstruida a las tasa de muestreo en (a) 12 Hz, (b) 6 Hz y (c) 3 .Hz. (Desde Müller, M. Fundamentals of music processing: Audio, analysis, algorithms, applications. Springer, 2015.)

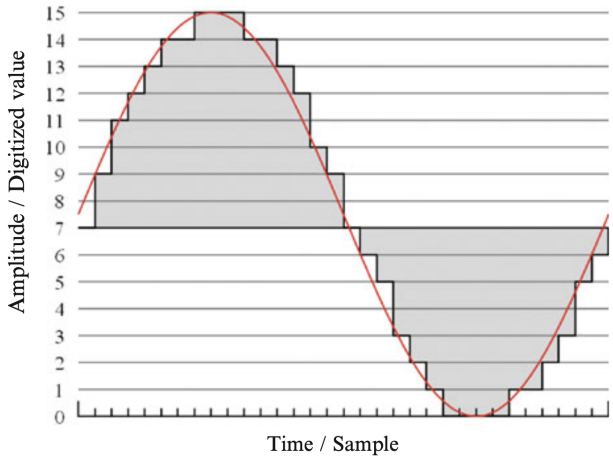


Figura 3.8: Ejemplo de cuantificación de una señal análoga. En este caso se están considerando 4 bits para la codificación (Desde Knees, P. and Schedl, M. Music similarity and retrieval: An introduction to audio-and web-based strategies. Springer, 2016.)

Al igual que en el muestreo, considerar valores discretos para describir una señal continua podría provocar problemas de interpretación y contenido; en este caso se puede tener un error de cuantificación, el cual es la diferencia entre el valor de la señal continua y el valor discreto. En la figura 3.8 este error se puede visualizar como la diferencia entre la curva roja y la gráfica escalonada de color negro. Este problema se soluciona estableciendo una resolución más grande. Por ejemplo, en el caso de los discos compactos, se considera una resolución de 16 bits, lo que permite una cuantificación con 65,536 valores discretos.

Lo visto en esta sección es utilizado cada vez que es grabado algún sonido. Por ejemplo, cuando un sonido es producido y captado por un micrófono, la presión es recibida y se crea una oscilación en el micrófono mediante una membrana que crea una señal eléctrica analógica; esta se canaliza a una tarjeta de sonido que actúa como un dispositivo ADC. Dicho dispositivo aplica un **filtro anti-aliasing**, el cual elimina todas las frecuencias que sobre pasan a la frecuencia de Nyquist establecida y aplica muestreo y cuantificación. Al final la señal es almacenada como una señal digital y esta puede ser reconstruida posteriormente.

3.3. Análisis de Fourier

3.3.1. Series de Fourier (FS)

Ya que todo sonido complejo puede descomponerse en múltiplos de su frecuencia fundamental, y estos son modelados mediante funciones senoidales, es necesario hacer una introducción sobre la transformada de Fourier; ya que con esta es posible obtener la descomposición del sonido en lo que se conoce como el espectro del sonido, el cual será una herramienta fundamental para el propósito de este trabajo.

Idealmente, lo que se desea analizar es una señal de audio en tiempo continuo, es decir una función $f : \mathbb{R} \to \mathbb{R}$. Un resultado de suma importancia para el análisis de señales establece que $f(x)$, una función como las de nuestro interés, puede ser escrita en términos de una **Serie de Fourier** (FS) bajo ciertas condiciones [7], en particular, así como lo son las funciones senoidales, si $f(x)$ es 2π-periódica, esta puede ser escrita de la siguiente manera:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos(kx) + b_k \sin(kx)), \tag{3.1}$$

donde a_k y b_k, llamados **coeficientes de fourier** de la función $f(x)$, están dados por:

$$\begin{aligned} a_k &= \tfrac{1}{\pi} \textstyle\int_{-\pi}^{\pi} f(x) \cos(kx) dx, \\ b_k &= \tfrac{1}{\pi} \textstyle\int_{-\pi}^{\pi} f(x) \sin(kx) dx. \end{aligned}$$

Estos coeficientes pueden ser interpretadas como las coordenadas que se obtienen de proyectar la función $f(x)$ en las bases ortogonales de senos y cosenos $\{\cos(kx), \sin(kx)\}_{k=0}^{\infty}$

[7] Es necesario que $f(x)$ sea periódica, es decir que $f(x) = f(x + \lambda)$ con λ el periodo, continua a trozos y acotada.

[9]. Esto debido a que los coeficientes a_k y b_x pueden ser calculados como el producto interno de la función $f(x)$ con las bases anteriores, es decir:

$$a_k = \frac{1}{||\cos(kx)||^2}\langle f(x)|\cos(kx)\rangle,$$
$$b_k = \frac{1}{||\sin(kx)||^2}\langle f(x)|\sin(kx)\rangle.$$

La integral del producto interior entre dos funciones puede ser utilizada como una medida de similaridad entre dichas funciones; así, al obtener altos valores en los coeficientes, las funciones senoidales en ciertas frecuencias, que representan tonos puros, indican un alto grado de similaridad entre la función o señal que estamos analizando. Un ejemplo de esto se puede visualizar en la figura 3.9 donde la función senoidal $4.2\sin(2\pi(250t))$ se ajusta a la nota C4, la cual tiene una frecuencia nominal de 261 Hz.

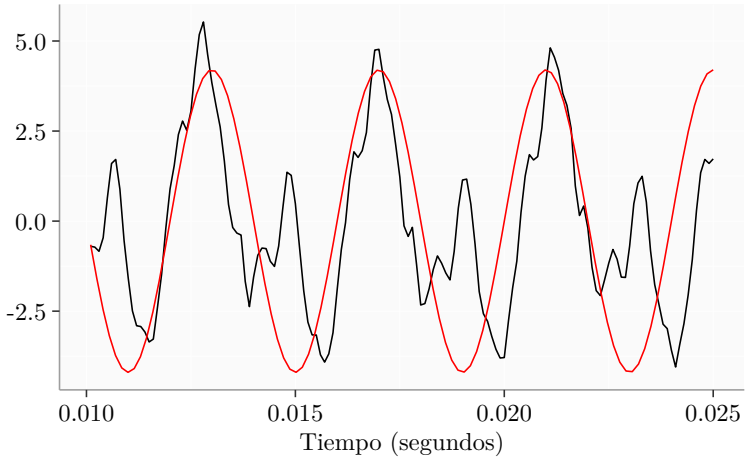


Figura 3.9: Señal de la nota C4 emitida por un piano comparada con una función senoidal con frecuencia de 250 Hz y una amplitud de 4.2.

Entonces, se deseará obtener los mejores ajustes de las funciones senoidales con la señal estudiada, esto para cada frecuencia. Para una frecuencia $\omega \in \mathbb{R}$, se definen

$$d_\omega = \max_{\phi\in[0,1)} \left(\int_{t\in\mathbb{R}} f(t)\cdot\sin(2\pi(ft-\phi))dt \right), \tag{3.2}$$

$$\phi_\omega = \arg\max_{\phi\in[0,1)} \left(\int_{t\in\mathbb{R}} f(t)\cdot\sin(2\pi(ft-\phi))dt \right), \tag{3.3}$$

donde d_ω representa la **magnitud**, la cual es la intensidad de la frecuencia ω dentro de la señal $f(t)$ y ϕ_ω, el coeficiente de **fase**, indica como debe ser desplazada la función seno en el tiempo para obtener el mejor ajuste con la señal anterior [43].

3.3.2. Transformaciones de Fourier (FT)

Basándonos en la ecuación (3.1), es natural considerar la identidad de Euler [8] en la serie de Fourier para que esta esté en términos complejos y así sintetizar la información;

[8] $e^{ikx} = \cos(kx) + i\sin(kx)$.

es decir, si consideramos la serie de Fourier en un dominio $[-L, L)$, la serie de Fourier queda expresada de la siguiente manera [9]:

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left[a_k \cos\left(\frac{k\pi t}{L}\right) + b_k \sin\left(\frac{k\pi t}{L}\right) \right] = \sum_{k=-\infty}^{\infty} c_k e^{ik\pi t/L},$$

donde los coeficientes están dados por

$$c_k = \frac{1}{2L}\langle f(t), \psi_k(t)\rangle = \frac{1}{2L}\int_{-L}^{L} f(t) e^{ik\pi t/L} dx.$$

Véase que la anterior representación involucra un conjunto discreto de frecuencias dadas por $\omega_k = k\pi/L$, por lo que, cuando $L \to \infty$, se está contemplando un rango continuo de frecuencias; en este caso, el producto interno[9] $\langle f(x), \psi_k(x)\rangle$ se convierte en una función en el dominio de la frecuencia $\hat{f} : \mathbb{R} \to \mathbb{C}$ conocida como la **Transformada de Fourier** (FT), denotada por $\hat{f}(\omega)$:

$$\hat{f}(\omega) = \mathcal{F}(f(t)) = \int_{-\infty}^{\infty} f(t) e^{-i\omega x} dx = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \omega x} dt.$$

Con base en lo anterior, se puede pensar que la transformada de Fourier es el límite de una serie de Fourier cuando la longitud del dominio se va al infinito; por lo mismo, no es necesario considerar funciones periódicas en la FT como sí en las FS, esto queda bien representado con la figura 3.10.

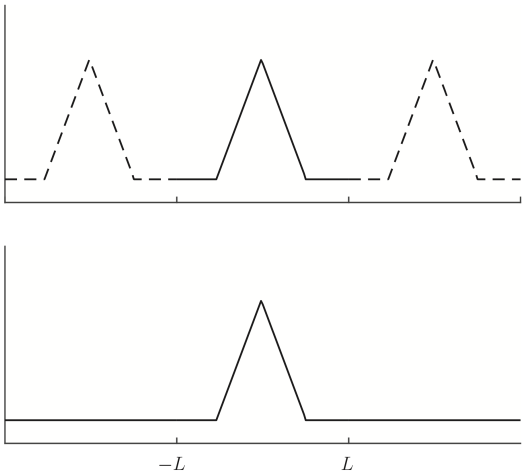


Figura 3.10: En la sub gráfica superior se da a entender que la serie de Fourier solo es válida para funciones periódicas; es decir, que cada cierto tiempo, se asume repetición. En el caso de la transformada de Fourier, esta es válida para funciones no periódicas (Desde Brunton, S.L. and Kutz, J.N. Data-driven science and engineering: Machine learning, dynamical systems, and control. Cambridge University Press, 2019).

Uno de los puntos más interesantes, además de las propiedades de la transformada de

[9]Por lo que podemos dar una interpretación análoga de la proyección de la función $f(t)$ pero ahora en el círculo unitario.

Fourier [10], es otra definición que podemos obtener de la transformada de Fourier. Se había mencionado que con el uso de los números complejos se lograba sintetizar la magnitud y la fase de las funciones senoidales en un solo número, este número puede ser calculado para cada frecuencia como:

$$c_\omega := \frac{d_\omega}{\sqrt{2}} e^{2\pi i(-\phi_\omega)}.$$

El anterior coeficiente utiliza las ecuaciones (3.2) y (3.3). La colección de estos coeficientes puede ser obtenida por una función compleja $\hat{f} : \mathbb{R} \to \mathbb{C}$, la cual asigna cada parámetro de frecuencia al coeficiente c_ω; es decir: $\hat{f}(\omega) := c_\omega$. La función $\hat{f}$ es nuevamente referida como la transformada de Fourier de f y a los valores c_ω se les conoce como **coeficientes de Fourier**.

Finalmente, $|\hat{f}(\omega)|$ será llamado la **magnitud** del coeficiente de Fourier. Este valor es utilizado en gráficas como la representada en la figura 3.11 donde se grafica la magnitud de los coeficientes con respecto a las frecuencias, así es posible detectar de manera sencilla aquellas frecuencias con mayor importancia en todo el espectro de frecuencias y determinar las componentes más importantes de la señal estudiada.

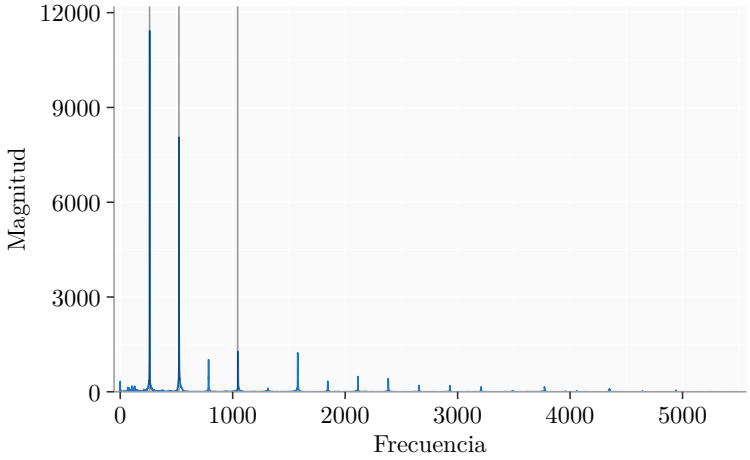


Figura 3.11: Espectro de frecuencias de una señal emitida por un piano en la nota C4. Las líneas grises ayudan a detectar la frecuencia fundamental (261 Hz) y los primeros dos armónicos (522 Hz y 1044 Hz) de este sonido.

En la figura 3.11, se puede visualizar que el coeficiente de Fourier con mayor magnitud es obtenido al rededor de 261 Hz, el cual, como ya se había mencionado, es la frecuencia modular de la nota C4. Es decir que con una función senoidal en dicha frecuencia, se tiene la mayor similaridad con señal emitida. A esta función senoidal se le denomina **parcial** y esta en particular se dice que se encuentra en la **frecuencia fundamental**. El tono de una nota musical es usualmente determinado por su frecuencia fundamental. Las restantes líneas grises en la gráfica reciben el nombre de **armónicos** (o armónicos parciales) los cuales son parciales que son obtenidos como un múltiplo de la frecuencia fundamental.

[10] La transformación de Fourier es un operador lineal, $\mathcal{F}\left(\frac{d}{dx}f(x)\right) = i\omega\mathcal{F}(f(x))$, $\int_{-\infty}^{\infty} |\hat{f}(\omega)|^2 d\omega = 2\pi \int_{-\infty}^{\infty} |f(x)|^2 dx$ y se tiene un comportamiento particular con la convolución de dos funciones: $(f \circ g)(x) = \int_{-\infty}^{\infty} f(x-\xi)g(\xi)d\xi$.

3.3.3. Transformaciones de Fourier Discretas (DFT)

Hasta el momento, con la transformada y la serie de Fourier, se han utilizado funciones continuas, pero como se mencionó en la sección 3.2, no se puede analizar directamente una señal análoga. Una versión discretizada de la transformada de Fourier para una señal en tiempo discreto $x : \mathbb{Z} \to \mathbb{R}$ está definida como [**muller2015fundamental?**] :

$$\hat{x}(\omega) := \sum_{n\in\mathbb{Z}} x(n)e^{-2\pi\omega n}.$$

Bajo este enfoque, existen aún dos problemas importantes ya que se tendrían que evaluar un número infinito de sumandos y se debería analizar la función anterior en el dominio de los reales, ya que ω es un parámetro continuo. En el primer caso, se puede asumir que la información relevante está limitada por la duración en minutos de la señal, como en el caso de las canciones; así, se asumiría que en una señal análoga, fuera del rango establecido por la duración empezando con $t = 0$, tendría un valor de 0. Con esto, podemos considerar un número finito de muestras $x(0), x(1), \ldots, x(N-1)$ con $N \in \mathbb{N}$.

Respecto al problema de la frecuencia, la solución radicará en considerar solo un número finito de frecuencias. De manera similar al muestreo del tiempo, en la frecuencia se suele realizar un muestreo de manera uniforme de acuerdo a una $M \in \mathbb{N}$ que determine la resolución, es decir: $\omega = k/M$ con $k \in [0, M-1]$. Generalmente se considera $M = N$ ya que esto garantiza que el resultado pueda ser invertido y sea computacionalmente eficiente [43]. Considerando los anteriores puntos, la **transformada de Fourier discreta** queda definida como

$$X(k) = \hat{x}(k/N) = \sum_{n=0}^{N-1} x(n)e^{-2\pi i n\frac{k}{N}}.$$

Un punto interesante de la anterior expresión es el hecho de que se puede obtener la frecuencia física, medida en Hz, correspondiente al índice k mediante la siguiente expresión:

$$F(k) := \frac{k}{N\cdot T} = \frac{k\cdot S_r}{N},$$

donde T y S_r son el periodo y la tasa de muestreo. Solo por considerar un ejemplo, la señal analizada en las figuras 3.9 y 3.11 contiene 29,750 muestras y fue muestreada a una tasa S_r = 11,025 Hz, por lo que la duración de la señal es de $29,750/11,025 = 2.698413$ segundos y así, la frecuencia analizada en $k = 1$ equivale a $F(1) = 11,025/29,750 = 0.3705882$ Hz.

De hecho, la figura 3.11 fue obtenida mediante los resultados de la aplicación de un algoritmo que implementa a la DFT llamado **transformada de Fourier rápida** (FFT). Este algoritmo reduce enormemente la complejidad computacional que tendría un algoritmo basado directamente de la definición de la DFT, ya que esté, de manera teórica, involucraría una matriz de $n \times n$ que requeriría $\mathcal{O}(N^2)$ operaciones. Con el algoritmo FFT, desarrollado en 1965 por James W. Cooley y John W. Tukey[11], la cantidad de operaciones se reduce a $\mathcal{O}(n\log(n))$.

[11] En realidad, el algoritmo fue formulado por Gauss en 1805 para aproximar las órbitas de los asteroides Pallas y Juno [9].

Todos los resultados obtenidos por el algoritmo FFT se pueden visualizar en la figura 3.12, donde se puede apreciar que la DFT tiene ciertas propiedades de simetría, lo cual sucede cuando $x(n)$ está valuada en números reales. Para evitar información redundante, se consideran solamente los coeficientes $X(k)$ con $k \in [0, N/2]$; de hecho, cuando N es par, el índice $k = N/2$ corresponde a $F(k) = S_r/2$, la cual es la frecuencia de Nyquist del proceso de muestreo.

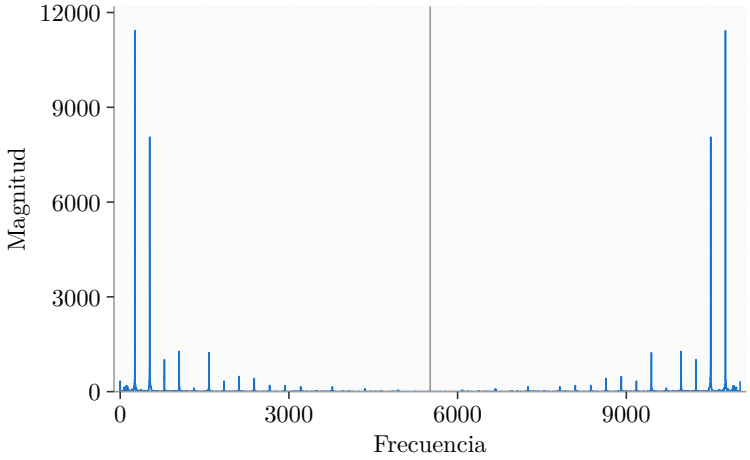


Figura 3.12: Espectro de frecuencias de una señal emitida por un piano en la nota C4. En este caso, la línea gris indica la frecuencia de Nyquist para esta señal.

3.3.4. STFT y Espectrograma

Con la transformada de Fourier o su versión discretizada, obtenemos información sobre la distribución de la frecuencia en la señal estudiada, pero se ha perdido la información de cuando dichas frecuencias tiene mayor presencial en la señal. Para solucionar dicho problema, en 1946 Dennis Gabor introdujo la **Transformada de Fourier de tiempo corto** (STFT) la cual considera pequeñas secciones de la señal para calcular la transformada de Fourier en dichos intervalos y así recuperar la información del tiempo.

Al considerar una pequeña sección de la señal al rededor de un punto t, esta es ponderada por una **función ventana** centrada en el mismo punto para todos los puntos de tiempo t que se consideren; usualmente estas funciones se eligen de tal manera que tengan una forma de campana, como un kernel gaussiano o la función **ventana de Hann** definida de la siguiente manera

$$g(u) := \begin{cases} (1 + cos(\pi u))/2 & \text{si } -0.5 \leq u \leq 0.5 \\ ,0 & \text{en otro caso.} \end{cases}$$

Entonces, la señal a la cual se le calculará la FT será descrita como

$$f_{g,t}(u) := f(u)g(u - t).$$

Por lo que, la Transformada de Fourier de tiempo corto es una función $f_g : \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ definida como

$$\tilde{f}_g(t,\omega) := \widehat{f_{g,t}}(\omega) = \int_{u\in\mathbb{R}} f(u)\bar{g}(u-t)e^{-i\omega u}du = \langle f, g_{t,\omega}\rangle,$$

donde en $\bar{g}$ se está considerando una función ventana $g : \mathbb{R} \to \mathbb{C}$. Un ejemplo de esto puede visualizarse en la figura 3.13.

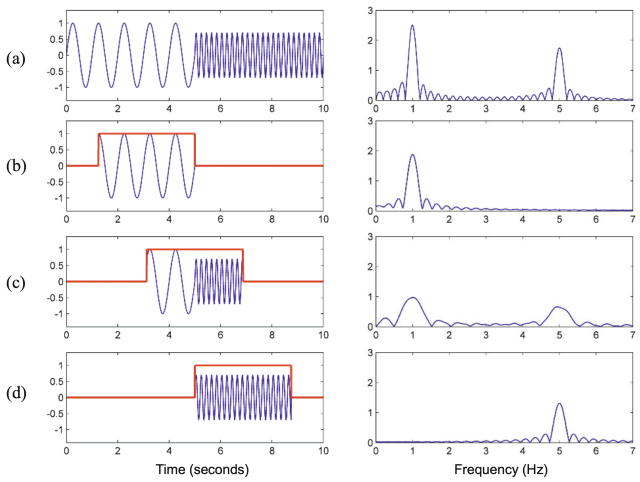


Figura 3.13: Interpretación de la transformada de Fourier en tiempo corto para una señal que consiste de dos funciones senoidales con frecuencias 1 Hz y 5 Hz. En (a), se muestra la FT de toda la señal; en el resto de sub gráficas, se obtiene la STFT con la señal aplicada a la función ventana rectangular centradas en (a) $t = 3$, (b) $t = 5$ y (d) $t = 7$ (Desde Müller, M. Fundamentals of music processing: Audio, analysis, algorithms, applications. Springer, 2015).

Nuevamente, lo anterior es considerando señales en tiempo continuo, por lo que se debe utilizar una versión discretizada para trabajar con la información que ha sido obtenida con la transformada de Fourier discreta. En este caso se debe discretizar el tiempo y la frecuencia.

Considere una señal muestreada por una tasa F_s dada en Hz, también una versión discretizada de una función ventana $w : [0, N-1] \to \mathbb{R}$; N determinará la cantidad de muestras consideradas en cada sección de la señal, por lo que cada sección (o frame) tendría una duración de N/F_s; a este parámetro se le conoce como ***frame size***. Otro parámetro necesario será H llamado ***hop size***, el cual determina la cantidad de muestras a la cual la ventana o sección se deberá mover provocando que existan traslapes entre las ventanas. Véase la figura 3.14 para un mejor entendimiento.

El parámetro H, junto con el uso de alguna función ventana, es importante ya que en caso de no considerarlo y tomar las ventanas sin que exista traslapes entre ellas, se puede generar un error conocido como **fuga espectral** (*Spectral Leakeage*), esto sucede por la desventaja de tomar frecuencias discretas en la DFT, ya que si la frecuencia de la señal no está en el conjunto de las frecuencias representadas en la DFT, no se puede aislar perfectamente la frecuencia de la señal.

Dicho error es común y solo puede ser eliminado en totalidad cuando una de las frecuencias consideradas en la DFT es exactamente igual a la frecuencia de la señal; en otros casos, este error solo puede ser reducido. En la figura 3.15 se puede ver el ejemplo donde no existe la fuga espectral, también cuando, por las razones anteriores, existe alguna discontinuidad y como su espectro es afectado después de aplicar alguna función ventana [52].

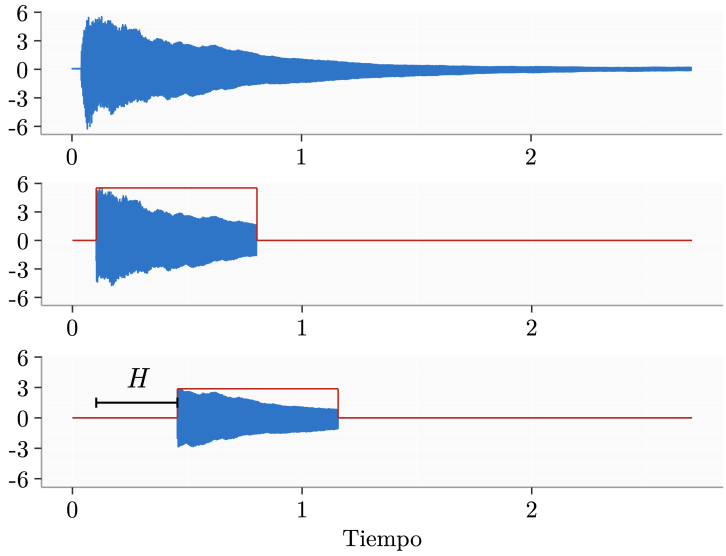


Figura 3.14: Uso de la función ventana rectangular sobre la señal correspondiente a la nota C4 emitida por un piano. H representa el parámetro hop size.

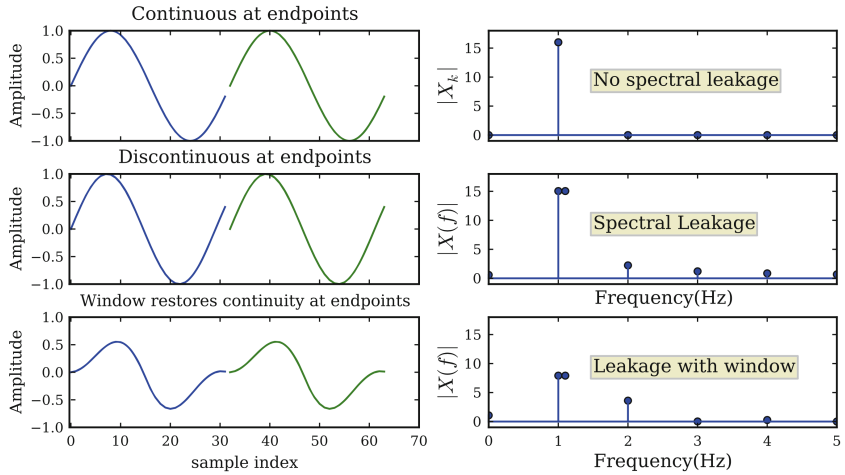


Figura 3.15: En la primera sub gráfica se tiene el caso donde no existe fuga espectral. Las gráficas intermedias muestran un caso clásico donde existe alguna discontinuidad provocando fuga espectral visible en el espectro de frecuencias. En la sección inferior, se aprecia la disminución de dicha fuga de energía con la aplicación de una función ventana (Desde Unpingco, J. Python for signal processing. Springer. 2016).

Con dichos parámetros, se define la **transformada de Fourier de tiempo corto discreta** (DSFTF) $\mathcal{X}$ de la señal x como

$$\mathcal{X}(m,k) := \sum_{n=0}^{N-1} x(n+mH)w(n)e^{-2\pi i n\frac{k}{N}},$$

donde $m \in \mathbb{Z}$ y $k \in [0, N/2]$ (considerando que N es par). $\mathcal{X}(m,k)$ es un número complejo, específicamente el k-ésimo coeficiente de Fourier para la m-ésima sección de tiempo. Véase que para cada sección de tiempo m, se obtiene un vector, llamado **vector espectral** de tamaño $N/2+1$, esta cantidad corresponde al número de secciones de frecuencia. Además, el número de secciones de tiempo es igual a (#muestras-N)/H + 1, por lo que al aplicar la DSFTF en una señal de tiempo discretizada se obtendrá una **matriz espectral** de dimensiones (# secciones de frecuencia, # secciones de tiempo) con los coeficientes de $\mathcal{X}(m,k)$. Respecto a las medidas físicas (tiempo en segundos y frecuencia en Hertz), se tienen las siguientes relaciones en base a los coeficientes obtenidos por $\mathcal{X}(m,k)$:

$$T(m) := \tfrac{m\cdot H}{F_s}; \quad F(k) := \tfrac{k\cdot F_s}{N}.$$

Una de las representaciones visuales más utilizadas con los resultados anteriores es el **espectrograma**, el cual es fácilmente obtenido con la siguiente relación:

$$\mathcal{Y}(m,k) := |\mathcal{X}(m,k)|^2.$$

En la figura 3.16 se pueden visualizar dos ejemplos del espectrograma aplicado a la señal que se ha utilizado continuamente en este capítulo. También es interesante ver que los espectrogramas tienen diferentes comportamientos de acuerdo a los parámetros considerados, esto se debe a que, de acuerdo a los valores de N y H, se obtendrá una mejor resolución de la frecuencia (con N) o del tiempo (con H) pero no ambos, por lo que se suele utilizar estos parámetros de acuerdo a la cantidad de información que se desee obtener. Los valores comunes que se utilizan en el análisis de audio son, para las secciones de tiempo: 512, 1024, o alguna otra potencia de 2. En el caso del hop size: 256, 512, 1024, alguna fracción de N, etc.

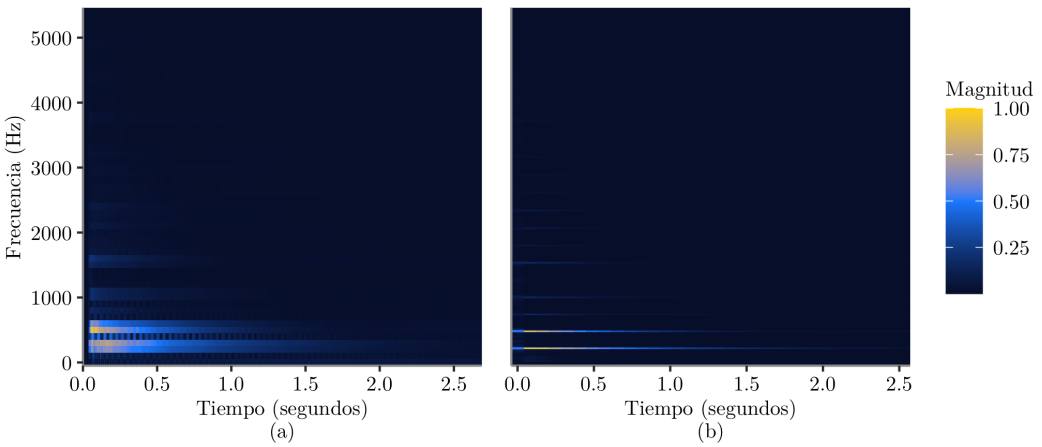


Figura 3.16: Espectrogramas de la nota C4. En (a): $N = 110$ y $H = 100$. En (b) $N = 1000$ y $H = 100$.

En las figuras anteriores, se puede visualizar que las frecuencias con mayor presencia en la señal analizada se encuentran antes de 1,000 Hz; recordemos que la señal analizada en este caso corresponde a la nota C4, con una frecuencia de 261 Hz. Al igual que en la figura 3.11, se pueden visualizar la frecuencia fundamental y el primer armónico, aunque en la figura 3.16 se aprecia como estas frecuencias se distribuye en el tiempo. La frecuencia fundamental tiene una mayor duración y energía que su primer armónico desde el inicio de la emisión, la cual va decayendo con el transcurso del tiempo.

3.4. MFCC

La figura 3.11 muestra que la mayoría de la información tiene muy poca magnitud o intensidad; esto se debe a que en dicha figura la escala de la intensidad es lineal y, como se había mencionado en la sección 3.1, la intensidad de una onda se entiende en términos logaritmos mediante la escala de los decibelios. Con dicha transformación, la figura 3.17 proporciona una mejor representación de la información musical y espectral de la señal estudiada, donde es posible ver la presencia de más armónicos presentes en la señal aunque con una menor sonoridad.

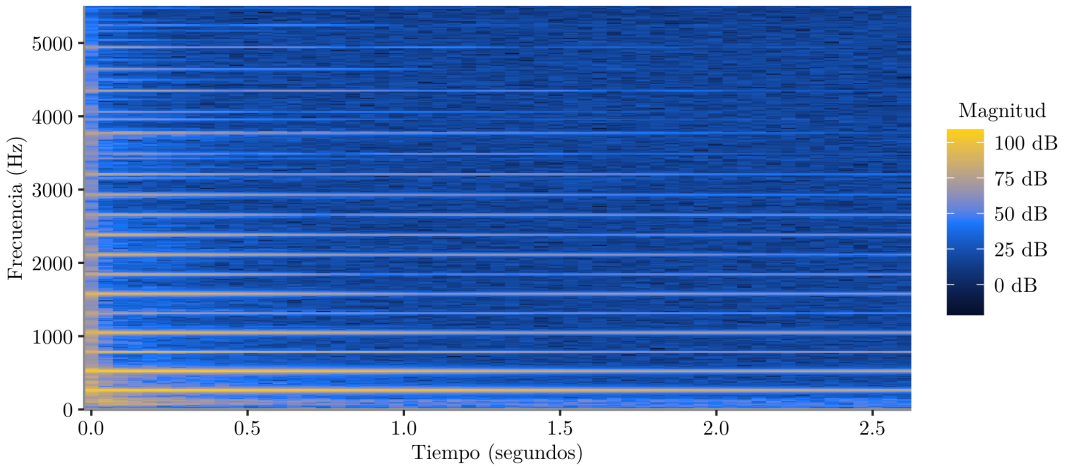


Figura 3.17: Espectrograma de la Nota C4 con escala en decibelios. $N = 1024$ y $H = 512$.

La frecuencia tiene un problema similar, ya que esta no es comprendida de manera lineal, sino de manera logarítmica mediante el tono; además, la percepción de los intervalos del tono no están linealmente correlacionados con los intervalos de frecuencia. Por lo anterior, se utilizan diferentes escalas que reflejan dicha relación no lineal mediante rangos que son juzgados como iguales por los humanos. Las escalas más comunes son la escala de Bark y la escala de **Mel**, las cuales fueron construidas mediante diversos experimentos auditivos. En este proyecto se considera solo la escala de Mel, la cual tiene la siguiente relación con la frecuencia f [32]:

$$m = 1127 \cdot \log\left(1 + \frac{f}{700}\right).$$

Aunque se tiene la anterior relación, la forma adecuada de transformar la escala de la frecuencia a escala de Mel es mediante los llamados **bancos de filtros**, ya que estos ayudan a tener una aproximación más robusta de la forma del espectro y concentran un

rango de frecuencias, o **bandas de frecuencias**, en un solo valor. Esto debido a que el ser humano no identifica los cambios entre las frecuencias adyacentes, sino hasta que dichas frecuencias se encuentran a ciertas distancias, las cuales aumentan con la frecuencia. Cada filtro o banda m pueden expresarse de la siguiente manera[12] considerando M bandas y que $f[x]$ es el centro de cada banda[13] [26]:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{(f[m+1]-k)}{(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases}$$

Estos bancos tienen filtros de forma triangular, llamados bandas de Mel, como en la figura 3.18. En estos se puede apreciar que algunos rangos de frecuencia medidos en Hertz equivalen a un solo valor en mels (estos son el centro de cada una de las bandas). Otro punto interesante es que cada frecuencia en esta escala está a una misma distancia, lo cual es el objetivo de esta escala.

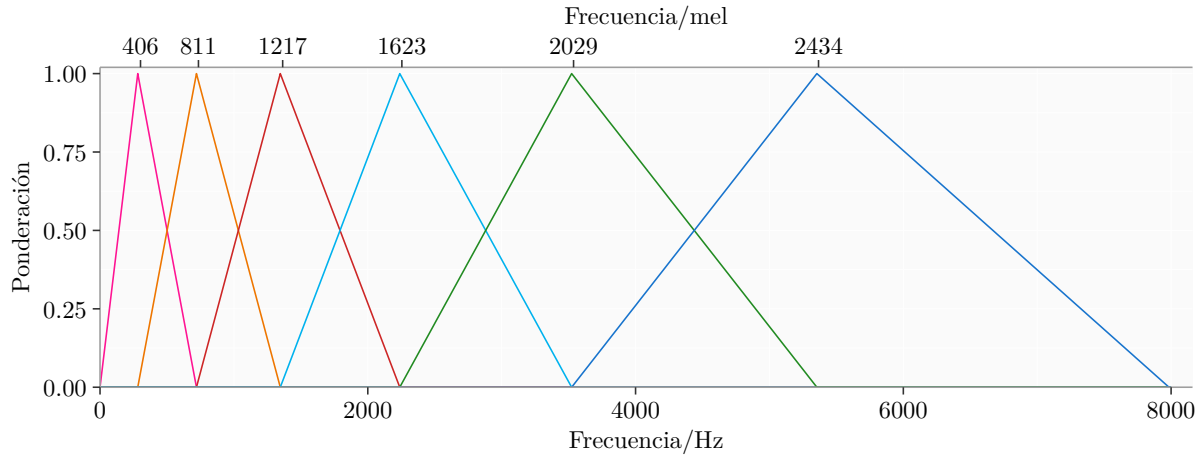


Figura 3.18: Banco de filtros con 6 bandas de Mel para la frecuencia de 10,000 Hz.

Para la construcción de las anteriores bandas es necesario conocer la frecuencia máxima f_h que se desea considerar (por ejemplo 44,100 Hz) y el número de bandas a construir, este último número dependerá del problema y puede ser considerado como un hiperparámetro para mejorar el modelo de acuerdo a la señal estudiada (generalmente se utilizan 40, 60, 90 u 128 bandas). La figura 3.18 solo es una representación visual ya que en realidad estos filtros quedan representados en una matriz M que será aplicada (multiplicada) al espectrograma. La matriz M tendrá un número de renglones igual al número de bandas de Mel y un número de columnas igual a número de secciones de frecuencia de la matriz espectral como resultado de la DSFTF.

Con lo anterior, se obtiene el **Espectrograma de Mel** que no es más que una matriz de dimensiones (# bandas de Mel, # secciones de tiempo) que tiene la información de

[12]En este caso, se está considerando que $\sum_{m=0}^{M-1} H_m[k] = 1$.

[13]Es decir: $f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_1) + m\frac{B(f_h)-B(f_1)}{M+1}\right)$ donde $B(f)$ es la frecuencia f en la escala de Mel, $B^{-1}(m)$ es la función inversa que lleva una frecuencia de la escala de Mel a hertz, f_1 la frecuencia más pequeña a considerar en el banco de filtros y f_h la más grande.

la señal considerando su duración y una escala de frecuencias como es percibida por el humano (en Hz[14]). El correspondiente espectrograma de Mel asociado a la figura 3.17 se puede visualizar en la figura 3.19, donde los colores (en una escala de decibelios) indican la fuerza de presencia que se consigue en una cierta banda de Mel en un cierto punto en el tiempo.

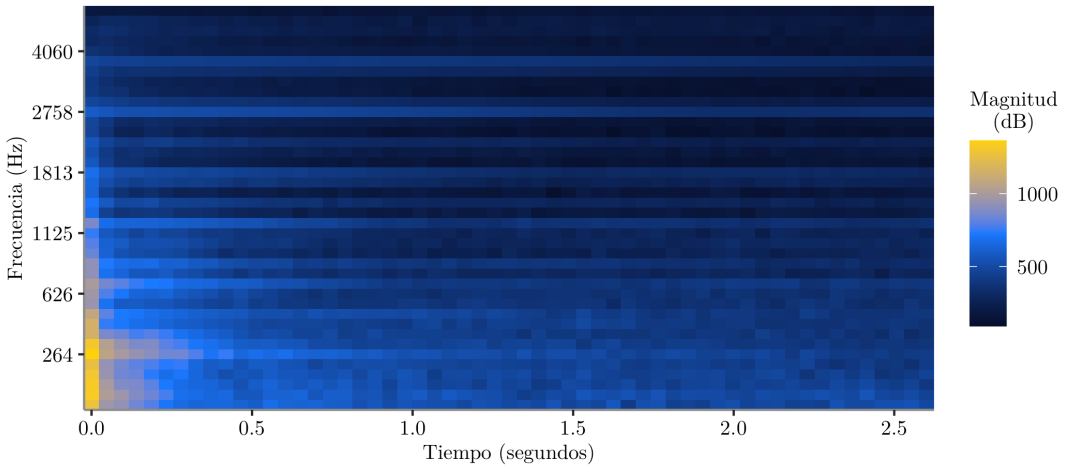


Figura 3.19: Espectrograma de Mel de la Nota C4 con escala en decibelios considerando el espectrograma de correspondiente a la figura 3.17 y 40 bandas de Mel.

Debido a la manera de construir los filtros, estos presentan un alto grado de correlación, lo cual puede repercutir drásticamente en la aplicación de ciertos modelos. Para solucionar esto, es común el uso de la **Transformada de coseno discreta** (DCT) ya que, además de eliminar la correlación producida, es utilizada como una función de comprensión, o de reducción de dimensionalidad tal como lo hace un método de componentes principales, concentrando la mayor información, en nuestro caso será la energía, en pocos coeficientes[15].

Esta transformación es una versión simplificada de la DFT utilizando solo números reales. La definición más general de la transformada de coseno discreta $F(k) : \mathbb{R} \to \mathbb{R}$ se puede encontrar en [12]; aunque, debido a que se opera en una secuencia discreta de puntos x_i y esto puede crear distintas condiciones de frontera (debido a problemas de paridad), se tienen distintos tipos de transformaciones. La más común, y la que será utilizada en este trabajo, es la transformada tipo 2 definida como:

$$F(k) = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right],$$

para $k = 0, 1, \ldots N - 1$.

En este punto es importante entender todo el procedimiento que se ha logrado para dar una interpretación a la información obtenida después de aplicar la DCT. Omitiendo la obtención del espectrograma, dada una señal que se desea analizar, después de pasar

[14]En la construcción de las bandas de Mel se llevan las frecuencias a la escala de Mel para ubicar los centros de dichas bandas y después se regresa, mediante una función inversa, a la unidad de medida de hertz; por lo que conseguimos considerar la frecuencia en otra escala pero en la misma unidad de medida.

[15]La motivación para la creación de esta transformada, propuesta en 1972 por Nasir Ahmed, es el hecho de que se puede demostrar que su conjunto base proporciona una buena aproximación a los eigenvectores de una clase particular de matrices [4].

un proceso de codificación, le es aplicada la transformada de Fourier discreta mediante el algoritmo FFT por secciones considerando traslapes entre estas y la aplicación de una función ventana (es decir, se aplica la STFT), obteniendo así el espectro de la señal respecto al tiempo y la frecuencia. Después le es aplicada una transformación logarítmica a dichos valores y nuevamente le es aplicada una transformación derivada de la transformada de Fourier. En resumen, podemos sintetizar lo anterior en la siguiente expresión:

$$C(x(t)) = \mathcal{F}\left\{\log\left(\mathcal{F}\left\{x(t)\right\}\right)\right\}.$$

A $C(x(t))$ se le conoce como el ***Cepstrum*** de la señal $x(t)$[16]. La palabra "*cepstrum*" es un anagrama de la palabra "*spectrum*", ya que al aplicar nuevamente una transformada $\mathcal{F}$, en este caso la DFT, se obtiene el espectro del espectro de la señal. Para entender mejor la aplicación de la última transformada véase la figura 3.20.

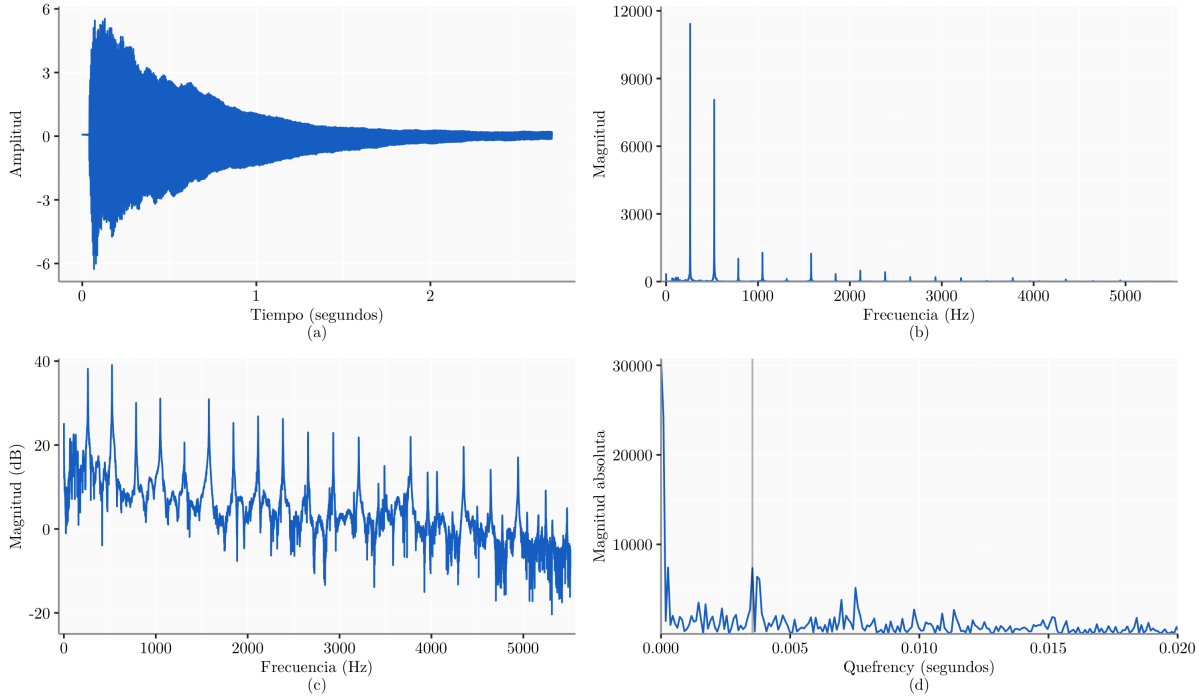


Figura 3.20: Representación de las estapas en la obtención del Cepstrum de la nota C4. La línea gris en (d) corresponde al primer rhamonico.

En dicha figura, se distinguen que las primeras dos sub gráficas corresponden a la señal en forma de onda y su espectro, en (c) se tiene el espectro aplicado a una transformación logarítmica con el fin de obtener una escala en dB; al hacer esto, se remueven los efectos multiplicativos resultando en componentes aditivas[17]. Véase que con esto se tiene una señal con componentes armónicas y comportamiento similar a la original y al obtener el espectro de esta se obtienen aquellas componentes con mayor importancia en la frecuencia. Ya que al aplicar la transformada en un dominio de la frecuencia (en la primera aplicación era en el tiempo), no se obtiene directamente esta, sino la denominada ***quefrency*** que

[16]Originalmente el cepstrum de una señal utiliza la transformada de Fourier inversa, es decir: $\mathcal{F}^{-1}\left\{\log\left(\left|\mathcal{F}\left\{x(t)\right\}\right|\right)\right\}$ pero se puede utilizar la anterior expresión con total libertad ya que son consistentes una con otra y la distribución de la frecuencia se mantiene igual, la única diferencia es un factor de escalamiento aunque es mejor utilizar la primera expresión [46].

[17]Recordar: $\log(a \cdot b) = \log(a) + \log(b)$.

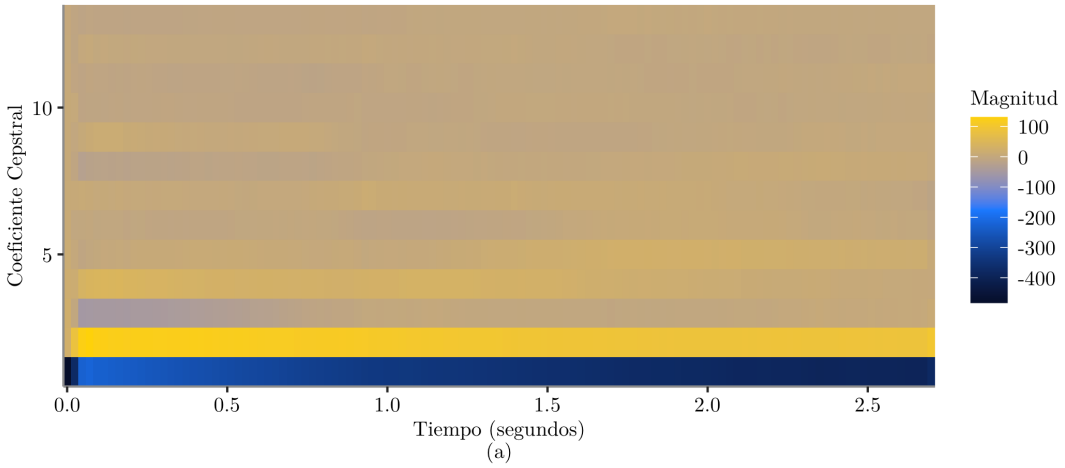
es una medida de tiempo aunque no con el mismo significado que el dominio de tiempo original[18].

En la misma figura, específicamente en (d), se puede ver el primer ***rhamonico***, el cual es el equivalente a la frecuencia fundamental de la señal original. Entonces podemos ver que la información de señal queda más condensada en el cepstrum que en el espectro y así, podemos interpretar al cepstrum como la tasa de cambio en las diferentes bandas del espectro. El cepstrum de una señal se diseño originalmente para el estudio de ecos en señales sísmicas y para la identificación de patrones en la voz[19]. En el año 2000, se comenzó a utilizar para el procesamiento musical.

Al igual que con el resultado de la STFT, podemos obtener un equivalente al espectrograma. En este caso, la potencia del cepstrum queda definida como:

$$C_p = |\mathcal{F}\left\{\log\left(|\mathcal{F}\left\{x(t)\right\}|^2\right)\right\}|^2.$$

Para conseguir lo anterior, basta con calcular el espectrograma de Mel y aplicar la DCT en cada columna y así obtenemos los **Coeficientes cepstrales en las frecuencias de Mel** o **MFCC**; los cuales, en el caso de la señal que ha sido estudiada en este capítulo, se pueden ver representados en la figura 3.21. En la aplicación de la DCT se debe determinar la cantidad de coeficientes cepstrales que se deseen obtener, regularmente 12 o 13 coeficientes son suficientes para capturar la mayoría de la información importante y el primero, regularmente, es omitido para análisis posteriores, ya que estos no contribuyen con información relevante. Es decir, que obtendremos una matriz donde los renglones representarán a cada uno de los coeficientes y las columnas el tiempo.



(a)

Figura 3.21: 13 MFCC de la nota C4.

Una de las principales desventajas que tienen estos coeficientes es la poca eficiencia que tienen para resumir información con el fin de obtener nuevamente la señal de audio aunque se ha visto que son muy adecuados para el procesamiento musical y el análisis de voz.

[18]Véase que las unidades de la *quefrency* son segundos, esto es debido a que la transformada inversa haría un mapéo en el tiempo.

[19]De hecho, al tener componentes aditivas, se puede descomponer la voz en la respuesta de frecuencia del tracto vocal y el pulso glótico y así se puede obtener las frecuencias a las que el tracto vocal se ajusta para producir distintos sonidos (el tracto vocal funciona como un filtro).

3.5. Otras estadísticas para resumir información auditiva

Lo anterior será utilizado para obtener vectores que representen a cada pieza musical, con lo que será posible comparar las canciones entre sí. Además, en este trabajo se consideraron otras alternativas comunes en el análisis musical para resumir la información. En este caso nos referimos a estadísticas de resumen enfocadas en el dominio de la frecuencia.

El primero de ellos es el **Centroide espectral** (SC) el cual representa el centro de gravedad del espectro; es decir, la banda de frecuencia donde la mayoría de la energía es concentrada. Muchas veces esta característica es utilizada como una medida de que tan "luminoso" es un sonido[20] ya que el timbre depende de la proporción de la intensidad de los armónicos, así como de los componentes de sonido similares al ruido [45].

Por ejemplo, una pieza de música clásica puede mostrar valores bajos y estables mientras que una canción del género electrónica o metal tendrá valores altos y fluctuantes, lo que evidencia la "claridad" o "brillo" del sonido [32]. Esta queda expresada de la siguiente manera:

$$SC_t = \frac{\sum_{n=1}^{N} m_t(n) \cdot n}{\sum_{n=1}^{N} m_t(n)},$$

donde $m_t(n)$ es la magnitud de la señal en el dominio de la frecuencia en la sección de tiempo t y la banda de frecuencia n.

Finalmente, el ancho de banda (BW) o **dispersión espectral** (SS), derivado del centroide espectral, indica el rango espectral de las partes interesantes de la señal; es decir, las partes alrededor del centroide. Puede interpretarse como la variación de la frecuencia media en la señal. El ancho de banda promedio de una pieza musical puede servir para describir su timbre percibido. Este queda definido como:

$$SC_t = \frac{\sum_{n=1}^{N} |n - SC_t| \cdot m_t(n)}{\sum_{n=1}^{N} m_t(n)}.$$

[20] El color del sonido resume el timbre.

Capítulo 4

Aplicación y resultados

Los capítulos anteriores tuvieron el objetivo de dar un marco teórico sobre los métodos de agrupamiento que fueron utilizados en este trabajo, así como ciertos detalles que podrían afectar los resultados, y una introducción al análisis de señales de audio mediante la aplicación de las variaciones de la transformada de Fourier hasta llegar a la obtención de los coeficientes cepstrales de Mel.

En este capítulo, se detallará todo el proceso que se realizó para la obtención de la información, su manipulación, la aplicación de los métodos de agrupamiento a esta base de datos y se mostrarán los resultados más relevantes, así como las abstracciones que se establecieron.

4.1. Construcción y características de la base de datos

La recolección de información es un tema relevante, ya que, de acuerdo a la información que pueda ser obtenida, la metodología tendrá que diseñarse. En este caso en concreto, lo que se buscaba era obtener archivos de audio de los que se pudieran extraer distintas características mencionadas en el capítulo 3. Esto presentó distintas dificultades por los siguientes motivos:

- La mayoría de las obras musicales requiere la solicitud de derechos mediante una remuneración económica.
- Diversas plataformas que proporcionan información de piezas musicales no otorgan la pieza musical digitalizada en ningún formato.
- Algunas bases de datos públicas no contienen una diversidad musical apropiada para este trabajo (es decir, que muchas de las bases de datos solo contenían obras musicales de muy pocos géneros o solo uno de ellos).

Tomando en consideración los puntos anteriores, y después de una búsqueda exhaustiva, las piezas musicales que se analizaron fueron obtenidas de la plataforma Free Music Archive (FMA), la cual fue fundada en 2009 y ofrece acceso gratuito a música original con licencia abierta; tal como se menciona en su página web oficial: *"Decenas de millones de visitantes cada mes descargan música para uso personal y muchos comparten y mezclan música de FMA en videos, podcasts, películas, juegos, aplicaciones e incluso proyectos escolares"*.

El contenido que está disponible en la plataforma ha sido proporcionado por artistas

independientes, colaboradores de otros medios como la radio, colectivos de artistas, museos y festivales de música con distintas restricciones de uso derivadas de las licencias que otorga la organización Creative Commons[1]. Para el uso del material otorgado en la plataforma FMA, se deben considerar las restricciones de cada tipo de licencia, toda la información utilizada está desglosada en el Apéndice 1, pero, en general, se puede hacer uso del material dando crédito al autor de su obra. En nuestro caso, no se está utilizando ninguna de las obras con fines de lucro o redistribuyendo las obras originales.

Ya que las piezas musicales que se encuentran en dicha plataforma no tienen la misma difusión que en otras plataformas populares de streaming enfocadas a la música, la información que se puede obtener de cada canción es limitada; de hecho, cada canción contiene la información de su autor, su nombre, álbum y un género musical[2], por lo que la única información registrada para este proyecto en la base de datos fue el enlace, por cada canción, para descargar el archivo con extensión mp3, esto con fines de optimización de recursos computaciones que más adelante serán explicados.

Para la creación de la base de datos, a la fecha de 1 de agosto del año 2021, se descargó la información pertinente de todas las canciones, disponibles hasta ese momento, mediante diversas técnicas de Web Scrapping utilizando funciones de los paquetes *request*, *BeautifulSoup*, *pandas* y *re* del lenguaje de programación Python de los géneros definidos en la plataforma como Soul RnB, International, Old Time/Historical, Folk, Country, Rock, Instrumental, Classical, Jazz, Hip Hop, Electronic, Blues y Pop; omitiendo así los géneros Experimental, Novelty y Spoken.

La información concentrada de dichos registros acumulaba más de 83,000 canciones. Por motivos de eficiencia y capacidad computacional, se consideró trabajar solo con una muestra de 300 registros seleccionada de manera pseudo aleatoria y sin remplazo obtenida por el método *sample* del paquete *pandas* utilizando una semilla igual a 92020.

4.2. Obtención de características y matrices de distancias

En el capítulo 3 se dio una introducción de diferentes técnicas para resumir información musical, las cuales son pocas considerando todas las metodologías que existen actualmente. Por esto, se decidió resumir cada canción considerando 4 enfoques distintos para comparar a cada uno de estos elementos desde diferentes perspectivas:

- Utilizando los coeficientes cepstrales de Mel.
- Enfoque similar utilizando el espectrograma.
- Utilizando el centroide espectral.
- Utilizando la dispersión espectral.

[1] De acuerdo a la información oficial de dicha organización, Creative Commons es una organización sin fines de lucro que ayuda a superar los obstáculos legales para compartir el conocimiento y la creatividad para abordar los desafíos apremiantes del mundo.

[2] Tal como se menciona en la información oficial de la plataforma: *a diferencia de otros sitios web, todo el audio ha sido seleccionado a mano por curadores de audio establecidos.* Por lo que se supone cierta consistencia en algunas características como el género.

A continuación se darán más detalles de lo anterior pero antes es importante mencionar que con cada enlace de la base de datos anterior, el proceso a seguir, independiente de la perspectiva de resumen que se esté considerando, fue descargar la canción y leerla con la función *load()* del paquete *torchaudio*[3] la cual identifica la tasa de muestreo a la que la canción fue previamente codificada y la onda es codificada en una serie de tiempo con valores estandarizados.

Obteniendo dicha tasa, con el propósito de homogeneizar todo el proceso y en caso de ser necesario, se realizó un re muestreo con la función *torchaudio.transforms.Resample()* para que todas las canciones tuvieran una tasa de muestreo de 44,100 Hz. Para evitar posibles problemas, en caso de tener una distribución en 2 canales[4], se calculó la media entre ambos canales y se trabajó con el audio monoaural. Con todo este tratamiento previo a cada canción, se obtuvo un resumen de cada una de ellas.

Todo lo anterior, y lo referente a esta sección, se ejecuto sobre diferentes máquinas virtuales proporcionadas por *Amazon Web Services* mediante su servicio *SageMaker* debido al alto costo computacional que implica realizar este tipo de procesos. Estos incluyen la descarga de la pieza musical, la carga de archivos, todas las operaciones realizadas en dichos elementos y la obtención de diferentes gráficos.

En el caso de los MFFC para cada canción, estos fueron obtenidos con la función *torchaudio.transforms.MFCC()* considerando un tamaño de ventana de 2048 muestras, un hop size de 512 muestras y el uso de la ventana Hann para el cálculo de la STFT. En cuanto a la cantidad de bandas de Mel se consideraron 128 y se estableció el número de coeficientes de Mel igual a 13.

Ya que al obtener los 12 coeficientes cepstrales más relevantes de Mel por cada canción se obtiene una matriz[5], con lo cual puede ser complicado tomar este tipo de elemento para ser comparado con otros, se consideró resumir toda la información de la matriz anterior en un solo vector, tal como se sugiere en [32], formado por la concatenación de las medias de cada coeficiente a lo largo de todas las ventanas de tiempo, junto con las varianzas y covarianzas distintas entre los 12 coeficientes.

En nuestro caso de 12 coeficientes, se consigue formar un vector de 90 entradas: las primeras 12 con las medias de cada coeficiente, las siguientes 12 por la varianza estimada de cada coeficiente y existen $12 \cdot (11)/2 = 66$ covarianzas distintas. Con estos vectores, se consigue resumir una canción sin importar su duración.

En este caso en específico, se decidió tomar lo distante que son dos piezas musicales con la función de distancia $D_{KL}(\mathcal{P}, \mathcal{Q})$ vista en el capítulo 2 donde

$$KL(\mathcal{P}||\mathcal{Q}) = \frac{1}{2}\left[\log\frac{|\Sigma_{\mathcal{P}}|}{|\Sigma_{\mathcal{Q}}|} + Tr(\Sigma_{\mathcal{P}}^{-1}\Sigma_{\mathcal{Q}}) + (\mu_{\mathcal{Q}} - \mu_{\mathcal{P}})^T\Sigma_{\mathcal{P}}^{-1}(\mu_{\mathcal{Q}} - \mu_{\mathcal{P}}) - d\right].$$

[3]Torchaudio es un paquete diseñado al procesamiento de señales y audio con PyTorch, el cual permite trabajar de manera más inteligente con información masiva permitiendo paralelismo, una mejor administración de memoria y un ecosistema compatible entre sus distintas librerías.

[4]Es común que existan muchas obras musicales en dos canales. Cuando esto sucede se está hablando de un sonido estereofónico o estéreo, el cual tiene la finalidad de otorgar una mejor experiencia auditiva. En caso de tener un solo canal, se considera que se tiene un sonido monoaural frecuentemente conocido como mono.

[5]Recordar que el primer coeficiente no es más que un promedio de toda la energía de los coeficientes, por lo que este es omitido generalmente cuando se analiza información musical.

Al realizar dicha función de distancia, se está asumiendo que las canciones pueden ser modeladas mediante un modelo de mezclas gaussianas (en realidad solo una distribución), con lo que se ha visto que se logra tener un comportamiento adecuado [32] [40] [49]. Véase, por ejemplo, la figura 4.1 donde se utiliza esta función de similaridad con distintas canciones de géneros variados.

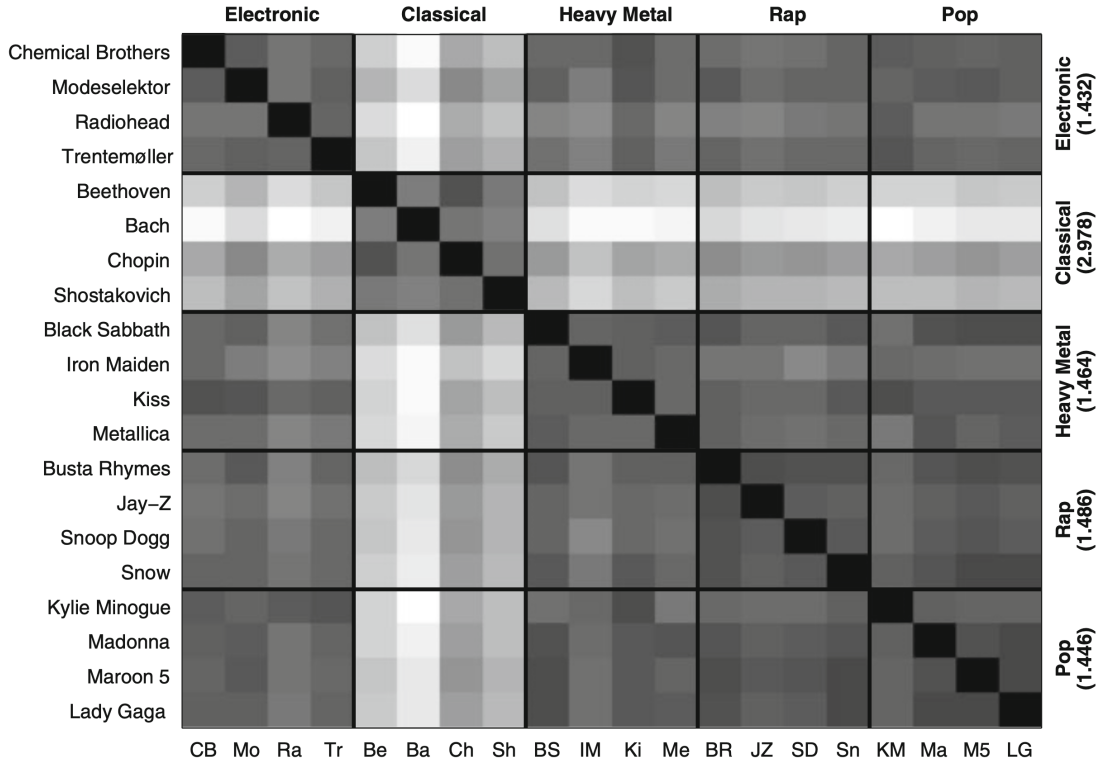


Figura 4.1: Ejemplo del uso de la función de distancia $D_{KL}(\mathcal{P}, \mathcal{Q})$ para diversas canciones utilizando los vectores resumen de los coeficientes cepstrales de Mel. Los colores más intensos representan mayor similitud (Desde Knees, P. and Schedl, M. Music similarity and retrieval: An introduction to audio-and web-based strategies. Springer. 2016).

Como un proceso final, se aplicó la anterior función a cada par de vectores resumen de cada canción para obtener una matriz de distancias con la cual trabajar en los métodos de agrupamiento.

Considerando este mismo enfoque de crear vectores resumen sobre una matriz, se propone hacer un proceso análogo con la matriz que representa al espectrograma. No se puede aplicar el mismo procedimiento directamente ya que en la matriz de los coeficientes de Mel se tiene un número fijo de coeficientes (renglones), mientras que en un espectrograma es variable la cantidad de renglones, por lo que se consideraron solo las frecuencias representadas en el espectrograma menores o iguales a 4,096 Hz[6] para, posteriormente, dividir todo el rango de frecuencias en 12 segmentos de igual tamaño donde se promediaron las frecuencias para formar cada bloque.

[6]Esto tiene dos justificaciones. La primera radica en que considerar frecuencias más grandes hace que se tengan, en promedio, más segmentos con valores negativos en el espectrograma lo que provocó problemas con los cálculos posteriores (especificamente en el cálculo del determinante). El segundo punto es el hecho de que la mayoría de las notas musicales hechas por distintos instrumentos, incluyendo las voces humanas, en general no sobre pasan una frecuencia igual a $2^{12} = 4,096$ Hz (las de un teclado común abarcan entre 27.50 Hz y 20.93 Hz).

Con estas matrices derivadas del espectrograma, se realizó la comparación de cada dos piezas musicales con la función D_{KL} hasta obtener la correspondiente matriz de distancias con las 300 canciones seleccionadas. Es importante mencionar que las funciones en este caso fueron *torchaudio.transforms.Spectrogram()* y *power_to_db()*; esta última del paquete *librosa*[7]. Referente a los parámetros, se consideró un tamaño de ventana de 2048 muestras, un hop size de 512 muestras y el uso de la ventana Hann.

En estos casos, la obtención de las estadísticas SC y SS son fácilmente obtenibles por las funciones *feature.spectral_centroid()* y *feature.spectral_bandwidth()* del paquete *librosa*; estas funciones fueron realizadas con los mismos parámetros que el caso anterior.

El único problema que surgió en este caso fue, que a diferencia de las matrices de distancias generadas por los vectores resumen de los MFCC y espectrogramas, el tiempo (el número de ventanas de tiempo) supeditaba al tamaño de los vectores y, por lo tanto, se tenían diferentes dimensiones para cada canción. Para solucionar dicho problema, se concatenaron ceros a todos los vectores resultantes hasta tener el mismo tamaño del vector cuya observación tenía la mayor duración. Por la naturaleza de las estadísticas SC y SS, el cero representa una ausencia de energía, por lo que es adecuado considerar que se tiene dicha energía cuando las piezas musicales hayan acabado.

En cuanto a las matrices de distancias generadas para estas estadísticas, se consideraron las distancias Manhattan, Euclideana y coseno; así la aplicación de los métodos de agrupamiento se realizó con matrices de distancias en lugar de las observaciones directamente para ahorrar cálculos internos. En resumen, se obtuvieron 8 matrices con las que se implementaron los métodos de agrupamiento.

4.3. Aplicación de métodos de agrupamiento y resumen de resultados

Para tener una mayor certidumbre de la presencia de grupos en los distintos conjuntos de vectores resumen, se aplicó la estadística de Hopkins a cada uno de los siguientes grupos y se obtuvieron los siguientes índices[8]:

Tipo de vector resumen	Valor promedio
Espectrograma	0.12
MFCC	0.17
SC	0.16
SS	0.15

Por lo que es estadísticamente factible realizar algún método de agrupamiento en los anteriores conjuntos de vectores resumen. El siguiente paso fue aplicar cada uno de los métodos de agrupamiento vistos en el capítulo 2 a los diferentes conjuntos de vectores resumen. Todos los resultados que se presentarán a continuación fueron obtenidos de la

[7]El paquete librosa es un paquete implementado para el lenguaje de programación Python enfocado al análisis musical.

[8]Las implementaciones comunes del índice de Hopkins necesitan como parámetro un número entero k para obtener el k-vecino que se tomarán en cuenta como el más cercano, este parámetro fue recorrido desde 1 hasta 299.

aplicación de distintas funciones del paquete *scikit-learn*, las cuales permiten ingresar como parámetro una matriz de distancias para realizar todas las implementaciones de los métodos.

Las métricas de rendimiento fueron obtenidas desde la misma paquetería a excepción del índice CVNN, ya que este fue obtenido de la función *cvnn()* del paquete *fcp* del lenguaje de programación R.

4.3.1. Métodos aglomerativos

Para los siguientes resultados, se utilizó la función *AgglomerativeClustering()* del paquete antes mencionado con diferentes parámetros sobre el tipo de enlace (promedio, único y completo) y diferentes números de grupos a buscar. El número de grupos en todos los resultados de esta y las demás secciones fueron seleccionados por obtener la mejor estadística en alguno de los cuatro criterios de rendimiento (los índices S, CH, DB y $CVNN$).

Metodología	Enlace	Número de grupos	S	CH	DB	$CVNN$
Espectrograma	Promedio	4	0.9211	3.3699	1.0634	1.3107
		5	0.8790	2.7359	0.8308	1.3105
	Completo	5	0.8823	3.0417	0.8364	1.5532
		10	0.1921	5.4067	2.4607	1.3041
	Único	8	0.8367	1.7793	0.7259	1.2807
MFCC	Promedio	6	0.8459	2.8125	1.9443	1.1654
	Completo	4	0.9081	2.3406	3.2826	0.7420
		16	0.1265	2.9567	2.4168	1.0211
	Único	4	0.9497	1.9767	1.4542	1.1860

Tabla 4.1: Resumen de la aplicación de diferentes métodos aglomerativos de agrupamiento en los vectores resumen basados en el espectrograma y los MFCC

Es interesante mencionar que en la mayoría de los resultados asociados con este tipo de método de agrupamiento se obtuvieron muchos grupos con pocos elementos y predominaba, regularmente, uno o dos grupos con más del 90 % de todas las observaciones a excepción de los resultados mostrados con el método Ward de las siguientes tablas. En los casos mostrados en la tabla 4.1, los mejores resultados, basados en el índice $CVNN$, se obtienen considerando 4 grupos en el método completo con los vectores de resumen MFCC, aunque con un grupo con más del 93 % de todas las observaciones.

En la tabla 4.2 se puede ver que, cuando se considera la métrica euclidiana, los índices tienen su mejor rendimiento, en especial con el índice CH. De hecho, con esta métrica se consiguen mínimos o máximos globales de los índices DB y CH si se realizan agrupamientos considerando desde 2 hasta 299 grupos (en algunos casos se consiguieron el segundo mínimo en DB) en los diferentes métodos. Además, se observó que la distribución de los grupos, en esta métrica, fue más homogénea que en los demás casos.

Es interesante ver que al aplicar la métrica euclidiana se favorece al índice CH pero, en general, es fácil ver que no se obtiene el mejor rendimiento en los índices con este método de agrupamiento, ya que la mayoría de los resultados correspondientes al índice $CVNN$ son altos, donde es importante recordar que este índice tiene una mayor precisión al determinar

Metodología	Enlace	Número de grupos	S	CH	DB	$CVNN$
SC. Métrica Euclidiana	Promedio	3	0.6110	12.1203	0.2623	1.9885
		10	0.3485	13.3230	0.6587	1.8754
	Completo	5	0.3553	21.5642	1.2738	1.8864
		8	0.3341	15.3374	0.8365	1.8859
	Único	4	0.5535	10.9663	0.2908	1.9800
		5	0.5143	10.0830	0.3080	1.9728
	Ward	3	0.1486	54.1985	2.1774	1.2535
		4	0.1576	44.8725	2.1295	1.3053
SC. Métrica Manhattan	Promedio	4	0.6284	10.4384	0.3077	1.9658
		7	0.5822	12.5943	0.9487	1.8462
		9	0.4442	11.1133	0.7413	1.8404
	Completo	3	0.5845	25.9616	1.1596	1.8268
		4	0.5670	20.2270	0.9916	1.8256
		6	0.4288	24.0242	1.3173	1.7176
	Único	7	0.5628	8.7773	8.0461	1.9305
SC. Métrica Coseno	Promedio	3	0.4190	12.3624	1.0103	1.9267
		5	0.3047	9.2635	1.4546	1.8889
	Completo	4	0.2740	29.1887	1.8892	0.9229
		5	0.2721	26.6650	1.5575	1.7661
	Único	3	0.4065	6.5809	0.6081	1.9952
SS. Métrica Euclidiana	Promedio	5	0.5106	14.7720	0.3902	1.9644
		7	0.3848	23.4858	0.7409	1.8635
	Completo	4	0.3177	49.9352	1.1292	1.8165
		8	0.2747	29.9688	0.8844	1.8048
	Único	5	0.5106	14.7720	0.3903	1.9644
		6	0.5072	12.0616	0.2944	1.9644
	Ward	4	0.1700	69.8008	1.7641	1.0356
		5	0.1757	68.1212	1.4383	1.8233
		6	0.1857	65.6906	1.4797	1.7901
SS. Métrica Manhattan	Promedio	3	0.6801	27.1900	0.8083	1.9017
		5	0.6425	16.0803	0.7316	1.9010
		6	0.5644	14.7378	0.6666	1.8919
	Completo	3	0.5891	38.6763	0.9841	1.8098
		4	0.5228	29.8142	1.2021	1.8061
		8	0.4728	24.3863	1.0240	1.7055
	Único	3	0.4728	20.3108	0.4399	1.9622
		5	0.6557	12.9081	0.2823	1.9486
SS. Métrica Coseno	Promedio	3	0.4521	15.3487	0.2773	1.9936
		7	0.3385	34.0088	1.1707	1.5709
	Completo	4	0.3687	49.2459	1.4895	0.8269
		5	0.3671	48.8419	1.2198	1.7710
	Único	4	0.5088	8.1171	0.6769	1.9831

Tabla 4.2: Resumen de la aplicación de diferentes métodos aglomerativos de agrupamiento en los vectores resumen basados en el centroide espectral considerando las distancias manhattan, euclidiana y coseno

la calidad de los grupos que los otros índices considerados. El mejor rendimiento en dicho índice se obtuvo al considerar la métrica coseno.

4.3.2. Método espectral

En el caso del método espectral, todas las matrices de adyacencia se construyeron en base a un gráfico totalmente conectado con un parámetro de escala establecido de manera local por el ancho de las vecindades, tal como se menciona en el capítulo 2. En cuanto a la implementación, se utilizó la función *SpectralClustering()* del paquete *scikit-learn* con una semilla igual a 21. Para la obtención de los mejores resultados, los cuales son reflejados en la tabla 4.3, se realizó un proceso iterativo donde se configuraron distintos números de vecinos (desde 2 hasta 299). Además se omitieron aquellos donde el número de grupos era pequeño (2 o 3) ya que se busca un mayor número de grupos en donde la cantidad de elementos en la mayoría de los conglomerados no sea escasa.

Metodología	k vecinos	Número de grupos	S	CH	DB	$CVNN$
Espectrograma	2	7	0.2208	7.8686	2.9487	1.195018
	271	6	0.1871	3.5243	5.7740	1.4576
MFCC	14	5	0.2814	5.0005	3.6110	0.4476
	60	7	0.1679	10.3460	3.8373	0.6638
	144	7	0.1761	10.0175	3.7098	0.5272
	73	10	0.1200	7.2270	4.2760	0.6273
SC. Métrica Euclidiana	2	7	0.0046	28.5389	1.9589	1.2675
	6	6	0.0298	33.5627	1.9675	1.1568
	196	4	0.1094	44.9990	2.0840	1.2886
	288	4	0.1286	47.0754	2.0282	1.2298
SC. Métrica Manhattan	2	4	-0.0216	33.0101	1.9356	1.0248
	12	4	0.0182	36.4016	2.0417	0.8972
	153	4	0.2165	45.6473	2.0808	1.1362
SC. Métrica Coseno	2	4	0.2694	30.2988	2.0081	0.8849
	2	5	0.2390	28.7821	2.1679	0.8465
	15	4	0.2637	34.6032	2.2890	1.0070
SS. Métrica Euclidiana	2	4	0.09193	54.7963	1.5857	0.9331
	10	4	0.1693	71.3705	1.6146	0.8879
	275	4	0.1840	72.0484	1.6385	1.0746
	293	6	0.1835	55.8989	1.8447	1.4113
SS. Métrica Manhattan	68	4	0.2063	74.1824	1.6221	0.9111
	281	7	0.3688	43.7183	1.2288	1.8356
SS. Métrica Coseno	3	5	0.3400	50.3565	1.6654	0.7070
	11	4	0.3167	59.3640	1.7492	0.7853
	11	8	0.2860	36.8600	2.0451	0.6073

Tabla 4.3: Resumen de la aplicación del método espectral de agrupamiento en todos los vectores resumen considerando diferentes funciones de distancias.

El uso de los vectores resumen basados en los MFCC son mejores en comparación con los obtenidos con las restantes metodologías, resaltando los casos con 5 y 7 conglomerados

considerando 14 y 144 vecinos, respectivamente, con la metodología de los MFCC, donde se obtuvieron los dos valores más bajos del índice $CVNN$ tomando en cuenta todos los resultados de este proyecto. Los siguientes resultados importantes son los obtenidos con la métrica coseno, en especial donde se utilizaron los vectores resumen basados en la dispersión espectral. En este caso también se obtienen valores pequeños del índice $CVNN$ y se obtienen los mejores valores correspondientes a los índices S y CH en comparación a los del caso anterior.

Analizando de manera mas exhaustiva los grupos con mejores resultados numéricos, se puede distinguir una clara diferencia entre los conglomerados formados por la metodología de los MFCC y la dispersión espectral, ya que estos últimos no muestran alguna segregación auditiva clara agrupando canciones de manera diversa sin que algún género, instrumento, calidad del sonido o con piezas auditivas que comparten segmentos idénticos (en nuestra muestra se encuentran dos elementos que comparten la misma introducción por unos segundos y el resto del contenido es diferente entre estas piezas. Estas observaciones si se encuentran en el mismo grupo en cualquiera de los resultados obtenidos por MFCC) tengan una clara diferencia entre los elementos de los distintos grupos.

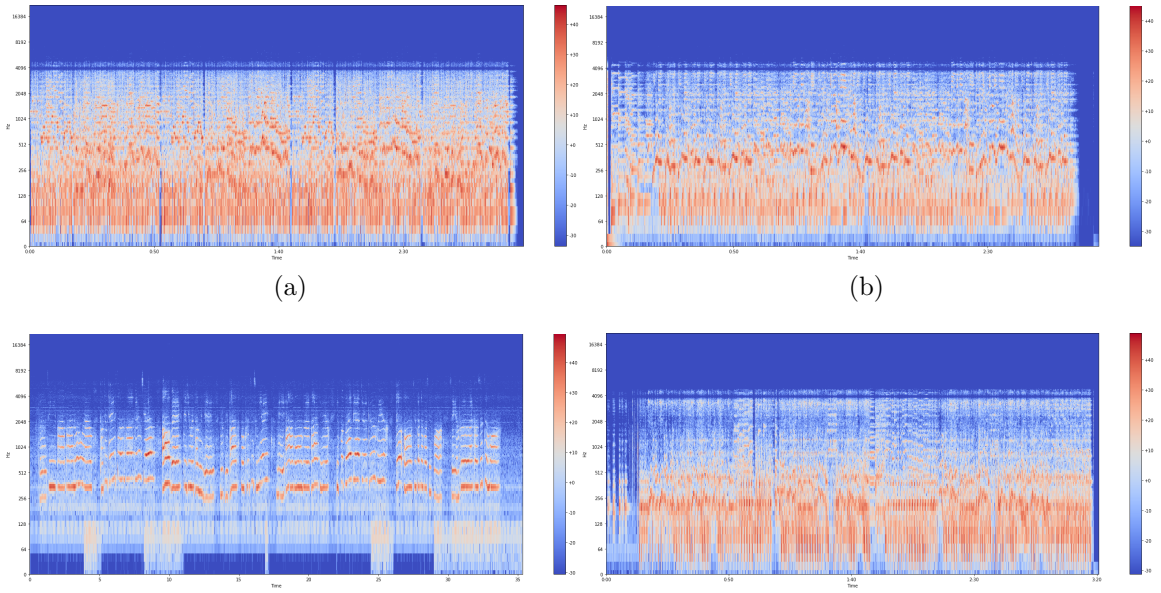
Estos resultados cambian considerando los conglomerados formados por la metodología MFCC ya que, en los casos específicos con 14 y 144 vecinos, se consiguen grupos con características comunes auditivas. Consideremos el conglomerado que contiene solo 5 grupos. En este se pueden percibir las siguientes similitudes entre la mayoría de las canciones de cada grupo:

- Grupo 1: Canciones del género pop y Old Time/Historical, música instrumental, canciones acústicas (voz con pocos instrumentos o solo escasos instrumentos) y canciones del genéro electrónica suaves.
- Grupo 2: Canciones de distintas variantes del rock y canciones del género electrónica con énfasis en sonidos de percusión.
- Grupo 3: Canciones que presentan estática o ruido de fondo. Este grupo contiene 6 canciones.
- Grupo 4: Piezas musicales con solo piano (1 de ellas se realizó con un órgano), del género Old Time/Historical con ruido de fondo y electrónica. Este grupo contiene 12 elementos.
- Grupo 5: Canciones realizadas con algún instrumento, como la guitarra acústica, con ruido de fondo, muy pocas sin este ruido, y algunas canciones del género Old Time/Historical con estática de fondo. Este grupo contiene 14 elementos.

Algo que es importante mencionar es el hecho de que en la mayoría de todos los grupos formados, independientemente de la metodología aplicada, se tienen diversos elementos del género electrónica asemejando sonidos distorsionados. Otro punto interesante radica en que la configuración de los últimos 3 grupos se mantiene (es deir que se tienen las mismas canciones), generalmente, como grupos aislados en todos los resultados que utilizan los MFCC. De hecho, el conglomerado con 7 grupos y 60 vecinos contiene estos 3 grupos de manera idéntica.

Otras características comunes pueden ser visualizadas desde los espectrogramas de las canciones que forman cada grupo. Por ejemplo, para el grupo 3, más del 80 % de las

canciones son algún tipo de canto con estática de fondo y para algunas de estas la figura 4.2 muestra sus espectrogramas. Para las canciones (a), (b) y (d) es posible visualizar que existen zonas con una concentración de energía predominante pero no tan elevada para resaltar en las canciones, lo cual hace que la voz quede atenuada por el ruido de fondo. En el caso de (c), la voz tiene una alta concentración de energía pero gran parte del resto del espectro tiene la presencia suficiente para agregar un fondo con sonido de estática.



(a) (b)

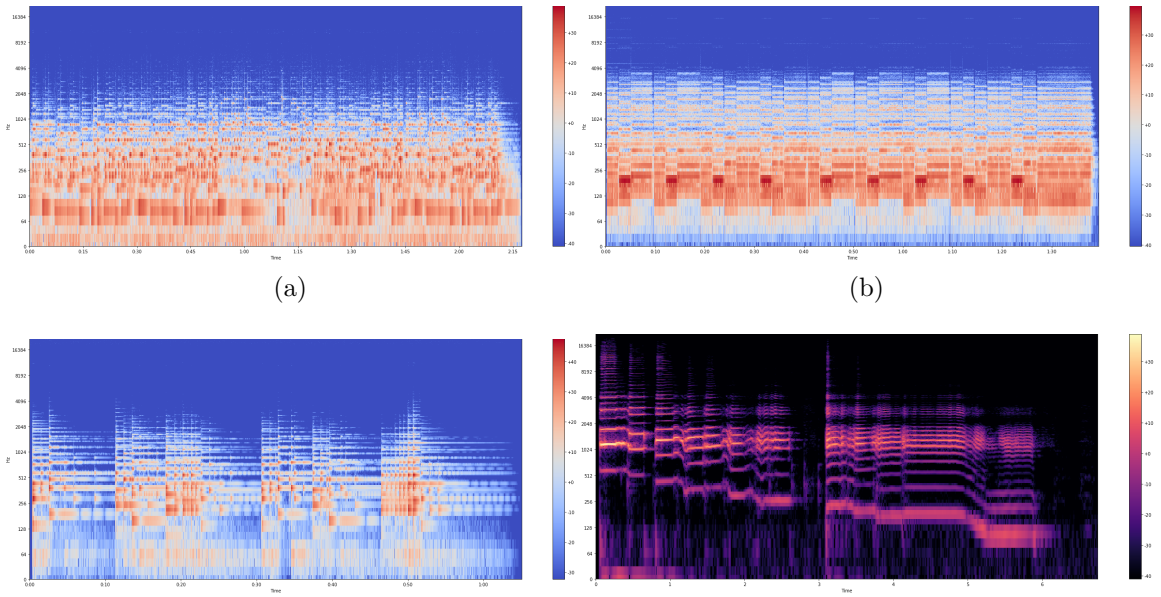
(c) (d)

Figura 4.2: Algunos espectrogramas del grupo 3 respecto al conglomerado formado por el método espectral con la metodología MFCC considerando 14 vecinos y 5 grupos.

Respecto al grupo 4, un poco más del 50 % de todas las canciones son efectuadas por solo un instrumento y en la figura 4.3 podemos notar algunos ejemplos de estos. En el caso de (a) y (c), se tienen canciones realizadas totalmente con un piano, (b) fue creada con un órgano y (d) con una trompeta. A pesar de los diversos instrumentos, es fácil distinguir la concentración y dispersión de energía en notas especificas y sus armónicos a medida que transcurre el tiempo.

El último grupo tiene una gran presencia de estática en sus canciones. En la figura 4.4, específicamente en las gráficas (a), (b) y (c), podemos ver los espectrogramas de algunos ejemplos de canciones donde se utiliza la voz y existe ruido de fondo; en estos se puede notar que la voz tiene poca energía sobre todo el espectro correspondiente haciendo que el ruido de fondo tenga tanta presencia como este elemento auditivo. En (d) y (e) se muestran ejemplos donde se utiliza alguna guitarra y existe ruido de fondo; en el caso de (e) las zonas con mayor energía solo corresponden a ciertos sonidos producidos por la guitarra. Finalmente, en (f) se tiene un ejemplo donde muchos sonidos son reproducidos a la vez con una intensidad similar.

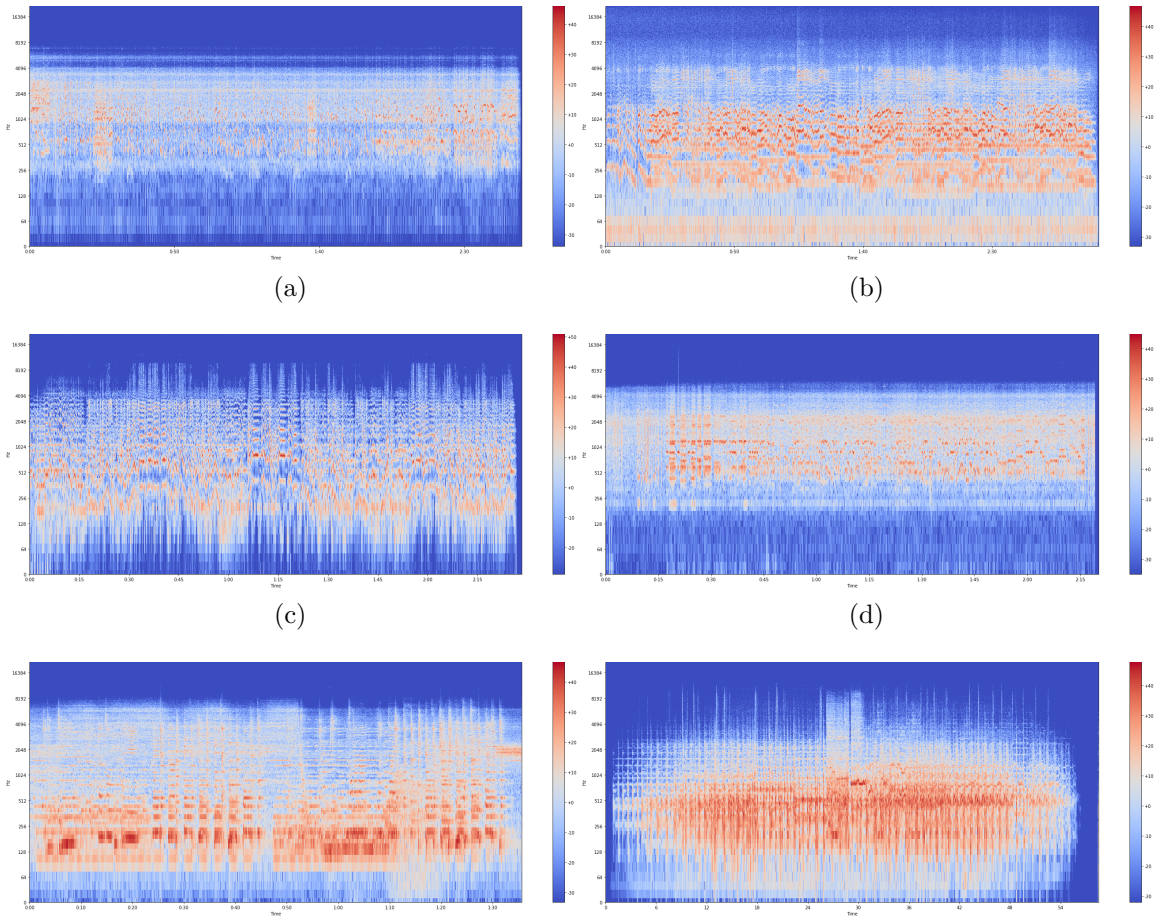
En el siguiente conglomerado, el cual tiene un comportamiento entre los grupos similar al obtenido con 7 grupos y 60 vecinos, se pueden apreciar, nuevamente, los últimos tres grupos del caso anterior salvo con muy pocas modificaciones (el grupo 3 permaneció intacto, el grupo 4 cambio una canción por otra con un comportamiento similar y el último grupo añadió 4 canciones extra de las cuales 2 tienen una gran presencia de algún tipo de guitarra). El resto de los grupos formados tienen las siguientes características importantes:



(a) (b)

(c) (d)

Figura 4.3: Algunos espectrogramas del grupo 4 respecto al conglomerado formado por el método espectral con la metodología MFCC considerando 14 vecinos y 5 grupos.



(a) (b)

(c) (d)

(e) (f)

Figura 4.4: Algunos espectrogramas del grupo 5 respecto al conglomerado formado por el método espectral con la metodología MFCC considerando 14 vecinos y 5 grupos.

- Grupo 4: Grupo donde la mayoría de las obras pertenecen a los géneros blues, electrónica y rock las cuales contienen una gran presencia de instrumentos de percusión. Este grupo contiene canciones del grupo 2 del caso anterior añadiendo 2 canciones extra que comparten las características aquí mencionadas. En total, este grupo tiene 18 elementos.

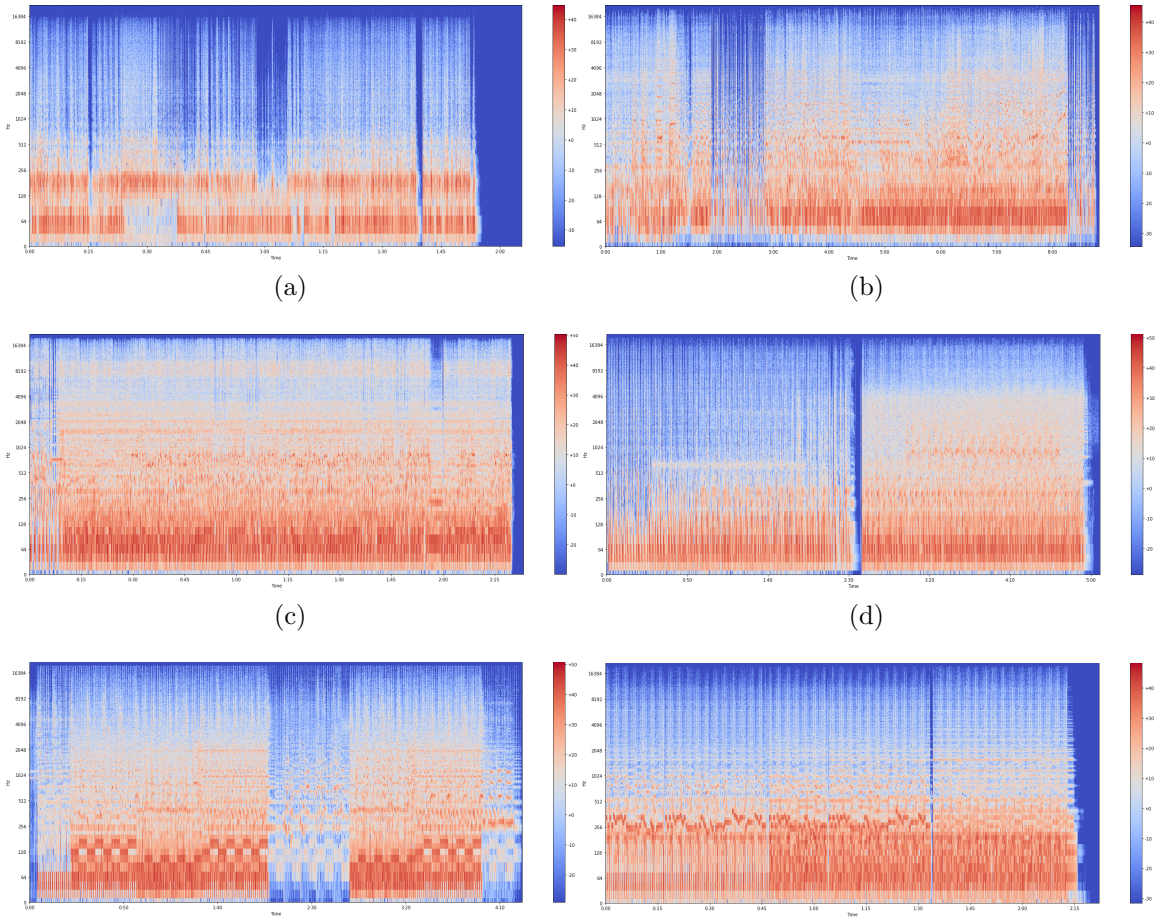
- Grupo 5: Grupo donde la presencia de diferentes variantes del rock caracterizan a un gran porcentaje de las canciones. Existen pocas pistas del género electrónica y música con distorsión además de pocas canciones efectuadas por escasos instrumentos. Podría pensarse que este grupo, junto con el anterior descrito, caracterizan al grupo 2 del caso anterior pero separando de manera adecuada las observaciones donde los sonidos de percusión aportan un mayor contenido energético a la canción.

- Grupo 6: Grupo donde la mayoría de las canciones solo han sido realizadas con pocos instrumentos (acústicas) y donde algunas de ellas son acompañadas con pocas voces (generalmente una). Además de obras con solo el uso del piano, órgano, guitarra, etc., se tienen algunas del género Old Time/Historical, electrónica con ruido de fondo y jazz.

- Grupo 7: El último grupo es el más diverso en cuanto a géneros musicales ya que se pueden encontrar canciones del género pop, rock, electrónica con y sin ruido y hip-hop, además de canciones hechas por escasos instrumentos. Considerando este y el previo grupo descrito, se tiene el comportamiento del grupo 1 del caso anterior.

Respecto a los espectrogramas generales de los grupos, la figura 4.5 contiene algunos de estos respecto al grupo 4. En las sub gráficas (a) y (b) se tienen los respectivos a dos canciones meramente efectuadas por una batería, en la primera de estas se pueden observar dos segmentos en el rango de las frecuencias que predominan por su energía, donde la banda inferior corresponde al sonido de un bombo y la superior a la efectuada por un redoblante y en (b) se observa solo la presencia de un tambor. Las sub gráficas (c) y (d) corresponden a canciones del género rock donde la batería tiene una presencia notable, lo cual se puede apreciar en bandas inferiores de frecuencia con gran potencia. Finalmente, en las restantes, se tienen dos ejemplos de música electrónica donde existe un fuerte sonido de percusión, por ejemplo en (f), una canción infantil, se hace uso de un tambor en toda la canción.

En la figura 4.6 se pueden apreciar algunos ejemplos del grupo 5. Los ejemplos (a) y (b) corresponden a canciones del género rock y rock 'n roll donde se aprecian varios rangos de frecuencia con alto contenido energético. Canciones de estos géneros musicales y algunas variantes tienen un comportamiento, en general, similar a los presentados y estas canciones abarcan un poco más del 70 % de todo el grupo. Las canciones respectivas a (c) y (d) son ejemplos donde varios instrumentos o un instrumento componen toda la canción con ausencia de voz; en (d) se muestra una canción dónde solo se tiene el sonido de un acordeón (en la muestra tomada solo existen cuatro canciones con este estilo y dos se encuentran en este grupo; en el conglomerado que considera 7 grupos y 60 vecinos, estas cuatro canciones están en el mismo grupo). Finalmente, en (e) y (f) se muestran ejemplos de música electrónica o pura distorsión de sonidos.

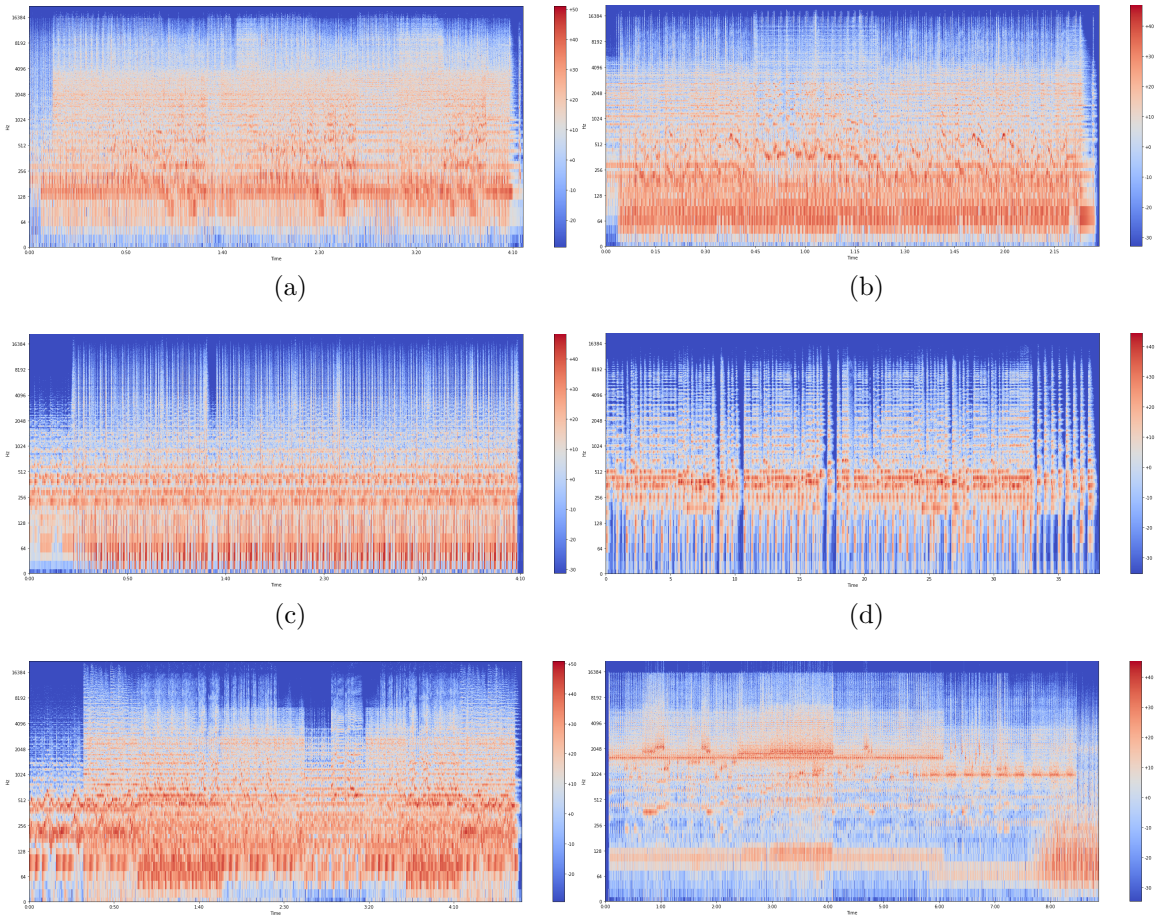


(a) (b) (c) (d) (e) (f)

Figura 4.5: Algunos espectrogramas del grupo 4 respecto al conglomerado formado por el método espectral con la metodología MFCC considerando 144 vecinos y 7 grupos.

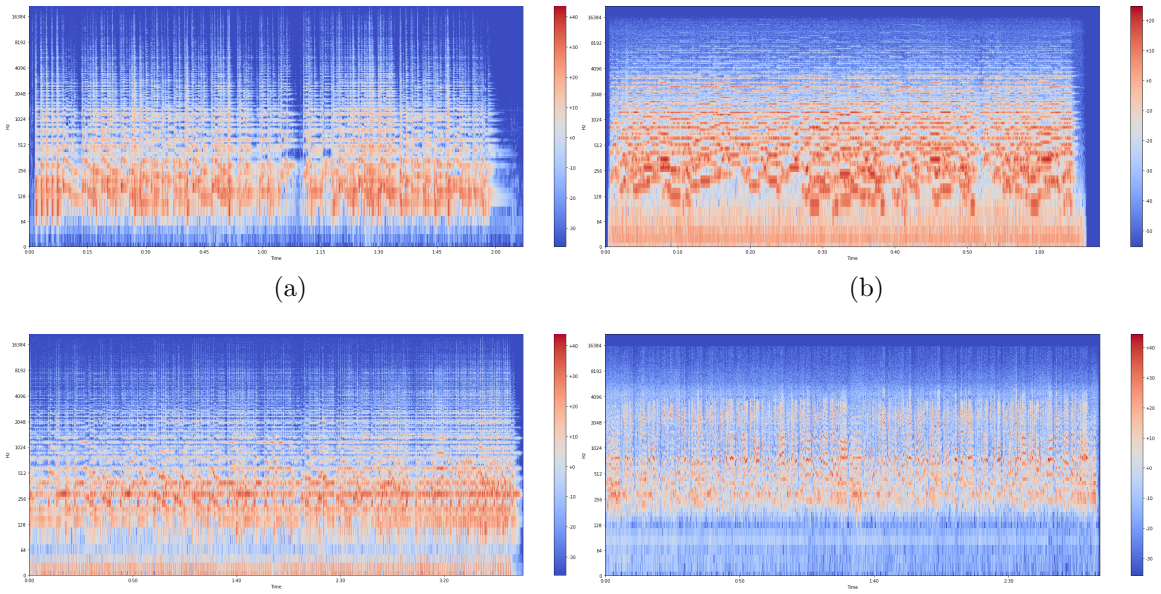
Para el caso del grupo 6, la figura 4.7 muestra varios espectrogramas que caracterizan a la mayoría del contenido en este grupo. En (a) y en (b) se tienen los casos para canciones que fueron creadas con un solo instrumento, un arpa y un órgano respectivamente y donde son visibles los armónicos de ciertas notas; este tipo de canciones representa un 30 % de los elementos de este grupo. Un 44 % del grupo son canciones donde más de un instrumento es utilizado, generalmente 2, y donde algunas tienen presencia de la voz. El caso de (c) representa una canción donde se utiliza solo una guitarra y un piano a la vez. El porcentaje restante corresponde a canciones con ruido de fondo o música electrónica. En (d) se puede visualizar un ejemplo de una canción del género Old Time/Historical, donde actúa principalmente la voz, con este tipo de ruido.

Finalmente, el último grupo de este caso tiene una gran presencia de música electrónica con y sin distorsión, lo que abarca al rededor de un 54 % de las canciones en este grupo. Ejemplos de los espectrogramas de esto se pueden visualizar en la figura 4.8, en especifico en las sub gráficas (a), música electrónica, y (b), sonidos distorsionados. Un comportamiento curioso de estas canciones es el hecho de mostrar una mayor energía entre un rango de hasta 128 Hz, como se visualiza en (a). En este grupo también se encuentran canciones donde la voz tiene un papel especial, como en una entrevista (el caso que muestra (c)) y en el hip hop (d)). Por último, este grupo tiene diversas canciones de pop (un ejemplo se presenta en (e)), rock, un rock más tranquilo y ligero que las canciones agrupadas en el



(a) (b) (c) (d) (e) (f)

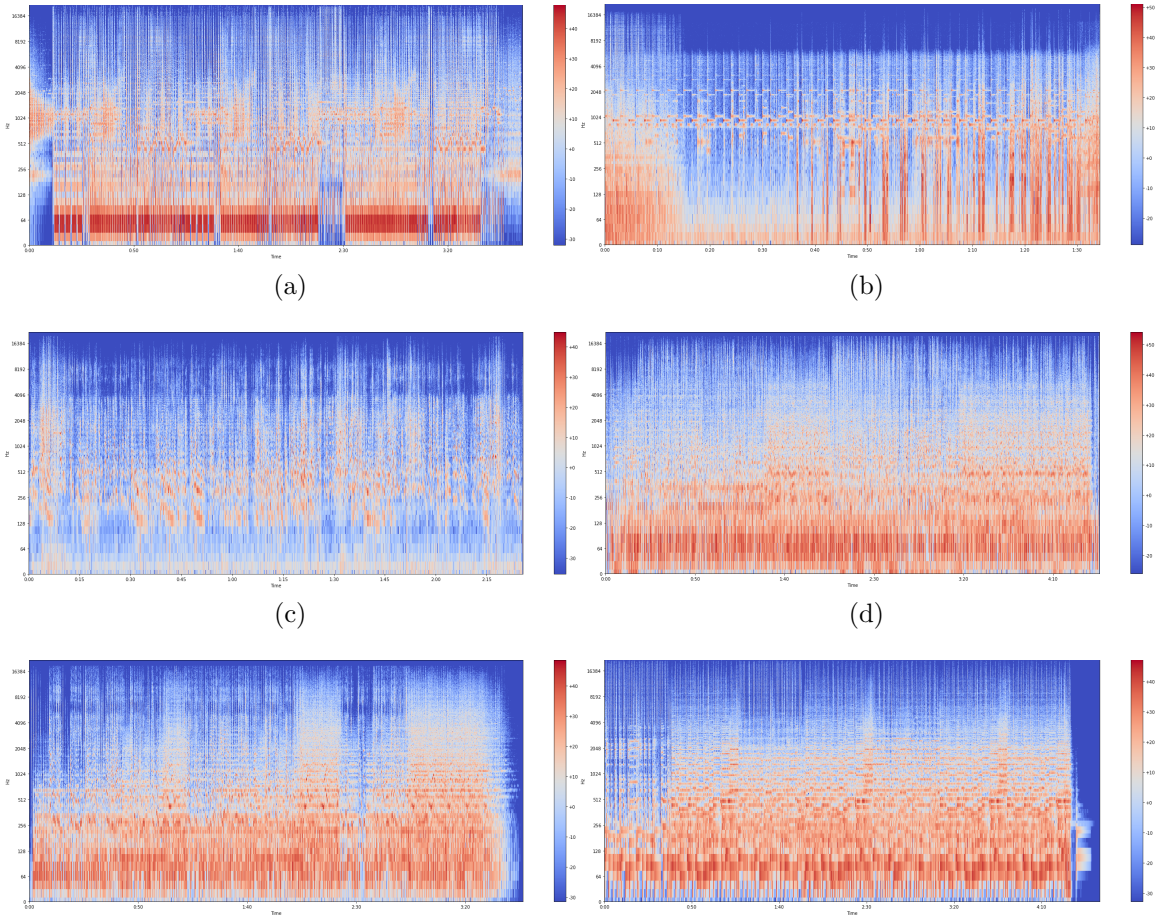
Figura 4.6: Algunos espectrogramas del grupo 5 respecto al conglomerado formado por el método espectral con la metodología MFCC considerando 144 vecinos y 7 grupos.



(a) (b) (c) (d)

Figura 4.7: Algunos espectrogramas del grupo 6 respecto al conglomerado formado por el método espectral con la metodología MFCC considerando 144 vecinos y 7 grupos.

grupo 5, y de blues (como se muestra en (f)).



(a) (b) (c) (d) (e) (f)

Figura 4.8: Algunos espectrogramas del grupo 7 respecto al conglomerado formado por el método espectral con la metodología MFCC considerando 144 vecinos y 7 grupos.

Es importante mencionar que en este conglomerado, en comparación al formado considerando 7 grupos y 60 vecinos, se obtienen, en general, mejores resultados organizando ciertas canciones en un mejor grupo que, auditivamente, comparten más características con dichas canciones.

El conglomerado que considera 10 grupos y 73 vecinos no obtuvo mejores resultados que los casos con 7 grupos, ya que varios de los nuevos grupos contienen observaciones que son bien separadas en los casos anteriores. Por ejemplo, algunas canciones del grupo 6 del caso anterior fueron establecidas en otro donde había canciones del género del rock y electrónica (más de los que ya contenía el propio grupo 6) además de que se crearon dos grupos con muchas canciones del género del rock muy similares entre ellos. La única ventaja que muestra el conglomerado con 10 grupos es una clasificación más refinada de música electrónica aunque sigue teniendo un comportamiento similar al grupo 7 del caso anterior.

4.3.3. DSBCAN

Al implementar este algoritmo con los diferentes conjuntos de vectores resumen se obtuvieron malos resultados; ya que en todos estos se tenía solo un grupo con la mayoría de las observaciones y un grupo representado como ruido. Para este método se utilizó la función *DBSCAN()* del paquete antes mencionado y se consideraron aquellas combinaciones de parámetros que lograran tener los mejores valores en los índices de rendimiento. Además se utilizó la heurística de la gráfica $k - dist$ con distintos valores de k, las cuales tenían un comportamiento similar con cada una de las matrices consideradas mostrando un punto de inflexión importante al considerar un número k grande, es decir cercano a 300.

Al generar solo dos grupos (en el cual uno de ellos era considerado como ruido), se genera un potencial problema para calcular el índice $CVNN$ ya que este necesita un conjunto de K agrupaciones. Por tal motivo, se omite la información recabada de las restantes estadísticas en los diferentes resultados obtenidos. Es importante mencionar que con ninguna combinación de parámetros, para este método, se lograron tener mejores índices en comparación a los obtenidos con las otras metodologías.

Ya que se clasificaba la información en dos grupos (ruido y no ruido), se realizó diversas veces, en todos los casos, la eliminación de las observaciones categorizadas como ruido y se ejecuto nuevamente el método sin dichas observaciones. Los resultados siguieron siendo ineficientes.

Capítulo 5

Discusión y conclusiones

Diferentes metodologías para la búsqueda de conglomerados se efectuaron sobre diversas formas de resumir información musical, las cuales fueron aplicadas a distintas canciones con ciertas licencias derivadas de Creative Commons tomadas desde la plataforma FMA. De las distintas metodologías de resumen, los mejores resultados fueron obtenidos al considerar los vectores resumen creados a partir de los coeficientes cepstrales de cada canción. Por otra parte, las restantes estrategias para resumir el contenido de una pieza musical no logran capturar diferencias auditivas perceptibles que ayuden crear grupos homogéneos entre diversas características auditivas.

Se debe tomar en cuenta que tres de las técnicas de resumen están basadas directamente del espectrograma, en el cual es aplicado solo una transformación necesaria para dar un acercamiento al como un humano entiende o percibe una señal sonora, en este caso estamos hablando de que solo la intensidad es tratada en términos logarítmicos, en una escala de decibelios, por lo que en los conglomerados que utilizan estas tres técnicas se está suponiendo que la escala de la frecuencia es lineal, lo cual puede ser beneficioso en otras aplicaciones pero en este caso en concreto, no favoreció en agrupar canciones con sonidos similares o alguna otra característica auditiva perceptible.

En cuanto a los métodos de agrupamiento que se utilizaron, el que mejores resultados mostró, tanto por sus métricas de desempeño, principalmente el índice CVNN, como por la coherencia de los grupos resultantes, fue el método espectral, resaltando el conglomerado que consideró 7 grupos y 144 vecinos respecto a las características mencionadas en la sección 2.2.3, y utilizando los coeficientes cepstrales para crear los vectores que resumieron a cada canción considerada. Los siguiente métodos que tuvieron un mejor comportamiento fueron los métodos de agrupación tradicionales, obteniendo buenas métricas, en especial en el índice de silueta, aunque la dispersión de las canciones dentro de los grupos no fuera adecuada.

Un punto interesante se puede observar al considerar el método espectral, ya que al crear las matrices de distancias con la distancia coseno y utilizando los vectores construidos por la dispersión espectral, se obtienen los siguientes mejores resultados, solo en términos del índice CVNN. Este hecho y considerando que los coeficientes cepstrales se pueden entender como una proyección del espectrograma, al igual que la métrica coseno tiene una natural interpretación de proyección, dan fuertes indicios de que las obras musicales, o al menos las que fueron consideradas en este proyecto, requieran el uso de proyecciones para lograr capturar características relevantes.

La observación anterior puede explicar el comportamiento obtenido por el método DBSCAN, ya que la información suministrada a este algoritmo dio resultados completa-

mente negativos al clasificar la información en ruido y no ruido. Al igual que en diversas metodologías que se han desarrollado en el aprendizaje supervisado, las proyecciones a otros espacios matemáticos ayudan significativamente a lograr buenos resultados de los algoritmos, por lo que, como trabajo futuro, se recomiendo hacer uso algún tipo de proyección (PCA o SVD) previo al uso del método. En otros artículos donde se utiliza una metodología similar referente a los MFCC, [13], [14] y [19], el uso de la función de similitud coseno resulta apropiada para encontrar segmentos similares en las canciones de manera interna. Además, se utiliza la metodología de descomposición de valores singulares con el mismo propósito de segmentar internamente una canción mostrando resultados favorables. Por lo que la proyección de este tipo de observaciones a mostrado un efecto favorable en el estudio de la música.

Otro punto a notar es el hecho de que la métrica euclidiana, así como el método Ward, propician a mejores resultados del índice CH y que el uso del método promedio en un método aglomerativo favorecen al índice DB. Esto debido a las deficiencias que pueden tener este tipo de métricas si no se tratan con datos que tengan formas convexas y esféricas. Los buenos resultados del índice CVNN resultaron en adecuados grupos para este estudio en particular.

Los resultados mostrados en este trabajo favorecen al uso de los MFCC y el método espectral, así como del uso del índice CVNN. Si bien los coeficientes cepstrales fueron diseñados para capturar de una mejor manera la voz humana, estos muestran una eficiente aplicación en otros ámbitos donde se tengan diferentes fuentes de sonido y donde la interpretación humana tenga relevancia para discernir grupos, lo cual podría no limitarse a la música. Con estos resultados, la aplicación más directa es la creación de algún método de recomendación musical que no se limite a clasificaciones de género, artista, etc. o como una parte importante para la creación de dicho producto. Otros trabajos que pueden resultar interesantes es el estudio de diferentes dialectos, o hasta de lenguas, con esta estrategia, y el estudio de reconocimiento de especies, como las aves, con la finalidad de determinan si algún segmento de la población estudiada tiene características auditivas comunes que nos ayuden en la toma de decisiones.

Apéndice A

Descripción de la base de datos

A continuación se enlistan las canciones utilizadas en los diferentes modelos descritos en el capítulo 4, agregando su autor y la licencia correspondiente.

Artista	Nombre	Licencia
Mar-Tie	Side B, Track 2 (Tribute to Betoven)	CC BY-NC 4.0
Kai Engel	October	CC BY-NC 4.0
Here Comes A Big Black Cloud!!	Death March	CC BY-NC-SA 3.0
Today's Man	Coming 'Round	CC BY-NC-SA 3.0 US
Ada Jones and Len Spencer	Mariutch At Coney Isle	public domain
Steven Arntson	Bones	CC BY-NC 3.0 US
Ludwigs Steirische Gaudi	Running Lucky Through The Village (ID 08)	CC BY-NC-ND 4.0
James Kibbie	BWV 579: Fugue in B Minor	CC BY-NC-ND 3.0
—	The No Bar Suite - Say Yes To Any New Drug	CC BY-NC-ND 4.0
Snow Monster	Intro Reprise	CC BY-NC-SA 3.0 US
King Ghidorah!	Mishima	CC BY-NC-ND 3.0
Rocketship Park	Cakes and Cookies	CC BY-NC-ND 3.0 US
Shilpa Ray	You're Fucking No One	CC BY-NC-ND 4.0
Simon Mathewson	This is Madness	CC BY-NC-SA 4.0
Glass Boy	WELP	CC BY-ND 3.0
Deniz Tek	Big Accumulator (Live at WFMU)	CC BY-NC-ND 3.0 US
The Suzan	Devils	CC BY-NC-ND 3.0 US
6th Sense & Mick Boogie	Intro	CC BY-NC-SA 3.0 US
Single Copy	Book of Job	CC BY-NC-ND 3.0
Humanfobia & The Adverse Event	Energía Residual	CC BY-NC-ND 4.0
Blear Moon	Certain force	CC BY-NC 4.0
The Hall Monitors	Be Your Man	CC BY-NC-SA 3.0
The Relatives	We Need Love	CC BY-NC-ND 3.0
Inaequalis	Vortex	CC BY-NC-SA 4.0
LndMKR	Free Pussy Riot LndRMK	CC BY-NC-ND 4.0
Blear Moon	Step out	CC BY-NC 4.0
2pol	neuer James Bond	CC BY-NC-SA 3.0 US
Sci Fi Industries	Howsway	CC BY-NC-SA 4.0

Artista	Nombre	Licencia
The United States Army Old Guard Fife and Drum Corps	Drum Feature: The Rage of Cornwallis from the George Washington Show	Public Domain Mark 1.0
etc.	endless #04	CC BY-NC-SA 3.0
Kurt Baker	Everybody Knows	CC BY-NC-ND 4.0
Animus Invidious	A Parting Gift	CC BY-NC-ND 4.0
Nihilore	Glimmer	CC BY 3.0
Arrogalla	Padenti secrets	CC BY-NC-ND 2.0 FR
Steve Combs	L'Enfer C'est Les Autres	CC BY 4.0
Cullah	Wander and Ramble	CC BY 4.0
Jamison Williams	deliberately anti-sex symbol, like Prospero in The Tempest, and this film	CC BY-NC-SA 3.0
Bacalao	Quatre Saisons Carrosse - Intro	CC BY-NC-ND 3.0
Janne Nummela	Kingdom of Scandalia	CC BY-NC-SA 3.0
Jahzzar	Invisible	CC BY-SA 4.0
Lovira	All things considered	CC BY-SA 4.0
Ljova and the Kontraband	Hochu Lyubit' (I Want To Love)	CC BY-NC-ND 4.0
Malani Bulathsinhala	Maa Hada Selena	CC BY-ND 4.0
Komiku	Fetch Land	(CC0 1.0) Public Domain Dedication
Deepika Priyadarshani Peiris	Mage Kandulu Wel	CC BY-ND 4.0
Allen Ratnayake	Pura Pura Shree.mp3	CC BY-ND 4.0
The Necks	live at WFMU 2/10/2009	FMA License
Vialka	En Attendant Tout S'Evapore	CC BY-NC 3.0
Il Culto dell'Annientamento	Amore Incondizionato	CC BY-NC-SA 4.0
Lobo Loco	Awakening Spring (ID 376)	CC BY-NC-SA 4.0
KieLoBot	Hounds of Darkmoor (ID 109)	CC BY-NC-ND 4.0
Albert Beger	Tales of Beelzebub	CC BY-NC-ND 3.0
Dirty Beaches	Sweet 17	CC BY-NC-ND 3.0 US
Karunarathna Divulgane	Gahaka Mal Pipila	CC BY-ND 4.0
SANMI	f_s_h_d	CC BY-NC-SA 4.0
Monplaisir	Ca avait l'air d'être facile	CC0 1.0
Vegaenduro	High Speed Flash	CC BY-NC 4.0
Mombojo	Splash	CC BY-NC 3.0 BR
Fallen Leaves	Happy Times	CC BY-NC-SA 3.0
Janne Nummela	Ichthys II	CC BY-NC-SA 3.0
Gigaboy	Hiak Toyaz	CC BY-NC 4.0
ИЗ-ПОД ЗЕМЛИ	Кристина	CC BY-NC-SA 4.0
So Cow	Youre Nice Mysteries	CC BY-NC-ND 3.0 US
Josh Woodward	Go (Instrumental)	CC BY 3.0
Lobo Loco	All has Just Begun (ID 1588)	CC BY-NC-SA 4.0
W.H.W.Y	I Say Goodbye	CC BY-NC-SA 4.0

Artista	Nombre	Licencia
Keshco	Technicolor Universe	CC BY-SA 4.0
Ed Schrader's Music Beat	Intro	CC BY-NC-ND 3.0 US
Broke For Free	Budding	CC BY-NC-ND 3.0
Cherubim	Kasm	CC BY-NC-ND 3.0 US
Leedian	08.model for us	CC BY-NC-SA 3.0
Harvey Milk	I've Got a Love	CC BY-NC-SA 3.0
POVALISHIN DIVISION	Крошка	CC BY-NC 4.0
American Ice Age	Crooked Numbers	CC BY-NC 3.0
—	Africa	CC BY-NC-ND 4.0
Jared C. Balogh	Micro Composition 28	CC BY-NC-SA 3.0
The Cute Lepers	Noisy Song	CC BY-NC-ND 3.0 US
Désir Decir	Words	CC BY-NC-ND 3.0 US
Tortue Super Sonic	Crash	CC BY-NC-SA 3.0
Forget the Whale	Clocks	CC BY-NC 4.0
Scott Holmes Music	Happy Ending	CC BY-NC 4.0
Coachwhips	evil son	CC BY-NC-ND 3.0 US
Trans Atlantic Rage	SPORADIC RANDOM ERRATIC	CC BY-NC-SA 3.0
Squire Tuck	Squire Tuck - Homage to John Carpenter	CC BY-NC-ND 4.0
Ash Turner	Leaping Leopards	CC BY-NC-ND 4.0
—	Ending Credits	CC BY-NC-ND 4.0
Marwood Williams	Mister Nobody Prelude	CC BY-NC-ND 4.0
Ryan Andersen	Easy Feels	CC BY-NC 4.0
Steve Combs	Sun is Rising	CC BY 4.0
Bradley The Buyer	Dragonaut	CC BY-NC-ND 4.0
James and the Ultrasounds	Robot Love	CC BY-NC-ND 4.0
Audiobinger	No More Trap	CC BY-NC 4.0
Podington Bear	Theme in G	CC BY-NC 3.0
Calexico	Crystal Frontier	CC BY-NC-SA 3.0 US
Coatsie	King	CC BY-NC-ND 2.0 FR
John Paul Keith and The One Four Fives	You Devil You	CC BY-NC-ND 3.0 US
A. A. Aalto	Entonces	CC BY-NC 4.0
Section 27 Netlabel	Colour Is Sound	CC BY-NC-ND 3.0
DUB DEPT	Вон(Out)	CC BY 4.0
Tan Low	Angela's Song	CC BY-NC-SA 4.0
Folk Festival Massacre	push pull	CC BY-NC-ND 3.0
Podington Bear	Flitter Key Backwards Beat	CC BY-NC 3.0
Inkubus Sukkubus	Away With the Faeries	CC BY-NC-ND 3.0 US
The Matt Kurz One	Unhappy People	CC BY-NC-SA 3.0 US
Mystery Mammal	Ah!	CC BY 4.0

Artista	Nombre	Licencia
Jared C. Balogh	Keepin' It Steady	CC BY-NC-SA 3.0
Allah-Las	I Had It All	CC BY-NC-ND 3.0
Brown Recluse	Margo, Left In Bed	CC BY-NC-SA 3.0 US
Podington Bear	The Speed Of Life	CC BY-NC 3.0
Noiserv	Dance	CC BY-NC-ND 3.0
Rex Hobart & the Misery Boys	Heartbreak To Hide	CC BY-NC-ND 3.0 US
Lobo Loco	Opener Village Feast (ID 586)	CC BY-NC-ND 4.0
Return to Normal	Like Victory at Marathon	CC BY-NC 4.0
Ketsa	Dancing-Dead	CC BY-NC-ND 4.0
The Womb	The Angle Of The Blade (demo)	FMA License
Paniks live	STE-027	CC BY-NC-SA 3.0
Dee Yan-Key	ave maria	CC BY-NC-ND 4.0
Ninnie	Pretty Polly (vinyl version)	CC BY-NC 4.0
Dariusz Jackowski & Filmy Ghost	Eye Disease	CC BY-NC-ND 4.0
Ade Hodges & Cousin Silas	[waag_rel042] Ade Hodges & Cousin Silas (Space Time Mantra) - 03.An Abandoned Cofee...	CC BY-NC-SA 3.0 US
Satori	Roam	CC BY-NC-ND 4.0
The Kyoto Connection	The last days of a Samurai Soul	CC BY-NC-ND 2.5 AR
Lobo Loco	Jessy Travel Gambler (ID 1143)	CC BY-NC-SA 4.0
Jody Pou	Udite amanti (B. Strozzi)	CC BY-NC-SA 3.0 US
Slicing Grandpa	Mystery Guest	CC BY-NC-ND 3.0
Snowboarder	Sled Dogs	CC BY-NC-SA 3.0 US
Squire Tuck	Something Borrowed, Something New	CC BY-NC-ND 4.0
Unknown Artist	Before The Dawn	CC BY-NC-SA 3.0
Jody Pou	FIN CHE TU SPIRI/Barbara Strozzi	CC BY-NC-SA 3.0 US
Box Elders	Ronnie Dean	CC BY-NC-ND 3.0 U
—	Julie's An Android	CC BY-NC-ND 4.0
Morgan Fisher	MO 30-1 CUT	CC BY-NC-ND 3.0
Box Elders	S+M Party	CC BY-NC-ND 3.0 US
Kai	Eyes are	CC BY-NC-ND 3.0
Nat M. Wills	Saving Up Coupons For Mother	public domain
Nonima	XIII (feat. Altered:Carbon)	CC BY-NC-ND 3.0
The Snow	Silent Parade	CC BY-NC-SA 3.0
Csum	Hypolink (loopy mix)	CC BY-NC-SA 3.0 US
Dee Yan-Key	The Dreaming Swan	CC BY-NC-SA 4.0
Infinite Livez	Lucky you (the Earlyman remix)	CC BY-NC-ND 3.0
Origami Repetika	Sunny Morning Exercise Club	CC BY 4.0
Vic Ruggiero & Jesse Wagner	Pray	CC BY-NC-ND 4.0
The Agrarians	And Love You Most of All	CC BY-NC-SA 3.0 US
Colin Langenus	Dos	CC BY-NC-ND 3.0 US
Roglok	Radetzky March 303	CC BY-NC-ND 4.0

Artista	Nombre	Licencia
The Laurels	Falling Away With You	CC BY-NC-ND 3.0
Jared C. Balogh	Micro Composition E	CC BY-NC-SA 3.0
Komiku	Treasure finding	CC0 1.0
Désir Decir	Laura Rose	CC BY-NC-ND 3.0 US
Welcome Wizard	Mathlab	CC BY-NC-ND 3.0 US
half cocked	It Happens	CC BY-NC-ND 4.0
No Monster Club	Ye Olde Head Shoppe	CC BY-NC-ND 3.0
Podington Bear	Désormais	CC BY-NC 3.0
Squire Tuck	Losing My Way	CC BY-NC-ND 4.0
Lee Maddeford, Roland Vouilloz	Yuyu	CC BY-NC-SA 3.0
Soft Serve	Let's Go Dutch	CC BY-NC-ND 3.0 US
Derek Clegg	It Ain't This Bed	CC BY-NC-SA 3.0 US
Born Loose	Whiskey Holiday	CC BY-NC-ND 3.0 US
The Spectacular Fantastic	Ufos	CC BY-NC 3.0
Bauchamp	140 in the forrest	CC0 1.0
Born Loose	Step Up To The Plate (Be A Runaway)	CC BY-NC-ND 3.0 US
The Split Squad	I've Got a Feeling	CC BY-NC-ND 3.0 US
Scott Holmes Music	Bouncy Fun Song	CC BY-NC 4.0
S.Bobrytskyy & M.Paramzin	Hello, Dora	CC BY-NC-SA 3.0
The Cynics	Interview w/ Terre	CC BY-NC-SA 3.0
Jahzzar	Candy	CC BY-SA 4.0
Minmae	—	CC BY-NC-SA 3.0 US
Jahzzar	The Wrong Way	CC BY-SA 4.0
Waylon Thornton	Twenty Four	CC BY-NC-SA 3.0 US
Trans Atlantic Rage	NO MATH 4	CC BY-NC-SA 3.0
King Imagine	Evening Hot Song	CC BY-NC-ND 4.0
W.H.W.Y	К Тебе	CC BY-NC-SA 4.0
Zero V	A Tourist in his Hometown	CC BY-NC-SA 3.0
Malani Bulathsinhala	Adarayaka Hengumak	CC BY-ND 4.0
James Kibbie	BWV 770: Partite diverse sopra il Corale Ach, was soll ich Sünder machen	CC BY-NC-ND 3.0
Sam Weinberg	Shrapnel Facsimiles	CC BY-NC-ND 4.0
Zeke Healy	September Song	CC BY-NC-ND 3.0 US
Bounar	last blow	CC BY-NC-SA 3.0 US
No Monster Club	Good Boy For Life	CC BY-NC-ND 3.0
Gablé	seminéoproantiantiantifolk	CC BY-NC-SA 3.0 US
Pianochocolate	Pianochocolate - little Princess	CC BY-NC-ND 4.0
Mors Ontologica	Voice Of Degeneration	CC BY-NC-SA 3.0 US
No Monster Club	Let's Crowdsurf Mike Stevens	CC BY-NC-ND 3.0
LE CHEVALIER DE RINCHY	midimomi	CC BY-NC-ND 4.0

Artista	Nombre	Licencia
Goto80	fonky spenat	CC BY-NC-SA 3.0
Dee Yan-Key	Adagio con anima	CC BY-NC-SA 4.0
Monplaisir	Brother	CC0 1.0
Lloyd Cole	Women's Studies	CC BY-NC-ND 4.0
King Ghidorah!	Sleeping Sickness	CC BY-NC-ND 3.0
Kiko Dinucci, Juçara Marçal, Thiago França	Ora Iê iê o	CC BY-NC-ND 3.0
James Kibbie	BWV 741: Ach Gott, vom Himmel sieh darein	CC BY-NC-ND 3.0
Deerhoof	Deerhoof7	CC BY-NC-ND 3.0
Dee Yan-Key	Four	CC BY-NC-SA 4.0
half cocked	Mind Control	CC BY-NC-ND 4.0
Maya DeVries	Petit Papa Noël	CC BY-NC-SA 4.0
Westy Reflector	Words Of Release	CC BY-NC-SA 3.0 US
LE CHEVALIER DE RINCHY	pur classik rock jingle	CC BY-NC-ND 4.0
ST37	Concrete Island	CC BY-NC-SA 3.0
Victor Herbert Orchestra	1911 - Jubel Overture (Weber)	Public Domain Mark 1.0
Jason Shaw	SOLO ACOUSTIC GUITAR	CC BY 3.0 US
Azevedo Silva	A Morte	CC BY-NC-ND 3.0
Hector 3	Bad Man	CC BY-NC-ND 3.0 US
—	No News Is Good News	CC BY-NC-ND 4.0
Mentz	Introduction	CC BY-NC-SA 3.0 US
Ludwigs Steirische Gaudi	You an Me (ID 19)	CC BY-NC-ND 4.0
Lobo Loco	Greatful World (ID 1549)	CC BY-NC-SA 4.0
Kimiko Ishizaka	Aria da Capo è Fine	CC0 1.0
D'r Sjaak	Boxeboam	CC BY-NC-SA 3.0
Sam Gas Can	Crazy Family	CC BY-NC-SA 3.0 US
Visciera	Fear	CC BY-NC-SA 4.0
—	Super-friendly	CC BY-NC-ND 4.0
Azevedo Silva	À Deriva	CC BY-NC-ND 3.0
Sergi Boal	la mar	CC BY-NC-ND 3.0
Siddhartha Corsus	Epiphany	CC BY-NC 4.0
Эксперимент	Крошка	CC BY-NC-ND 3.0
James Kibbie	BWV 530: Trio Sonata VI in G Major - 2. Lente	CC BY-NC-ND 3.0
Borrtex	Perception	CC BY-NC 4.0
Ken Fury	Creature Filth	CC BY-NC-ND 4.0
Mar-Tie	Love Of Mine (Navajo Lady)	CC BY-NC 4.0
Dee Yan-Key	buen viaje (montuno)	CC BY-NC-SA 4.0
äNACRUSä	Menudencias Asunción	CC BY-NC-SA 3.0 US
Josh Woodward	Little Tomcat (Instrumental Version)	CC BY 3.0
Folk Festival Massacre	Impulsive	CC BY-NC-ND 3.0

Artista	Nombre	Licencia
—	People	CC BY-NC-ND 4.0
Blue Dot Sessions	An Opus in Ab	CC BY-NC 4.0
SANMI	Funki Tonki	CC BY-NC 4.0
Podington Bear	Spring Comes Early	CC BY-NC 3.0
Pau Riba	Noia de Porcellana (Porcelain Girl)	CC BY-NC-ND 4.0
—	Etude No.1 - 4. Lydien	CC BY-NC-ND 4.0
The Wrong Words	What Went Wrong	CC BY-NC-ND 3.0 US
Crno dete	Isuse	CC BY-NC-SA 3.0 US
NiCad	Paradise	CC BY-NC-ND 3.0 US
The Money Shot	Stoked	CC BY-NC-ND 4.0
—	Crawling	CC BY-NC-ND 4.0
Virginia Pipe	Basically Drove The Guy Crazy	CC BY-NC-SA 2.0 UK
Lesvicon Soul feat. Kourisha	The One in Dub	CC BY-NC-SA 3.0
Keshco	Got Lot Of Stuff	CC BY 4.0
Cyminology	Nachofteam	FMA License
Nanda Malini	Ran Giri Giri Gigiri.mp3	CC BY-ND 4.0
Custodian of Records	Lifes Blunder	CC BY-NC-SA 3.0 US
Dagos	Not so long time ago	CC BY-NC-SA 3.0
Bleeding Rainbow	Pink Ruff	CC BY-NC-ND 3.0 US
half cocked	Little Snub Nose	CC BY-NC-ND 4.0
HR Jothipala	Dili Dili Dilisevi	CC BY-ND 4.0
Dorothy Kingsley	Call Round Any Old Time	public domain
U.S. Army Blues	Barbara	Public Domain Mark 1.0
Cath and Phil Tyler	Abbeville	CC BY-NC-SA 3.0 US
The Polish Ambassador	Unrequited Droid Love	CC BY-NC-SA 3.0 US
Podington Bear	Epiphany	CC BY-NC 3.0
Pavlosiuk & the Dudes	Tarpinė	CC BY-NC-SA 4.0
Turku, Nomads of the Silk Road	Majnun Nabudom	CC BY-NC-SA 3.0 US
Thorn & Shout	Trouble	CC BY-NC-SA 3.0 US
OsO El roTo	petakita part one	CC BY-NC-SA 3.0 US
Lucas Gonze	Talk About Suffering	CC0 1.0
Podington Bear	SproutJam	CC BY-NC 3.0
Tagirijus	Break The Keys	CC BY-NC-SA 4.0
Jared C. Balogh	Micro Composition 12	CC BY-NC-SA 3.0
Captive Portal	Have You Ever Been In Here?	CC BY-SA 4.0
A. A. Aalto	Rack Focus	CC BY-NC 4.0
Monplaisir	Au delà de la magie	CC0 1.0
Ergo Phizmiz	Lottie on the Streets	CC BY-NC-SA 4.0
Komiku	In the restaurant	CC0 1.0
It's Notherground Music!!	_	CC BY-NC-SA 3.0 US
Jahzzar	Playtime	CC BY-SA 4.0

Artista	Nombre	Licencia
King Imagine	Maple Leaf	CC BY-NC-ND 4.0
James Kibbie	BWV 764: Wie schön leuchtet der Morgenstern	CC BY-NC-ND 3.0
Samuel Segal	Storyteller Waltz	public domain
Georgian State Folk Song and Dance Ensemble	Chakrulo	CC BY-NC 4.0
Ludwigs Steirische Gaudi	Lucky Hannes (ID 26)	CC BY-NC-ND 4.0
Anamanaguchi	Helix Nebula	CC BY-NC-ND 3.0
Cory Gray	Technological 2	CC BY-NC 3.0
Fallen Leaves	Days of Summer	CC BY-NC-SA 3.0
Slowdance	Je Compte Pour Toi	CC BY-NC-ND 3.0 US
Josh Woodward	Pompeii (No Vocals)	CC BY 3.0
Blue Wave Theory	22 Hornet	CC BY-SA 4.0
Kenny Tudrick	The River	CC BY-NC-ND 3.0
Flux Without Pause	Project 5am 2007-2014	CC BY-NC-ND 4.0
Daniel Barbiero - Cristiano Bocci	gli alberi, a gennaio	CC BY-NC-ND 3.0
Edward Shallow	Inhibitor	CC BY-NC-SA 3.0 US
Hall Of Fame	Rival	CC BY-NC-ND 3.0 US
The Kyoto Connection	River of Hope	CC BY-NC-ND 2.5 AR
Pheasant	Debtors	CC BY-NC-SA 4.0
Yung Kartz	Yuh	CC BY-NC-ND 4.0
Jane Weaver	Modern Kosmology	CC BY-NC-ND 4.0
The Agrarians	As Crimson As Sunrise	CC BY-NC-SA 3.0 US
Muck and the Mires	Do It All Over Again	CC BY-NC-ND 3.0 US
Evan Schaeffer	Turnaround	CC BY 4.0
O+YN	uduada	CC BY-NC-SA 3.0 US
Jesse Spillane	Ringer	CC BY 4.0
Demoiselle Döner	Frère	CC0 1.0
Jesse Spillane	G1	CC BY 4.0
Carsie Blanton	Buoy	CC BY-NC-ND 3.0
Toxic Lipstick	stabbed	CC BY-NC-SA 3.0 US
Seazo	Drunk Clown	CC BY-NC-SA 4.0
Archers	Evil City Music	CC BY-NC-SA 3.0 US
The Joy Drops	NotDrunk-snippet-trumpet	CC BY 4.0
—	Have It	CC BY-NC-ND 4.0
Marty Ehrlich	Unison	CC BY-NC-ND 3.0 US
Burnt Ones	Fountain Of Youth/ Bury Me In Smoke	CC BY-NC-ND 3.0
Salakapakka Sound System	One Minute 10	CC BY 4.0

Bibliografía

[1] Adolfsson, A., Ackerman, M. and Brownstein, N.C. 2019. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*. 88, (2019), 13–26.

[2] Aggarwal, C.C. and Reddy, C.K. 2014. *Data clustering: Algorithms and application*. CRC Press.

[3] Aherne, F.J., Thacker, N.A. and Rockett, P.I. 1998. The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*. 34, 4 (1998), 363–368.

[4] Ahmed, N., Natarajan, T. and Rao, K.R. 1974. Discrete cosine transform. *IEEE transactions on Computers*. 100, 1 (1974), 90–93.

[5] Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W. and Iannone, R. 2021. *Rmarkdown: Dynamic documents for r*.

[6] Axler, S.J. 1997. *Linear algebra done right*. Springer.

[7] Banerjee, A. and Dave, R.N. 2004. Validating clusters using the hopkins statistic. *2004 IEEE international conference on fuzzy systems (IEEE cat. No. 04CH37542)* (2004), 149–153.

[8] Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V. and Pebesma, E.J. 2008. *Applied spatial data analysis with r*. Springer.

[9] Brunton, S.L. and Kutz, J.N. 2019. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press.

[10] Buitinck, L. et al. 2013. API design for machine learning software: Experiences from the scikit-learn project. *ECML PKDD workshop: Languages for data mining and machine learning* (2013), 108–122.

[11] Caliński, T. and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*. 3, 1 (1974), 1–27.

[12] Chen, W.-H., Smith, C. and Fralick, S. 1977. A fast computational algorithm for the discrete cosine transform. *IEEE Transactions on communications*. 25, 9 (1977), 1004–1009.

[13] Cooper, M. and Foote, J. 2002. Automatic music summarization via similarity analysis. *ISMIR* (2002).

[14] Cooper, M. and Foote, J. 2003. Summarizing popular music via structural similarity analysis. *2003 IEEE workshop on applications of signal processing to audio and acoustics (IEEE cat. No. 03TH8684)* (2003), 127–130.

[15] Cover, T.M. and Thomas, J.A. 2006. *Elements of information theory*. Wiley-Interscience.

[16] Davies, D.L. and Bouldin, D.W. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*. 2 (1979), 224–227.

[17] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* (1996), 226–231.

[18] Everitt, B.S., Landau, S., Leese, M. and Stahl, D. 2011. *Cluster analysis*. John Wiley.

[19] Foote, J. 2000. Automatic audio segmentation using a measure of audio novelty. *2000 ieee international conference on multimedia and expo. icme2000. Proceedings. Latest advances in the fast changing world of multimedia (cat. No. 00th8532)* (2000), 452–455.

[20] Friedman, J., Hastie, T. and Tibshirani, R. 2001. *The elements of statistical learning*. Springer series in statistics New York.

[21] Gan, G., Ma, C. and Wu, J. 2007. *Data clustering: Theory, algorithms, and applications*. SIAM.

[22] Giancoli, D.C. 2009. *Física para ciencias e ingeniería con física moderna*.

[23] Gower, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics*. (1971), 857–871.

[24] Han, J., Pei, J. and Kamber, M. 2011. *Data mining: Concepts and techniques*. Elsevier.

[25] Hennig, C., Meila, M., Murtagh, F. and Rocci, R. 2015. *Handbook of cluster analysis*. CRC Press.

[26] Huang, X., Acero, A., Hon, H.-W. and Reddy, R. 2001. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

[27] Jain, A.K., Murty, M.N. and Flynn, P.J. 1999. Data clustering: A review. *ACM computing surveys (CSUR)*. 31, 3 (1999), 264–323.

[28] Johnson, R.A., Wichern, D.W., et al. 2002. *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ.

[29] Kassambara, A. 2017. *Practical guide to cluster analysis in r: Unsupervised machine learning*. Sthda.

[30] Kassambara, A. and Mundt, F. 2020. *Factoextra: Extract and visualize the results of multivariate data analyses*.

[31] Kaufman, L. and Rousseeuw, P.J. 2009. *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

[32] Knees, P. and Schedl, M. 2016. *Music similarity and retrieval: An introduction to audio-and web-based strategies*. Springer.

[33] Koch, I. 2013. *Analysis of multivariate and high-dimensional data*. Cambridge University Press.

[34] Legendre, P. and Legendre, L.F. 2012. *Numerical ecology*. Elsevier.

[35] Ligges, U., Krey, S., Mersmann, O. and Schnackenberg, S. 2018. *tuneR: Analysis of music and speech*.

[36] Ligges, U., Short, T. and Kienzle, P. 2021. *Signal: Signal processing*.

[37] Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J. and Wu, S. 2013. Understanding and enhancement of internal clustering validation measures. *IEEE transactions on cybernetics*. 43, 3 (2013), 982–994.

[38] Luque-Calvo, P.L. 2017. *Escribir un trabajo fin de estudios con r markdown*. Disponible en http://destio.us.es/calvo.

[39] Maimon, O. and Rokach, L. 2010. *Data mining and knowledge discovery handbook*. Springer.

[40] Mandel, M.I. and Ellis, D.P. 2005. Song-level features and support vector machines for music classification. (2005).

[41] Maulik, U. and Bandyopadhyay, S. 2002. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*. 24, 12 (2002), 1650–1654.

[42] McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E. and Nieto, O. 2015. Librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference* (2015), 18–25.

[43] Müller, M. 2015. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer.

[44] Müller, M. 2021. *Fundamentals of music processing: Using python and jupyter notebooks*. Springer.

[45] Müller, M. 2007. *Information retrieval for music and motion*. Springer.

[46] Norton, M.P. and Karczub, D.G. 2003. *Fundamentals of noise and vibration analysis for engineers*. Cambridge university press.

[47] Olver, P.J., Shakiban, C. and Shakiban, C. 2006. *Applied linear algebra*. Springer.

[48] Orlóci, L. 1967. An agglomerative method for classification of plant communities. *The Journal of Ecology*. (1967), 193–206.

[49] Pampalk, E. 2006. *Computational models of music similarity and their application in music information retrieval*.

[50] Sueur, J., Aubin, T. and Simonis, C. 2008. Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics*. 18, (2008), 213–226.

[51] Sundararajan, D. 2018. *Fourier analysis–a signal processing approach*. Springer.

[52] Unpingco, J. 2016. *Python for signal processing*. Springer.

[53] Ward Jr, J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*. 58, 301 (1963), 236–244.

[54] Wickham, H. et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software*. 4, 43 (2019), 1686.

[55] Wierzchoń, S.T. and Kłopotek, M.A. 2018. *Modern algorithms of cluster analysis*. Springer.

[56] Wilson, J.D. and Buffa, A.J. 2002. *Física*. Pearson Educación.

[57] Winston Chang 2014. *Extrafont: Tools for using fonts*.

[58] Yang, Y.-Y. et al. 2021. TorchAudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*. (2021).

[59] Zelnik-manor, L. and Perona, P. 2005. Self-tuning spectral clustering. *Advances in neural information processing systems* (2005).