Clasificación de géneros musicales con técnicas de *Clustering*

Bianca Picchetti
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Buenos Aires, Argentina
bianca.picchetti@gmail.com

Federico Moreno
Departamento de Economía
Universidad Torcuato Di Tella
Buenos Aires, Argentina
fede_moreno613@hotmail.com

***Abstract*—Un género musical es una categoría que reúne composiciones musicales que comparten distintos criterios de afinidad. Se busca en el presente trabajo determinar si las pistas musicales se agrupan naturalmente por géneros a partir de atributos de alto y bajo nivel. Para ello se aplicaron las técnicas de agrupamiento jerárquico aglomerativo, k-medias y PAM. El primer algoritmo mostró una agrupación más homogénea de los atributos de alto nivel (*audio_features*) en relación a los de bajo nivel (*audio_analysis*). Consecuentemente, al aplicar los algoritmos de k-medias las métricas de validación interna permitieron seleccionar un número mayor de grupos para *audio_features* que para los demás, obteniéndose un índice de Van Dongen de 0.688. Al aplicar el logaritmo PAM ocurrió lo mismo; sin embargo, el índice obtenido fue de 0.700 para *audio_analysis* y de 0.643 para *audio_features*.**
En todos los casos se observó que los géneros *ambient y classical* y *jazz y world-music* eran mayoritariamente asignados al mismo *cluster*. Esto está en consistencia con los observado al graficar las clases reales en baja dimensión (TSNE), dado que dichos géneros presentan superposición.
Se realizó también análisis espectral de los atributos de bajo nivel (timbre) de un *track*, pudiendo reconocer en la matríz de afinidad distintas partes de la canción. Al aplicar k-medias (*k*=3) a los datos, el estribillo y las estrofas fueron asignados a grupos diferentes.
Dada la elección de trabajar únicamente con variables contínuas -y considerando que la clasificación según género musical de un tema no depende exclusivamente de variables cuantificables- los géneros no se separan completamente unos de otros y por consiguiente los algoritmos aplicados no lograron una agrupación adecuada. Sin embargo, conociendo esta superposición de clases, pudo obtenerse una clasificación consistente con ella.

***Index Terms*—clustering, clustering jerárquico aglomerativo, K-Medias, K-Medoides, PAM, DBScan, TSNE, clustering espectral**

I. Introducción

Un género musical es una categoría que reúne composiciones musicales que comparten distintos criterios de afinidad [1] tales como su función (música de danza, música religiosa, música de cine), su instrumentación (música vocal, música instrumental, música electrónica), el contexto social en que es producida o el contenido de su letra. Mientras que las tradicionales clasificaciones académicas en géneros musicales han atendido fundamentalmente a la función de la composición musical (para qué es compuesta la pieza, como en los ejemplos anteriores), las clasificaciones por géneros de la música moderna, usadas por la industria discográfica, han atendido más a criterios específicamente musicales (ritmo, instrumentación, armonía) y a características culturales, como el contexto geográfico, histórico o social.
Sin embargo, uno de los inconvenientes al agrupar música por géneros reside en que se trata de un proceso subjetivo, influido por el conocimiento personal y la forma de cada uno de sentir y escuchar la música. Gracias al gran avance en los últimas años en tecnologías de aprendizaje automático y a la alta disponibilidad de datos gracias a servicios de *streaming* de música, se han estudiados diversos algoritmos para realizar esta clasificación en forma automática, basándose en distintos atributos de los temas musicales (tales como descriptores de bajo nivel [2] y basados en análisis espectral [3]).
Por ende, se propone en el presente estudio determinar si las pistas musicales (*tracks*) se agrupan naturalmente por géneros a partir de atributos a distintos niveles de observación del tema (alto y bajo nivel). Para ello se evaluarán distintos algoritmos de *clustering* y set de datos buscando encontrar una concordancia entre los géneros musicales y los grupos hallados de forma no supervisada.

II. Dataset de trabajo

Para el presente trabajo se cuenta con tres *datasets*, obtenidos a través de consultas a la API de *Spotify*: *metadata*, *audio_features* y *audio_analysis*. El listado completo de las variables que componen a cada uno de ellos se puede encontrar en el Apéndice A.
El set de datos *metadata* consta de 2206 registros y 17 variables. Corresponde a datos en su mayoría del tipo texto que identifican al *track*: artista, álbum, número de disco, género, etc. La variable *genre* (género) de este *dataset* será utilizada como la clase real para la validación externa de los distintos algoritmos de *clustering*. Existen en total 5 géneros musicales distintos en los datos estudiados aquí: *jazz, classical, drum-and-bass, world-music, ambient*.
El set de datos *audio_features* consta de 2206 registros y 17 variables. Se trata de descriptores de alto nivel tales como tempo, volumen, energía, etc. Las variables no numéricas y las binarias fueron descartadas (ver Apéndice A) así como la variable *speechiness* por presentar poca variación en sus

valores; también se descartaron aquellas variables altamenta correlacionadas: *acousticness* y *loudness*, conservando la variable *energy* (ver Apéndice B). Se conservó un total de 7 variables para el estudio de *clustering*.
El set de datos *audio_analysis* contiene las variables continuas de bajo nivel, *timbre* (timbre) y *pitch* (tono), estimadas en ventanas temporales. Cada una de ellas consta de 12 datos, correspondientes a los 12 valores tímbricos y de tono que brinda *Spotify*[1]. Los datos fueron resumidos en su media y su desvío standard, obteniendo así para cada *track* 12 medias y desvíos standard para la variable *timbre* y 12 medias y desvíos standard para la variable *pitch*.
Finalmente, todos los datos fueron normalizados en el rango [0-1]. Se utilizan para el análisis los *datasets* de *audio_features* (AF), *audio_analysis* (AA) y la unión de ambos (AA+AF).

III. Tendencia al clustering: Índice de Hopkins

El índice de Hopkins es una manera de medir la tendencia que tienen los datos a formar *clusters*. Para ello, se calcula la distancia a los vecinos más próximos y se la compara con la distancia entre puntos generados al azar; si el índice presenta un valor cercano a 0.5 significa que los datos de trabajo presentan una distribución similar a los datos distribuídos de forma *random*. En cambio, si los datos de trabajo se encuentran más agrupados, el índice será mucho menor a 0.5. Se observan en la Tabla I los valores obtenidos. En los tres casos se concluye que existe tendencia al *clustering*.

TABLE I: Índices de Hopkins obtenidos

Dataset	Índice de Hopkins
AA+AF	0.08
AF	0.07
AA	0.06

IV. Clustering Jerárquico Aglomerativo

A. Método

Como primer paso y a modo exploratorio, se aplica la técnica de agrupamiento jerárquico aglomerativo: se trata de un acercamiento ascendente; cada observación comienza en su propio grupo y los pares de grupos son mezclados mientras uno sube en la jerarquía. En todos los casos se utiliza la matriz de distancia Euclídea -dado que sólo se trabajará con datos contínuos y normalizados- y se calcula el índice cofenético para determinar el método a aplicar. Para los casos AA+AF y AF se utilizó el método de enlace *average* (promedio) mientras que para AA se utilizó *weighted* (pesado).

[1]https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-analysis/

B. Resultados y discusión

Se observa en la Figura 1 el dendrograma obtenido para cada uno de los set de datos. Para el caso de AA+AF se observan dos grandes grupos y al menos tres de mucho menor tamaño. En el caso de AA es apreciable un grupo que comprende la mayoría de los registros. En cambio, para el *dataset* AF se observan cinco grupos de tamaño más homogéneo en comparación a los dendrogramas previamente mencionados.
En principio, observando los gráficos de la Figura 1 y conociendo la cantidad de clases reales (cinco), se esperaría que el *dataset* AF sea el que agrupe mejor según el género musical.

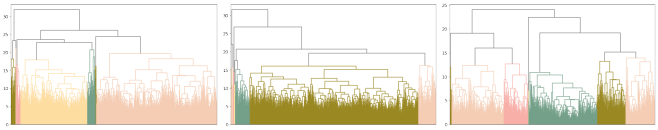

Fig. 1: Dendogramas obtenidos para el *dataset* AA+AF (izquierda, método *average*), AA (medio, método *weighted*) y AF (derecha, método *average*)

V. K-Medias (*K-Means*)

A. Método

K-medias es un método de agrupamiento que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Uno de los requisitos para aplicar este algoritmo es indicar el número apropiado de *clusters* (k). En el presente trabajo el mismo fue estimado por dos técnicas: el *"bastón roto"*, en la que se grafica la suma de los errores cuadrados (SSE) en función del k; este método es heurístico y se busca elegir una cantidad de grupos tal que el agregado de uno más no reduzca de forma significativa la varianza explicada. Dado que esta técnica suele ser ambigua, se complementa con la técnica de validación interna de *silhouette*, en la que se calcula y se grafica dicho coeficiente para distinta cantidad de grupos. Este valor es una medida de cuán parecidos son los datos dentro de un grupo comparado con otros *clusters*. Un nivel cercano a 1 indica que los datos dentro de un grupo son muy parecidos, pero muy distintos a los de otros grupos. Se considera como k adecuado aquel que minimiza SSE y maximiza el índice de *silhouette*; dado que estos valores no siempre coinciden, se busca un compromiso entre ambas condiciones.
Una vez seleccionado k se realizó la validación externa mediante la matriz de confusión y los índices de Van Dongen (VDI). Se elige esta métrica dado que los grupos obtenidos no coinciden con la cantidad de clases reales. También se comparan los agrupamientos obtenidos para cada *dataset* con el fin de estudiar su similaridad.

B. Resultados y discusión

Se muestran en la Figura 2 los gráficos de SSE y coeficiente de *silhouette* en función de *k*. Para AA+AF se observa un quiebre en la curva de SSE para 2 *clusters*, coincidente con el valor máximo del coeficiente de *silhouette*. El gráfico de *silhouette* muestra el coeficiente para cada caso, obteniéndose unos pocos mal clasificados en el *cluster* de etiqueta 1 (coeficiente negativo). El *dataset* AA muestra un comportamiento similar, obteniéndose también 2 grupos. En cambio, el *dataset* AF no muestra un descenso rápido del coeficiente de *silhouette* a medida que aumenta el valor de *k*, manteniéndose en valores cercanos al máximo hasta *k*=6. El gráfico de *silhouette* indica una buena separación en 4 grupos (Figura 2, abajo, derecha), mejor que para 5 grupos (no mostrado).

En la Tabla II se muestran las matrices de confusión

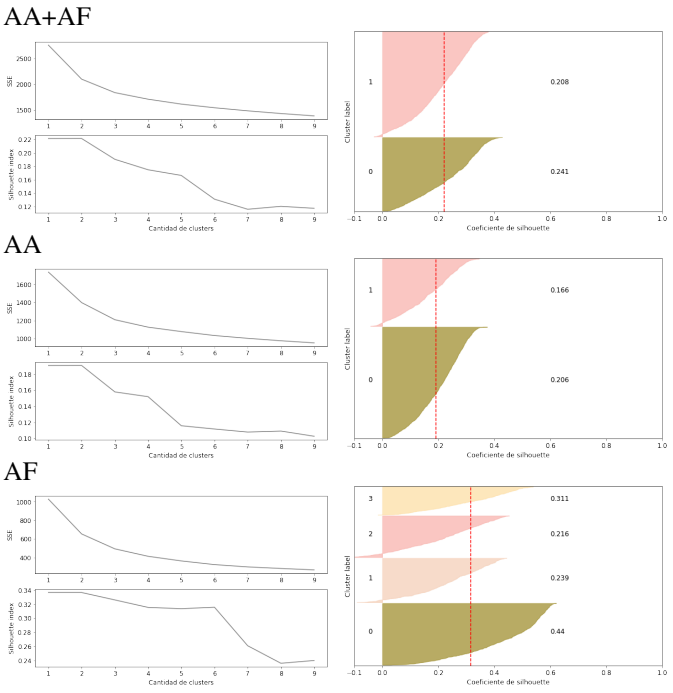

Fig. 2: SSE y coeficiente de *silhouette* en función de la cantidad de *clusters* obtenido con K-Medias para cada *dataset*; se muestra el gráfico de *silhouette* para el *k* elegido. (a) AA+AF. (b) AA.(c) AF. No es posible calcular el valor del coeficiente de *silhouette* para un único *cluster*, por lo que se asigna el coeficiente de *k*=2 para su visualización.

obtenidas y el índice de Van Dongen calculado. Se observa que tanto AA+AF como AA agrupan de forma similar a los mismos géneros: *ambient* y *classical* por un lado, *drum-and-bass*, *jazz* y *world-music* por el otro, con casos asignados al otro grupo. Los grupos obtenidos para AF, si bien presenta el mayor índice de Van Dongen de los tres, se encuentran los *tracks* de un mismo género distribuídos entre los cuatro grupos. En el caso de *ambient* y *classical* la mayoría de los temas de cada uno se encuentran en el mismo *cluster*; *drum-and-bass* se divide en cantidades similares entre dos grupos distintos: uno de ellos compartido con la mayoría de los temas del género *world-music* y el otro con una cantidad de casos significativos de *ambient* y *jazz*.

Estos resultados muestran una clasificación en grupos similar

TABLE II: Matrices de confusión de las clases reales contra las distribución en grupos obtenidas con k-medias

	AA+AF		AA		AF			
ambient	376	84	351	109	312	7	105	36
classical	381	24	366	39	314	4	12	75
drum-and-bass	2	449	16	435	0	207	242	2
jazz	77	349	39	387	86	88	101	151
world-music	72	391	62	401	62	250	62	89
VDI	0.617		0.648		0.688			

para AA+AF y AA -lo cual es razonable considerando que un *dataset* está contenido en el otro y representa la mayoría de sus variables. Los datos de AF si bien las métricas indican que podrían separarse en cuatro grupos, al observar la matríz de confusión se ve una gran superposición entre ellos. El método de K-medias es una técnica sensible a *outliers*; dado que no se realizó un trabajo exhaustivo para detectarlos y eliminarlos es esperable obtener una clasificación poco óptima. En la siguiente sección se aplica el algoritmo de K-Medoides el cual presenta mayor robustez a datos atípicos. Además de la validación externa, se compararon las matrices de confusión obtenidas (Tabla III) con cada *dataset* y se calcularon los correspondiente índices. En el caso de AA+AF y AA se pueden considerar agrupaciones similares, con un índice de Van Dongen cercano a 0.75. Para las otras dos duplas el índice obtenido se ubica por debajo de 0.6. En principio, y consistente con la validación externa, las agrupaciones obtenidas para AA y AA+AF son similares.

Por último, se evaluó el ”efecto uniforme”: el método de

TABLE III: Matrices de confusión comparativa para la agrupación de cada dataset obtenida con k-medias

	AA+AF		AF			
AA	797	111	663	20	58	93
	37	1260	111	536	464	260
	VDI: 0.749		VDI: 0.568			
AF	748	26				
	0	556				
	66	456				
	94	259				
	VDI: 0.429					

k-medias tiende a formar grupos de tamaño uniforme, aun cuando las clases sean claramente no balanceadas. Para ello se calcula el coeficiente de variación (CV = desvío estándar/media) de la distribución del tamaño de las clases. En forma empírica se mostró que si las clases presentan un CV mayor que 0.85 es bastante posible que el método de k-medias introduzca alguna distorsión en el resultado [4]. En todos los casos se obtuvieron valores por debajo de 0.85.

Según las métricas obtenidas en esta sección, el set de datos que clasifica mejor los géneros musicales es AF, presentando AA un valor de VDI cercano.

TABLE IV: Matrices de confusión de las clases reales contra las distribución en grupos obtenidas con PAM

	AA+AF				AA			AF				
ambient	27	351	61	21	90	274	96	37	352	66	5	0
classical	18	365	7	15	34	313	58	75	320	3	8	1
drum-and-bass	41	0	249	161	391	60	0	7	1	237	9	197
jazz	266	80	51	29	369	47	10	127	112	65	110	12
world-music	311	67	38	47	391	40	32	72	73	45	238	35
VDI	0.578				0.700			0.643				

VI. K-Medoides (PAM)

A. Método

En contraste con el algoritmo k-medias, PAM escoge un punto de los datos como centros y trabaja con una métrica de distancias entre los puntos. Es más robusto ante el ruido que k-medias porque minimiza una suma de disimilaridades (entre pares de puntos) en vez de una suma de distancias euclidianas cuadradas. Así, un medoide puede ser definido como el objeto de un grupo cuya disimilaridad media a todos los objetos en el grupo es mínima: es el punto ubicado más hacia el centro en todo el grupo. Al igual que k-medias, es necesario indicar el número de grupos *k*.
El procedimiento para la selección del número *k* para cada *dataset* es análogo al utilizado para K-medias: se estudia SSE y el índice *silhouette* en función de la cantidad de *clusters*.

B. Resultados y discusión

En la Figura 3 se muestran los gráficos de SSE y coeficiente de *silhouette* en función de *k*, así como el gráfico de *silhouette* para cada *dataset*. En los casos de AA+AF y AA (Figura 3, a y b) se optó por 4 y 3 grupos respectivamente; observando el gráfico de *silhouette* se ven casos mal clasificados (índice negativo) en todos los *clusters*, tanto para AA+AF como para AA. El *dataset* AF fue agrupado en 5 *clusters*, obteniendo el índice de *silhouette* mayor de los tres set de datos y casos mal clasificados en todos sus grupos.

En la Tabla IV se muestran las matrices de confusión contra las clases reales. Para los tres *datasets* se observa que una gran parte de los temas de los géneros *ambient* y *classical* se encuentran en el mismo grupo; de forma similar ocurre con los géneros *jazz* y *world-music* para AA+AF y AA, seaprándose un poco mejor en AF. AA presenta un índice de Van Dongen superior a los demás.
En la Tabla V se muestran las matrices de confusión comparativas para las agrupaciones de cada *dataset*. Según los índices obtenidos las agrupaciones de AA y AF son las que presentan mayor similaridad, con un valor cercano al obtenido al comparar AA+AF con AF. Sin embargo, todo los índices se encuentran por debajo de 0.6.
A diferencia de k-medias con la técnica de PAM fue posible separar en más grupos, observándose las misma tendencias con ambos algoritmos (mismo géneros musicales asignados al mismo *cluster*). En este caso el set de datos que presenta mejor desempeño -según la validación externa- es AA, presentando AF un valor de VDI cercano.

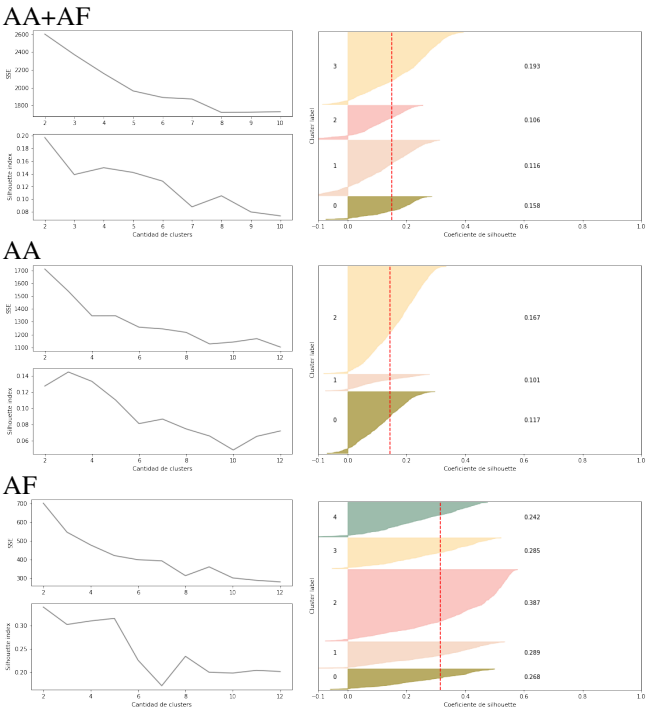

Fig. 3: SSE y coeficiente de *silhouette* en función de la cantidad de *clusters* obtenido con PAM para cada *dataset*; se muestra el gráfico de *silhouette* para el *k* elegido.

TABLE V: Matrices de confusión comparativa para la agrupación de cada dataset obtenida con PAM

	AA+AF				AF				
AA	655	92	342	186	208	141	354	362	210
	8	580	61	85	106	527	60	8	33
	0	191	3	2	2	190	2	0	2
	VDI: 0.537				VDI: 0.587				
AF	179	64	1	72					
	34	793	28	3					
	59	4	347	6					
	346	2	2	20					
	45	0	28	172					
	VDI: 0.340								

VII. DBSCAN

A. Método

DBSCAN (*Density-based spatial clustering of applications with noise*) es un algoritmo de agrupamiento basado en densidad que encuentra un número de grupos comenzando por una estimación de la distribución de densidad de los nodos correspondientes. En este método no es necesario indicar la cantidad de *clusters*, sino que es determinado por el propio algoritmo. Trabaja con dos parámetros que deben ser

ingresados por el investigador: *eps*, la distancia máxima entre dos muestra para que una sea considerada en la vecindad de la otra; y *MinPts*, el número mínimo de muestras en la vecindad de un punto para ser considerado un nodo.
Siguiendo las recomendaciones de Schubert et al. [5] se probaron como valores de *MinPts* el doble de la cantidad de variables del *dataset* y el logaritmo natural de la cantidad de registros. Para estimar el parámetro *eps* se graficaron las distancias obtenidas por KNN (*K nearest neighbour*) con $K = (2xdimensión) - 1$ [5] y se seleccionaron aquellas distancias correspondientes a los puntos de mayor curvatura.

B. Resultados y discusión

Todas las combinaciones probadas para cada uno de los set de datos dio como resultado un único *cluster*, no pudiendo separarlos y obteniendo un grupo etiquetado con valor negativo: estos datos son considerados *outliers* por este algoritmo. Un posible motivo de esta falla podría ser que los datos en estudio presenten alta densidades en su conjunto o densidad similares entre los distintos grupos.

VIII. Visualización en baja dimensión

A. Método

Cuando se trabajo con set de datos de tantas dimensiones es habitual quedarse con un número reducido de autovectores de la matriz que descompone a la covarianza, para así facilitar el análisis y la visualización, o incluso encontrar información que no se evidencia en las dimensiones originales.
Para visualizar en baja dimensión las clasificaciones obtenidas por las diferentes técnicas aquí mostradas, se optó por utilizar la técnica de reducción *T-distributed Stochastic Neighbor Embedding*, conocida por sus siglas TSNE. Este método busca crear una proyección a un espacio de dimensión menor tratando de preservar la distribución de puntos en el espacio original manteniendo los agrupamientos locales. Uno de los parámetros libres de TSNE es *perplexity* (literalmente, perplejidad) que es una forma de medir la cantidad de vecinos que se consideran. Para los tres *datasets* en estudio se obtuvieron los dos componentes principales fijando una perplejidad de 30.

B. Resultados y discusión

En la Figura 4 se muestran para cada set de trabajo los datos graficados en sus dos componentes obtenidos. Se colorean según la clase real, la distribución de clases obtenidas por K-medias y la distribución de clases obtenida por PAM. Además, se detallan en el recuadro de referencia los géneros musicales mayoritarios en cada clase (Según lo mostrado en las tablas de confusión, Tablas II y IV).
Para los tres *datasets* se observa superposición de las clases reales (Figura 4, columna izquierda) en esta proyección. En ninguno de los casos se encuentran 5 grupos bien definidos, observando tan solo uno (*drum-and-bass*) que pareciera estar separado del resto. Los 4 restantes se encuentran superpuestos, observando mayor superposición entre *ambient* y *classical* y *jazz* y *world-music*. Esto está en línea con lo obtenido en las matrices de confusión contra las clases reales (Tablas II y IV).
El algoritmo K-medias (Figura 4, columna del medio) pareciera separar correctamente en dos grupos a AA+AF y AA, con una zona con clases mezcladas en las cercanías entre ellos. En el caso de AF se lo considera, al menos visualmente, el más exitoso: los *tracks* presentan una mayor separación que los otros set de datos y la clasificación obtenida con el algoritmo parece correcta. Se observa que los agrupamientos obtenidos, si bien están compuestos por puntos de más de un género musical, son consistentes con la superposición de las clases reales.
El algoritmo PAM (Figura 4, columna derecha) para el *dataset* AA+AF se considera el menos exitoso de los tres; el género *drum-and-bass* fue distribuido entre dos *clusters*, el cual uno (verde) coincide con la clase real, mientras que el otro (amarillo) presenta una sección bien clasificada y otra no. Este género, que es el mejor separado de los demás en esta proyección, ha sido separado en dos grupos considerados diferentes y extendido a temas de otros géneros. Para el set de datos de AA se observa un gran grupo que abarca los géneros *drum-and-bass, jazz* y *world-music*, no pudiéndose distinguir el primero de los otros dos (pero coincidente con la región en la que se ubican estos tres géneros en el gráfico de las clases reales). Sin embargo, dentro del grupo de *ambient* y *classical*, se distingue una región correspondiente a la primera, bien clasificadas. El *dataset* AF fue agrupado de forma similar al algoritmo k-medias, obteniendo un *cluster* más que separa *drum-and-bass* de *jazz* y *world-music* (según la clasificación de k-medias). Aquí también el género *drum-and-bass* fue separado en dos *clusters* diferentes. Esta separación de un mismo género en dos podría indicar que el mismo presenta variedad en sus descriptores de alto nivel pudiéndose pensar en la existencia de sub-géneros dentro del mismo.
Se considera, tras estos resultados, que el algoritmo PAM aplicado al set de datos AF es el más informativo de los aquí mostrados.

IX. Clustering Espectral

A. Método

Una técnica para analizar el contenido de una serie temporal es comparar los valores de la serie en un instante con otros instantes. Si esto se realiza para los pares de punto de una serie, se puede obtener una matriz de distancias, también llamada de recurrencia en el caso de series de tiempo. Para esta sección se obtuvieron a través de la API de *Spotify* los datos de timbre de *audio_analysis* del tema *AUSLÄNDER* de la banda *Rammstein*.
La serie de tiempo que entrega *Spotify* no tiene una frecuencia de muestreo constante; para corregirlo se interpola la serie para tener una representación lineal en el tiempo. Luego se normalizan los datos y se obtiene la matriz de distancia con la

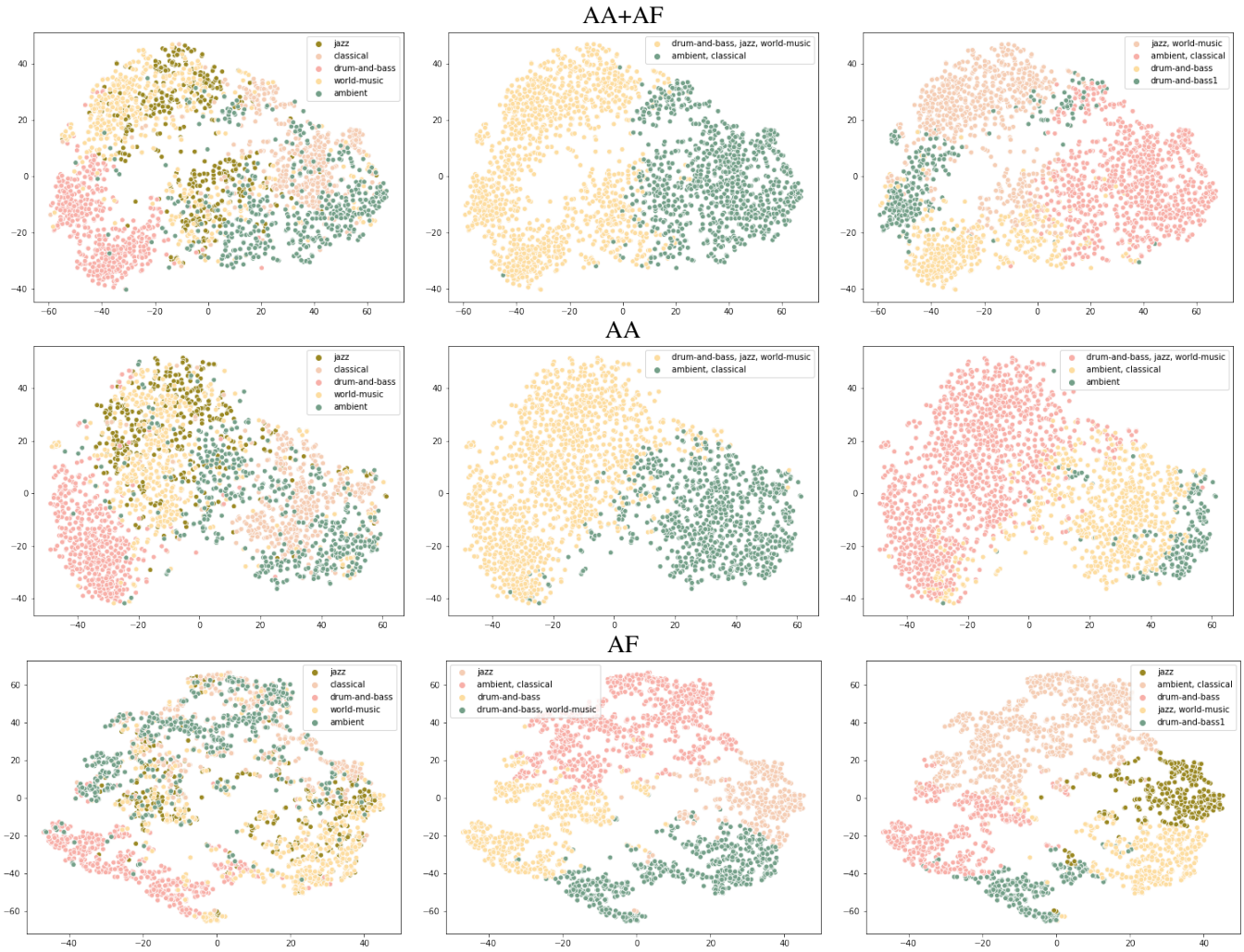

Fig. 4: Visualización en las dimensiones obtenidas con la técnica TSNE. Se observan coloreadas las clases reales (izquierda), las obtenidas por k-medias (medio) y por PAM (derecha). En el recuadro de referencias se explicitan los géneros reales mayoritarios en la clase de ese color.

métrica coseno. Finalmente, para mejorar la representación de esta matriz se indican cuantos vecinos se quieren considerar y se transforman los valores de distancias (disimilaridades) a afinidad.

McFee et al. [6] sugieren realizar *clustering* sobre los primeros autovectores de la matriz de recurrencia; así, se toman los primeros autovectores y se aplica el algoritmo de k-medias para hallar los grupos.

B. Resultados y discusión

Se muestra en la Figura 4 la matriz de afinidad obtenida: arriba la original, con las etiquetas de las distintas secciones del tema identificadas; abajo, coloreada según los grupos encontrados por el algoritmo con k=3. En este último gráfico se observan los tres *clusters* identificados que corresponden a las estrofas (coloreado en rosa) y a los estribillos (coloreados en ocre y verde). Si bien los dos estribillos son ”iguales” el segundo coincide con el final de la canción (sección conocida como *outro*) que presenta característica diferentes al resto del tema; además, en este segundo estribillo pueden

TABLE VI: Secciones distinguibles del tema *AUSLÄNDER* de *Rammstein*. Se muestra el comienzo de cada una en minutos y en segundos. Las letras A y B indican diferencias musicales dentro de la misma sección.

Inicio (m)	Inicio (s)	Descripción
00:00	0	Intro suave
00:16	16	Intro fuerte
00:31	31	Corte
00:33	33	Estrofa 1A
00:50	50	Estrofa 1B
01:06	66	Puente 1A
01:21	81	Puente 1B
01:29	89	Estribillo 1
02:07	127	Estrofa 2A
02:24	144	Estrofa 2B
02:39	159	Puente 2A
02:53	173	Puente 2B
03:01	181	Estribillo 2
03:31	211	Outro

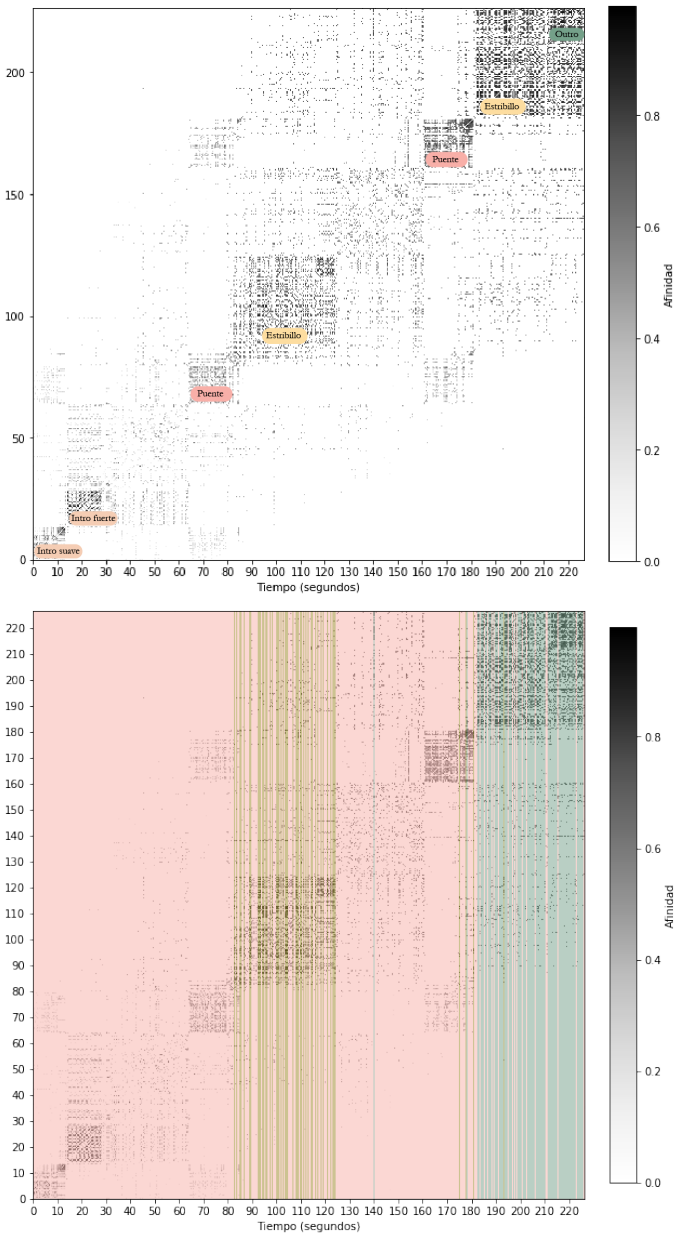

Fig. 5: Matriz de afinidad obtenida para los datos de timbre del tema *AUSLÄNDER* de *Rammstein*. Arriba: matríz original con las secciones del tema etiquetadas. Abajo: se colorean sobre ella los *clusters* obtenidos tras en análisis espectral.

encontrarse líneas color ocre, correspondientes a la clase asignada al primer estribillo. En la Tabla VI se indican, en minutos y en segundos, el comienzo de cada sección que compone al tema. Fue posible identificar, validándolo escuchando la pista, 5 partes fundamentales: intro suave y fuerte, un corte de transición, dos estrofas A y B (estas letras identifican características musicales distintas dentro de la misma sección), dos puentes A y B (una sección de contraste habitualmente posicionada previa al estribillo), estribillos 1 y 2 y outro.

Si bien no es posible identificar todas las partes de un tema musical con esta técnica se logra exitosamente reconocer aquellas partes que se repiten en el tema y presentan diferencias musicales, tales como estribillos y estrofas.

X. Conclusiones

En el presente trabajo se aplicaron algoritmos de *clustering* a tres set de datos de características de alto y bajo nivel de temas musicales: agrupamiento jerárquico aglomerativo, k-medias, PAM y DBSCAN. Con la técnica de *clustering* jerárquico pudo tenerse un primer acercamiento a cómo se agrupan naturalmente los datos de trabajo, observando que AF es el que se distribuye de forma más homogénea en al menos 5 grupos, mientras que los otros *datasets* de trabajo presentan un gran grupo mayoritario. Un comportamiento consistente con esta observación se obtuvo con k-medias, siendo AF el único set de trabajo que presentó las métricas aceptables para dividirlo en 4 grupos. El índice de Van Dongen, métrica de validación externa, fue mayor para este *dataset* que para los otros. Según este mismo índice las clasificaciones obtenidas para AA y AA+AF presentan similaridad. Las métricas de validación interna obtenidas con PAM permitieron separar en más grupos que K-medias, obteniendo, tras la validación externa, un mejor desempeño de AA. La técnica de DBSCAN no pudo ser aplicada ya que, a pesar de buscar los parámetros óptimos, no fue posible detectar más de un grupo.

A pesar de encontrarse desempeños y clasificaciones distintas para cada algoritmo se observó un comportamiento en común para todos: en general distintos géneros musicales eran asignados al mismo grupo, manteniéndose en todos estas mismas duplas (*ambient y classical*, *jazz y world-music*). Esto se vio evidenciado al proyectar los datos y clases reales en dos dimensiones (con la técnica TSNE), en la que se ve la mezcla de géneros. Basado en estas observaciones era esperable que los *clusters* obtenidos por los distintos algoritmos aplicados aquí no coincidan estrictamente con el género musical; sin embargo este resultado no debe considerarse necesariamente desfavorable dado que si la separación en grupos es buena podría tratarse de nuevo conocimiento sobre las formas de caracterizar y asignar un género. Incluso podría, en casos en que un género es asignado a distintos grupos, tratarse de sub-géneros.

En todos los casos AA+AF presentó el peor desempeño, considerándose una mala opción la unión de estos dos *datasets*. Además, los datos de *audio_analysis* fueron resumidos en tan solo dos valores (media y desvío standard), siendo tal vez poca información (o información errónea) para clasificar. El *dataset* sin resumir -de un solo *track*- fue utilizado en la sección de *clustering* espectral, obteniendo los resultados esperados y correctamente validados. Quedará para futuros estudios realizar un preprocesamiento más apropiado con los datos de análisis de audio.

Respecto al algoritmo de DBSCAN, si bien no se obtuvo un agrupamiento, el resultado obtenido podría ser de utilidad al realizar un análisis más exhaustivo de *outliers*. En el presente trabajo se realizó un estudio de selección de variables sin

ahondar en la búsqueda de datos atípicos.

Finalmente, dado que la asignación de géneros no está basada únicamente en métricas cuantitativas, podrían presentarse a estas técnicas como clasificadoras exclusivas del sonido musical. O bien, sería interesante para futuras investigaciones, trabajar con distintos tipos de variables: conservar las binarias e incorporar variables categóricas que introduzcan características no cuantificables que habitualmente se utilizan para la asignación de géneros (tales como país de origen, periodo histórico, instrumentación, temática de su *lyric*, etc).

Se concluye así que, dadas las limitaciones de trabajar únicamente con variables contínuas, los géneros no se separan completamente unos de otros y por consiguiente los algoritmos aplicados no lograron una agrupación adecuada. Sin embargo, conociendo la superposición de clases real, pudo obtenerse una clasificación consistente con ella.

REFERENCES

[1] Van der Merwe, Peter (1989). Origins of the Popular Style: The Antecedents of Twentieth-Century Popular Music. Oxford: Clarendon Press. ISBN 0-19-316121-4.

[2] C. N. Silla Jr., C. A. A. Kaestner and A. L. Koerich (2007). "Automatic music genre classification using ensemble of classifiers," 2007 IEEE International Conference on Systems, Man and Cybernetics, Montreal, Que., 2007, pp. 1687-1692.. doi: 10.1109/ICSMC.2007.4414136

[3] C. Lee, J. Shih, K. Yu and H. Lin (2009). "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features," in IEEE Transactions on Multimedia, vol. 11, no. 4, pp. 670-682, June 2009. doi: 10.1109/TMM.2009.2017635

[4] Wu J. (2012). The Uniform Effect of K-means Clustering. In: Advances in K-means Clustering. Springer Theses (Recognizing Outstanding Ph.D. Research). Springer, Berlin, Heidelberg

[5] Schubert E., Sander J., Ester M., Kriegel H.P., Xu X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Transactions on Database Systems (TODS) Volume 42 Issue 3

[6] [1] McFee, B., Ellis, D. (2014). Analyzing Song Structure with Spectral Clustering. In ISMIR (pp. 405-410)

Appendix A
Listado de variables

Metadata

- album (char)
- artists (char)
- available_markets (char)
- disc_number (int)
- duration_ms (int)
- explicit (bool)
- external_ids (char)
- external_urls (char)
- href (char)
- is_local (bool)
- name (char)
- popularity (int)
- preview_url (char)
- track_number (int)
- type (char)
- uri (char)
- genre (char)

Audio_features

- acousticness (float)
- analysis_url (char)
- danceability (float)
- duration_ms (int)
- energy (float)
- instrumentalness (float)
- key (int)
- liveness (float)
- loudness (float)
- mode (int)
- speechiness (float)
- tempo (float)
- time_signature (int)
- track_href (char)
- type (char)
- uri (char)
- valence (float)

Audio_analysis

- timbre_start (float)
- timbre_0 (float)
- timbre_1 (float)
- timbre_2 (float)
- timbre_3 (float)
- timbre_4 (float)
- timbre_5 (float)
- timbre_6 (float)
- timbre_7 (float)
- timbre_8 (float)
- timbre_9 (float)
- timbre_10 (float)
- timbre_11 (float)
- timbre_filename (char)
- pitch_start (float)
- pitch_0 (float)
- pitch_1 (float)
- pitch_2 (float)
- pitch_3 (float)
- pitch_4 (float)
- pitch_5 (float)
- pitch_6 (float)
- pitch_7 (float)
- pitch_8 (float)
- pitch_9 (float)
- pitch_10 (float)
- pitch_11 (float)
- pitch_filename (char)

Appendix B
Gráficos

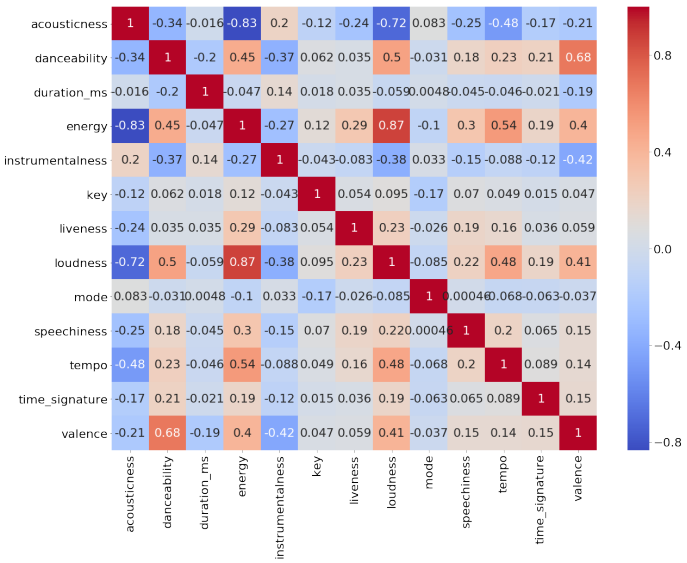

Fig. B.1: Correlograma para el set de datos *audio_feature*. Se observa alta correlación entre las variables *energy*, *accousticness* y *loudness*. Se conserva sólo *energy*.