Del Color del Crimen

Victorien Bigarré

Probabilidad y Estadísticas para la Análisis de Datos

Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Otoño 2025

Resumen

Si existen tiene muchas maneras estudiar la criminalidad, el análisis de datos es una herramienta poderosa porque permite probar rigorosamente las hipótesis confrontándolas con la realidad a una larga escala. Una entre ellas es la idea que existen colores del crimen es decir que existen categorías específicas de crímenes propias a cada clase socioeconómica.

Este informe, estudiando la criminalidad de 360 ciudades francesas en 2021, intenta verificar si existen esos colores. Para eso, dos enfoques son útiles: un clustering de las ciudades (según las características socioeconómicas) para separar las diferentes clases socioeconómica, y estudiar si aparece en cada clúster una categoría especifica de criminalidad. Los resultados mostraran que, si las clases rurales y urbanas tienen más o menos las mismas criminalidades, la de las clases desfavorecidas es mucha más violente mientras que la de la población estudiantina es menos. El gradiente de la tasa de crimen en cada ciudad permite encontrar cual son las características socioeconómicas que favorecen una categoría de criminalidad, y medir la contribución de cada factor. Un segundo enfoque de clustering de los crímenes (según las características socioeconómicas) para ver como se concentran las diferentes categorías de criminalidad en el paisaje socioeconómico, aunque los resultados no sean significativos.

Plan

1. Introducción

Si el crimen siempre ocupaba un lugar importante en las sociedades, hay desde el fin del previo siglo una voluntad de objetivar el estudio de la criminalidad transformándola en una ciencia. Paralelamente, hay un *populismo penal* que se desarrolla en las sociedades occidentales que afecta durablemente las políticas de la prevención y de la represión de la criminalidad. En esa atmosfera aparece relevante usar las herramientas matemáticas del análisis de datos para contribuir a las conclusiones de la criminología. Así, el enfoque de este estudio es estudiar la hipótesis académica que existe una *coloración* del crimen definida como una sobre-representación de un tipo de criminalidad en un tipo de clase social dado (coloración roja si hay muchos crímenes de sangre, blanco si hay muchos crímenes financieros etc.…)

Para tratar este tema, empezaremos con una presentación de los datos y una exploración breve de estos para entender que son los sesgos que pueden parasitar las conclusiones. Después, se propondrá un método de análisis de los datos que consiste en un clustering de las ciudades con los indicadores socioeconómicos de manera a separar las diferentes clases sociales, y estudiar las proporciones de cada categoría de criminalidad en eses clústeres para identificar las coloraciones posibles de la criminalidad. Además, será construido un gradiente que permitirá identificar los factores que contribuyen a cada categoría de criminalidad. Para profundizar, se propondrá un segundo enfoque de clustering de los crímenes para estudiar cómo se distribuyen en las sociedades.

Al final, se presentarán las conclusiones que muestran si los niveles económicos tienen una coloración de criminalidad. Además, una comparación con otros algoritmos de clustering será hecho siguiendo el *Common Task Framework* para explorar posibilidades de mejoramiento.

Además, a lo largo del informe, los parágrafos al fin de cada sección explicaran los problemas encontrados y las soluciones construidas.

2. Presentación de los datos:

Dataset

Por sus naturalezas, los datos a propósito de los crímenes son filtrados por las instituciones que los conservan. Así, el objetivo es cruzar los datos de criminalidad por zonas con el perfil socioeconómico esas zonas.

Así el dataset es constituido de:

- TABLA_1: números de crímenes por Comisaria General de Provincias (después llamados como CGP) y por crímenes; Fuente: Ministerio del Interior, 2021.
    - Las columnas: 102 crímenes y delitos mezclados: son el conjunto el más grande y variados encontrado. (Vea Anexo)
    - Las líneas: 372 Comisarias General Departamental de Francia
- TABLA-2: estadísticas de las circunscripciones. Fuente: Instituto Nacional de la Estadística y de los Estudios Económicos.
    - Las columnas: 100 variables diferentes sobre el paisaje socioeconómico de cada circunscripción Vea Anexo)
    - Las líneas: 577 circunscripciones de Francia
- TABLA 3: vinculador de las ciudades a las circunscripciones. Fuente: Instituto Nacional de la Estadística y de los Estudios Económicos.
    - Las columnas: 6 características administrativas (número de circunscripción, de región…) y una variable binaria E/P que indica si la ciudad en la línea es integralmente en la circunscripción o parcialmente.
    - Las líneas: 35133 ciudades de Francia

La fusión de estas tres tablas será utilizada como dataset para obtener:

- 359 líneas que son las ciudades donde hay los Comisarias General de Provincias
- 100 primeras columnas que son las características socioeconómicas
- 102 columnas después que son las tasas de crímenes

Clasificación de McDonald:

La cuestión de la clasificación de los crímenes no es una cuestión que es académicamente fácil a contestar para dos razones. La primer es que las disciplinas que lo intentaron lo hicieron con sus propios puntos de vista por ejemplo el derecho tiene una escala de gravedad más que de manera de hacerlo. La segunda es que categorizar los crímenes implica poner en casilla un fenómeno que tiene multitud de formas: ¿por ejemplo, como categorizar la ayuda a los inmigrantes ilegal? En 1933, un profesor del derecho a la *Cornell Law University*, propuso una clasificación general de los crímenes basándose sobre los *métodos de acción* de los crímenes, con el objetivo para permitir una clasificación que sea utilizable y académicamente construida. Lo usé para construir un diccionario que grupa los tipos de criminalidad de mi dataset según las categorías definidas por McDonald.

Problemas encontrados

Aquí lo más difícil fue encontrar datos que permiten estudiar el fenómeno de la distribución del crimen. Buscar los datos implica pensar a como usar las, que son los modelos que pueden ser relevante y a qué tipo de conclusión pueden llegar. Es por lo que usé 3 tablas diferentes: no existe dataset ya construido para este tipo de estudio entonces tuve que construirlo yo basándome sobre los datos oficiales.

Además, seguir la clasificación de McDonald subraya el problema de organizar el crimen según los métodos utilizados: no todos los crímenes pertenecen siempre a una sola

categoría, y las categorías no siempre existen. Decidí concentrarme en quién sufre más del crimen para decidir en qué categoría clasificarlo.

3. Sesgos, observaciones e hipótesis

Como muchos estudios societales, hay dificultades a entender lo que subrayan las cifras. Las conclusiones de ese papel deben ser entendida y manipulada con precaución dado que existen sesgos e hipótesis.

- Un sesgo de realidad: Los datos analizados provienen de delitos registrados por la policía y la gendarmería francesa, pero no siempre fueron objeto de una investigación judicial. Por lo tanto, pueden contener falsos positivos. Aun así, se consideran como la mejor aproximación disponible para captar las dinámicas ilegales en la sociedad.

- Un sesgo de detección: Ciertas poblaciones son más controladas que otras por las fuerzas del orden, debido a prejuicios estructurales como el racismo. Esto genera una sobrerrepresentación de crímenes en los estratos sociales bajos, no necesariamente porque cometan más delitos, sino porque son más vigilados.

- Un sesgo de denunciación: Algunos delitos, como los crímenes sexuales, son menos denunciados por razones culturales o psicológicas (tabú, miedo). Esto provoca una sobrerrepresentación estadística que debe ser tenida en cuenta al interpretar los resultados.

- Hipótesis de zona: Dado que los datos criminales están asociados a comisarías que cubren zonas algo más amplias que las ciudades mismas, se ha optado por utilizar las circunscripciones legislativas como unidad de referencia. Estas ofrecen un compromiso razonable entre la escala municipal y la departamental para analizar el contexto socioeconómico.

Problemas encontrados

Esa sección pidió una reflexión y un conocimiento mínima sobre los sesgos que pueden existir en la detección de los crímenes que hay en la literatura. Si usé lo que ya sabía, añadí algunos sesgos que encontraron otros alumnos cuando presenté mi dataset, y hablando con ellos mejoré la comprensión de los sesgos que pueden parasitar mi estudio.

4. Metodología

Para crear el dataset final tuve que vincular cada ciudad con las circunscripciones a las que pertenece, utilizando la información proporcionada en la tabla 3. En los casos en que una ciudad se encuentra totalmente contenida dentro de una única circunscripción, se asumió que el perfil socioeconómico de dicha circunscripción representa directamente el perfil

socioeconómico de la zona que cubre el CGP de esa ciudad. Por otro lado, cuando una ciudad se extiende sobre varias circunscripciones, se consideró que su paisaje socioeconómico corresponde al promedio de los perfiles de todas las circunscripciones que la abarcan.

Para garantizar la comparabilidad entre las distintas ciudades, se aplicó un tratamiento específico a los datos. En primer lugar, reemplacé los datos brutos de criminalidad por las tasas de crímenes por habitante, lo que permite corregir el sesgo asociado al tamaño poblacional y obtener un indicador proporcional de la actividad delictiva. Estas tasas se calcularon dividiendo el número de delitos registrados en cada comisaría general de provincia por la población correspondiente de su circunscripción. En segundo lugar, todas las variables, tanto socioeconómicas como delictivas, fueron estandarizadas mediante la transformación z-score —es decir, con media cero y desviación estándar igual a uno— con el fin de eliminar las diferencias de escala entre variables y evitar que alguna de ellas domine en los análisis multivariados posteriores.

El primero enfoque se centró en agrupar las ciudades del dataset mediante un algoritmo de *k-means*, utilizando como base sus perfiles socioeconómicos promedio. Así, dos ciudades se consideran próximas si presentan características socioeconómicas similares. Una vez definidos los clústeres, se analizó dentro de cada ellos la proporción de los distintos categoría de crímenes siguiendo la clasificación de J.W.McDonald, con el fin de identificar posibles patrones delictivos asociados a contextos sociales y económicos específicos. A partir de ello, las tendencias predominantes de criminalidad aparecen en las diferencias de proporción de cada categoría de crimen.

Para estudiar el gradiente de criminalidad, se proyectaron las ciudades en el espacio definido por las componentes principales obtenidas mediante un análisis de componentes principales (PCA). A continuación, se realizó una regresión lineal de los índices de criminalidad por categoría en función de estas componentes principales. Este enfoque permitió identificar cómo varían los niveles de criminalidad según las distintas dimensiones latentes del perfil socioeconómico, y así determinar qué factores estructurales contribuyen al aumento o a la disminución de cada categoría delictiva.

El segundo enfoque, más experimental, consistió en aplicar un algoritmo de *k-means* sobre los distintos tipos de crímenes del dataset, utilizando como variables los perfiles socioeconómicos promedio de las ciudades en las que cada tipo de delito es más frecuente. De este modo, dos tipos de crímenes se consideran similares si tienden a concentrarse en contextos socioeconómicos con características comparables. Una vez definidos los clústeres, se comparó la agrupación obtenida con la clasificación propuesta por McDonald, con el objetivo de evaluar el grado de correspondencia entre ambos enfoques. Esta comparación permitió analizar si los tipos de delitos pueden efectivamente vincularse con perfiles socioeconómicos específicos y, por tanto, si ciertas condiciones sociales y económicas predisponen a determinadas formas de criminalidad.

En cada enfoque, estudié la eficacia de los algoritmos utilizando diferentes métricas de evaluación (Silhouette, Inertia, Davies-Bouldin, Calinski-Harabasz, etc.) para optimizar sus parámetros. Probé sistemáticamente un rango amplio de valores para el número de clústeres, y todos los indicadores coincidieron en que el valor óptimo era 4. Además, proyecté los resultados sobre las componentes principales para confirmar visualmente la separación de los grupos, y verifiqué la estabilidad de los clústeres con múltiples inicializaciones del algoritmo.

Problemas encontrados:

Esa etapa de estructuración y de tratamiento de los datos me pidió 40% del tiempo de trabajo. En un primer tiempo porque encontrar datos relevantes sobre la criminalidad es difícil porque las instituciones no divulgan mucho esos tipos de datos. Así, del inicio del proyecto pensé a como estructurar en su globalidad el análisis para ser cierto que los datos puedan ser útiles a lo que querría buscar. Además, una vez halladas, tuve que fusionar las tablas con claves que no fueron las mismas: por ejemplo, las ciudades con caracteres especiales o acentos o que empiezan con "Saint-"fueron en la tabla de los crímenes abreviada en "St". Así, si para 80% de las ciudades he fusionado las tablas fácilmente con códigos, para esas ciudades especiales tuve que verificar una por una los datos que los fueron asignados. Esa experiencia me enseno la importancia tener datos que sea buen estructurados con claves que sean las mismas entre las diferentes tablas.

Para las otras etapas, el más difícil fue entender loque hace el algoritmo para asegurarme que interpreto bien los resultados. Por eso, elegí desplegar cada vez los crímenes los mas frecuentes para cada cluster, y el perfil económico para analizar con múltiples puntos de vista lo que hace el código.

5. Análisis exploratorio

Repartición de los crímenes

Dos aspectos de los datos muy relevante aparecen en el análisis exploratorio: primero los crímenes los más frecuente (acuerdo a la clasificación de McDonald) son los contra las cosas (705,000), la criminalidad de sangre (140,000) para un promedio de 100,000. Todos los otros son abajo 50,000 ocurrencias (Fig1). Así esperamos tener una gran proporción de crímenes de esas categorías en cada clúster. Además, para los crímenes los menos numerosos, los resultados pueden ser más varios porque un pequeño tamaño no permite suprimir los sesgos. además, aproximadamente 75% de las ciudades son en una región rural. Así el paisaje criminalista que tenemos es sesgado para representar más la criminalidad rural.

Análisis de los Componentes Principales

Antes de realizar el clustering, realice un análisis de componentes principales que subraya las 3 dimensiones principales que resumen los datos socioeconómicos. Así obtuve: PC1 con un escoré de 30% que abarca los perfiles de las clases rurales: los criterios que tienen una

contribución positiva la más importante describen un ámbito rural. La PC2 con un escore de 17% abarca los perfiles de las clases urbanas favorecidas. La PC3 abarca las estratas desfavorecidas de la sociedad urbana. Esas tres componentes explican 62% del fenómeno y servirán después a identificar las características que influyen sobre el gradiente de criminalidad.

6. Resultados

Enfoque 1

En una primera etapa, logré construir una cartografía de la criminalidad en función de las estratificaciones sociales. Para ello, realicé un clustering de las ciudades utilizando el algoritmo k-means, con el fin de reagruparlas según similitudes en sus perfiles socioeconómicos. Tras haber evaluado diferentes valores de k entre 1 y 10, el valor óptimo se identificó con $k = 4$, alcanzando una puntuación de *silhouette = 0.33*, *Davies–Bouldin = 1.004* y *Calinski–Harabasz = 177*. Estos resultados pueden considerarse razonablemente buenos, aunque los escores indicen que los clústeres no están perfectamente separados.

Basándome en las características que más contribuyeron a la formación de cada clúster (evaluadas por su posición en el espacio de las tres componentes principales), pude identificar cuatro grandes grupos sociales: población favorecida urbana, población rural, población desfavorecida urbana y población estudiantil/joven activa urbana.

Para cada clúster, calculé la proporción de cada categoría de crimen —según la clasificación de McDonald— con el objetivo de observar si ciertas formas de criminalidad predominan en determinadas estratificaciones sociales.

Los resultados muestran que todas las categorías de crimen están presentes en cada clúster, pero con variaciones significativas en sus proporciones, lo cual puede señalar ciertas tendencias. Así, los dos primeros clústeres —población rural y población favorecida urbana— presentan perfiles criminales muy similares: aproximadamente un 12% de crímenes de sangre y un 59% de crímenes contra las cosas. El clúster de estudiantes y jóvenes activos urbanos muestra también un perfil cercano, con un 10% de crímenes de sangre y un 62% contra las cosas.

Lo más destacable es el clúster asociado a las clases desfavorecidas urbanas, donde se observa una caída de 10 puntos porcentuales en los crímenes contra las cosas (50%) y un aumento equivalente en los crímenes de sangre (21%). Esto sugiere que esta última estratificación social está más expuesta a una criminalidad violenta, mientras que en los otros grupos prevalece una criminalidad más relacionada con daños materiales o delitos contra la propiedad. (Fig 2)

Gradientes

Para identificar los factores que catalizan o frenan una determinada categoría de criminalidad, proyecté las ciudades en el espacio definido por las tres primeras componentes principales y analicé el gradiente de las tasas de criminalidad por categoría. Esto me permitió construir un vector cuyos componentes corresponden a los parámetros socioeconómicos: un coeficiente positivo indica que un aumento del parámetro está asociado a un incremento en esa categoría de crimen, mientras que un coeficiente negativo sugiere lo contrario.

Los catalizadores son aquellos factores con mayor contribución positiva al gradiente, mientras que los factores de freno son aquellos con coeficientes negativos de gran magnitud. Esta herramienta permite identificar de manera más precisa los factores criminógenos —o al menos significativamente correlacionados con los crímenes— y puede ser útil para orientar con mayor eficacia las políticas públicas dirigidas a la prevención y reducción del delito. (Ejemplo Fig 3)

Enfoque 2:

Por el contrario, intenté aplicar el enfoque inverso: en lugar de agrupar las ciudades según los crímenes, quise agrupar los crímenes según el perfil socioeconómico de las ciudades donde ocurren. Es decir, dos tipos de crímenes se consideran similares si tienden a producirse en contextos socioeconómicos parecidos. Para ello, realicé un clustering con K-means que dio lugar a 4 clústeres (silhouette = 0.11, Davies–Bouldin = 2.19 y Calinski–Harabasz = 55.8), resultados que no son particularmente fiables.

Sin embargo, los clústeres obtenidos parecen corresponder a: uno de criminalidad baja, otro centrado en delitos contra la propiedad, un tercero con crímenes violentos, y un último con crímenes muy violentos. Es posible relacionarlos con las clases sociales observadas anteriormente, pero para ello es necesario analizar al menos los 10 o 15 principales factores socioeconómicos, ya que los más relevantes se asemejan entre sí y no permiten una distinción clara.

Esto puede deberse a un sesgo no identificado, o simplemente a que no existe una relación biunívoca clara entre clase social y tipo de criminalidad, sino únicamente ciertas tendencias difíciles de captar. Por tanto, este enfoque inverso no aporta mucho más al análisis ya realizado con el primer enfoque.(Fig 4)

Problemas encontrados:

Entre los distintos desafíos que presentó este proceso, lo más difícil para mí fue diseñar el modelo: cuál es la mejor manera de poner a prueba la hipótesis? Los resultados obtenidos evalúan realmente lo que quiero demostrar, o no? Comprender lo que realmente hace el algoritmo para poder interpretar correctamente los resultados fue lo que más tiempo me tomó.

Para ello, analicé las características socioeconómicas de cada clúster y los crímenes más representativos, con el fin de entender el núcleo de cada grupo. Esa fue precisamente la razón por la cual paralelamente al primero desarrollé un segundo enfoque: quería comprobar si era posible medir el fenómeno desde otra perspectiva, y si abordarlo de otra manera podía aportar más claridad. Es solo usando indicadores (como el *silhouette score*) me ayudó para elegir el cual es el mejor entonces me concentré en el enfoque más interesante: el clustering de las estratificaciones sociales, el análisis de la distribución de las categorías de crimen según la clasificación de McDonald, y finalmente el uso del gradiente para perfeccionar el estudio y afinar la comprensión del fenómeno.

7. Pista de Mejoramiento

Después estos resultados, intenté mejorarlos resultados para tener un modelo más interesante. Querría cambiar el algoritmo de clustering siguiendo el método del Common Task Framework fijé:

- Tarea: clusterizar las ciudades según las variables socioeconómicas para separar las clases sociales.
- Fijé los datos: las ciudades con las características socioeconómicas
- Las medidas de eficacidad: silhouette, Davies–Bouldin y Calinski–Harabasz
- No fijé el número de clúster, la idea es evaluar a su optimo cada algoritmo de clustering

Así, intenté comparar el K-mean con el agglomerative que puede ser más interesante para las formas de clúster no esféricas. El agglomerativo utilizado aquí puede ser parado a un numero de k cluster elegido a priori. ¨Porque queremos estudiar las performances de cada algoritmo a su optimo, comparamos para todo k el K-Mean y el *agglomerative*. Lo que aparece es que para según el escore silhouette el *agglomerative* es mejor para 6 clústeres, pero según la medida de Davies-Bouldin es mejor solo después 8 clústeres y es menos interesante para todos los k según el de Calinski-Harabasz.

Así, usando el Comon Task Framework para comparar los algoritmos de clustering a sus óptimos, aparece que el k-mean con 4 clústeres da los resultados con escore de compacidad mejores. Fig 5

8. Conclusiones

Este estudio pone de relieve dos principios relevantes en el estudio de los crímenes. Primero, no hay predisposiciones intrínsecas a las clases sociales que vinculan estrictamente las clases sociales a una categoría de criminalidad, es decir que todos los crímenes existen en todas las estratas de la sociedad. Segundo, existen factores que catalizan la criminalidad orientándola hacia una categoría: la población estudiantina sufre más probablemente de criminalidad de cosas mientras que las poblaciones desfavorecidas sufren más de los crímenes de sangre. De manera general, la pobreza, una situación personal complicada y un nivel de estudio bajo son catalizadores de criminalidad mientras que el acceso a un empleo y a los servicios públicos son correlacionados a un bajo nivel de criminalidad.

Además de los sesgos que afectan los crímenes, el estudio puede ser mejor con un base de datos más variados (considerando todos los crímenes por zonas y no solo los de las comisarias generales) y un número más grande de clústeres para tener perfiles más precisos de las clases sociales, aunque eso altere la cualidad del clustering. Un punto de vista de un profesional de la criminalidad puede ser muy relevante para interpretar las variaciones de las proporciones las categorías que no sé cómo interpretar como la sobrerrepresentación de delictivos “pequeño” en las clases desfavorecidas (+20% que el promedio de los otros clústeres) o contra la policía (+80%) que pueden resultar de un contacto más frecuente con la policía que favorece las interacciones hostiles.

9.Graficos

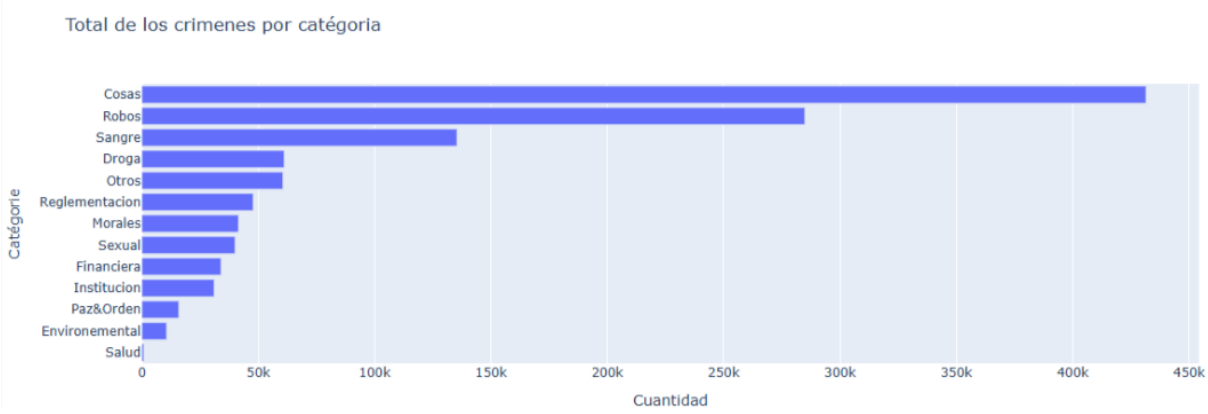

Fig 1

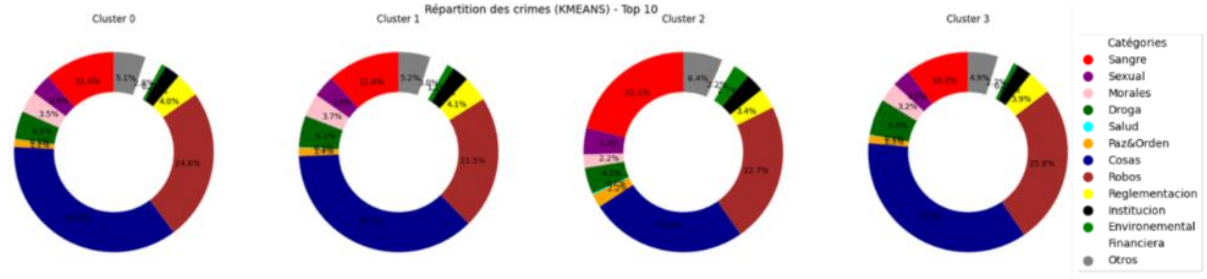

Fig 2

Fig 3

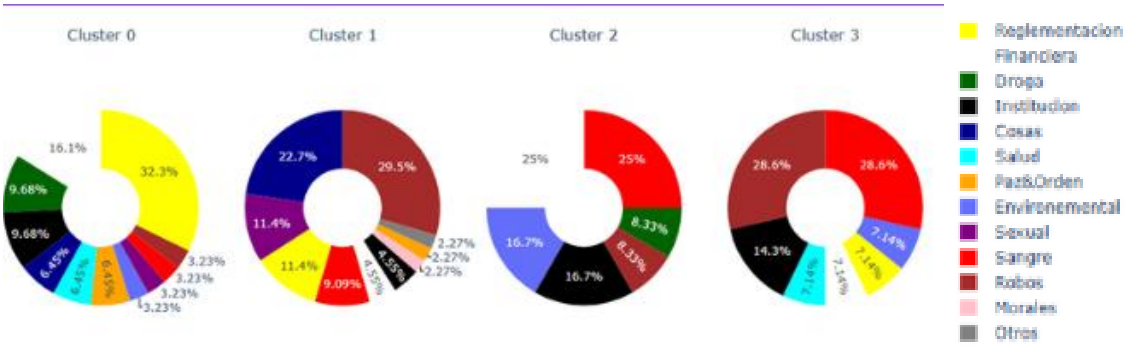
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Reglamentacion
Financiera
Drogo
Institucion
Cosas
Salud
Paz&Orden
Environemental
Sexual
Sangre
Robos
Morales
Otros

Fig 4

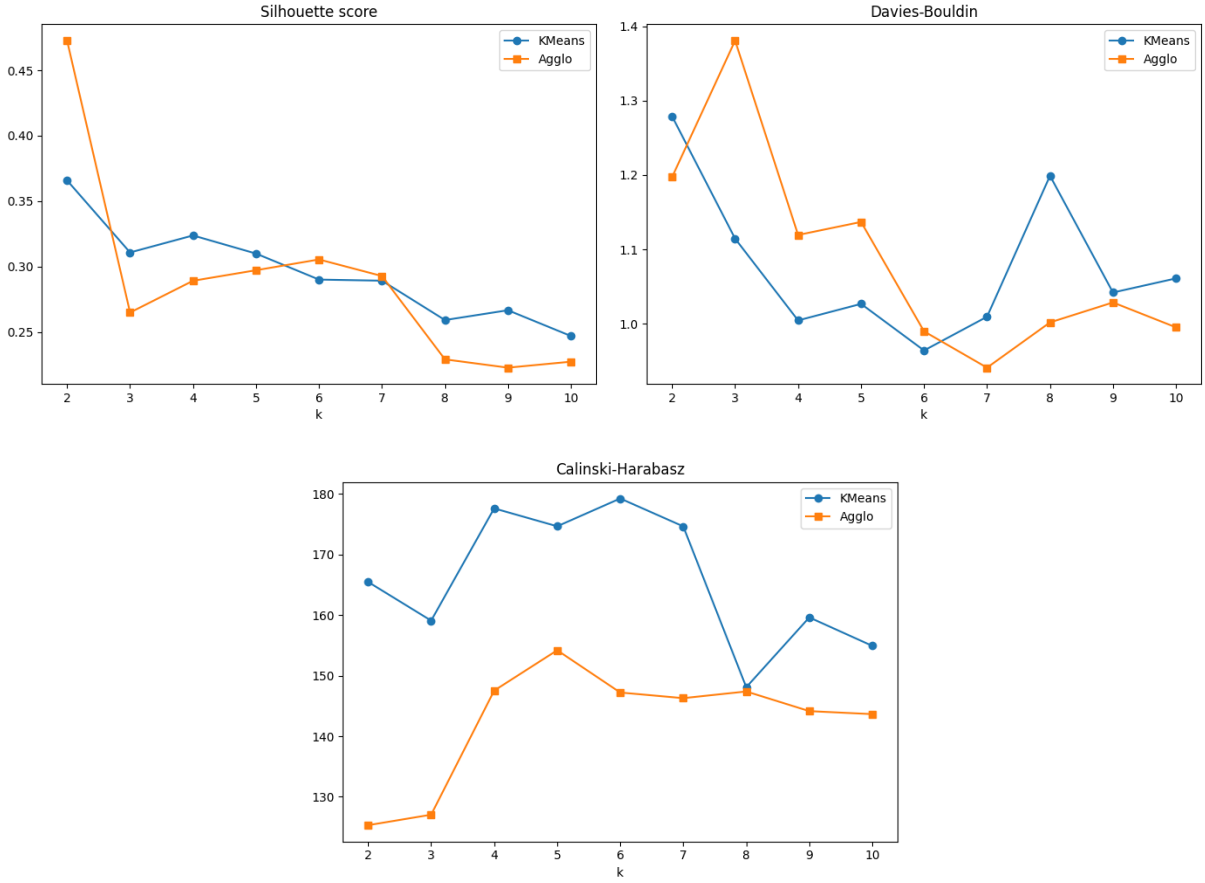
Silhouette score
Davies-Bouldin
Calinski-Harabasz
KMeans
Agglo
k

Fig 5