Inteligencia de Negocio

P2 -Segmentación para Análisis Empresarial

Grado Ing. Informática

Francisco Carrillo Pérez
Grupo 1 martes a las 09:30

Contents

List of Figures

List of Tables

1 Introducción

En este caso se va a trabajar con un conjunto de datos que correspondes a los accidentes ocurridos en España durante el año 2013. El número de accidentes es de 89.519.

Cada valor de este conjunto de datos se va a ver definido por las siguientes variables:

- **MES**: Variable numérica que nos indica el mes del accidente.
- **HORA**: Variable numérica que nos indica la hora del accidente.
- **DIASEMANA**: Variable numérica que nos indica el día de la semana.
- **PROVINCIA**: Variable categórica con el nombre de la provincia donde se produjo el accidente.
- **COMUNIDAD_AUTONOMA**: Variable categórica con el nombre de la comunidad autónoma donde se produjo el accidente.
- **ISLA**: Variable categórica donde se indica si donde se produjo el accidente es una isla o no.
- **TOT_VICTIMAS**: Variable numérica que indica el total de víctimas en el accidente.
- **TOT_VICTIMAS30D**: Variable numérica que indica el número de víctimas que se esperan en los 30 días posteriores al accidente.
- **TOT_MUERTOS**: Variable numérica que indica el total de muertos en el accidente.
- **TOT_MUERTOS30D**: Variable numérica que indica el número de meurtos que se esperan en los 30 días posteriores al accidente.
- **PRIORIDAD**: Variable categórica que indica si existía algún tipo de prioridad.
- **SUPERFICIE_CALZADA**: Variable categórica que indica el estado de la calzada en el momento del accidente.
- **LUMINOSIDAD**: Variable categórica que indica el estado de luminosidad en el momento del accidente.
- **FACTORES_ATMOSFERICOS**: Variable categórica que indica el estado atmosférico en el momento del accidente.
- **VISIBILILDAD_RESTRINGIDA**: Variable categórica que indica si la visibilidad se encontraba restringida.
- **OTRA_CIRCUSTANCIA**: Variable categórica que indica si existía alguna otra circustancia para el acaecimiento del accidente.
- **ACERAS**: Variable categórica que indica si existía acera en el lugar del accidente.
- **TIPO_ACCIDENTE**: Variable categórica que indica que tipo de accidente ha sido.
- **DENSIDAD_CIRCULACION**: Variable categórica que indica cuál era la densidad de la circulación a la hora del accidente.
- **MEDIDAD_ESPECIALES**: Variable categórica que indica si se ha producido alguna mediada especial.

1.1 Algoritmos de clustering utilizados

Se han elegido cinco algoritmos diferentes para analizar los distintos casos de uso.

Estos cinco algoritmos son los siguientes:

- **K-means**: algoritmo de clustering que agrupa los datos en K grupos en el que cada dato pertenece al cluster cuyo valor medio es el más cercano.
- **Agglomerative Clustering con método Ward**: algoritmo de agglomerative clustering el cuál utiliza el método Ward para elegir que dos clusters se mezclan en cada paso basándose en el valor óptimo de la función objetivo.
- **BIRCH**: algoritmo de clustering en el que se construye una estructura de árbol en el cuál los centroids de los clusters se leen en las hojas. Estos pueden ser usados como los centroids finales de los cluster o se pueden utilizar como la entrada para un algoritmo de Agglomerative clustering.
- **DBSCAN(Density-Based Spatial Clustering of Applications with Noise)**: algoritmo de clustering que encuentra datos base con densidad alta y expande los clusters a partir de estos datos. Es un buen método para datos que contienen clusters con densidad similar.
- **Spectral Clustering**: algoritmo de clustering el cual hace uso de los eigenvalues de la matriz de similaridad de los datos para realizar una reducción dimensional antes de crear los cluster con menos dimensiones.

2 Caso 1: accidentes de colisión de vehículos en una vía convencional

En este caso se han decidido analizar aquellos accidentes de colisión de vehículos que se han producido en una vía convencional.

Table 1: Resultados en el caso 1 para los distintos algoritmos

Nombre	K	Calinski-Harabaz Index	Silhouette Coefficient
k-means	8	41525.010	0.95303
ward	2	0.0	0.0
birch	1	0.0	0.0
dbsscan	6	6267.617	0.94433
spectral	8	32.792	-0.92600

Los valores de las métricas que podemos observar en la Tabla 2 nos indican que en el caso de la métrica Silhouette los algoritmos **k-means y dbsscan** son los que mejores resultados obtienen, quedando mejor posicionado el k-means. Se observa también como el algoritmo **spectral** obtiene el peor resultado. Esto lo que nos indica es que sus parámetros no están bien definidos.

2.1 Análisis de parámetros

En esta sección se va a hacer un pequeño estudio modificando los parámetros de los algoritmos para observar si se consiguen mejoras.

Table 2: Resultados de las modificaciones en el caso 1 para los distintos algoritmos

Nombre	Modificación	K	Calinski-Harabaz Index	Silhouette Coefficient
k-means	k=3	3	4428.446	0.94926
ward	k=3	3	0.0	0.0
birch	k=3	3	4428.446	0.94926
dbsscan	eps=0.01	6	3445.617	0.94331
spectral	k=3	3	44.719	-0.87794

En la Figura 1 podemos observar una comparación entre los valores del **Calinski-Harabaz Index** entre cada algoritmo con sus valores por defecto y la modificación realizada que se puede observar en la Tabla 2.1. Con la modificación, todos los algoritmos obtienen peores valores excepto el algoritmo birch, que mejora notablemente su puntuación con esta métrica.

A continuación, se ha deseado comparar los valores en la métrica **Silhouette Coefficient** en la Figura 2. Se puede observar como la modificación y el por defecto en el algoritmo k-means y en el algoritmo dbscan no varían mucho. Sin embargo se puede observar una mejoría con las modificaciones en los algoritmos birch y spectral.

Figure 1: Comparación de los valores del Calinski-Harabaz Index en el caso1

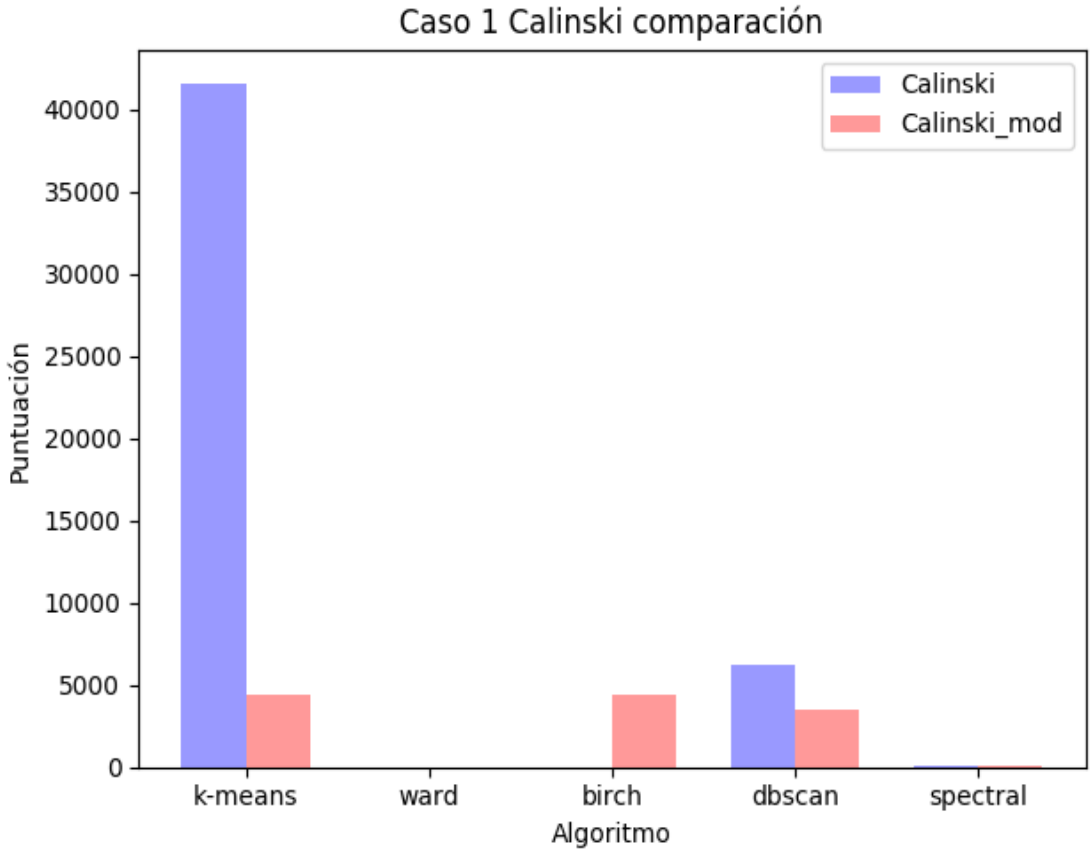
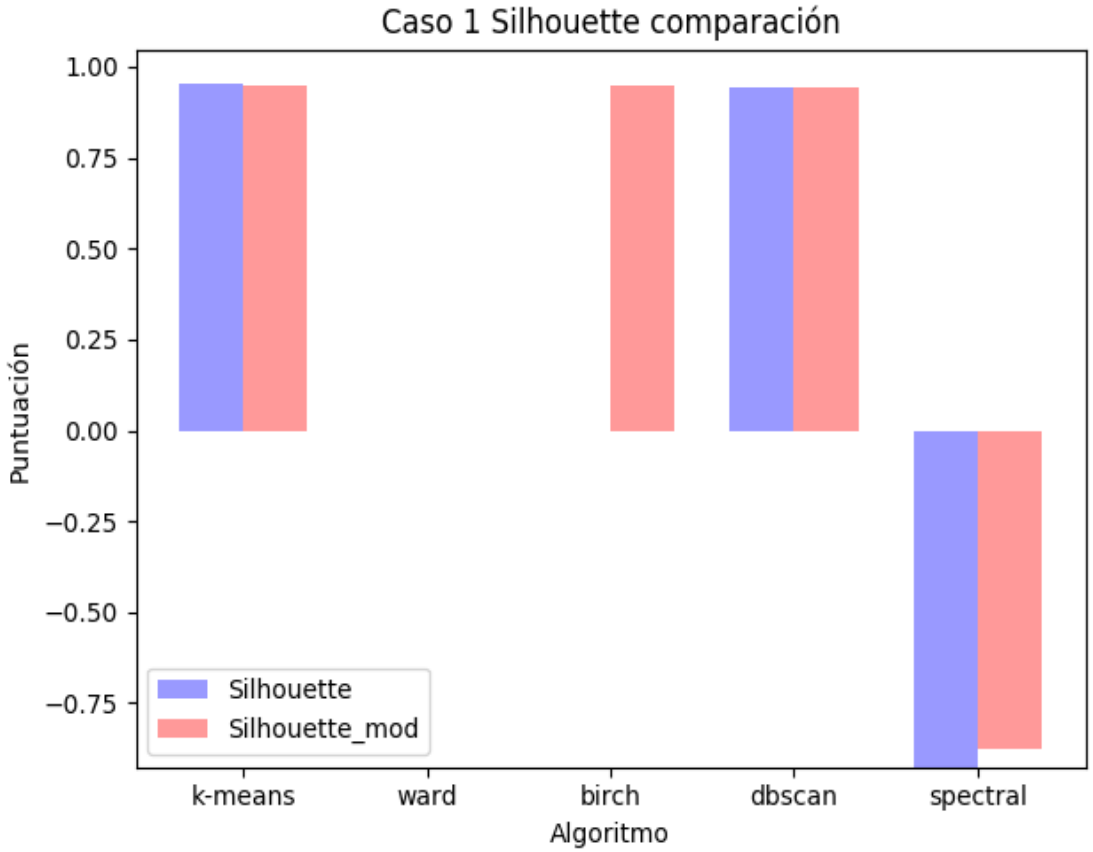


Figure 2: Comparación de los valores del Silhouette Coefficient en el caso 1



2.2 Interpretación de la segmentación

En la Figura 3 podemos observar el scatter matrix en el caso 1 para el algoritmo K-means. En este caso, el número de clusters que determinan de forma dinámica. Como se puede observar, el número de clusters es grande ya que **k=8**, esto lo que nos puede indicar es que el problema es u problema complejo donde la separación no es clara o los datos están muy mezclados y por ello se necesita un número grande de clusters para poder hacer una separación.

También se puede observar en la Figura 3 el **cluster 1** (el de color azul oscuro en el scatter matrix) es el cluster predominante dentro de nuestros datos. Esto lo que puede significar es que la mayor parte de nuestros datos se encuentran muy juntos y por ello pertenecen a este cluster, pero luego se encuentran puntos lo bastante separados y distantes entre sí como para que el algoritmo determine que deben pertenecer a un cluster a parte. Como se ha decidido no realizar ningún filtrado del número de puntos mínimos que deben pertenecer a cluster para que se le considere como tal, puede ser que estos clusters estén formados por muy pocos puntos comparados con el cluster 1.

En la Figura 4 se puede observar una visualización en 3D de los distintos clusters. He de decir que el número de puntos es inferior al que realmente se encuentran en el subset del caso utilizado ya que muchos puntos tenían los mismos valores. Se puede observar como hay un número de puntos que se aglomeran mucho en la valores cercanos a 0 para las variables **TOT_MUERTOS y TOT_HERIDOS_GRAVES** pero que varían con respecto a **TOT_HERIDOS_LEVES**. Por desgracia la escala de colores no es la misma que en la Figura 3 pero se podría diferir que esos puntos corresponden a lo que sería el cluster 1 en dicha figura. A partir de ahí se puede observar como el resto de puntos se distribuyen por todo el espacio, lo que da pie a la creación de muchos

clusters, que es lo que realiza el algoritmo finalmente.

También se deseaba hacer un análisis de este caso de estudio con otras variables, es por ello que se usaron otras variables para sacar el número de cluster con k-means. Las variables que se deseaban utilizar era el total de heridos leves, el mes y el día de la semana. Se puede observar en la Figura 5 como se siguen encontrando 8 clusters y como no existe una relación clara entre estas variables. Parece que existen el mismo número heridos leves indiferentemente del día de la semana o del mes. Yo pensaba que los últimos días de la semana existirían más heridos leves y es cierto que parece que existe un poco más de densidad, pero no esta clara esta relación.

Figure 3: Scatter Matrix en el caso 1 con el algoritmo K-Means

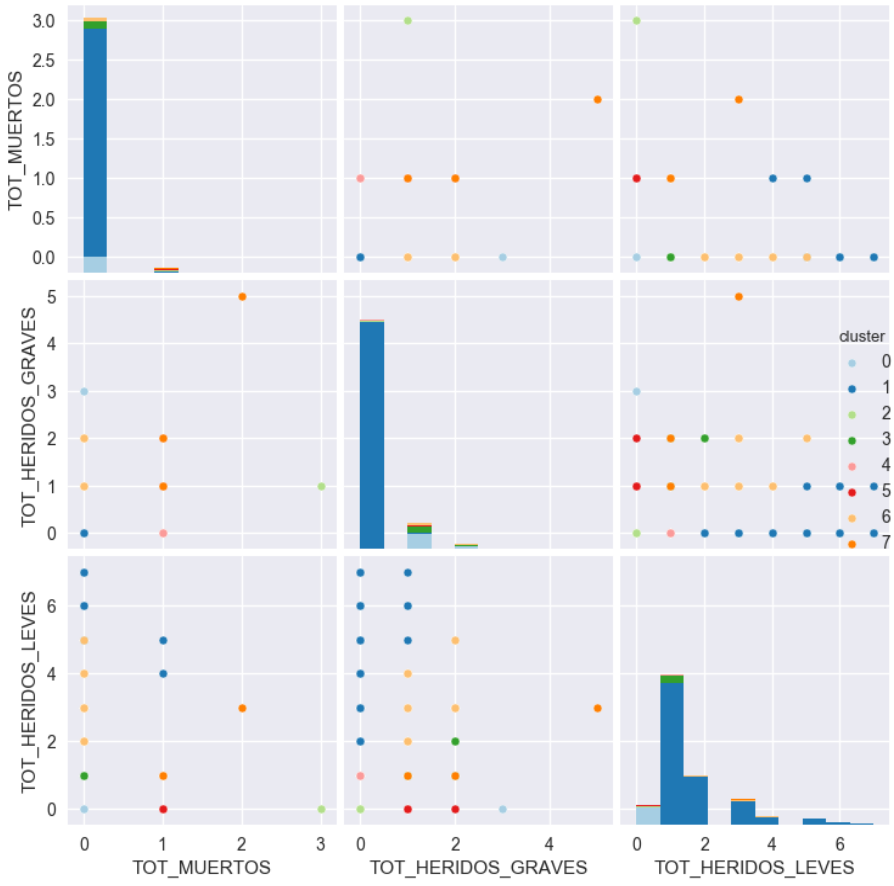


Figure 4: Visualización 3D en el caso 1 con el algoritmo K-Means

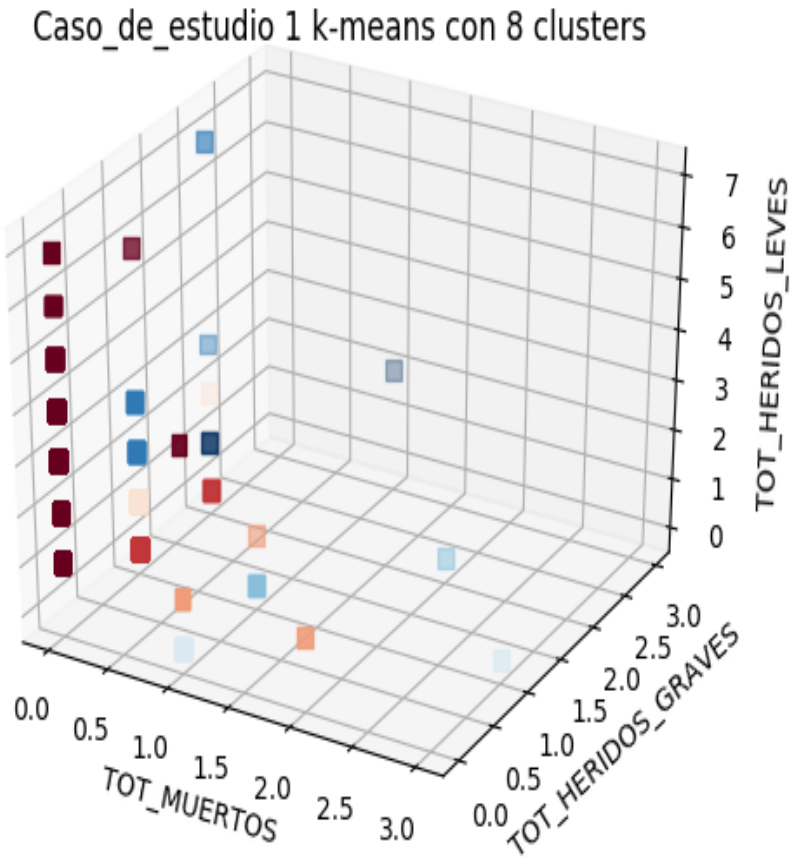
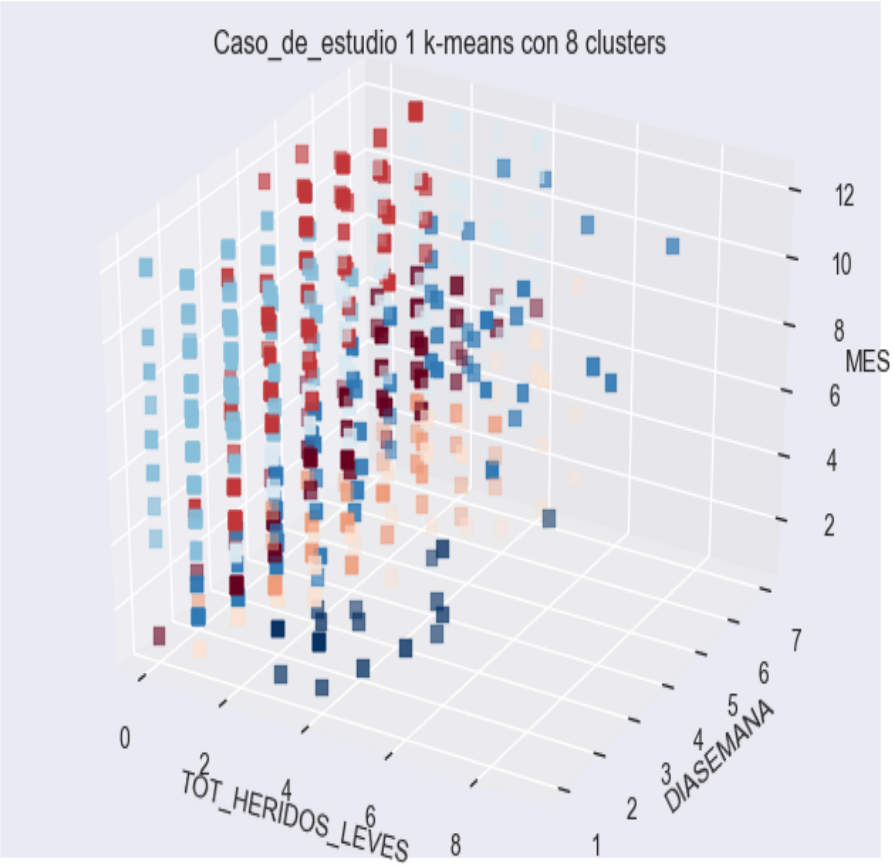


Figure 5: Visualización 3D en el caso 1 con el algoritmo K-Means con otras variables



3 Caso 2: accidentes de vuelco en la calzada con el suelo mojado

En este caso se decidió evaluar aquellos accidentes de vuelco en la calzada en los que el suelo se encontrase mojado.

Table 3: Resultados en el caso 2 para los distintos algoritmos

Nombre	K	Calinski-Harabaz Index	Silhouette Coefficient
k-means	4	1.000	0.97959
ward	2	0.0	0.0
birch	1	0.0	0.0
dbsscan	3	7458.376	0.94433
spectral	8	24.991	NaN

El algoritmo dbsscan es el que parece que ofrece mejores resultados en ambas métricas. Se puede observar como se obtuvo un valor de NaN con el algoritmo spectral para la métrica Silhouette, lo cuál no se ha conseguido descubrir a qué es debido.

3.1 Análisis de parámetros

En esta sección se va a hacer un pequeño estudio modificando los parámetros de los algoritmos para observar si se consiguen mejoras.

Table 4: Resultados de las modificaciones en el caso 2 para los distintos algoritmos

Nombre	**Modificación**	**K**	**Calinski-Harabaz Index**	**Silhouette Coefficient**
k-means	k=3	3	16390.905	0.96855
ward	k=3	3	0.0	0.0
birch	k=3	3	0.0	0.0
dbsscan	eps=0.01	3	7458.376	0.97959
spectral	k=3	3	74.622	-0.81633

En la Figura 1 podemos observar una comparación entre los valores del **Calinski-Harabaz Index** entre cada algoritmo con sus valores por defecto y la modificación realizada que se puede observar en la Tabla 4.El algoritmo que obtiene una mayor mejora es el algoritmo k-means. Sin embargo en el resto de algoritmos o se mantiene o mejora muy levemente.

A continuación, se ha deseado comparar los valores en la métrica **Silhouette Coefficient** en la Figura 2. En este caso, el único algoritmo que mejora es el dbscan, mientras que el algoritmo spectral y el k-means empeoran con respecto a la ejecución con los valores por defecto.

Figure 6: Comparación de los valores del Calinski-Harabaz Index en el caso2

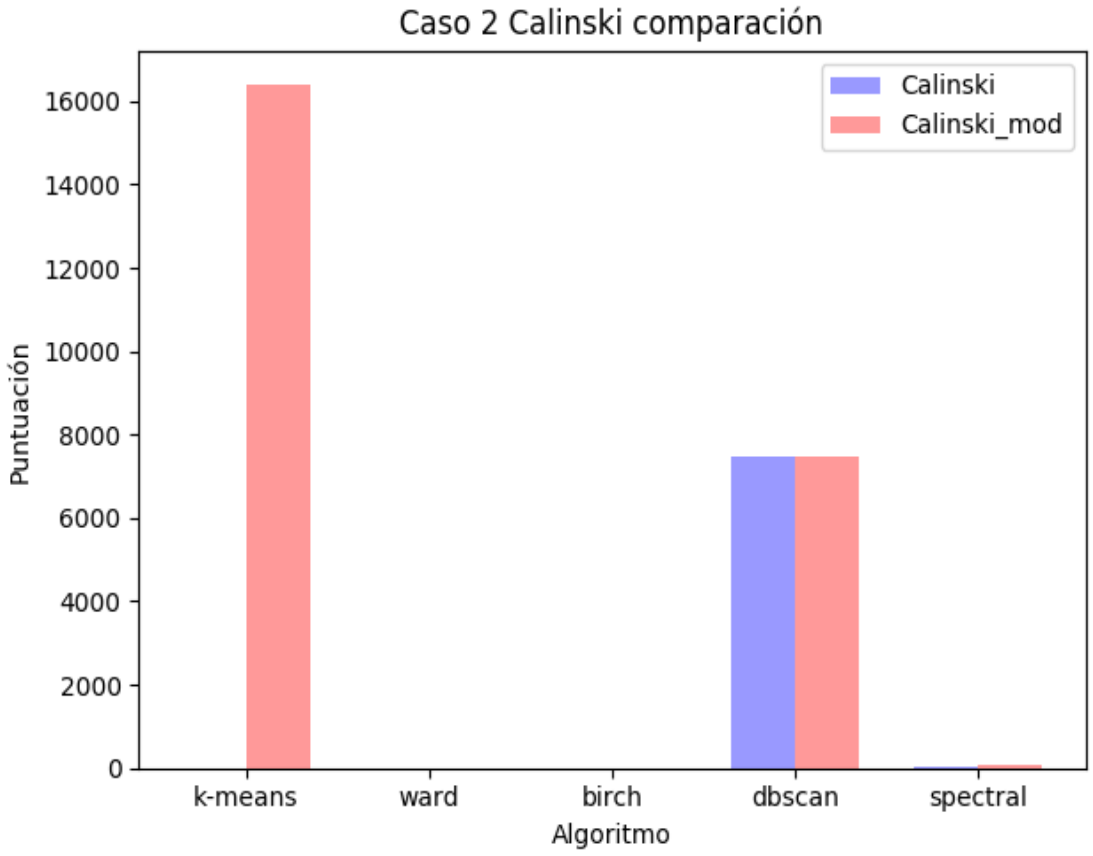
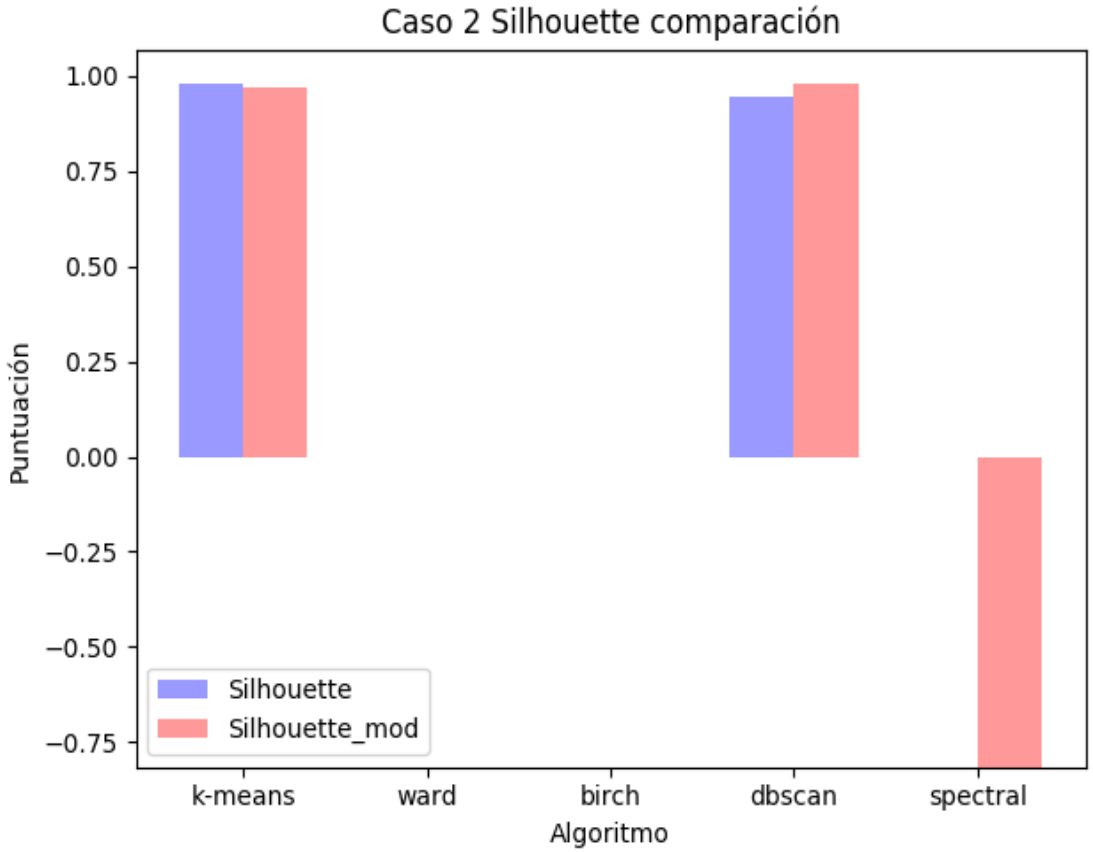


Figure 7: Comparación de los valores del Silhouette Coefficient en el caso 2



3.2 Interpretación de la segmentación

En este caso, en la Figura 8 se puede observar el scatter matrix que produce el algoritmo k-means para el caso 2. Nos encontramos con que el número de clusters que se genera es 4, el cuál es un número bastante reducido. Se puede observar como existe una predominancia del cluster 0 sobre el resto de clusters seguido del cluster 1. El cluster 2 y 3 parece que son cluster con una menor concentración de muestras.

En la Figura 9 se puede observar como la distancia entre los puntos es muy grande. Al igual que anteriormente, el número de muestras en el subset es superior, pero muchos puntos tiene el mismo valor y por ello no se muestran todos. La distancia espacial es grande entre puntos y esto es lo que provoca que se formen 4 clusters distintos. También se puede observar como en cluster predominante que era el cluster 0 la Figura 8 corresponda al cluster de color rojo en la FIgura 9, ya que es aquel con más puntos y más concentrado.

Al igual que en el caso 1, también se ha decidido analizar el problema usando otras variables. Para ello se han vuelto a coger las variables de total de heridos leves, el mes y el día de la semana. Como se puede observar la concentración de los puntos es menos clara que con las variables anteriormente estudiados y las muestras se distribuyen más uniformemente en el espacio. Por ello en este caso tampoco se pueden observar que exista una correlación más clara entre ciertos días de la semana en los que los accidentes provoquen más heridos leves.

Figure 8: Scatter Matrix en el caso 2 con el algoritmo K-Means

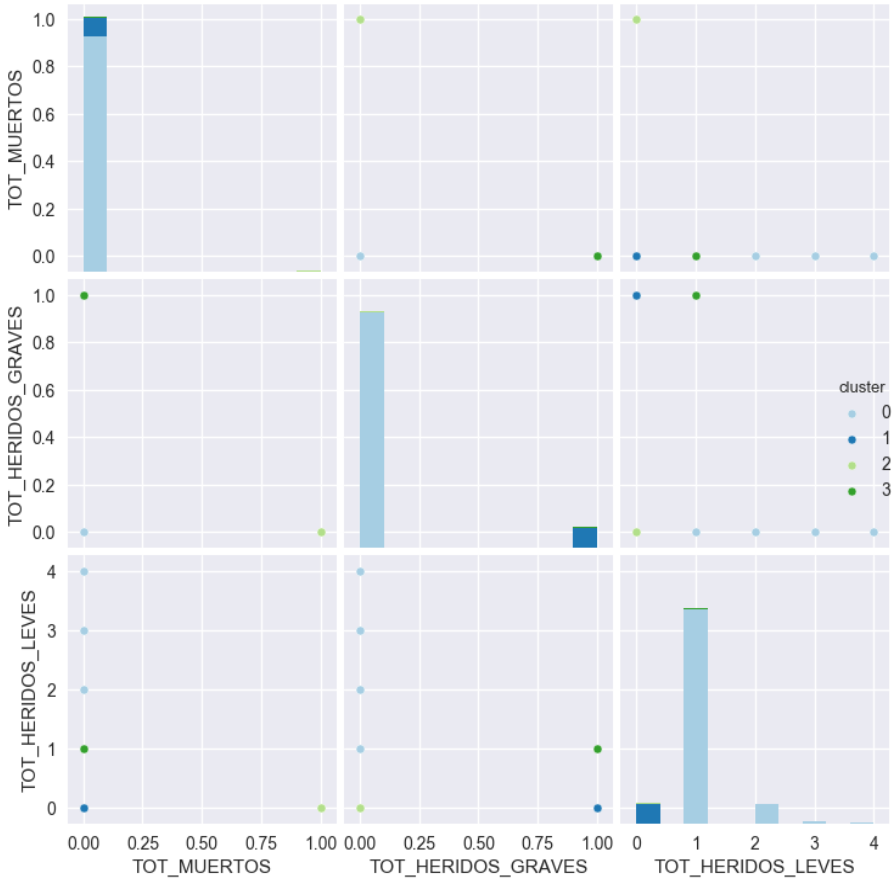


Figure 9: Visualización 3D en el caso 2 con el algoritmo K-Means

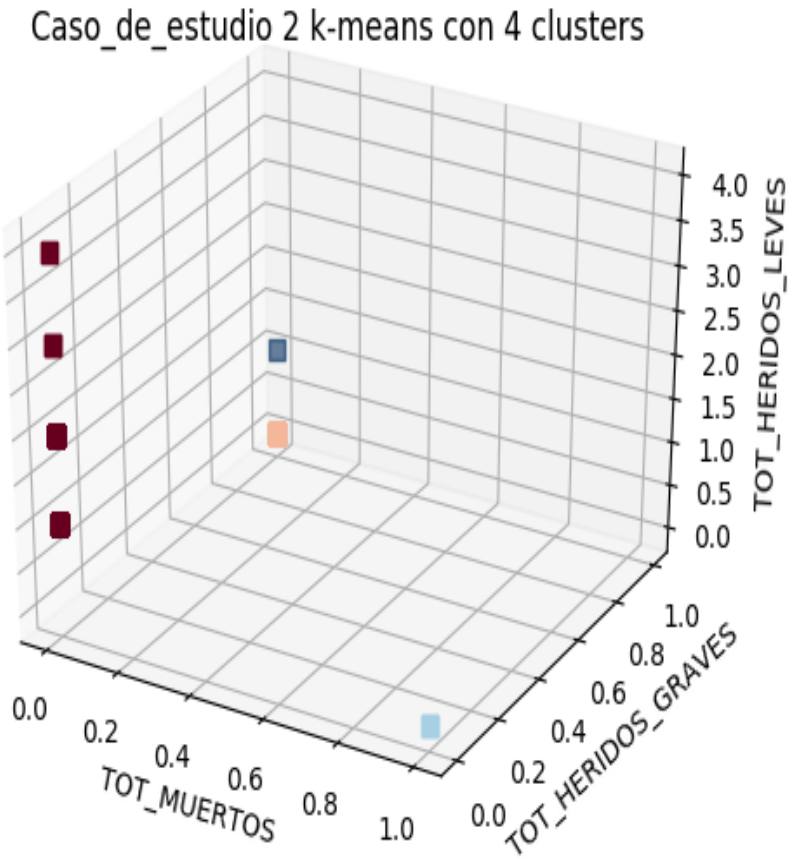
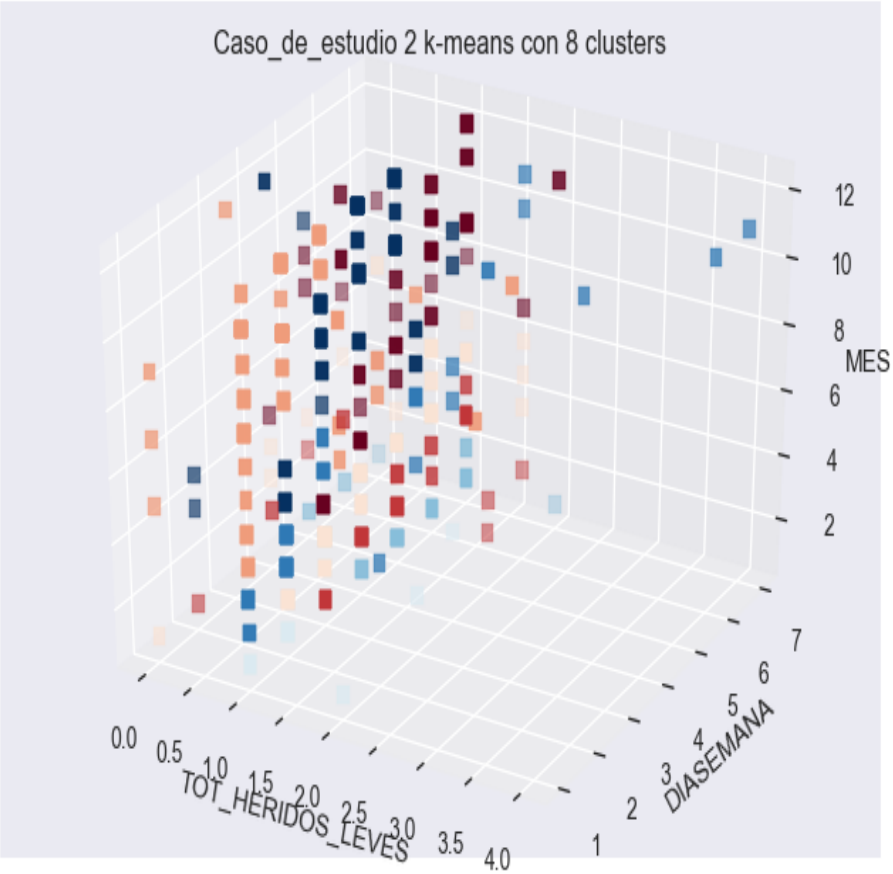


Figure 10: Visualización 3D en el caso 2 con el algoritmo K-Means con otras variables



4 Caso 3: atropello a peatón aislado o en grupo en Granada

En este caso, se han decidido estudiar aquellos datos cuyo accidente ha sido atropello a peatón aislado o en grupo en la provincia de Granada.

Table 5: Resultados en el caso 3 para los distintos algoritmos

Nombre	K	Calinski-Harabaz Index	Silhouette Coefficient
k-means	6	1.000	0.94444
ward	2	0.0	0.0
birch	3	2410.828	0.94444
dbsscan	4	4310.887	0.94444
spectral	8	26.101	NaN

En este caso, en la Tabla 5 se puede observar como los algoritmos no se diferencian en su puntuación en la métrica Silhouette pero sin embargo, con respecto al Calinski-Harabaz Index el algoritmo que mejor resultados nos proporciona es el algoritmo dbsscan.

4.1 Análisis de parámetros

En esta sección se va a hacer un pequeño estudio modificando los parámetros de los algoritmos para observar si se consiguen mejoras.

Table 6: Resultados de las modificaciones en el caso 3 para los distintos algoritmos

Nombre	Modificación	K	Calinski-Harabaz Index	Silhouette Coefficient
k-means	k=3	3	2496.436	0.94444
ward	k=3	3	0.0	0.0
birch	k=3	3	2410.828	0.94444
dbsscan	eps=0.01	3	1050.108	0.94444
spectral	k=3	3	37.645	-0.70139

En la Figura 11 podemos observar una comparación entre los valores del **Calinski-Harabaz Index** entre cada algoritmo con sus valores por defecto y la modificación realizada que se puede observar en la Tabla 6.El algoritmo que obtiene una mayor mejora es el algoritmo k-means, al igual que ocurría en la Figura 6 del caso 2. En el resto de algoritmos no mejora, excepto en el birch que se mantiene.

A continuación, se ha deseado comparar los valores en la métrica **Silhouette Coefficient** en la Figura 2. Aquí se puede observar como las modificaciones no han afectado a los algoritmos, ya que su puntuación ha sido la misma, excepto en el caso del algoritmo spectral, que disminuye bastante.

Figure 11: Comparación de los valores del Calinski-Harabaz Index en el caso3

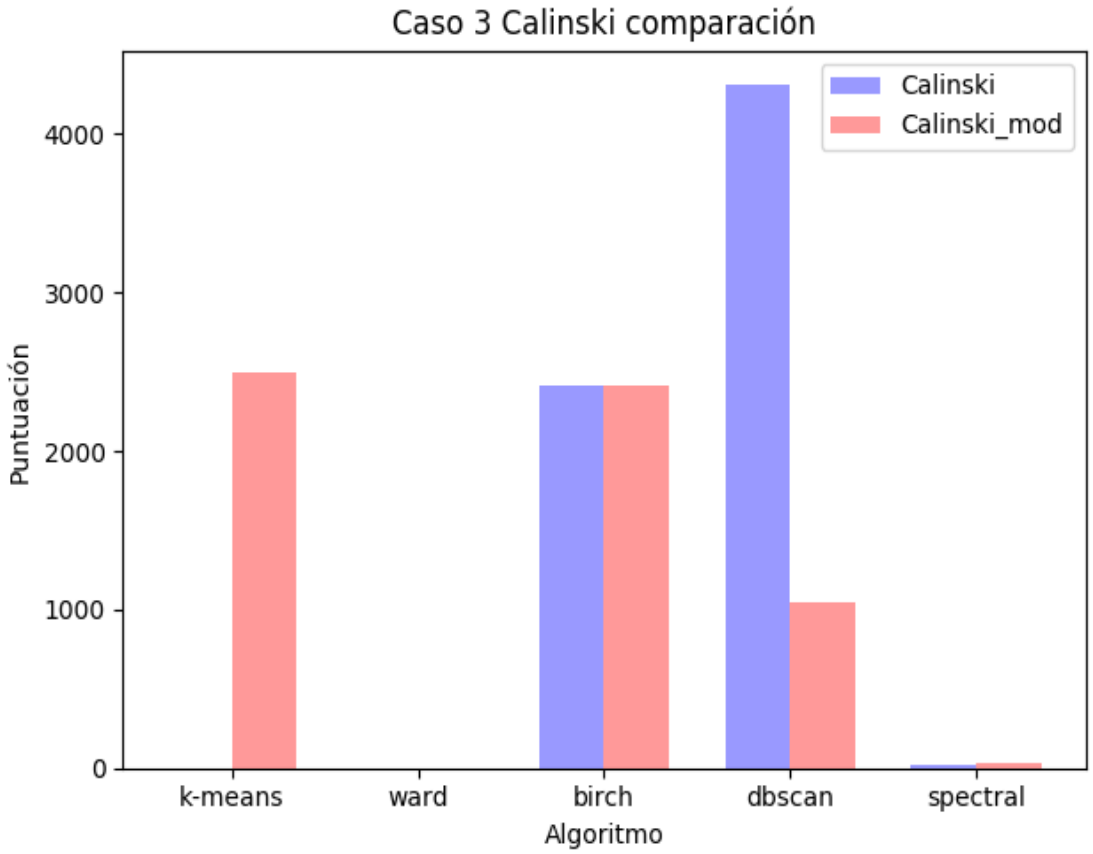
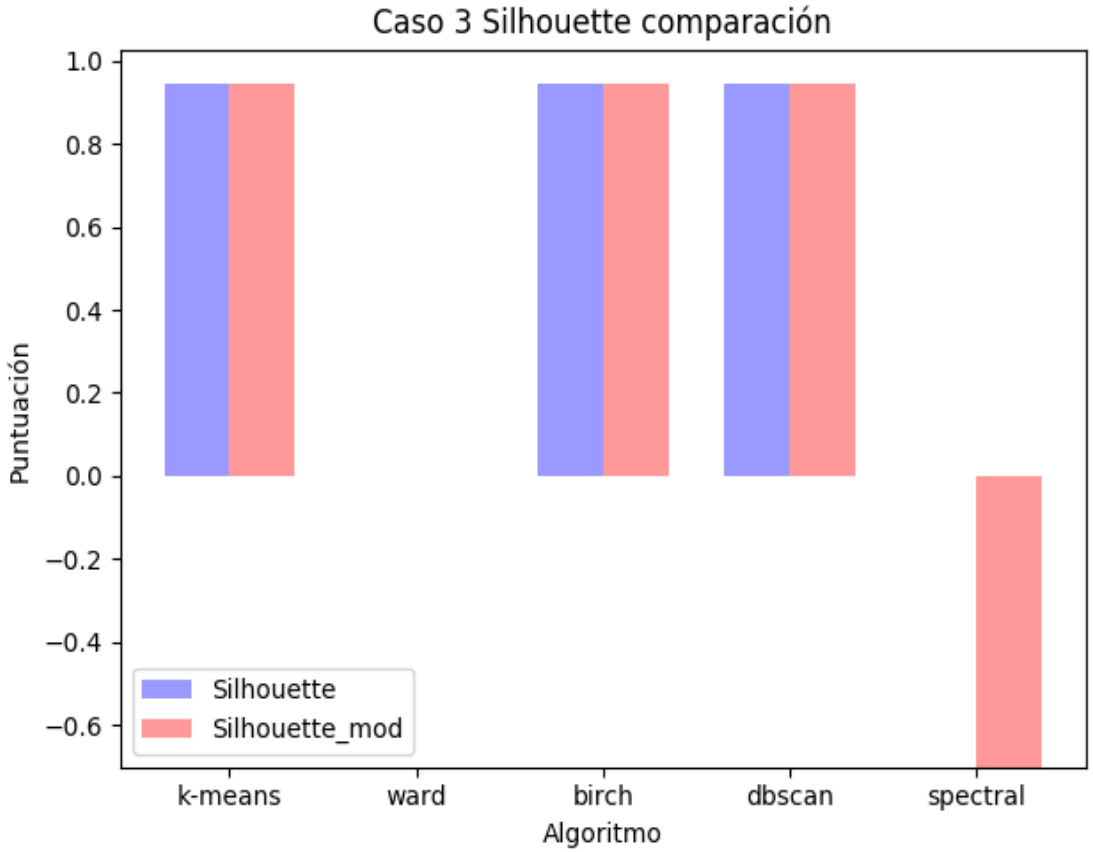


Figure 12: Comparación de los valores del Silhouette Coefficient en el caso 3



4.2 Interpretación de la segmentación

En este caso, al tratarse de un subset pequeño ya que solo contiene 186 datos, un dendograma junto con el heatmap puede resultar muy interesante a la hora de visualizar como se agrupan los datos.

En la Figura 13 se puede observar el dendograma junto con el heatmap. Los valores arriba no se observan normalizados ya que para la visualización se utilizó el subset original, mientras que para la ejecución de los algoritmos se utilizó el subset con sus valores normalizados entre 0 y 1. En el dendograma se puede observar como se van formando primero cluster pequeños con algunas muestras qu se encuentran entre distintos valores del total de heridos leves y luego se van agrupando en clusters más grandes. Al inicio, se agrupan entre los que tienen menor número de heridos leves desde 0.0 a 0.8 para luego juntarse con el cluster que se forma con los valores entre 0.8 a 2.4. Luego, esta cluster finalmente se junta con el cluster con mayor valor en la variable de total de heridos leves para formar un único cluster. Este tipo de agrupación lo que nos indica es que los datos se encuentran diferenciados más o menos bien por unas fronteras que los separan dependiendo del número total de heridos leves y que se pueden ir separando por su rango de heridos entre 3 cluster más diferenciados que sería donde yo me quedaría, entre pocos heridos leves, un número medio de heridos leves, y muchos heridos leves.

En la Figura 14 se puede observar el dendograma y el heatmap usando variables distintas para el análisis como se ha realizado anteriormente para el caso 1 y el caso 2. Se puede observar como en este caso, los puntos se encuentran más distribuidos espacialmente ya que se forman muchísimos cluster con pocos puntos, al contrario de lo que pasaba en la Figura 13. Aun así, se puede observar como la variable total de heridos leves se encuentra distribuida de forma parecida en todos

los meses y días de la semana. Los clusters se van formando poco a poco para luego hacer una separación en 3 cluster grandes por lo que se puede observar. Según se observan los valores se podría decir que un cluster es de los primeros meses del año, otro de los meses intermedios y el tercero de los últimos meses del año, por lo que la gama cromática nos da a entender. Pero lo interesante y destacable es que no parece que haya una correlación entre el mes o el día de la semana con que exista un mayor número de heridos leves en los accidentes que pertenecen a este subset.

Una hipótesis que existía en mi mente a la hora de realizar este análisis con la variable mes y día de la semana es que se producirían más atropellos en los últimos meses del año y el primero por ser fiesta como por ejemplo la Navidad, pero parece ser que este análisis no valida esta hipótesis.

Figure 13: Dendogram y Heatmap en el caso 3

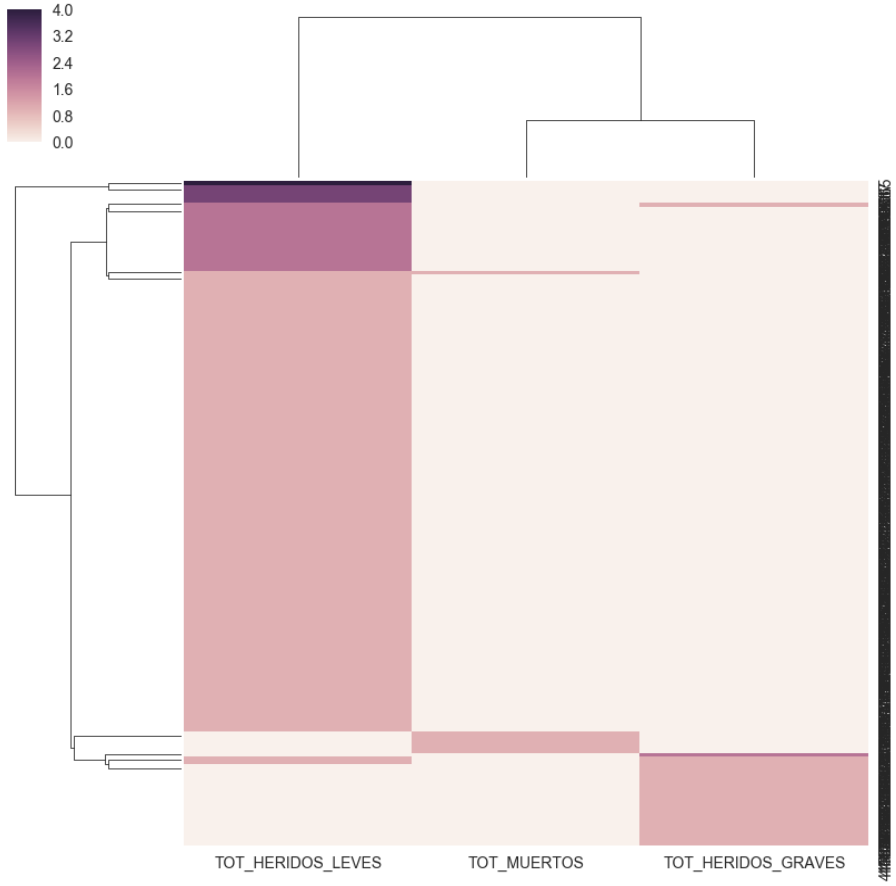
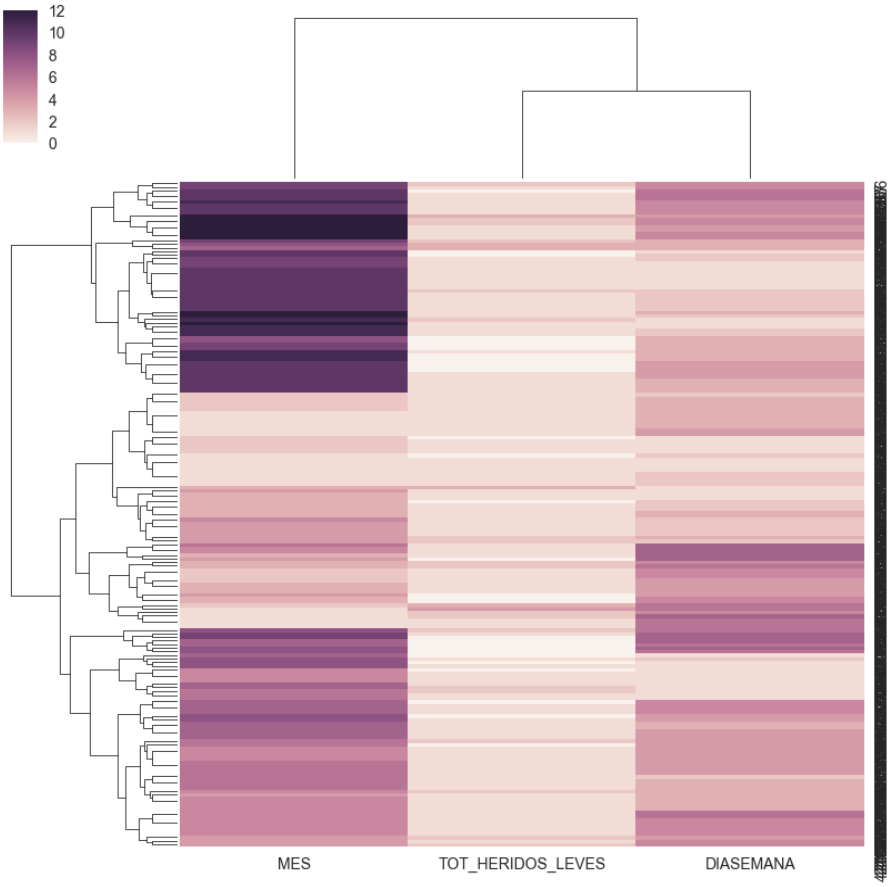


Figure 14: Dendogram y Heatmap en el caso 3 con otras variables



5 Contenido Adicional

6 Bibliografía
References