Inteligencia de Negocio

Miguel Morales Castillo

miguemc@correo.ugr.es
Grupo Lunes

Noviembre 2018

Índice

1. Introducción

En esta práctica aplicaremos el uso de técnicas de aprendizaje no supervisado para análisis empresarial. Trabajaremos con un conjunto de datos del censo de Granada en 2011. Sobre este conjunto de datos aplicaremos 5 algoritmos de clustering y analizaremos los resultados obtenidos.

El conjunto de datos cuenta con 83.499 instancias. Como esta cifra es bastante alta para trabajar en ordenadores personales ya que los algoritmos de clustering necesitan mucha memoria RAM, hemos escogido 3 casos de uso que filtrarán el conjunto de datos para reducir este número.

En cuanto a los algoritmos elegidos, nos hemos basado para la elección en aquellos que son mas eficientes en términos de gestión de memoria y Tiempo (s) de uso para los casos de uso que queríamos estudiar. Los algoritmos escogidos son K-Means, DBSCAN, Agglomerative Clustering, Birch y MeanShift. Una breve introducción a ellos sería la siguiente:

K-means almacena k centros que son usados para definir los cluster. Un punto pertenecerá al cluster cuyo centro este mas cercano al punto en cuestión. el centro va cambiando en mientras se añadan puntos al cluster, y se finaliza cuando ya no hay mas cambios en los centros.

DBSCAN es un algoritmo de clustering basado en densidad porque encuentra un número de clusters comenzando por una estimación de la distribución de densidad de los nodos correspondientes.

Agglomerative Clustering es un algoritmo en el que cada punto es un cluster, entonces, usando la distancia definida, los cluster mas cercanos se unen en uno solo, y este proceso se repite hasta tener un solo cluster.

BIRCH es un algoritmo usado para clustering jerárquico sobre conjuntos grandes de datos. Una ventaja de BIRCH es su capacidad para agrupar de forma incremental y dinámica los puntos entrantes en un intento de producir la agrupación en clusters de la mejor calidad posible. En la mayoría de los casos, BIRCH solo requiere un solo análisis de la base de datos.

MeanShift es otro algoritmo basado en densidad, asigna los puntos de datos a los clusters de manera iterativa al desplazar los puntos hacia el modo. El modo puede entenderse como la densidad más alta de puntos de datos.

2. Casos de Estudio

2.1. Personas entre 30 y 55 años con mas de 3 personas en su familia

En este caso de estudio queremos estudiar las familias numerosas residentes en Granada, centrándonos en las personas entre 30 y 55 años, para observar de que estudios disponen y si son familias de aquí o vienen de fuera.

Veamos los resultados que obtenemos con este caso :

2.1.1. K-means

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
K-Means	0.11	6610.102	0.34163

nº cluster	Tamaño	Porcentaje sobre el total
0	6074	40.87
1	4819	32.43
2	2056	13.84
3	1911	12.86

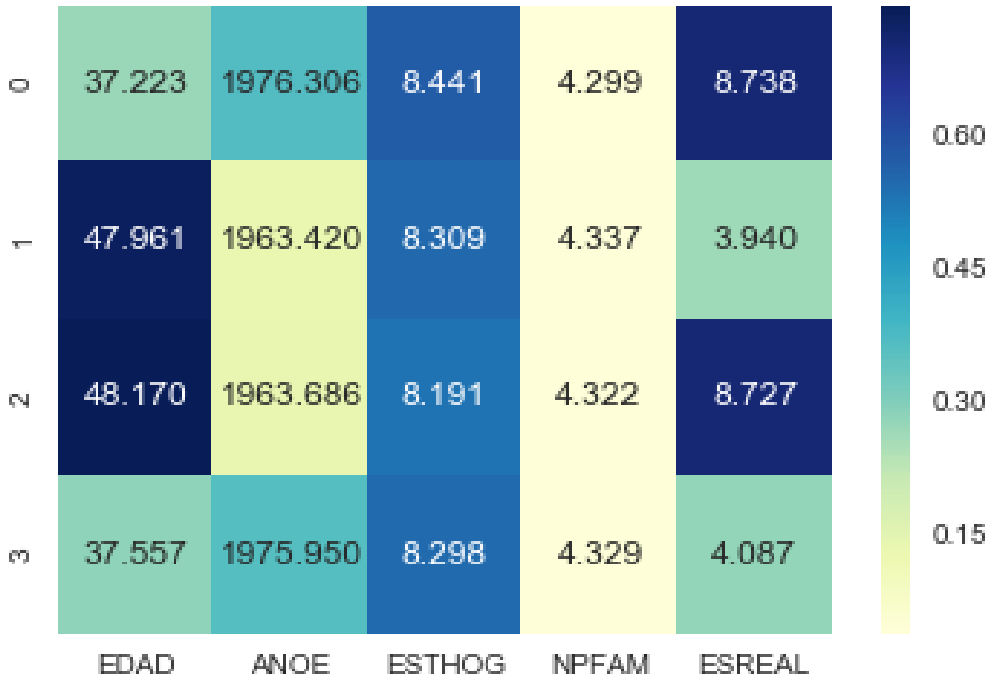

Figura 1: Heat Map de *K-Means*

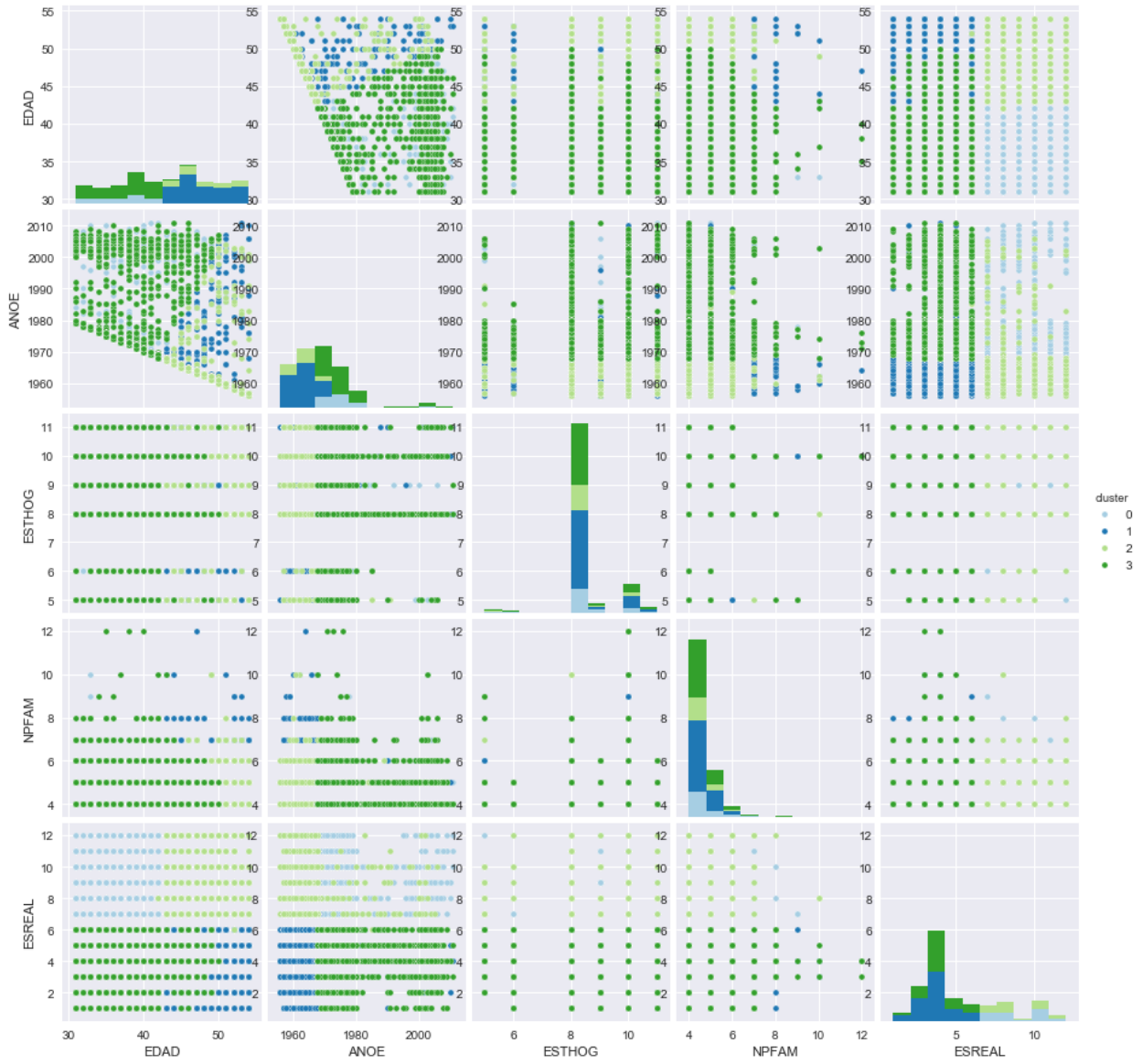

Figura 2: Scatter Matrix de *K-Means*

2.1.2. DBSCAN

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
DBSCAN	1.80	562.221	-0.02535

nº cluster	Tamaño	Porcentaje sobre el total
-1	2172	14.62
0	1137	7.65
1	11331	76.25
2	111	0.75
3	109	0.73

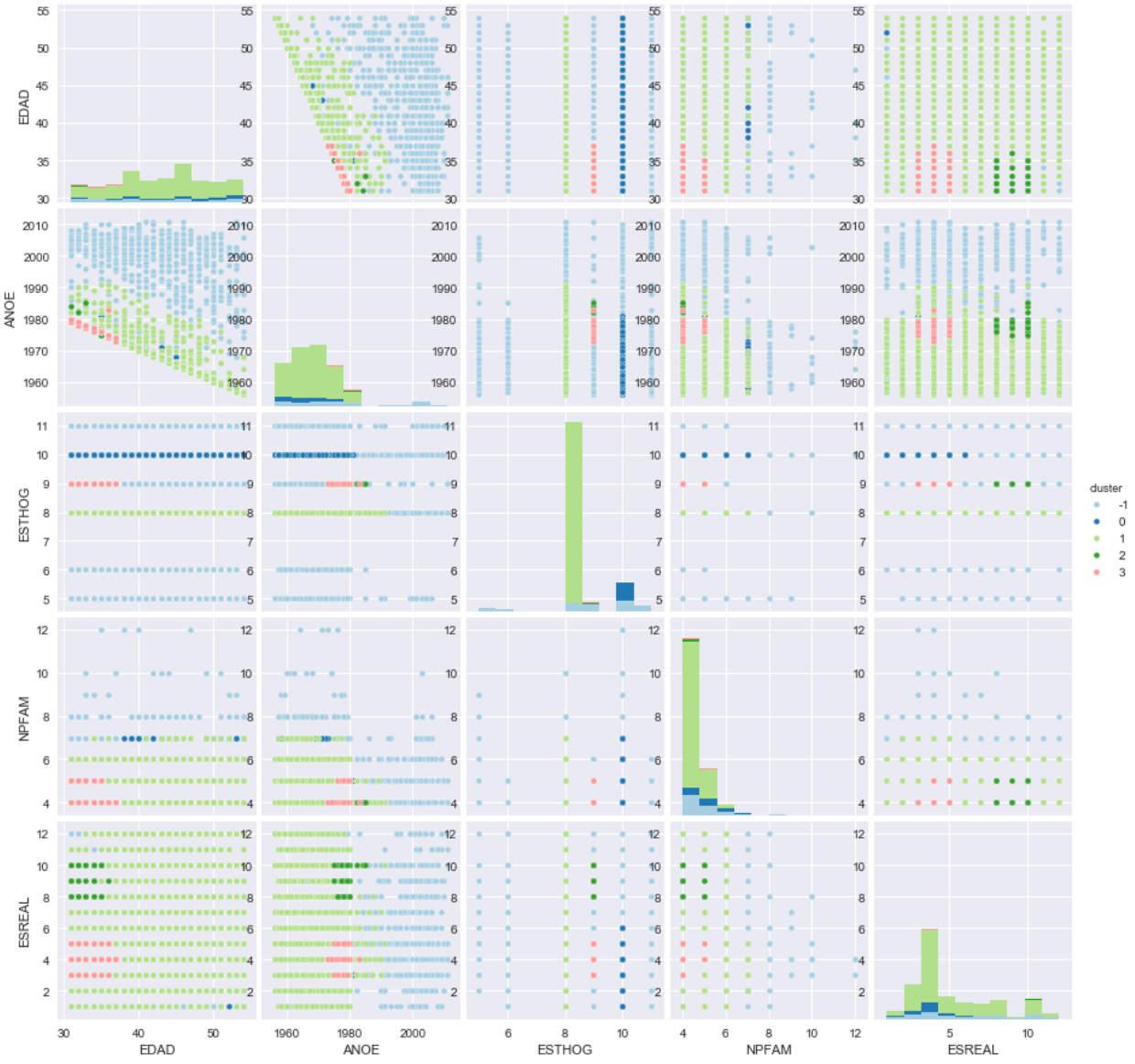

Figura 3: Scatter Matrix de *DBSCAN*

2.1.3. Agglomerative Clustering

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
Agglomerative	11.14	5161.279	0.30503

nº cluster	Tamaño	Porcentaje sobre el total
0	1759	11.84
1	5349	35.82
2	2429	16.35
3	5323	35.82

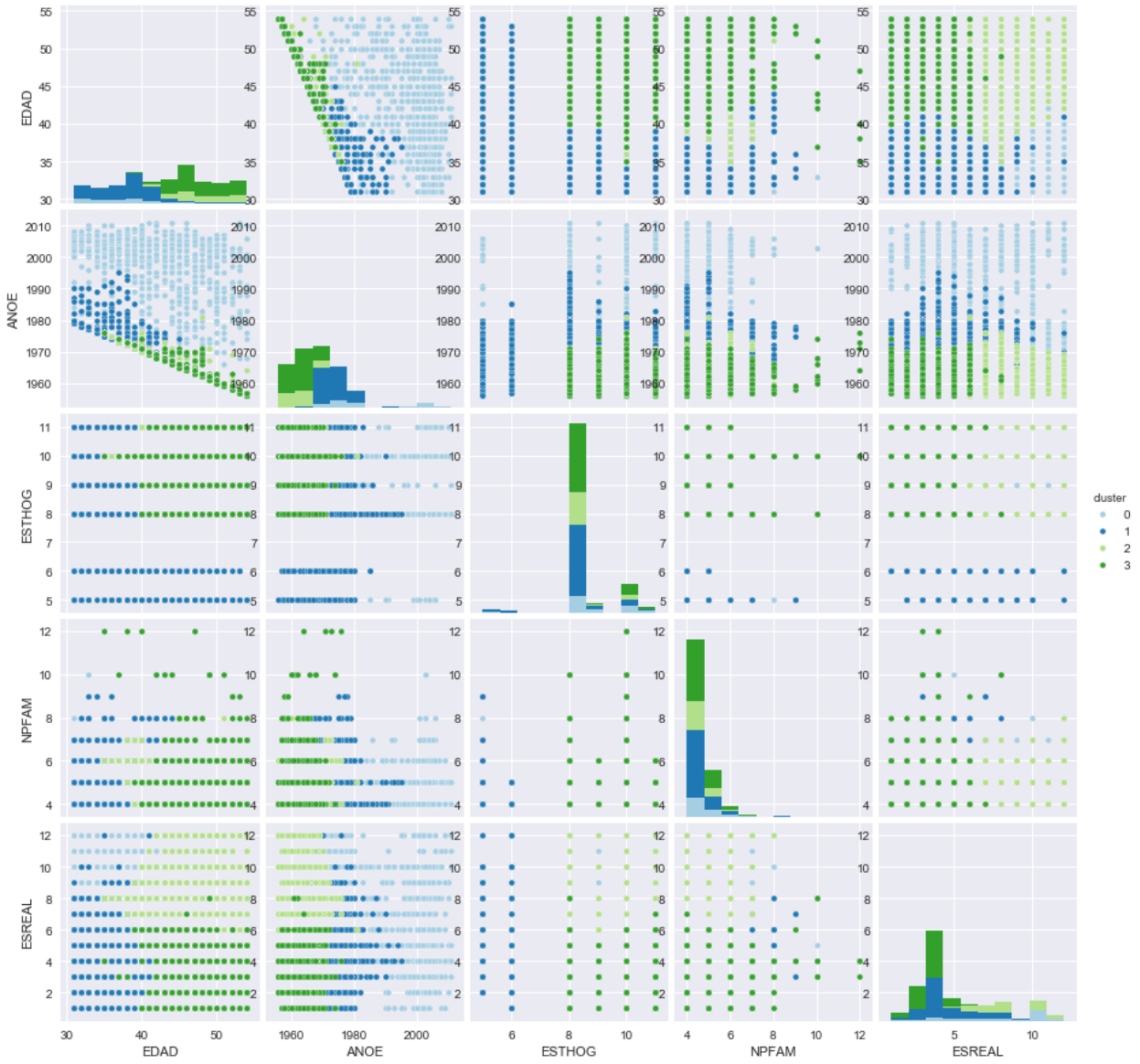

Figura 4: Scatter Matrix de *Agglomerative Clustering*

2.1.4. Birch

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
BIRCH	0.41	3736.522	0.34273

nº cluster	Tamaño	Porcentaje sobre el total
0	722	4.86
1	247	1.66
2	8858	59.61
3	5033	33.87

Figura 5: Scatter Matrix de *Birch*

2.1.5. MeanShift

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
MeanShift	191.28	906.989	0.24562

nº cluster	Tamaño	Porcentaje sobre el total
0	12335	83.01
1	2273	15.30
2	205	1.38
3	30	0.20
4	16	0.11
5	1	0.01

Figura 6: Scatter Matrix de *MeanShift*

A la Luz de los resultados arrojados por los distintos algoritmos podemos sacar varias conclusiones de este caso de estudio.

Por un lado tenemos un grupo mayoritario de personas entre 30 y 50 años que la estructura de su

hogar esta formada por pareja hijos y otras personas y por hijos mayores de 25 años. Estas personas tienes como máximo bachillerato o equivalente.

Por otra parte en un grupo mucho menor se encuentran personas entre 30 y 40 años, que disponen de estudios entre bachillerato y licenciatura universitaria. Estas personas se encuentran en una estructura familiar compuesta por hijos mayores de 25 años.

Esto nos puede llevar a concluir que no es tan determinante la estructura familiar actual de la persona como la edad. Esto tiene sentido ya que con los años la facilidad para estudiar es cada vez mayor.

Si nos fijamos en los resultados aportados por DBSCAN o Agglomerative Clustering, una gran cantidad de personas provienen de fuera de España, ya que su edad no concuerda con el año de llegada a España. Este grupo si nos fijamos en el tipo de estudios realizados, vemos que hay de todo tipo, por lo que no parece que haya una relación entre ser extranjero y el tipo de estudios.

2.2. Jóvenes menores de 25 años

En este caso, veremos que ocurre con los jóvenes menores de 25 años, que estructura familiar tienen y los estudios realizados para ver si tienen alguna relación.

Veamos los resultados que obtenemos con este caso :

2.2.1. K-means

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
K-Means	0.19	29807.228	0.37916

nº cluster	Tamaño	Porcentaje sobre el total
0	6552	28.96
1	3808	16.83
2	7020	31.03
3	5244	23.18

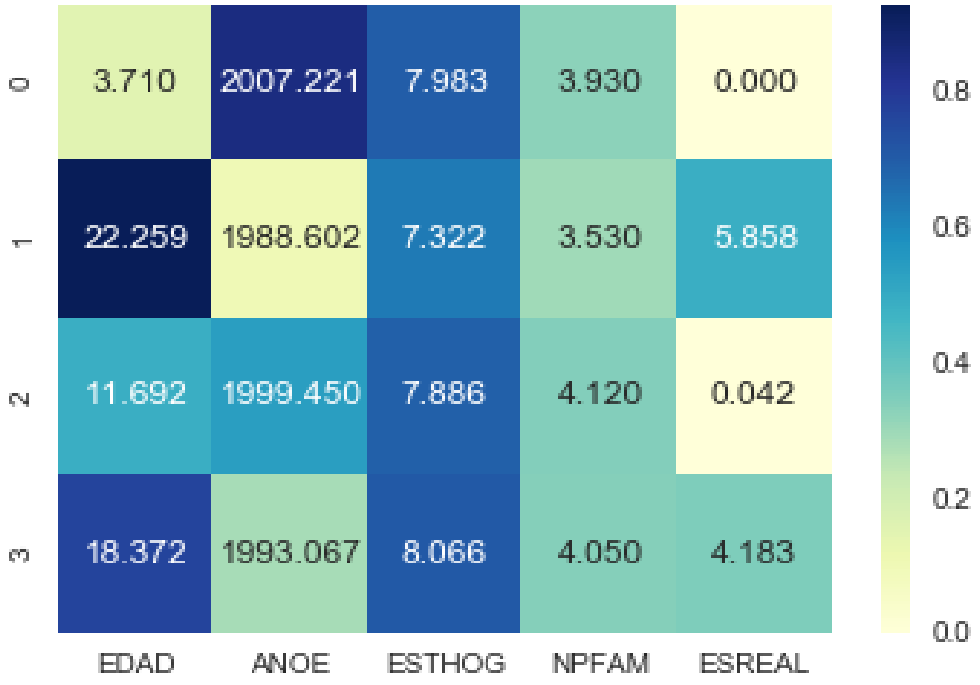

Figura 7: Scatter Matrix de *K-Means*

Figura 8: Scatter Matrix de *K-Means*

2.2.2. DBSCAN

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
DBSCAN	3.75	1081.888	-0.09447

nº cluster	Tamaño	Porcentaje sobre el total
-1	583	2.58
0	17934	79.27
1	1477	6.53
2	1154	5.10
3	1370	6.06
4	106	0.47

Figura 9: Scatter Matrix de *DBSCAN*

2.2.3. Agglomerative Clustering

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
Agglomerative Clustering	29.65	28476.394	0.41637

nº cluster	Tamaño	Porcentaje sobre el total
0	7743	34.22
1	7695	34.01
2	5793	25.61
3	1393	6.16

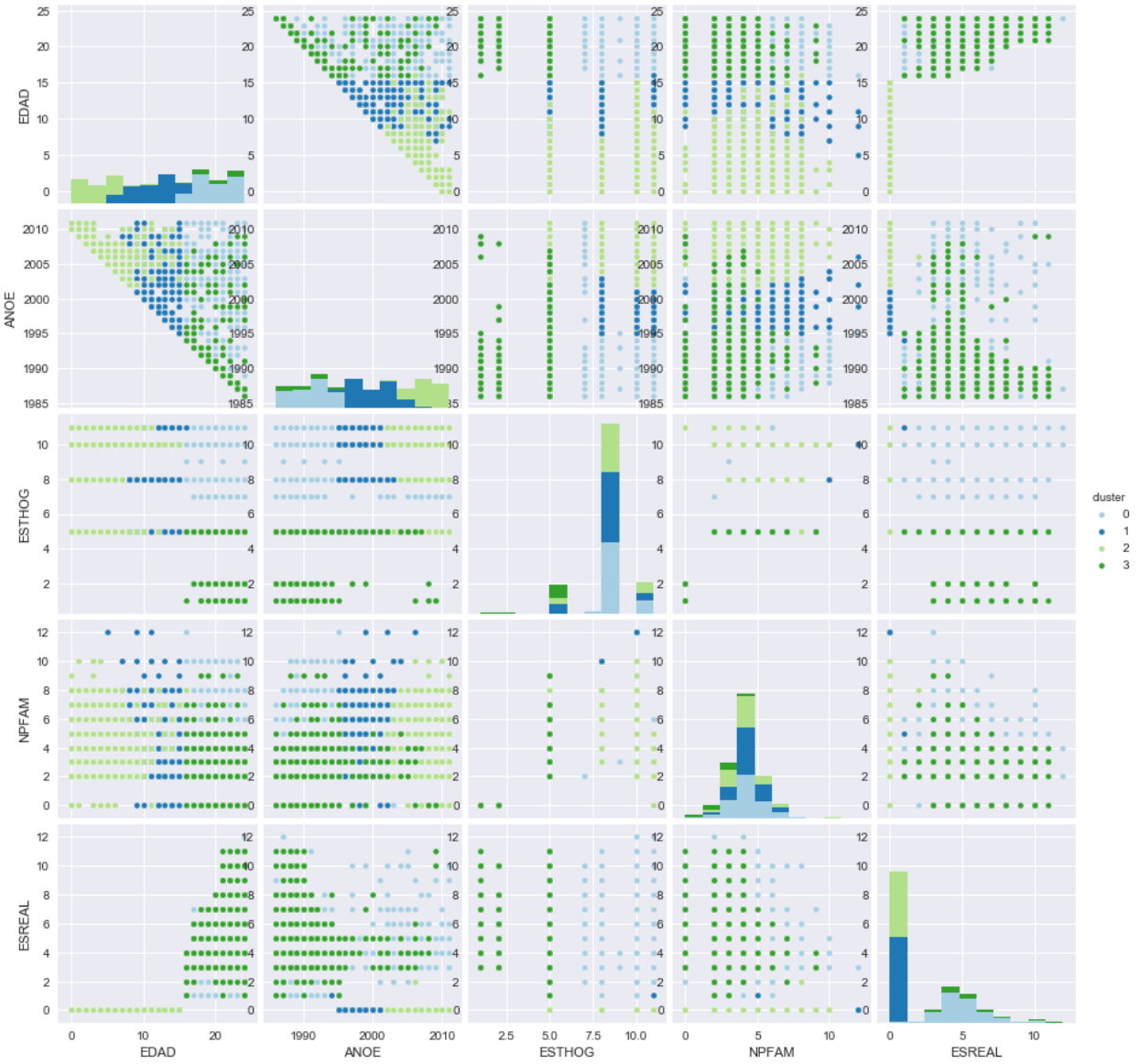

Figura 10: Scatter Matrix de *Agglomerative Clustering*

2.2.4. Birch

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
BIRCH	0.50	27357.944	0.37687

nº cluster	Tamaño	Porcentaje sobre el total
0	4508	39.52
1	3059	27.03
2	6116	19.93
3	8941	13.52

Figura 11: Scatter Matrix de *Birch*

2.2.5. MeanShift

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
MeanShift	223.26	12924.314	0.45890

nº cluster	Tamaño	Porcentaje sobre el total
0	10937	50.68
1	11465	48.34
2	214	0.95
3	8	0.04

Figura 12: Scatter Matrix de *MeanShift*

Podemos observar de los resultados obtenidos que hay tres grupos marcados. Primero, tenemos jóvenes de entre 0 y 10 años, después entre 10 y 20 años, y finalemente entre 20 y 25. Respecto a los estudios realizados, vemos que muy pocos jóvenes entre 20 y 25 años disponen de una carrera universitaria, la mayoría solo disponen de bachillerato o similar. Vemos que hay un número elevado dentro de esos pocos jóvenes que su estructura familiar es uniparental.
De esto podemos inferir que hay muy pocos jóvenes que realicen carrera universitaria, y que no parece tener relación alguna con la estructura familiar, ya que incluso de los que tienen carrera son mayoritariamente de familias uniparentales.

2.3. Personas menores de 30 años cuyos padres solo tienen estudios básicos

En este caso de estudio queremos ver si tiene alguna relación el desempeño escolar de los padres, con el camino seguido por los hijos, para ello, hemos considerado aquellos padres de menores de 30

que solo tienen estudios básicos.

Veamos los resultados que obtenemos con este caso :

2.3.1. K-means

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
K-Means	0.06	4616.607	0.34962

nº cluster	Tamaño	Porcentaje sobre el total
0	1819	27.80
1	1473	22.51
2	576	8.80
3	2675	40.88

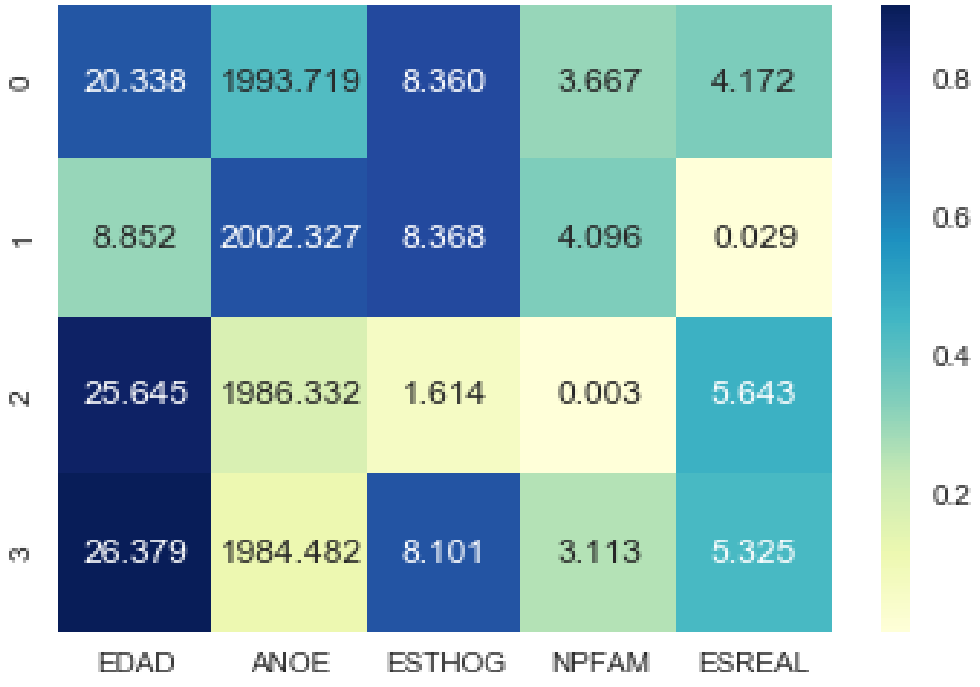

Figura 13: Scatter Matrix de *K-Means*

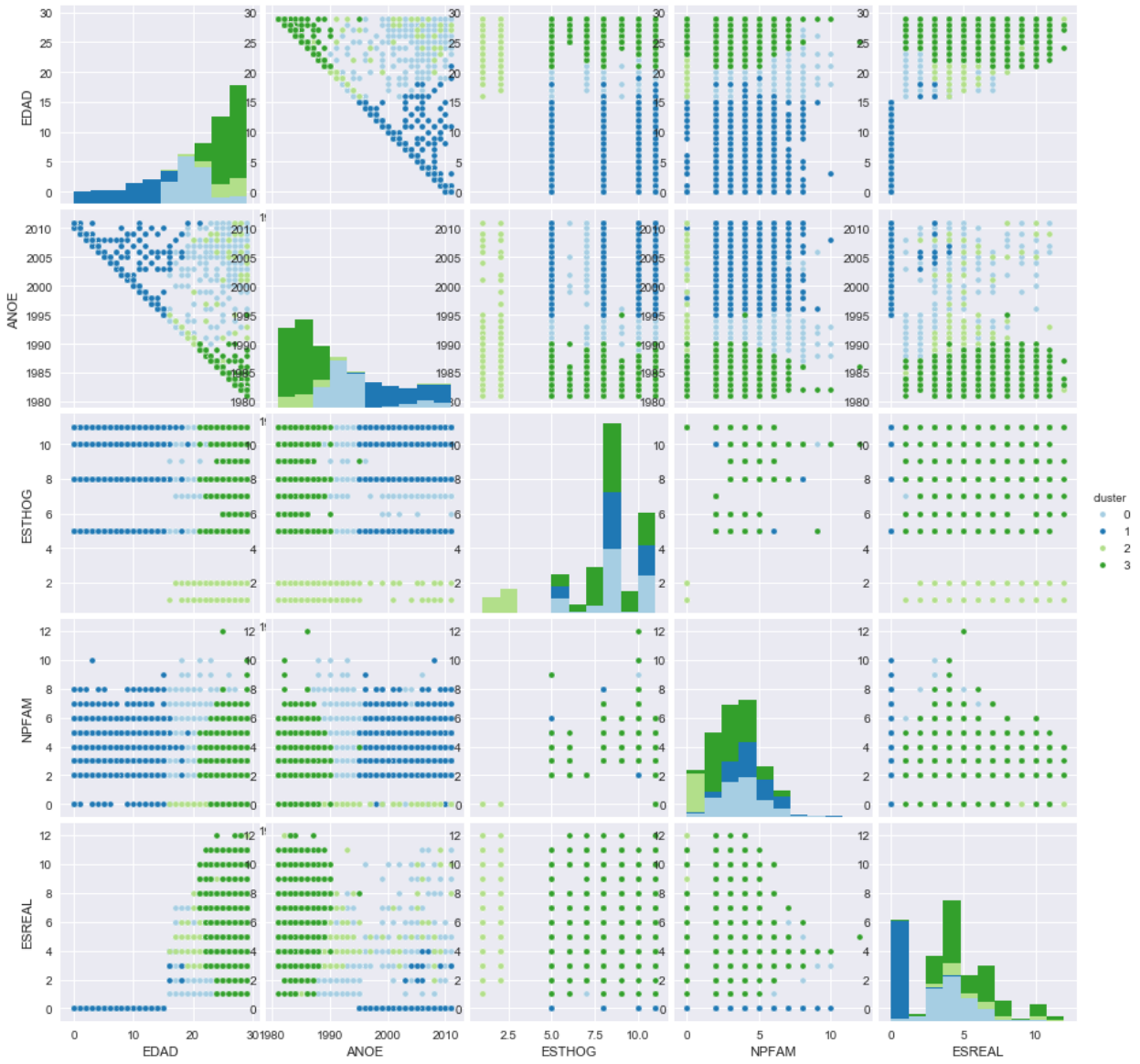

Figura 14: Scatter Matrix de *K-Means*

2.3.2. DBSCAN

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
DBSCAN	0.77	658.636	0.11997

nº cluster	Tamaño	Porcentaje sobre el total
-1	923	14.11
0	4723	72.18
1	344	5.26
2	458	7.00
3	95	1.45

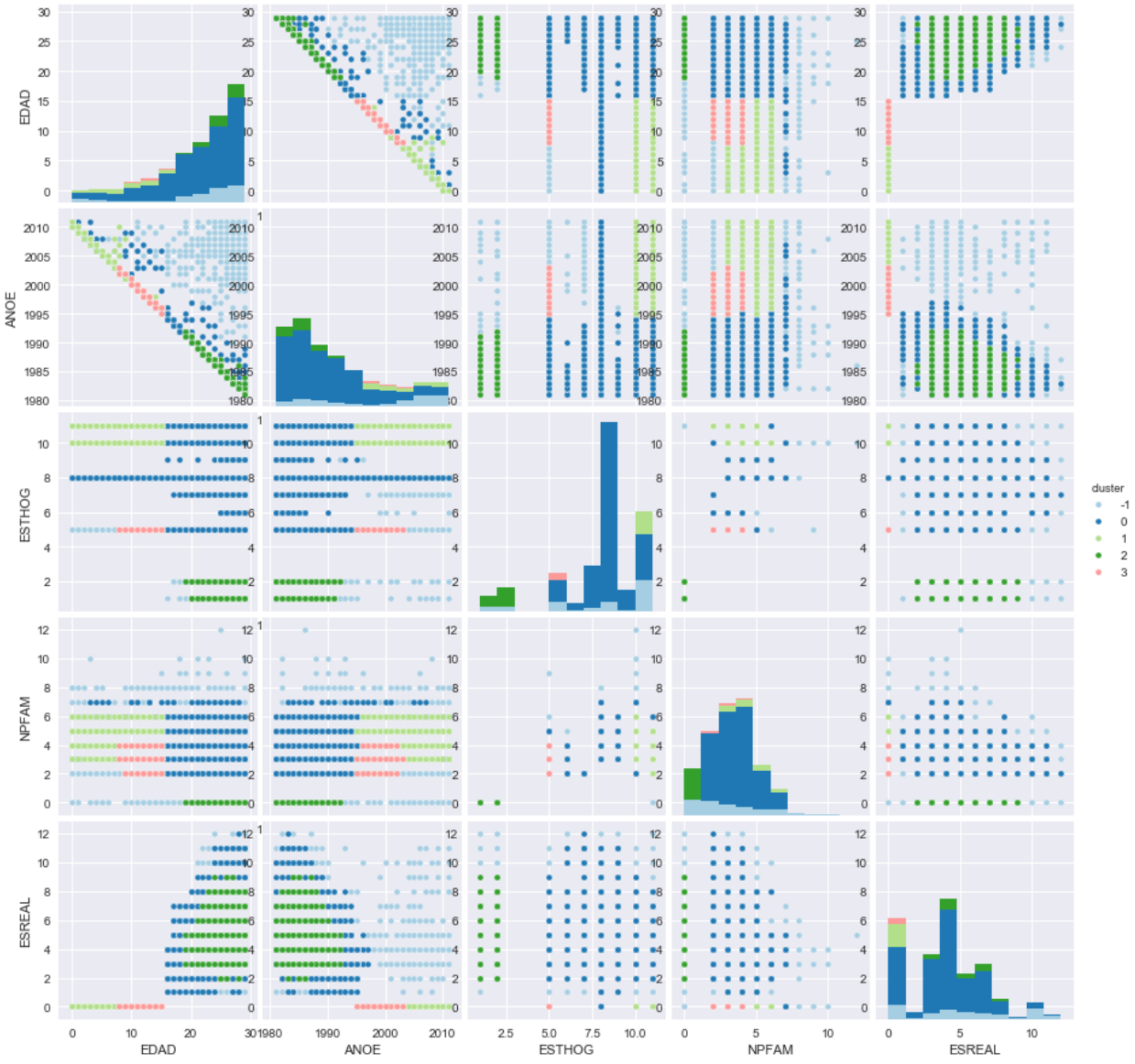

Figura 15: Scatter Matrix de *DBSCAN*

2.3.3. Agglomerative Clustering

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
Agglomerative Clustering	1.63	4250.701	0.31158

nº cluster	Tamaño	Porcentaje sobre el total
0	2613	39.94
1	1460	22.31
2	574	8.77
3	1896	28.98

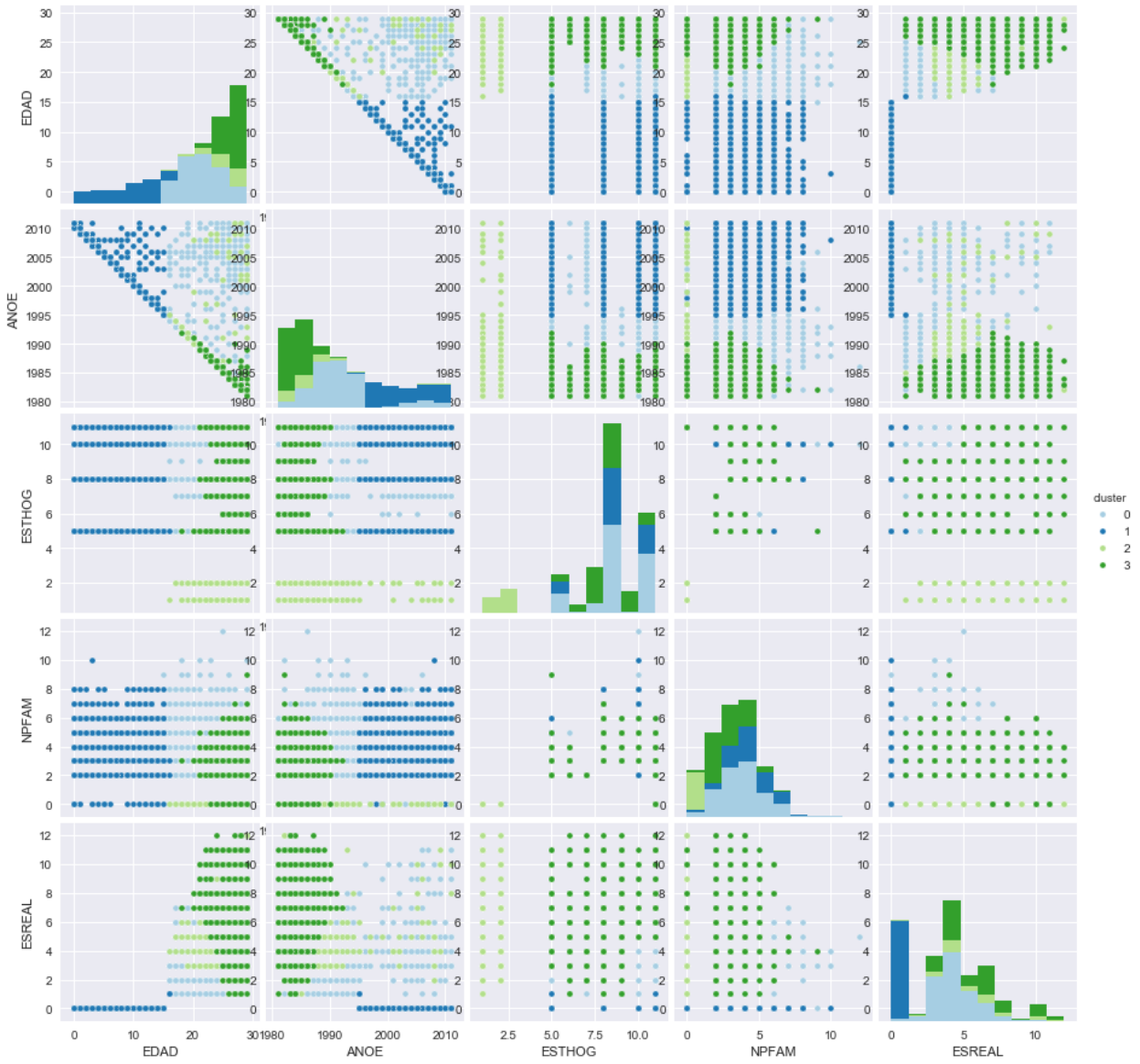

Figura 16: Scatter Matrix de *Agglomerative Clustering*

2.3.4. Birch

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
BIRCH	0.14	3855.577	0.42072

nº cluster	Tamaño	Porcentaje sobre el total
0	575	8.79
1	294	4.49
2	4261	65.12
3	1413	21.60

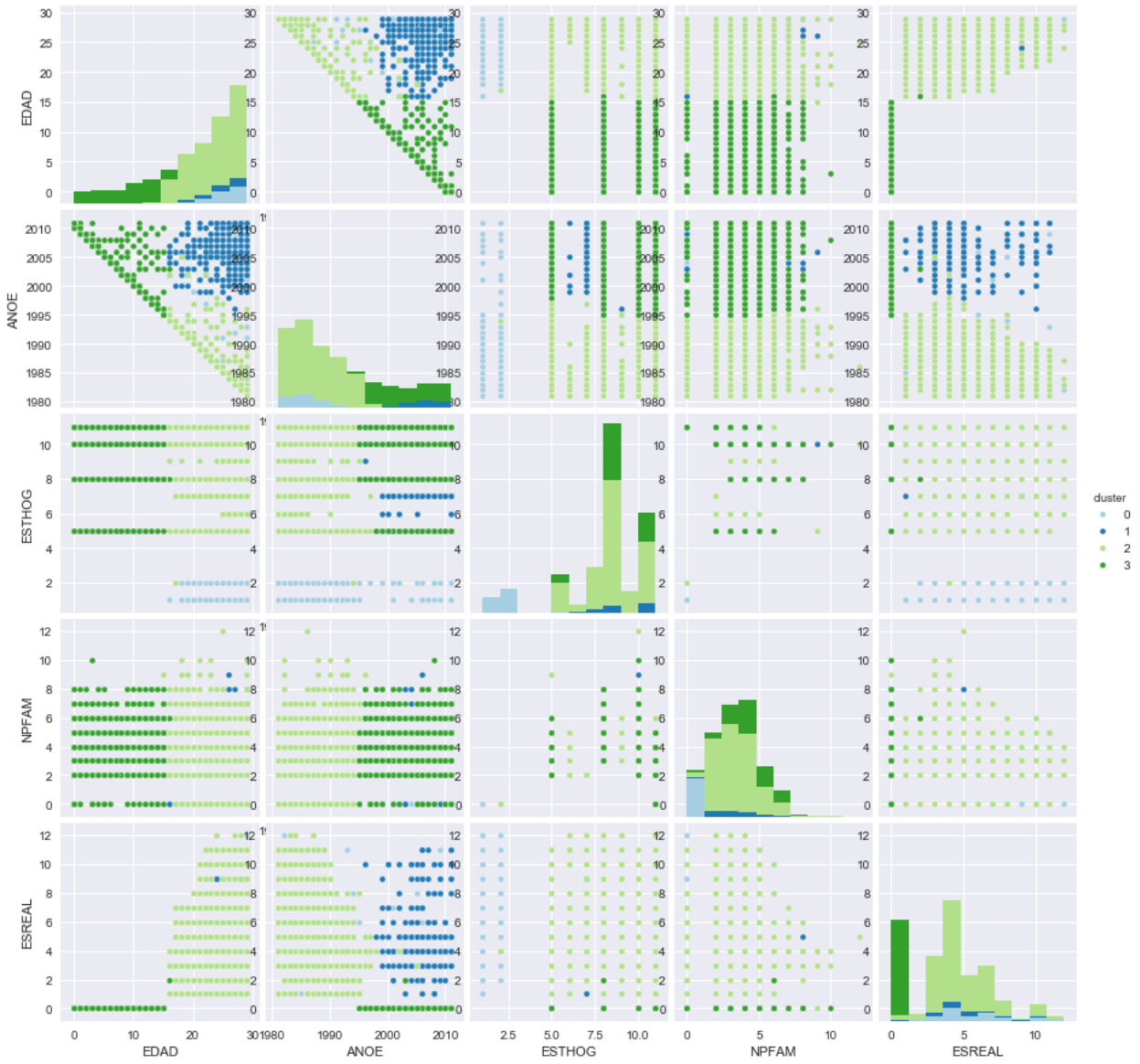

Figura 17: Scatter Matrix de *Birch*

2.3.5. MeanShift

Modelo	Tiempo (s)	Calinski-Harabaz Index	Coeficiente Silhouette
MeanShift	56.72	3396.625	0.39370

nº cluster	Tamaño	Porcentaje sobre el total
0	4159	63.56
1	1654	25.28
2	429	6.56
3	301	4.60

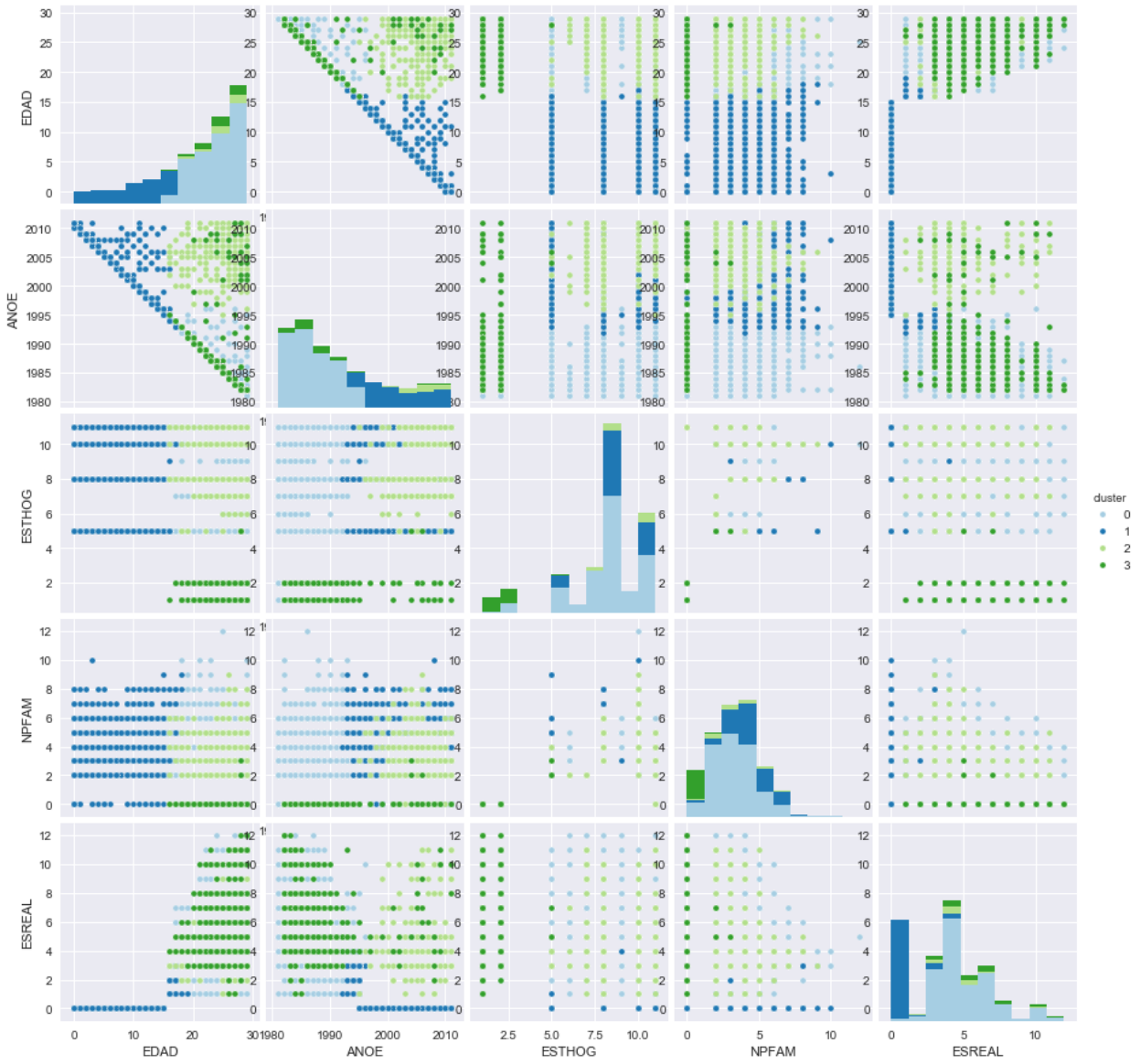

Figura 18: Scatter Matrix de *MeanShift*

En este caso de estudio tenemos un grupo marcado de jóvenes menores de 20 años de los cuales una mínima parte tiene estudios superiores a bachillerato, lo cual podría tener relación con que los padres solo tengan estudios básicos. Si nos fijamos en la franja de edad entre 20 y 30 años, vemos que hay tres grupos mezclados. Uno de ellos se restringe entre 25 y 30 años y estos jóvenes disponen de bachillerato o superior, incluyendo carrera universitaria. por otra parte un gran número de jóvenes entre 20 y 30 años solo disponen de bachillerato o menos.

De esto podemos inferir que podría haber cierta relación entre los estudios de los padres y que los hijos a una edad temprana abandone los estudios para dedicarse a otras cosas, sin embargo los que continúan, aunque a una edad mas tardía logran obtener estudios superiores.

Como comentario final en relación a los tres casos, observamos que los jóvenes granadinos en su

mayoría no disponen de estudios universitarios o superior. Las causas como hemos visto pueden ser que sus padres solo disponen de estudios básicos, pero no depende de la estructura de la que se compone el hogar como podría la intuición decirnos.

Por otro lado, que los padres no dispongan de estudios hemos visto que depende sobre todo de la generación a la que pertenecen y no tanto de que tuviesen muchos hijos por ejemplo. Esto podríamos decir que es resultado del avance en el sistema educativo y la cantidad de ayudas y becas disponibles para las familias.