Trabajo Practico 1

Informe Final

Aplicaciones de Datamining en Ciencia y Tecnología

Integrantes: Mario Rossi, Fernando Menéndez, Fabio Zilberman y Juan Ignacio Etcheberry Mason

28 de Octubre de 2018.

CRISP DM

Introducción

El presente trabajo practico consiste en aplicar la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) para el análisis e identificación de estrellas pertenecientes al clúster abierto de la *Hiades*. CRISP-DM es una de la metodología más utilizada en los proyectos de implementación de minería de datos y está dividida en cuatro niveles de abstracción organizados en forma jerárquica y consta en seis fases distintas, algunas de las cuales son bidireccionales. A continuación, se listan y se describen los distintos pasos y los correspondientes resultados obtenidos:

1. **Fase de Comprensión de Dominio**

- *Objetivos*

El objetivo del presente proyecto es hallar de la forma más fiable estrellas que pertenezcan al clúster *Hiades* a partir de información obtenida de los catálogos *Hipparcos* y *Tycho* que representan los productos primarios de la misión astrométrica *Hipparcos* de la Agencia Espacial Europea (ESA).

- *Evaluar situación*

Para el desarrollo de este proyecto se utilizará información obtenida a partir de un análisis preliminar (pre-informe) realizado cruzando datos obtenidos de la base de datos astronómica *SIMBAD* con los datos del catálogo *Hipparcos*. Esta comparación nos permitió identificar una lista de 50 estrellas presentes en el catálogo de *Hipparcos* que corresponden a elementos del clúster abierto *Hiades*. Estas estrellas serán utilizadas como validación para identificar elementos similares basados en sus características de posición, movimiento y espectro de emisión.

- *Objetivos del Data Mining*

En particular para este proyecto se aplicarán distintos algoritmos de aprendizaje no supervisado. Se utilizará clustering por particiones de modo de agrupar los elementos de los distintos *Datasets* analizados en conjuntos lo más homogéneos posible. Esperamos que esta estrategia nos permita identificar uno o pocos clústeres que contengan a la gran mayoría de las estrellas *Hiades* y que además engloben a otras estrellas que puedan representar candidatas a

pertenecer al cluster de las *Hiades*. En este sentido, si bien los algoritmos de clustering son métodos de aprendizaje no supervisado, en el contexto del presente proyecto se aplicará una estrategia de aprendizaje semi supervisado que nos permita utilizar los algoritmos de clustering como clasificadores.

- Plan de proyecto

El proyecto se dividirá en las siguientes etapas para facilitar su organización:

• *Etapa 1*: Análisis de la estructura de los datos y la información de las bases de datos.

• *Etapa 2*: Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la minería de datos sobre ellos.

• *Etapa 3*: Aplicación de distintas técnicas de modelado (en nuestro caso utilizaremos sólo clustering) y ejecución de las mismas sobre los datos. En una primera instancia se analizará el *Dataset Hipparcos* y luego el *Dataset Thyco*.

• *Etapa 4*: Análisis de los resultados obtenidos en la etapa anterior, si fuera necesario repetir la etapa 3.

• *Etapa 5*: Producción de un informe con los resultados obtenidos en función de los objetivos propuestos.

2. Fase de Comprensión inicial de Datos

Esta fase consiste en familiarizarse con los datos y estudiar su calidad y estructura, así como identificar las relaciones más evidentes para formular las primeras hipótesis

- *Colección inicial de Datos*

Los datos utilizados para el desarrollo del trabajo derivan de dos catálogos de estrellas (*Hipparcos* y *Thyco*) que contienen información relativa a estrellas obtenidas por la misión astrométrica *Hipparcos* en el marco del programa científico de la Agencia Europea Espacial (ESA). En particular los datos disponibles corresponden a una región acotada del espacio donde se sabe se encuentran las estrella del clúster abierto de las *Hiades*.

- *Describir los datos*

Cada uno de los *Datasets* contiene distintas variables relativas a la posición, movimiento y espectro de emisión de las distintas estrellas: RA= ascensión recta, DE = declinación, Plx = paralaje, pmRA = movimiento propio en ascensión recta, pmDE= movimiento propio en declinación, B-V = diferencia de magnitud entre filtro Johnson B (400 - 500 nm rango del azul) y filtro Johnson V (500 - 600 nm corresponde con lo que ve el ojo humano), BT = magnitud entre 400 - 500 nm del sistema fotométrico de *Tycho*, VT= Magnitud entre 500 - 600 nm del sistema fotométrico de *Tycho*. Las mediciones del catálogo *Hipparcos* (alrededor de 2600 estrellas) fueron realizadas con mayor precisión que las obtenidas en el catálogo *Thyco* (alrededor de 16000). Los datos del catálogo *Hipparcos* fueron utilizados en un análisis preliminar (pre-informe), donde, utilizando sólo criterios de semejanza posicional basados en los parámetros: ascensión recta (RA) y declinación (DE). Mediante este análisis se identificaron y clasificaron una lista de 50 estrellas que correspondían al clúster abierto de las *Hiades* que representan nuestros patrones de referencia para la identificación de candidatas *Hiades* en los dos *Datasets*.

- *Exploración de los datos*

En una primera instancia realizamos una análisis exploratorio para ver la distribución y tipo de variables presentes en los dos *Datasets*.

```
### Data Frame Summary
```

hip
N: 2655

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	HIP\ [numeric]	mean (sd) : 21210.44 (4320.41)\ min < med < max :\ 13510 < 21089 < 28887\ IQR (CV) : 7310 (0.2)	2655 distinct values	2655\ (100%)	0\ (0%)
2	RA_J2000\ [numeric]	mean (sd) : 4.54 (0.91)\ min < med < max :\ 2.9 < 4.52 < 6.1\ IQR (CV) : 1.57 (0.2)	2655 distinct values	2655\ (100%)	0\ (0%)
3	DE_J2000\ [numeric]	mean (sd) : 13.9 (6.16)\ min < med < max :\ 3.41 < 13.89 < 24.5\ IQR (CV) : 10.72 (0.44)	2655 distinct values	2655\ (100%)	0\ (0%)
4	Plx\ [numeric]	mean (sd) : 8.83 (10.07)\ min < med < max :\ 0.03 < 6.06 < 172.78\ IQR (CV) : 6.42 (1.14)	1361 distinct values	2655\ (100%)	0\ (0%)
5	pmRA\ [numeric]	mean (sd) : 22.5 (74.69)\ min < med < max :\ -330.66 < 4.78 < 1999.05\ IQR (CV) : 28.99 (3.32)	2245 distinct values	2655\ (100%)	0\ (0%)
6	pmDE\ [numeric]	mean (sd) : -29.4 (72.37)\ min < med < max :\ -1570.64 < -13.19 < 238.42\ IQR (CV) : 26.9 (-2.46)	2196 distinct values	2655\ (100%)	0\ (0%)
7	Vmag\ [numeric]	mean (sd) : 8.34 (1.51)\ min < med < max :\ 0.45 < 8.42 < 12.66\ IQR (CV) : 1.8 (0.18)	656 distinct values	2655\ (100%)	0\ (0%)
8	B-V\ [numeric]	mean (sd) : 0.63 (0.46)\ min < med < max :\ -0.22 < 0.56 < 3.1\ IQR (CV) : 0.66 (0.73)	1269 distinct values	2640\ (99.44%)	15\ (0.56%)
9	Symbad_Hyades\ [logical]		2605 (98.1%)\ 50 (1.9%)	2655\ (100%)	0\ (0%)

Data Frame Summary
tyc
N:
16258

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	recno\ [numeric]	mean (sd) : 73094.02 (45031.95)\ min < med < max :\ 3569 < 57089.5 < 156305\ IQR (CV) : 52262.5 (0.62)	16258 distinct values	16258\ (100%)	0\ (0%)
2	TYCID1\ [numeric]	mean (sd) : 897.7 (535.94)\ min < med < max :\ 51 < 723 < 1868\ IQR (CV) : 638 (0.6)	283 distinct values	16258\ (100%)	0\ (0%)

3	TYCID2\ [numeric]	mean (sd) : 893.51 (635.11)\ min < med < max :\ 1 < 789 < 3944\ IQR (CV) : 842 (0.71)	2616 distinct values	16258\ (100%)	0\ (0%)
4	TYCID3\ [numeric]	mean (sd) : 1 (0.02)\ min < med < max :\ 1 < 1 < 2\ IQR (CV) : 0 (0.02)	1 : 16254 (100.0%)\ 2 : 4 (0.0%)	16258\ (100%)	0\ (0%)
5	RA_J2000_24\ [numeric]	mean (sd) : 4.86 (0.91)\ min < med < max :\ 2.9 < 5.06 < 6.1\ IQR (CV) : 1.49 (0.19)	16258 distinct values	16258\ (100%)	0\ (0%)
6	DE_J2000\ [numeric]	mean (sd) : 13.97 (6.24)\ min < med < max :\ 3.4 < 14.05 < 24.5\ IQR (CV) : 11.02 (0.45)	16258 distinct values	16258\ (100%)	0\ (0%)
7	pmRA\ [numeric]	mean (sd) : 6.26 (23.53)\ min < med < max :\ -149.3 < 1.8 < 198.2\ IQR (CV) : 13 (3.76)	1480 distinct values	16258\ (100%)	0\ (0%)
8	pmDE\ [numeric]	mean (sd) : -12.44 (20.92)\ min < med < max :\ -199.9 < -7.6 < 177\ IQR (CV) : 14.2 (-1.68)	1349 distinct values	16258\ (100%)	0\ (0%)
9	BT\ [numeric]	mean (sd) : 10.74 (1.22)\ min < med < max :\ 2.79 < 11.04 < 12.85\ IQR (CV) : 1.36 (0.11)	4343 distinct values	16258\ (100%)	0\ (0%)
10	VT\ [numeric]	mean (sd) : 9.94 (1.1)\ min < med < max :\ 0.77 < 10.22 < 11.95\ IQR (CV) : 1.23 (0.11)	4048 distinct values	16258\ (100%)	0\ (0%)
11	V\ [numeric]	mean (sd) : 9.87 (1.1)\ min < med < max :\ 0.58 < 10.14 < 11.93\ IQR (CV) : 1.24 (0.11)	15844 distinct values	16258\ (100%)	0\ (0%)
12	B-V\ [numeric]	mean (sd) : 0.68 (0.45)\ min < med < max :\ -0.46 < 0.56 < 3.39\ IQR (CV) : 0.66 (0.67)	3439 distinct values	16258\ (100%)	0\ (0%)
13	HD\ [numeric]	mean (sd) : 137082.43 (117395.4)\ min < med < max :\ 18019 < 39727.5 < 287468\ IQR (CV) : 219645 (0.86)	5751 distinct values	5758\ (35.42%)	10500\ (64.58%)
14	HIP\ [numeric]	mean (sd) : 21327.99 (4326.46)\ min < med < max :\ 13526 < 21257 < 28882\ IQR (CV) : 7368.5 (0.2)	2470 distinct values	2483\ (15.27%)	13775\ (84.73%)
15	Plx\ [numeric]	mean (sd) : 7.99 (6.45)\ min < med < max :\ 1.01 < 6.09 < 87.9\ IQR (CV) : 5.52 (0.81)	1135 distinct values	2253\ (13.86%)	14005\ (86.14%)

- *Verificar calidad de los datos*

Todas las variables de los dos *Datasets* analizados (salvo "*Symbad_Hyades*" del *Dataset Hipparcos* que creamos nosotros) son numéricas y la gran mayoría presenta el 100% de datos validos. Existen solo 15 registros de *Dataset Hipparcos* (0.56%) que presentan datos faltantes en la variable ***B-V***. Por el contrario, en el *Dataset Thyco* observamos que la variable ***Plx*** presenta un número muy alto de faltantes (86.14%). Una situación similar presentan las variables ***HD*** y ***HIP*** del *Dataset Thyco*. Estas dos ultimas variables representan la ID cruzada del catálogo *Thyco* con el catálogo HD y Hipparcos respectivamente, y en el caso de la ultima sirve de referencia para poder combinar los datos de los dos *Datasets* analizados.

3. Fase de Preparación de los datos (/02.Preprocesamiento/pre_procesamiento.rmd)

En esta fase de la metodología se trata de preparar los datos para adecuarlos a las técnicas de minería de datos que se van a emplear sobre ellos. Esto implica seleccionar el subconjunto de datos que se va a utilizar, limpiarlos para mejorar su calidad, añadir nuevos datos a partir de los existentes y darles el formato requerido por la herramienta de modelado.

- *Obtener conjunto inicial de datos*

Los *Datasets* antes descriptos representan nuestro conjunto de datos iniciales y cabe aclarar que para la clusterización no se incluyen las variables que dan información de ID cruzada. Para ambos casos sin embargo, preservamos el campo *symbad hyade*s para luego mapear las *Hiades* por clúster cómo método de validación.

- *Limpieza de los datos*

En análisis de clúster, un método usualmente aceptado es remover instancias que posean faltantes de la clusterización inicial. Luego con clúster bien definidos, se suele intentar mapear las instancias con faltantes a los clúster más parecidos.

Con esta posibilidad en mente, removemos 15 estrellas del *Dataset* de *Hipparcos* para trabajar con un conjunto de datos completos a la hora de clusterizar. En este sentido, para el caso del *Dataset Thyco*, la variable ***Plx*** presenta un número muy elevado de faltantes (86.14%) y por lo tanto no se la consideró para el análisis de clusterización. Asimismo, se procedió a remover las columnas de IDs otros catálogos del *Dataset Thyco*, ***TYCID1***, ***TYCID2*** y ***TYCID3***, ***HD*** y ***HIP***.

- *Integración de los datos*

Hemos considerado que no es necesario la creación de nuevas estructuras (campos, registros, etc). Sin embargo, con el fin de no repetir el procedimiento para los dos *Datasets*, eliminamos las estrellas del catálogo *Hipparcos* del *Dataset Thyco*, excepto las estrellas *Hiades* que servirán como grupo de control para el análisis.

- *Formateo de los datos*

Para ambos *Datasets*, se procedió a escalar y estandarizar las variables para evitar que las dimensiones de estas incidan o afecten las medidas de distancia dentro del análisis de clúster.

Para las variables RA_J2000 y DE_J2000, se realizará un escalado especial que permita conservar las proporciones entre ambas variables, ya que consideramos que una "estandarización" para estas dos podría resultar en una posible distorsión de la medición que se posee de la proyección de la estrella en el espacio.

4. Fase de Modelado (/03.Modelo_Evaluacion/modelo_evaluacion.Rmd para **Hipparcos y /03.Modelo_Evaluacion/tycocloud.R** para ***Thyco*)**

En esta fase de la metodología se escogerá la técnica (o técnicas) más apropiadas para los objetivos marcados de la minería de datos.

– *Seleccionar la técnica del modelo*

Debido a que las consignas especificas del trabajo práctico nos piden realizar un análisis de clusterización no jerarquica utilizamos en una primera instancia pruebas sólo con el *Dataset Hipparcos* con el método de K-medias y El método de clustering por prototipos PAM (*Partition around Medoids*).

– *Generar diseño de prueba*

Para probar la calidad y validez de los modelos se utilizarán criterios de validación interna. En particular se utilizaron el análisis de Silhouette y la suma de los errores al cuadrado (SSE).

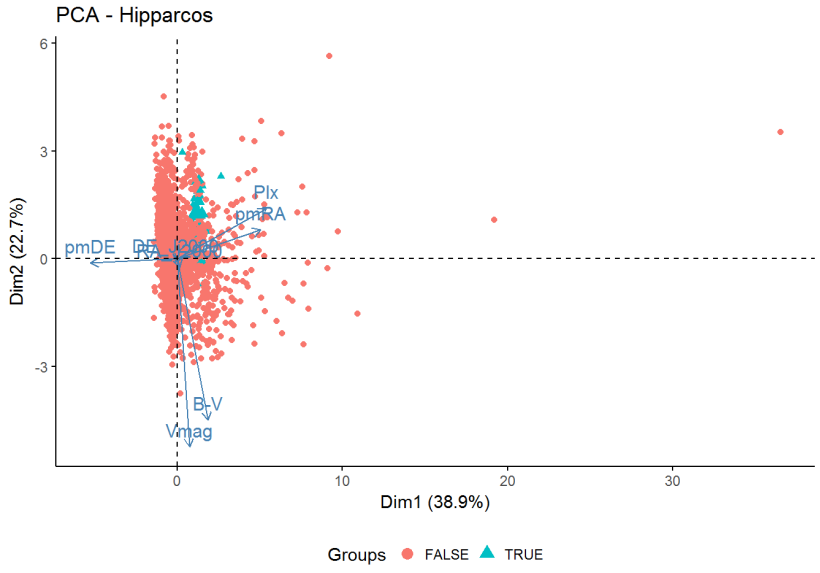
– *Construir el modelo*

A continuación se procederá a ejecutar el modelo elegido sobre los datos de entrenamiento. En este apartado se describirán los ajustes de parámetros del modelo que se eligen en la herramienta de minería de datos, así como la salida de dicho modelo y su descripción.

En un principio trabajaremos con el *Dataset* de *Hipparcos* porque contiene información muy similar a la de *Tychos* y es de menor tamaño permitiendo analizar y probar distintas métricas utilizando menos recursos computacionales.

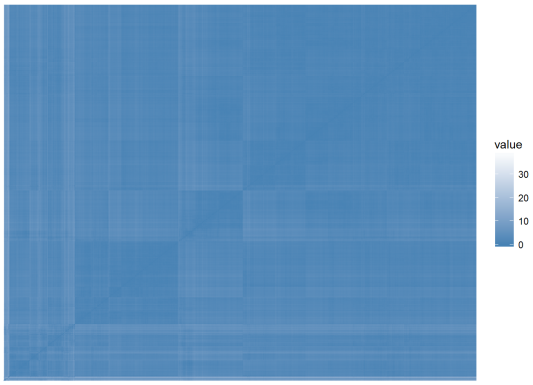
Análisis de tendencia de clusterización

Realizamos un análisis de PCA para estudiar si existen indicios para pensar que el *Dataset* de Hipparcos podría ser clusterizable.



Las *Hiades* se indican como triángulos verdes y se observar que se encuentran concentradas. Además se observa que las variables RA_J2000 y DE_J2000 no contribuyen demasiado a explicar la variabilidad observada.

A continuación se procedió a hacer un análisis de tendencia a la clusterización del conjunto de datos. Para esto se utilizará la función “get_clust_tendency” que tiene implementado el cálculo del estadístico de Hopkins.



El estadístico de Hopkins nos indicar que este conjunto no es clusterizable, ya que su valor es cercano a 0. (Un conjunto con fuerte tendencia a la clusterización debería tener un estadístico >0.5)

Sin embargo para este caso no queremos clusterizar perfectamente todo el *Dataset* y por el contrario estamos solamente interesado en clusterizar las Hiades. Por lo tanto decidimos avanzar con la creación de los modelos.

A continuación realizamos el análisis de Cluster tomando dos métodos distintos: -Kmeans (o K-medias) -PAM (Partition around Medoids)

- *Analizar el modelo*

Análisis utilizando K-medias

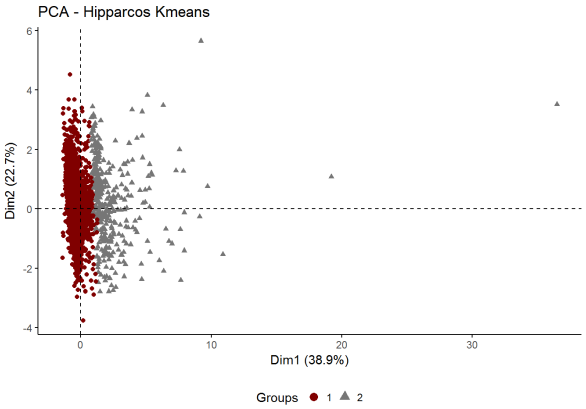
Para el análisis de K-medias se llamará a la librería NbClust que cuenta con un método de validación por votación a partir del cálculo de 30 índices y métricas distintas de ajuste de los clústers, para determinar cual debería ser la cantidad adecuada de k a utilizar.

Como la función NbClust de su librería original poseía ciertas restricciones respecto a la cantidad de parámetros que se pueden pasar al algoritmo Kmeans, se generó una función auxiliar “NbClust2” a partir del código fuente original de la librería (/03.Modelo_Evaluacion/NbClust2_kmeans.R). Con esta pequeña modificación se puedo extender la cantidad iteraciones máximas posibles del K-medias hasta hallar la convergencia.

```
## *** : The D index is a graphical method of determining the number of clusters.
##       In the plot of D index, we seek a significant knee (the significant peak in Dindex
##       second differences plot) that corresponds to a significant increase of the value of
##       the measure.
##
## *******************************************************************
## * Among all indices:
## * 7 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 5 proposed 4 as the best number of clusters
## * 2 proposed 5 as the best number of clusters
## * 1 proposed 8 as the best number of clusters
## * 3 proposed 9 as the best number of clusters
## * 3 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
```

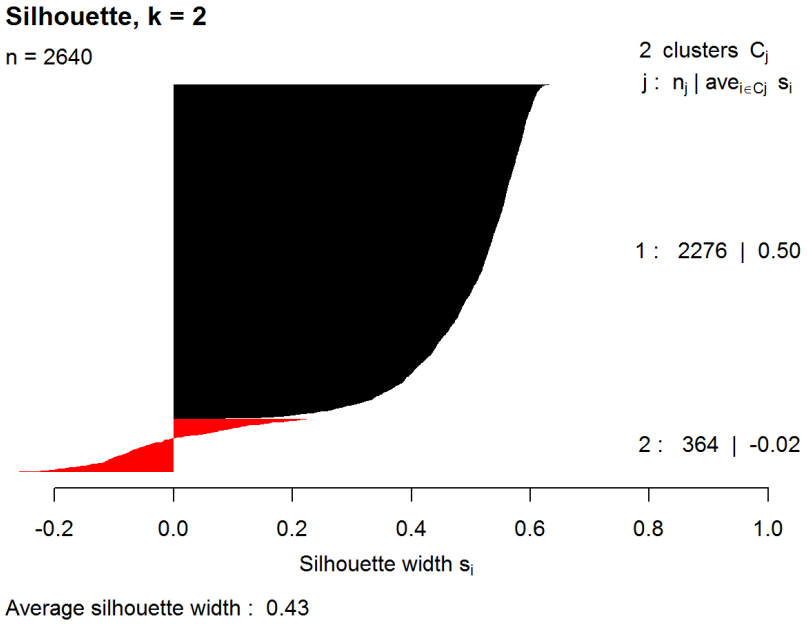
De todos los índice y métricas que el NbClust corrió, 7 de ellos propusieron 2 clústers como la partición que mejor ajusta a los datos.

A continuación se guardó el modelo de K-medias para k=2 y se procedió a graficar la proyección de los clúster en los ejes de las componentes principales.



De los resultados del análisis de Kmeans se extrae que de la clusterización, 49 Hiades han sido clasificadas dentro de un mismo clúster: el número 2. Dentro del clúster número 2 han sido agrupadas otras 315 estrellas que podrían resultar en potenciales candidatas a formar parte del clúster de estrellas *Hiades*.

A continuación analizamos los resultados de Silhoutte y en particular nos focalizamos en el resultado del clúster numero 2.



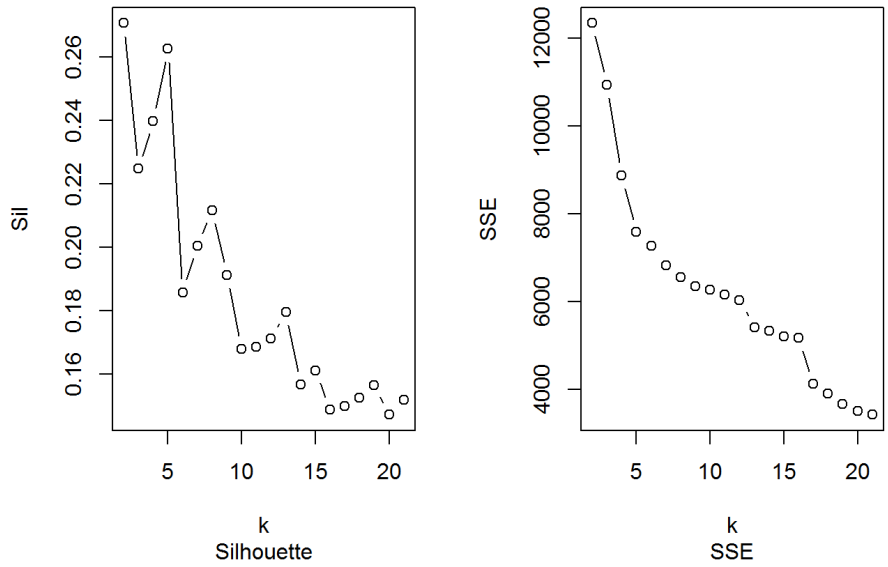
De observar el gráfico de *Silhouette*, surge que en el clúster número 2, no todas las estrellas parecerían ser similares hacia dentro del clúster. De hecho, parecería ser que un poco más de la mitad de ellas debería pertenecer al clúster 1 en vez del clúster 2 (poseen valores de Silhouette negativos).

A continuación se procedió a filtrar por todas las estrellas pertenecientes al clúster 2 que además posean un score de *Silhouette* positivo para verificar cuantas *Hiades* estaban contenidas dentro de este grupo.

Del listado del clúster 2 se desprende que las estrellas que poseen *Silhouette* más bajo son aquellas que han sido identificadas como verdaderas *Hiades* y por lo tanto es probable que el clúster hasta aquí analizado posea algún grupo de estrellas no compatible con las características de la *Hiades* o mucho más cercanas entre sí que las *Hiades* mismas.

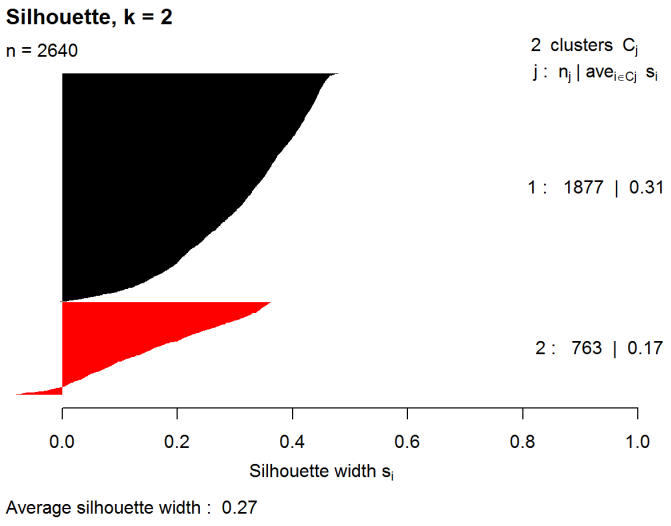
Análisis utilizando PAM (Partition Around Medoids)

Como primer paso realizamos una evaluación de los valores de *Silhouette* promedio y la suma de los errores al cuadrado utilizando un numero variable de k entre 2 y 21 (/03.Modelo_Evaluacion/PAM_loop.r)



El análisis conjunto de los gráficos de *Silhouette* y SSE parecería indicar que el numero optimo de clúster es 2 o 5. En este sentido se observa que para estos valores el valor de *Silhouette* promedio es máximo y además para el grafico de SSE se observa un quiebre en la pendiente para k igual a 5.

<u>Análisis PAM para K=2</u>

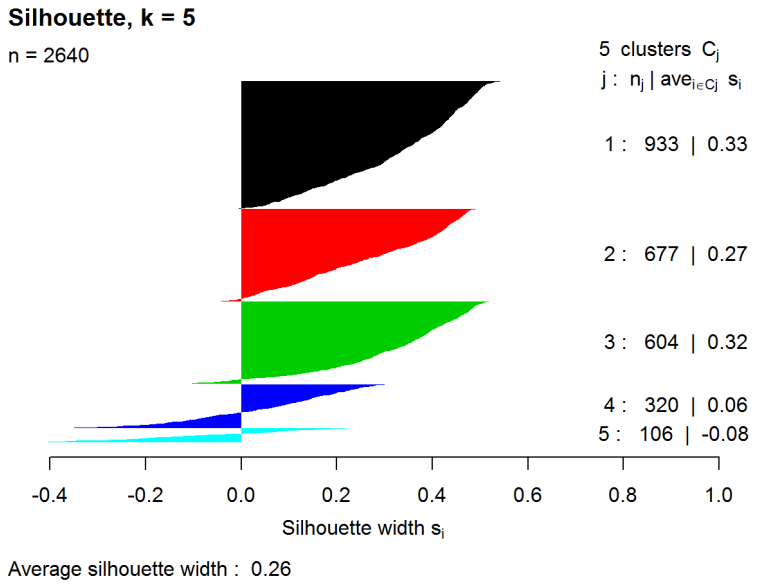


Si ambos cluster poseen un valor de *Silhouette* medio positivo, la gran mayoría de la *Hiades* (45 de 50) están clusterizadas en el clúster numero 1 donde se encuentran el 70% de los elementos del *Dataset*.

```
##         Cluster
## Hiades     1    2
##   FALSE 1832  758
##   TRUE    45    5
```

Se conoce que el número de *Hiades* en esa región del espacio es pequeña con respecto a todas las estrellas que están contenida y por lo tanto concluimos que no podemos utilizar la clusterización con k =2 para clasificar nuevas posibles candidatas a *Hiades*.

Análisis PAM para K=5



En este caso, observamos que existen también varios clúster con valores de *Silhouette* adecuados y por lo tanto estudiamos cuantas y en cuales clúster quedaron clasificadas las *Hiades*.

```
##         Cluster
## Hiades     1   2   3   4   5
##   FALSE  933 676 604 271 106
##   TRUE     0   1   0  49   0
```

Como se observa en la matriz construida en base a los resultados obtenidos de la clusterización, existe un unico clúster (el numero 4) que contiene a casi todas las *Hiades* y además todas ellas poseen un valor de *Silhouette* positivo.

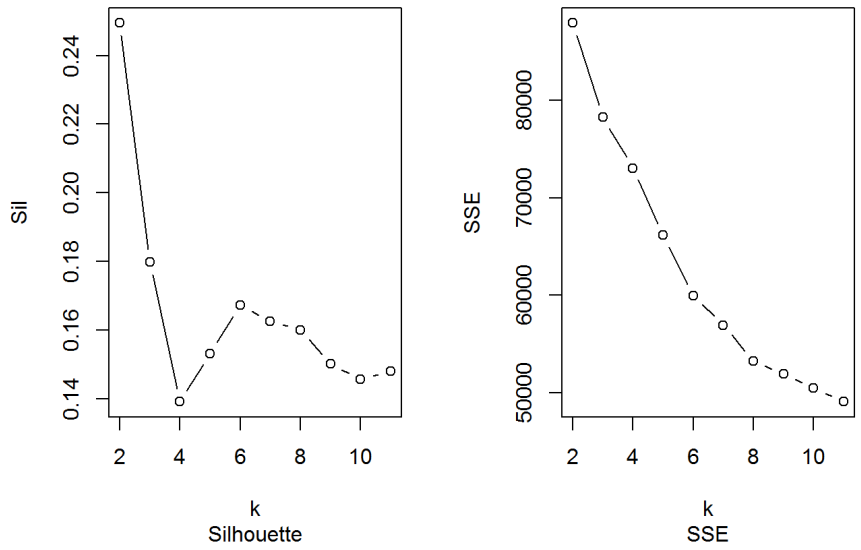
Por lo tanto este clúster es el mejor clúster donde buscar estrellas candidatas a ser *Hiades*.

Análisis del *Dataset Thyco* utilizando PAM (Partition Around Medoids)

A continuación realizamos el análisis de *Dataset Thyco* del cual, como se indico anteriormente, se habían eliminado todos los elementos en común con el *Dataset Hipparcos*. Considerando que los dos *Datasets* poseen información

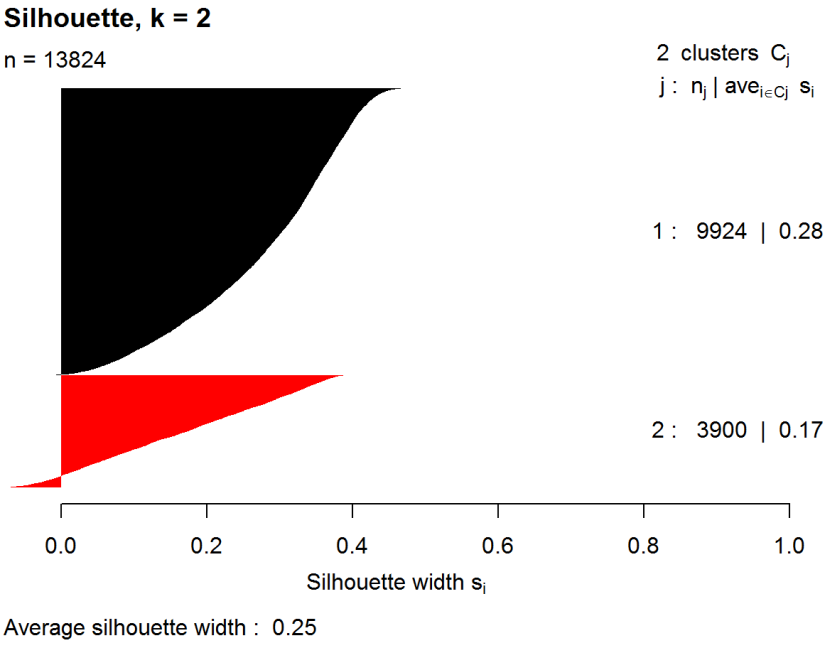
similar decidimos evaluar sólo el método PAM que nos había permitido obtener mejores resultados cuando lo aplicamos al *Dataset* de *Hipparcos*.

Como primer paso realizamos una evaluación de los valores de *Silhouette* promedio y la suma de los errores al cuadrado (SSE) utilizando un numero variable de k entre 2 y 11 (/03.Modelo_Evaluacion/PAM_loop.r).



El análisis conjunto de los gráficos de *Silhouette* y SSE parecería indicar que el número óptimo de clústers es 2 o 6. En este sentido se observa que para estos valores el valor de *Silhouette* promedio es máximo y además para el grafico de SSE se observa un leve quiebre en la pendiente para k igual a 6.

Análisis PAM para K=2



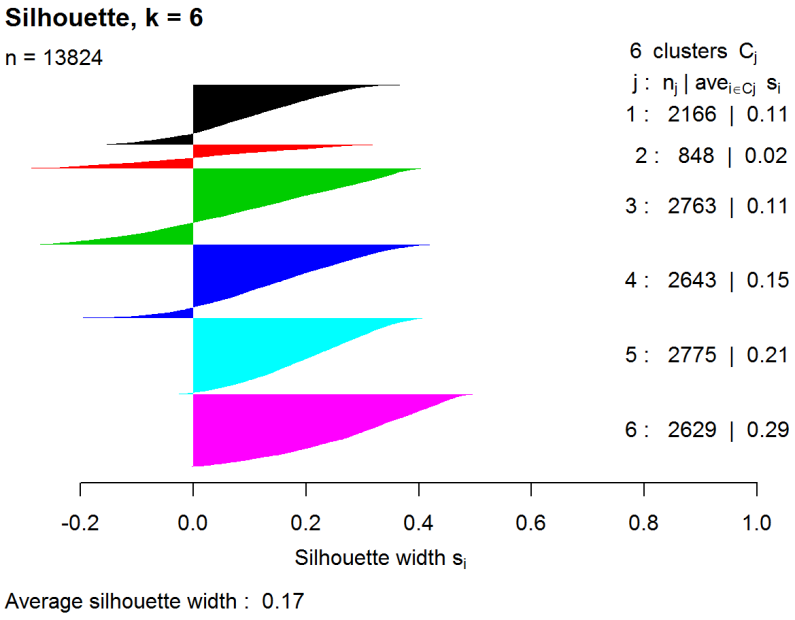
Se observa que los dos clúster identificados tienen valores aceptables de *Silhouette* medio y por lo tanto decidimos estudiar la distribución de las *Hiades* en ambos clúster.

Como se observa en la siguiente tabla todas las *Hiades* menos una se encuentran en el clúster más pequeño (número 2) y además todas poseen valores de *Silhouette* positivos.

```
##         Cluster
## Symbad      1    2
##   FALSE  9924 3851
##   TRUE      0   49
```

Análisis PAM para K=6

Considerando que la clusterización con k igual a 2 nos agrupó un número muy grande de estrellas junto con las *Hiades*, analizamos si aumentando el valor de k podemos mejorar la clusterización.



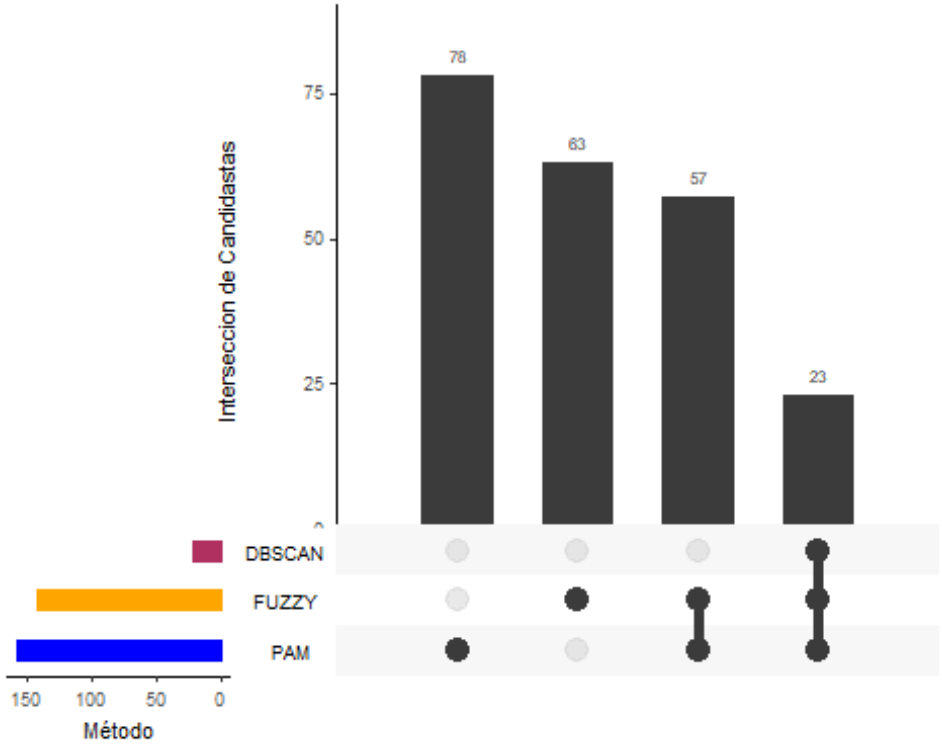

```
##         Cluster
## Symbad      1    2    3    4    5    6
##   FALSE  2166  840 2722 2643 2775 2629
##   TRUE      0    8   41    0    0    0
```

Observamos que la mayoría de las *Hiades* clusterizan en el clúster numero 3 (41 sobre un total de 50). También en este caso todas poseen valor de *Silhouette* positivas. Sin embargo, si bien se reduce el número de estrellas clusterizadas junto con las *Hiades* (de 3850 a 2722) un gran numero de *Hiades* (8) se encuentran en el clúster 2 y asimismo los valores de *Silhouette* observados son menores comparados a aquellos registrados utilizando un k igual 2. Por todo esto consideramos que obtuvimos un mejor resultados con k igual a 2.

5. Conclusiones

Desde el punto de vista de la utilización de metodología de clúster, pudimos observar que en el caso especifico de nuestros *Datasets*, de todas las estrategia de clusterizacion utilizadas (k-medias, PAM, DB-scan y Fuzzy), la

metodología de PAM parece haber brindado los mejores resultados. SI bien la aplicación de los distintos métodos de clusterización no permitieron una definición de clúster con características marcadamente distintas entre si, permitieron acotar un espacio de búsqueda que se le puede proporcionar a expertos en el campo para su ulterior análisis. Es importante destacar que la aplicación de distintas metodologías de clusterización permitió identificar un cierto numero de candidatas en común. En el siguiente gráfico se puede observar que las estrellas detectadas en DBSCAN fueron identificadas tanto en PAM como en clustering difuso. Los 3 métodos comparten un total de 80 estrellas, entendemos a partir de esta validación que estas tienen mayor chance de pertenecer efectivamente al grupo de *Hiades*.



Los documentos "*Candidatas Hyades - Hipparcos.csv*" y "*Candidatas Hyades - Tycho.csv*" se adjuntan por separado y contienen la lista de estrellas identificadas como candidatas *Hiades* candidatas presentes en el catálogo *Hipparcos* y *Thyco* respectivamente. Asimismos todos los códigos utilizados para el desarrollo del presetne proyecto se puede consultar en ***https://github.com/JMason88/Cluster_Estrellas***

a. Punto adicional I – Estrategia de búsqueda y complejidad algorítmica

En un proceso de búsqueda de equivalencias entre estrellas de dos lotes, donde el patrón de equivalencia es la similitud de coordenadas, utilizaremos la distancia entre las coordenadas de ambos lotes para determinar cuáles son los pares correspondientes.

El algoritmo más trivial es medir las distancias de todas las estrellas de un lote contra todas las del otro. Elegir el par con la menor distancia de todas. Se quitan ambas y se realiza toda la comparación nuevamente.

Dos lotes de estrellas: A y B

La cantidad de estrellas en el lote A es NA, y la cantidad de estrellas del lote B es NB

1. Calculamos NA x NB distancias, armamos una tabla con ID de estrella A, ID de estrella B y la distancia
2. Ordenamos la tabla por distancias de menor a mayor
3. Elegimos el primer elemento. Marcamos la estrella del lote A y la del lote B como "par"
4. Quitamos de la tabla todas las filas que contengan esas estrellas, con el resto de la tabla volvemos al punto 3

La complejidad del proceso es:

A. un cálculo de distancias de tamaño NA * NB
B. recorrido de una tabla de distancias de NA * NB filas

El algoritmo de comparación por grillas propuesto en el TP, propone asignar a cada estrella de ambos a la celda de una grilla utilizando la discretización de los valores de las coordenadas para determinar la celda.

Luego realizaremos los cálculos de distancias entre estrellas de ambos lotes solo de las celdas próximas.

Dos lotes de estrellas: A y B

La cantidad de estrellas en el lote A es NA, y la cantidad de estrellas del lote B es NB

La grilla es de K filas y K columnas

La zona de vecindad es una grilla pequeña de L * L celdas

1. Discretizamos las coordenadas de las NA + NB estrellas y las asignamos a la celda correspondiente, en promedio cada celda recibe NA/K^2 y NB/K^2 estrellas

2. Las zonas de vecindad son de a lo sumo L * L celdas. Ya que las de los bordes tienen menos celdas.

3. Calculamos las distancias entre cada estrella del Lote A, y las estrellas de las celdas vecinas a esta. La tabla de distancias tiene NA * NB/K^2 * L^2 filas

4. Elegimos el primer elemento. Marcamos la estrella del lote A y la del lote B como "par"

5. Quitamos de la tabla todas las filas que contengan esas estrellas, con el resto de la tabla volvemos al punto 6

La complejidad del proceso es:

a. discretización de (NA + NB) * 2 coordenadas
b. un cálculo de distancias de tamaño NA * NB * L^2 / K^2
c. recorrido de una tabla de distancias de NA * NB * L^2 / K^2 filas

Por lo tanto, cuanto mayor sea la relación entre K y L (K/L), más eficiente y menos costoso en tiempo de cálculo resultara este segundo algoritmo.

Generalizando para un tamaño teórico de lotes de N, el proceso de comparación por grilla va a ser de complejidad O(N^2 * L^2 / K^2) comparado contra O(N^2)

Comparando Symbad (NA = 178) con Hipparcos (NB = 2655)

Todos contra Todos:

Time difference of 5.795763 mins

472590 distancias

Time difference of 5.858445 mins total

50x50 celdas / 3x3 celdas

Time difference of 4.019846 secs

2394 distancias

Time difference of 4.202339 secs

197 veces más rápido en cantidad de distancias calculadas (comparado vs Todos contra Todos)

83.6 veces más rápido en tiempo de computo (comparado vs Todos contra Todos)

30x30 celdas / 3x3 celdas

Time difference of 5.907489 secs

6281 distancias

Time difference of 5.998217 secs

75 veces más rápido en cantidad de distancias calculadas (comparado vs Todos contra Todos)

58.7 veces más rápido en tiempo de computo (comparado vs Todos contra Todos)

50x50 celdas / 5x5 celdas

Time difference of 7.251834 secs

6292 distancias

Time difference of 7.328629 secs

75 veces más rápido en cantidad de distancias calculadas (comparado vs Todos contra Todos)

47.9 veces más rápido en tiempo de computo (comparado vs Todos contra Todos)

b. Punto adicional II – Análisis Datos Faltantes de Paralaje en Tycho

Al observar el conjunto de datos en Tycho, se puede observar que existen en total 14005 estrella con datos faltantes de paralaje (Plx) dentro del catálogo de Tycho.

recno	TYCID1	TYCID2	TYCID3	RA_J2000_24	DE_J2000	pmRA
0	0	0	0	0	0	0
pmDE	BT	VT	V	B-V	HD	HIP
0	0	0	0	0	10500	13775
Plx	Symbad_Hyades					
14005	0					

Interesa observar si existe algún patrón en los faltantes y es por eso que se decide ver con detenimiento si existe una relación entre los datos faltantes de paralaje y las estrellas que no fueron incluidas dentro de los catálogos HIP y HD.

Si se filtra el conjunto de datos solamente por el indicador de catálogo HIP y HD indicando las estrellas que pertenecen a los dos catálogos al mismo tiempo, la cantidad de datos faltantes para la variable Plx desciende a unas 164 estrellas únicamente, lo cual significa que el dato de paralaje parecería ser una medición que no fue tenida en cuenta por el catálogo Tycho.

recno	TYCID1	TYCID2	TYCID3	RA_J2000_24	DE_J2000	pmRA
0	0	0	0	0	0	0
pmDE	BT	VT	V	B-V	HD	HIP
0	0	0	0	0	0	0
Plx	Symbad_Hyades					
164	0					

Como estrategia de imputación se podría buscar el dato puntual de Plx directamente al catálogo de Hipparcos. Al hacer Left join entre los dos catálogos, se podrían reducir esas 164 estrellas con datos de paralaje faltantes a un total de 58 estrellas como se desprende de la siguiente tabla:

recno	TYCID1	TYCID2	TYCID3	RA_J2000_24	DE_J2000	pmRA
0	0	0	0	0	0	0
pmDE	BT	VT	V	B-V	HD	HIP
0	0	0	0	0	0	0
Plx.Tyc	Symbad_Hyades	Plx.Hip				
164	0	58				

Por último, no creemos que exista una estrategia de imputación adecuada para el resto de las estrellas que no aparecen los catálogos HIP y HD. Esto se debe principalmente a dos cuestiones que vale la pena marcar:

a) La proporción de datos faltantes con respecto al total de estrellas del catálogo Tycho es demasiado alta y ronda alrededor del 85%.
b) Como ya hemos marcado anteriormente, la mayoría de los métodos de imputación de faltantes resultan adecuados cuando no existe un mecanismo o

lógica subyacente a la distribución de los faltantes. En este caso el valor de Plx responde a que los datos de la misión Tycho no consideraron esta variable para sus mediciones mientras que los otros catálogos sí.

De no existir estas dos limitaciones, se podría proponer un método de imputación a través de una heurística de Vecinos más cercanos o inclusive se podría considerar algún tipo de regresión que permita aproximar los valores de Plx para este caso, siempre y cuando exista algún tipo de relación entre las demás variables y la variable paralaje.

c. Punto adicional IV – Agrupamientos utilizando DBScan y Fuzzy Clustering

DBSCAN

Introducción

Dado el dataset de estrellas de Hiparcos y el set de Hyades identificadas previamente se procede a realizar experimentos de clustering por densidad mediante el algoritmo DBSCAN de la librería fpc. En los mismos se hará énfasis en los siguientes puntos:

- Seleccion de Variables del dataset origen
- Hiperparametrizacion de eps y minPts
- Revision de la presencia de Hyades en los clusters

Para este último punto se intentará identificar cluster relativamente pequeños que contengan la mayor cantidad de las Hyades previamente señaladas como tales.

Construcción de los experimentos

Los experimentos con el algoritmo se construyen a partir de los siguientes pasos:

1. Generación de knnDistPlots sobre los minPts candidatos a fin de identificar los limites superior e inferior de eps para la parametrización.
2. Ejecución y extracción de resultados identificando para cada cluster de cada ejecución la cantidad de Hyades (definidas).
3. Ilustración de máxima cantidad de hyades en un cluster por ejecución, tamaño del cluster y tamaño del cluster en relación con la cantidad de hyades presentes (ratio). Se buscarán los clusters con mayor concentración de Hyades conocidas y de tamaño relativamente pequeño (no más de 500 estrellas)
4. Conclusión y extracción de candidatas cuando corresponda.

Con excepción del primer experimento, los siguientes se realizan con las variables normalizadas

Definiciones

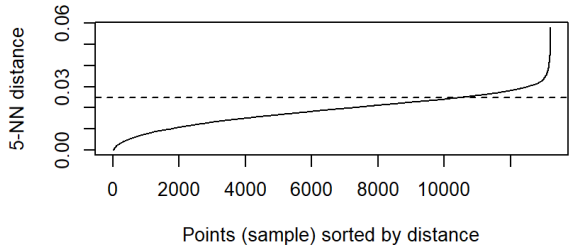
Se definió como rango de minPts de 5 a 50 arbitrariamente. Se establecieron 4 grandes experimentos variando las variables seleccionadas:

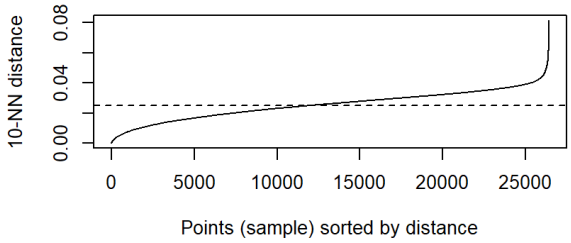
1. Medidas de Posicion (RA y DE)
2. Vmag y B.V
3. Todas las variables de Hiparcos
4. Todas las variables, con excepción de las medidas de posición

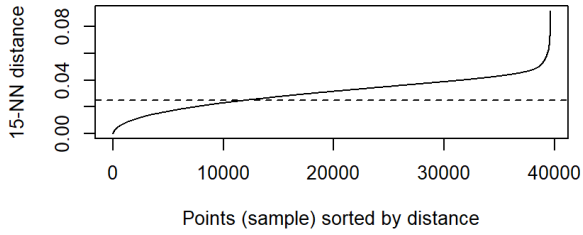
(*) Para cada salida de la hyperparametrizacion, se identifica visualmente si alguno de los test es aceptable para producir candidatos

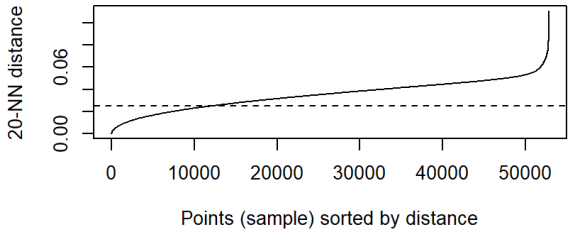
Experimento 1 - Medidas de Posición

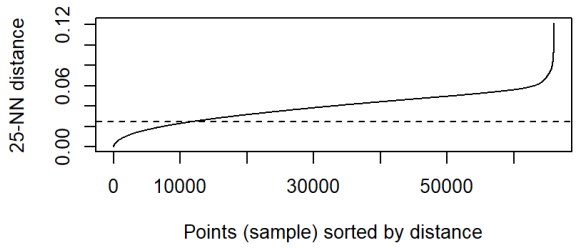
Identificación de EPS:

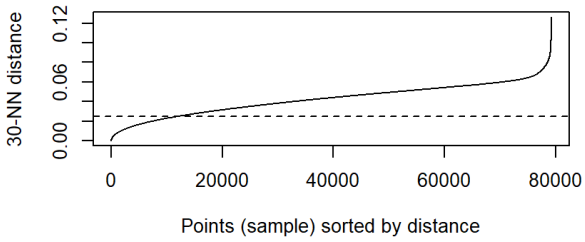


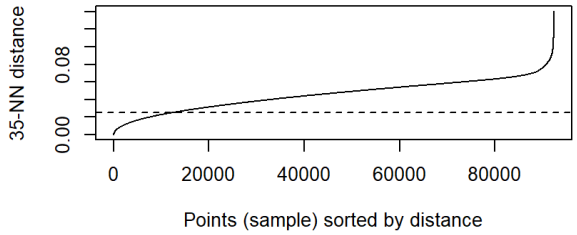


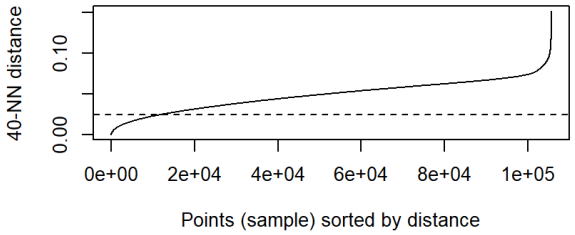


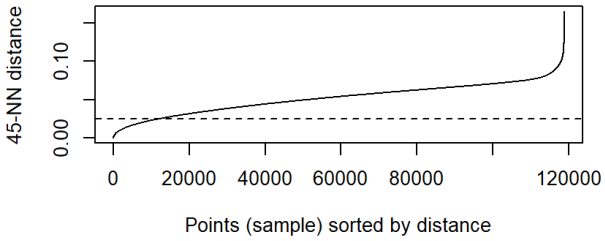


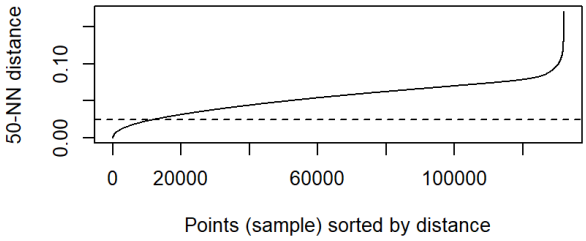




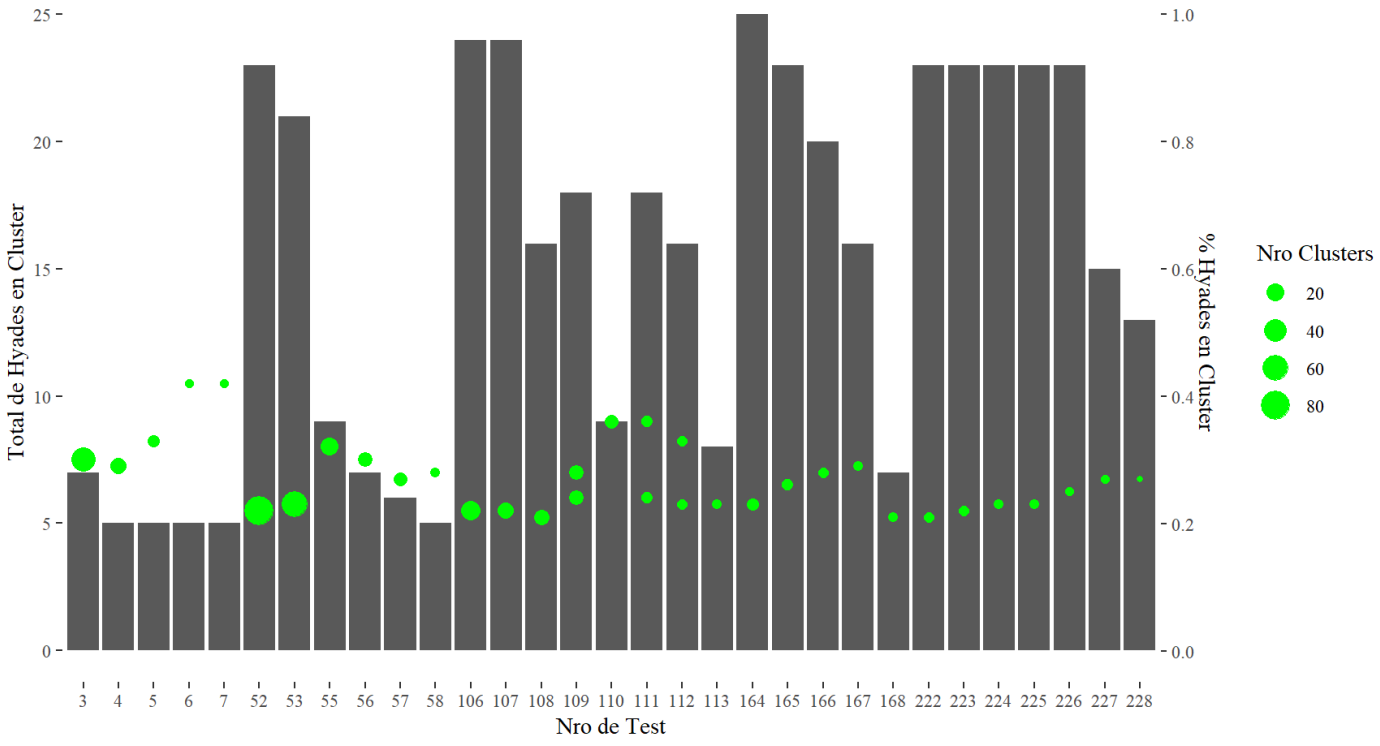








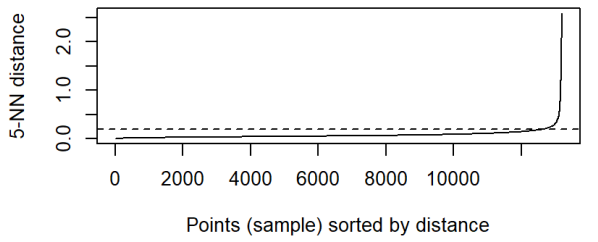
OBS: Basado en los gráficos anteriores se define el rango de *eps* entre 0.02 y 0.1, ejecutándose para todos los *minPts* entre 5 y 50

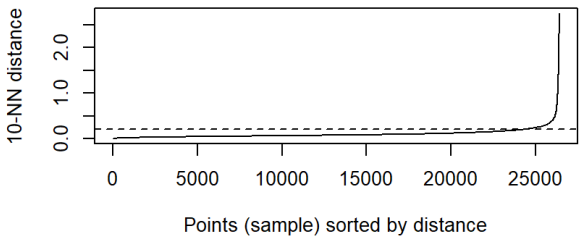


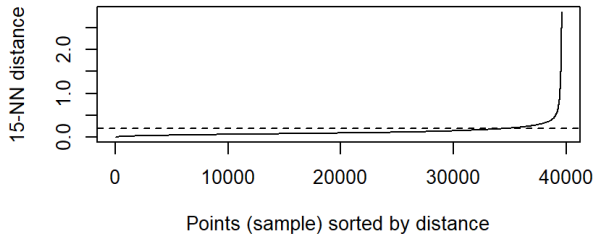
OBS: No se visualizan clusters que presenten suficientes Hyades para tomar otras candidatas

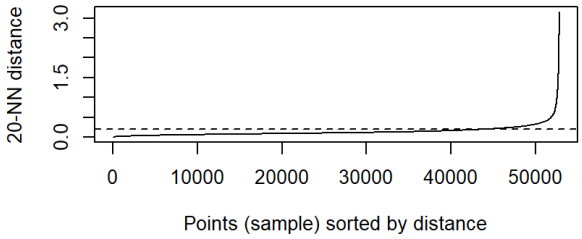
Experimento 2 - B.V y Vmag

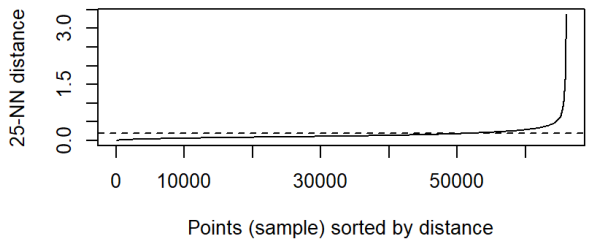
Identificación de EPS:

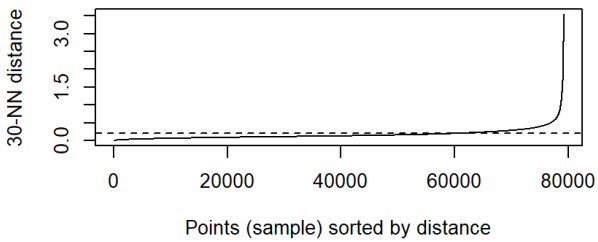


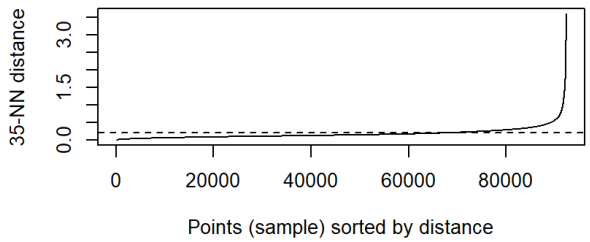


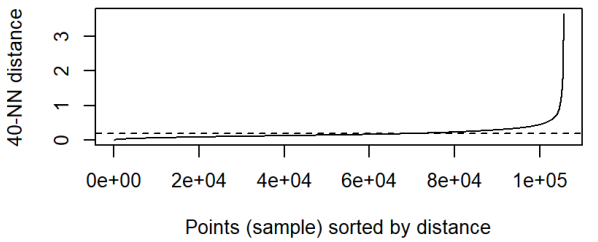


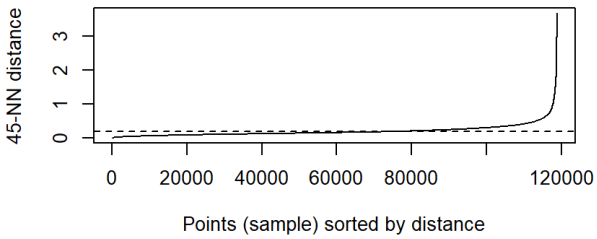


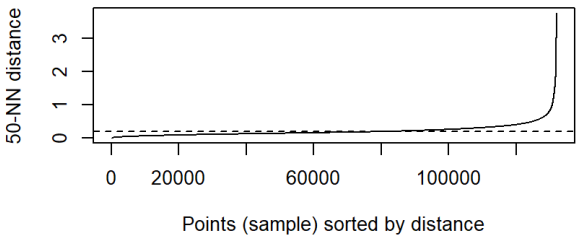




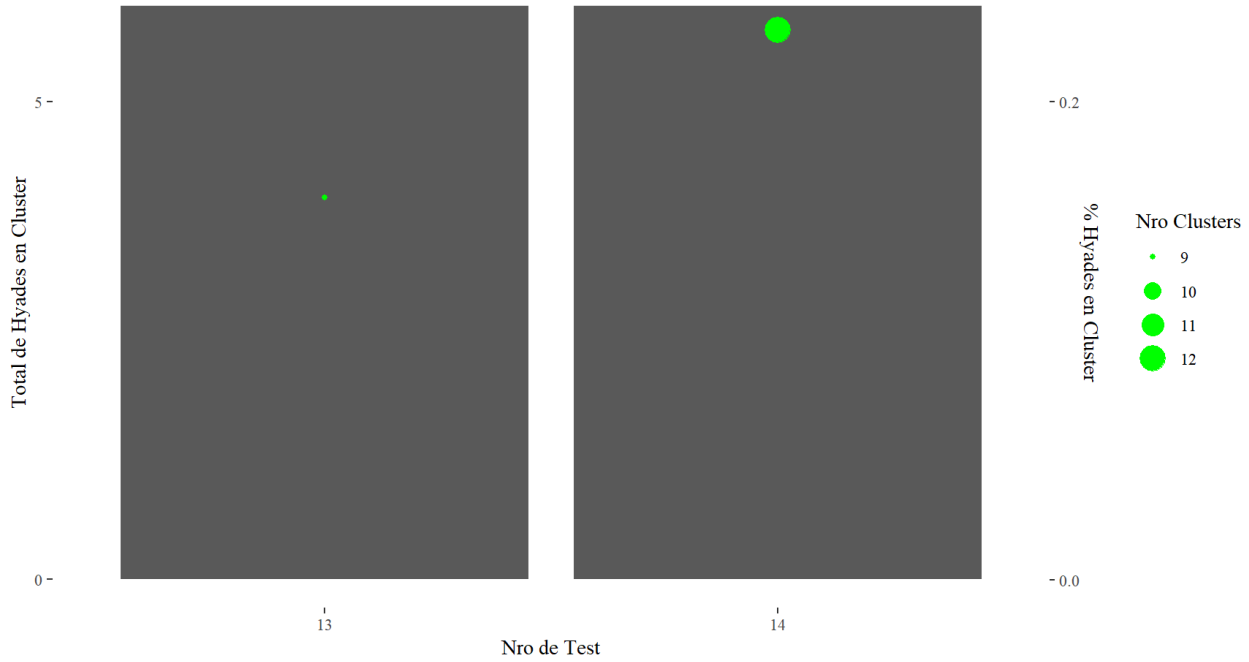








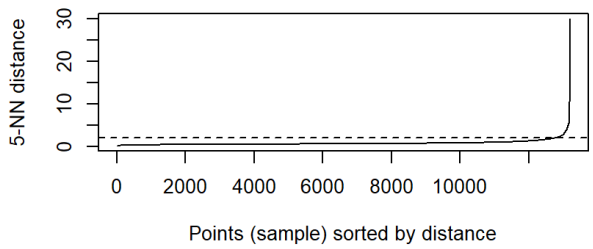
OBS: Basado en los gráficos anteriores se define el rango de *eps* entre 0.1 y 1, ejecutándose para todos los *minPts* entre 5 y 50

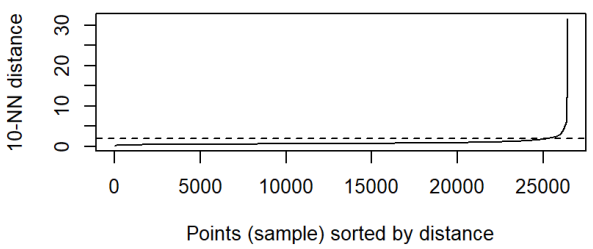


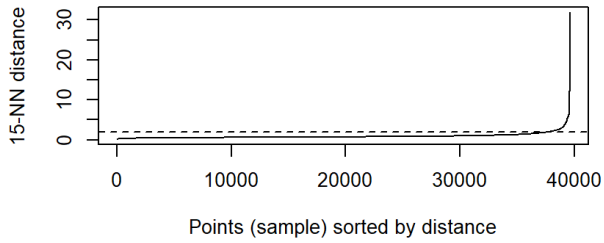
OBS: Nuevamente, no se visualizan clusters que presenten suficientes Hyades para tomar otras candidatas. En este caso los mejores clusters solo presentan 6 de las Hyades conocidas

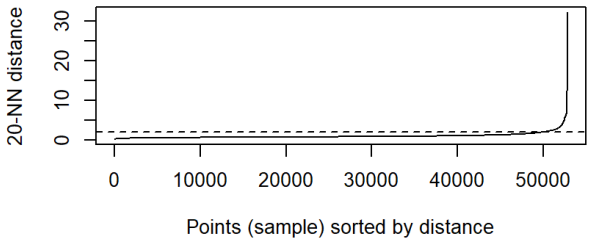
Experimento 3 - Full dataset

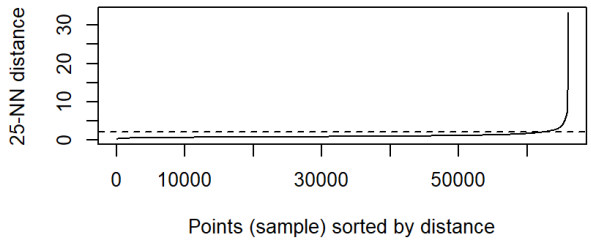
Identificación de EPS:

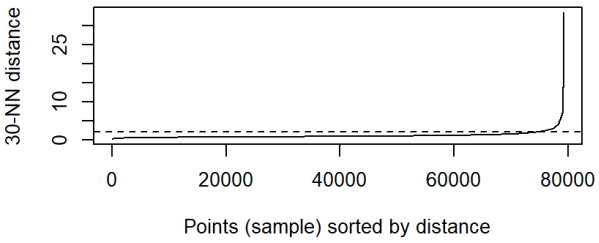


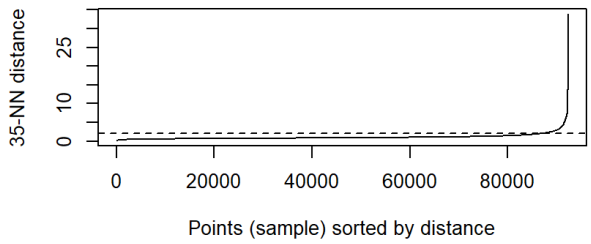


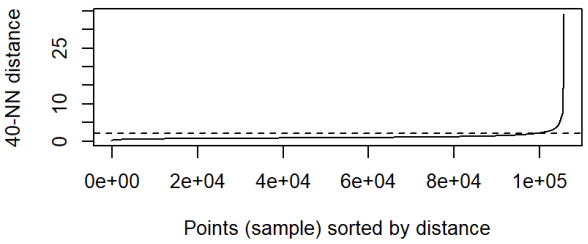


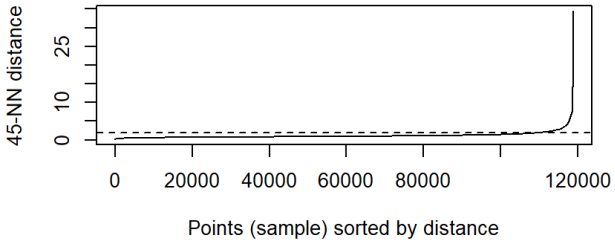


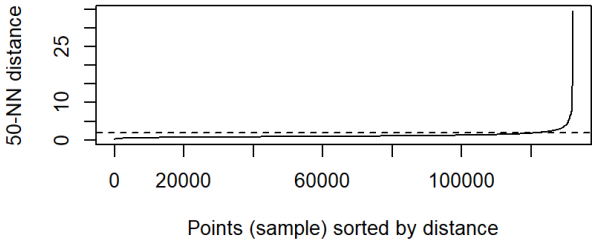




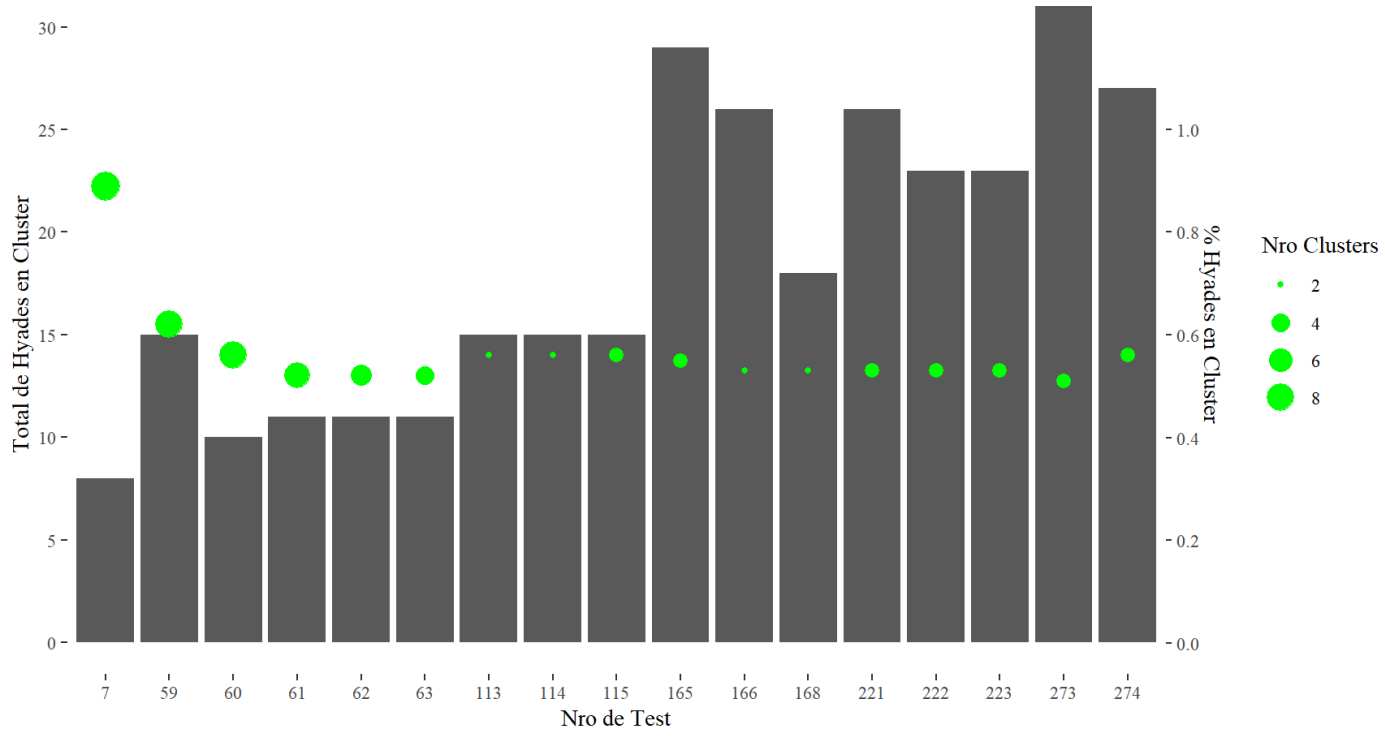








OBS: Basado en los gráficos anteriores se define el rango de *eps* entre 0.5 y 2, ejecutándose para todos los *minPts* entre 5 y 50



OBS: La ejecución 273 podría mostrar candidatas. Se repite esa en particular

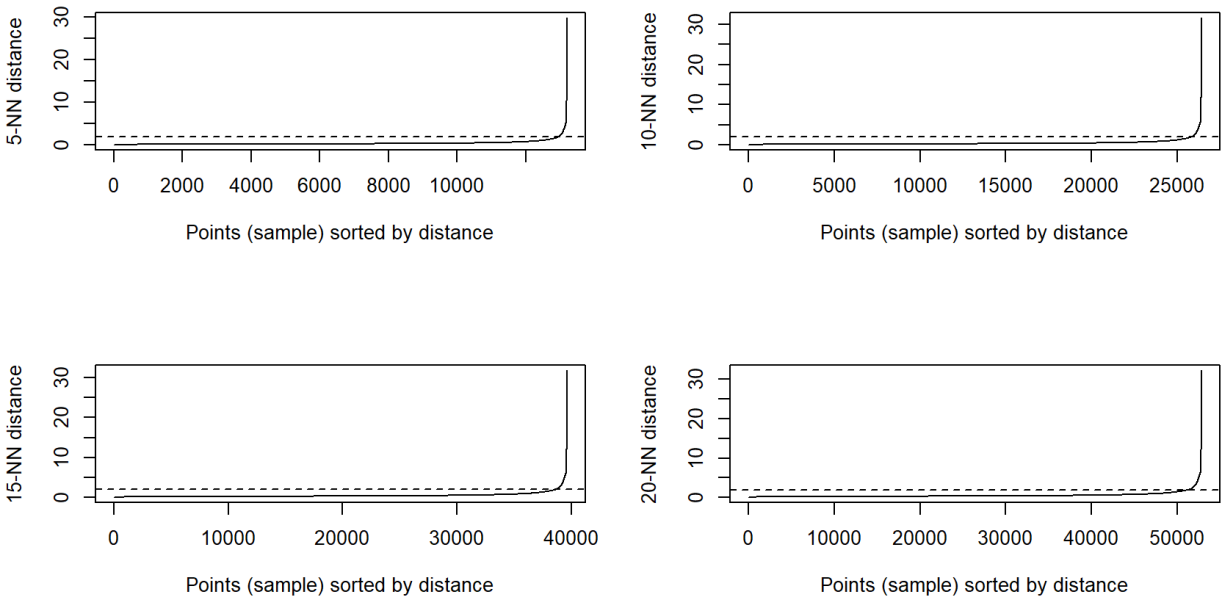
```
set.seed(123)
db <- fpc::dbscan(df3, eps = 1, MinPts = 47)
db
```

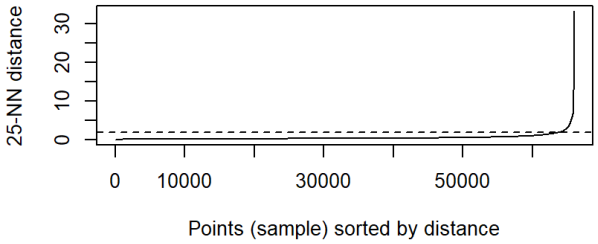
```
## dbscan Pts=2640 MinPts=47 eps=1
##          0    1  2  3
## border 938  708 81 58
## seed     0  848  4  3
## total  938 1556 85 61
```

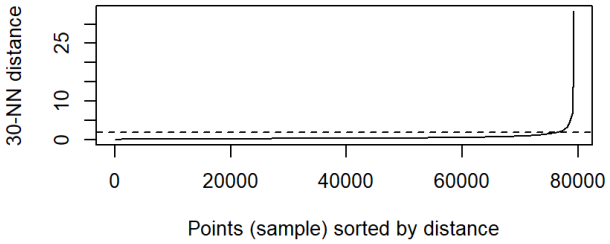
Indicado el cluster número 3 con un total de 61 estrellas. Se reservan estas como candidatas.

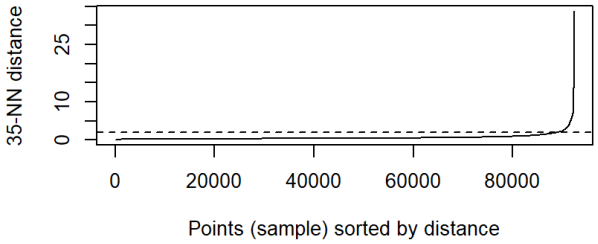
Experimento 4 - Sin distancias

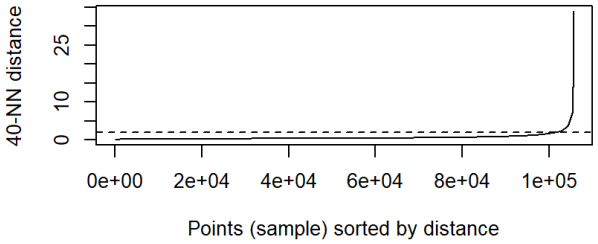
Identificación de EPS:

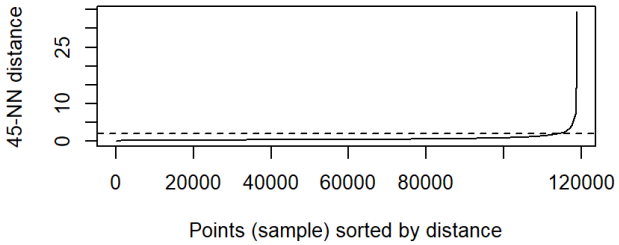


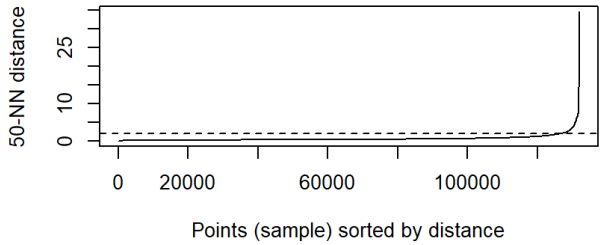




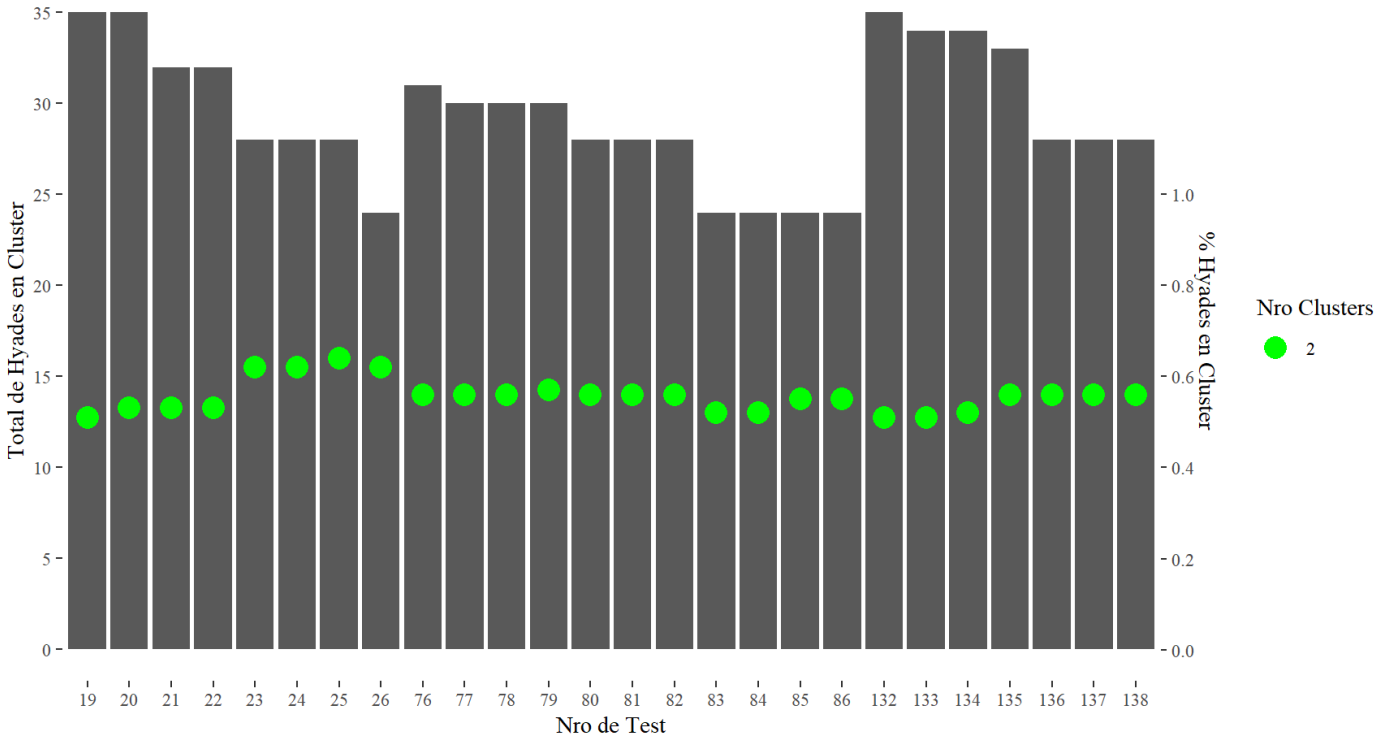








OBS: Basado en los gráficos anteriores se define el rango de *eps* entre 0.5 y 2, ejecutándose para todos los *minPts* entre 5 y 50



OBS: Resaltan las ejecuciones 19 y 132. Nuevamente, repetimos los experimentos para encontrar candidatas

```
## # A tibble: 2 x 9
##   cls   Hyades No_Hyades   eps minPts test  total ratio total_cls
##   <chr>  <dbl>     <dbl> <dbl>  <dbl> <fct> <dbl> <dbl>     <int>
## 1 2         35        34   0.5     23 19       69  0.51         2
## 2 2         35        34   0.7     44 132      69  0.51         2
```

```
set.seed(123)
db <- fpc::dbscan(df4, eps = .5, MinPts = 23)
db
```

```
## dbscan Pts=2640 MinPts=23 eps=0.5
##          0    1  2
## border 699  374 49
## seed     0 1498 20
## total  699 1872 69
```

```
set.seed(123)
db <- fpc::dbscan(df4, eps = .7, MinPts = 44)
db
```

```
## dbscan Pts=2640 MinPts=44 eps=0.7
##          0    1  2
## border 523  303 55
## seed     0 1745 14
## total  523 2048 69
```

Ambos casos, generan un cluster con 69 estrellas de las cuales 35 son las Hyades conocidas

A continuación, y con motivo de refinar la selección, se unen las salidas de los experimentos para intentar confirmas las estrellas presentes que respondieron de igual manera con relación a las Hyades

```
cand.fil<-cand %>% filter(Symbad_Hyades==FALSE)
cand2.fil<-cand2 %>% filter(Symbad_Hyades==FALSE)
cand3.fil<-cand3 %>% filter(Symbad_Hyades==FALSE)
cand.list <- cand.fil %>% inner_join(cand2.fil, by = "HIP") %>% inner_join(cand3.fil, by = "HIP") %>% select(HIP)
cand.list
```

```
##      HIP
## 1  19148
## 2  19786
## 3  20019
## 4  20056
## 5  20146
## 6  20237
## 7  20255
## 8  20284
## 9  20440
## 10 20553
## 11 20614
## 12 20686
## 13 20719
## 14 20826
## 15 21112
## 16 21137
## 17 21543
## 18 21654
## 19 22203
## 20 22422
## 21 22496
## 22 22505
## 23 22826
```

Clustering Difuso

Introducción

Dado el dataset de estrellas de Hiparcos y el set de Hyades identificadas previamente se procede a realizar experimentos de clustering difuso mediante la función *fanny* del paquete cluster. En los mismos se hará énfasis en los siguientes puntos:

- Selección de Variables del dataset origen Hiperparametrizacion de k y memb.exp
- Revisión de la presencia de Hyades en los clusters

Para este último punto se intentará identificar cluster relativamente pequeños que contengan la mayor cantidad de las Hyades previamente señaladas como tales.

Construcción de los experimentos

Los experimentos con el algoritmo se construyen a partir de los siguientes pasos:

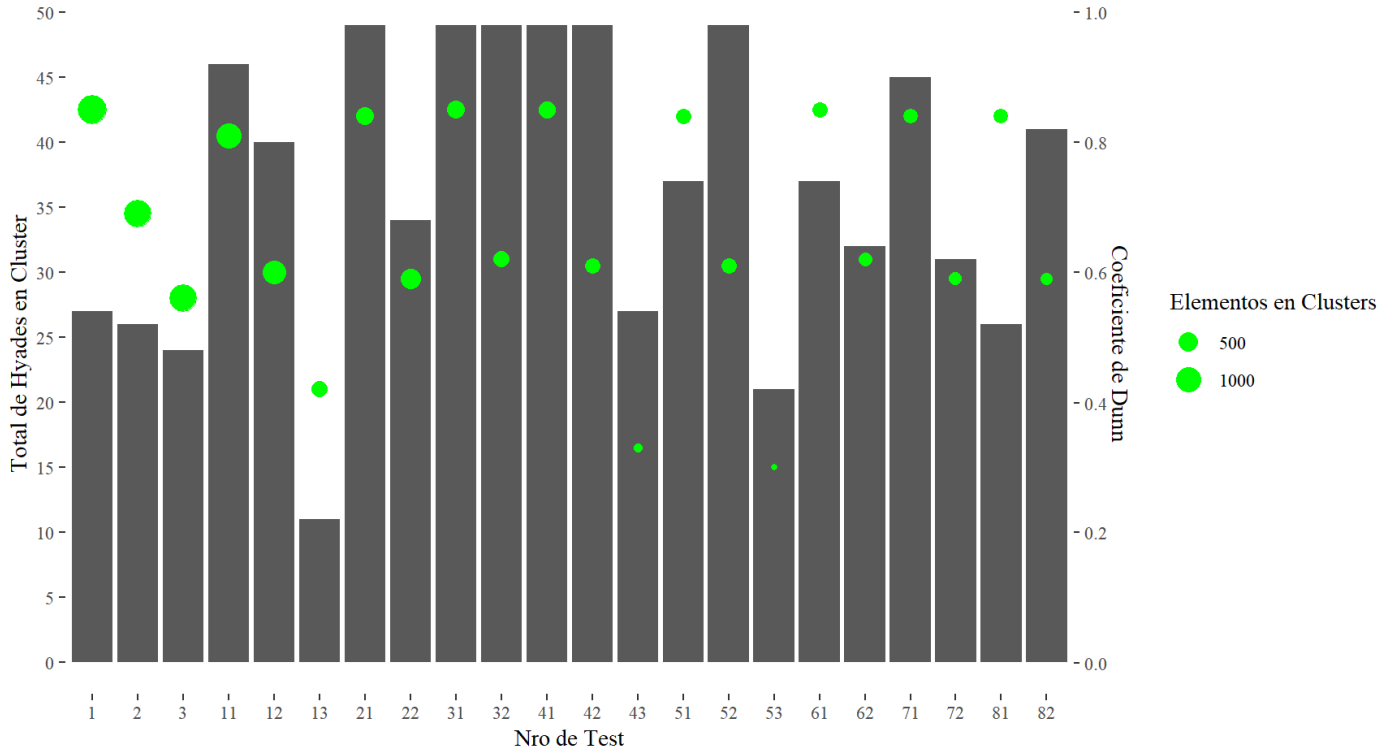
1. Ejecución y extracción de resultados identificando para cada cluster de cada ejecución la cantidad de Hyades (definidas)
2. ilustración de máxima cantidad de hyades en un cluster por ejecución, tamaño del cluster y coeficiente de Dunn asociado a la ejecución.
3. Conclusión y extracción de candidatas cuando corresponda.

Definiciones

- Se definió como rango de experimentación de k entre 2 y 10 ya que en las ejecuciones de los otros algoritmos se relativizó la mayor tasa de clasificación en valores de k pequeños (entre 2 y 3), con el fin de no sesgar el potencial del método se extiende el rango hasta 10 clusters.
- Arbitrariamente, se seleccionó el rango de memb.exp entre 1.1 y 2, dejando fuera el valor 1 ya que en ese caso el algoritmo se comporta como k-means que ya fue abordado previamente.
- A fin de identificar la pertenencia a los clusters según los resultados de la función de membresía se establece como representativo un valor mayor a 0.5.
- Se establecieron 2 grandes experimentos con selección de variables basados en los resultados precios de la ejecución de cluster por densidad, siendo:
    1. Todas las variables de Hiparcos
    2. Todas las variables, con excepción de las medidas de posición (RA y DE)
- Nuevamente, los experimentos se realizan con las variables normalizadas.

Para cada salida de la hyperparametrizacion, se identifica visualmente si alguno de los test es aceptable para producir candidatos

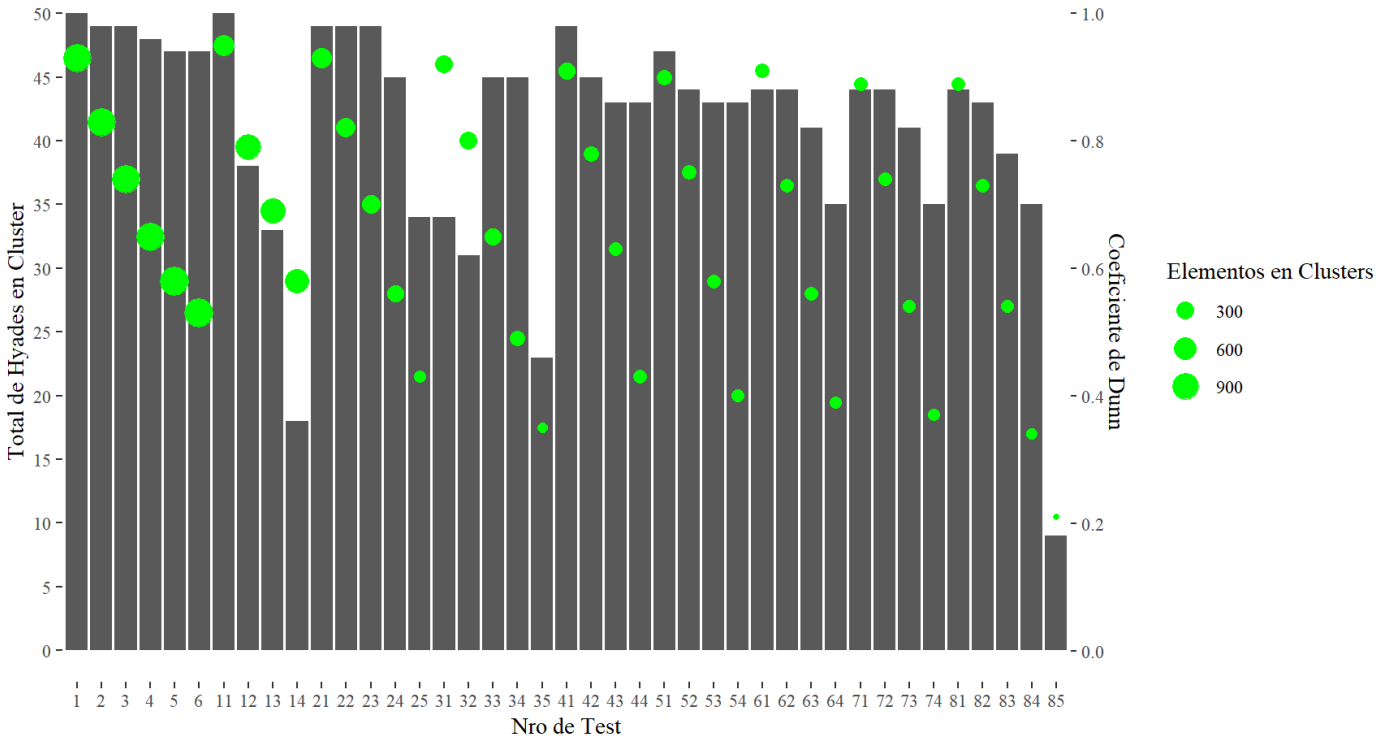
Experimento 1: Todas las Variables



OBS: Los experimentos 21, 31 y 41 presentan todas las Hyades en un cluster de tamaño pequeño en relación a los demás y un coeficiente de Dunn mayor a 0.8, los repetimos para extraer los clusters candidatos y se verifican las estrellas en común.

```
## # A tibble: 3 x 10
##   Cluster Hyades No_Hyades   k_n memb.exp test  total ratio dunn_coeff
##   <chr>    <dbl>     <dbl> <int>    <dbl> <fct> <dbl> <dbl>      <dbl>
## 1 3           49       325     4      1.1 21      374  0.13       0.84
## 2 4           49       323     5      1.1 31      372  0.13       0.85
## 3 3           49       311     6      1.1 41      360  0.14       0.85
## # ... with 1 more variable: normalized <dbl>
```

Experimento 2: Sin medidas de posición



OBS: Los experimentos 1, 11, 21 y 41 muestran resultados similares, variando solo en ellos la cantidad de clusters, revisamos 11, 21 y 41 ya que muestran clusters más pequeños. Así mismo, la única diferencia entre estas ejecuciones es el parámetro k, por lo que se puede deducir que, aunque aumento el paramento de segmentación, se encontró el mismo cluster en todos los casos.

```
## # A tibble: 4 x 10
##   Cluster Hyades No_Hyades   k_n memb.exp test  total ratio dunn_coeff
##   <chr>    <dbl>     <dbl> <int>    <dbl> <fct> <dbl> <dbl>      <dbl>
## 1 2           50      1037     2      1.1 1      1087  0.05       0.93
## 2 3           50       430     3      1.1 11      480  0.1        0.95
## 3 4           49       381     4      1.1 21      430  0.11       0.93
## 4 5           49       228     6      1.1 41      277  0.18       0.91
## # ... with 1 more variable: normalized <dbl>
```

A continuación, y con motivo de refinar la selección, se unen las salidas de los experimentos para intentar confirmar las estrellas presentes que respondieron de igual manera con relación a las Hyades, dando como resultado 143 estrellas candidatas.

```
cand.list.eucl <- cand.eucl %>%
  inner_join(cand.eucl2, by = "HIP")
cand.list.eucl
```

```
## # A tibble: 143 x 1
##      HIP
##    <dbl>
##  1 13589
##  2 13600
##  3 13601
##  4 13679
##  5 13982
##  6 14054
##  7 14098
##  8 14230
##  9 14258
## 10 14721
## # ... with 133 more rows
```

6. Bibliografía

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Consultado el 7 de abril de 2009.

Kaufman, Leonard & J. Rousseeuw, Peter. (1990). Partitioning Around Medoids (Program PAM). Finding Groups in Data: An Introduction to Cluster Analysis. 68 - 125. 10.1002/9780470316801.ch2.

Charrad, M., Ghazzali, ., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1 - 36. doi:http://dx.doi.org/10.18637/jss.v061.i06

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 1987, vol. 20, p. 53-65.

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). «A density-based algorithm for discovering clusters in large spatial databases with noise». En Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226-231. ISBN 1-57735-004-9.

BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984, vol. 10, no 2-3, p. 191-203.