APRENDIZAJE NO SUPERVISADO

FACULTAD DE ESTUDIOS ESTADÍSTICOS

PRÁCTICA 1: Análisis Clúster y Análisis Discriminante.

Pablo Galarón Mateo

ÍNDICE

1. **Introducción**
2. **Selección de la muestra y variables**
    2.1 Descripción del conjunto de datos
    2.2 División de la muestra
    2.3 Selección y reducción de variables
3. **Análisis clúster jerárquico**
    3.1 Análisis preliminar y representaciones iniciales
    3.2 Comparación de métodos jerárquicos
    3.3 Selección del número de clústeres
    3.4 Interpretación y representación de los clústeres
4. **Análisis clúster no jerárquico**
    4.1 FASTCLUS con semilla
    4.2 FASTCLUS con inicialización aleatoria
    4.3 Aplicación del método DRIFT
    4.4 Comparación con el análisis jerárquico
    4.5 Caracterización de los clústeres finales
5. **Análisis discriminante**
    5.1 Selección de variables discriminantes
    5.2 Comprobación de supuestos
    5.3 Estimación del modelo discriminante
    5.4 Evaluación del modelo y capacidad de generalización
6. **Conclusiones**

1 Introducción inicial y selección de variables.

Como en esta práctica se emplea el **mismo conjunto de datos que en la Práctica 1**, la selección de variables se realizará en base a los **factores obtenidos previamente** en el análisis factorial. En la práctica anterior se retuvieron **7 factores**, pero, con el fin de construir un modelo de clustering interpretable y evitar la creación de un número excesivo de grupos, en este trabajo se ha decidido comenzar utilizando **únicamente los dos primeros factores**, ya que son los que explican mayor proporción de la variabilidad total y presentan una interpretación más clara.

En la Práctica 1, el **Factor 1** quedó interpretado como **rendimiento académico**, al estar asociado principalmente con las calificaciones G1, G2 y G3, mientras que el **Factor 2** se relacionó con el **consumo de alcohol y la vida social** del estudiante (variables Walc, Dalc y goout). Por tanto, estos dos factores recogen las dimensiones más relevantes de la estructura del dataset, y constituyen una base adecuada para comenzar el estudio de clúster.

En esta fase inicial se analizará si los clusters formados mediante **F1 y F2** presentan una estructura clara y bien diferenciada. Si se observa que los grupos no quedan suficientemente separados o que el dendrograma no es estable, se considerará la posibilidad de refinar el modelo aplicando el procedimiento **ACECLUS**, que genera variables canónicas específicamente diseñadas para la formación de clusters. Esto permitirá comparar ambas aproximaciones y seleccionar la que proporcione la agrupación más coherente.

A continuación se recuerda brevemente qué representa cada uno de los factores utilizados:

Nombre	Descripción
Factor 1	Representa el rendimiento académico general, altamente asociado a las calificaciones del estudiante (G1, G2 y G3) y a variables relacionadas con el estudio.
Factor 2	Representa la dimensión de vida social y consumo de alcohol, destacando las variables Walc, Dalc y goout.

Así, el análisis de clúster se inicia utilizando estas dos dimensiones fundamentales del comportamiento estudiantil. A lo largo del trabajo, se evaluará si esta elección es suficiente o si resulta necesario complementar el modelo con otras transformaciones (como ACECLUS), en función de los resultados.

2 Selección de la muestra.

Para esta práctica partimos del mismo conjunto de datos ya depurado en la Práctica 1. Con el fin de construir los modelos de clustering y, posteriormente, evaluar su estabilidad, se decidió dividir el dataset en una proporción **70%–30%**.

El **70%** se utilizará para generar los clusters, mientras que el **30% restante** servirá como muestra de validación. Esta partición resulta adecuada porque contamos con un número suficiente de observaciones y nos permite comprobar si, al introducir datos nuevos no utilizados en la fase de entrenamiento, se obtiene la misma estructura de agrupamiento. De este modo podremos evaluar si los clusters formados son consistentes y generalizables.

PROGRAMACIÓN SAS-SELECCIÓN DE MUESTRA

```
/* División del dataset limpio en 70%(SAMPLE) y 30% (RESTANTE) */

PROC SURVEYSELECT DATA=practic2.Afact_QUARTIMAX OUT=SAMPLE_RAW
     METHOD=SRS          /* Muestreo aleatorio simple */
     SAMPRATE=0.7        /* 70% para clúster */
     SEED=12345
     OUTALL noprint;
RUN;

/* 70%  */
DATA practic2.SAMPLE;
    SET SAMPLE_RAW;
    IF Selected = 1;
RUN;

/* 30% */
DATA practic2.RESTANTE;
    SET SAMPLE_RAW;
    IF Selected = 0;
RUN;
```

Realizo un proc print sobre el conjunto de factores obtenido en la práctica anterior para revisar su estructura y observar un poco.

Obs	Medu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G3	ID	G1	G2	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
1	4	2	2	0	4	3	4	1	1	3	6	6	1	5	6	-1.85113	-0.54563	0.15283	-0.88627	-0.47090	-1.15445	-0.13059
2	1	1	2	0	5	3	3	1	1	3	4	6	2	5	5	-2.06515	-1.08025	0.17298	-0.15582	-1.00213	-0.45923	-0.04051
3	1	1	2	3	4	3	2	2	3	3	10	10	3	7	8	-1.13934	-0.49572	0.00225	1.06204	2.82090	0.87045	0.61790
4	4	1	3	0	3	2	2	1	1	5	2	15	4	15	14	0.89190	-0.48246	-1.44743	-0.98042	0.05076	1.05495	1.15988
5	3	1	2	0	4	3	2	1	2	5	4	10	5	6	10	-0.94752	-0.49748	-0.65874	-0.92740	-0.39303	0.97771	-0.12495
6	4	1	2	0	5	4	2	1	2	5	10	15	6	15	15	1.20140	-0.75064	0.61270	-1.18915	0.59013	1.26606	-0.74314

3 Análisis Jerárquico.

En esta parte del trabajo comenzaré analizando si los **factores** obtenidos en la Práctica 1 son adecuados para formar grupos diferenciados. Para ello utilizaré **Factor 1** (relacionado con el rendimiento académico) y **Factor 2** (asociado a la vida social y el consumo de alcohol). Mi objetivo inicial es comprobar si únicamente con estos dos factores es posible obtener una estructura de clusters coherente. En caso de que no consiga una separación clara, más adelante aplicaré el procedimiento **ACECLUS** para generar variables canónicas que puedan mejorar la calidad del agrupamiento.

Como punto de partida utilizaré el método de Ward, ya que es uno de los más eficaces para formar clusters compactos y con menor variabilidad interna. Más adelante compararé estos resultados con otros métodos (enlace promedio, centroides, enlace completo, etc.) para evaluar si alguno ofrece una mejora. Para determinar cuál es el número óptimo de clusters y qué método funciona mejor, evaluaremos varias métricas que proporciona PROC CLUSTER:

● **R^2 (Proporción de variabilidad explicada)**

Indica cuánta variabilidad total explican los clusters formados.

- Quiero maximizarlo, porque valores altos significan que los grupos separan bien a las observaciones.
- Si el R^2 apenas aumenta al añadir más clusters, no compensa seguir subdividiendo.

● **Pseudo-F**

Compara la variabilidad entre clusters con la variabilidad dentro de los clusters.

- Quiero valores altos.
- Máximos relativos suelen indicar buenas opciones para seleccionar el número de clusters, ya que en esos puntos la separación es mayor.

● **Pseudo-T²**

Este índice evalúa si dos clusters que se acaban de fusionar son realmente parecidos.

- En este caso, me fijaré en los mínimos, porque un valor bajo indica que los clusters fusionados no eran muy distintos, y por tanto la fusión es razonable.
- No busco máximos, ya que valores muy altos significan lo contrario: que la fusión junta grupos demasiado diferentes y empeora el modelo. Por eso, me interesa localizar descensos importantes del Pseudo-T², ya que suelen marcar puntos donde la estructura del clustering es más estable.

● **CCC (Cubic Clustering Criterion)**

Ayuda a determinar si existe una estructura clara de agrupamiento.

- Este índice solo es interpretable cuando CCC > 3.
- Si supera ese valor, cuanto más alto mejor, ya que sugiere que el número de clusters es estadísticamente justificable.
- Si el CCC es menor de 3, puedo interpretarlo como que los datos están muy dispersos o que no siguen una estructura de clúster bien definida.

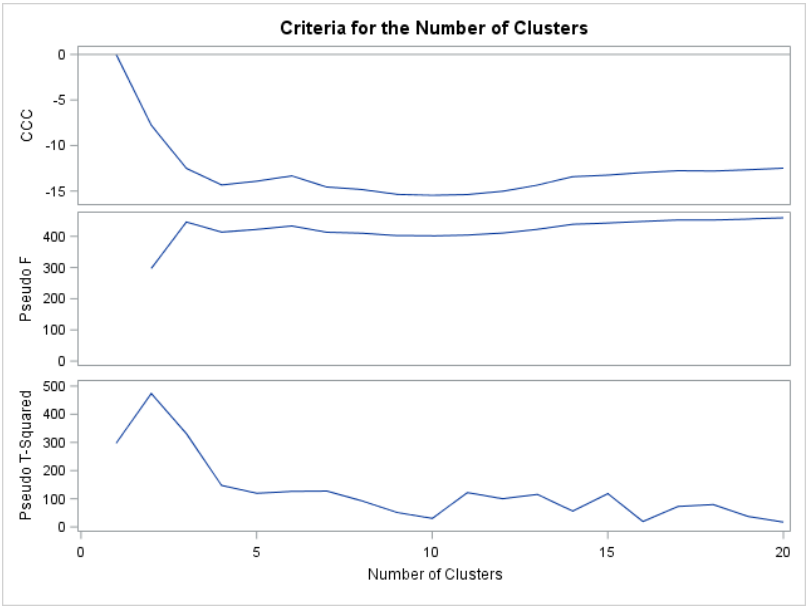
Con el objetivo de evaluar qué representación resulta más adecuada para la construcción de clústeres, se analizó la capacidad discriminante de distintas combinaciones de variables latentes: **Factor 1 y Factor 2**, **Factor 1 en solitario**, y posteriormente las **variables canónicas Can1 y Can2**, así como Can1 de forma individual.

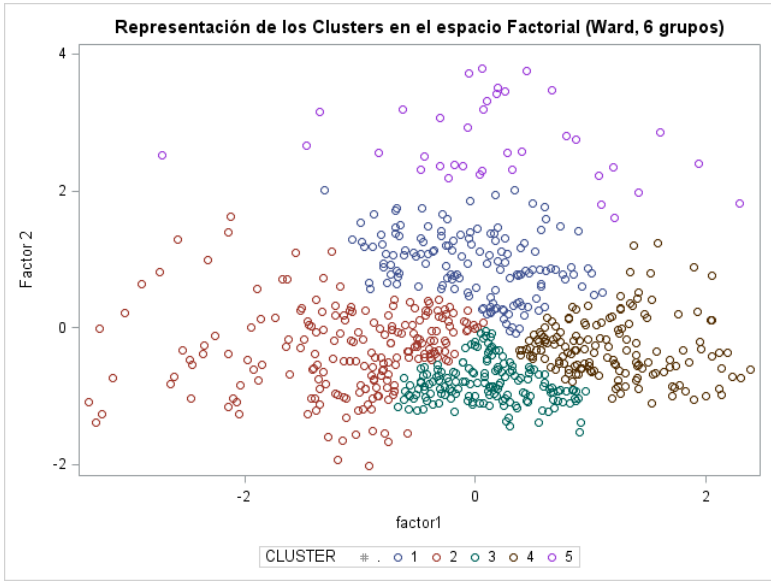
3.1 Análisis preliminar con Factor 1 y Factor 2.

PROGRAMACIÓN SAS-FACTOR1 Y FACTOR2

```
/*----------------------Observo que pasa con los primeros factores---------------------
*/
PROC CLUSTER DATA=practic2.SAMPLE
method=ward
std nonorm pseudo RSQUARE ccc /*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_ward_factores;
var factor1 factor2;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu;
id id;
RUN;
```

```
proc tree data=salida_ward_factores ncl=5 out=clusters_factores noprint;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor1 factor2;
id id;
run;

proc sgplot data=clusters_factores;
    scatter x=factor1 y=factor2/ group=cluster ;
    xaxis label="factor1";
    yaxis label="Factor 2";
    title "Representación de los Clusters en el espacio Factorial (Ward, 6 grupos)";
run;
```





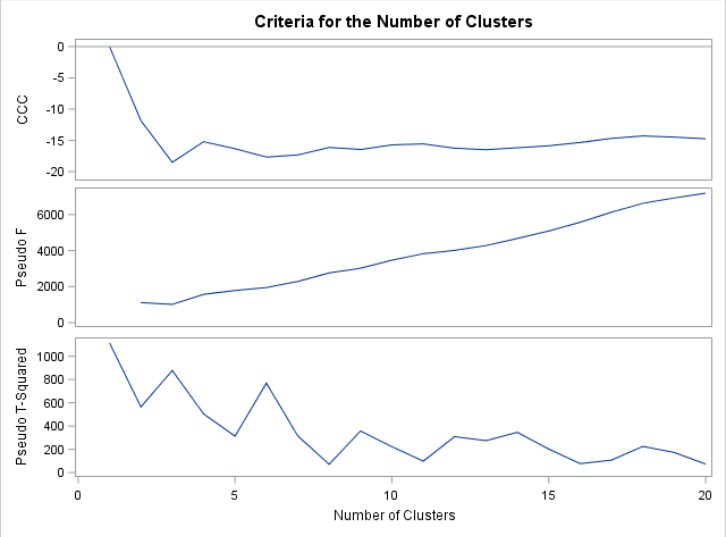
En la representación de los clústeres en el plano formado por **Factor 1 y Factor 2**, se observa que la estructura de los grupos **no es mala en términos generales**, ya que algunos clústeres aparecen relativamente diferenciados, especialmente a lo largo del eje del Factor 1.

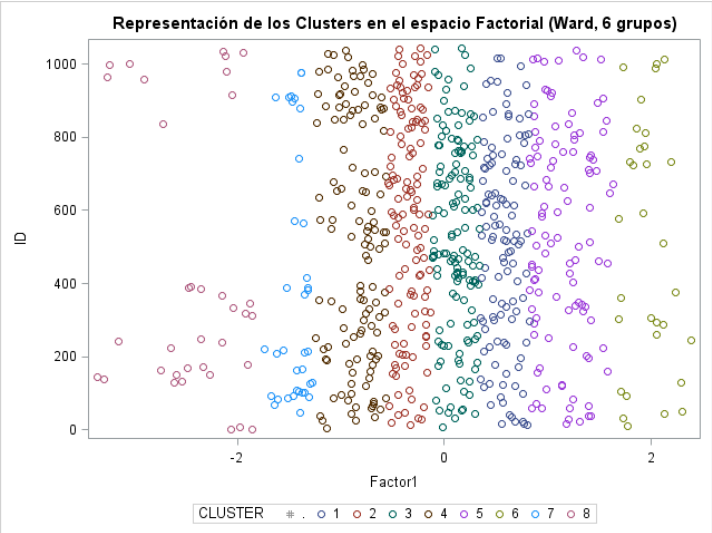
No obstante, el **Factor 2 presenta una menor capacidad discriminante**, ya que existe un solapamiento considerable entre clústeres en el eje vertical. Esto indica que, aunque el segundo factor aporta cierta información adicional, **no contribuye de forma clara a la separación entre grupos**, sino que introduce dispersión adicional en la representación. Además no coincide ningún máximo en la F con un mínimo en la T.

3.2 Análisis preliminar con factor 1.

PROGRAMACIÓN SAS-FACTOR1

```
/*-----------------------Voy a hacerlo solo para factor1--------------------------------
-------------------------------------------------------
*/

PROC CLUSTER DATA=practic2.SAMPLE
method=ward
std nonorm pseudo RSQUARE ccc /*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_ward_factor1;
var factor1;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor2;
id id;
RUN;

proc tree data=salida_ward_factor1 ncl=8 out=clusters_factor1 noprint;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor1 factor2;
id id;
run;


proc sgplot data=clusters_factor1;
    scatter x=factor1 y=id/ group=cluster ;
run;
```





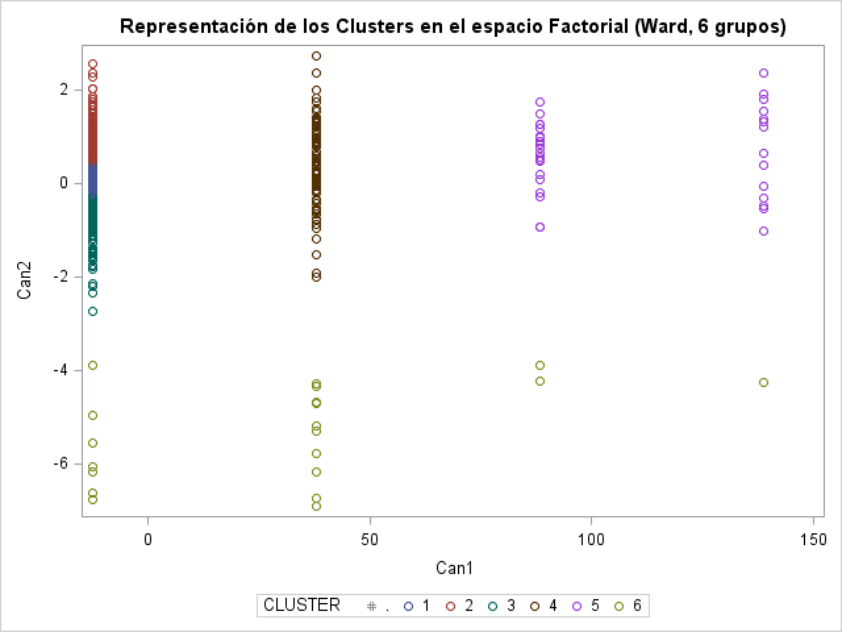
Al representar los clústeres empleando exclusivamente **Factor 1**, se observa una **mejora notable en la separación entre grupos**. Los clústeres aparecen claramente ordenados a lo largo de este eje, con fronteras más definidas y menor solapamiento entre observaciones pertenecientes a distintos grupos.

Este resultado indica que **Factor 1 concentra la mayor parte de la información relevante para la discriminación**, mientras que el Factor 2 aporta poco valor añadido. Desde un punto de vista interpretativo y estadístico, esta solución resulta **más parsimoniosa y más clara**, cumpliendo mejor el objetivo del análisis clúster.

3.3 Análisis preliminar can1 y can2.

PROGRAMACIÓN SAS-CAN1 Y CAN2

```
/*-----------------------Voy a hacer aceclus para obtener can1 y can2-------------------
------------------------------------------------------------------------
*/

PROC ACECLUS DATA=practic2.SAMPLE
             OUT=practic2.SAMPLE_TRANSFORMED
             P=.03;
    VAR traveltime studytime failures famrel freetime goout dalc walc health absences G1
G2 G3 Medu factor1 factor2;
run;

PROC CLUSTER DATA=practic2.SAMPLE_transformed /*datos no estandarizados*/
method=ward
std nonorm pseudo RSQUARE ccc/*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_ward_cans;
var  Can1 Can2; /*especificar las variables transformadas en esta forma*/
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor1 factor2;
id id;
RUN;

proc tree data=salida_ward_cans ncl=6 out=clusters_cans noprint;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu can1 can2 factor1 factor2;
id id;
run;

proc sgplot data= clusters_cans;
scatter x=can1 y=can2 / group=cluster;
run;
```



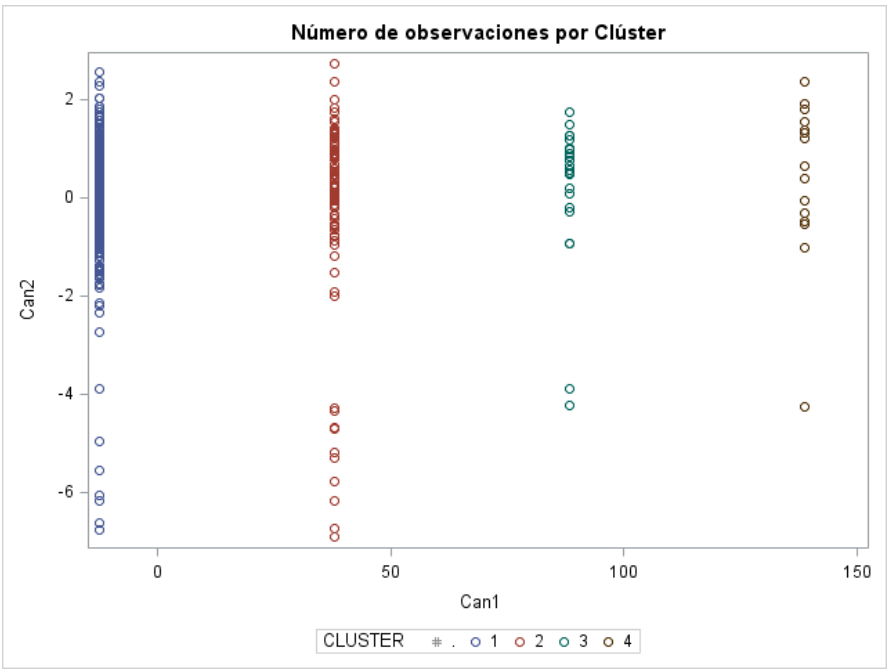
A continuación, se analizó la representación de los clústeres en el espacio formado por las variables canónicas **Can1 y Can2**. En este caso, la discriminación entre clústeres es **razonablemente buena**, especialmente en Can1, mientras que Can2 vuelve a mostrar una capacidad discriminante más limitada.

Aunque esta representación no es incorrecta, se observa que **la mayor parte de la separación se produce nuevamente en una única dimensión**, lo que sugiere que el segundo eje canónico no resulta esencial para identificar la estructura de los grupos.

3.4 Análisis preliminar con can1.

PROGRAMACIÓN SAS-CAN1

```
/*-----------------------Voy a probar solo con can 1 a ver si mejora algo----------------
--------------------------------------------------------------------------------
*/

PROC CLUSTER DATA=practic2.SAMPLE_transformed /*datos no estandarizados*/
method=ward
std nonorm pseudo RSQUARE ccc/*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_ward_cans1;
var  Can1; /*especificar las variables transformadas en esta forma*/
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor1 factor2 can2;
id id;
RUN;

proc tree data=salida_ward_cans1 ncl=4 out=clusters_can1 noprint;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu can1 can2 factor1 factor2;
id id;
run;

proc sgplot data= clusters_can1;
scatter x=can1 y=can2 / group=cluster;
run;
```



Cuando se representa el clúster únicamente en función de **Can1**, la separación entre grupos parece **prácticamente perfecta**, con clústeres muy bien definidos y sin apenas solapamiento.

Sin embargo, este resultado resulta **excesivamente bueno**, lo que lleva a analizar con mayor detalle la composición de los clústeres mediante un **PROC FREQ**. Al examinar la distribución de individuos por clúster, se observa que **uno de los clústeres concentra la inmensa mayoría de las observaciones**, mientras que el resto contienen un número muy reducido de individuos.

Esto indica que, aunque visualmente Can1 discrimina de forma extrema, **no se está obteniendo una partición equilibrada ni informativa**, sino un agrupamiento sesgado en el que un único clúster aglutina casi toda la muestra. Por tanto, esta solución **no representa adecuadamente la estructura interna de los datos** y no cumple el objetivo del análisis clúster.

CLUSTER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	588	82.58	588	82.58
2	86	12.08	674	94.66
3	23	3.23	697	97.89
4	15	2.11	712	100.00
Frequency Missing = 1				

3.5 Conclusión final.

Tras analizar todas las representaciones, se concluye que:

- La combinación **Factor 1 y Factor 2** no ofrece una estructura de clúster claramente definida debido a la escasa capacidad discriminante del segundo factor.
- El uso exclusivo de **Can1**, aunque visualmente muy separador, genera un clúster dominante que concentra la mayoría de las observaciones, lo que invalida la solución desde un punto de vista práctico.
- La solución más adecuada es **trabajar únicamente con Factor 1**, ya que proporciona una separación clara entre clústeres, una estructura equilibrada y una interpretación más coherente de los grupos.
- En consecuencia, el análisis clúster definitivo se realizará **utilizando exclusivamente Factor 1**, al considerarse la opción más robusta. Vamos a probar con otros métodos para ver si nos mejora algo.

3.6 Análisis con otros métodos.

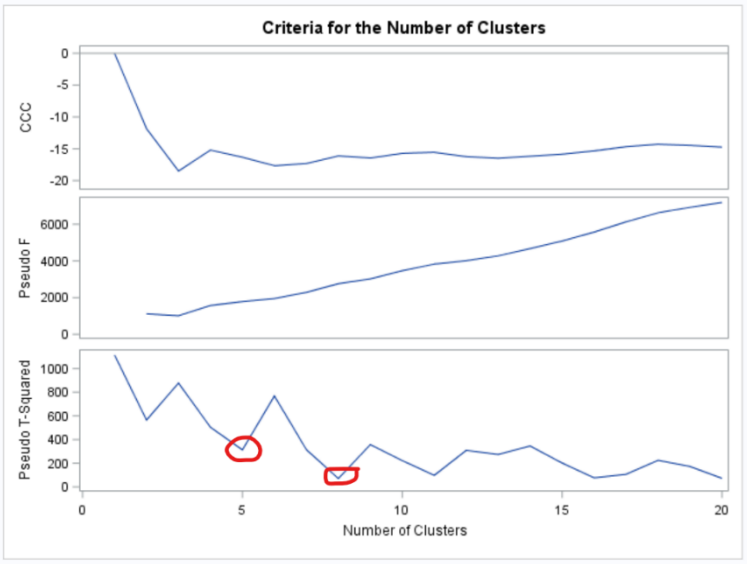
Distancia Ward(Mínima Varianza).

PROGRAMACIÓN SAS-WARD-INICIAL(ESTÁ AHORA REPETIDO)

```
PROC CLUSTER DATA=practic2.SAMPLE
method=ward
std nonorm pseudo RSQUARE ccc /*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_ward_factor1;
var factor1;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor2;
id id;
RUN;
```

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Between Cluster Sum of Squares	Tie
20	CL46	CL32	36	0.0005	.995	.998	-15	7190	72.7	0.3382	
19	CL34	CL53	68	0.0005	.994	.997	-14	6924	173	0.3483	
18	CL45	CL31	95	0.0006	.994	.997	-14	6625	225	0.4231	
17	CL48	CL23	55	0.0009	.993	.997	-15	6133	107	0.6458	
16	CL28	CL29	31	0.0012	.992	.996	-15	5575	76.9	0.8686	
15	CL37	CL25	72	0.0014	.990	.996	-16	5088	202	1.021	
14	CL22	CL35	125	0.0017	.989	.995	-16	4676	346	1.1817	
13	CL19	CL33	114	0.0021	.987	.994	-16	4277	274	1.481	
12	CL36	CL18	139	0.0022	.984	.993	-16	4014	309	1.542	
11	CL27	CL24	28	0.0024	.982	.992	-16	3828	97.5	1.6897	
10	CL15	CL26	125	0.0040	.978	.990	-16	3470	223	2.8485	
9	CL17	CL21	107	0.0063	.972	.988	-16	3021	358	4.4618	
8	CL11	CL30	35	0.0069	.965	.985	-16	2758	70.8	4.9176	
7	CL10	CL20	161	0.0137	.951	.980	-17	2287	315	9.7337	
6	CL12	CL13	253	0.0186	.933	.973	-18	1953	769	13.204	
5	CL16	CL9	138	0.0225	.910	.961	-16	1788	313	16.017	
4	CL7	CL14	286	0.0405	.869	.938	-15	1572	504	28.83	
3	CL6	CL5	391	0.1285	.741	.890	-19	1014	877	91.359	
2	CL8	CL4	321	0.1303	.611	.751	-12	1114	565	92.612	
1	CL2	CL3	712	0.6107	.000	.000	0.00	.	1114	434.23	

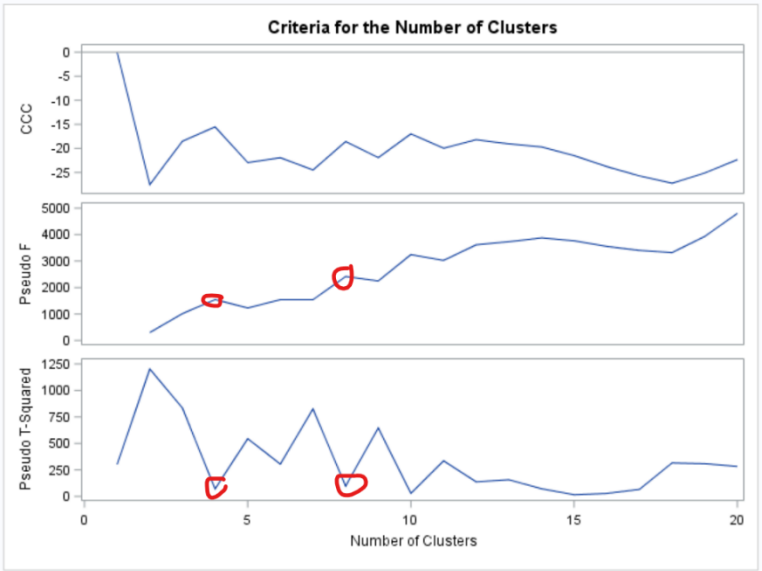


Enlace Promedio (Distancia Media).

PROGRAMACIÓN SAS-AVERAGE

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	RMS Distance	Tie
20	CL32	CL41	104	0.0016	.992	.998	-22	4800	282	0.2268	
19	CL50	CL28	139	0.0022	.990	.997	-25	3931	309	0.244	
18	CL29	CL36	141	0.0024	.988	.997	-27	3320	315	0.2513	
17	CL30	CL38	36	0.0005	.987	.997	-26	3402	66.1	0.2698	
16	CL35	CL26	13	0.0003	.987	.996	-24	3554	27.5	0.276	
15	CL40	CL23	7	0.0002	.987	.996	-21	3765	14.6	0.2805	
14	CL48	CL25	26	0.0006	.986	.995	-20	3876	72.2	0.2843	
13	CL27	CL24	66	0.0017	.985	.994	-19	3730	157	0.2979	
12	CL33	CL17	62	0.0019	.983	.993	-18	3615	137	0.3291	
11	CL34	CL20	155	0.0053	.977	.992	-20	3026	336	0.3604	
10	CL14	CL31	31	0.0008	.977	.990	-17	3244	28.4	0.4097	
9	CL18	CL21	224	0.0141	.962	.988	-22	2249	646	0.4705	
8	CL22	CL16	28	0.0024	.960	.985	-19	2415	97.5	0.5249	
7	CL19	CL11	294	0.0308	.929	.980	-24	1542	825	0.5925	
6	CL10	CL13	97	0.0131	.916	.973	-22	1541	305	0.7119	
5	CL12	CL9	286	0.0417	.874	.961	-23	1230	544	0.8379	
4	CL8	CL15	35	0.0069	.867	.938	-16	1545	70.8	0.9881	
3	CL7	CL6	391	0.1265	.741	.890	-19	1014	834	1.2084	
2	CL5	CL3	677	0.4434	.298	.751	-28	301	1203	1.5504	
1	CL4	CL2	712	0.2975	.000	.000	0.00	.	301	2.7	



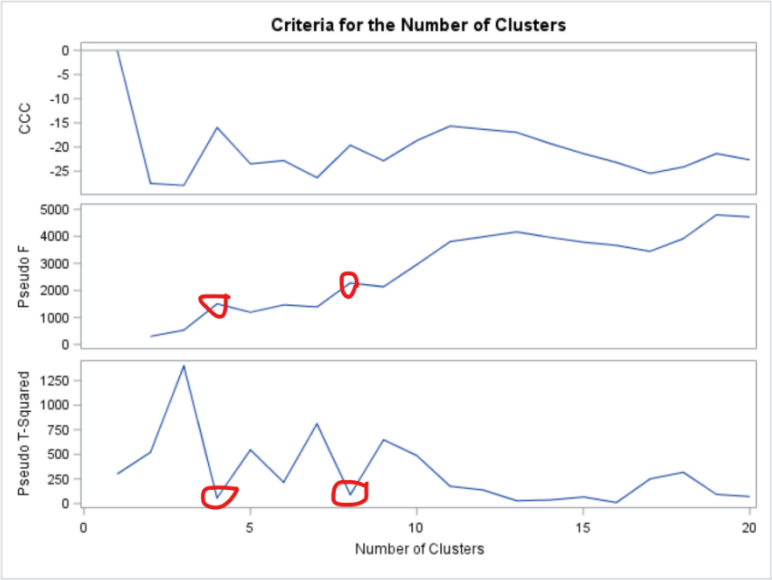
Distancia entre centroides (Vector Medio).

PROGRAMACIÓN SAS-CENTROID

```
PROC CLUSTER DATA=practic2.SAMPLE
method= centroid
std nonorm pseudo RSQUARE ccc /*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_centroid_factor1;
var factor1;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor2;
id id;
RUN;
```

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Centroid Distance	Tie
20	CL54	CL51	15	0.0002	.992	.998	-23	4717	70.8	0.1969	
19	CL52	CL48	22	0.0003	.992	.997	-21	4799	92.6	0.197	
18	CL29	CL34	139	0.0024	.990	.997	-24	3917	318	0.2268	
17	CL58	CL28	141	0.0021	.988	.997	-25	3449	250	0.2332	
16	CL32	959	6	0.0001	.987	.996	-23	3664	10.5	0.2384	
15	CL31	CL36	36	0.0005	.987	.996	-21	3790	66.1	0.256	
14	CL26	CL30	21	0.0004	.987	.995	-19	3964	36.4	0.2731	
13	CL16	CL25	14	0.0004	.986	.994	-17	4168	28.0	0.2943	
12	CL35	CL15	62	0.0019	.984	.993	-16	3988	137	0.3009	
11	CL19	CL27	65	0.0024	.982	.992	-16	3808	175	0.3404	
10	CL21	CL24	166	0.0076	.974	.990	-19	2961	488	0.3623	
9	CL18	CL22	222	0.0138	.961	.988	-23	2137	648	0.435	
8	CL20	CL13	29	0.0028	.958	.985	-20	2279	89.0	0.5208	
7	CL17	CL10	307	0.0357	.922	.980	-26	1389	811	0.5773	
6	CL11	CL14	86	0.0098	.912	.973	-23	1466	216	0.6642	
5	CL12	CL9	284	0.0412	.871	.961	-23	1193	546	0.7775	
4	CL8	CL23	35	0.0065	.864	.938	-16	1506	57.8	0.9607	
3	CL5	CL7	591	0.2651	.599	.890	-28	530	1398	1.1303	
2	CL3	CL6	677	0.3019	.298	.751	-28	301	522	1.6908	
1	CL4	CL2	712	0.2975	.000	.000	0.00	.	301	2.5213	



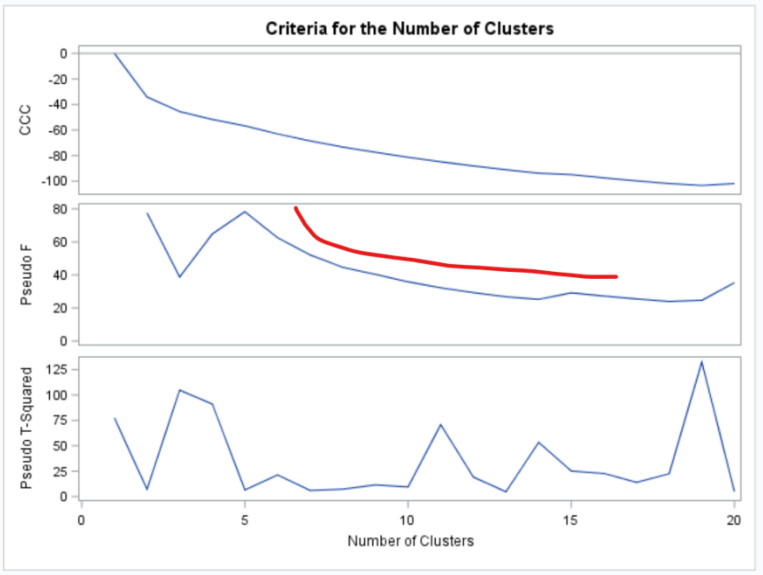
Enlace Simple (Vecino más cercano).

PROGRAMACIÓN SAS-SINGLE

```
PROC CLUSTER DATA=practic2.SAMPLE
method= single
std nonorm pseudo RSQUARE ccc /*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_single_factor1;
var factor1;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor2;
id id;
RUN;
```

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Min Dist	Tie
20	CL35	132	4	0.0000	.492	.998	-102	35.3	5.0	0.0455	
19	CL25	CL21	662	0.1018	.390	.997	-103	24.7	132	0.0482	
18	CL19	CL26	667	0.0206	.370	.997	-102	24.0	22.5	0.0497	
17	CL156	CL22	4	0.0000	.370	.997	-100	25.5	14.0	0.0565	
16	CL20	CL90	7	0.0000	.370	.996	-97	27.2	22.7	0.0575	
15	CL103	CL17	8	0.0000	.370	.996	-95	29.2	25.2	0.0589	
14	CL18	CL15	675	0.0500	.320	.995	-94	25.2	53.4	0.0591	
13	CL14	221	676	0.0047	.315	.994	-91	26.8	4.7	0.0635	
12	149	CL57	4	0.0000	.315	.993	-88	29.3	19.2	0.0635	
11	CL36	CL27	15	0.0002	.315	.992	-85	32.2	70.8	0.0652	
10	CL34	145	4	0.0000	.315	.990	-81	35.8	9.5	0.0657	
9	CL16	CL61	9	0.0001	.315	.988	-77	40.4	11.5	0.0779	
8	CL13	246	677	0.0074	.307	.985	-73	44.6	7.3	0.0863	
7	CL10	243	5	0.0000	.307	.980	-69	52.1	6.0	0.0928	
6	CL9	CL12	13	0.0003	.307	.973	-63	62.6	21.3	0.0949	
5	CL7	1001	6	0.0001	.307	.961	-57	78.3	6.5	0.1026	
4	CL11	CL8	692	0.0915	.215	.938	-52	64.8	91.1	0.1105	
3	CL4	CL6	705	0.1169	.099	.890	-46	38.8	105	0.1142	
2	CL5	959	7	0.0001	.098	.751	-34	77.6	7.2	0.1451	
1	CL3	CL2	712	0.0985	.000	.000	0.00	.	77.6	0.1586	



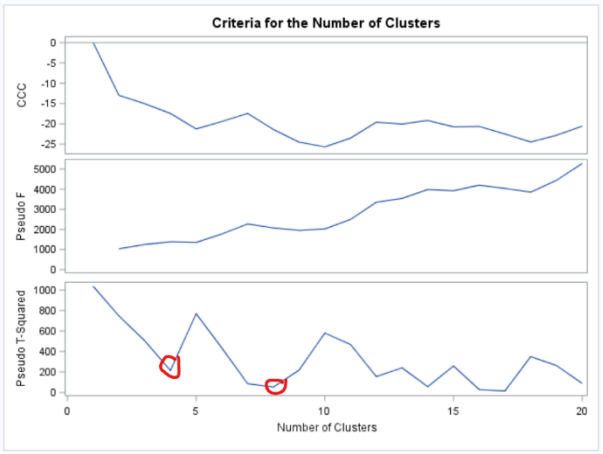
Enlace completo (Vecino más lejano).

PROGRAMACIÓN SAS-COMPLETE

```
PROC CLUSTER DATA=practic2.SAMPLE
method= complete
std nonorm pseudo RSQUARE ccc /*incluir "std" para estandarizar los datos*/
print=20 outtree=salida_complete_factor1;
var factor1;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu factor2;
id id;
RUN;
```

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Maximum Distance	Tie
20	CL41	CL35	40	0.0004	.993	.998	-21	5275	88.7	0.3853	
19	CL29	CL45	115	0.0017	.991	.997	-23	4442	263	0.4001	
18	CL33	CL39	113	0.0019	.990	.997	-24	3853	350	0.4074	
17	CL56	CL28	7	0.0002	.989	.997	-23	4037	14.6	0.4585	
16	CL42	CL26	13	0.0003	.989	.996	-21	4201	27.5	0.4758	
15	CL32	CL34	65	0.0016	.987	.996	-21	3923	259	0.4832	
14	CL24	CL47	21	0.0008	.987	.995	-19	3984	56.2	0.5414	
13	CL31	CL22	115	0.0029	.984	.994	-20	3539	241	0.5852	
12	CL21	CL30	46	0.0024	.981	.993	-20	3351	154	0.6627	
11	CL25	CL18	185	0.0087	.973	.992	-24	2499	468	0.7077	
10	CL19	CL23	211	0.0098	.963	.990	-26	2027	580	0.7119	
9	CL14	CL20	61	0.0060	.957	.988	-25	1952	218	0.9765	
8	CL12	CL27	55	0.0032	.954	.985	-21	2073	52.0	0.994	
7	CL16	CL17	20	0.0029	.951	.980	-17	2272	85.9	1.0929	
6	CL11	CL15	250	0.0246	.926	.973	-19	1774	434	1.2003	
5	CL13	CL10	326	0.0416	.885	.961	-21	1355	769	1.3143	
4	CL9	CL7	81	0.0303	.854	.938	-17	1384	214	2.1836	
3	CL8	CL6	305	0.0751	.779	.890	-15	1252	505	2.2172	
2	CL4	CL5	407	0.1857	.594	.751	-13	1037	748	3.5219	
1	CL2	CL3	712	0.5936	.000	.000	0.00	.	1037	5.7435	



¿Qué método es el mejor?.

A partir de las salidas obtenidas en SAS, los **candidatos relevantes** para cada método son los siguientes:

Método de Ward (mínima varianza)

- **8 clústeres**
  - R^2 = **0.965** (96.5%)
  - Pseudo-F = **2758**
  - Pseudo-t^2 = **70.8**
- **5 clústeres**
  - R^2 = **0.910** (91.0%)
  - Pseudo-F = **1788**
  - Pseudo-t^2 = **313**

Ambas soluciones cumplen sobradamente el criterio de $R^2 \geq 70\%$.
No obstante, 8 clústeres resultan excesivos para el objetivo de segmentación.

Enlace promedio (distancia media)

- **8 clústeres**
  - R^2 = **0.960** (96.0%)
  - Pseudo-F = **2415**
  - Pseudo-t^2 = **97.5**
- **4 clústeres**
  - R^2 = **0.867** (86.7%)
  - Pseudo-F = **1545**
  - Pseudo-t^2 = **70.8**

Ambas soluciones cumplen el umbral de R^2.
Destaca especialmente la solución de **4 clústeres**, ya que combina:

- R^2 claramente superior al 70%,
- **Pseudo-F muy elevado**,
- **mínimo claro de la Pseudo-t^2**

Distancia entre centroides

- **8 clústeres**
  - R^2 = **0.958** (95.8%)
  - Pseudo-F = **2279**
  - Pseudo-t^2 = **89.0**
- **4 clústeres**
  - R^2 = **0.864** (86.4%)
  - Pseudo-F = **1506**
  - Pseudo-t^2 = **57.8**

Aunque los valores son aceptables y cumplen $R^2 \geq 70\%$, los estadísticos Pseudo-F son **inferiores** a los obtenidos con el enlace promedio para el mismo número de clústeres.

Enlace completo (vecino más lejano)

- **8 clústeres**
  - R^2 = **0.954** (95.4%)
  - Pseudo-F = **2073**
  - Pseudo-t^2 = **52.0**
- **4 clústeres**
  - R^2 = **0.864** (86.4%)
  - Pseudo-F = **1384**
  - Pseudo-t^2 = **214**

Aunque cumple el criterio de R^2, el valor de Pseudo-F para 4 clústeres es **menor que en otros métodos**, lo que indica una peor separación relativa.

Enlace simple (vecino más cercano)

Este método se descarta completamente:

- $R^2 \approx$ **0.49** (49.2%), **muy por debajo del 70%**
- Pseudo-F bajo y decreciente
- Ausencia de mínimos claros en la Pseudo-t^2
- Presencia del efecto cadena

No cumple el criterio mínimo de calidad.

Método	Nº clústeres	R^2	Pseudo-F	Pseudo-t^2	¿Cumple $R^2 \geq 70\%$?
Ward	8	0.965	2758	70.8	Sí
Ward	5	0.910	1788	313	Sí
Average	8	0.960	2415	97.5	Sí
Average	**4**	**0.867**	**1545**	**70.8**	**Sí**
Centroides	8	0.958	2279	89.0	Sí
Centroides	4	0.864	1506	57.8	Sí
Completo	8	0.954	2073	52.0	Sí
Completo	4	0.864	1384	214	Sí
Simple	—	0.492	—	—	**No**

Tras comparar los distintos métodos jerárquicos utilizando el Factor 1, se observa que todos los métodos, excepto el enlace simple, cumplen el criterio mínimo de calidad establecido ($R^2 \geq 70\%$).

Aunque las soluciones con 8 clústeres presentan valores de R^2 y Pseudo-F muy elevados, el número de grupos resulta excesivo para el objetivo de segmentación de la población. Por este motivo, se priorizan soluciones con **4 clústeres**, que mantienen un buen ajuste y una interpretación más clara.

Entre las soluciones con 4 clústeres, el **método de enlace promedio** destaca por presentar el **mayor valor del estadístico Pseudo-F**, junto con un **mínimo claro de la Pseudo-t²** y un R^2 claramente superior al 70%.

En consecuencia, se selecciona como solución final el **método de enlace promedio con 4 clústeres**, al considerarse la opción más equilibrada desde el punto de vista estadístico e interpretativo.

3.7 Composición de clusters. Representación—DEFINITIVO.

A continuación, procedemos a obtener los clústeres correspondientes al número de grupos seleccionado, que en nuestro caso ha resultado ser **n = 4**, empleando la salida del método de distancia elegido, que ha sido el **método de enlace promedio (distancia media)**.

Para ello, utilizamos la salida generada previamente por el procedimiento PROC CLUSTER y aplicamos el procedimiento PROC TREE, fijando el número de clústeres en cuatro:

```
proc tree data=salida_average_factor1 ncl=4 out=clusters_avg_factor1 noprint;
copy traveltime studytime failures famrel freetime goout dalc walc health absences G1 G2
G3 Medu  factor1;
id id;
run;
```

Mediante esta salida se obtiene el dendrograma que permite visualizar el proceso de formación de los clústeres en función de las observaciones. No obstante, dado que el número de observaciones es elevado, dicho dendrograma resulta **poco interpretable**, por lo que no se representa gráficamente en esta práctica.

Seguidamente, ordenamos la salida obtenida en función del clúster asignado a cada observación y visualizamos el resultado:

```
proc sort data = clusters_avg_factor1; by cluster; run;
proc print data = clusters_avg_factor1; run;
```

De esta forma, se obtiene una tabla que indica el clúster al que pertenece cada observación, identificada mediante la variable id.

A continuación, se construye la tabla de frecuencias para analizar la distribución de las observaciones entre los distintos clústeres formados:

```
PROC FREQ DATA=clusters_avg_factor1;
   TABLES CLUSTER;
   TITLE "Número de observaciones por Clúster";
RUN;
```

Número de observaciones por Clúster

The FREQ Procedure

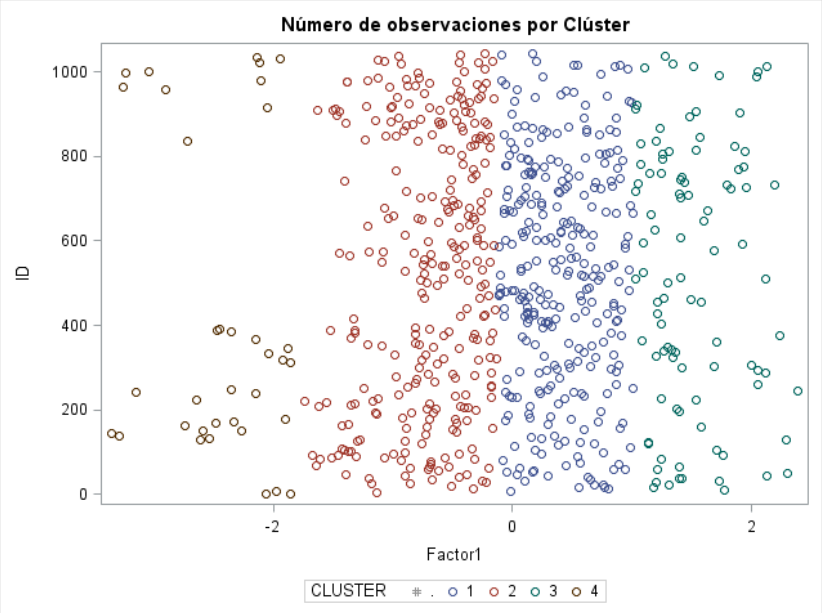
CLUSTER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	294	41.29	294	41.29
2	286	40.17	580	81.46
3	97	13.62	677	95.08
4	35	4.92	712	100.00

Frequency Missing = 1

A partir de esta salida se observa que las observaciones se distribuyen entre los cuatro clústeres de forma razonablemente equilibrada, sin que ninguno de ellos concentre de manera excesiva la mayoría de la muestra.

Por último, se representa gráficamente la asignación de las observaciones a los clústeres con el fin de comprobar si se han construido grupos **heterogéneos entre sí y homogéneos internamente**:

```
proc sgplot data= clusters_avg_factor1;
scatter x=factor1 y=id / group=cluster;
run;
proc print data = clusters_avg_factor1; run;
```



En el gráfico se representan las observaciones en función del **Factor 1**, diferenciadas según el clúster al que pertenecen. Se observa que los clústeres quedan claramente ordenados a lo largo de este eje, lo que indica que el **Factor 1 discrimina adecuadamente entre los grupos**.

Aunque existen algunas observaciones próximas en las fronteras entre clústeres, no se aprecia una mezcla significativa entre ellos. Por tanto, los clústeres obtenidos presentan una separación clara, siendo relativamente homogéneos internamente y heterogéneos entre sí.

4 Análisis No Jerárquico.

Una vez obtenida la partición definitiva mediante el análisis clúster jerárquico, se procede a realizar un **análisis clúster no jerárquico** con el objetivo de **comprobar si es posible mejorar la segmentación obtenida previamente**.

El análisis no jerárquico requiere fijar de antemano el número de clústeres, que en este caso se establece en **cuatro**, de acuerdo con la solución seleccionada en el análisis jerárquico. De este modo, se pretende evaluar si este enfoque permite obtener una asignación más precisa de las observaciones, manteniendo la coherencia con los resultados anteriores.

Al igual que en el análisis jerárquico, se emplea únicamente el **Factor 1**, con el fin de facilitar la comparación entre ambos métodos y valorar posibles mejoras en la estructura de los clústeres.

4.1 Estandarización y centroides.

Con el fin de facilitar la convergencia del algoritmo no jerárquico y evitar una selección aleatoria de los centros iniciales, se procede previamente a la **estandarización de las variables** y al **cálculo de los centroides iniciales** de los clústeres obtenidos mediante el método jerárquico. Estos centroides representan el valor medio del Factor 1 en cada clúster y se utilizan como punto de partida en el análisis no jerárquico.

```
proc standard data=clusters_avg_factor1 MEAN=0 STD=1 OUT=SALIDA_stan;
VAR g1 g2 g3 factor1;
RUN;

proc sort data = salida_stan; by cluster; run;

proc means data=SALIDA_stan NOPRINT;
BY cluster;
Var g1 g2 g3 factor1;
OUTPUT OUT=CENTINIC MEAN= factor1;/* ESCRIBIR NOMBRE DE  LAS VARIABLES*/
RUN;
PROC PRINT DATA=CENTINIC;
RUN;
```

Obs	CLUSTER	_TYPE_	_FREQ_	factor1
1	.	0	1	.
2	1	0	294	0.40959
3	2	0	286	-0.75911
4	3	0	97	1.58181
5	4	0	35	-1.62139

La tabla de centroides obtenida recoge, para cada clúster, el número de observaciones que lo componen y el valor medio del Factor 1, lo que permite identificar la posición de cada grupo en el espacio de la variable utilizada. A partir de estos valores, se aplicará el procedimiento de clúster no jerárquico con el objetivo de evaluar si la reasignación iterativa de las observaciones permite obtener una partición más precisa o estable que la obtenida mediante el análisis jerárquico.

- Clúster 1: 294 observaciones, centro en factor1 ≈ 0.41
- Clúster 2: 286 observaciones, centro en factor1 ≈ −0.76
- Clúster 3: 97 observaciones, centro en factor1 ≈ 1.58
- Clúster 4: 35 observaciones, centro en factor1 ≈ −1.62

4.2 Aplicación del algoritmo FASTCLUS y comparación de soluciones.

Una vez calculados los centroides iniciales a partir de la solución jerárquica, se procede a aplicar el **análisis clúster no jerárquico mediante el procedimiento PROC FASTCLUS**, con el objetivo de evaluar si este método permite **mejorar la segmentación obtenida previamente**.

En primer lugar, se ejecuta el algoritmo no jerárquico utilizando como **centros iniciales los centroides calculados a partir del método jerárquico**. De este modo, el procedimiento FASTCLUS parte de una solución informada, lo que permite comprobar si la reasignación iterativa de las observaciones conduce a una mejora respecto a la partición jerárquica original.

```
PROC FASTCLUS DATA=clusters_avg_factor1 SEED=CENTINIC RADIUS=0 REPLACE=FULL DISTANCE
MAXCLUSTERS=4 OUT=CLUSTER_definitivo MAXITER=20;
VAR factor1;
run;
```

En primer lugar, se comparan los resultados del análisis clúster jerárquico y del análisis clúster no jerárquico con semilla, atendiendo a los dos criterios principales de evaluación: el R-cuadrado aproximado y el estadístico Pseudo F.

En el análisis jerárquico, el valor del R-cuadrado aproximado es:

- **Approximate Expected Overall R-Squared** ≈ 0.938

y el valor del estadístico Pseudo F es:

- **Pseudo F** = 1545

Por su parte, en el análisis no jerárquico con semilla, el R-cuadrado aproximado toma prácticamente el mismo valor:

- **Approximate Expected Overall R-Squared** ≈ 0.938

lo que indica que ambos métodos explican una proporción muy similar de la variabilidad total.

Sin embargo, al comparar el estadístico Pseudo F, se observa que el valor obtenido en el análisis no jerárquico es superior al del análisis jerárquico:

- **Pseudo F (no jerárquico con semilla)** = 1674.84

Dado que uno de los criterios fundamentales de selección es **maximizar el estadístico Pseudo F**, este resultado indica una **mejor separación entre clústeres** en la solución no jerárquica con semilla. En consecuencia, puede afirmarse que el análisis clúster no jerárquico con semilla **mejora ligeramente la solución jerárquica**, manteniendo el mismo nivel de ajuste global y aumentando la diferenciación entre grupos.

Posteriormente, se repite el análisis clúster no jerárquico **sin fijar centros iniciales**, permitiendo que el algoritmo seleccione los centroides de forma **aleatoria**. Esta segunda ejecución tiene como finalidad contrastar si una inicialización aleatoria es capaz de producir una solución comparable o superior a la obtenida utilizando los centroides derivados del análisis jerárquico.

Para la ejecución del procedimiento PROC FASTCLUS se fija el número de clústeres en **cuatro**, manteniendo el uso exclusivo del **Factor 1**. Asimismo, fue necesario **incrementar el número máximo de iteraciones a 30**, dado que con el valor por defecto (MAXITER = 20) el algoritmo **no alcanzaba la convergencia**, lo que indica una mayor dificultad en la estabilización de los clústeres cuando los centros iniciales se eligen de forma aleatoria.

```
PROC FASTCLUS DATA=clusters_avg_factor1 random = 12345678 RADIUS=0 REPLACE=random
DISTANCE
MAXCLUSTERS=4 OUT=CLUSTER_definitivo_random MAXITER=30;
VAR factor1;
run;
```

Se obtiene nuevamente un **R-cuadrado aproximado** en torno a **0.938**, similar al observado en las soluciones anteriores. Asimismo, el valor del estadístico **Pseudo F** obtenido en este caso (**Pseudo F = 1675.02**) es muy próximo al alcanzado mediante el análisis no jerárquico con semilla.

Estos resultados indican que el análisis no jerárquico con inicialización aleatoria ofrece una **calidad de ajuste y una separación entre clústeres similares** a las obtenidas con la inicialización basada en centroides. No obstante, el hecho de haber requerido un mayor número de iteraciones para alcanzar la convergencia sugiere que esta solución es **menos estable** que la obtenida utilizando centroides como semilla.

4.3 Análisis clúster no jerárquico con actualización de centroides (DRIFT).

Una vez analizadas las soluciones obtenidas mediante el análisis clúster no jerárquico con semilla y con inicialización aleatoria, se procede a repetir ambos análisis incorporando el **criterio DRIFT**, con el objetivo de evaluar si la **actualización iterativa de los centroides** permite mejorar la calidad de la partición obtenida.

El parámetro **DRIFT** permite que los centroides se **ajusten progresivamente durante la ejecución del algoritmo**, en lugar de permanecer fijos, facilitando así una mejor adaptación de los clústeres a la estructura real de los datos. De este modo, se pretende comprobar si esta mayor flexibilidad en la actualización de los centros conduce a una **mejor separación entre clústeres** y a una solución más estable.

- **Drift y semilla.**

En primer lugar, se aplica el algoritmo FASTCLUS utilizando como centros iniciales los **centroides obtenidos previamente** y permitiendo su actualización mediante DRIFT. En este caso, el algoritmo alcanza correctamente la convergencia con un número máximo de iteraciones igual a 20.

```
PROC FASTCLUS DATA=clusters_avg_factor1 SEED=CENTINIC RADIUS=0 REPLACE=FULL DISTANCE
DRIFT
MAXCLUSTERS=4 OUT=CLUSTER_definitivo_drift MAXITER=20;
VAR factor1;
run;
```

Convergence criterion is satisfied.

Pseudo F Statistic =	1679.75
Approximate Expected Over-All R-Squared =	0.93820
Cubic Clustering Criterion =	-14.033

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
Factor1	0.99939	0.35151	0.876810	7.117569
OVER-ALL	0.99939	0.35151	0.876810	7.117569

Los resultados obtenidos son los siguientes:

- **Pseudo F = 1679.75**
- **Approximate Expected Overall R-Squared = 0.93820**
- **CCC = −14.033**

El valor del R-cuadrado aproximado es muy elevado y coincide con el obtenido en los análisis anteriores, lo que indica que se mantiene una alta proporción de variabilidad explicada. Asimismo, el valor del estadístico **Pseudo F es el más alto de todas las soluciones analizadas**, lo que refleja una **mejor separación entre clústeres**. Al igual que en los casos anteriores, el criterio CCC toma valores negativos y no resulta interpretable.

- **Drift y aleatorio.**

A continuación, se repite el análisis utilizando **inicialización aleatoria de los centroides**, permitiendo igualmente su actualización mediante DRIFT. En este caso, fue necesario aumentar el número máximo de iteraciones a 30 para alcanzar la convergencia.

```
PROC FASTCLUS DATA=clusters_avg_factor1 random = 12345678 RADIUS=0 REPLACE=random
DISTANCE DRIFT
MAXCLUSTERS=4 OUT=CLUSTER_definitivo_random_drift MAXITER=30;
VAR factor1;
run;
```

Convergence criterion is satisfied.

Pseudo F Statistic =	1674.84
Approximate Expected Over-All R-Squared =	0.93820
Cubic Clustering Criterion =	-14.085

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
Factor1	0.99939	0.35196	0.876494	7.096794
OVER-ALL	0.99939	0.35196	0.876494	7.096794

Los resultados obtenidos son:

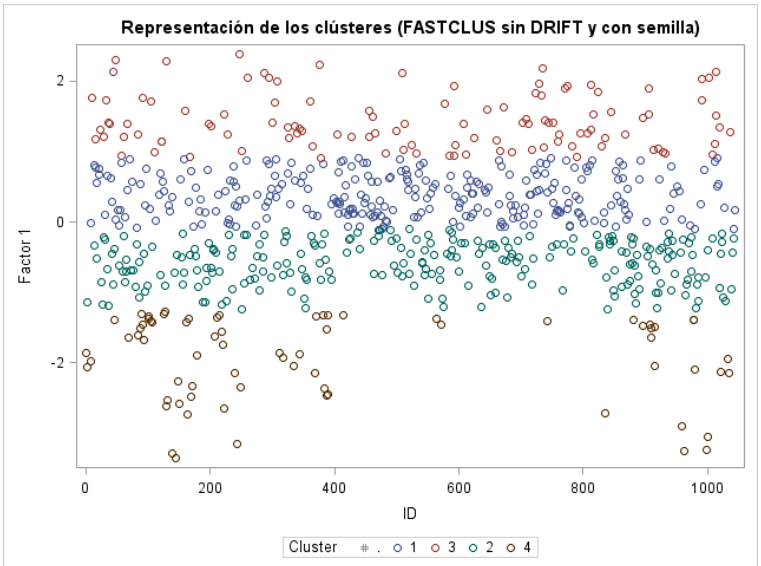
- **Pseudo F = 1674.84**
- **Approximate Expected Overall R-Squared = 0.93820**
- **CCC = −14.085**

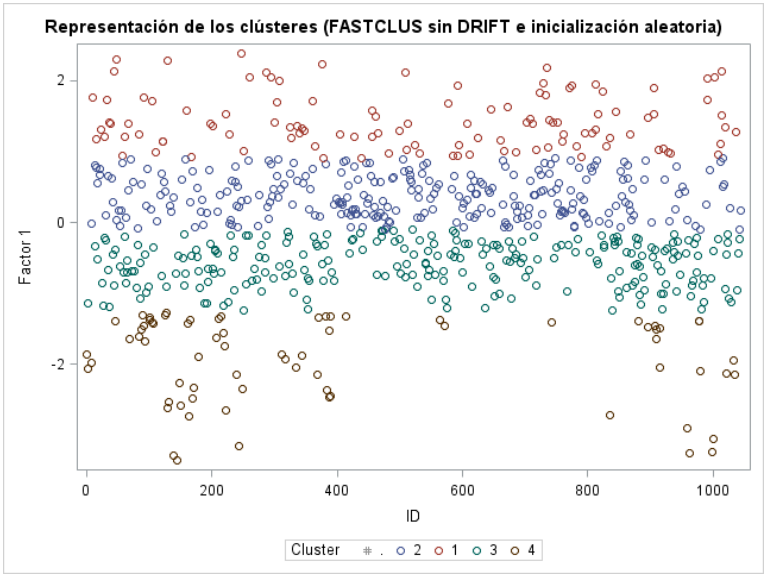
De nuevo, el R-cuadrado aproximado mantiene el mismo valor elevado, indicando una calidad de ajuste similar. No obstante, el valor del estadístico Pseudo F es **ligeramente inferior** al obtenido en la solución con DRIFT y semilla, lo que sugiere una separación entre clústeres algo menor.

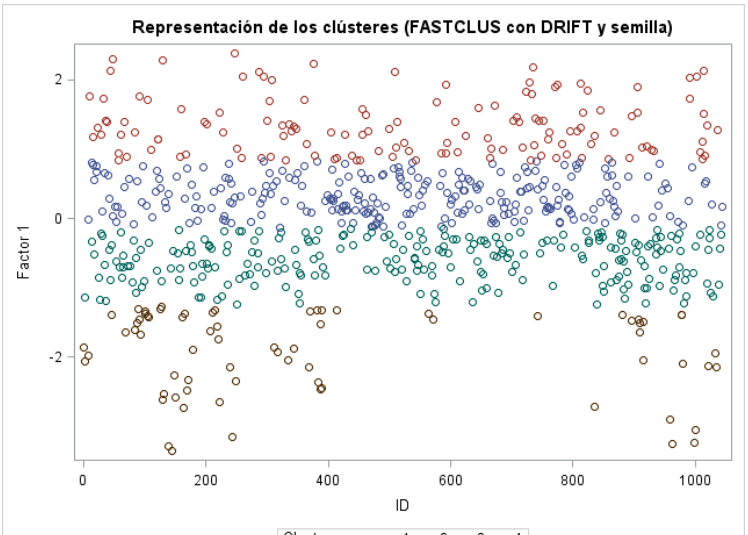
Por tanto, La incorporación del parámetro DRIFT permite mejorar la separación entre clústeres en el análisis no jerárquico. En particular, la solución obtenida mediante **FASTCLUS con DRIFT y semilla** presenta el **mayor valor del estadístico Pseudo F**, manteniendo un R-cuadrado aproximado elevado y constante. En consecuencia, esta solución se considera la **mejor de todas las analizadas** y se selecciona como partición final. Vamos a verlo gráficamente a ver como quedaría con cada opción ya que tengo valores muy similares y así lo comprobamos mejor.

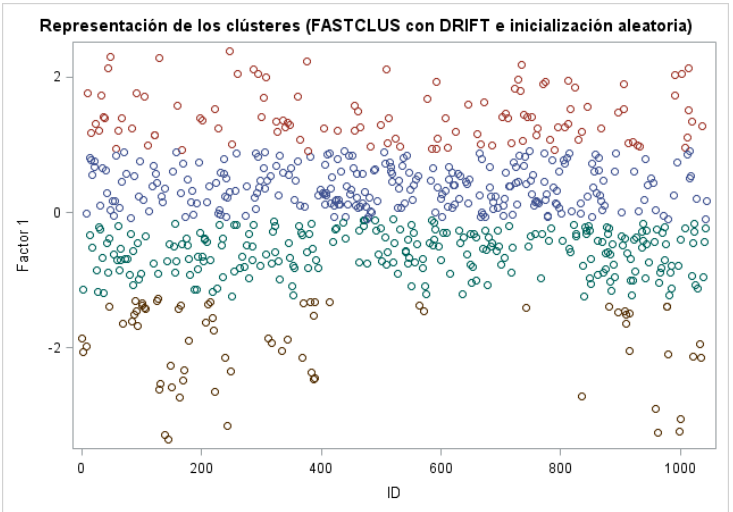
```
proc sgplot data=CLUSTER_definitivo;
   scatter y=factor1 x=id / group=cluster;
   xaxis label="ID";
   yaxis label="Factor 1";
   title "Representación de los clústeres (FASTCLUS sin DRIFT y con semilla)";
   keylegend / title="Cluster";
run;

proc sgplot data=CLUSTER_definitivo_random;
```

```
    scatter y=factor1 x=id / group=cluster;
    xaxis label="ID";
    yaxis label="Factor 1";
    title "Representación de los clústeres (FASTCLUS sin DRIFT e inicialización
aleatoria)";
    keylegend / title="Cluster";
run;


proc sgplot data=CLUSTER_definitivo_drift;
    scatter y=factor1 x=id / group=cluster;
    xaxis label="ID";
    yaxis label="Factor 1";
    title "Representación de los clústeres (FASTCLUS con DRIFT y semilla)";
    keylegend / title="Cluster";
run;


proc sgplot data=CLUSTER_definitivo_random_drift;
    scatter y=factor1 x=id / group=cluster;
    xaxis label="ID";
    yaxis label="Factor 1";
    title "Representación de los clústeres (FASTCLUS con DRIFT e inicialización
aleatoria)";
    keylegend / title="Cluster";
run;
```









Las representaciones gráficas de las distintas soluciones no jerárquicas son muy similares, ya que en todos los casos los clústeres se ordenan claramente a lo largo del Factor 1. Dado que el R-cuadrado aproximado es prácticamente idéntico en todas las soluciones, la elección final se basa en el estadístico Pseudo F. Por ello, se selecciona la solución con **mayor Pseudo F**, que en este caso corresponde a **FASTCLUS con DRIFT y semilla**.

4.4 Caracterización de los clusters obtenidos.

En este apartado se procede a la **caracterización de los clústeres obtenidos**, con el objetivo de analizar el comportamiento interno de cada grupo y comprobar su grado de homogeneidad y diferenciación.

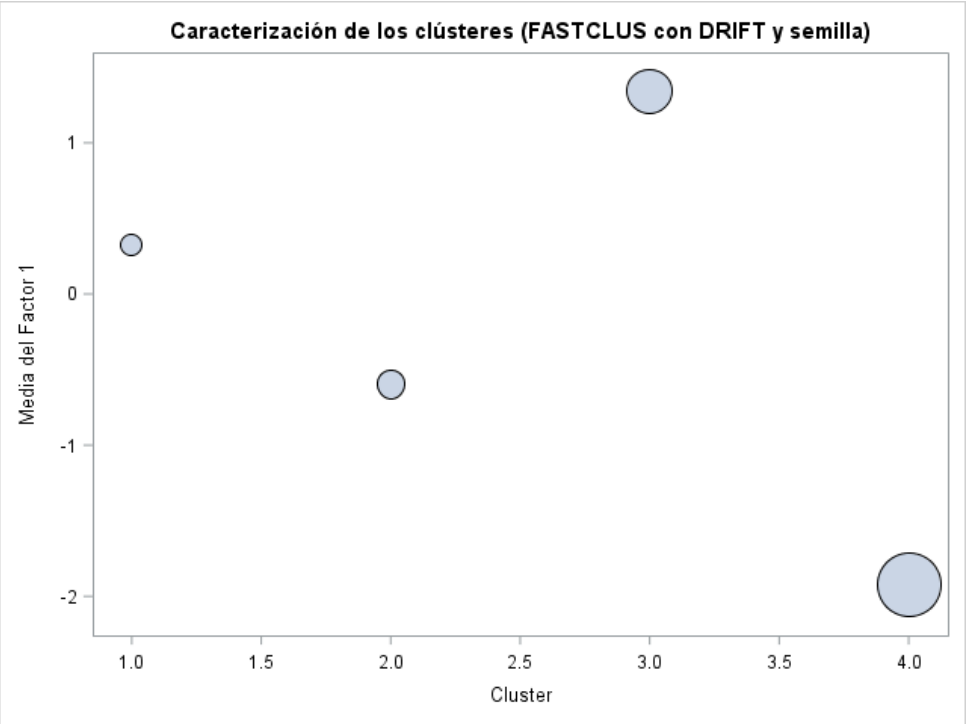
Medias y desviaciones típicas por clúster.

```
proc means data=CLUSTER_definitivo_drift mean std nway;
   class cluster;
   var factor1;
   output out=stats_cluster mean=mean_f1 std=std_f1;
run;
```

Analysis Variable : Factor1			
Cluster	N Obs	Mean	Std Dev
1	256	0.3238680	0.2672421
2	249	-0.5979898	0.2972667
3	136	1.3399237	0.4077459
4	71	-1.9252492	0.5965220

Gráfico de burbujas: caracterización de los clústeres.

```
proc sgplot data=stats_cluster;
   bubble x=cluster y=mean_f1 size=std_f1;
   xaxis label="Cluster";
   yaxis label="Media del Factor 1";
   title "Caracterización de los clústeres (FASTCLUS con DRIFT y semilla)";
run;
```



A partir de las medias y desviaciones típicas del **Factor 1**, así como del gráfico de burbujas, se observa una clara diferenciación entre los clústeres obtenidos. El **clúster 3** presenta el valor medio más elevado del Factor 1 (media ≈ 1.34), lo que indica un perfil claramente alto en esta dimensión. Por el contrario, el **clúster 4** muestra el valor medio más bajo (media ≈ −1.93), representando el perfil opuesto.

Los **clústeres 1 y 2** presentan valores intermedios: el clúster 1 se sitúa ligeramente por encima de la media global (media ≈ 0.32), mientras que el clúster 2 presenta valores moderadamente negativos (media ≈ −0.60). Esto sugiere una gradación clara de los individuos a lo largo del Factor 1, desde perfiles bajos hasta perfiles altos.

En cuanto a la variabilidad interna, las desviaciones típicas son relativamente reducidas en todos los clústeres, especialmente en los clústeres 1 y 2, lo que indica una elevada **homogeneidad interna**. El clúster 4 presenta una desviación algo mayor, reflejando una mayor dispersión, aunque sigue manteniendo una diferenciación clara respecto al resto de grupos.

En conjunto, tanto las medias como el gráfico de burbujas confirman que los clústeres están **bien definidos, homogéneos internamente y claramente diferenciados entre sí**, lo que valida la calidad de la segmentación obtenida.

5 Análisis Discriminante.

Una vez obtenidos los clústeres mediante el análisis clúster, realizamos un **análisis discriminante** con el objetivo de **comprobar si los grupos formados están bien diferenciados** y validar la calidad de la segmentación obtenida.

Este análisis permite evaluar hasta qué punto las variables utilizadas son capaces de **clasificar correctamente a los individuos en su clúster correspondiente**, funcionando así como una herramienta de contraste del análisis no supervisado.

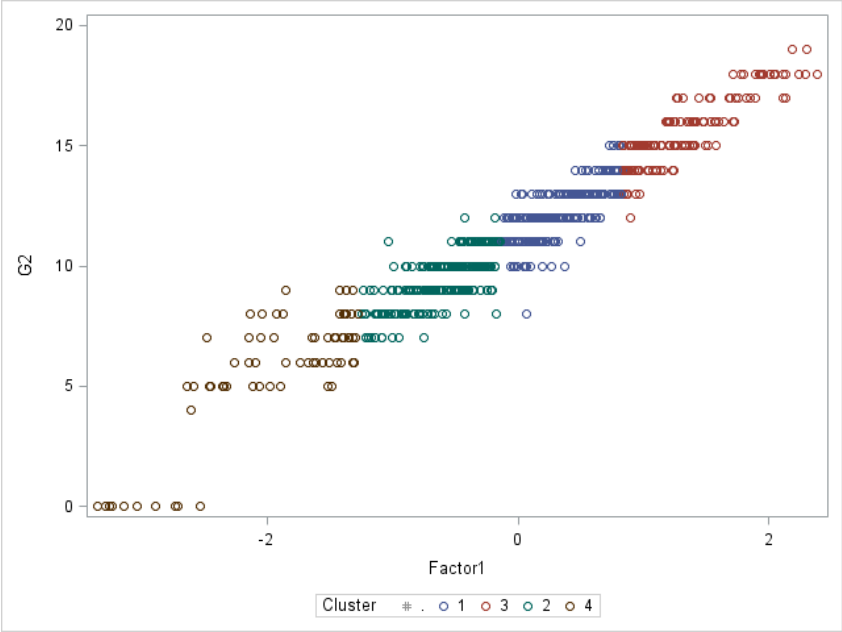
5.1 Selección de variables discriminantes.

En una primera fase, se aplicó el procedimiento **PROC STEPDISC** con el método *stepwise* con el objetivo de evaluar de forma exploratoria qué variables presentaban mayor capacidad para diferenciar los clústeres obtenidos. No obstante, yo solo tengo factor1, voy a meter todas las variables para comprobar que efectivamente las variables que más discriminan son las representadas por el factor1.

```
proc stepdisc data=CLUSTER_definitivo_drift method=stepwise sle=0.1;
   var traveltime studytime failures famrel freetime goout dalc walc health absences G1
G2 G3 Medu;
   class cluster;
run;
```

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	G2		0.8060	980.23	<.0001	0.19404163	<.0001	0.26865279	<.0001
2	2	G1		0.2669	85.79	<.0001	0.14225681	<.0001	0.29562828	<.0001
3	3	G3		0.1944	56.78	<.0001	0.11460704	<.0001	0.33986381	<.0001
4	4	traveltime		0.0342	8.33	<.0001	0.11068456	<.0001	0.34189702	<.0001
5	5	studytime		0.0293	7.08	0.0001	0.10744322	<.0001	0.34403947	<.0001
6	6	absences		0.0244	5.86	0.0006	0.10482029	<.0001	0.34686684	<.0001
7	7	failures		0.0218	5.22	0.0014	0.10253157	<.0001	0.35381963	<.0001
8	8	health		0.0206	4.91	0.0022	0.10042296	<.0001	0.35723795	<.0001
9	9	Medu		0.0170	4.03	0.0074	0.09871952	<.0001	0.35995262	<.0001
10	10	freetime		0.0181	4.30	0.0051	0.09693223	<.0001	0.36272835	<.0001
11	11	Dalc		0.0137	3.23	0.0220	0.09560495	<.0001	0.36428063	<.0001
12	12	famrel		0.0100	2.34	0.0722	0.09465151	<.0001	0.36546833	<.0001



Observamos que discrimina muy bien. Es por ello que para el proc discrim utilizaré estas variables en vez de usar factor1. No vamos a incluir más variables aunque salgan significativas porque si nos fijamos en el r cuadrado sale muy bajo para las siguientes, aportan muy poco. Por tanto nos pasamos a hacer el proc discrim con estas tres variables que son las que representan al factor1.

5.2 Comprobación de normalidad.

Antes de aplicar el análisis discriminante, es necesario comprobar el **supuesto de normalidad** de las variables discriminantes dentro de cada grupo. Aunque el análisis discriminante es relativamente robusto ante desviaciones de la normalidad cuando el tamaño muestral es elevado, resulta conveniente verificar este supuesto para justificar adecuadamente el modelo.

```
%NORMAL_MULT(DATA=CLUSTER_definitivo_drift, VAR=g1 g2 g3);
```

Obs	Y1	Y2	ACUMULADO	DIFERENCIA	REFERENCIA	HIPOTESIS
1	1.21253	4.10834	0	0.5	0.056175	Norm. rechazada

El contraste de normalidad indica que la **hipótesis de normalidad es rechazada**, ya que la diferencia observada es superior al valor de referencia (p-valor < 0.05). Por tanto, el Factor 1(g1,g2,g3) no sigue estrictamente una distribución normal dentro de los clústeres**.**

5.3 Análisis Discriminante-proc discrim.

Una vez comprobado que el supuesto de normalidad no se cumple estrictamente, se procede a realizar el **análisis discriminante.**

- **PRIORS PROPORTIONAL**
  Se utiliza porque los clústeres tienen tamaños distintos, asignando probabilidades a priori proporcionales al número de observaciones en cada grupo.
- **TESTDATA = practic2.restante**
  Se emplea el conjunto de datos reservado previamente (30%) para evaluar la capacidad predictiva del modelo sobre observaciones no utilizadas en su estimación.
- **CROSSVALIDATE**
  Se incluye para obtener una medida adicional y más robusta de la tasa de clasificación correcta mediante validación cruzada.
- **POOL = TEST**
  Se incorpora para contrastar la igualdad de las matrices de covarianzas y permitir que el modelo utilice automáticamente una función discriminante lineal o cuadrática. Al salirnos la normalidad rechazada nos vemos obligados a comprobarlo así.

PROGRAMACIÓN SAS-PROC DISCRIM

```
PROC DISCRIM DATA=cluster_definitivo
      OUT=SOL
      CROSSVALIDATE
      OUTSTAT=ESTADISTICOS
      POOL=TEST
      TESTDATA=practic2.restante;

      PRIORS PROPORTIONAL;

      CLASS CLUSTER;

      VAR g1 g2 g3;
RUN;
```

Test de homogeneidad de las matrices de covarianzas.

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
578.843208	18	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

El test de homogeneidad de las matrices de covarianzas es significativo ($p < 0.0001$), por lo que se rechaza la igualdad de covarianzas entre los grupos. En consecuencia, el análisis discriminante emplea **funciones cuadráticas**, al no cumplirse el supuesto de homocedasticidad.

Miramos el modelo en resustitución.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CLUSTER_DEFINITIVO
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into CLUSTER					
From CLUSTER	1	2	3	4	Total
1	249 91.21	16 5.86	8 2.93	0 0.00	273 100.00
2	20 7.84	230 90.20	0 0.00	5 1.96	255 100.00
3	11 9.73	0 0.00	102 90.27	0 0.00	113 100.00
4	0 0.00	16 22.54	0 0.00	55 77.46	71 100.00
Total	280 39.33	262 36.80	110 15.45	60 8.43	712 100.00
Priors	0.38343	0.35815	0.15871	0.09972	

Error Count Estimates for CLUSTER					
	1	2	3	4	Total
Rate	0.0879	0.0980	0.0973	0.2254	0.1067
Priors	0.3834	0.3581	0.1587	0.0997	

La matriz de clasificación muestra un **alto porcentaje de aciertos global**, con una **tasa de error total del 10.67%**, lo que implica que aproximadamente el **89.3% de las observaciones** se clasifican correctamente. Este resultado indica una **buena capacidad discriminante del modelo** sobre los datos de calibración.

Por clúster, los grupos **1, 2 y 3** presentan porcentajes de clasificación correcta superiores al **90%**, lo que refleja una separación clara entre estos grupos. El **clúster 4** muestra un mayor porcentaje de error (22.54%), lo que puede explicarse por su **menor tamaño muestral**, aunque sigue manteniendo una tasa de acierto razonable.

Miramos el modelo en validación cruzada.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CLUSTER_DEFINITIVO
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into CLUSTER					
From CLUSTER	1	2	3	4	Total
1	247 90.48	17 6.23	9 3.30	0 0.00	273 100.00
2	20 7.84	230 90.20	0 0.00	5 1.96	255 100.00
3	11 9.73	0 0.00	102 90.27	0 0.00	113 100.00
4	0 0.00	18 25.35	0 0.00	53 74.65	71 100.00
Total	278 39.04	265 37.22	111 15.59	58 8.15	712 100.00
Priors	0.38343	0.35815	0.15871	0.09972	

Error Count Estimates for CLUSTER					
	1	2	3	4	Total
Rate	0.0952	0.0980	0.0973	0.2535	0.1124
Priors	0.3834	0.3581	0.1587	0.0997	

La matriz de clasificación obtenida mediante **validación cruzada** muestra una **tasa de error total del 11.24%**, ligeramente superior a la obtenida en la resustitución, lo cual es esperable y refleja una evaluación más realista del modelo. En consecuencia, el porcentaje global de clasificación correcta se sitúa en torno al **88.8%**.

Los **clústeres 1, 2 y 3** mantienen porcentajes de acierto cercanos o superiores al **90%**, lo que confirma una buena capacidad discriminante del modelo. El **clúster 4** presenta nuevamente un mayor porcentaje de error (25.35%), explicado en parte por su menor tamaño y mayor heterogeneidad.

Miramos el modelo en datos no vistos(RESTANTE).

The DISCRIM Procedure
Classification Summary for Test Data: PRACTIC2.RESTANTE
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	305
Number of Observations Used	305

Number of Observations and Percent Classified into CLUSTER

	1	2	3	4	Total
Total	121 39.67	102 33.44	59 19.34	23 7.54	305 100.00
Priors	0.38343	0.35815	0.15871	0.09972	

- Cluster 1: 39.67%
- Cluster 2: 33.44%
- Cluster 3: 19.34%
- Cluster 4: 7.54%

Estas proporciones son muy similares a los priors:

- Priors: 0.383 – 0.358 – 0.159 – 0.100

Esto indica que el modelo generaliza bien y no está sesgado al clasificar datos nuevos.

La clasificación sobre el conjunto de test muestra una distribución de observaciones muy coherente con los priors estimados en la muestra de calibración, lo que sugiere que el modelo discriminante mantiene un comportamiento estable al aplicarse a datos nuevos.

Funciones Discriminantes.

1	QUAD	G1	-0.35414573	0.1836298464	0.0122152706
1	QUAD	G2	0.1836298464	-0.937565859	0.5923582988
1	QUAD	G3	0.0122152706	0.5923582988	-0.723631013
1	QUAD	_LINEAR_	3.8555564404	3.5678989503	3.5413221213
1	QUAD	_CONST_	-69.72437877	-69.72437877	-69.72437877
2	QUAD	G1	-0.367249039	0.0997238476	0.0959333709
2	QUAD	G2	0.0997238476	-0.838171447	0.5324302136
2	QUAD	G3	0.0959333709	0.5324302136	-0.754183242
2	QUAD	_LINEAR_	3.0388911104	3.5856940253	2.9147064638
2	QUAD	_CONST_	-46.42272926	-46.42272926	-46.42272926
3	QUAD	G1	-0.405664213	0.1751097379	0.1307845633
3	QUAD	G2	0.1751097379	-0.958585506	0.7296774647
3	QUAD	G3	0.1307845633	0.7296774647	-1.03106014
3	QUAD	_LINEAR_	2.8516886713	1.1612253921	6.372855394
3	QUAD	_CONST_	-85.36021391	-85.36021391	-85.36021391
4	QUAD	G1	-0.447131049	0.0888311805	-0.022207916
4	QUAD	G2	0.0888311805	-0.118673755	0.0419672201
4	QUAD	G3	-0.022207916	0.0419672201	-0.053838547
4	QUAD	_LINEAR_	5.1022257741	-0.169684949	0.2553383517
4	QUAD	_CONST_	-21.399793	-21.399793	-21.399793

$$D_1 = -69.72 + 3.86\,G1 + 3.57\,G2 + 3.54\,G3 - 0.35\,G1^2 - 0.94\,G2^2 - 0.72\,G3^2$$

$$D_2 = -46.42 + 3.04\,G1 + 3.59\,G2 + 2.91\,G3 - 0.37\,G1^2 - 0.84\,G2^2 - 0.75\,G3^2$$

$$D_3 = -85.36 + 2.85\,G1 + 1.16\,G2 + 6.37\,G3 - 0.41\,G1^2 - 0.96\,G2^2 - 1.03\,G3^2$$

$$D_4 = -21.40 + 5.10\,G1 - 0.17\,G2 + 0.26\,G3 - 0.45\,G1^2 - 0.12\,G2^2 - 0.05\,G3^2$$

6 Conclusiones.

En este trabajo se ha llevado a cabo un análisis clúster seguido de un análisis discriminante utilizando como base el **Factor 1**, construido a partir de las variables **G1, G2 y G3**. En una primera fase se exploraron distintas representaciones de los datos, analizando combinaciones como **Factor 1 y Factor 2**, **Factor 1 en solitario** y también las variables canónicas (**Can 1 y Can 2**). Tras comparar los resultados, se concluyó que **Factor 1** era la dimensión que mejor discriminaba entre individuos. A partir de ahí, se probaron distintos métodos de agrupación, tanto jerárquicos como no jerárquicos, seleccionando el mejor método jerárquico y comparándolo posteriormente con las soluciones no jerárquicas.

Finalmente, se decidió trabajar con una solución **no jerárquica**, concretamente el método **FASTCLUS con DRIFT y semilla**, al ser el que presenta un mejor equilibrio entre separación entre grupos y estabilidad de los resultados. La solución final consta de **4 clústeres**, que muestran perfiles bien diferenciados y una elevada homogeneidad interna.

Para validar esta segmentación, se dividieron los datos en un conjunto de calibración y un conjunto de validación. El análisis discriminante confirma que el modelo mantiene tasas de clasificación similares en resustitución y validación cruzada, lo que indica una buena capacidad de generalización y ausencia de sobreajuste. Además, al aplicar el modelo a los datos nuevos, la distribución de las observaciones entre clústeres resulta coherente con las proporciones originales, reforzando la estabilidad de la segmentación obtenida.

En conjunto, los resultados muestran que la solución no jerárquica seleccionada es consistente, interpretable y generaliza correctamente a nuevas observaciones.