Examen 2: Análisis de Factores + Análisis de Clusters

Néstor Iván Martínez Ostoa - 315618648
Análisis Multivariado

9 de diciembre del 2021

1 Pregunta: Análisis de Factores

Descripción: use los datos correspondientes a gastos realizados por familias francesas de diferentes niveles socio-económicos en bienes comestibles para responder lo siguiente:

- ¿Qué información puede obtener al hacer un análisis de factores?
- ¿Cómo se relaciona esta con respecto a la información proveniente de componentes principales?
- ¿Hay relación alguna con la pregunta 3 del examen 1?

1.1 Análisis de Factores

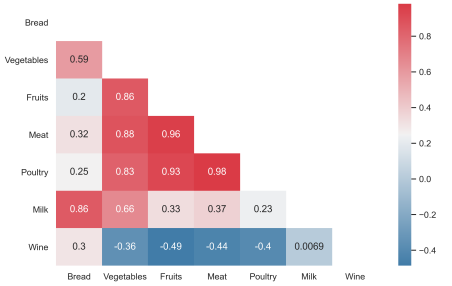
La metodología que emplearé será la siguiente:

1. Análisis de la estructura de la matriz de correlaciones R para dar un número inicial de factores k
2. Análisis de los grados de libertad d
3. Análisis de factores
4. Información obtenida

1.1.1 Matriz de correlaciones R

Comenzamos realizando un análisis de correlación para determinar un número óptimo de factores. A continuación se muestra la matriz de correlaciones en un mapa de calor y de manera numérica

(figura 1) para los datos de las familias francesas.



	Bread	Vegetables	Fruits	Meat	Poultry	Milk	Wine
Bread	1.000000	0.593110	0.196139	0.321269	0.248008	0.855575	0.303761
Vegetables	0.593110	1.000000	0.856250	0.881081	0.826777	0.662799	-0.356468
Fruits	0.196139	0.856250	1.000000	0.959477	0.925542	0.332189	-0.486281
Meat	0.321269	0.881081	0.959477	1.000000	0.981785	0.374591	-0.437235
Poultry	0.248008	0.826777	0.925542	0.981785	1.000000	0.232892	-0.400157
Milk	0.855575	0.662799	0.332189	0.374591	0.232892	1.000000	0.006880
Wine	0.303761	-0.356468	-0.486281	-0.437235	-0.400157	0.006880	1.000000

Figura 1: Matriz de correlaciones para los datos de gastos en comida de familias francesas

De la figura 1 podemos observar lo siguiente:

- Las familias que compraron pan también compraron leche (las variables *bread* y *milk* tienen una correlación de 0.86)
- Las familias que compraron verduras, también compraron frutas, carne y pollo (correlación de 0.86, 0.88 y 0.83 respectivamente)
- Las familias que compraron carne, también compraron pollo (correlación de 0.98)

Derivado de las observaciones anteriores, podemos plantear un modelo con $k = 3$ factores o incluso $k = 2$. Esto se debe al hecho de que tenemos tres grupos de correlaciones:

1. Variables pan y leche
2. Variables verduras, frutas, carne, pollo
3. Variable de vino

1.1.2 Grados de libertad d

p	k	$d = 0.5(p-k)^2 + 0.5(p+k)$
7	2	8
7	3	3
7	4	-1

Cuadro 1: Grados de libertad

De la tabla 1 podemos observar que en ambos casos se trata del caso en donde no exista una solución única ($d > 0$) por lo que podemos tomar hasta 3 factores para nuestro análisis. El análisis lo haré empezando con $k = 3$ dado que esto implica que el modelo tendrá 3 parámetros, si usáramos $k = 2$, tendríamos que considerar 8 parámetros lo cual no tendría sentido porque uno de los objetivos de análisis de factores es reducir la dimensionalidad de los datos originales.

1.1.3 Análisis de factores

Para realizar el análisis de factores consideré cuatro algoritmos distintos utilizando $k = 3$ factores. Para cada uno de ellos muestro los valores de los factores, comunalidades y varianza específica; así como las gráficas de factores. A continuación éste análisis:

1. Método de Máxima Verosimulitud (MLM por sus siglas en inglés) [1]

   - **Análisis**: este modelo implica que podemos utilizar los primeros dos factores para explicar la varianza de los datos (valores en **negritas**). Sin embargo, analizando las comunalidades, podemos observar que ningún factor logra explicar la varianza suficiente de ninguna variable. Es valor más alto es 0.2641 para la variable de frutas, sin embargo, es muy bajo considerando que queremos que sea lo más cercano a 1
   - **Tabla de valores**

Load1	Load2	Load3	Comunalidades	Varianzas especificas	Variables
0.1135	**-0.2671**	0.0558	0.0873	0.9127	Bread
-0.377	-0.1942	-0.0147	0.1801	0.8199	Vegetables
-0.5093	-0.0512	-0.0459	0.2641	0.7359	Fruits
-0.4902	-0.0684	-0.0127	0.2452	0.7548	Meat
-0.4926	-0.0172	0.0078	0.243	0.757	Poultry
0.008	**-0.3283**	-0.0302	0.1087	0.8913	Milk
0.4901	0.0906	0.0835	0.2553	0.7447	Wine

Cuadro 2: Método de Máxima Verosimulitud

   - **Gráfica de factores**

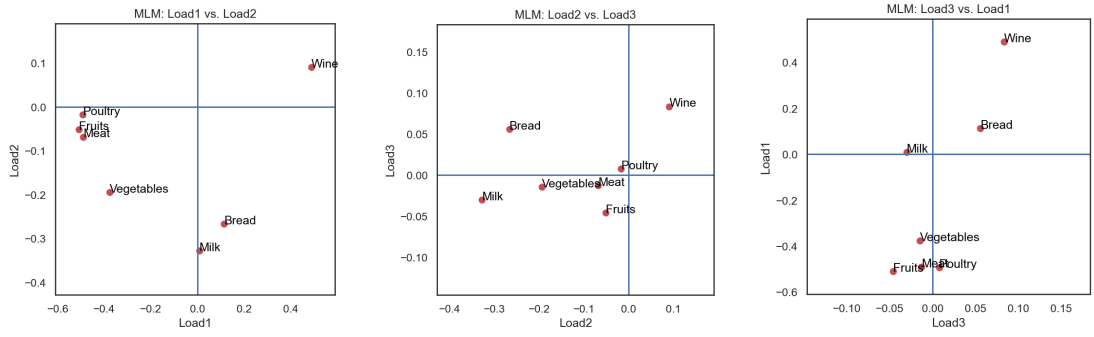


Figura 2: Método de Máxima Verosimilitud

2. Método de Máxima Verosimulitud (MLM) con rotación *varimax*

   - **Análisis**: de igual forma que con el modelo presentado en la tabla 2, si aplicamos una rotación *varimax*, los resultados siguen siendo muy similares. El método de máxima

[1] *Maximum Likelihood Method*

verosimilitud (MLM) no logra explicar la varianza de los datos de manera efectiva. De igual forma, los factores 2 y 3 logran explicar todas las variables pero con valores de varianza muy bajo. La precisión de este modelo es idéntica a la del punto anterior (ver tabla 2) 0.18 pues la rotación no altera la comunalidad

- **Tabla de valores**

Load1	Load2	Load3	Comunalidades	Varianzas especificas	Variables
-0.0499	**-0.28**	-0.0804	0.0873	0.9127	Bread
-0.0099	-0.1425	**0.3996**	0.1801	0.8199	Vegetables
0.0137	0.0169	**0.5135**	0.2641	0.7359	Fruits
-0.0184	-0.0028	**0.4948**	0.2452	0.7548	Meat
-0.0387	0.0482	**0.4891**	0.243	0.757	Poultry
0.0291	**-0.3263**	0.0374	0.1087	0.8913	Milk
-0.0523	0.0245	**-0.502**	0.2553	0.7447	Wine

Cuadro 3: Método de Máxima Verosimulitud con rotación *varimax*

- **Gráfica de factores**

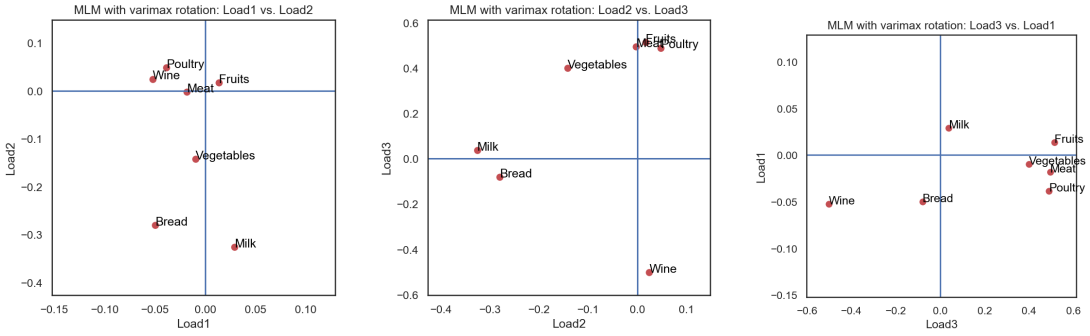


Figura 3: Método de Máxima Verosimilitud con rotación *varimax*

3. Método de Componentes Principales (PCM por sus sigás en inglés) después de la rotación *varimax*

   - **Análisis**: a diferencia de MLM, el métodoo de componentes principales -de entrada- ya utiliza los tres factores para explicar la varianza de las 7 variables. El factor 1 [2] explica las variables de verduras, frutas, carne y pollo, mientras que el factor 2 explica las variables de pan y leche y el factor 3 la variable de vino. De manera análoga, la varianza explicada por este modelo es del 97%, un valor casi 9 veces más grande que los modelos de MLM. Esto es un buen valor porque indica que PCM es un modelo que con los tres factores logra explicar el 97% de la variabilidad de los datos.

   - **Tabla de valores**

[2] En la sección **Información obtenida**, explico qué significa cada uno de los factores

Load1	Load2	Load3	Comunalidades	Varianzas especificas	Variables
-0.1869	**0.9123**	0.2997	0.9571	0.0429	Bread
-0.781	0.5582	-0.1975	0.9606	0.0394	Vegetables
-0.9313	0.1439	-0.2565	0.9539	0.0461	Fruits
-0.9552	0.2114	-0.1795	0.9893	0.0107	Meat
-0.981	0.0843	-0.1069	0.981	0.019	Poultry
-0.1586	**0.9668**	-0.0742	0.9653	0.0347	Milk
0.302	0.111	**0.9401**	**0.9873**	**0.0127**	Wine

Cuadro 4: Método de componentes principales

- **Gráfica de factores**

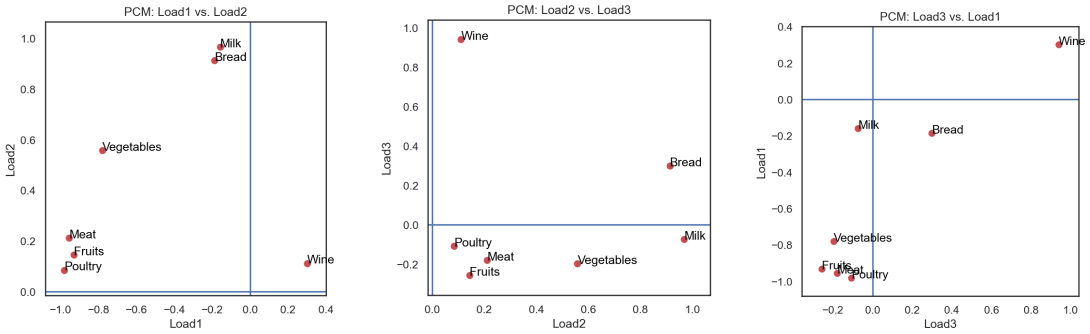


Figura 4: Método de componentes principales

4. Método de Factores Principales (FCM por sus sigás en inglés) después de la rotación *varimax*

- **Análisis**: finalmente, el método de factores principales, de igual forma que PCM, es un buen modelo pues los tres factores explican el 92% de la variabilidad de los datos. De igual forma, el factor 1 explica las variables de verduras, frutas, carne y pollo, el factor 2 las variables de pan y leche mientras que el factor 3 el de vino. La diferencia principal contra PCM, es el porcentaje de varianza que explica para la variable de vino, 0.7483 vs 0.9873 de PCM
- **Tabla de valores**

Load1	Load2	Load3	Comunalidades	Varianzas especificas	Variables
-0.179	**0.91**	0.3043	0.9528	0.0472	Bread
-0.7784	0.5534	-0.1971	0.951	0.049	Vegetables
-0.9177	0.1478	-0.2463	0.9246	0.0754	Fruits
-0.9648	0.2085	-0.1472	0.9961	0.0039	Meat
-0.9891	0.0812	-0.0538	0.9877	0.0123	Poultry
-0.1641	**0.9435**	-0.1052	0.9282	0.0718	Milk
0.3668	0.153	**0.7684**	**0.7483**	**0.2517**	Wine

Cuadro 5: Método de factores principales

- **Gráfica de factores**

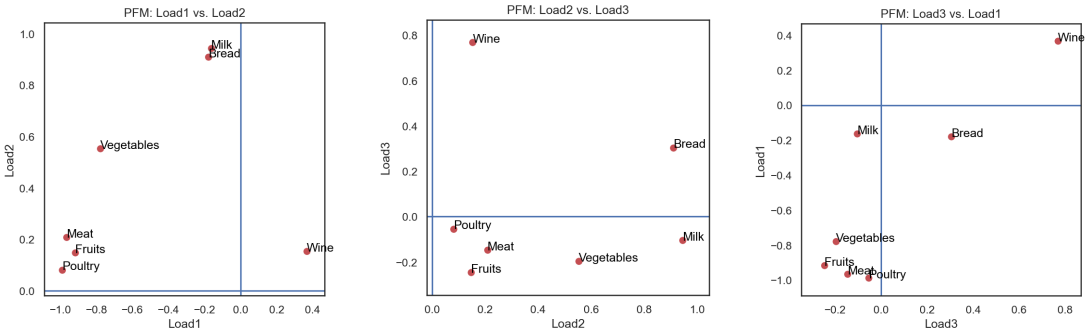


Figura 5: Método de factores principales

1.1.4 Información obtenida

Con base en la subsección de **Análisis** para cada uno de los modelos obtenidos, opté por elegir el modelo de componentes principales (tabla 4 y figura 4), pues es el que mayor varianza explica de los datos 97%. Ahora sí, con respecto a la información obtenida de este modelo, podemos notar lo siguiente:

- El factor 1 explica con una correlación altamente negativa las variables de verduras, frutas, carne y pollo. Por lo que podría considerarse como un factor de comida de lujo. Mientras menos sea el gasto en comida de lujo, mayor será el gasto en comida esencial y viceversa
- El factor 2 explica con una correlación altamente positiva las variables de pan y leche. Por lo que éste factor puede ser considerado como un factor de merienda puesto que usualmente en la merienda es cuando más se consume la leche y pan. Esto debido a la correlación positiva entre el factor 2 con el pan y leche
- El factor 3 explica con una correlación altamente positiva la variable de vino por lo que puede considerarse como el factor de bebidas de lujo
- Con respecto a las gráficas (tomando las gráficas de la izquierda y derecha de la figura 4), podemos observar que el factor 1 (gráfica de hasta la izquierda) representado en el eje X, es capaz de dividir a las comidas esenciales (verduras, frutas, pollo y carne) de las de lujo (vino)
- Al final de éste análisis, por las interpretaciones de los factores y el porcentaje de varianza explicada, podemos también concluir que utilizar únicamente dos factores es suficiente para explicar la varianza de los datos puesto que el tercer factor no es muy claro lo que hace considerando los otros dos factores que hacen la distinción entre comidas esenciales y de lujo

1.2 Relación con Componentes Principales

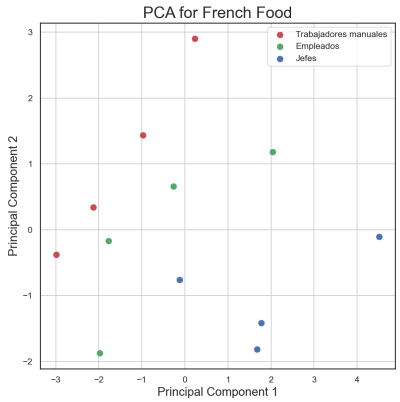


Figura 6: Análisis de componentes principales

En la figura 6 se observa el resultado de componentes principales aplicados a los mismo datos. Si bien análisis de factores y componentes principales utilizan las mismas herramientas matemáticas, la diferencia está en cómo se construye el modelo. La información obtenida en análisis de factores (sobre el factor que implica que hay una diferencia sustancial entre comidas esenciales como verduras, frutas, carne y pollo y las comidas de lujo como el vino) se relaciona con componentes principales en cuánto al dinero que cada familia invierte en comida. En componentes principales, podemos observar que en el cuadrante superior izquierdo se agrupan las familias de jefes (que son las que más invierten en comida) mientras que en el cuadrante superior e inferior derecho se agrupan las otras familias (empleados y trabajadores manuales). Este patrón también se presenta en análisis de factores puesto que existen dos factores que separan a las comidas esenciales (usualmente las más baratas) de las comidas de lujo (como el vino).

1.3 Relación con el análisis de gastos (P3 examen 1)

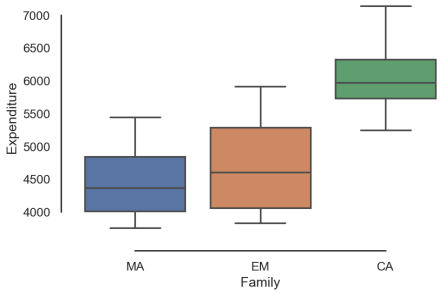


Figura 7: Distribución de gastos en comida por familia - pregunta 3 examen 1

Con base en la figura 7, podemos concluir que sí existe una relación entre el análisis de factores y esta figura puesto que seguimos observando dos grupos: gastos bajos en comida (familias de trabajadores manuales y empleados y comidas esenciales) y gastos altos en comida (familias de jefes y comidas de lujo).

2 Pregunta: Análisis de *Clusters*

Descripción: para los datos de arrestos en los estados de E.U.A., confirme (o establezca un contraejemplo) las afirmaciones en la columna 5 de la tabla 4.1 (página 9 de la nota "CA2.pdf")

Table 4.1 Standard agglomerative hierarchical clustering methods.

Method	Alternative name[a]	Usually used with:	Distance between clusters defined as:	Remarks
Single linkage Sneath (1957)	Nearest neighbour	Similarity or distance	Minimum distance between pair of objects, one in one cluster, one in the other	Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure.
Complete linkage Sorensen (1948)	Furthest neighbour	Similarity or distance	Maximum distance between pair of objects, one in one cluster, one in the other	Tends to find compact clusters with equal diameters (maximum distance between objects). Does not take account of cluster structure.
(Group) Average linkage Sokal and Michener (1958)	UPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	Tends to join clusters with small variances. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust.
Centroid linkage Sokal and Michener (1958)	UPGMC	Distance (requires raw data)	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged cluster. Subject to reversals.
Weighted average linkage McQuitty (1966)	WPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	As for UPGMA, but points in small clusters weighted more highly than points in large clusters (useful if cluster sizes are likely to be uneven).
Median linkage Gower (1967)	WPGMC	Distance (requires raw data)	Squared Euclidean distance between weighted centroids	Assumes points can be represented in Euclidean space for geometrical interpretation. New group is intermediate in position between merged groups. Subject to reversals.
Ward's method Ward (1963)	Minimum sum of squares	Distance (requires raw data)	Increase in sum of squares within clusters, after fusion, summed over all variables	Assumes points can be represented in Euclidean space for geometrical interpretation. Tends to find same-size, spherical clusters. Sensitive to outliers.

[a]U = unweighted; W = weighted; PG = pair group; A = average; C = centroid.

Figure 8: Tabla 4.1

2.1 Single Linkage

- **Análisis**: Se confirma la afirmación de producir *clusters* desbalanceados. Claramente podemos observar una tendencia hacía la izquierda. El estado 47 (de hasta la izquierda) solo está agrupado en un cluster con la raíz, mientras que otros estados de la derecha pertenecen a más clusters
- **Afirmación**: *"Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure."*
- **Dendrograma**:

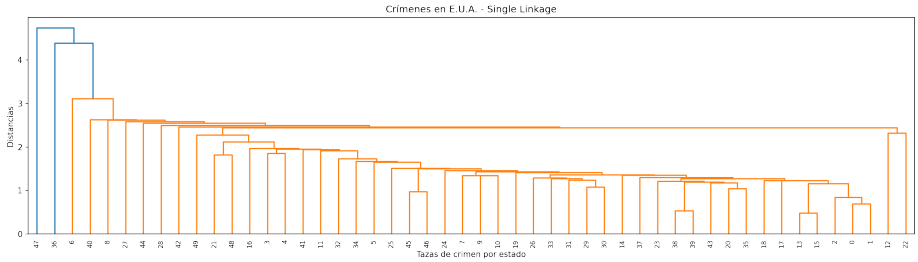


Figura 9: Single Linkage

2.2 Complete Linkage

- **Análisis**: a diferencia de *single linkage*, aquí podemos ver que las distancias entre *clusters* tiende a ser más balanceada. Se confirma la afirmación puesto que el nodo del dendrograma (representado por la línea azul) se encuentra justo en el centro de los estados mostrados
- **Afirmación**: *"Tends to find compact clusters with equal diameters (maximum distance between objects)."*
- **Dendrograma**:

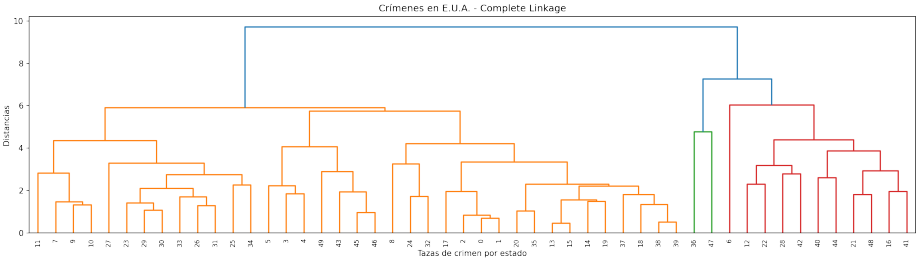


Figura 10: Complete Linkage

2.3 Group Average Linkage

- **Análisis**: se confirma la afirmación de que éste método es un intermedio entre *single linkage* y *complete linkage* puesto que podemos observar que loos primeros estados (36 y 47) están agrupados hasta la izqiuerda mientras que los demás estados sí tienen un diámetro similar y se podría decir que están balanceados
- **Afirmación**: *"Tends to join clusters with small variances. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust."*

- **Dendrograma**:

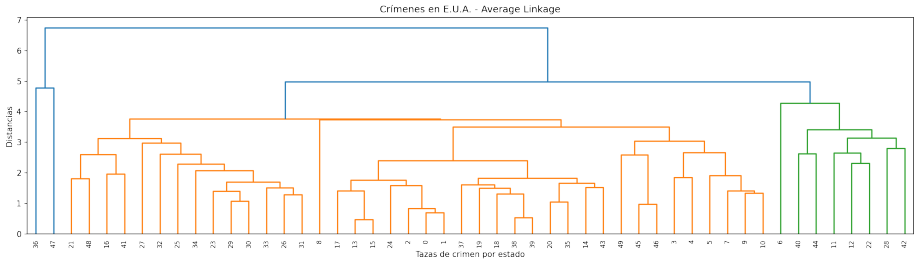


Figura 11: Average Linkage

2.4 Centroid Linkage

- **Análisis**: a diferencia de otros métodos, éste método no requiere de los datos estandarizados. Con base en esto, podemos observar que en efecto, el cluster más numeroso será el que predomine de los dos clusters asociados. Puntualmente, observamos que hace referencia al estado 36 (Texas) el cual es un estado con altos índices delictivos por la frontera con México
- **Afirmación**: *"Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged cluster. Subject to reversals."*
- **Dendrograma**:

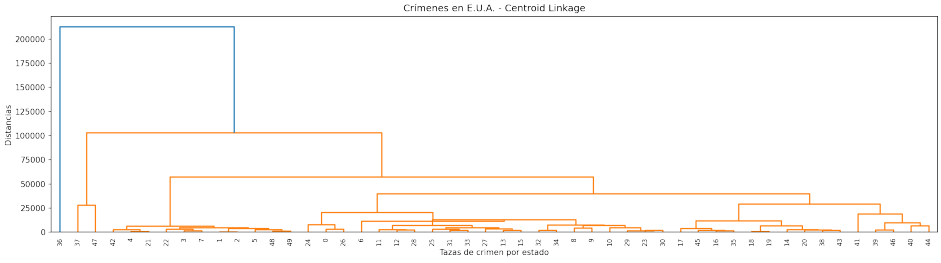


Figura 12: Centroid Linkage

2.5 Weighted Average Linkage

- **Análisis**: podemos confirmar la afirmación porque el resultado final de éste método es un dendrograma bastante balanceado a pesar de que los tamaños de los *clusters* no lo son (como se puede ver en *centroid linkage* con el estado de Texas - 36). De igual forma los *clusters* más pequeños (rojos y verdes) tienen una altura de cluster más grande que algunos *clusters*

más grandes (azul), es decir, no se nota la diferencia en tamaños de *cluster* como sí en otros métodos (centroid, median, Ward)

- **Afirmación**: *"As for UPGMA, but points in small clusters weighted more highly than points in large clusters (useful if cluster sizes are likely to be uneven)."*
- **Dendrograma**:

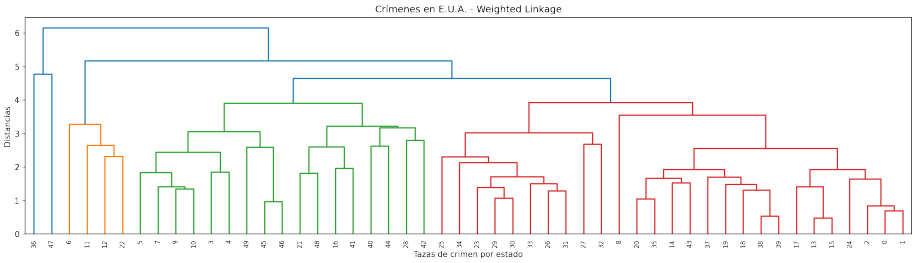


Figura 13: Weighted Average Linkage

2.6 Median Linkage

- **Análisis**: Se confirma la afirmación puesto que observamos que los nuevos *clusters* se posicionan en medio de los grupos combinados. En la figura 14 claramente podemos observar que los *clusters* están posicionados al medio a diferencia del método de *centroid linkage* donde los *clusters* tienen una tendencia hacía la izquierda
- **Afirmación**: *"Assumes points can be represented in Euclidean space for geometrical interpretation. New group is intermediate in position between merged groups. Subject to reversals."*
- **Dendrograma**:

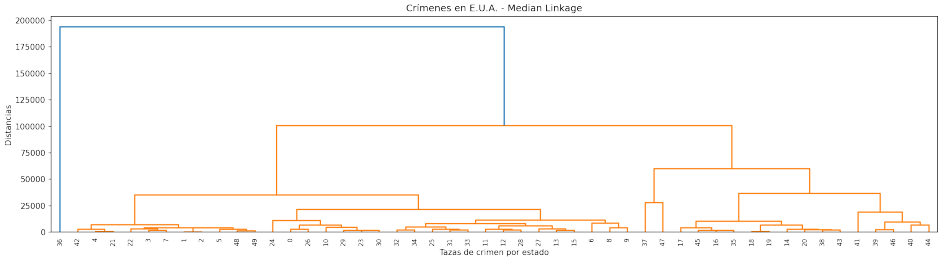


Figura 14: Median Linkage

2.7 Ward's Method

- **Análisis**: la afirmación nos dice que éste método tiende a encontrar *clusters* del mismo tamaño. Esta afirmación se confirma puesto que si analizamos el estado 36 (Texas), podemos observar que a diferencia de *centroid* y *median linkage*, Texas ya no se encuentra aislado, ahora pertenece a los *clusters* de color verde de la derecha
- **Afirmación**: *"Assumes points can be represented in Euclidean space for geometrical interpretation. Tends to find same-size, spherical clusters. Sensitive to outliers."*
- **Dendrograma**:

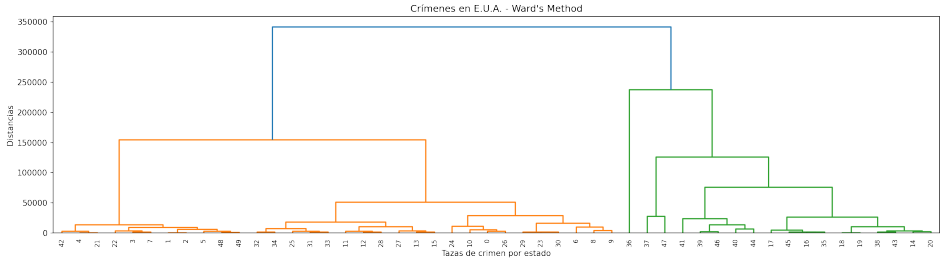


Figura 15: Ward's Method