ESTUDIO DE LA AGRUPACIÓN SOCIOECONÓMICA DE LAS COMUNIDADES AUTÓNOMAS DE ESPAÑA

MINERÍA DE DATOS EN NEGOCIOS

Ágatha del Olmo Tirado | 2ºBIA | 28/10/2023

VNIVERSITAT DE VALÈNCIA

INTELIGENCIA Y ANALÍTICA DE NEGOCIOS

ÍNDICE

INTRODUCCIÓN

Este informe presenta un análisis detallado de la agrupación de Comunidades Autónomas españolas, basado en datos socioeconómicos recopilados en el archivo "comaut.csv". El objetivo de este estudio es utilizar técnicas de clustering tanto directas como jerárquicas para agrupar las comunidades en grupos homogéneos pero heterogéneros entre sí, lo que nos permitirá comprender mejor las similitudes y diferencias entre las diversas regiones españolas.

Hemos empleado las técnicas directas “farthest first” y “canopy” a través del programa Weka, y las jerárquicas “encadenamiento simple”, “encadenamiento completo”, “Ward” y “centroide” en RStudio, así como un agrupamiento de tipo red “SOM” (Self-Organised Map).

Hemos realizado un análisis detallado de las variables socioeconómicas que podrían estar influyendo en la formación de grupos, y hemos explorado las diferencias en las medias de las variables entre los grupos. Tras identificar las medias más diferenciadoras a primera vista, hemos realizado un análisis de la varianza (ANOVA) para identificar diferencias significativas entre los grupos en las variables más relevantes, y un análisis múltiple de estas para poder interpretar más eficazmente los resultados.

Los resultados y las conclusiones derivados de este análisis ofrecen una comprensión más profunda de la diversidad dentro del país y pueden ser de utilidad para la toma de decisiones y la planificación estratégica de la zona.

Para el seguimiento del trabajo recomendamos ver el Link al script de Rstudio.

SELECCIÓN Y MODIFICACIÓN DE VARIABLES

Un paso esencial previo es la selección correcta de variables interesantes para el proyecto y la modificación de estas a través del programa Rstudio.

Es importante destacar que no se encuentran en la base de datos las ciudades autonómas de Ceuta y Melilla ya que no consideramos que aún relativizándolas, lo que eliminaría el efecto de ser tan pequeñas en relación con el resto, sean lo suficientemente interesantes para el estudio teniendo en cuenta su situación tan extraordinaria.

Primero establecimos una semilla -***set.seed(16)***- (para garantizar la reproducibilidad de los datos (especialmente teniendo en cuenta que algunas agrupaciones tienen rasgos aleatorios que después comentaremos).

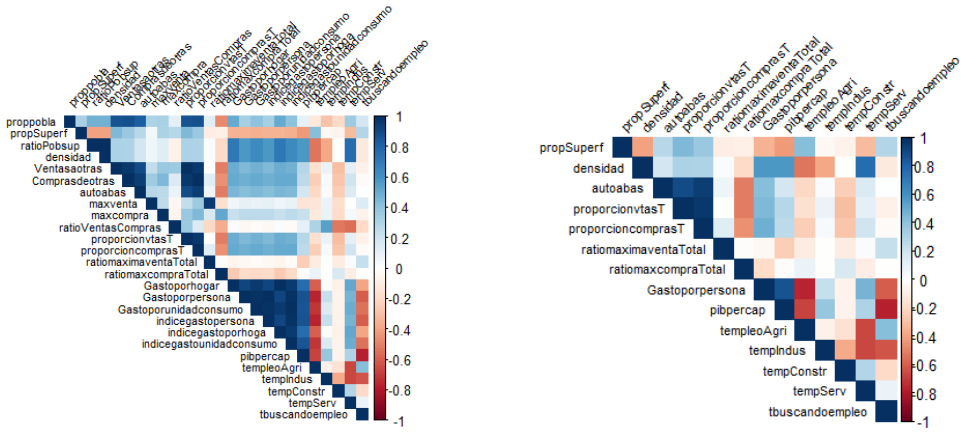
Posteriormente cargamos las bibliotecas que necesitaremos para el estudio, que en nuestro caso son ***haven*** para la importación y exportación de ficheros, ***kohonen*** para la posterior creación del SOM, ***dplyr*** para la manipulación de los datos, ***corrplot*** para la visualización de las correlaciones, ***DescTools*** para el análisis de medias, ***gridExtra*** para la visualización de plots y Ggplot2 para la creación de plots complejos.

Ya cargadas las bibliotecas que nos interesan, cargamos los datos en R con la instrucción ***datos <- read.csv2("./comaut.csv", dec = ",")*** (importante indicar en read.csv2 los decimales, porque en Excel se abrió con decimales por comas y R solo los entiende por puntos). Después se ordenó el conjunto de datos por orden alfabético según la variable Cautonoma, que contiene el nombre de cada comunidad autónoma ***datos <- datos %>% arrange(Cautonoma)***.

Con todos los datos cargados en R ya podemos realizar un estudio de las variables del dataframe “datos”. Dos pasos que consideramos esenciales son la eliminación de variables redundantes y la selección o modificación para que tengan su rango de valores entre 0 y 1 y no entre -1 y 1 porque su comportamiento es “mejor”. Cabe destacar podríamos usar para “descorrelacionar” y tipificar desde un principio la distancia euclídea aplicada a las componentes principales o la distancia de Mahalanobis, e incluso tras eliminar redundancias podríamos usar directamente la distancia normalizada, pero

utilizaremos la distancia euclídea, de forma que debemos eliminar altas correlaciones y tipificar.

Para ver la correlación entre variables primero configuramos los datos de forma que la variable "Cautonoma" pase a ser el nombre de las filas del dataframe con la instrucción ***row.names(datos)=datos$Cautonoma*** ya que necesitamos que todas las variables sean numéricas, además guardamos la variable como un objeto ***cautonoma*** <- ***datos$Cautonoma*** para poder eliminarla ***datos$Cautonoma*** <- ***NULL*** pero poder usarla en un futuro. Ahora que tenemos el formato de base de datos necesario podemos crear una matriz de correlaciones con ***correlation_matrix*** <- ***cor(datos)*** y visualizarla a través de ***corrplot(correlation_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45)***. Así podemos ver claramente los valores en la matriz de correlaciones representados por un mapa de calor, como vemos en la primera imagen.



Las variables con alta correlación y que consideramos menos interesantes para el estudio son ratioPobsup, proppobla, Ventasaotras, Comprasdeotras, maxventa, maxcompra, ratioVentasCompras, Gastoporhogar, Gastoporunidadconsumo, indicegastopersona, indicegastoporhoga, indicegastounidadconsumo, y la matriz de correlaciones queda tal como vemos en la segunda imagen.

Tras eliminar redundancias evitando multicolinealidad, debemos relativizar las variables ya que según su escala unas tienen más o menos peso, (lo hacemos para todas las variables que no sean proporciones, tasas o ratios, es decir, autoabas, Gastoporpersona, pibpercap y densidad) asando la instrucción siguiente ***datos_rel$variable<-datos$variable/sum(datos$variable)***.

Ahora, como estamos usando distancia euclídea, tipificamos las variables a media=0 y varianza=1, con la instrucción ***datos_tipificados*** <- ***as.data.frame(scale(datos_rel))***, y podemos comprobar que ha funcionado correctamente con la media y la varianza ***mean(datos_tipificados$variable)*** y ***var(datos_tipificados$variable)***, que deben dar respectivamente 0 y 1.

De esta forma ya hemos preparado el conjunto de datos de forma que se pueda llevar a cabo correctamente el análisis de agrupación, así que guardamos la nueva base de datos en un fichero csv para su uso en Weka. Para esto, primero volvemos a añadir la variable

cautonoma que antes hemos guardado como objeto en el environment ***datos_save$Cautonoma <- cautonoma*** y lo guardamos en nuestro directorio usando la instrucción ***write.csv(datos_save, file ="comaut_mod.csv", row.names = FALSE)***.

EL PROCESO DE CLUSTERING EN RSTUDIO

TAREAS PREVIAS

Primero, realizamos la red del SOM, y después, creamos las agrupaciones jeráquicas (Complete, Simple, Ward y Centroide), ya que así podremos elegir el número de clústeres más adecuado, y realizamos entonces el clustering del SOM.

Con el conjunto de datos ya modificado,creamos la matriz modelo que llamamos Z sin incluir ni "Cautonoma" ni el intercepto (para esto usamos -1 al inicio). ***Z=model.matrix(~-1+densidad+propSuperf+autoabas+proporcionvtasT+proporcioncomprasT+ratiomaximaventaTotal+ratiomaxcompraTotal+Gastoporpersona+pibpercap+templeoAgri+tempIndus+tempConstr+tempServ+tbuscandoempleo, datos_tipificados).***

Como en la base de datos datos_tipificados las filas eran las CCAA, en la matriz también lo son así que no hace falta indicárselo.

CREACIÓN DEL SOM

Para crear el mapa auto-organizativo, creamos el conjunto de entrenamiento, ***data_train <- datos_tipificados[, c(1:14)]***, cogiendo la totalidad de los datos que tenemos, y lo convertimos a matriz ***data_train_matrix <- as.matrix(data_train)***. Después, indicamos el número de perfiles que queremos obtener, que en nuestro caso son 9, con un 3x3 en forma hexagonal para que haya más vecinos por perfil ***som_grid<-somgrid(xdim = 3, ydim=3, topo="hexagonal")***. De esta forma ya podemos realizar el propio SOM con la instrucción

```
som_model <- som(data_train_matrix,
        grid=som_grid,
        rlen=1000,
        alpha=c(0.05,0.01),
        keep.data = TRUE )
```

Podemos analizar diferentes métricas de la red. Para empezar, podemos ver en qué perfil cayó cada individuo, utilizando "unit.classif", que nos dice en orden el número del perfil, podemos hacer un for que nos de la información de forma más compacta:

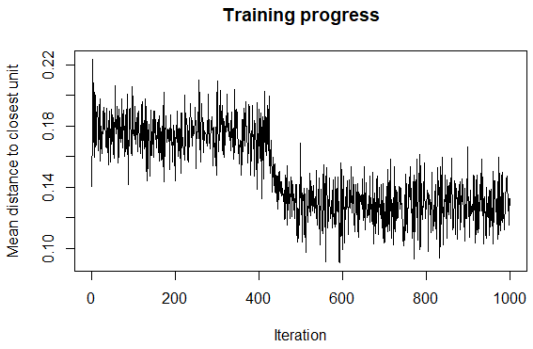
```
for (i in 1:length(unique(som_model$unit.classif))) {
```

```
 perfil <- unique(som_model$unit.classif)[i]

 cat("Perfil:", perfil, "\n")

 cat("Comunidades Autónomas:", cautonoma[som_model$unit.classif == perfil],
"\n\n")

}
```

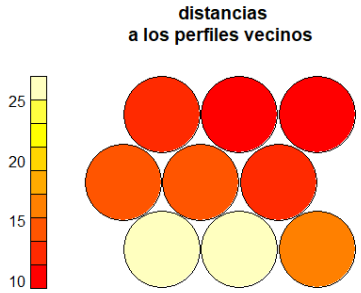
El output es el siguiente:

Perfil: 3
Comunidades Autónomas: Andalucía Castilla y León
Perfil: 7
Comunidades Autónomas: Aragón Cantabria Illes Balears
Perfil: 9
Comunidades Autónomas: Canarias Extremadura Región de Murcia
Perfil: 6
Comunidades Autónomas: Castilla-La Mancha
Perfil: 1
Comunidades Autónomas: Cataluña Comunidad de Madrid
Perfil: 4
Comunidades Autónomas: Comunidad Foral De Navarra País Vasco
Perfil: 2
Comunidades Autónomas: Comunidad Valenciana
Perfil: 5
Comunidades Autónomas: Galicia
Perfil: 8
Comunidades Autónomas: La Rioja Principado De Asturias

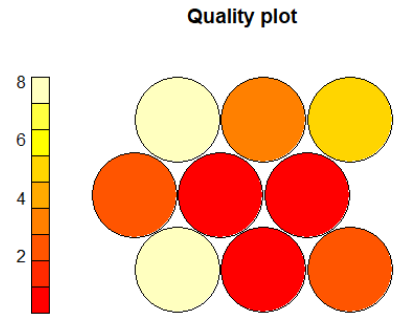
También podemos ver el progreso del entrenamiento de la red con la instrucción ***plot(som_model, type="changes")***. El eje X representa las iteraciones durante el entrenamiento, mientras que el eje Y muestra la distancia media al vector del libro de códigos (codebook) más cercano.Este plot representa la distancia media entre cada dato de entrada y su vector de pesos correspondiente. En la red, cada vector de pesos se ajusta durante el proceso de entrenamiento para acercarse a los datos de entrada, por lo tanto, a medida que aumenta el número de iteraciones, esperaríamos que la distancia en conjunto disminuya. Aquí vemos una caída muy grande a partir de las 400 iteraciones, pero a partir de ahí llega a un mínimo del que no podrá bajar porque la red no consigue mejorarse más, en este caso parece ser alrededor del valor 0.12.



También podemos ver la suma de distancias entre perfiles a los vecinos inmediatos con la instrucción ***plot(som_model, type="dist.neighbours", main = " distancias a los***



perfiles vecinos"). La red no solo es competitiva en el sentido de que solo "una neurona gana", sino también cooperativa, ya que cada vez que un perfil actualiza sus pesos, también lo hacen sus perfiles vecinos, por eso es interesante usar la forma hexagonal, que maximiza el número de vecinos. En este plot podemos ver que los perfiles 6,7 y 8 (de izquierda abajo a derecha arriba) son los más parecidos entre sí, siendo el 1 y 2 los

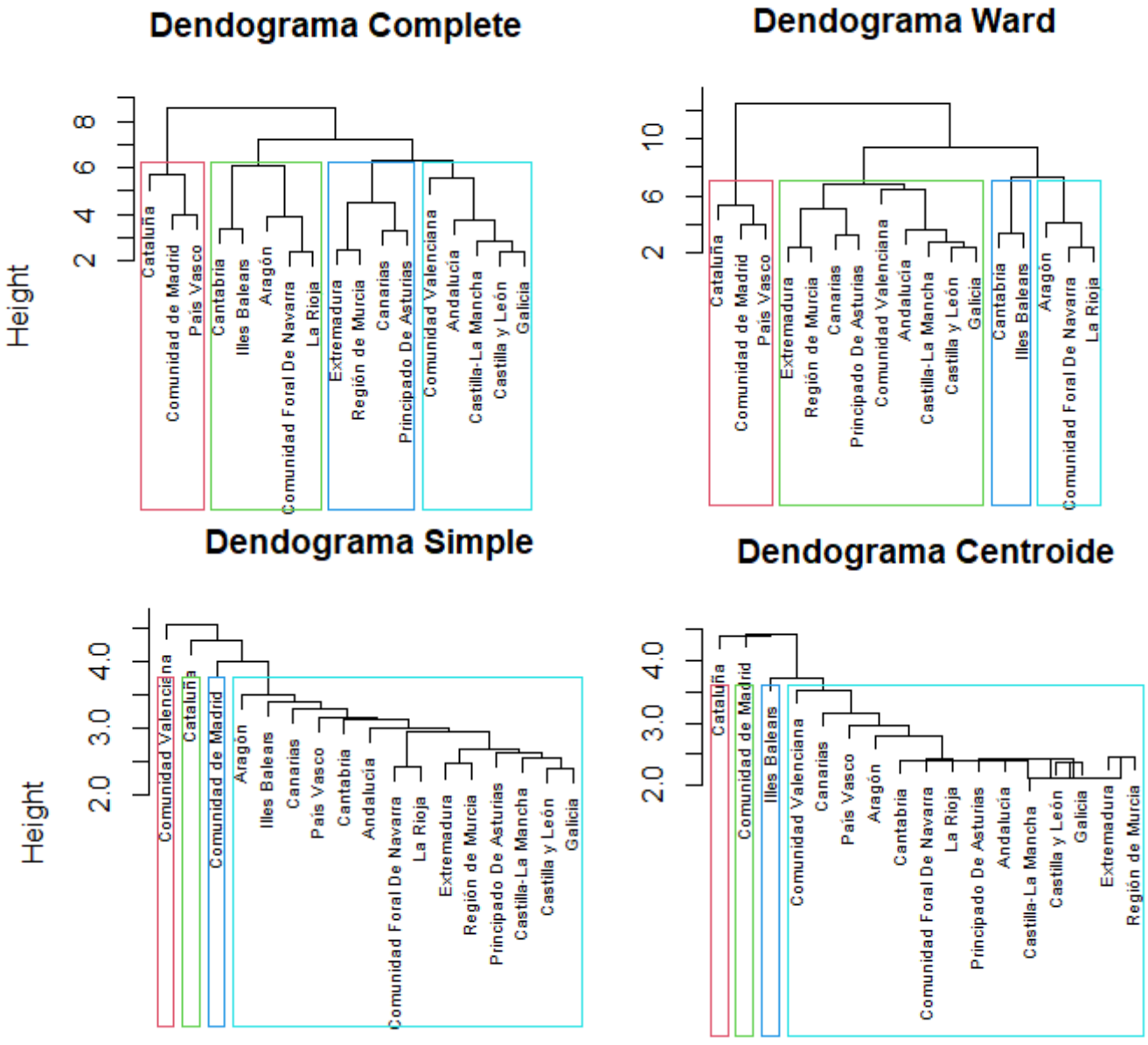


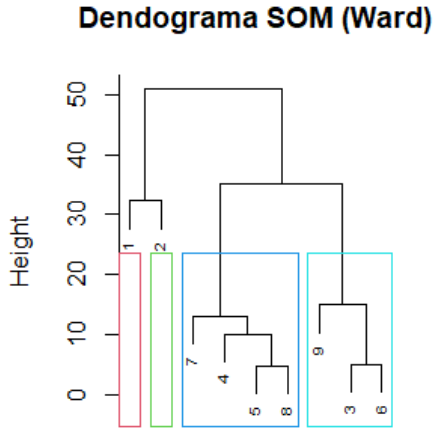
Además, podemos ver "quality", que nos muestra la distancia media de las instancias asignadas al nodo a cada perfil. Cuanto más pequeñas estas distancias mejor representadas estarán las instancias por el perfil al que se le ha asignado. Como vemos, los nodos que más se parecen a sus instancias son el 2, el 5 y el 6. (Se puede volver a la página anterior a ver qué CCAA hay en cada uno de los nodos).

Ya analizada la red y sus características, podemos agrupar los jerárquicos y el SOM y elegir un número de clústeres.

CONFIGURACIÓN DE LA AGRUPACIÓN

Los clusterers jerárquicos que vamos a realizar son los siguientes:



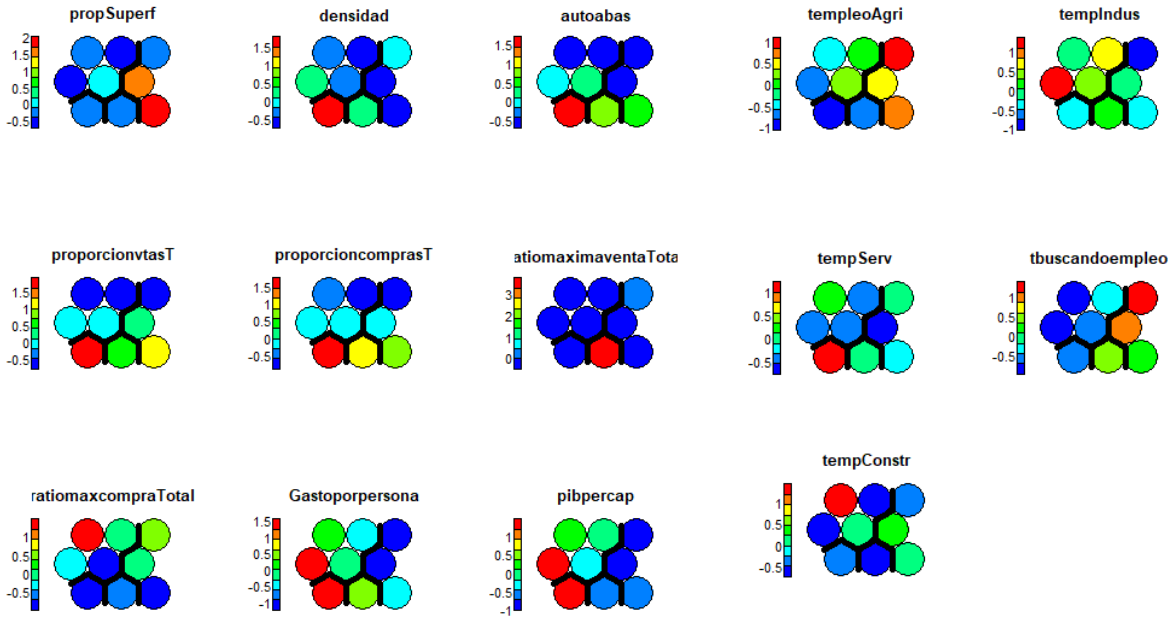


Para todos usamos la instrucción ***clusX<-hclust(dist(Z), method= "X")***, y luego ***plot(clusX, cex = 0.6)*** para verlo y poder elegir al final una cardinalidad de clusters.

El eje Y representa la pérdida de homogeneidad interna, y no nos interesa ver saltos muy grandes, por esto mismo creemos que separar los clusters en 4, ya que no pierde demasiada homogeneidad aún reduciendo en mucho los grupos. Para ver claramente cómo se separan añadimos ***rect.hclust(clusJCe, k = 4, border = 2:5)***.

Ya elegida la cardinalidad, podemos ejecutar clusterX <- cutree(clusX, k = 4) para todos, y podemos analizar qué caracteriza a cada grupo en el SOM con las instrucciones:

```
par(mfrow = c(3, 3))

plot_list <- list()

for (i in 1:14) {

  current_plot <- plot(som_model, type = "property", property =
getCodes(som_model)[, i],

                       main = colnames(getCodes(som_model))[i], palette.name =
coolBlueHotRed)

  add.cluster.boundaries(som_model, clusterSOM)

  plot_list[[length(plot_list) + 1]] <- current_plot

}
```



Como podemos ver en los gráficos, los grupos más diferenciadores en las variables son:
-**propSuperf**: El Grupo 4, con la excepción del Perfil 9, exhibe los valores más elevados
-**densidad**: El Perfil 1 supera significativamente a los demás en términos de densidad
-**autoabas**: De nuevo, el Perfil 1 se destaca como el más alto
-**proporcionvtasT** y **proporcioncomprasT**: Perfiles 1 y 2 tienen los valores más altos
-**ratiomaximaventaTotal**: En esta variable, el Perfil 2 es el más destacado
-**ratiomaxcompraTotal**: Los Clústeres 1 y 2 muestran los valores más bajos
-**gastoporpersona**: El Clúster 3 registra el valor más bajo
-**pibpercap**: Los Clústeres 1 y 4 presentan los valores más altos
-**templeoAgri, tempConstr y Tbuscandoempleo**: El Clúster 3 supera a todos los de más en estas variables
-**tempServ**, el primer perfil es el más alto por mucha diferencia

EL PROCESO DE CLUSTERING EN WEKA

Pasamos a usar Weka, entorno gráfico que nos permite crear y analizar experimentos sobre tareas de clustering, y en concreto su interfaz Explorer. Aquí realizamos las agrupaciones con los métodos directos FarthestFirst y Canopy.

TAREAS PREVIAS

El punto de partida es el archivo " comaut_mod.csv", que acabamos de guardar y contiene información socioeconómica de las 17 comunidades autónomas en el área metropolitana valenciana. Al trabajar con este archivo en formato CSV, es necesario realizar ciertos ajustes iniciales. Para ello, abrimos la ventana de diálogo "Invoke Options Dialog" y cambiamos el separador de campo de "," a ";" para asegurarnos de que los datos se importen correctamente. Dado que estamos buscando agrupar las comunidades y no clasificarlas, configuramos la clase en "No Class" para que no se incluya un atributo de clase.

CONFIGURACIÓN DE LAS AGRUPACIONES

La forma de agrupar de las agrupaciones que hemos seleccionado son las siguientes:

-FarthestFirst: escoge aleatoriamente un elemento de los datos como primer centroide y calcula su distancia con el resto. El punto más alejado pasa a ser el nuevo controide del conjunto de datos y repite el cálculo de distancias, y va realizando asignaciones hasta tener las agrupaciones necesarias.

-Canopy: se fija de manera aleatoria el primer centroide y se generan grupos provisionales que llamamos canopies según unos umbrales, y los elementos que no estén en ningún canopy en común son candidatos a centroides. Después, se realiza una clasificación del algoritmo K-means, pero no se calcula la distancia entre los elementos del mismo canopy.

Para añadir las agrupaciones entramos en Preprocess y seleccionamos el filtro dentro de “unsupervised” ya que el clustering es no supervisado (no tiene una variable de respuesta ya que no existen clases previas en las que clasificar), y dentro de “attribute” seleccionamos “AddCluster”, y pinchamos con el botón derecho encima entrando en “Show properties”. Ahí, elegimos el clusterer Canopy, seleccionamos el atributo a ignorar que es Cautonoma y dándole clic derecho encima del clusterer ponemos a 4 el número de clústeres. Después de aplicarlo, usamos el filtro “RenameAttribute” y cambiamos este por clusterCanopy para luego poder identificarlo correctamente y para permitir crear la siguiente agrupación. Repetimos el proceso para FarthestFirst, pero esta vez ignoramos también el clúster que acabamos de realizar, y lo renombramos a clusterFF.

Ya realizadas las agrupaciones podemos guardar la base de datos con save como comaut_directos.csv para poder importarla a R con la instrucción de abajo.

datos_dir <- read.delim("./comaut_directos.csv", sep=",")

ANÁLISIS DE LAS AGRUPACIONES

ELEMENTOS DE CADA AGRUPACIÓN

Analizamos los elementos de las agrupaciones directas ya que las jerárquicas las tenemos en los dendogramas de antes.

FARTHEST FIRST:

GRUPO 1: Comunidad Valenciana, Galicia, Aragón, Cantabria, Comunidad Foral de Navarra, Illes Balears, La Rioja, Canarias, Extremadura, principado de Asturias y Región de Murcia.

GRUPO 2: Cataluña

GRUPO 3: Comunidad de Madrid y País Vasco

GRUPO 4: Andalucía, Castilla y León y Castilla-La Mancha

CANOPY

GRUPO 1: Comunidad Valenciana, Galicia, Aragón, illes Balears, Principado de Asturias, Región de Murcia, País Vasco y Castilla y León

GRUPO 2: Canarias, Extremadura, Andalucía y Castilla-La Mancha

GRUPO 3: Cataluña y Comunidad de Madrid

GRUPO 4: Comunidad Foral de Navarra y La Rioja

Como podemos ver, lo más característico es la marginación de Cataluña, Madrid, el País Vasco y la Comunidad Valenciana, así que tendremos en cuenta al analizar sus medias que estaremos hablando de unos grupos concretos y lo suficientemente diferentes o extraños como para que se hayan exiliado en diferentes grupos.

ANÁLISIS DE LAS MEDIAS

Para poder estudiar qué diferencia a cada grupo, miramos las medias y detectamos las que tengan una diferencia mayor entre grupos dentro de una misma agrupación. Para esto usamos la función aggregate() con el modelo previo al que llamamos Z, el modelo correspondiente y con FUN=mean. Al analizarlas, detectamos diferencias en las medias de las siguientes variables:

-Jerárquico simple: densidad, autoabas, proporcionventasT, proporcioncomprasT, ratiomaxventatotal, ratiomaximacompraTotal, gastoporpersona, pibpercap, tempServ

-Jerárquico complete: densidad, autoabas, proporcionventasT, proporcioncomprasT, ratiomaximacompraTotal, gastoporpersona, pibpercap, tbuscandoempleo

-Jerárquico Ward: densidad, autoabas, proporcionventasT, proporcioncomprasT, ratiomaximacompraTotal, gastoporpersona, pibpercap, templeoAgri, tempIndus, tempConstr, tempServ, tbuscandoempleo

-Jerárquico Centroide: densidad, autoabas, proporcionventasT y proporcioncomprasT, ratiomaximacompraTotal, pibpercap, templeoAgri, tempIndus, tempconstr y tempServ.

-Canopy: densidad, autoabas, proporcionventasT, proporcioncomprasT, gastoporpersona, pibpercap, templeoagri, tempindus, tbuscandoempleo

-FarthestFirst: densidad, autoabas, proporcionventasT, proporcioncomprasT, ratiomaximacompraTotal, gastoporpersona, pibpercap, templeoAgri, tempServ, tbuscandoempleo

-SOM: densidad, autoabas, proporcionventasT, proporcioncomprasT, ratiomaxventatotal, gastoporpersona, pibpercap, templeoAgri, tempServ, tbuscandoempleo

ANÁLISIS ANOVA

El análisis de la varianza (ANOVA) es una técnica estadística utilizada para evaluar si existen diferencias significativas entre los grupos en función de una variable o factor.

En concreto, en este caso, detectaremos diferencias significativas entre los grupos de CCAA españolas en las variables en las que hemos encontrado aparentes diferencias entre las medias de los grupos (apartado anterior) con la función *aov(variable-clúster)*.Para hacerlo, antes hemos convertido las variables de los clústeres a factor con *as.factor()* porque en este caso se refiere a categorías discretas que representan distintos grupos, y no tiene un significado numérico continuo. Ahora ANOVA lo tratará como niveles y no como variables con una interpretación numérica. Los resultados han sido, en resumen, los siguientes:

Variable \ Método	Jer. Centroide	Jer. Complete	Jer. Simple	Jer. Ward	Jer. Canopy	FarthestFirst	SOM
Densidad	Todos los niveles	0.01	Cualquier nivel	0.01	0.01	0.001	0.01
Autoabas	0.001	0.01	Cualquier nivel	0.05	0.05	Cualquier nivel	0.05
ProporcionventasT	0.001	Cualquier nivel	0.001	0.01	0.001	Cualquier nivel	0.01
ProporcioncomprasT	0.01	0.001	0.001	0.01	0.01	Cualquier nivel	0.01
RatiomaxcompraTotal	-	0.01	Cualquier nivel	0.05	-	-	Cualquier nivel
Gastoporpersona	-	0.001	-	0.001	0.001	0.01	0.001
Pibpercap	-	Cualquier nivel	-	Cualquier nivel	Cualquier nivel	0.001	0.001
TempleoAgri	-	0.001	-	0.05	-	-	0.01
TempIndus	-	0.01	-	0.01	0.01	-	No
TempConstr	0.01	Cualquier nivel	Cualquier nivel	Cualquier nivel	Cualquier nivel	Cualquier nivel	-
TempServ	-	0.05	-	0.05	No	No	No
Tbuscandoempleo	-	0.001	-	0.001	Cualquier nivel	-	0.001

COMPARACIONES MÚLTIPLES

Para acabar, realizamos el análisis de Scheffe, que es el más duro al encontrar diferencias significativas, con la función ScheffeTest(anova) sobre las variables en las que hemos detectado diferencias significativas. Este análisis es útil cuando tienes más de dos grupos y deseas determinar cuáles de ellos son significativamente diferentes entre sí, cosa que con el ANOVA no podemos conseguir. De esta forma podremos entender qué diferencia a los grupos en cada agrupación realmente.

JERÁRQUICO CENTROIDE

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 3	6.3e-05	Grupo 3 > Grupo 1
Densidad	2 vs 3	0.0061	Grupo 3 > Grupo 2
Densidad	3 vs 4	0.0054	Grupo 3 < Grupo 4
Autoabas	1 vs 2	0.0022	Grupo 2 > Grupo 1
Autoabas	2 vs 3	0.0317	Grupo 3 < Grupo 2
Autoabas	2 vs 4	0.0084	Grupo 4 < Grupo 2
ProporcionvtasT	1 vs 2	0.0190	Grupo 2 > Grupo 1
ProporcionvtasT	2 vs 4	0.0181	Grupo 4 < Grupo 2
ProporcioncomprasT	1 vs 2	0.0264	Grupo 2 > Grupo 1
ProporcioncomprasT	2 vs 4	0.0426	Grupo 4 < Grupo 2
TempConstr	1 vs 4	0.0129	Grupo 4 > Grupo 1
TempConstr	2 vs 4	0.0447	Grupo 4 > Grupo 2
TempConstr	3 vs 4	0.0773	Grupo 4 > Grupo 3
TempServ	1 vs 3	0.0501	Grupo 3 > Grupo 1
TempServ	1 vs 4	0.0930	Grupo 4 > Grupo 1

Grupo 1: Este grupo tiene una densidad y un nivel de autoabastecimiento (Autobaas) más bajos en comparación con otros grupos, lo que podría indicar una menor concentración de población y una menor capacidad de producción interna. Además, sus índices de compra y venta (ProporcionvtaasT y ProporcioncomprasT) son más bajos en comparación con el grupo 2, lo que podría sugerir una capacidad económica más baja.

Grupo 2: Este grupo tiene una densidad más baja que el grupo 3, pero un nivel de autoabastecimiento más alto que el grupo 1, lo que podría indicar una mayor capacidad de producción interna. Sus índices de compra y venta son más altos que los del grupo 1, lo que podría indicar una mayor capacidad económica. Sin embargo, estos índices son más bajos en comparación con el grupo 4, lo que sugiere que su capacidad económica es menor que la del grupo 4.

Grupo 3: Este grupo tiene la mayor densidad en comparación con los grupos 1 y 2, lo que podría indicar una alta concentración de población. Su nivel de autoabastecimiento es más alto que el del grupo 2, lo que podría indicar una alta capacidad de producción interna.

Grupo 4: Este grupo tiene la mayor densidad en comparación con el grupo 3, lo que podría indicar una muy alta concentración de población. Su nivel de autoabastecimiento es más bajo que el del grupo 2, pero sus índices de compra y venta son más altos, lo que podría indicar una alta capacidad económica.

JERÁRQUICO COMPLETE

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 4	0.0399	Grupo 4 > Grupo 1
Densidad	2 vs 4	0.0450	Grupo 4 > Grupo 2
Densidad	3 vs 4	0.0961	Grupo 4 > Grupo 3
Autoabas	4 vs 2	0.0426	Grupo 4 > Grupo 2
Proporción Ventas T	1 vs 2	0.0320	Grupo 1 > Grupo 2
Proporción Ventas T	1 vs 3	0.0381	Grupo 1 > Grupo 3
Proporción Ventas T	2 vs 4	0.0034	Grupo 4 > Grupo 2
Proporción Ventas T	3 vs 4	0.0042	Grupo 4 > Grupo 3
Proporción Compras T	1 vs 2	0.0731	Grupo 1 > Grupo 2
Proporción Compras T	1 vs 3	0.0826	Grupo 1 > Grupo 3
Proporción Compras T	2 vs 4	0.0068	Grupo 4 > Grupo 2
Proporción Compras T	3 vs 4	0.0082	Grupo 4 > Grupo 3
Ratio Max Compra Total	1 vs 2	0.0690	Grupo 1 > Grupo 2
Ratio Max Compra Total	2 vs 4	0.0982	Grupo 4 > Grupo 2
Gasto por Persona	1 vs 4	0.0192	Grupo 4 > Grupo 1
Gasto por Persona	2 vs 4	0.1187	Grupo 4 > Grupo 2
Gasto por Persona	3 vs 4	0.0082	Grupo 4 > Grupo 3
PIB per Capita	1 vs 4	0.00015	Grupo 4 > Grupo 1
PIB per Capita	3 vs 4	0.00011	Grupo 4 > Grupo 3
T Buscando Empleo	1 vs 2	0.0537	Grupo 1 > Grupo 2
T Buscando Empleo	2 vs 3	0.0034	Grupo 3 > Grupo 2
T Buscando Empleo	2 vs 4	0.9628	Grupo 4 > Grupo 2
T Buscando Empleo	3 vs 4	0.0212	Grupo 4 > Grupo 3

Grupo 1: Este grupo se caracteriza por tener una densidad significativamente menor en comparación con el Grupo 4, sugiriendo una distribución espacial diferente de las variables. Además, presenta una proporción de ventas totales significativamente mayor que el Grupo 2 y el Grupo 3, indicando que las ventas desempeñan un papel más

destacado en su actividad económica. Asimismo, exhibe un gasto por persona y un PIB per cápita significativamente superiores en comparación con el Grupo 4, lo que sugiere una mayor capacidad económica.

Grupo 2: En contraste, este grupo no muestra diferencias significativas en densidad ni en proporciones de compras o ventas en comparación con otros grupos. Sin embargo, destaca por una tasa buscando empleo significativamente menor que el Grupo 3. El gasto por persona y el PIB per cápita no difieren significativamente de otros grupos.

Grupo 3: El Grupo 3 se caracteriza por una densidad significativamente mayor en comparación con el Grupo 4, pero menor proporción de ventas totales, gasto por persona y PIB per cápita. Además, presenta una tasa buscando empleo significativamente mayor que el Grupo 2 y el Grupo 4.

Grupo 4: Este grupo destaca por tener la mayor densidad y un autoabastecimiento significativamente mayor que el Grupo 2. Aunque muestra proporciones de ventas y compras totales menores en comparación con el Grupo 1 y el Grupo 2, presenta un gasto por persona y un PIB per cápita significativamente mayores que el Grupo 2 y menores que el Grupo 1 y el Grupo 3. La tasa de búsqueda de empleo es significativamente menor que el Grupo 3 y mayor que el Grupo 2.

JERÁRQUICO SIMPLE

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 4	7.4e-05	Grupo 4 > Grupo 1
Densidad	2 vs 4	0.0068	Grupo 4 > Grupo 2
Densidad	3 vs 4	0.0050	Grupo 4 > Grupo 3
Autoabas	1 vs 2	0.0009	Grupo 2 > Grupo 1
Autoabas	1 vs 4	0.0199	Grupo 4 > Grupo 1
Autoabas	2 vs 4	0.0199	Grupo 4 > Grupo 2
Proporción Ventas	1 vs 2	0.0175	Grupo 2 > Grupo 1
Proporción Compras	1 vs 2	0.0117	Grupo 2 > Grupo 1
Ratio Máxima Venta Total	1 vs 3	1.2e-08	Grupo 3 > Grupo 1
Ratio Máxima Venta Total	2 vs 3	1.9e-07	Grupo 3 > Grupo 2
Ratio Máxima Venta Total	3 vs 4	2.4e-07	Grupo 3 > Grupo 4

Grupo 1: Este grupo tiene una densidad más baja en comparación con el grupo 4, lo que podría indicar una menor concentración de población. Su nivel de autoabastecimiento también es más bajo que los grupos 2 y 4, lo que podría reflejar una menor capacidad de producción interna. Además, sus proporciones de ventas y compras son menores que el grupo 2.

Grupo 2: Aunque este grupo tiene una densidad más baja en comparación con los grupos 3 y 4, su nivel de autoabastecimiento es superior al del grupo 1, indicando una mayor capacidad de producción interna. Sus proporciones de ventas y compras son superiores a las del Grupo 1 pero no se comparan con los Grupos 3 o 4.

Grupo 3: Este grupo se destaca por tener la mayor ratio máxima total, superando a los grupos 2 y 4 en este aspecto. Aunque su densidad es inferior a la del Grupo 4, sigue siendo superior a la del Grupo 2. Esto podría indicar un equilibrio entre la concentración poblacional y la eficiencia económica.

Grupo 4: Con la mayor densidad entre todos los grupos, este indica una alta concentración poblacional. Aunque su nivel de autoabastecimiento es inferior al del Grupo2 , sus proporciones en ventas y compras no se ven afectadas negativamente, manteniendo un equilibrio económico sólido.

JERÁQUICO WARD

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 4	0.0288	Grupo 4 > Grupo 1
Densidad	2 vs 4	0.0373	Grupo 4 > Grupo 2
Densidad	3 vs 4	7.4e-05	Grupo 4 > Grupo 3
Autoabas	1 vs 4	0.2547	Grupo 4 > Grupo 1
Autoabas	2 vs 4	0.1425	Grupo 4 > Grupo 2
Autoabas	3 vs 4	0.1603	Grupo 4 > Grupo 3
Proporción Ventas	1 vs 4	0.1310	Grupo 4 > Grupo 1
Proporción Ventas	2 vs 4	0.0385	Grupo 4 > Grupo 2
Proporción Ventas	3 vs 4	0.0639	Grupo 4 > Grupo 3
Proporción Compras	1 vs 4	0.1351	Grupo 4 > Grupo 1
Proporción Compras	2 vs 4	0.0639	Grupo 4 > Grupo 2
Gasto por Persona	1 vs 4	0.0069	Grupo 4 > Grupo 1
PIB per Cápita	1 vs 2	0.0011	Grupo 2 > Grupo 1
PIB per Cápita	1 vs 4	1.6e-05	Grupo 4 > Grupo 1
PIB per Cápita	3 vs 4	0.0171	Grupo 4 > Grupo 3
Tasa de Empleo en Industria	1 vs 2	0.0393	Grupo 2 > Grupo 1

Grupo 1: Este grupo tiene una densidad más baja que el grupo 4, lo que podría indicar una menor concentración de población. Su nivel de autoabastecimiento también es más bajo que el grupo 4, lo que podría reflejar una menor capacidad de producción interna, pero su PIB es más alto que el 2 y el 4 y tiene una mayor tasa de empleo en industria que el 2, lo cual indica una capacidad económica alta.

Grupo 2: Aunque este grupo tiene una densidad más baja en comparación con el grupo 4, su nivel de autoabastecimiento es superior al del grupo 1, indicando una mayor capacidad de producción interna.

Grupo 3: Este grupo tiene una densidad más baja en comparación con el grupo 4. Aunque su nivel de autoabastecimiento es más bajo que el del grupo 4, esto no parece afectar negativamente a sus proporciones de ventas y compras, sobre todo teniendo en cuenta que tiene un PIB más alto que el 4.

Grupo 4: Este grupo tiene la mayor densidad entre todos los grupos, lo que indica una alta concentración de población. Aunque su nivel de autoabastecimiento es más bajo que el del grupo 2, sus proporciones de ventas y compras no se ven afectadas negativamente, lo que indica un equilibrio económico sólido.

CANOPY

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	3 vs 1	0.0366	Grupo 3 > Grupo 1
Densidad	3 vs 2	0.0443	Grupo 3 > Grupo 2
Densidad	4 vs 3	0.0551	Grupo 4 > Grupo 3
Autoabas	3 vs 1	0.1394	Grupo 3 > Grupo 1
Autoabas	3 vs 2	0.2273	Grupo 3 > Grupo 2
Proporción Ventas	3 vs 1	0.0404	Grupo 3 > Grupo 1
Proporción Ventas	3 vs 4	0.0452	Grupo 3 > Grupo 4
Proporción Compras	3 vs 1	0.0954	Grupo 3 > Grupo 1
Proporción Compras	3 vs 4	0.0561	Grupo 3 > Grupo 4
Gasto por Persona	2 vs 1	0.0699	Grupo 2 > Grupo 1
Gasto por Persona	3 vs 2	0.0062	Grupo 3 > Grupo 2
PIB per Cápita	3 vs 1	0.0731	Grupo 3 > Grupo 1
PIB per Cápita	3 vs 2	0.0374	Grupo 3 > Grupo 2
PIB per Cápita	3 vs 4	0.7773	Grupo 3 > Grupo 4
Tasa de Empleo en Industria	4 vs 2	0.0148	Grupo 4 > Grupo 2
Tasa de Empleo en Industria	4 vs 3	0.1270	Grupo 4 > Grupo 3
Tasa de Empleo en Industria	4 vs 1	0.00088	Grupo 4 > Grupo 1
En búsqueda de empleo	2 vs 1	0.00104	Grupo 2 > Grupo 1
En búsqueda de empleo	3 vs 2	0.00822	Grupo 3 > Grupo 2
En búsqueda de empleo	4 vs 2	0.00088	Grupo 4 > Grupo 2
En búsqueda de empleo	4 vs 3	0.71930	Grupo 4 > Grupo 3

Grupo 1: Este grupo tiene una densidad más alta que los grupos 2 y 3, pero más baja que el grupo 4. Aunque su nivel de autoabastecimiento es menor que el del grupo 3, supera al grupo 4 en este aspecto. En términos de proporción de ventas y compras, se encuentra por debajo del grupo 3 pero por encima del grupo 4. El Grupo 1 tiene un gasto por persona inferior al del Grupo 2 pero superior al del Grupo 3. Su PIB per cápita es menor que el de los grupos 2 y 3, y su tasa de empleo en la industria es la más baja entre todos los grupos.

Grupo 2: Este grupo se caracteriza por tener una densidad menor que los grupos 1 y 4, pero mayor que el grupo 3. Aunque su nivel de autoabastecimiento es inferior al del Grupo 1, supera a los Grupos 3 y 4 en este aspecto. En términos de proporción de ventas y compras, se encuentra por debajo del Grupo 1, pero supera a los Grupos 3 y 4. El Grupo 2 tiene el gasto por persona más elevado entre todos los grupos. Su PIB per cápita es mayor que los de los Grupos 1 y4,y su tasa de empleo en la industria es mayor que la del Grupo 1, pero menor que la del Grupo4.

Grupo 3: Este grupo tiene una densidad más baja que el grupo 1 pero mayor que el grupo 2. Su nivel de autoabastecimiento es mayor que el del grupo 4 pero menor que el del grupo 2. En términos de proporción de ventas y compras, supera a todos los demás grupos. El Grupo 3 tiene un gasto por persona menor que el del Grupo 2 pero mayor que el del Grupo 1. Su PIB per cápita es el más elevado entre todos los grupos, y su tasa de empleo en la industria es superior a la del Grupo 2.

Grupo 4: Este grupo tiene una densidad más alta que cualquier otro grupo, aunque su nivel de autoabastecimiento es más bajo que el del grupo 2, sus proporciones de ventas y compras no se ven afectadas negativamente, lo que indica un equilibrio económico sólido.

FARTHESTFIRST

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	3 vs 1	0.0056	Grupo 3 > Grupo 1
Densidad	4 vs 3	0.0065	Grupo 4 > Grupo 3
Autoabas	2 vs 1	0.0004	Grupo 2 > Grupo 1
Autoabas	3 vs 2	0.0078	Grupo 3 < Grupo 2
Autoabas	4 vs 2	0.0061	Grupo 4 < Grupo 2
Autoabas	4 vs 3	0.9990	Grupo 4 = Grupo 3
Proporción Ventas	2 vs 1	0.00076	Grupo 2 > Grupo 1
Proporción Ventas	3 vs 1	0.05910	Grupo 3 > Grupo 1
Proporción Ventas	3 vs 4	0.01895	Grupo 3 > Grupo 4
Proporción Compras	2 vs 1	0.0026	Grupo 2 > Grupo 1
Proporción Compras	3 vs 2	0.1204	Grupo 3 < Grupo 2
Proporción Compras	3 vs 4	0.0663	Grupo 3 < Grupo 4
Gasto por Persona	4 vs 3	0.0486	Grupo 4 > Grupo 3
PIB per Cápita	3 vs 1	0.0231	Grupo 3 > Grupo 1
PIB per Cápita	3 vs 2	0.0115	Grupo 3 > Grupo 2
PIB per Cápita	4 vs 3	0.1401	Grupo 4 > Grupo 3

Grupo 1: Este grupo tiene una densidad más alta que los grupos 2 y 3, pero más baja que el grupo 4. Su nivel de autoabastecimiento es menor que el del grupo 3, pero supera al grupo 4. En términos de proporción de ventas y compras, se encuentra por debajo del grupo 3 pero por encima del grupo 4.

Grupo 2: Este grupo tiene una densidad menor que los grupos 1 y 4, pero mayor que el grupo 3. Su nivel de autoabastecimiento es inferior al del Grupo 1, pero supera a los Grupos 3 y 4. En términos de proporción de ventas y compras, se encuentra por debajo del Grupo 1 pero supera a los Grupos 3 y 4.

Grupo 3: Este grupo tiene una densidad más baja que el grupo 1 pero mayor que el grupo 2. Su nivel de autoabastecimiento es mayor que el del grupo 4 pero menor que el del grupo 2. En términos de proporción de ventas y compras, supera a todos los demás grupos.

Grupo 4: Este grupo tiene una densidad más alta que cualquier otro grupo. Aunque su nivel de autoabastecimiento es más bajo que el del grupo 2, sus proporciones de ventas y compras no se ven afectadas negativamente, lo que indica un equilibrio económico sólido.

SOM

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	3 vs 1	0.0314	Grupo 3 < Grupo 1
Densidad	4 vs 1	0.0270	Grupo 4 < Grupo 1
Autoabas	2 vs 1	0.0749	Grupo 2 < Grupo 1
Autoabas	3 vs 1	0.1513	Grupo 3 < Grupo 1
Proporción Ventas	3 vs 1	0.0124	Grupo 3 < Grupo 1
Proporción Ventas	4 vs 1	0.0595	Grupo 4 < Grupo 1
Proporción Compras	2 vs 1	0.0307	Grupo 2 < Grupo 1
Proporción Compras	3 vs 1	0.0940	Grupo 3 < Grupo 1
Proporción Compras	4 vs 1	0.8626	Grupo 4 = Grupo 1
Ratio Máxima Venta Total	3 vs 2	8.2e-09	Grupo 3 < Grupo 2
Ratio Máxima Venta Total	4 vs 2	1.6e-08	Grupo 4 < Grupo 2
Ratio Máxima Venta Total	4 vs 3	0.7576	Grupo 4 = Grupo 3
Gasto por Persona	4 vs 1	0.0041	Grupo 4 < Grupo 1
PIB per Cápita	4 vs 1	0.0044	Grupo 4 < Grupo 1
PIB per Cápita	4 vs 3	0.0196	Grupo 4 < Grupo 3
Temp. Empleo Agrícola	4 vs 1	0.0439	Grupo 4 < Grupo 1
Temp. Empleo Agrícola	4 vs 3	0.0539	Grupo 4 < Grupo 3
T. Buscando Empleo	4 vs 3	0.0078	Grupo 4 > Grupo 3

Grupo 1: Este grupo tiene una mayor densidad y nivel de autoabastecimiento en comparación con los Grupos 3 y 4. Esto podría indicar que las Comunidades Autónomas en este grupo tienen una economía más fuerte y son más autosuficientes.

Grupo 2: La única comparación disponible para el Grupo 2 en la tabla es con el Grupo 1 en términos de autoabastecimiento. El Grupo 2 tiene un nivel de autoabastecimiento inferior al del Grupo 1. Esto podría sugerir diferencias en la estructura económica o de empleo entre estos dos grupos.

Grupo 3: El Grupo 3 tiene una densidad inferior a la del Grupo 1 pero un nivel de autoabastecimiento superior al del Grupo 4. Esto podría indicar que, aunque estas Comunidades Autónomas pueden tener una menor concentración de ciertos atributos o características (representados por "Densidad"), son capaces de autoabastecerse de manera más eficiente que el Grupo 4.

Grupo 4: Aunque el Grupo 4 tiene una densidad superior a la de cualquier otro grupo, su nivel de autoabastecimiento es inferior al del Grupo 2. Esto podría sugerir que, aunque estas Comunidades Autónomas tienen una alta concentración de densidad, dependen más de fuentes externas para su abastecimiento.

CONCLUSIONES

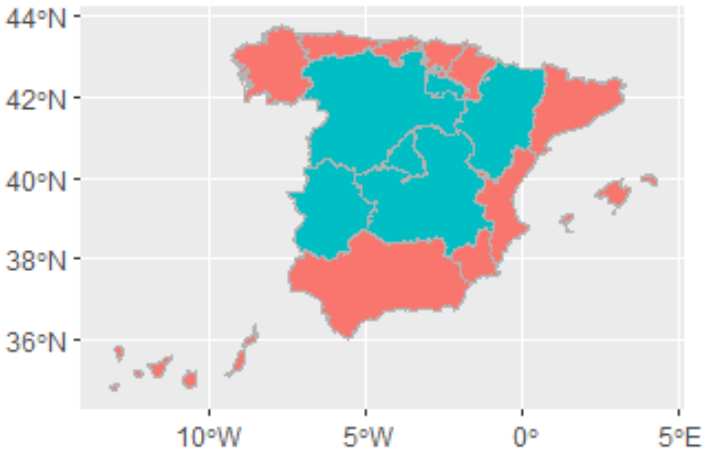
Tras analizar las distintas agrupaciones generadas, hemos observado que presentan muchas diferencias entre sí. Sin embargo, nuestra elección sería la agrupación obtenida mediante el algoritmo Jerárquico Complete. Esta selección se justifica por la mayor claridad y distinción de los grupos, con excepción del Grupo 1, que es menos diferenciado. El resto de las agrupaciones son más complicadas de explicar y no son tan uniformes. Ahora ya podemos comparar nuestros clusters con los previos.

COMPARACIONES CON CLASIFICACIONES PRE-EXISTENTES

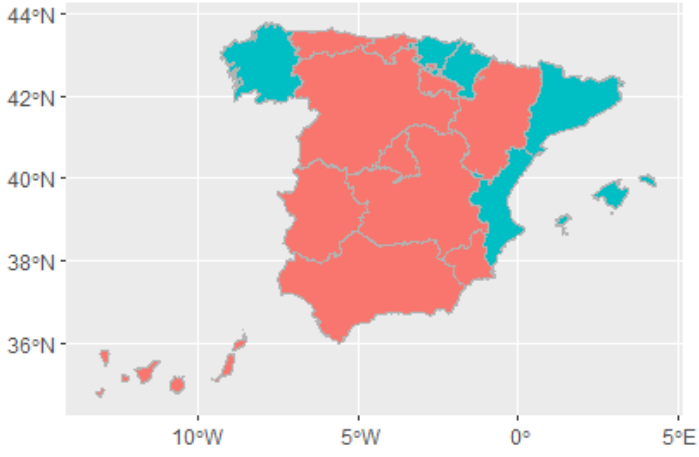
Las agrupaciones previas las hemos graficado con la librería mapSpain, para esto extrajimos una base de datos ya incluida en el paquete, *esp_codelist*. y creamos artificialmente listas con el número del grupo como factores de cada criterio usado para los mapas. Los criterios son los siguientes, en orden, por periferia o interior (ROJO=periferia, AZUL=interior), por lengua propia o no (ROJO=no lengua propia, AZUL=lengua propia), por norte o sur (ROJO=Norte, AZUL=Sur), y según el nivel de paro (ROJO= más del 15%, AZUL= menos del 10%, VERDE= entre el 10% y el 15%).

Para realizar los gráficos hemos usado *ggplot(datos)+geom_sf(aes=fill=cluster)...)*.

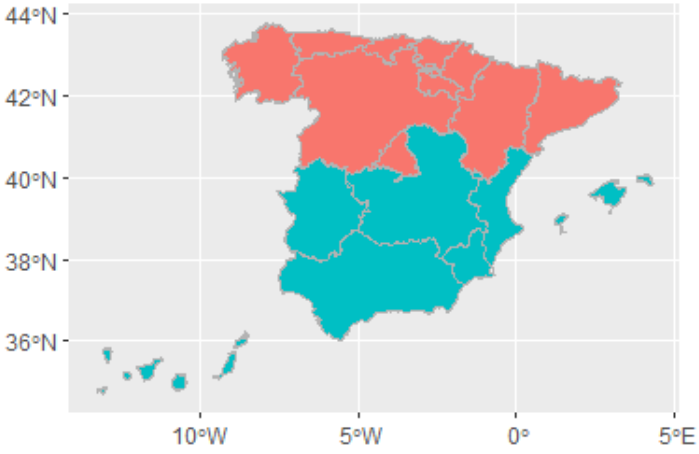
AGRUPACIÓN DE LAS CCAA ESPAÑOLAS SEGÚN SI SON PERIFERIA O INTERIOR



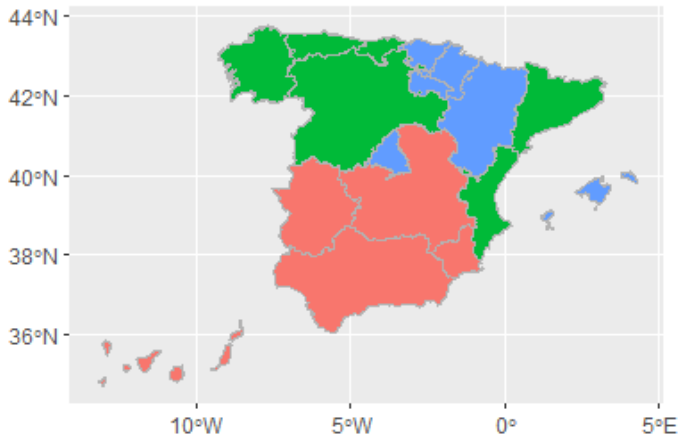
AGRUPACIÓN DE LAS CCAA ESPAÑOLAS SEGÚN SI TIENEN LENGUA PROPIA O NO



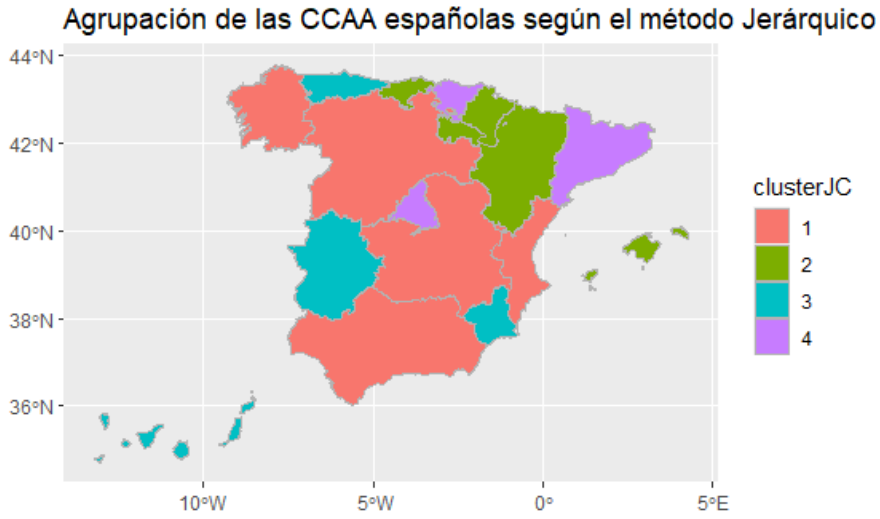
AGRUPACIÓN DE LAS CCAA ESPAÑOLAS SEGÚN SI SON NORTE O SUR



AGRUPACIÓN DE LAS CCAA ESPAÑOLAS SEGÚN SU NIVEL DEL PARO



Nuestro clusterer, en cambio, queda de la siguiente forma:



Las coincidencias que podemos ver son pocas, tal vez en cuanto a lengua propia en el cluster 4, que tiene el País Vasco y Cataluña. Además, podríamos ver el grupo 1 como interior si nos fijamos en la zona del Norte. Podríamos ver que el grupo 2 se encuentran aquellos con menos del 10% de paro menos por Madrid y el País Vasco. Pero, además de esto no vemos unas claras similitudes, especialmente con el grupo 3.

Esto puede indicar que nuestro país es muy variopinto y porr lo tanto podría clasificarse de muchas formas según el objetivo del estudio. En el nuestro, hemos optado por un enfoque más bien económico, pero cualquier otro de los anteriores sería igual de válido.

BIBLIOGRAFÍA

Hernangómez, D. (n.d.). Get started.

https://cran.rproject.org/web/packages/mapSpain/vignettes/mapSpain.html

RPUBS - ML-Assignment 2 Clustering. (n.d.).
https://rpubs.com/sushantgote/ml2clustering

som. (n.d.). https://www.uv.es/mlejarza/datamine/som.html

RPUBS - Good Practice Toolbox for Analyzing Data using Unsupervised Learning Methods.(n.d.).

https://rstudiopubsstatic.s3.amazonaws.com/548392_de6cd39746a641718eb3412342

1cb11f.html

GeeksforGeeks. (2023, March 21). Making Maps with R.

https://www.geeksforgeeks.org/making-maps-with-r/

RPUBS-Análisis exploratorio por Componentes Principales y Clúster
https://edimer.github.io/Stat/6_MetodosExploratorios.html

GeeksforGeeks. (2023, April 18). Self organizing maps Kohonen maps.

https://www.geeksforgeeks.org/self-organising-maps-kohonen-maps/