Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Análisis Multivariado
Cali, abril de 2019

Análisis de Correspondencias Múltiples (Parcial Aplicado)

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

Para esta práctica decidimos trabajar sobre la base de datos correspondiente a la "Encuesta Nacional de Uso del Tiempo" (ENUT: 2016 - 2017). Al ser una encuesta tan extensa decidimos tomar solo los datos correspondientes al "CAPITULO B: CONDICIONES DE LA VIVIENDA", el cual cuenta con más de 44 mil datos de hogares y 14 variables cualitativas o categóricas.
Luego de observar las posibles variables que podíamos utilizar para realizar el análisis de correspondencias múltiples, nos quedamos con 9 variables, de las cuales 8 corresponden a preguntas sobre las condiciones de vivienda y una es la variable "REGIÓN" correspondiente a las 6 regiones de Colombia en las cuales se realizó el estudio.

Las variables son las siguientes:

P424: Tipo de vivienda(1.Casa,2.Apartamento,3.Cuarto,4.Vivienda indígena,5.Otra)

P4030S1: Cuenta la vivienda con energía eléctrica? (1.Si , 2.No)

P4030S1A1: Estrato socio económico(Bajo - Bajo /1, Bajo /2, Medio - Bajo /3, Medio /4, Medio - Alto /5, Alto /6)

P4030S5: Cuenta la vivienda con Acueducto? (1.Si , 2.No)

P4030S3: Cuenta la vivienda con Alcantarillado? (1.Si , 2.No)

P4030S4: Cuenta la vivienda con recolección de basuras? (1.Si , 2.No)

P4030S4A1: Cuántos días de la semana se recolecta basura?(1,2,3,4,5,6,7,8)

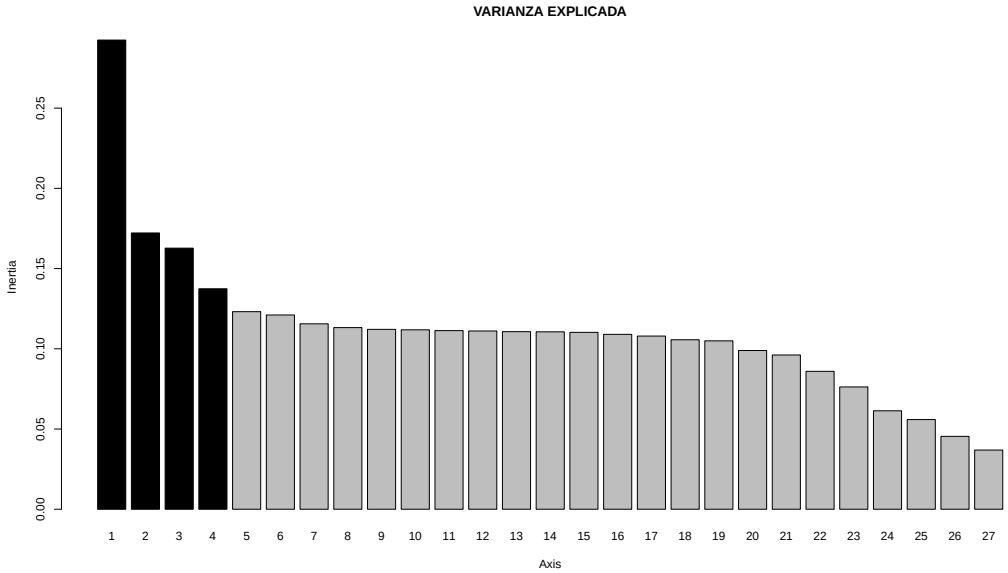
P4030S2: Cuenta la vivienda con Gas natural? (1.Si,2.No)

REGION: Región de la encuesta (6.San Andrés , 5.Bogotá , 4.Pacífica , 3.Oriental , 2.Central , 1.Atlántica)

2. Punto 1. ¿Cuántos ejes se generan en el ACM?

Se generaron en total 27 ejes, en la siguiente gráfica podemos observar los 27 ejes y su respectiva varianza explicada.

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co



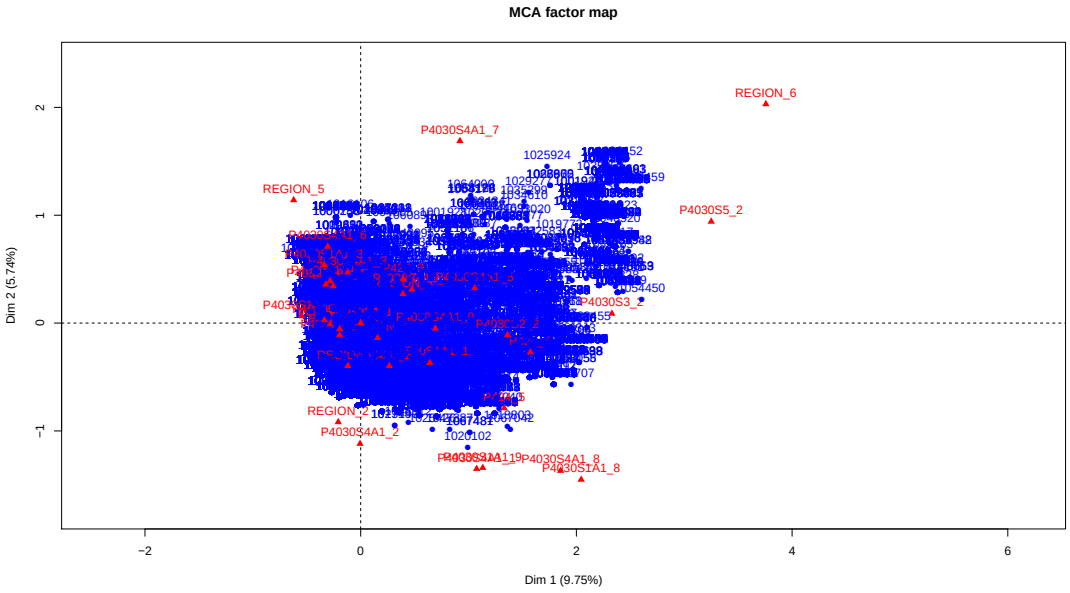
3. Punto 2. ¿Qué relación existe entre la Inercia Total, el número de modalidades y el número de variables cualitativas?

La relación que existe es que el número de ejes que resultan en un ACM es igual al número de modalidades menos el número de variables, en nuestro caso contamos con un total de 36 modalidades y 9 variables o preguntas por lo que tenemos un total de 27 $(36 - 9)$ ejes, cada eje explica un porcentaje de la inercia y el total de inercia es igual a la suma de la inercia por modalidades o por variables(preguntas):

$$I = \sum_{q=1}^{s} I_q = \sum_{j=1}^{p} I_j; \;\; Donde: \;\; I_q = Inercia \;\; por \;\; pregunta \;\; y \;\; I_j = Inercia \;\; por \;\; modalidad$$

4. Punto 3. ¿Qué similitudes existen entre las filas?

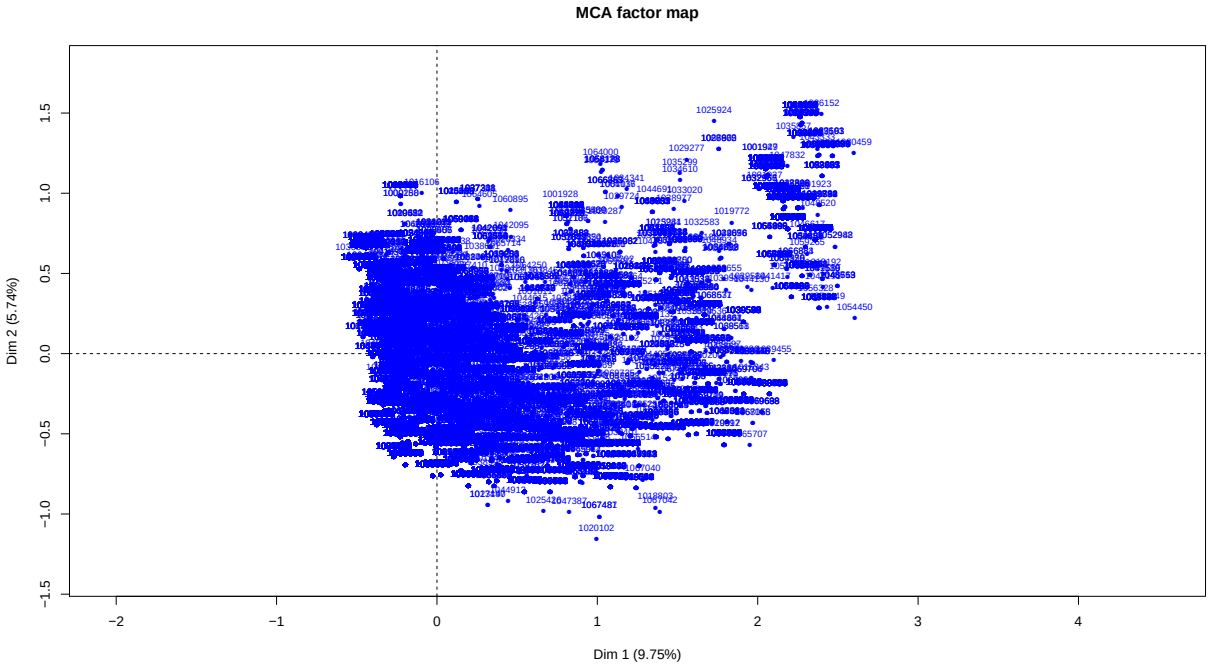
Las filas están dadas por los hogares encuestados, en nuestro caso dos hogares son similares o se parecen si ambos han seleccionado las mismas modalidades, por ende pasamos a evaluar que similitudes hay entre los hogares:



En principio las gráficas se ven algo cargadas ya que estamos trabajando con más de 38 mil hogares o individuos y 36 modalidades, sin embargo podemos ver que varios hogares comparten similitudes entre sus modalidades. En el gráfico se puede apreciar que se forman 3 grupos, la gran mayoría de hogares se ven reunidos en el grupo más grande, el cual es el que está más cerca del centro, en este se encuentran los hogares que tuvieron respuestas afirmativas a varias preguntas sobre si se cuenta con servicios públicos y que corresponden a las regiones más grandes que son 1.Atlántico y 3.Oriental, dado lo anterior, podemos decir que la gran mayoría de hogares encuestados tiene todos los servicios públicos y de los cuales una gran parte pertenece a las regiones de Atlántico y Oriental. Observamos que los individuos mas alejados están relacionados con la región 6 que es San Andrés, esta región al ser una isla es la que tiene menos hogares, al igual que los estratos por encima del 3, resultan ser los menos frecuentes.

5. Punto 4. Con base a la nube de individuos ¿Cuáles son los valores excéntricos?

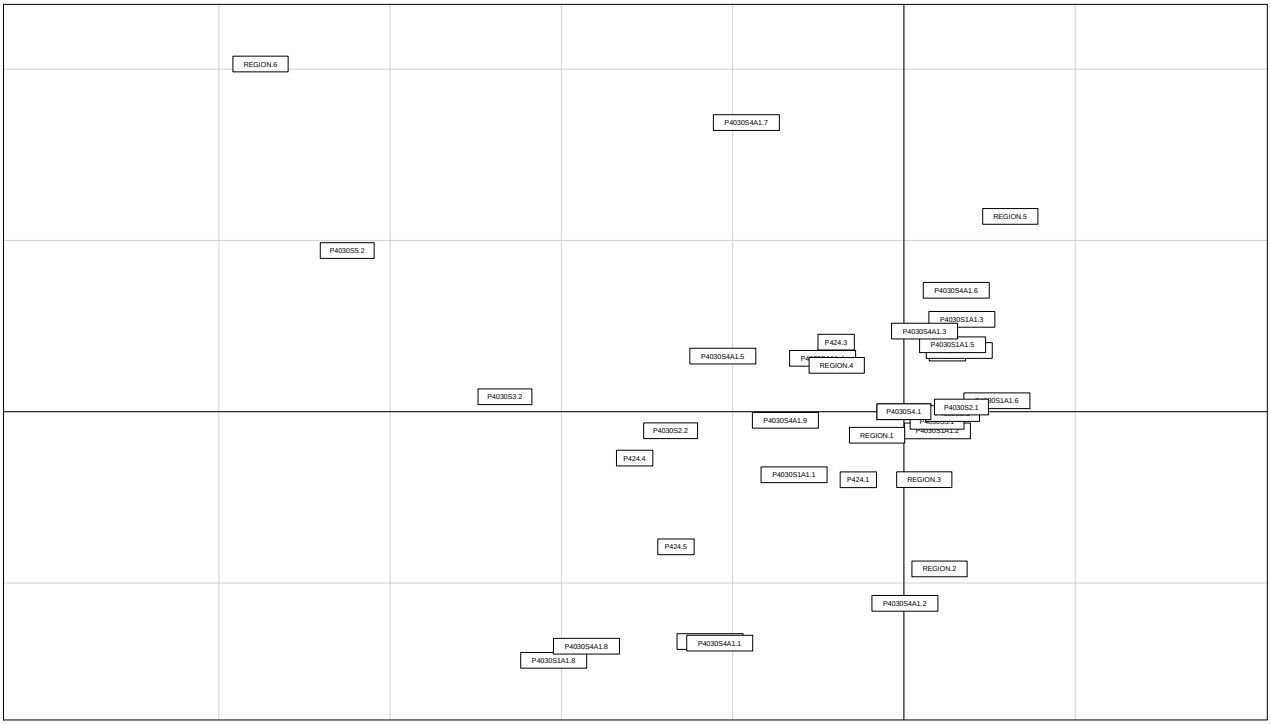
Evaluamos el gráfico de la nube de individuos para conocer los valores que se alejan del centro.



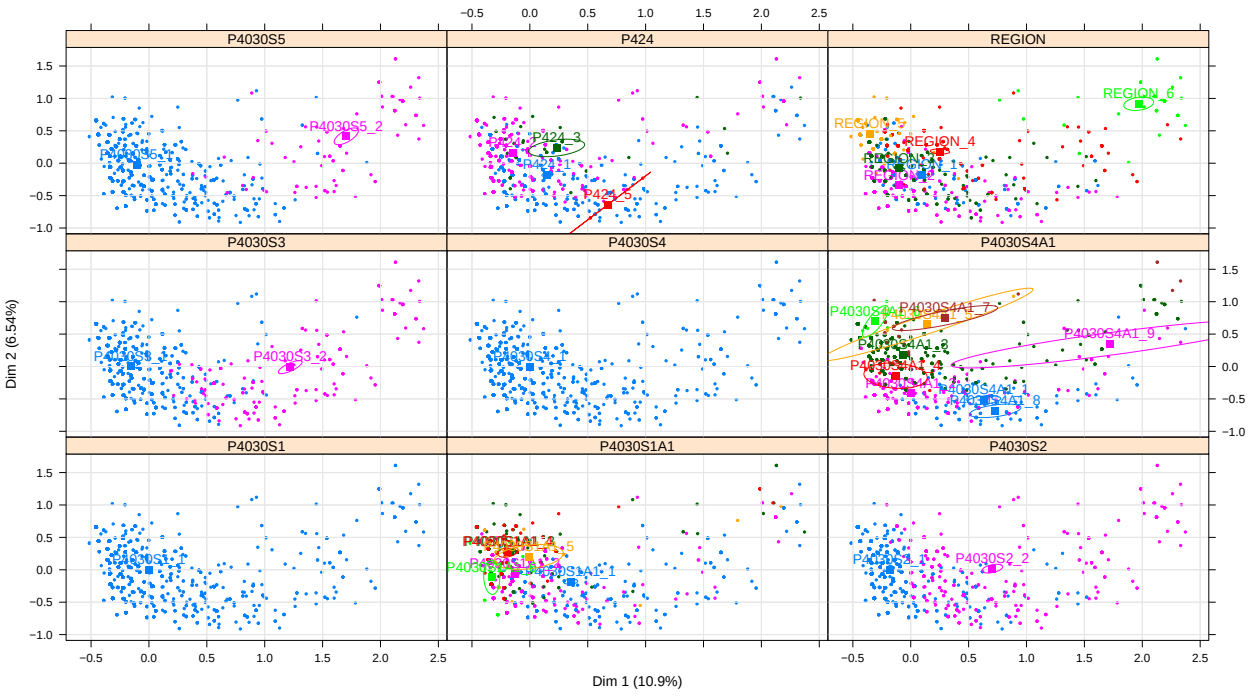
Según vemos en el gráfico y pudimos apreciar en el punto anterior los valores excéntricos son los que están mas relacionados con las modalidades "no" de las preguntas sobre si se cuenta con servicios públicos, podemos observar hogares como 102988, 1041524,1038213,1019294,1039511,... entre muchos otros, que además de no contar con varios servicios están mas asociados a la región de San Andrés que es la que menos habitantes tiene.

6. Punto 5. Con base a la nube de variables ¿Cuáles son las modalidades más asociadas?

Evaluamos el gráfico de la nube de variables para conocer que modalidades están más cerca entre sí. Como se logra apreciar, el formato del gráfico cambió, esto se debe a que fue realizado con el paquete *ade*4 y los anteriores fueron realizados con el paquete "FactoMineR", aun así las interpretaciones son las mismas, sin embargo *ade*4 resulta ser mas amigable en algunos gráficos a la hora de interpretar.
En el gráfico siguiente, correspondiente a la nube de las variables, podemos observar que las personas con hogares tipo casa, apartamento y cuarto suelen tener la mayoría de servicios públicos, también los hogares que cuentan con todos sus servicios, o sea los que marcaron "si" en estas preguntas



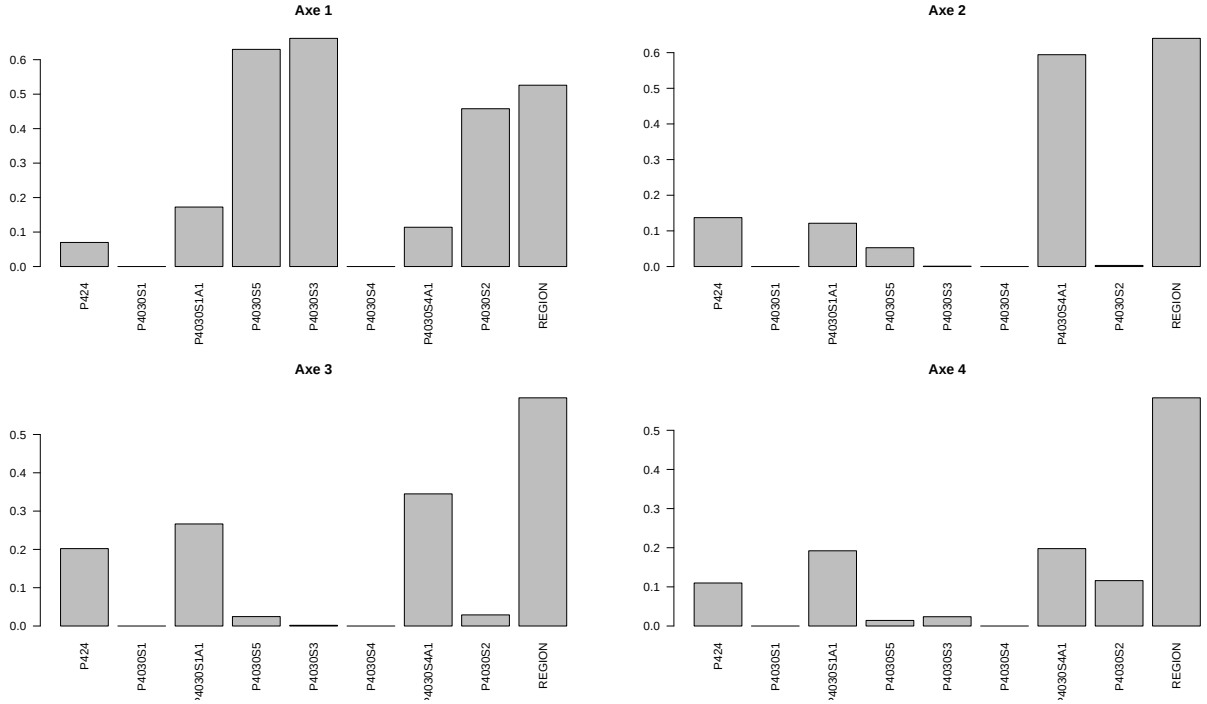
suelen estar cerca de las modalidades referentes a la pregunta estrato que tienen valores del 2 al 6, por lo que gran parte de las personas encuestadas a priori se puede decir que cuentan con varios de los servicios públicos. Algo curioso a tener en cuenta es que la región Pacifica esta muy relacionada con cuartos, por lo que se puede decir que las personas en el pacifico suelen vivir en cuartos con sus familias y dado esto, se podría hablar de una posible sobre población y pobreza en esa región.



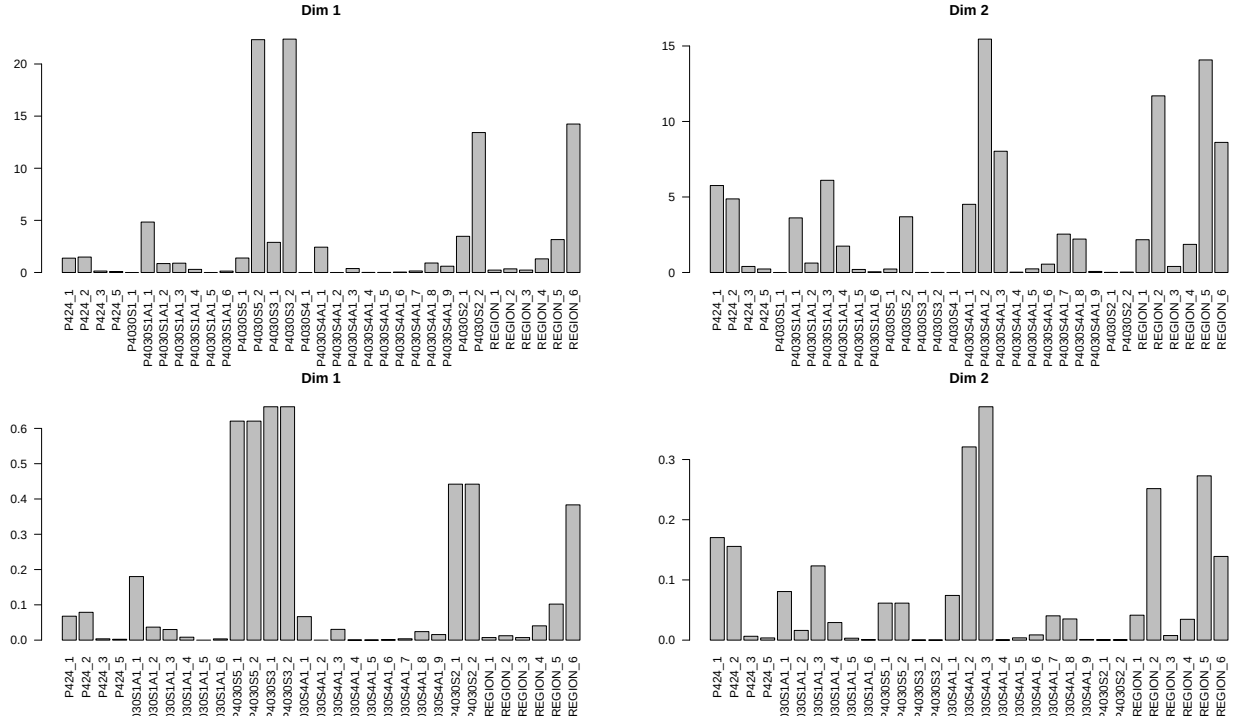
El gráfico anterior nos muestra las modalidades que tienen diferencias significativas, lo primero que notamos es que la región 6(San Andrés) es la más alejada del resto de las regiones, las regiones 1(Atlántica),2(Central) y 3(Oriental) son las más parecidas. Vemos que la mayoría de hogares cuenta con acueducto, alcantarillado y todos cuentan con recolección de basuras y energía eléctrica.

7. Punto 6. Realice una interpretación de las contribuciones absolutas y cosenos cuadrados para las modalidades y variables. ¿Los valores excéntricos tienen contribuciones altas?

El siguiente gráfico nos muestra cada pregunta como se ve representada en cada uno de los ejes, podemos ver que la mayoría de preguntas se ven mejor representada en el eje 1, por otra parte la región se ve muy bien representada en los primeros 4 ejes, el eje 3 muestra mejor el tipo de hogar.



Ya que tenemos 36 modalidades y mas de 38 mil datos, revisar cada uno de los valores se vuelve tedioso y complicado, por ende para el análisis por modalidades se optó por realizar un gráfico de barras de los cosenos cuadrados y las contribuciones en los dos primeros ejes.



Este dio como resultado para los dos primeros ejes que las modalidades que más contribuyen a la formación del eje 1 son la respuestas negativas a si se cuenta con servicio de acueducto y alcantarillado, mientras que para el eje dos las regiones 5 ,6 y 2 contribuyen a su formación.
Por la parte de los cosenos, cuyos gráficos están en la segunda fila, podemos ver que las modalidades que mejor se ven representadas son las referentes a la pregunta sobre si se cuenta con acueducto y si se cuenta con alcantarillado, para el eje dos las modalidades que se ven mejor representadas son en las que se recolecta basura los días 1 y 3 de la semana así como varias de las regiones.

En el caso de los individuos es muy difícil lograr visualizar estas contribuciones y cosenos por la magnitud de los datos, sin embargo podemos ver que tanto contribuyen y se ven representados los hogares que tuvieron valores excéntricos.

Teníamos que algunos valores excéntricos son: (1041524,1038213,1019294,1039511), de los cuales en el eje 1 todas dieron contribuciones inferiores a 1, por lo que estos individuos excéntricos no contribuyen a su formación, para el eje 2 son aun más bajas. Los cosenos cuadrados para los 3 primeros hogares en el eje 1 rondan valores entre 0.30 y 0.45, se puede decir que se ven bien representados, por otra parte el tercer hogar tiene un valor de aproximadamente 0.08 el cual es bastante bajo.

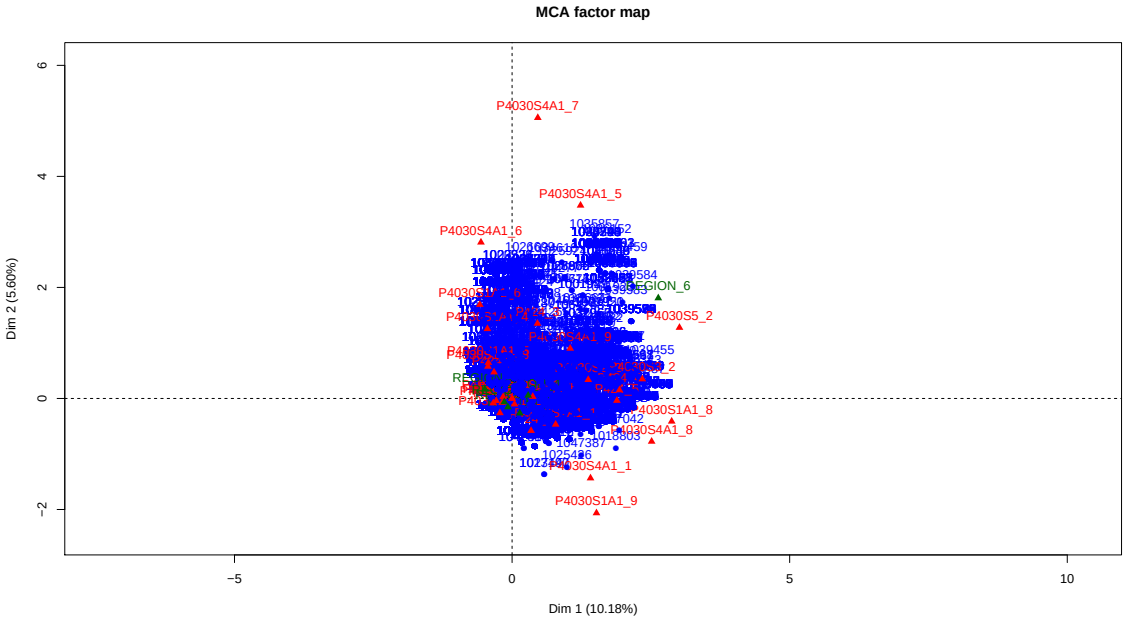
CONTRIBUCIONES			
Hogar	Dim 1	Dim 2	Dim 3
1041524	0.9665660	0.5168859	0.4833209
1038213	0.90955947	0.40907346	0.49802097
1019294	0.97264253	0.29979671	0.10706549
1039511	0.96740190	0.04157307	0.14073777

COSENOS			
Hogar	Dim 1	Dim 2	Dim 3
1041524	0.33623659	0.10779526	0.09316037
1038213	0.317828028	0.085694690	0.096425279
1019294	0.426899316	0.078884300	0.026037776
1039511	0.071953936	0.001853750	0.005800167

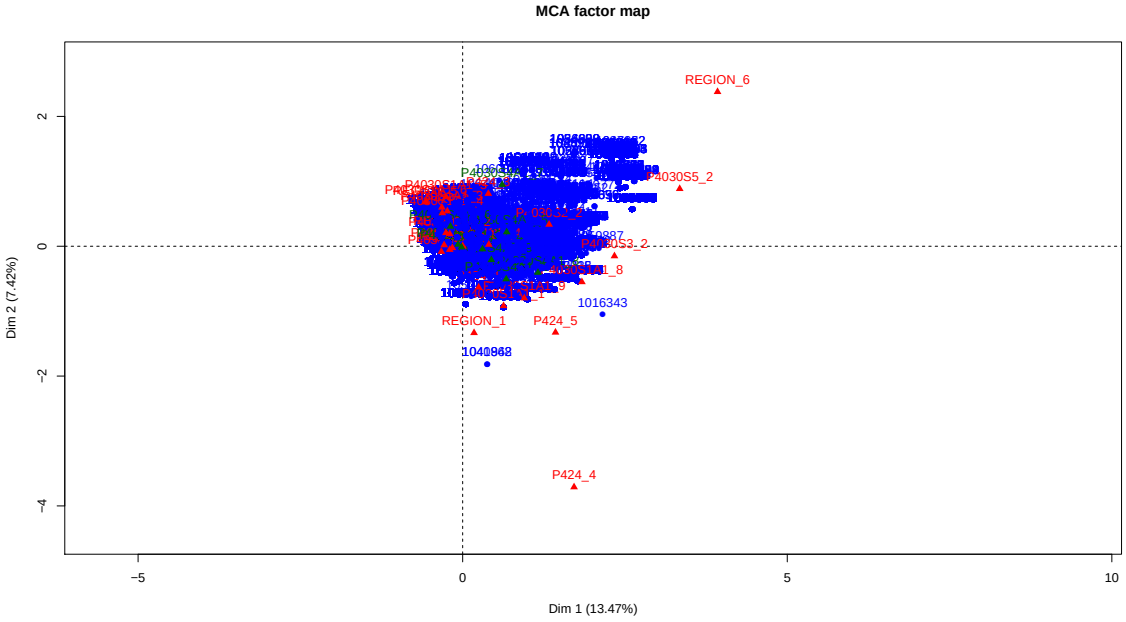
La conclusión principal que se obtiene es que los valores excéntricos no contribuyen de gran forma a la formación de los ejes, ya que estos presentan frecuencias bajas en la matriz F.

8. Punto 7. Utilizar variables suplementarias o individuos suplementarios si se requiere

El ACM se realizó teniendo en cuenta todas la variables e individuos sin embargo se podría haber tomado varias preguntas como suplementarias, ente este punto veremos como se comporta el ACM si tomamos la pregunta Región y la Pregunta referente al día en que se recolecta la basura como variables suplementaria.



Tomando la región como variable suplementaria vemos una menor dispersión en el gráfico, por lo que se puede decir que las modalidades de esta variable hacían que varios hogares se diferenciaran de otros, lo cual es lógico ya que dependiendo de la región también se tenían diferentes cantidades de hogares. El ACM se pudo haber trabajado de esta manera y se hubieran tal vez obtenido conclusiones mas generales como que la gran mayoría del país cuenta con varios de los servicios públicos.

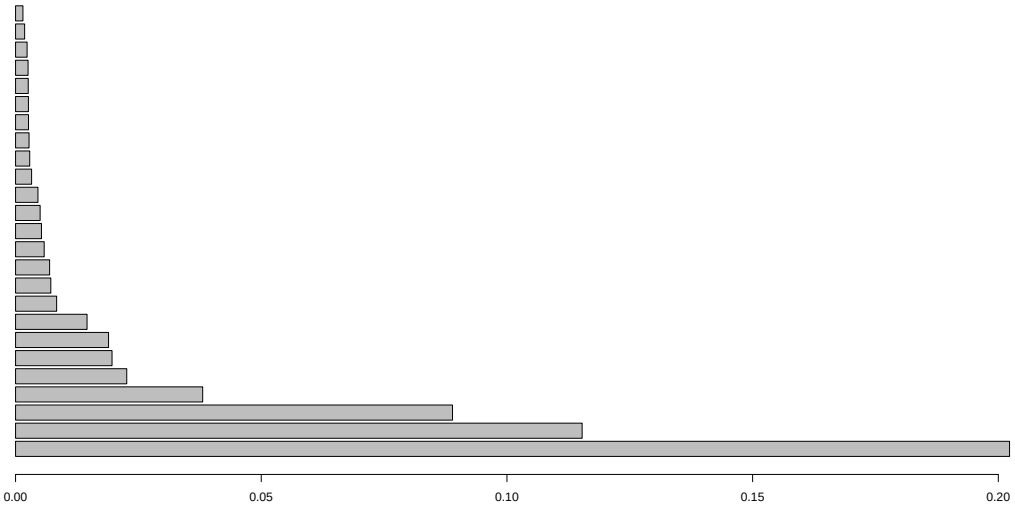


Esta pregunta referente al día que se recolectan las basuras, quisimos verla como variable suplementaria ya que es algo conflictiva, tiene muchas modalidades y la interpretación de cada una es algo confusa, ademas de que no aporta mucho al análisis, sin embargo decidimos incluirla en el ACM ya

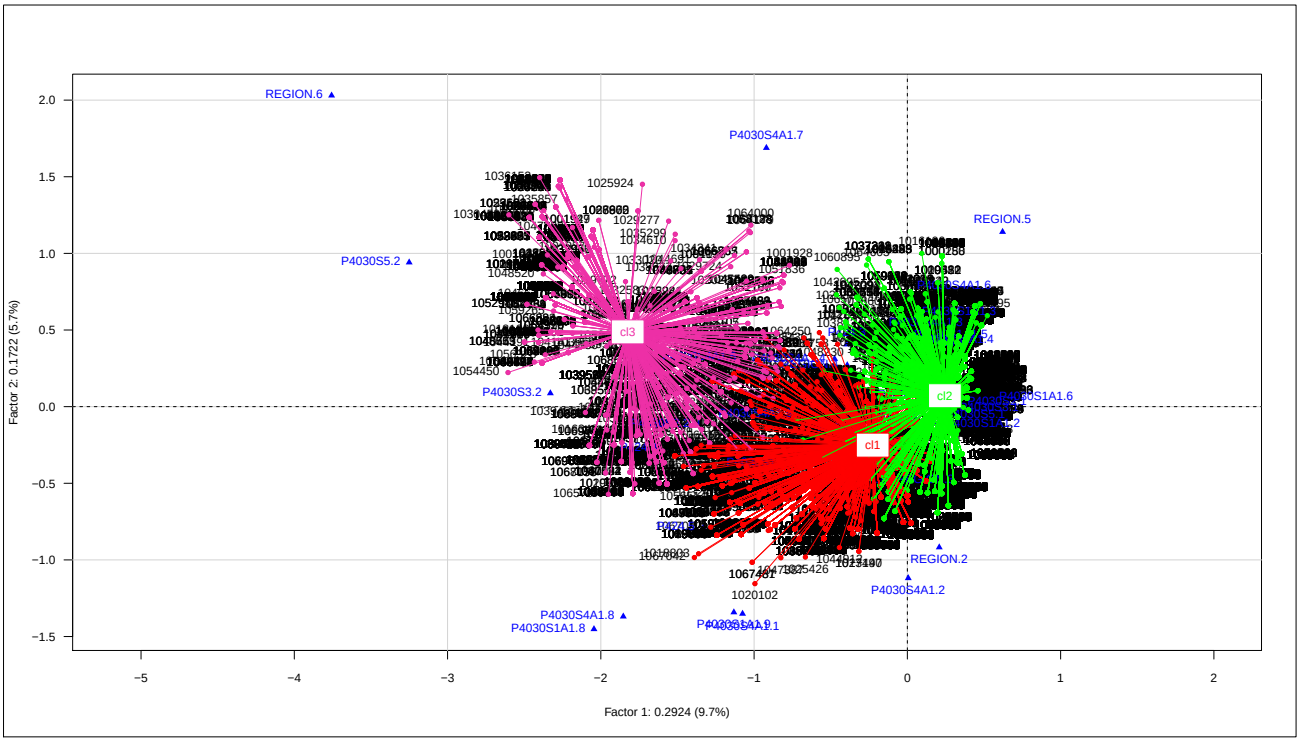
que el trabajo es con fines académicos.
Vemos que de haberla tomado como variable suplementaria o de no haberla incluido los datos logran verse mas agrupados en el centro, sin embargo la región sigue llevándose gran parte de los hogares, en especial la región 6.

9. Punto 8. Acompañar los resultados con el clúster mixto

Se opta por la utilización de un clúster mixto ya que se están trabajando con variables categóricas, ademas de que se tiene un volumen de datos relativamente grande.
Primero miramos cuantos posibles clúster realizar, o cuantos grupos se pueden formar:



Observando la gráfica anterior y también como se distribuyen los puntos en la nube de individuos podemos decir que se conforman 3 grupos, así pues pasamos a realizar el clúster.



El gráfico nos muestra como se conforman los 3 grupos, donde los grupos de color verde y rojo

recogen los hogares con mas frecuencia y el otro recogería los hogares faltantes que tienen menos frecuencia.
Podemos relacionar los grupos con ciertas variable,el grupo verde y rojo tienen hogares con varios de los servicios públicos, pero el grupo rojo tiene hogares con estratos bajos, mientras que el verde tiende mas a l clase media-alta,el grupo 3 o rosa tiene mayor cantidad de hogares en la región de San Andrés, y esta mas relacionado con modalidades poco frecuentes.

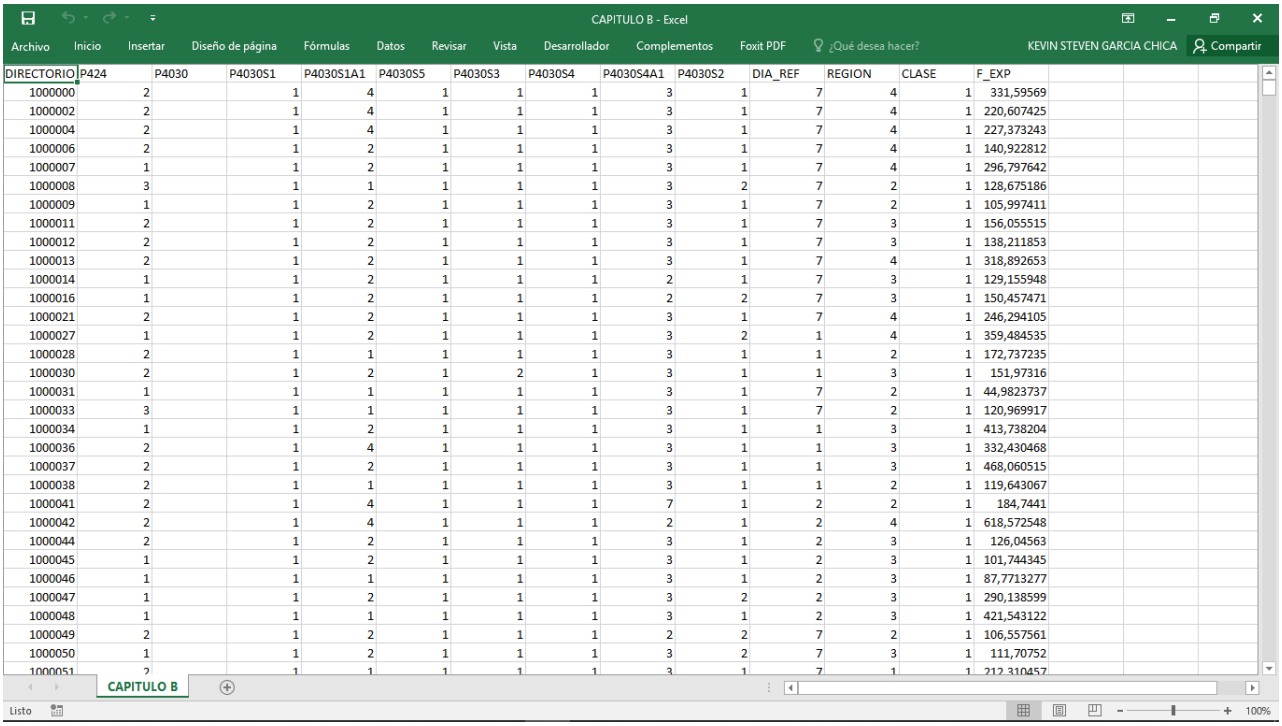
10. Conclusiones

Para concluir podemos decir que la mayoría de los hogares en Colombia cuentan con todos los servicios públicos, además que los estratos 2 y 3 aunque son clase "media - baja" curiosamente cuentan con la mayoría de los servicios públicos.
Podemos también decir que la mayoría de hogares en la región Pacifica son cuartos, lo cual como se mencionó anteriormente, es un fuerte indicio de aglomeración o pobreza en esa región. También se notó que todos los hogares encuestados en todo el país cuentan con servicios de recolección de basuras y acueducto.
Con respecto a las variables suplementarias; se vio que aunque se decidiera incluir la variable referente a la pregunta sobre el día en que se recolecta la basura esta demostró no ser de mucho aporte en el análisis por lo que podría eliminarse del análisis.

Para complementar las conclusiones decidimos realizar adicionalmente un análisis descriptivo de la base de datos, se obtuvo que, con respecto a la primera variable P424 la cual corresponde al tipo de vivienda, el 97.96 % de los hogares viven en casa o apartamento, el porcentaje restante (2.04 %) se divide entre cuarto, vivienda indígena y otra. Con respecto a las variables P4030S1, P4030S5, P4030S3, P4030S4 y P4030S2 se obtuvo respectivamente, que el 98.89 % de las viviendas cuentan con energía eléctrica, el 89.53 % de las viviendas cuentan con acueducto, el 79.34 % de las viviendas cuentan con alcantarillado, el 87.28 % de las viviendas cuentan con recolección de basuras y el 71.52 % de las viviendas cuentan con gas natural. Adicionalmente, con respecto a la variable P4030S1A1 que corresponde al estrato socio económico, se obtuvo un resultado preocupante y es que el 71.83 % de las viviendas son de estrato bajo - bajo y de bajo, es decir, estratos 1 y 2, el resto del porcentaje (28.16 %) se distribuyó en los 4 estratos restantes de manera decreciente, siendo el estrato 3 (medio - bajo) el que mayor porcentaje presenta.

Aquí se pudo evidenciar una de las ventajas fuertes del análisis multivariado en general y del análisis de correspondencias múltiples específicamente, ya que en las estadísticas descriptivas dadas anteriormente no se pueden establecer relaciones entre las variables, es decir, no se puede decir por ejemplo, a que región pertenecen mayormente las viviendas de estratos bajos, o en que regiones la mayoría de las viviendas no cuentan con algún servicio público, entre otras relaciones. Por lo que el análisis multivariado saca una gran ventaja en este aspecto, permitiendo no solo obtener conclusiones generales como con las descriptivas, si no también permitiendo encontrar posibles relaciones causales entre las variables de estudio.

Apéndice A. Base de datos utilizada



CAPITULO B - Excel

Archivo Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista Desarrollador Complementos Foxit PDF ¿Qué desea hacer? KEVIN STEVEN GARCIA CHICA Compartir

DIRECTORIO	P424	P4030	P4030S1	P4030S1A1	P4030S5	P4030S3	P4030S4	P4030S4A1	P4030S2	DIA_REF	REGION	CLASE	F_EXP
1000000	2		1	4	1	1	1	3	1	7	4	1	331,59569
1000002	2		1	4	1	1	1	3	1	7	4	1	220,607425
1000004	2		1	4	1	1	1	3	1	7	4	1	227,373243
1000006	2		1	2	1	1	1	3	1	7	4	1	140,922812
1000007	1		1	2	1	1	1	3	1	7	4	1	296,797642
1000008	3		1	1	1	1	1	3	2	7	2	1	128,675186
1000009	1		1	2	1	1	1	3	1	7	2	1	105,997411
1000011	2		1	2	1	1	1	3	1	7	3	1	156,055515
1000012	2		1	2	1	1	1	3	1	7	3	1	138,211853
1000013	2		1	2	1	1	1	3	1	7	4	1	318,892653
1000014	1		1	2	1	1	1	2	1	7	3	1	129,155948
1000016	1		1	2	1	1	1	2	2	7	3	1	150,457471
1000021	2		1	2	1	1	1	3	1	7	4	1	246,294105
1000027	1		1	2	1	1	1	3	2	1	4	1	359,484535
1000028	2		1	1	1	1	1	3	1	1	2	1	172,737235
1000030	2		1	2	1	2	1	3	1	1	3	1	151,97316
1000031	1		1	1	1	1	1	3	1	7	2	1	44,9823737
1000033	3		1	1	1	1	1	3	1	7	2	1	120,969917
1000034	1		1	2	1	1	1	3	1	1	3	1	413,738204
1000036	2		1	4	1	1	1	3	1	1	3	1	332,430468
1000037	2		1	2	1	1	1	3	1	1	3	1	468,060515
1000038	2		1	1	1	1	1	3	1	1	2	1	119,643067
1000041	2		1	4	1	1	1	7	1	2	2	1	184,7441
1000042	2		1	4	1	1	1	2	1	2	4	1	618,572548
1000044	2		1	2	1	1	1	3	1	2	3	1	126,04563
1000045	1		1	2	1	1	1	3	1	2	3	1	101,744345
1000046	1		1	1	1	1	1	3	1	2	3	1	87,7713277
1000047	1		1	2	1	1	1	3	2	2	3	1	290,138599
1000048	1		1	1	1	1	1	3	1	2	3	1	421,543122
1000049	2		1	2	1	1	1	2	2	7	2	1	106,557561
1000050	1		1	2	1	1	1	3	2	7	3	1	111,70752
1000051	2		1	1	1	1	1	3	1	7	1	1	212,310457

CAPITULO B

Listo 100%

Apéndice B. Código R

```
#Parcial Aplicado ACM
#Kevin García - 1533173
#Alejandro Vargas - 1525953
datos=read.table (file = "clipboard", header = TRUE, row.names = 1, sep = "\t", dec = ".")
View(datos)
#quitamos los datos faltantes
datos=na.omit(datos)
#covertir las variables en factores
datos$P424=as.factor(datos$P424)
datos$P4030S1=as.factor(datos$P4030S1)
datos$P4030S1A1=as.factor(datos$P4030S1A1)
datos$P4030S5=as.factor(datos$P4030S5)
datos$P4030S3=as.factor(datos$P4030S3)
datos$P4030S4=as.factor(datos$P4030S4)
datos$P4030S4A1=as.factor(datos$P4030S4A1)
datos$P4030S2=as.factor(datos$P4030S2)
datos$REGION=as.factor(datos$REGION)
str(datos)
length(datos$P424)#numero de individuos
#####__###____#____#___#
#quitamos el 60% de los datos
#g=c(sample(1:length(datos$P424),0.95*length(datos$P424)))
#datos=datos[-g,]
#datos=datos[,-4]
#####__###____#____#___#
#descritivas
#graficar cada una de las modalidades segun sus ocurrencias
x11()
par(mfrow=c(3,3))
for (i in 1:length(datos)) {
```

```
  plot(datos[,i])

}

#para encontrar la incercia de la modalidad si en la pregunta del gas natural
j=c()
for (i in 1:length(datos$P4030S2)) {
  if(datos$P4030S4A1[i]==4){
    j=c(j,datos$P4030S4A1[i])
  }

}

sum(j)/length(datos$P424)
#Para realizar el ACM
install.packages("ade4")
install.packages("FactoMineR")
library("ade4")
mod.mca=dudi.acm(datos,scannf = FALSE, nf = 4)#modelo ACM
#grafico de valores propios
x11()
screeplot(mod.mca,main = "VARIANZA EXPLICADA")
#valores propios
summary(mod.mca)

mod.mca$cr

sum(mod.mca$eig)

#grafico de variables
x11()
s.corcircle(mod.mca$co, 1, 2, clabel = 0.7)
#grafico de individuos
x11()
s.corcircle(mod.mca$li, 1, 2, clabel = 0.7)
#correlacion graficos
x11()
par(mfrow = c(2, 2))
for (i in 1:4) barplot(mod.mca$cr[, i], names.arg = row.names(mod.mca$cr), las = 2, main = paste("A

#representar la modalidades
x11()
s.label(mod.mca$co, clabel = 0.5)

#representacion de individuos
x11()
s.label(mod.mca$li, clabel = 0, pch = 20)

#Con FactoMineR seria
install.packages("FactoMineR")
library("FactoMineR")
#Modelo
head(datos)
acm2 <- MCA(datos, ncp = 5, graph = FALSE )
acm2 <- MCA(datos, ncp = 5, graph = FALSE,quali.sup =7 )#con region como SUPL
acm2
#Contribuciones y cosenos para las variables y modalidades
acm2$var$contrib
acm2$var$cos2
na.omit(acm2$var$cos2)
```

```
#para las contribuciones
x11()
par(mfrow = c(2,2 ))
for (i in 1:2) barplot(acm2$var$contrib[, i], names.arg = row.names(acm2$var$contrib), las = 2, mai
for (i in 1:2) barplot(na.omit(acm2$var$cos2)[, i], names.arg = row.names(na.omit(acm2$var$cos2)), l

#para los cosenos
x11()
par(mfrow = c(3,2 ))
for (i in 1:2) barplot(na.omit(acm2$var$cos2)[, i], names.arg = row.names(na.omit(acm2$var$cos2)), l


#Contribuciones y cosenos para los individuos
acm2$ind$contrib
acm2$ind$cos2
x11()
par(mfrow = c(2,2 ))
for (i in 1:2) barplot(acm2$ind$contrib[, i], names.arg = row.names(acm2$ind$contrib), las = 2, mai
for (i in 1:2) barplot(na.omit(acm2$ind$cos2)[, i], names.arg = row.names(na.omit(acm2$ind$cos2)), l

u=c(1041524,10282113,1019294,1039511)
u=as.character(u)
for (i in 1:4) {
  acm2$ind$contrib[u[i],]
}
acm2$ind$cos2["1041524",]
acm2$ind$cos2["1038213",]
acm2$ind$cos2["1019294",]
acm2$ind$cos2["1039511",]


#grafica de individuos y variables
x11()
plot(acm2)

#solo individuos
x11()
plot.MCA(acm2, invisible=c("var","quali.sup"), cex=0.7)

#solo variables
x11()
plot.MCA(acm2, invisible=c("ind","quali.sup"), cex=0.7)

#para observar si existen diferencias significativas entre modalidades
x11()
plotellipses(acm2)

#valores propios
acm2$eig

####CLUSTER
##metodo k-means
install.packages("cluster")
install.packages("factoextra")
install.packages("FatoClass")
library(cluster)
library(factoextra)
library(FactoClass)
#determinacion de los cluster
```

```
#####-##
x11()
Y3D <- scatterplot3d (datos, main ="Y",type="h",color ="darkblue",box=FALSE)
Y3D$points3d(datos,pch=1)

#cluster
cluster1=FactoClass(datos,dudi.acm,nf = 3,nfcl = 3,scanFC = F,k.clust = 3)

x11()
plot(cluster1$cluster)

#representacion de individuos
x11()
clusplot(mod.mca$li,cluster2$cluster)


#otro grafico
x11()
plotFactoClass(cluster1,cframe=1,col.row=c("red","green","maroon2","orchid4","pink"))
write.table(data.frame(cluster1$cluster), file="salidas.txt")

head(datos)


#usando k-means
cluster2=kmeans(datos,centers = 3)
```