Universidad del Valle Facultad de Ingeniería Escuela de Estadística Programa Académico de Estadística Diseño y Análisis de Experimentos Cali, marzo de 2019

PARCIAL PRACTICO

KEVIN STEVEN GARCÍA^a, ALEJANDRO VARGAS^b

1. DESARROLLO

INTRODUCCION:

Para esta practica decidimos trabajar sobre la base de datos correspondiente a la .^{En}cuesta Nacional de Uso del Tiempo"(ENUT: 2016 - 2017), al ser una encuesta tan extensa decidimos tomar solo los datos correspondientes a el ÇAPITULO B: CONDICIONES DE LA VIVIENDA". el cual cuenta con mas de 44 mil datos de hogares y 14 variables cualitativas o categóricas.

luego de observar las posibles variables que podíamos utilizar nos quedamos con 9 variables de las cuales 8 corresponde a preguntas sobre las condiciones de vivienda y 1 es la variable REGIÓNçorrespondiente a las 6 regiones de Colombia en las cuales se realizo 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 recoleccion de basuras? (1.Si , 2.No)

P4030S4A1: cuantos 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)

1.

¿Cuantos ejes se generan en el ACM?

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

2.

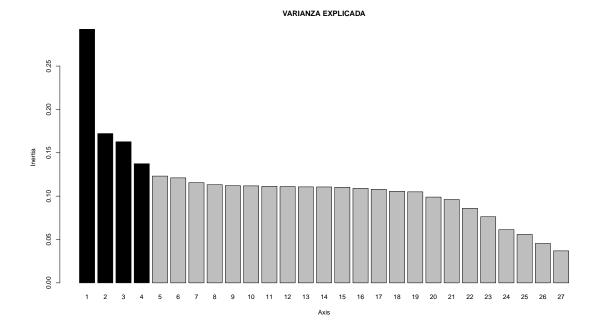
¿Qué relación existe entre la Inercia Total, el número de modalidades y el número de variables cualitativas?

La relación que existe es que el numero de ejes que resultan en un ACM es igual al numero de modalidades menos el numero 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 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: $I_q = Inercia \ por \ pregunta \ y \ I_j = Inercia \ por \ modalidad$

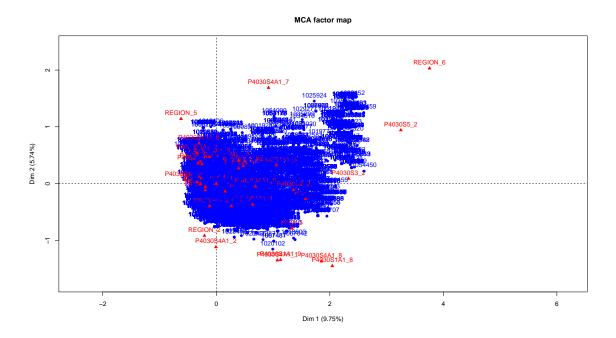
^aCódigo: 1533173. E-mail: kevin.chica@correounivalle.edu.co

^bCódigo: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co



3. ¿Qué similitudes existen entre las filas?

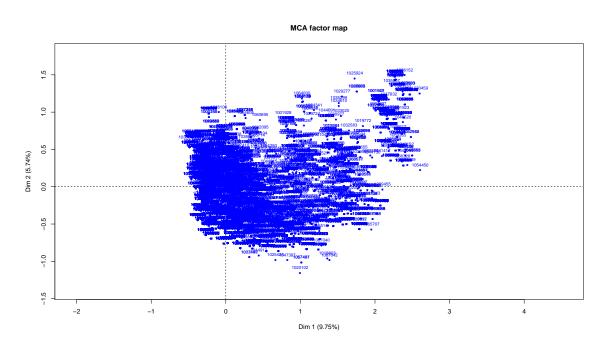
Las filas están dadas por los hogares encuestados, en nuestro caso dos hogares son similares o se parecen si ambos han seleccionado las mismas modalidades, por ende pasamos a evaluar que similitudes hay entre los hogares:



En principio las gráficas se ven algo cargadas ya que estamos trabajando con mas 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 mas grande, el cual es el que esta mas cerca del centro, en este se encuentran las respuestas afirmativas a varias preguntas sobre si se cuenta con servicios públicos y las regiones mas grandes que son 1. Atlántico y 3. Oriental, 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 Andres, esta región al

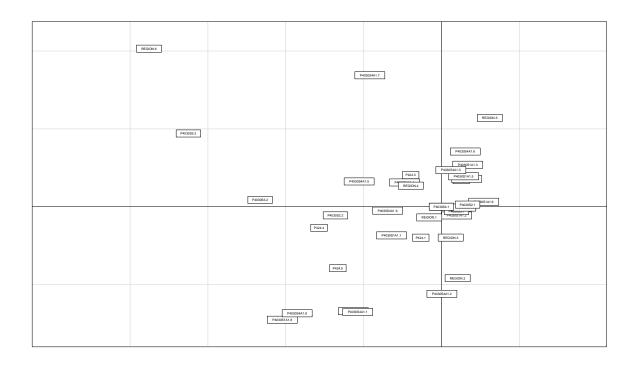
ser una isla es la que tiene menos hogares, al igual que los estratos por encima del 3, resultan ser los menos frecuentes.

4. Con base a la nube de individuos cuales son los valores excéntricos? Evaluamos el gráfico 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 otros varios que ademas de no contar con varios servicios están mas asociados a la región de San Andres ay que es la que menos habitantes tiene.

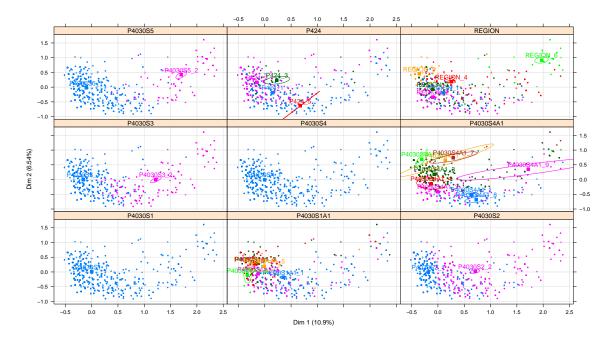
5. Con base a la nube de individuos cuales son los valores excéntricos?



Como se aprecia el formato del gráfico cambio, esto se debe a que fue realizado con el paquete

ade4 y los anteriores fueron realizados con el paquete "FactoMnineR", aun así las interpretaciones son las mismas sin embargo ade4 resulta ser mas amigable en algunos gráficos a la hora de interpretar.

En el gráfico de las modalidades 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, osea 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.



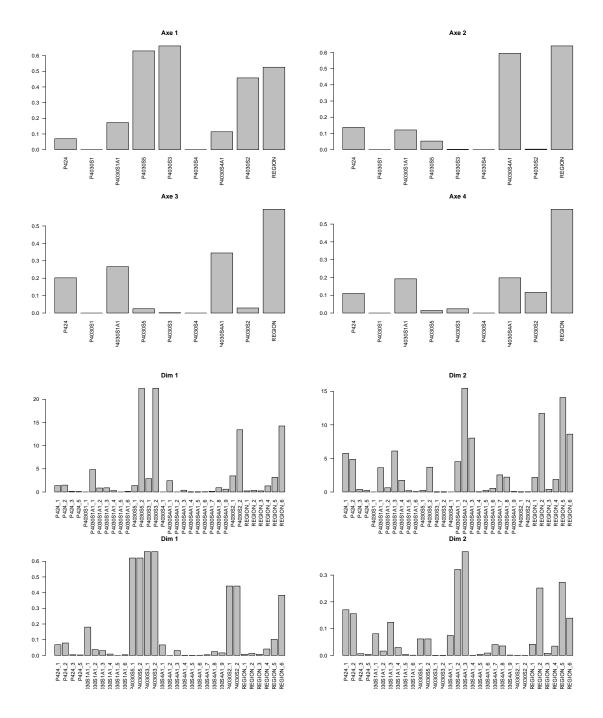
El gráfico anterior nos muestra las modalidades que tienen diferencias significativas, lo primero que notamos es que la región 6 es la mas alejada del resto de las regiones, la región 1,2 y 3 son las mas parecidas. Vemos que la mayoría de hogares cuenta con acueducto, alcantarillado y todos cuentan con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

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 modalidad 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 opto por Barplot de los cosenos cuadrados y las contribuciones.

Este dio como resultado para los dos primeros ejes que las modalidades que mas 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 mas 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 3er hogar tiene un valor de aproximadamente 0.08 el cual es bajo.

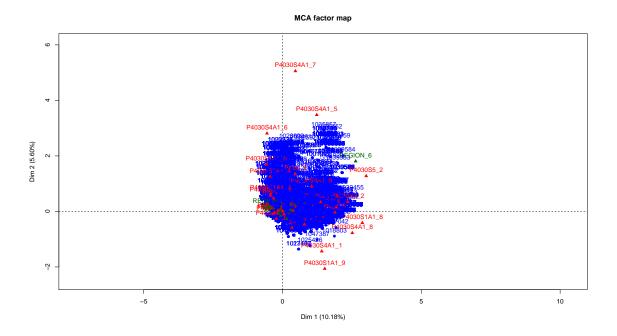
CONTRIBUCIONES			
Hogar	Dim 1	Dim 2	Dim 3
1041524	0.9665660	0.5168859	0.4833209
1038213	0.90955947	0.40907346	0.49802097
1019294	0.97264253	0.29979671	0.10706549
1039511	0.96740190	0.04157307	0.14073777

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

Utilizar variables suplementarias o individuos suplementarios si se requiere.

7.

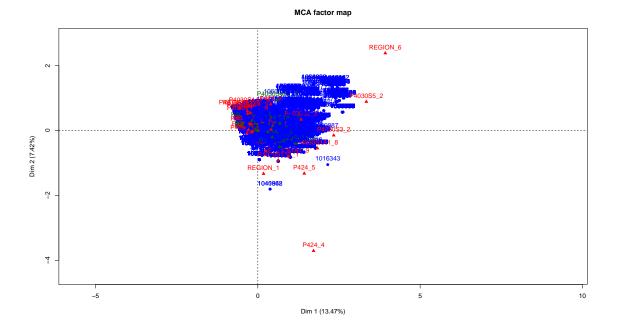
El ACM se realizó teniendo en cuenta todas la variable se individuos sin embargo se podría haber tomado varias preguntas como suplementarias, ente este punto veremos como se comporta el ACM si tomamos la pregunta Región y la Pregunta referente al día en que se recolecta la basura como variables suplementaria.



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 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

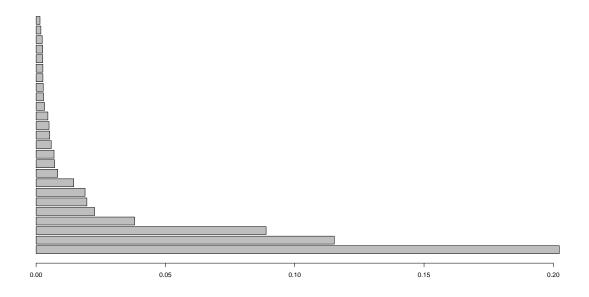
vemos que de haberla tomado como variable suplementaria o de no haberla incluido los datos logran verse mas agrupados en el centro,, sin embargo la región sigue llevándose gran parte de los hogares, en especial la región 6.



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.

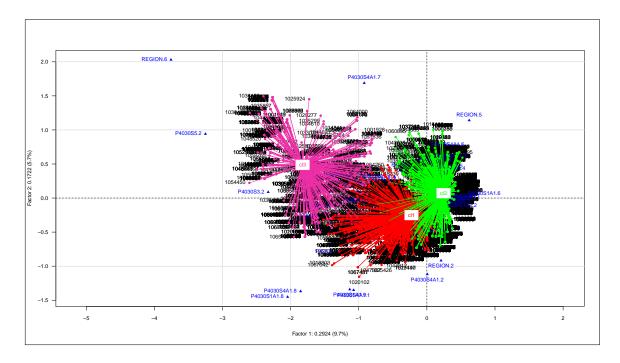
Primero miramos cuantos posibles cluster realizar, o cuantos grupos se pueden formar:



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 cluster.

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 Andres, y esta mas relacionado con modalidades poco frecuentes.

CONCLUSIONES.

Para concluir podemos decir que la mayoría de los hogares en Colombia cuentan con todos los servicios públicos, ademas que los estratos 2 y 3 aunque son çlase 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 y que todos los hogares encuestados en todo el país cuentan con servicios de recolección de basuras y acueducto.

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 analisis por lo que podría quitarse.