Análisis Factorial de Correspondencias

Hugo César Octavio del Sueldo

11/14/2020

Introducción

Pretendemos conocer la relación o asociación existente entre los partidos políticos y las `clases de trabajadores` divididos en quintas partes según su situación laboral de la siguinte forma:

- Trabaja: Trabajador por cuenta ajena
- Doméstico
- Parado
- Estudiante
- Jubilado.

Los `partidos políticos` a analizar seran los siguientes:

- PP: Partido Popular
- PSOE: Partido Socialista Obrero
- UP: Unidas Podemos
- Cs: Ciudadanos
- Resto: Otros Partidos Políticos

Para ello, haremos uso de una encuesta de 11.610 españoles mayores de edad, separados en virtud de su partido preferido y su situación laboral (nota: a partir de los datos de SocioMétrica - El Español, enero de 2018).

Exploratory Data Analysis

Comenzaremos con un análisis exploratorio de datos para comprender mejor la relación existente entre las variables del fichero a utilizar.

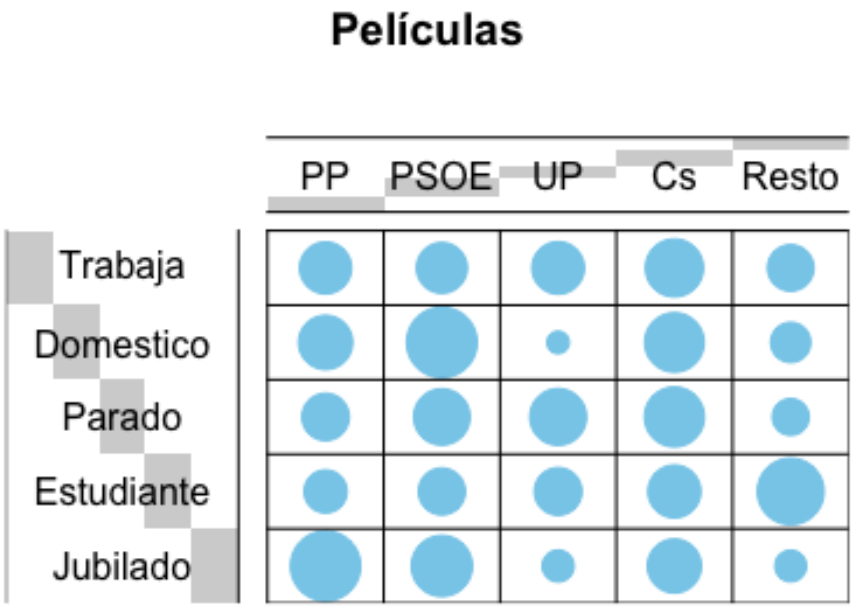
```
            PP PSOE  UP  Cs Resto
Trabaja    462  441 471 576   369
Domestico  502  857  83 606   274
Parado     383  544 551 616   230
Estudiante 316  376 388 478   762
Jubilado   846  639 172 499   169
```

De la inspeccion visual del fichero podemos observar que los jubilados tienden mayoritariamente a elegir el partido PP. A su vez, los empleados domesticos suelen elegir

en mayor proporcion al partido PSOE. Por otro lado, los estudiantes apoyan mucho más a otro partidos por fuera de PP, PSOE, UP y Cs.

Plots

Vamos a realizar un grafico para poder interpretar mejor los datos exhibidos arriba. Para esto utilizaremos balloonplot que nos permite visualizar muy bien estas relaciones.



Con el baloonplot podemos observar, por el tamaño de los circulos, la importancia relativa de los tipos de trabajadores con los partidos políticos. Asi vemos que Jubilados tienen mucha relacion con el partido PP, Personal Domestico se lo puede asociar con PSOE y los estudiantes con el Resto de partidos politicos. A su vez, observamos que los trabajadores por cuenta ajena no estan muy relacionados a ningún partido político en concreto

Análisis de Factorial de Correspondencias

Una vez observada la existencia de relaciones entre partidos politicos y sus votantes, el análisis de correspondencias nos permitirá identificar cuáles son de una manera sencilla en un espacio de dos dimensiones. El primer paso será, realizar la prueba Chi-cuadrado para saber si hay o no independencia entre las variables a análizar.

Contraste de independencia Chi cuadrado

```
 Pearson's Chi-squared test

data:  data
X-squared = 1704.3, df = 16, p-value < 2.2e-16
```

Al ser el p-value menor a 0.05 podemos concluir que se rechaza la hipotesis de independencia que plantea la prueba y confirmar la existencia de alguna relación entre filas y columnas.

Análisis de correspondencias

Su objetivo principal es representar en un espacio de dos dimensiones un conjunto de observaciones dadas en un espacio de dimensión mayor respetando las posiciones relativas entre los elementos; esas posiciones relativas están relacionadas con el grado de similitud de las variables, esto es, con su grado de asociación. El análisis de correspondencias busca determinar las combinaciones de las variables originales que pemitan la mejor representación posible en un espacio de dos dimensiones.

```
summary(votantes.afc, nb.dec = 2, ncp = 2)
```

```
##
## Call:
## CA(X = data, graph = FALSE)
##
## The chi square of independence between the two variables is equal to 1704.
298 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4
## Variance                0.09   0.04   0.02   0.00
## % of var.              64.65  24.45  10.79   0.11
## Cumulative % of var.   64.65  89.10  99.89 100.00
##
## Rows
##              Iner*1000   Dim.1   ctr  cos2   Dim.2   ctr  cos2
## Trabaja    |      7.86 |  0.14  4.33  0.52 |  0.13  9.20  0.42 |
## Domestico  |     30.66 | -0.30 18.80  0.58 | -0.18 17.11  0.20 |
## Parado     |     19.47 |  0.09  1.61  0.08 |  0.29 45.65  0.84 |
## Estudiante |     51.85 |  0.46 43.70  0.80 | -0.22 27.90  0.19 |
## Jubilado   |     36.95 | -0.39 31.56  0.81 | -0.02  0.14  0.00 |
##
## Columns
##              Iner*1000   Dim.1   ctr  cos2   Dim.2   ctr  cos2
## PP         |     28.84 | -0.30 20.77  0.68 | -0.01  0.03  0.00 |
## PSOE       |     21.35 | -0.25 16.35  0.73 | -0.06  2.41  0.04 |
## UP         |     40.90 |  0.39 23.25  0.54 |  0.36 52.25  0.46 |
## Cs         |      2.45 | -0.02  0.13  0.05 |  0.06  2.05  0.30 |
## Resto      |     53.25 |  0.49 39.51  0.70 | -0.32 43.26  0.29 |
```

Tras el cálculo del análisis de correspondencias podemos observar que en dos dimensiones podemos explicar el 89.10% de la variabilidad total.

Interpretación del análisis de correspondencias

Nivel de asociación entre filas y columnas

La primera etapa del ANACOR pasa por conocer si se da o no una asociación significativa entre filas y columnas; para ello, podemos emplear dos métodos alternativos:

1. La traza;
2. El estadístico chi cuadrado La traza, o inercia total de la tabla, es la suma de todos los autovalores; su raíz cuadrada puede interpretarse como el coeficiente de correlación entre filas y columnas; se calculará como sigue:

```
autov = get_eigenvalue(votantes.afc)
traza = sum(autov[,1])
cor.coef = sqrt(traza)
cor.coef
```

```
## [1] 0.3831393
```

En general, como regla empírica, suele emplearse 0.2 como umbral por encima del cual la correlación puede considerarse como importante. En nuestro caso, el valor alcanzado de 0.38 señala una asociación pero cercana al umbral. En este caso, emplearemos como complemente el estadístico chi-cuadrado para confirmar esta asociación.

El estadístico chi cuadrado mostrado previamente en el resumen ofrecido en summary.CA() , que con un valor en nuestro caso de 1704 con un nivel de significación de 0 nos lleva a rechazar la hipótesis de independencia de filas y columnas, permitiendo continuar con el análisis.

Autovalores y gráfico de sedimentación

Las dimensiones con las que finalmente debemos trabajar en la solución se determinará a partir del examen de los autovalores. Dado que la traza, como señalamos previamente, es la suma total de los autovalores, para un eje determinado la relación del valor propio respecto de la traza proporciona lo que denominamos el porcentaje de varianza (o inercia total, o chi cuadrado) explicada por el eje.

En nuestro caso, dos únicas dimensiones o ejes permiten explicar el 89.10% de la varianza;

Una posibilidad para determinar el número de dimensiones a retener es emplear el screeplot o gráfico de sedimentación:

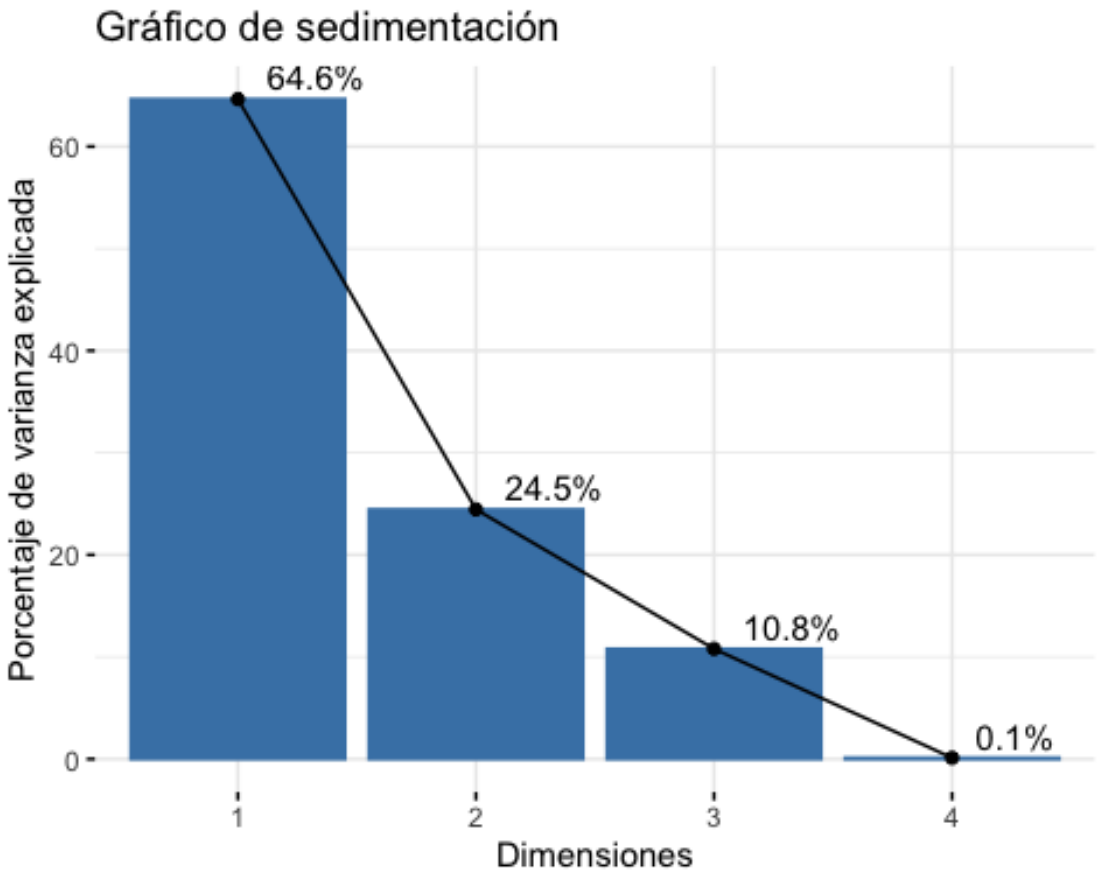


Grafico de sedimentacion

La idea con el grafico de sedimentacion es elegir en base a la regla del angulo o del codo. El criterio es quedarse con un numero de componentes que excluya los asociados a valores pequenos y aproximadamente del mismo tamaño. La regla consiste en seleccionar todas las componentes hasta el primer codo (sin incluirlo).

En nuestro caso, eligiriamos hasta la segunda componente principal. Aqui la primera CP representa 64,6% de la variabilidad total de las variables y la segunda componente el 24.5%. Es decir, en total estamos explicando el 89.1% de la variabilidad total solo con dos dimensiones.

Gráfico de dispersión del análisis de correspondencias entre filas y columnas.

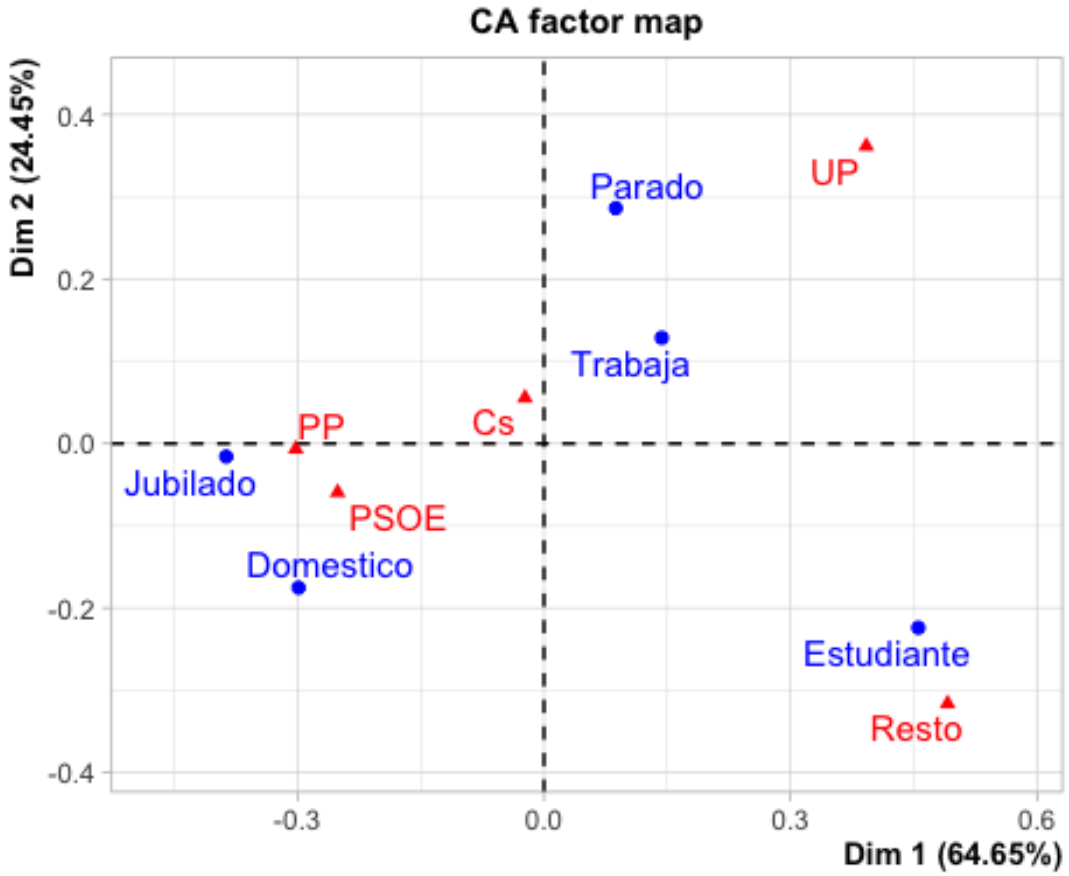


Grafico de disperson del AC entre filas y columnas

Este tipo de representación se denomina *simétrica*; las distancias entre los elementos representados (de filas o de columnas) dan idea de su similaridad/disimilaridad, de forma que puntos cercanos indican relaciones más fuertes que puntos lejanos. Es importante indicar que la mayor o menor cercanía entre puntos de filas y de columnas no puede interpretarse del mismo modo; para conocer la asociación entre filas y columnas debe acudirse a la representación asimétrica, en la que las filas se representan en el espacio de las columnas y viceversa.

A traves de este grafico, si consideramos las filas, podemos analizar que Estudiantes y personal Domestico estan un poco en tierra de nadie al no estar cerca de ninguno de los ejes. Es decir, que no se representa definidamente ni por la dimension 1, ni por la dimension 2. Por otro lado, podemos identificar que los Jubilados estan mejor representados por la dimension 1. A su vez, Los parados y Trabajadores por cuenta propia se los podria identificar mejor por la dimension 2

Por otro lado, por las columnas, podemos observar que UP y Resto de partidos politicos no se ven muy definidos por ninguna de las dimensiones elegidas. Al mismo tiempo, PP y PSOE estan mas asociados a la dimension 1 mientras que Cs a la dimension 2.

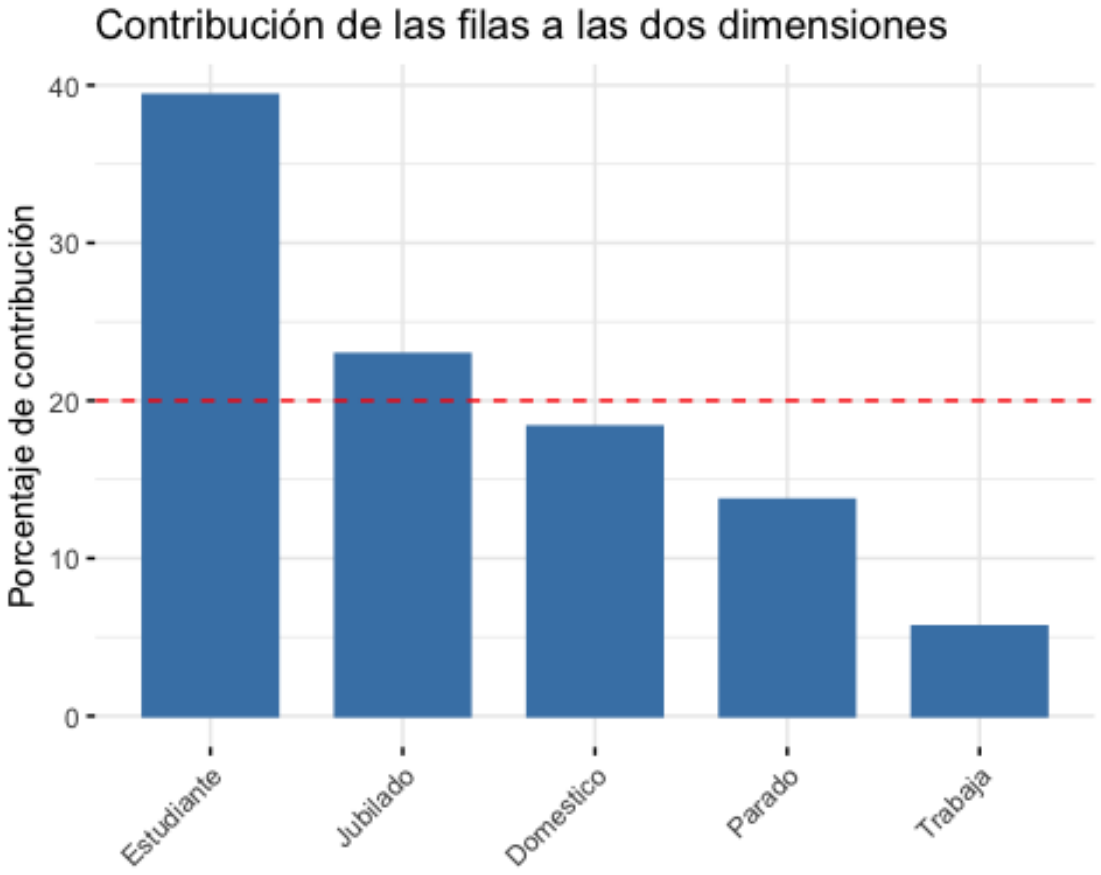
Contribución de filas y columnas

Necesitamos conocer qué filas (y columnas) son las que más y mejor definen las dimensiones o ejes que hemos retenido en el análisis.

Contribución de las filas a cada dimensión

Las filas con mayor peso / mayor indicador más contribuyen a la explicación del eje; en nuestro caso, Estudiantes y Jubilados, con un 43% y 31%, son las que más contribuyen a la explicación del eje 1. Mientras que los Parados seguidos por los estudiantes son los que mas contribuyen a la explicacion de la dimensión 2

Podemos visualizar esta contribución a los dos ejes mediante un gráfico de barras

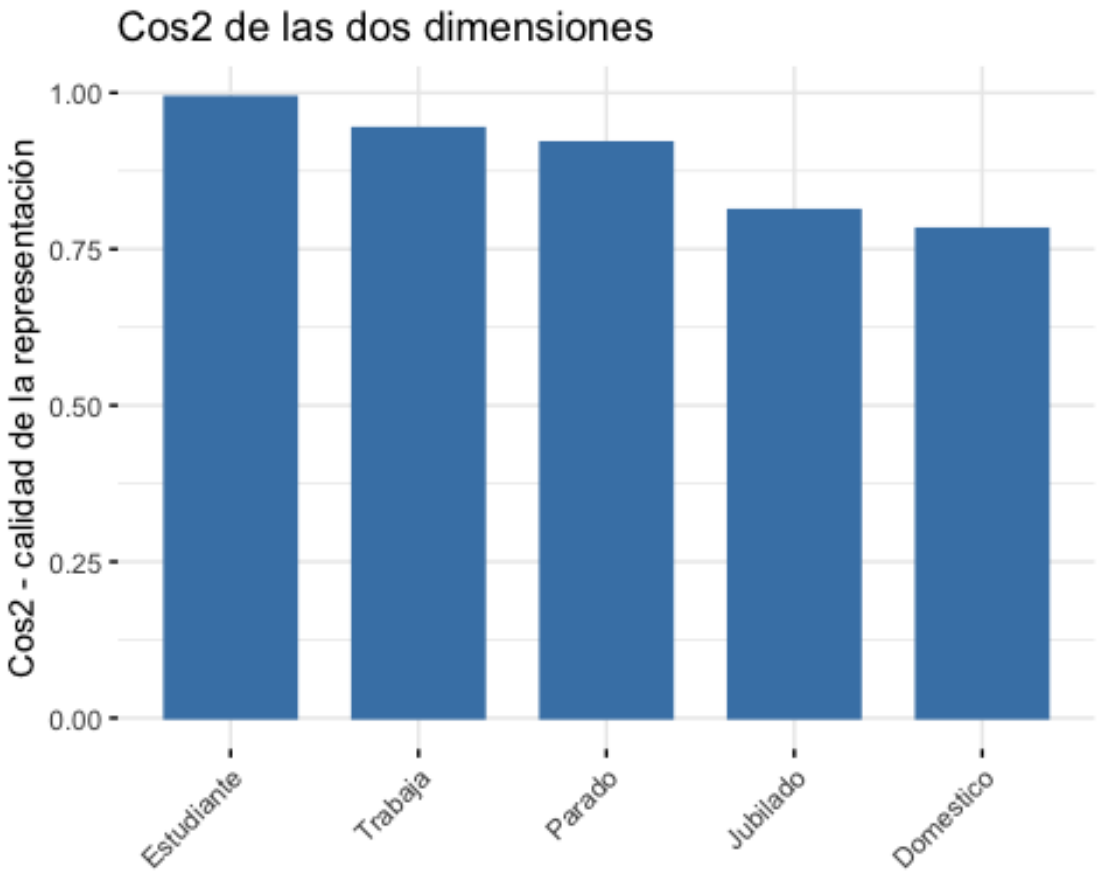


Aqui podemos apreciar que no existe independencia entre las varibles ya que en ese caso todas las dimensiones estarian sobre la linea punteada de color rojo. Aqui vemos que los estudiantes concentran la mayor contribucion sobre las dos dimensiones seguidas mas abajo por los jubilados.

Calidad de la representación de las filas: el Cos2

El cos2 o cuadrado del coseno o cuadrado de las correlaciones es la principal medida de la calidad de la representación alcanzada; mide la asociación entre filas (o columnas) y un determinado eje.

En este caso, Jubilados y Estudiantes están muy asociadas a la primera dimensión, como Parados lo están a la segunda. La suma de los cos2 de cada punto de fila (o columna) a lo largo de las distintas dimensiones es 1. La calidad de la representación de un punto de fila (o columna) en un espacio de n dimensiones viene dada por la suma de los cos2 de ese punto de fila a lo largo de las n dimensiones; es evidente que, al elegir un total de 2 dimensiones en nuestro caso, la representación no es perfecta.



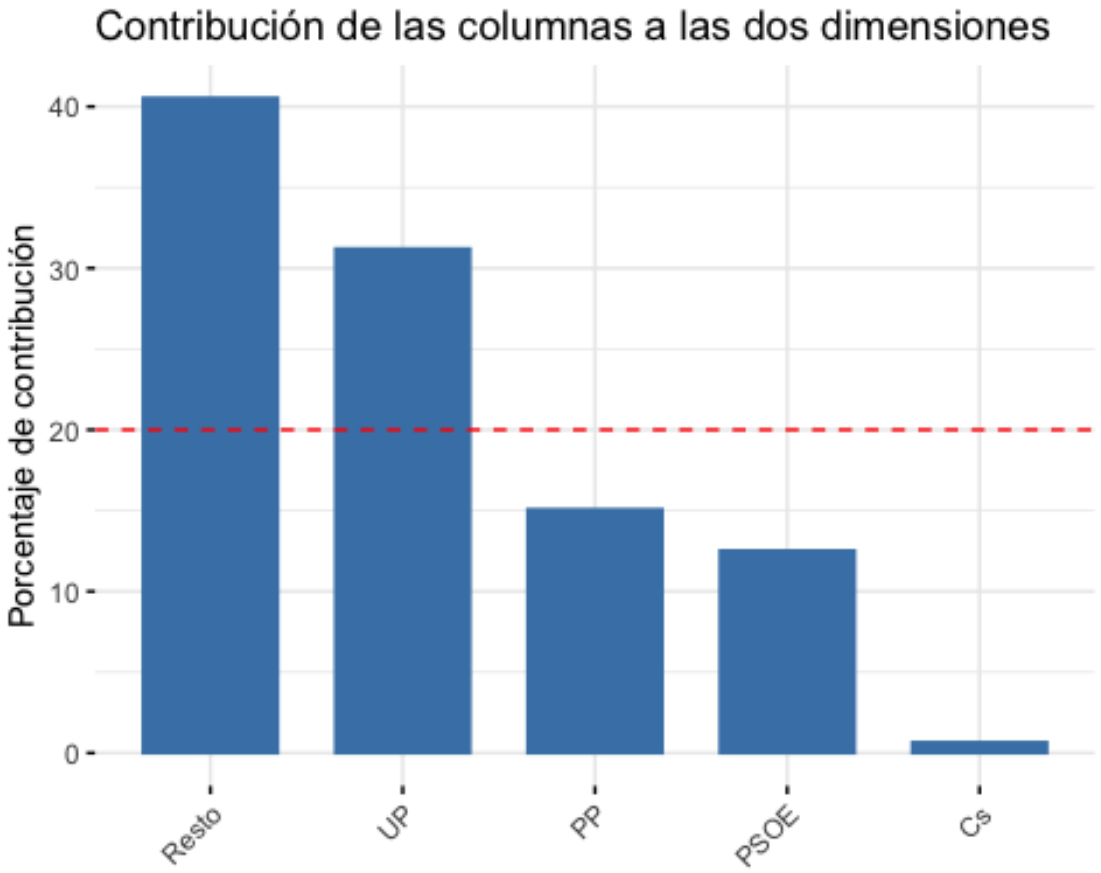
Aqui podemos apreciar que la calidad de la representacion de los estudiantes es casi maxima en las dos primeras dimensiones ya que esta variable se ve explicada casi en su totalidad por estas dos dimensiones.

En el caso del resto de variables no se llega al maximo porque su explicación esta más dispersa entre el resto de componentes principales.

Contribución de las columnas a cada dimensión

Del análisis realizados observamos que UP y el Resto de partidos politicos son las que más contribuyen a la dimension 2. Por otro lado, el Resto de partidos politicos, seguido por UP, PP y PSOE contribuyen a la dimension 1 pero ninguna con gran diferenciación.

Podemos visualizar esta contribución mediante un gráfico de barras:



Contribuciones de las columnas en las dos dimensiones

Aca podemos observar que el resto de partidos politicos son los que mas contribuyen a las dos primeras dimensiones, seguido por UP.

Calidad de la representación de las columnas: el Cos2

De nuestro análisis observamos que PSOE, Resto y PP están bastante asociadas a la primera dimensión, pero ninguno de los Partidos politicos están muy asociados a la segunda.

La suma de los cos2 de cada punto de fila (o columna) a lo largo de las distintas dimensiones es 1. La calidad de la representación de un punto de fila (o columna) en un espacio de n dimensiones viene dada por la suma de los cos2 de ese punto de fila a lo largo de las n dimensiones; es evidente que, al elegir un total de 2 dimensiones en nuestro caso, la representación no es perfecta.

Podemos visualizarlo con un grafico de barras como lo hicimos con las filas

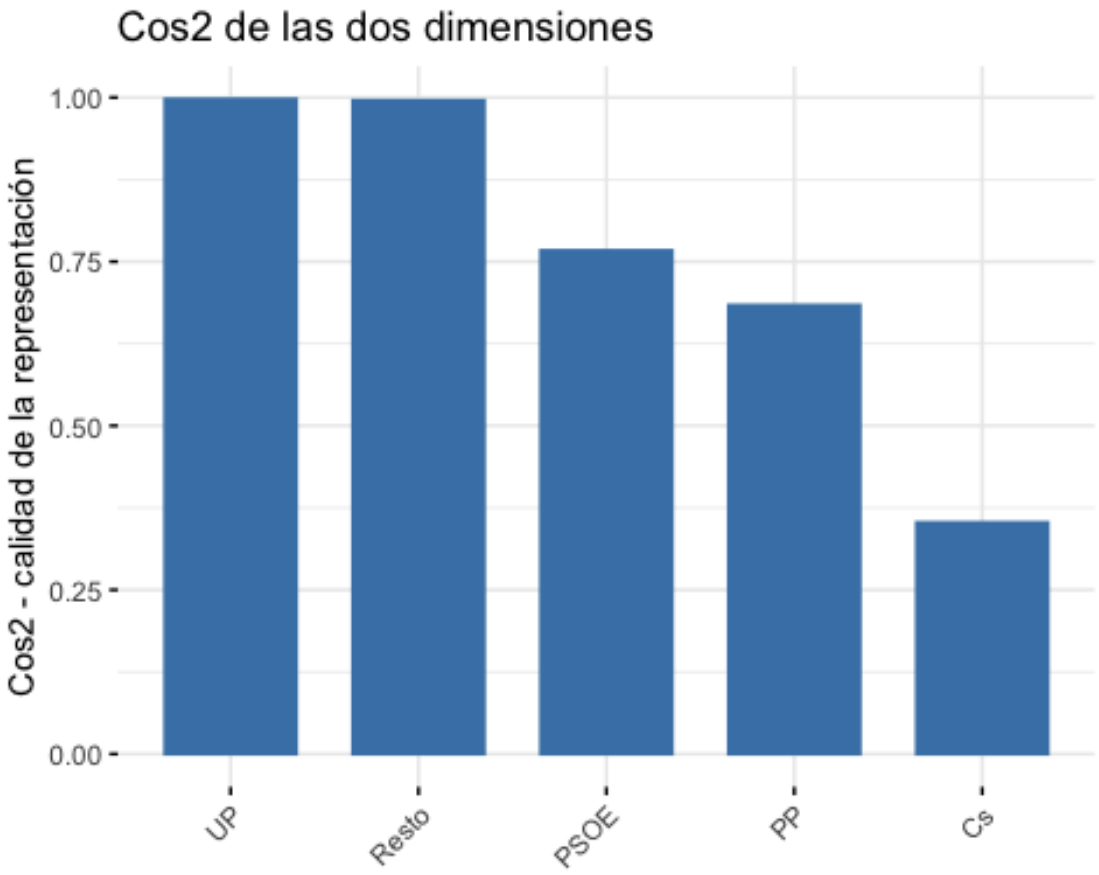


Grafico de barras del cos2 de las columnas en las dimensiones

Aqui podemos observar que la calidad de la representacion de cada punto columna en las dimensiones no es perfecta. Solo UP y Resto llegan casi al maximo de la representacion pero el resto de partidos politicos no ya que al elegir solo dos dimensiones solo estamos explicando un 89% de la variabilidad total. Es evidente que en el resto de dimensiones se explican el resto de las variables que no estan explicadas en su totalidad en las primeras dos dimensiones.

Gráfico asimétrico

En el caso de un gráfico *asimétrico*, los puntos de filas (o de columnas) se representan a partir de las coordenadas estándar, S, y los perfiles de la otra parte a partir de las coordenadas principales, P. Para un determinado eje, la relación entre S y P viene dada por

P=(autovalor^1/2)×S

siendo P la coordenada principal de la fila (o la columna) en el eje, y autovalor el correspondiente del eje.

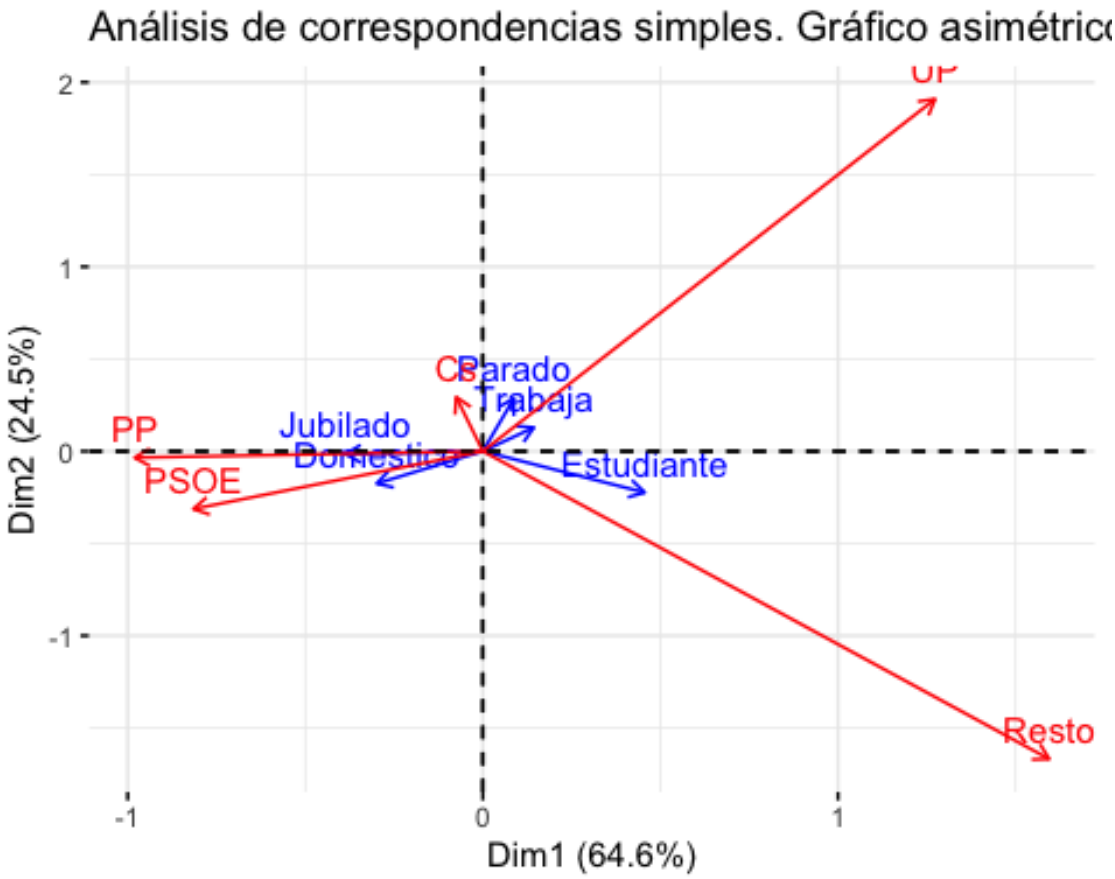


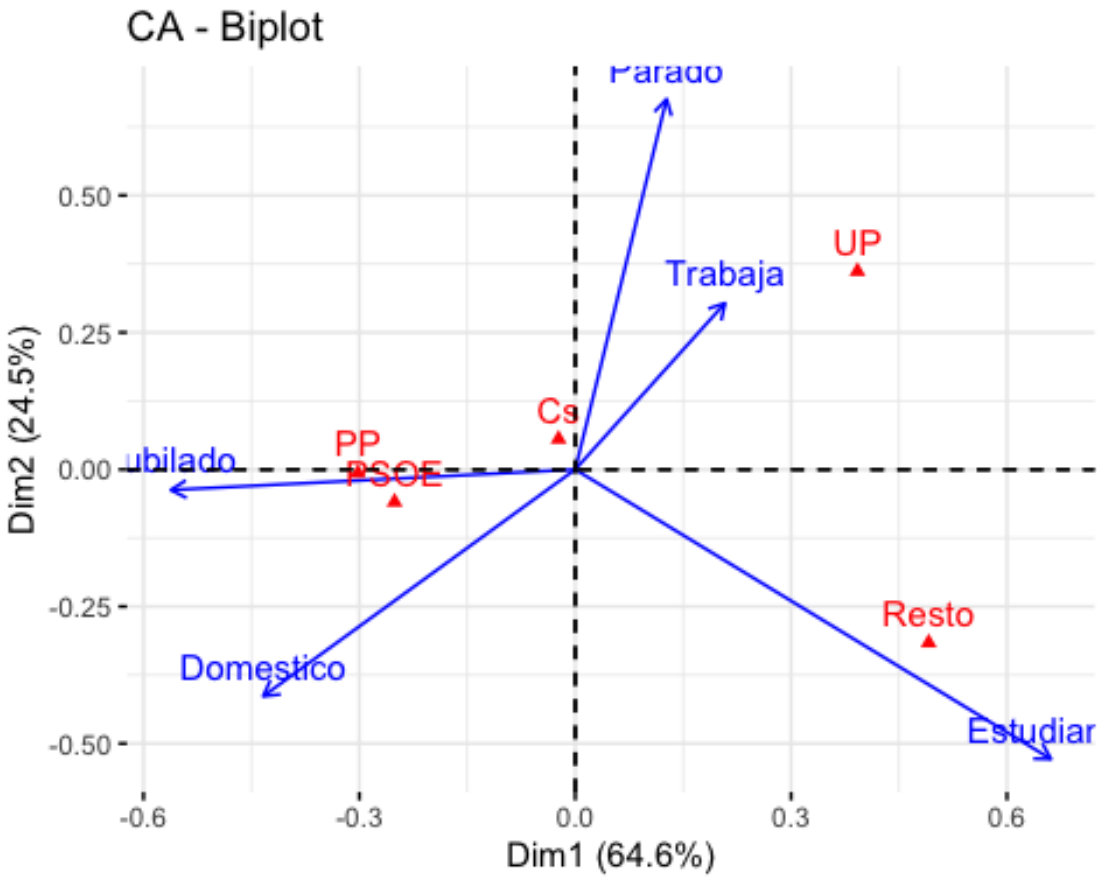
Grafico asimetrico de las filas y columnas en las dimensiones

Un ángulo agudo señala una alta asociación entre el punto de fila y el de columna; la distancia entre filas y columnas se interpreta mediante la proyección ortogonal de los puntos de fila en la flecha de la columna.

Explicado esto, podemos decir que existe una alta correlacion entre jubilado y PP, personal domestico y PSOE, Trabajadores por cuenta ajena y UP y finalmente estudiantes y el Resto de partidos politicos.

Gráfico de contribuciones

Esta representación permite conocer o, mejor, visualizar, la contribución de los puntos de fila y/o columna a los ejes, algo complicado en la solución ofrecida por el gráfico simétrico habitual.



La posición de los puntos de columna en el gráfico anterior no se ve modificada en relación con el gráfico tradicional; sin embargo, las distancias de los puntos de fila están ahora relacionadas con sus contribuciones al mapa factorial de dos dimensiones. Cuanto menor sea la distancia (angular) de un punto de fila respecto a algún eje, mayor será la contribución del mismo a la definición del eje. Asimismo, una posición intermedia entre dos ejes señala una contribución similar a ambos.

En el caso que nos ocupa, claramente Jubilado contribuyen fundamentalmente a la definición del eje 2, de forma negativa; por su parte, Parados lo hacen respecto del eje 1, de forma positiva; Estudiantes, personal Domestico, Trabajadores por cuenta ajena, por último, contribuyen de forma similar a ambos.

Conclusiones

Finalizado el análisis factorial de correspondencias podemos confirmar que existen relaciones entre la situación laboral y los partídos políticos. Es decir, si fueramos parte de estos partidos políticos ahora seríamos capaces de saber a que clase de trabajadores apuntar a la hora de las campañas. Si fueramos del PP, nos dirijiriamos principalmente a Jubilados. El PSOE debería hacerlo con el personal doméstico y por último el resto de partídos políticos deberían centrarse en los estudiantes. Por último, al no estar definidios los trabajadores por cuenta ajena por ninguno de los partidos políticos se podría decir que todos los partidos pueden sacar votos de esta clase de trabajadores.