Práctica Final

Estadística Multivariante

Análisis Exploratorio Multivariante

Juan José Herrera Aranda - juanjoha@correo.ugr.es
Pilar Navarro Ramírez - pilarnavarro@correo.ugr.es
Adrián Rodríguez Montero - adrianrm6@correo.ugr.es

14 de enero de 2022

Índice

1. Abstract

El problema escogido es el correspondiente a la tercera base de datos. Se ha realizado un análisis exploratorio univarariante y un análisis exploratorio multivariante. En el primero se han estudiado cada una de las variables de forma independiente incluyendo un análisis descriptivo numérico clásico, y exploración gráfica, entre otras técnicas. En el segundo se han estudiado las variables en conjunto, comprobando además la posible existencia de normalidad multivariante y aplicando otras técnicas de análisis de datos para reducir la dimensionalidad. Finalmente se hace un análisis discriminante para poder llevar a cabo una clasificación.

Para nuestro ejemplo concluiremos que con ACP se obtiene una explicación de la varianza del 80 % frente al 74 % de AF, por tanto, con ACP tenemos mejores resultados. Por otra parte, también se llegará a la conclusión de que el análisis discriminante realizado no tiene mucho sentido ya que se debería de haber hecho con una base de datos etiquetada.

2. Introducción

Por un lado, se ha realizado un análisis exploratorio univariante de los datos, analizando cada una de las variables de forma independiente. Se han mostrado los datos con *summary*, comprobando que hay un valor perdido, pero al suponer menos del 5 % de la cantidad total, lo hemos sustituido por la media correspondiente a ese atributo en vez de eliminarlo, pues al tener pocos datos no es recomendable. Después, se ha realizado un breve análisis descriptivo numérico clásico por cada variable, explicando cómo son cada uno de los 11 histogramas atendiendo a las medidas de centralidad, el coeficiente skewness, la curtosis, etc. Posteriormente se ha realizado una exploración gráfica comprobando que existen outliers mediante la representación de un diagrama de cajas (*boxplot*) y se han sustituido por la media. Por último, se concluye la sección comprobando la normalidad de cada variable para poder aplicar posteriormente ciertas técnicas estadísticas empleándose los gráficos *qqplot* y en caso de duda en alguna variable se emplea el test de **Kolmogorov-Smirnov** para ver si se distribuye según una normal.

Por otro lado, se lleva a cabo un análisis exploratorio multivariante, se comprueba la correlación entre variables. Además con el test de Bartlett se verifica si las muestras provienen de poblaciones con la misma varianza. También se comprueba si hay normalidad multivariante en la muestra aplicando el test de *Henze-Zirkler* y el test de *Mardia*. Por otra parte se analiza si se verifica la homogeneidad de la varianza mediante el test de *Box M*. Tras esto, mediante ACP se reduce la dimensión de los datos buscando combinaciones lineales de las variables observables que maximicen la varianza en cada dirección. Para elegir el número óptimo de componentes principales se aplica la regla de *Abdi et al.* y la regla del *codo*. Posteriormente, se realiza el análisis factorial obteniendo el número óptimo de factores y para ello se aplican los siguientes métodos: el *criterio Scree plot*, el *análisis paralelo* y el *test de hipótesis* mediante la orden *factanal*. A continuación se añade una variable cualitativa con dos categorías para poder efectuar un análisis discriminante. Tras lo anterior, se realiza una exploración gráfica de los datos para saber qué pares de variables separan mejor las dos clases. Por último, se lleva a cabo el análisis discriminante lineal y otro cuadrático obteniendo la función discriminante, la cual determina la categoría de cada observación y se emplea validación cruzada para estimar las probabilidades de clasificaciones erróneas.

El objetivo del trabajo consiste en descubrir patrones, perfiles y tendencias a partir del análisis de los datos utilizando las técnicas estadísticas del análisis multivariante de datos aprendidas a lo largo del curso.

3. Materiales y Métodos

3.1. Materiales

El conjunto de datos (*tercera base de datos*) está constituido por 34 estados del mundo donde se han observado 11 variables cuantitativas, las cuales están estandarizadas:

- Ztlibrop: Número de libros publicados.
- Ztejerci: Cociente entre el número de individuos en ejército de tierra y población total del estado.
- Ztpobact: Cociente entre población activa y total.
- Ztenergi: Tasa de consumo energético.
- Zpservi: Población del sector servicios.
- Zpagricu: Población del sector agrícola.
- Ztmedico: Tasa de médicos por habitante.
- Zespvida: Esperanza de vida.
- Ztminfan: Tasa de mortalidad infantil.
- Zpobdens: Densidad de población.
- Zpoburb: Porcentaje de población urbana.

```
   ZPOBDENS          ZTMINFAN          ZESPVIDA          ZPOBURB      
Min.   :-1.0778   Min.   :-1.1026   Min.   :-2.1453   Min.   :-1.7697  
1st Qu.:-0.8497   1st Qu.:-0.9586   1st Qu.:-0.7460   1st Qu.:-0.7320  
Median :-0.1616   Median :-0.3931   Median : 0.2781   Median : 0.1268  
Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
3rd Qu.: 0.5547   3rd Qu.: 0.8985   3rd Qu.: 0.9033   3rd Qu.: 0.8148  
Max.   : 2.8616   Max.   : 1.9048   Max.   : 1.2486   Max.   : 1.5096  

   ZTMEDICO          ZPAGRICU          ZPSERVI           ZTLIBROP      
Min.   :-1.1473   Min.   :-1.2342   Min.   :-1.88521   Min.   :-0.9696  
1st Qu.:-0.8829   1st Qu.:-0.8480   1st Qu.:-0.72858   1st Qu.:-0.9240  
Median :-0.2916   Median :-0.2134   Median : 0.03541   Median :-0.3237  
Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000  
3rd Qu.: 0.8694   3rd Qu.: 0.7115   3rd Qu.: 0.85176   3rd Qu.: 0.7736  
Max.   : 2.3717   Max.   : 1.9052   Max.   : 1.62885   Max.   : 2.4024  
                                                       NA's   :1        

   ZTEJERCI          ZTPOBACT          ZTENERGI      
Min.   :-0.86586   Min.   :-2.1341   Min.   :-0.9507  
1st Qu.:-0.59889   1st Qu.:-0.6735   1st Qu.:-0.7813  
Median :-0.20626   Median :-0.1067   Median :-0.3900  
Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000  
3rd Qu.: 0.07996   3rd Qu.: 0.8789   3rd Qu.: 0.5828  
Max.   : 4.42620   Max.   : 1.7045   Max.   : 2.7498  
```

Figura 1: Se presentan información estadísticas sobre las variables del problema. Esta información incluye: valor mínimo y máximo, media, mediana, primer y tercer cuantil y existencia o no de valores perdidos.

3.2. Métodos estadísticos

Las técnicas estadísticas empleadas han sido:

- Exploración descriptiva: Estudio de las distribuciones de las variables unidimensionales
- Exploración gráfica usando boxplots: para eliminación de outliers
- Método gráficos de comparación de cuantiles, qqplot: para comprobar la normalidad de las variables de manera visual
- Test de Kolmogorov-Smirnov: para comprobar la normalidad de las variables.
- Análisis de correlaciones: para comprobar qué variables están correladas y con cuales
  - Test de Barlett
- Test de Henze-Zirkler: para determinar si hay normalidad multivariante en la muestra
- Test de Mardia: igual que el anterior
- Test de Box M.: para contrastar la hipótesis *La matriz de covarianzas es constante en todas las clases.*
- Análisis de Componentes Principales: Para reducir la dimensión de los datos
  - Regla de Abdi et al.: Para obtener las componentes principales
  - Scree Plot con el método del Codo: Para comprobar la validad de la Regla anterior
- Análisis Factorial: Para reducir la dimensionalidad.
  - Scree Plot con el método del codo: Para obtener el número óptimo de factores
  - Análisis paralelo: Igual al anterior
  - Test de hipótesis: Igual a los anteriores.
- Análisis discriminante: Para clasificación

4. Resultados

4.1. Análisis univariante

En este análisis se ha encontrado un valor perdido en el atributo *Ztlibprop* que representa el 3 % de la cantidad total y se ha sustituido por la media de dicho atributo. Las medidas resistentes de centralidad de casi todas las variables están desplazadas ligeramente a la izquierda a excepción de *Zpservi, Zpesvida y Zpoburbs*, que están suavemente desplazada hacia la derecha. Tenemos también que todas las variables son platicúrticas a excepción de *Zpobdens, Ztjerci y Ztenergi* que son leptocúrticas. Además, *Ztenergi, Ztejerci, Ztlibrop, Zpagricu, Ztmedico, Ztminfan y Ztpobdens* tienen una cola simétrica extendida hacia los valores positivos mientras que el resto la tienen extendida hacia los valores negativos. También se ha detectado la presencia de outliers en los atributos *Zpobdens y Ztejerci* y se han sustituido por la media puesto que no hay suficientes datos como para eliminarlos.

Finalmente, se ha comprobado también que todas las variables siguen distribuciones normales univariantes de manera visual usando el método gráfico de comparación de cuantiles. Para el atributo *Ztejerci* el método anterior generaba cierta controversia, por lo que se empleó el test de Kolmogorov-Smirnov. En dicho test se ha obtenido un p-valor de 0.02069, por tanto al 95 %

de confianza se rechaza la hipótesis nula y la distribución no sigue una distribución normal. Pero se si aumenta el nivel de confianza hasta el 99 % no se puede rechazar dicha hipótesis y por tanto se podría admitir la normalidad.

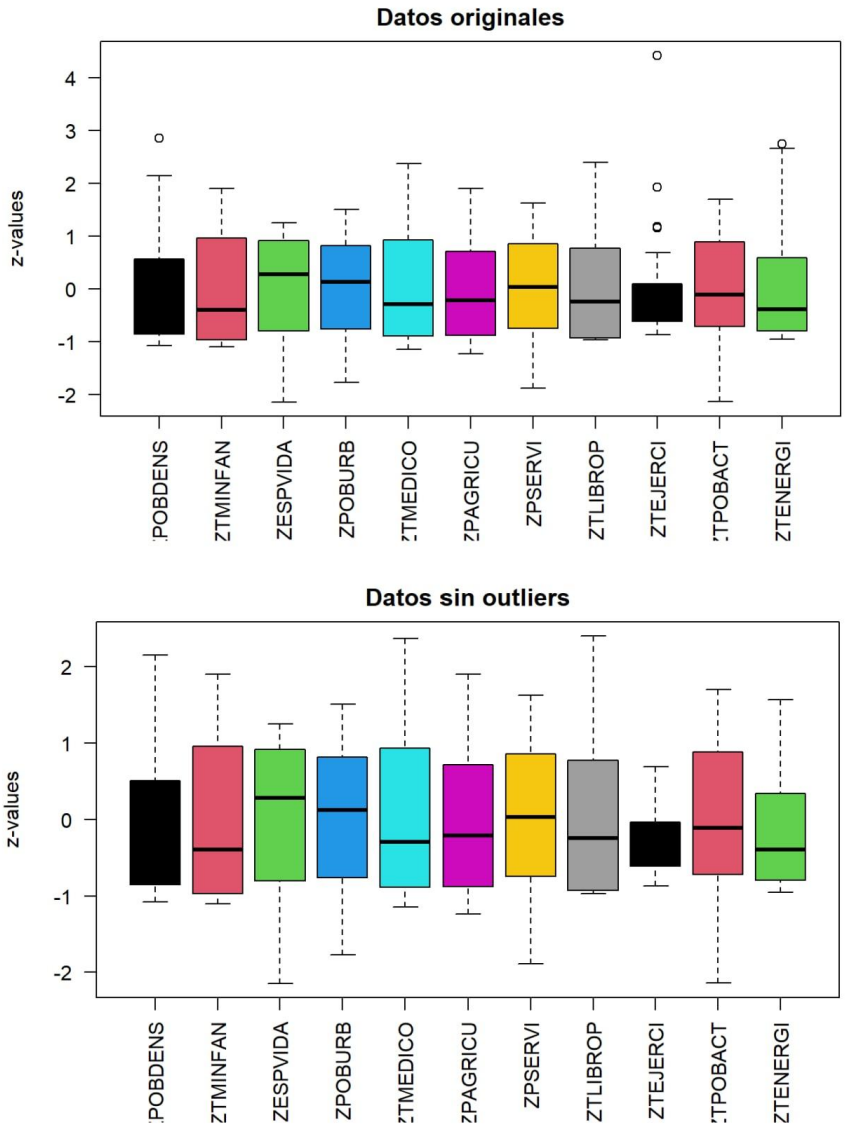

Figura 2: BoxPlots. En la imagen superior se muestran con los outliers y en la inferior sin ellos

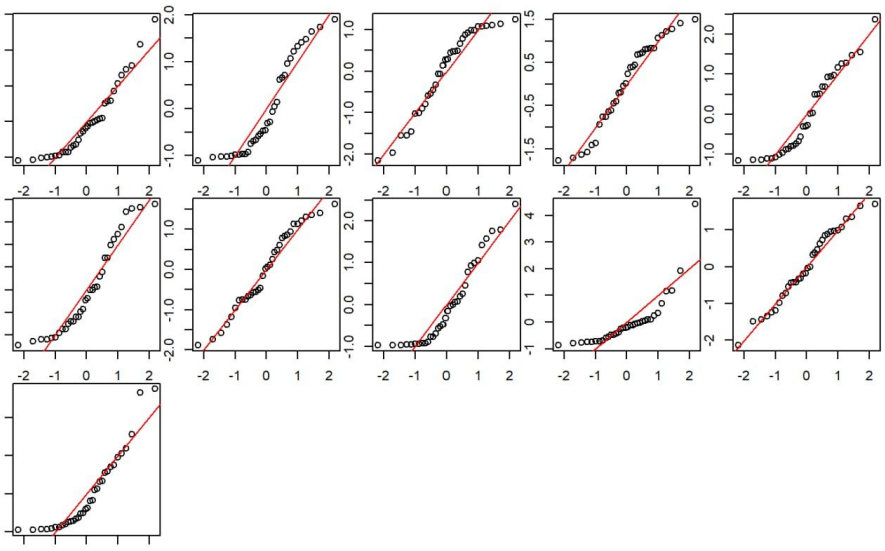

Figura 3: Gráficas generadas al aplicar a las variables el método gráfico de comparación de cuantiles qq-plot. Se observa de manera visual que todas las variables siguen una normal a excepción de la novena, que no queda muy clara.

4.2. Análisis de componentes principales (ACP)

Para llevar a cabo un análisis de componentes principales es necesario que haya correlación entre las variables. Con el fin de determinar esto representamos la matriz de correlaciones:

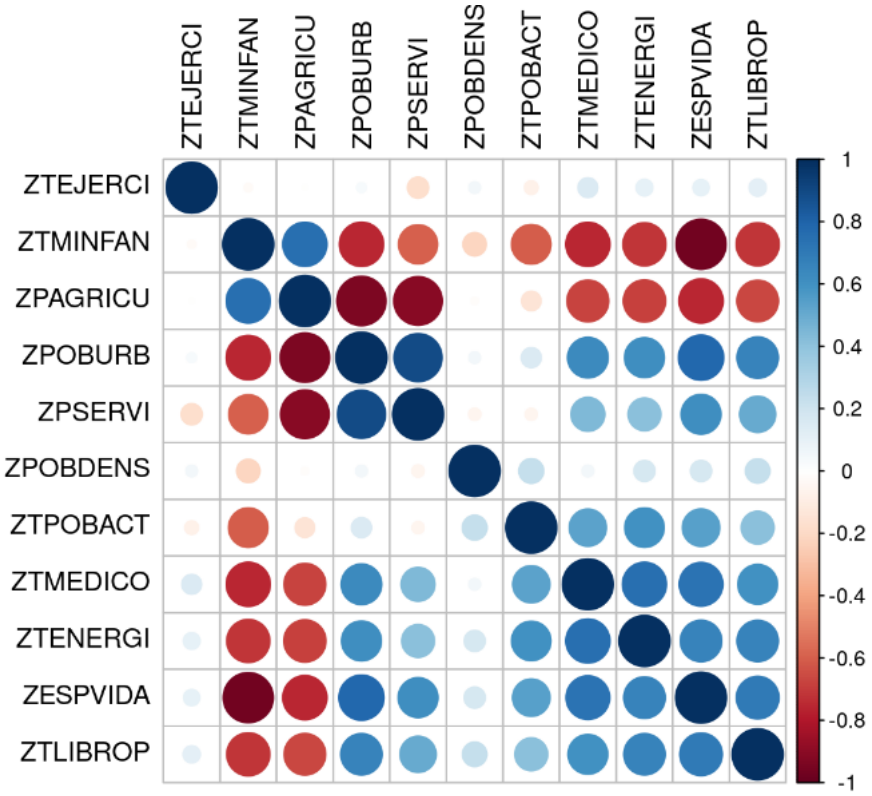

Figura 4: Matriz de correlaciones

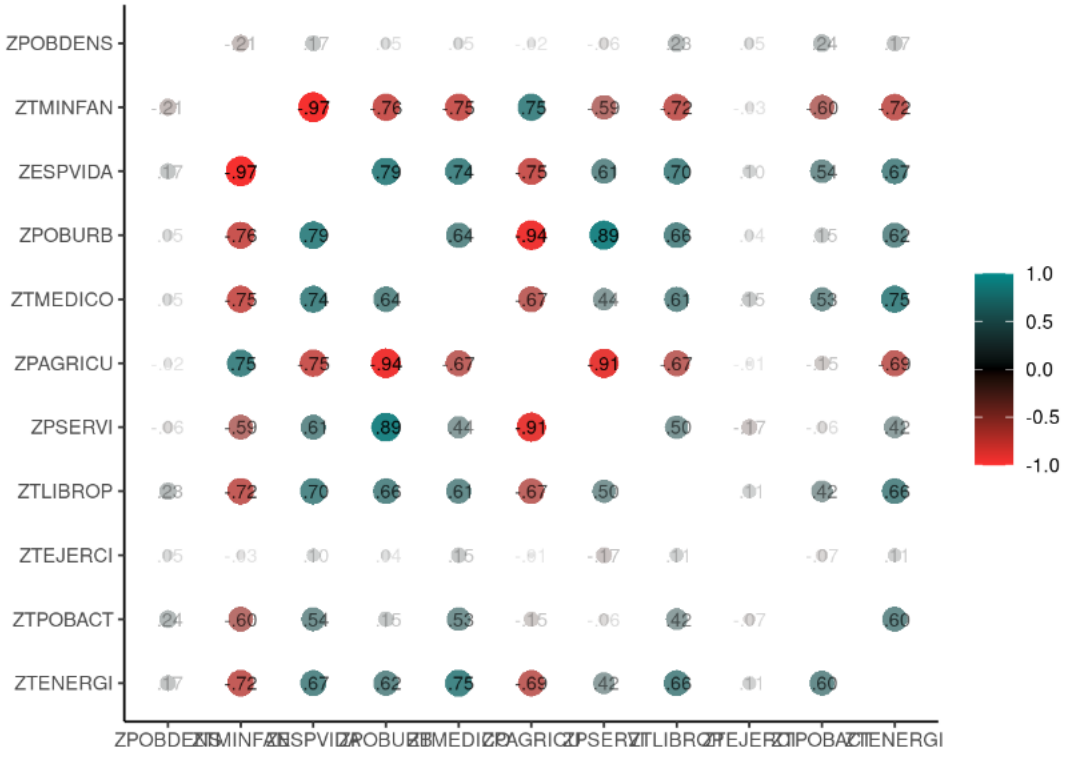

Figura 5: Matriz de correlaciones

Hay cierta correlación entre las variables, ya que la matriz de correlaciones dista de la matriz identidad. Por ejemplo vemos que la variable ZTMINFAN está correlada con la variable ZESPVIDA en un 96.7 % o la variable ZPAGRICU con ZPOBURB en un 93.8 %.

Para comprobar que a nivel poblacional efectivamente están correladas las variables usamos el test de Bartlett. El resultado del mismo nos lleva a rechazar la hipótesis nula, es decir, que las variables son incorreladas, de manera que podemos concluir que las variables están correladas y tiene sentido realizar un análisis de componentes principales.

Tras llevar a cabo un análisis de componentes principales se obtienen 11 componentes, cuya desviación típica, y varianza acumulada se muestran en la siguiente imagen:

```
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7
Standard deviation     2.4818 1.2806 1.03112 0.95154 0.6448 0.60272 0.48187
Proportion of Variance 0.5599 0.1491 0.09665 0.08231 0.0378 0.03302 0.02111
Cumulative Proportion  0.5599 0.7090 0.80568 0.88799 0.9258 0.95881 0.97992
                          PC8     PC9    PC10    PC11
Standard deviation     0.33542 0.24916 0.16420 0.13908
Proportion of Variance 0.01023 0.00564 0.00245 0.00176
Cumulative Proportion  0.99015 0.99579 0.99824 1.00000
```

Figura 6: Componentes principales obtenidas

Vemos a continuación la varianza explicada y varianza acumulada gráficamente.

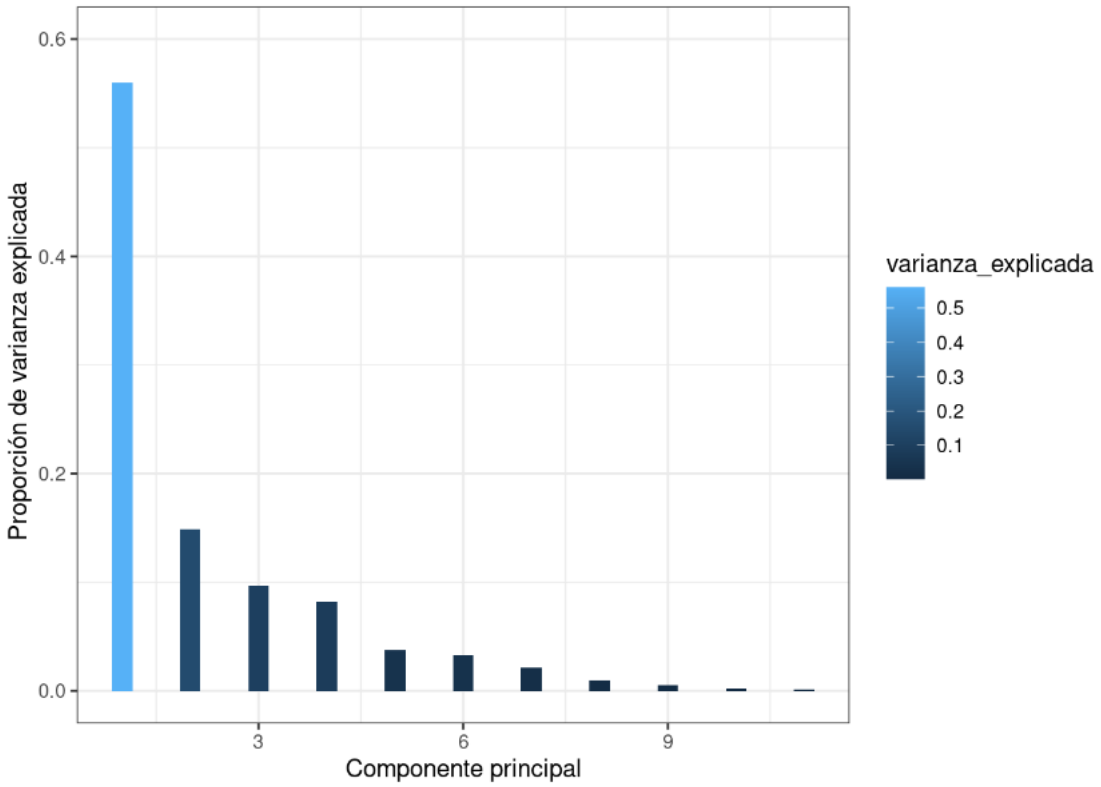

Figura 7: Varianza explicada

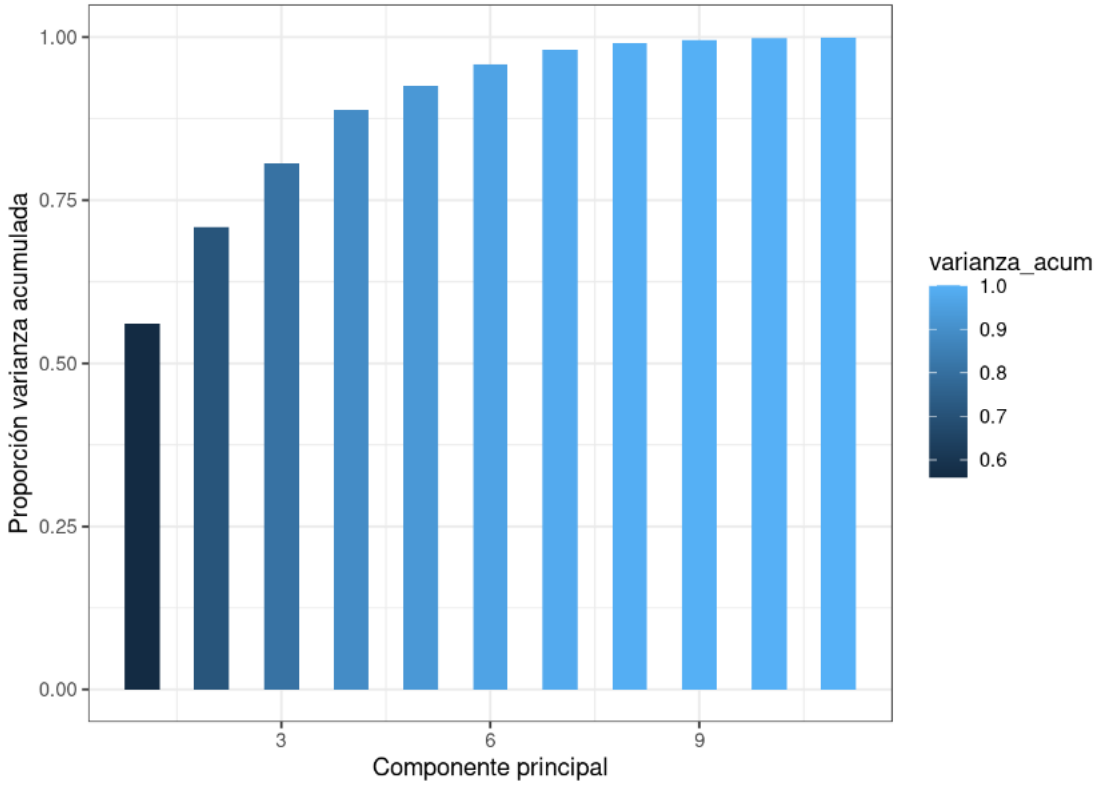

Figura 8: Varianza acumulada

Para elegir el número óptimo de componentes principales usamos la regla de Abdi et al. (2010). La media en nuestro caso es 1, por estar las variables tipificadas, y obtenemos 3 componentes principales con una proporción de varianza explicada por cada una de ellas superior a 1, de manera que nos quedamos con las tres primeras componentes principales. La varianza acumulada es 0.8056758, luego estas componentes pricipales explican el 80,57 % de la varianza total.

Además de esta regla, usaremos una técnica gráfica. En concreto aplicamos el método del codo con ScreePlot:

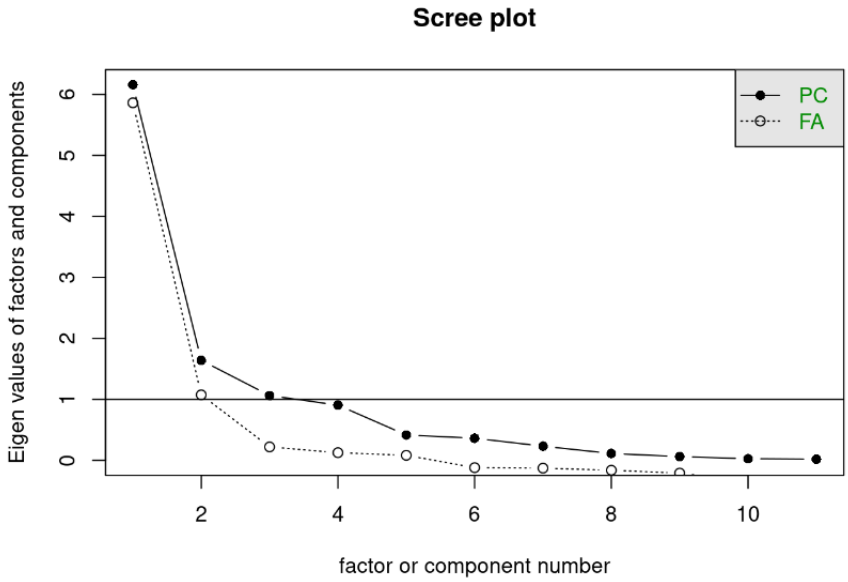

Figura 9

Esta técnica confirma que el número de componentes principales a considerar debería ser 3.

Nos quedamos entonces con tres componentes principales, y se muestran a continuación tres gráficos donde se representan conjuntamente las variables y las observaciones relacionando visualmente las posibles relaciones entre las observaciones, las contribuciones de estas a las varianzas de las componentes y el peso que tienen las variables en la definición de las componentes principales:

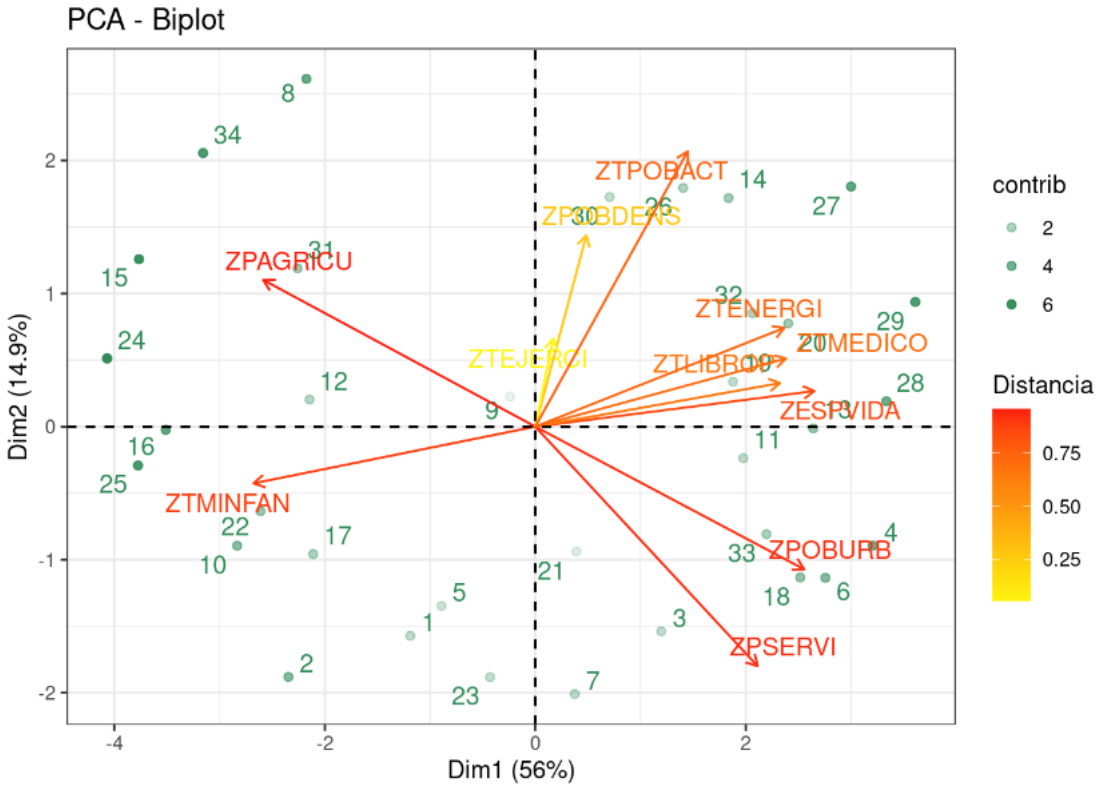

Figura 10: Variables y observaciones en la 1º y 2ª componente principal

Vemos que algunas variables como ZPSERVI o ZTPOBACT están casi correladas por igual con ambas componentes principales. Sin embargo, hay muchas que están más correladas con la primera componente principal como ZTLIBROP, ZTMEDICO y ZESPVIDA, siendo esta última la que presenta una mayor correlación positiva con dicha componente principal, mientras que ZTMINFAN presenta una mayor correlación negativa con la primera componente. Por otro lado tenemos un par de variables, ZTEJERC y ZPOBDENS, que están más correladas con la segunda componente principal.

Notamos que la observación que menos influye en la varianza explicada es la 9 pues está prácticamente sobre el origen de coordenadas. Las observaciones que más afectan a la primera componente principal son la 24 y la 29, ya que están más en los extremos del eje horizontal. Por otra parte, en la segunda componente principal, tienen más peso la observación 7 y 8.

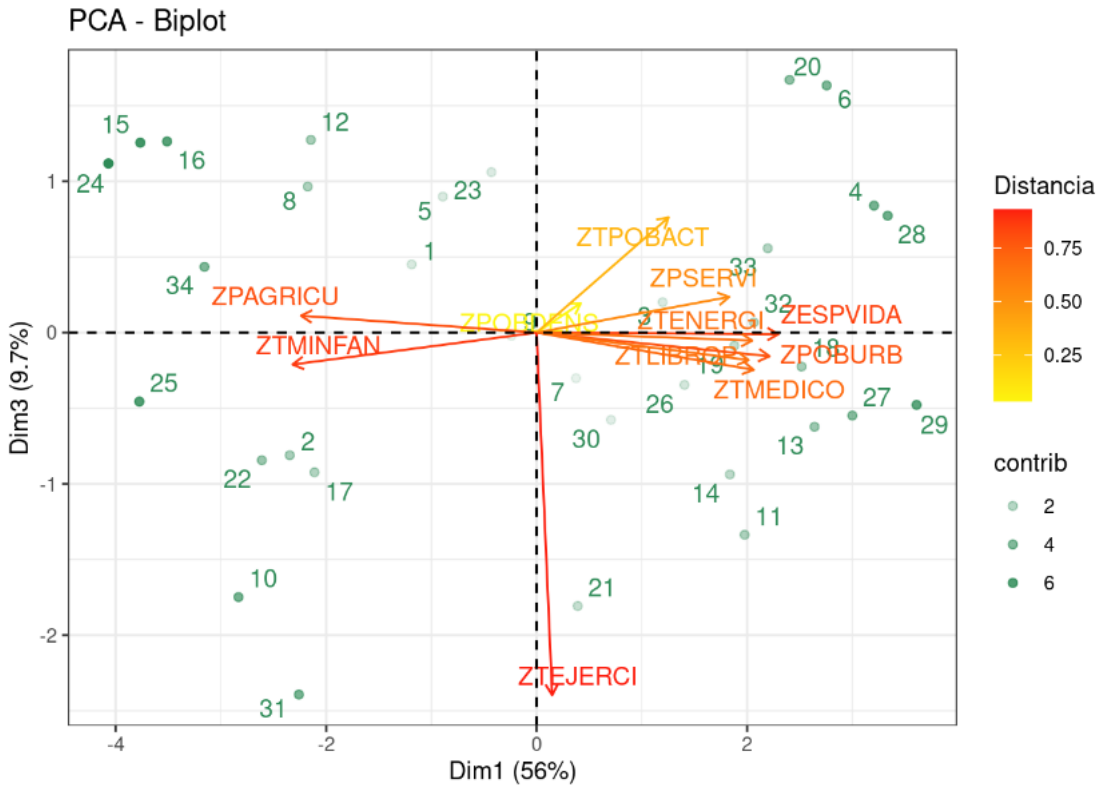

Figura 11: Variables y observaciones en la 1º y 3ª componente principal

En este caso la mayoría de las variables están más correladas con la primera componente principal, tanto positiva como negativamente. Unicamente vemos una variable, ZTEJERCI que está casi perfectamente correlada negativamente con la tercera componente principal. Apreciamos además una variable, ZTPOBACT, que presenta prácticamente la misma correlación positiva con ambas componentes principales. Además, notamos que hay una mayor correlación de las variables con la primera componente principal respecto a la tercera de la que había con la primera componente con respecto a la segunda en el primer gráfico, ya que las flechas están más pegadas al eje horizontal.

En este caso observamos que los datos 9 y 7 son los que menos afectan a las primera y tercera componente principal. Al igual que observamos en el gráfico anterior, los datos 29 y 24 son los que mayor influencia tienen en la primera componente, mientras que las observaciones 20 y 31 afectan en mayor medida a la tercera componente.

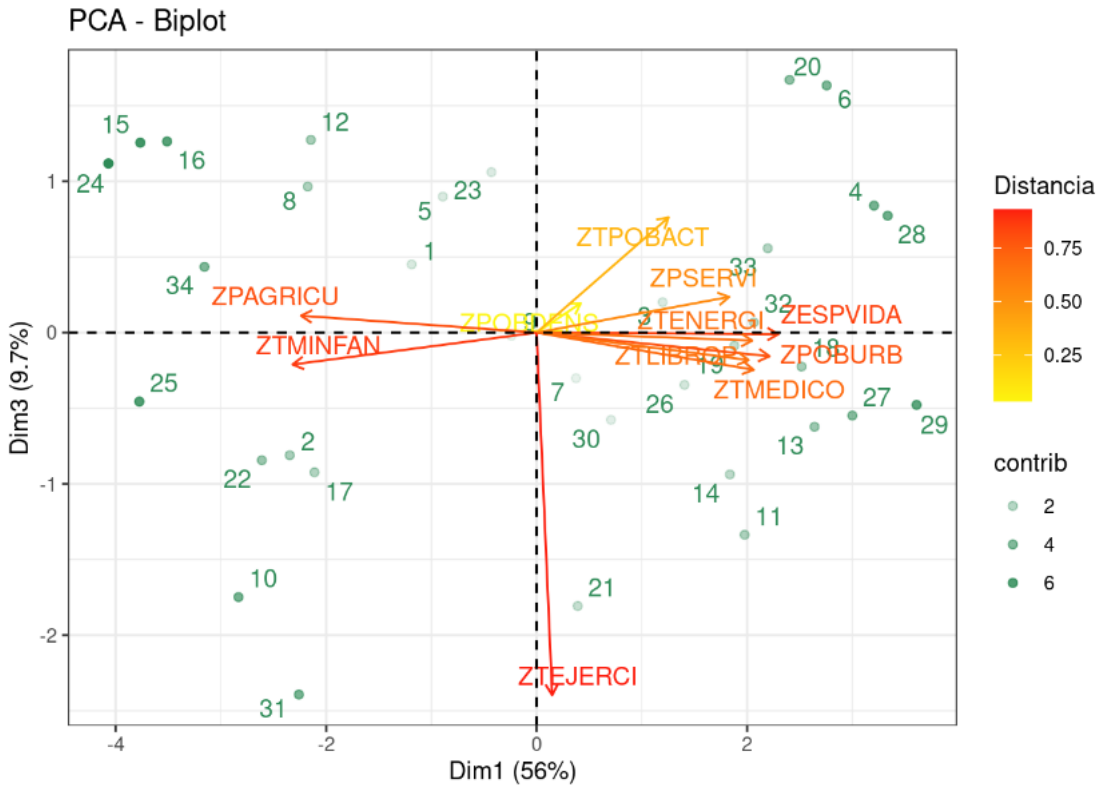

Figura 12: Variables y observaciones en la 2ª y 3ª componente principal

Como ocurría en el caso anterior, solo la variable ZTEJERCI está fuertemente correlada con la tercera componente principal, aunque aquí esta variable tiene algo más de peso en la segunda componente principal de la que tenía en la primera. Vemos que el resto de variables están más correladas con la segunda componente principal, aunque en general no en gran medida, pues las flechas son pequeñas y de color oscuro. La variable ZTPOBACT, al igual que en el gráfico anterior, influye casi por igual en ambas componentes principales, siendo en este gráfico la correlación positiva mayor, pues la flecha es de mayor longitud.

La observación 9 vuelve a ser la que menos afecta a las componentes principales, seguida de la 19 y la 13. La segunda componente se ve más influenciada, como ya sabíamos, por las observaciones 7 y 8. Con respecto a la tercera componente, el dato 31 es el más influyente, seguido por el 20, lo cual coincide con lo visto en el gráfico anterior.

4.3. Análisis Factorial (AF)

Para determinar el número óptimo de factores a considerar usaremos dos técnicas gráficas (scree plot y análisis paralelo) y un test de hipótesis (factanal). Mostramos en primer lugar los resultados de las técnicas gráficas:

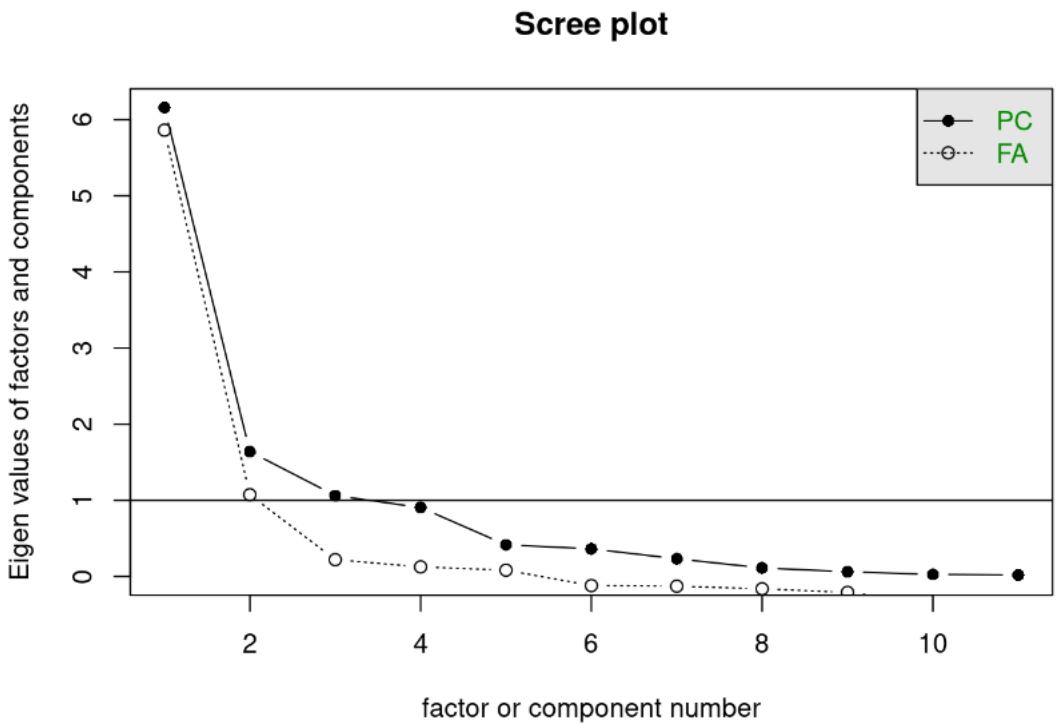

Figura 13: Scree plot

Por el método del codo deducimos que el número de factores a considerar debería estar entre 2 y 3.

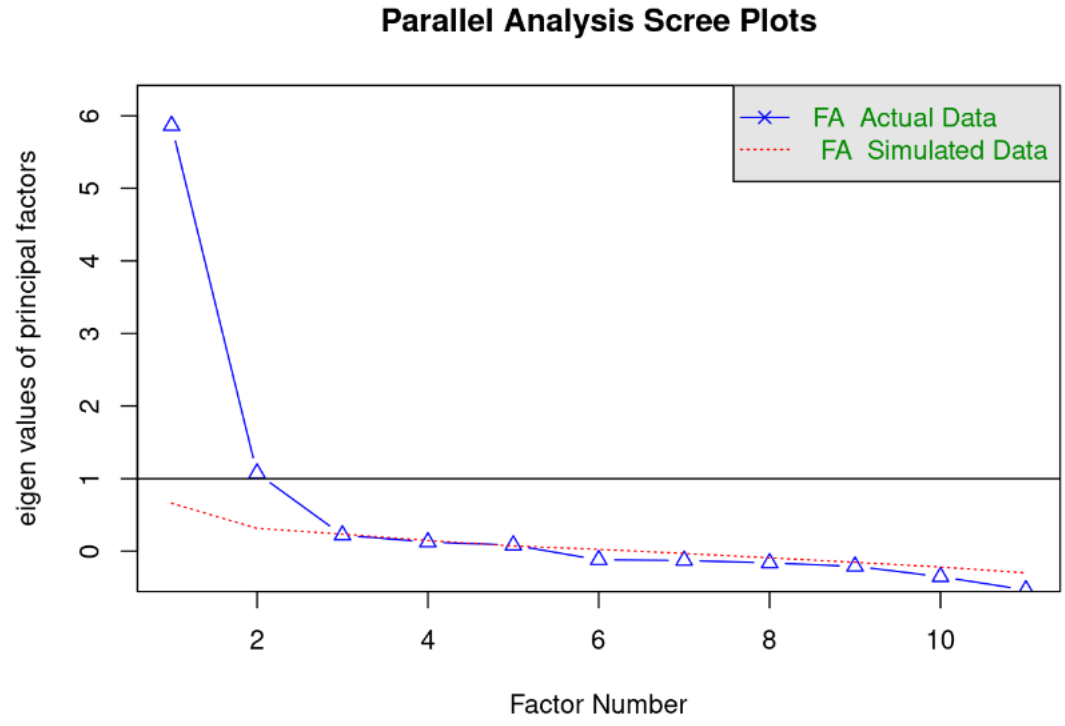

Figura 14: Análisis paralelo

El análisis paralelo indica que hay al menos dos factores, siendo posible considerar un tercer factor, pues la línea de puntos del análisis paralelo cruza la línea sólida del análisis factorial en el tercer factor, o incluso en el cuarto y quinto.

El test de hipótesis indica que 2 factores no son suficientes, pero 3 sí que lo son, de modo que tomaremos 3 factores para llevar a cabo el análisis factorial.

Así pues, estimamos el modelo factorial con 3 factores usando una rotación con el enfoque **varimax**.

La matriz de pesos factorial obtenida , así como la proporción de varianza explicada y acumulada para cada factor, ha sido la siguiente:

```
          MR1    MR2    MR3
ZPOBDENS  0.013  0.270  0.060
ZTMINFAN -0.725 -0.608 -0.009
ZESPVIDA  0.742  0.535  0.084
ZPOBURB   0.955  0.112  0.041
ZTMEDICO  0.607  0.531  0.179
ZPAGRICU -0.975 -0.110 -0.023
ZPSERVI   0.956 -0.159 -0.242
ZTLIBROP  0.628  0.436  0.145
ZTEJERCI  0.000  0.039  0.702
ZTPOBACT  0.055  0.962 -0.153
ZTENERGI  0.575  0.584  0.135

                 MR1   MR2   MR3
SS loadings    4.951 2.518 0.659
Proportion Var 0.450 0.229 0.060
Cumulative Var 0.450 0.679 0.739
```

Figura 15: Resultados del análisis factorial

La varianza acumulada explicada es de 0.739, es decir, el 74 % de la varianza original queda explicada por los tres factores determinados.

Ahora mostramos gráficamente con qué variables correlaciona cada uno de los factores obtenidos:

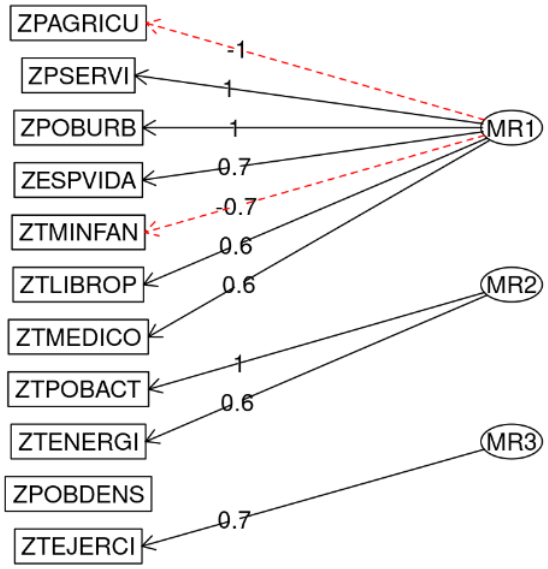

Figura 16

Podemos ver que el primer factor tiene una alta correlación con un gran número de variables, tanto de forma positiva como negativa. En cambio, el segundo factor sólo se correlaciona en gran medida con dos variables, ZTPOBACT y ZTENERGI, siendo la correlación con la variable ZTPOBACT perfecta y positiva. El tercer factor tiene una correlación alta únicamente con una variable, ZTEJERCI. Hay una variable, ZPOBDENS, que presenta correlación baja con los tres factores.

Notamos además que, el hecho de aplicar una rotación ortogonal, ha permitido que estemos cerca de una situación ideal, en la que tendríamos una alta correlación de cada uno de los factores con un grupo de variables observables concretas y prácticamente nula con el resto. Podemos ver por ejemplo que la variable observable ZTEJERCI tiene una alta correlación con el tercer factor y casi nula con los otros dos factores. Lo mismo ocurre con ZTPOBACT, que está casi perfectamente correlada con el segundo factor y presenta una correlación baja con el resto de factores. El primer factor tiene una alta correlación con un grupo mayor de variables observables, siendo por ejemplo la correlación casi perfecta con las variables ZPSERVI, ZPAGRICU y ZPOBURB, las cuales no presentan a penas correlación con los otros factores.

4.4. Análisis discriminante

Para realizar este análisis hemos tenido que etiquetar los datos. Se ha etiquetado de forma binaria a las instancias dividiendo a las que tienen el valor del atributo *ZespVida* por encima de la mediana de los que los tienen por debajo.

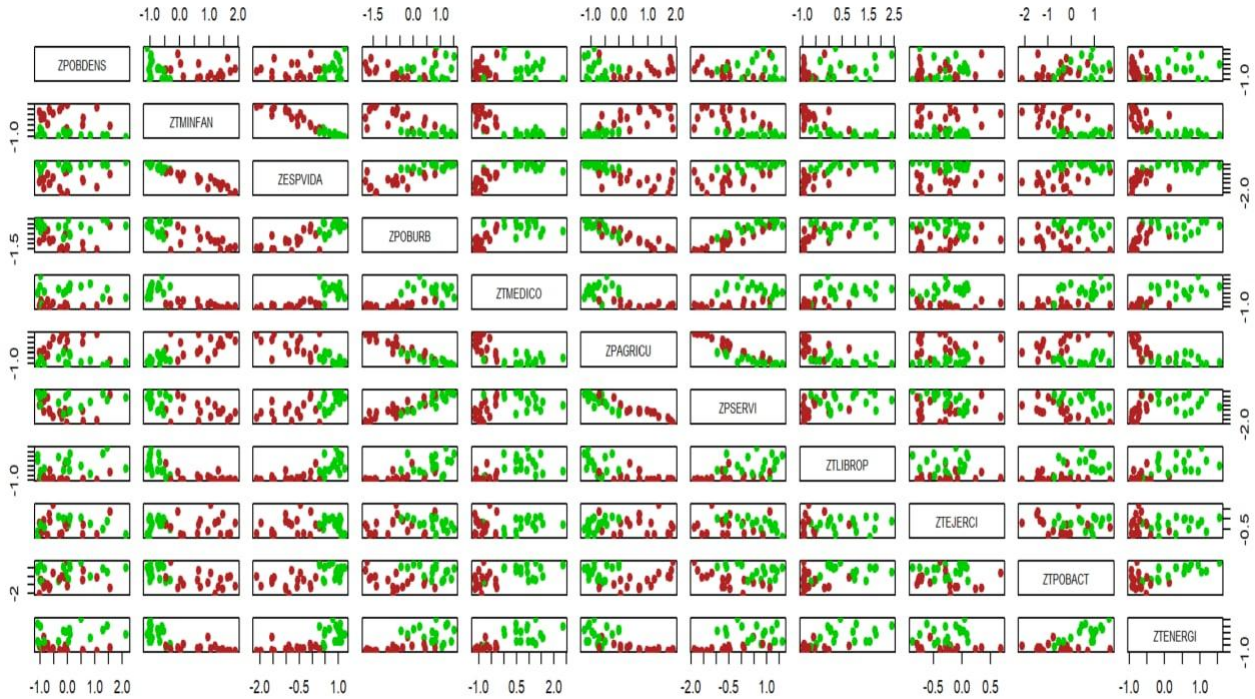

Figura 17: Separación de clases entre pares de variables

También, es necesario que haya normalidad multivariante en los datos. Para comprobarlo se usan los tests de *Henze-Zirkler* y de *Mardia*. Ambos tests muestran evidencias para suponer que hay normalidad multivariante al 95 % de confianza. Por otra parte es necesario contrastar la hipótesis de que la matriz de covarianzas es constante en todas las clases y para ello se usa el Test de *Box M*, dicho test es sensible a que los datos se distribuyan según una normal multivariante, que efectivamente se cumple. El resultados del test nos da un p-valor menor que 0.001 y rechazamos entonces la hipótesis nula, esto es, asumimos la no homogeneidad de varianzas, por lo que es recomendable usar un análisis discriminante cuadrático (QDA), aunque también se ha hecho el lineal (LDA).

En las tablas 1 y 2 presentamos los resultados del entrenamiento en forma de matriz de confusión. El *error de entrenamiento* para LDA es de 0 % mientras que para QDA es de aproximadamente un 2.9 %. Ambos modelos presentan una probabilidad a priori del 50 %.

Tabla 1: QDA: Resultados entrenamiento

QDA	**a**	**b**
a	17	0
b	1	16

Tabla 2: LDA: Resultados entrenamiento

LDA	**a**	**b**
a	17	0
b	0	17

Se creó un nuevo registro para clasificarlo con los siguientes valores: *ZPOBDENS=0.3, ZTMINFAN=-0.2, ZESPVIDA=0.1, ZPOBURB=-0.1, ZTMEDICO=0.5, ZPAGRICU=-0.5, ZPSERVI=0.5, ZTLIBROP=0.3, ZTEJERCI=-0.8, ZTPOBACT=0.2, ZTENERGI=0.1.* Los resultados se muestran en la tabla 3.

Tabla 3: Predicción

	a	**b**
LDA	0.003296763	0.9967032
QDA	0.917501	0.08249898

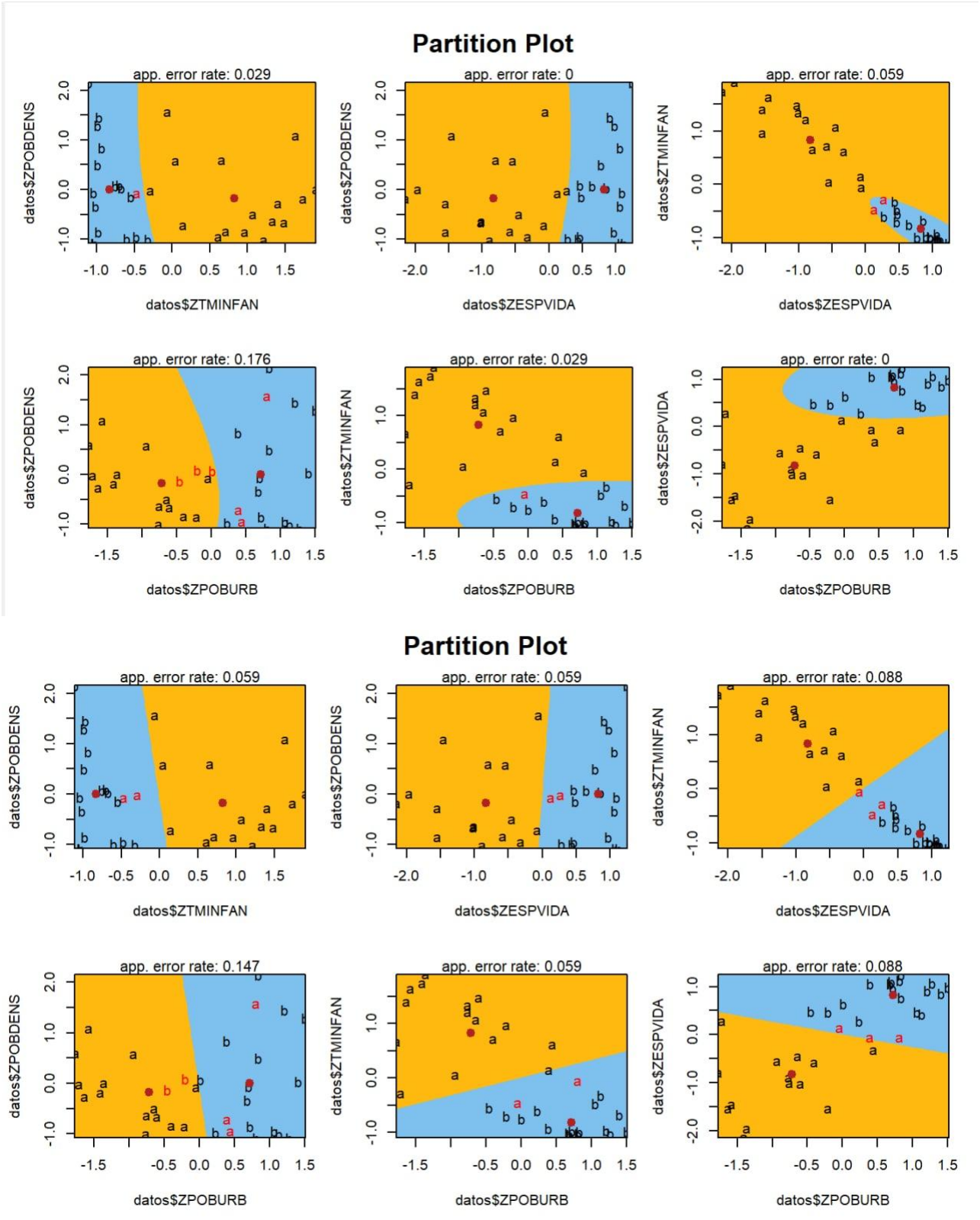

Figura 18: Representación de los límites del modelo para cada par de predicciones. Cada color representa una región de clasificación acorde al modelo, se muestra el centroide de cada región y el valor real de las observaciones. La imagen superior corresponde al QDA y la inferior al LDA

5. Discusión

El objetivo del trabajo como ya hemos mencionado, consiste en descubrir patrones, perfiles y tendencias a partir del análisis de los datos utilizando las técnicas estadísticas del análisis multivariante de datos aprendidas a lo largo del curso.

En efecto, en el **Análisis Univariante** podemos interpretar que todas las variables se distribuyen según una normal univariante, siendo una de ellas un poco conflictiva.

De los resultados obtenidos para el **análisis de componentes principales**, podemos decir, viendo las Figuras 8, 9 y 10 que la componente principal con la que están menos correladas las variables es la tercera, siendo únicamente ZTEJERCI la variable que más peso tiene en la tercera componente, aunque el resto también influyen, como por ejemplo, ZTPOBACT, que tiene un peso considerable en dicha componente. El resto de variables tienen, en general, más peso en la primera componente, aunque su influencia en la segunda también es considerable, estando algunas variables más correladas con la segunda que con la primera.

Además, podríamos considerar eliminar la observación 9 para este análisis, pues no parece afectar a las componentes principales seleccionadas. Por su parte, la observación 31 parece decisiva en la elección de la tercera componente principal. El resto de datos afectan por igual a la primera y segunda componente, con algunas excepciones de datos extremos que influyen en mayor medida en una u otra, como por ejemplo la observación 7, con gran peso en la segunda pero casi nulo en la primera.

Por su parte, en el **análisis factorial** el primer factor es el que está más correlacionado con la mayoría de las variables. Los otros dos factores están poco correlacionados con casi todas las variables, sobre todo el tercero, el cual vimos que sólo está correlacionado de manera considerable con una variable y su correlación con el resto de variables es prácticamente nula. Este hecho explica que, a la hora de seleccionar el número de factores óptimos estuviéramos en duda entre 2 y 3, pues el tercero parece aportar poco. De hecho, la proporción de varianza explicada del mismo es solamente del 6 %.

Si comparamos PCA con AF, en ambos casos hemos concluido que el número de componentes y factores óptimos a considerar era de 3. Sin embargo, las componentes principales explicaban un mayor porcentaje de la varianza total que los factores de AF, 80,57 % frente a 74 %. Además, con AF teníamos una variable, ZPOBDENS, que casi no estaba correlacionada con ninguno de los factores obtenidos. Así pues, podemos decir entonces que PCA obtiene en este caso mejores resultados, por lo que si tuviéramos que elegir entre reducir la dimensionalidad con PCA frente a AF, quizás sería más conveniente elegir PCA.

Por una parte, el **análisis discriminante** mediante QDA presenta un error de entrenamiento perfecto, del 0 %, esto es, el modelo se ajusta totalmente a los datos de entrenamiento. Mientras que el modelo de LDA se adapta también muy bien pero no es perfecto. Luego podemos incurrir en un posible sobreajuste, sobre todo en el QDA. Pese a que tenemos muy pocos ejemplos de entrenamiento, y los valores de las etiquetas han sido *"inventados"* no podemos esclarecer con total seguridad lo anterior. Aunque en la figura (Fig. 18) podemos ver por ejemplo que hay una partición en concreto en la que visualmente se produce sobreajuste (*Zpobdens-Zespvida*).

Tras ingresar un nuevo registro, obtenemos que la clasificación por parte de ambos modelos da lugar a desconcierto. Pues cada uno afirma con casi total seguridad que pertenece a una

etiqueta la cual es negada por el otro. Volviendo a reiterar lo anteriormente comentado, y al no tener constancia de qué etiqueta tiene realmente el registro inventado no podemos saber cual es la predicción errónea. Aunque, según el criterio usado para etiquetar, es más lógico la etiqueta que predice QDA, puesto que el valor de *Zespvida* está por debajo de la mediana.

6. Conclusión

En nuestro estudio hemos realizado un análisis univariante de los datos, sustituyendo tanto los outliers como los valores perdidos por la media. Este tratamiento que hemos hecho de los outliers puede que no sea el mejor, pues podrían eliminarse directamente, sustituirse por la mediana en lugar de la media, o trabajar con ellos sin modificarlos. Sería necesario analizar la naturaleza de los mismos para saber cómo lidiar con ellos de un modo más conveniente, pues puede ser que se deban simplemente a ruido en las mediciones, en cuyo caso sí que habría que eliminarlos, o que realmente tengan importancia en la población y serían tratados con la misma importancia que el resto de los datos.

Hemos conseguido reducir la dimensionalidad, tanto mediante variables observables (PCA) como mediante variables latentes (AF), pasando de las 11 variables originales a solamente 3, lo cual facilita la interpretabilidad. Con esta reducción perdemos parte de la varianza explicada, pues las 3 componentes principales explicarían el 80 % de la varianza, y, con las variables latentes, se explica el 74 %. Así, con PCA no se pierde demasiada varianza, pero con AF sí, de modo que dependiendo de la aplicación que se le quiera dar a los datos convendría realizar esta reducción o no. Por ejemplo, para clasificación con técnicas de Machine Learning (ML) la reducción de la dimensionalidad suele ser conveniente para dar mayor rapidez al entrenamiento y evitar el sobreajuste al reducir la varianza.

En cuanto al análisis discriminante comentar que no tiene mucho sentido pues las etiquetas han sido inventadas y no tiene ningún tipo de validez ni de caso práctico. En su lugar, se debería de haber hecho con una base de datos etiquetada. Sin embargo, con nuestros datos hemos conseguido llevar a cabo una clasificación casi perfecta de los datos de entrenamiento, con un clasificador tanto lineal como cuadrático.

Podría realizarse una clasificación con técnicas de Machine Learning más avanzadas para obtener mejores resultados y evitar sobreajuste, pero para esto necesitaríamos más datos, y etiquetas reales. Podría llevarse a cabo un aumento de los datos, modificando ligeramente los datos de los que disponemos con determinadas técnicas existentes en ML para aumentar el número de datos.

7. Anexo

Participación:

- Análisis univariante: Juanjo y Adrian
- ACP: Pilar
- AF: Pilar y Adrián
- AD: Juanjo
- Memoria: Adrián, Pilar y Juanjo.

Lo parte que ha realizado cada uno es algo relativo, pues a lo largo del trabajo nos hemos ido constantemente retroalimentando, proponiendo ideas, corrigiendo erratas y verificando que lo que hizo cada uno era correcto. De este modo, aunque Pilar haya hecho el ACP, Adrían y Juanjo comprobaron que estaba todo correcto, así como también Pilar y Adrián comprobaron que el análisis discriminante no tuviera errores.