Análisis de Datos Energéticos Globales

Modelos descriptivos y predictivos I : Proyecto

Grupo A1-12

Pablo Pertusa

Marc Romeu

Marc Siquier

Yasmín Serena

Índice

- Introducción
- Objetivos
- Limpieza y transformación
- Análisis PCA
- Análisis Clustering
- Análisis PLS
- Conclusiones del estudio
- Anexos

Introducción

La interacción entre la economía, la población y las fuentes de energía de un país desempeña un papel crucial en el desarrollo sostenible y la política energética global. El conjunto de datos estudiado proporciona una visión de cómo distintos países aprovechan y gestionan la energía teniendo en cuenta sus características económicas y demográficas. Los datos analizados contienen distintas variables que detallan la producción y el uso de energía, así como aspectos económicos como el PIB, y la población para diversos países a lo largo de varios años.

Contexto del estudio:

La base de datos ha sido obtenida de github. Esta tiene en origen 21591 observaciones de distintos países y regiones entre años variables 1900-2022, contiene 128 variables numéricas, y 2 categóricas que identifican el país.

Después de realizar el tratamiento y la limpieza de datos se ha llegado a un dataset con 24 variables numéricas y 1 categórica con 2028 observaciones en las que se encuentran los datos de los países más importantes económicamente con las observaciones desde el año 2000 hasta 2021, algunos datos se han imputado y se ha realizado una transformación logarítmica junto la division entre el numero de habitantes por país para cada variable numérica que no fuera año. El proceso se muestra en el Anexo A.

Estas son las variables que contiene el dataset con la limpieza realizada:

-**country**: País o región al que pertenecen los datos. (Variable Categórica)

-**year**: Año al que pertenecen los datos. (Variable Numérica)

-**population**: Población total del país o región. (Variable Numérica)

-**gdp**: Producto Interno Bruto total del país o región. (Variable Numérica)

-**carbon_intensity_elec**: Intensidad de carbono de la electricidad generada. (Variable Numérica)

-**coal_electricity**: Electricidad total generada por carbón. (Variable Numérica)

-**coal_production**: Producción total de carbón. (Variable Numérica)

-**electricity_demand**: Demanda total de electricidad. (Variable Numérica)

-**electricity_generation**: Generación total de electricidad. (Variable Numérica)

-**energy_per_gdp**: Consumo de energía por unidad de PIB. (Variable Numérica)

-**fossil_electricity**: Electricidad total generada por combustibles fósiles. (Variable Numérica)

-**gas_electricity**: Electricidad total generada por gas. (Variable Numérica)

-**gas_production**: Producción total de gas. (Variable Numérica)

-**greenhouse_gas_emissions**: Emisiones totales de gases de efecto invernadero. (Variable Numérica)

-**hydro_electricity**: Electricidad total generada por hidroeléctricas. (Variable Numérica)

-**net_elec_imports**: Total de importaciones netas de electricidad. (Variable Numérica)

-**nuclear_consumption**: Consumo total de energía nuclear. (Variable Numérica)

-**nuclear_electricity**: Electricidad total generada por energía nuclear. (Variable Numérica)

-**oil_electricity**: Electricidad total generada por petróleo. (Variable Numérica)

-**oil_production**: Producción total de petróleo. (Variable Numérica)

-**other_renewable_electricity**: Electricidad total generada por otras energías renovables (excluyendo biocombustibles). (Variable Numérica)

-**primary_energy_consumption**: Consumo total de energía primaria. (Variable Numérica)

-**renewables_electricity**: Electricidad total generada por energías renovables. (Variable Numérica)

-**solar_electricity**: Electricidad total generada por energía solar. (Variable Numérica)

-**wind_electricity**: Electricidad total generada por energía eólica. (Variable Numérica)

Objetivos

El análisis de este conjunto de datos busca entender mejor cómo las variables económicas, demográficas y de producción energética interactúan y afectan el desarrollo y la sostenibilidad energética de los países. Por ello se ha realizado un PCA, un clustering y un modelo PLS.

Limpieza y transformación

Tal y como se muestra en el anexo A comenzamos con muchas observaciones con poca información, por lo que nos centramos en seleccionar las observaciones a partir del 2000 y de los países más importantes. Tras esto, seleccionamos las columnas con menos del 20% de valores faltantes. Ahora seleccionamos las variables que más nos interesan, ya que muchas están relacionadas con otras, por ejemplo hay variables que representan la variación de otra respecto al año anterior. Cambiamos las unidades de medida variables “population” y “gdp” a millones de habitantes y miles de millones de doláres respectivamente. Una vez realizado esto y con muchos menos datos faltantes, utilizamos la libreria “mice” para imputar.

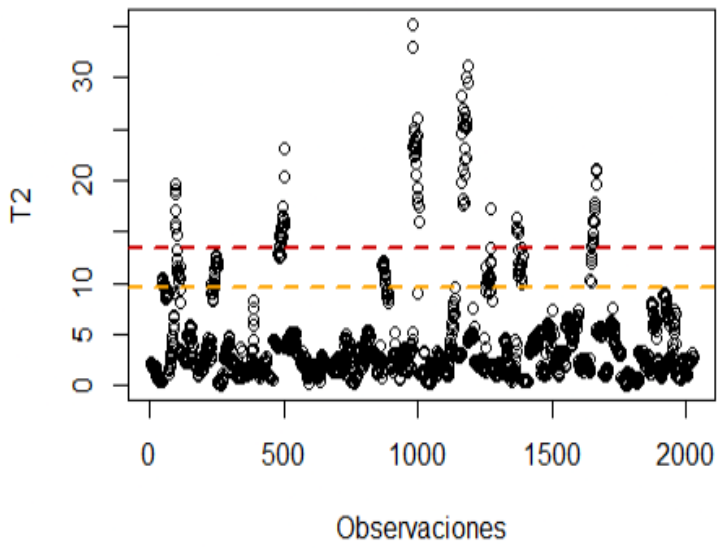
Una vez con los datos listos para comenzar, durante los primeros análisis nos dimos cuenta de la cantidad de datos anómalos que detectábamos, en su mayoría provenientes de variables con algunos valores muy superiores al resto, es por esto que decidimos aplicar una primera transformación para expresar las variables en millones de habitantes. Así, el consumo de los países es comparable. Aún así, fue necesaria una transformación para aplanar los datos, en este caso logarítmica. Para ello se excluyó la variable de “net_elec_imports” que cuantifica las importaciones de electricidad y puede ser negativa. La transformación es $\log(1 + x)$. Esto es debido a que estos datos no proceden de un entorno controlado y la disparidad entre países es

algo intrínseco. Por tanto, buscando suavizar estas diferencias con los valores más altos, se aplica la transformación por logaritmo. En el anexo A se estudia la distribución de las variables tras este proceso. Continuamos ahora con los datos listos para el análisis.

Análisis PCA

Comenzamos el Análisis en Componentes Principales con el objetivo de ver cómo se relacionan las variables entre sí. Entre las preguntamos que nos formulamos están: **¿Qué países son los que más consumen?**, **¿Cuál es la forma más común de producir energía?**, **¿Existe relación entre el patrón de consumo y las emisiones?**, **¿Hay países con políticas verdes?**, **¿Cuáles son?**

Realizamos nuestro modelo PCA con 4 componentes y usando como variables suplementarias "country", "year" y "population". Para verificar la validez de nuestro modelo, estudiamos el gráfico de la T2 de Hotelling para buscar observaciones anómalas.



Vemos claramente que hay más valores anómalos de los esperados, sin embargo, proseguiremos con el análisis para entender el por qué. Además podemos apreciar que las observaciones anómalas pertenecen a una serie de países al estar juntas entre sí, ya que las observaciones en los datos están ordenadas por país y año. Por lo tanto, estos países presentarán unas características que el modelo no ajusta completamente. Hay un total de 192 valores por encima del intervalo del 95% y 95 por encima del 99%.

Vamos a ver de qué países se trata:

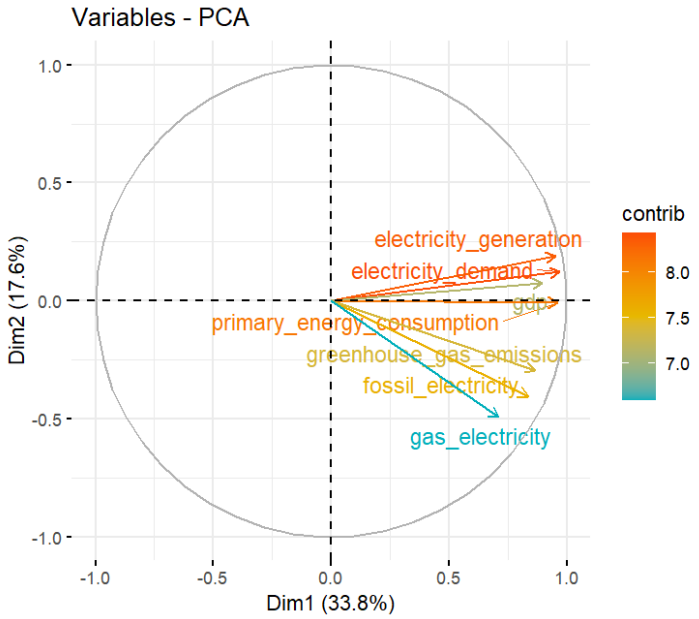
```
##   [1] "Australia"    "Bahrain"       "Canada"         "Finland"
"Kuwait"
```

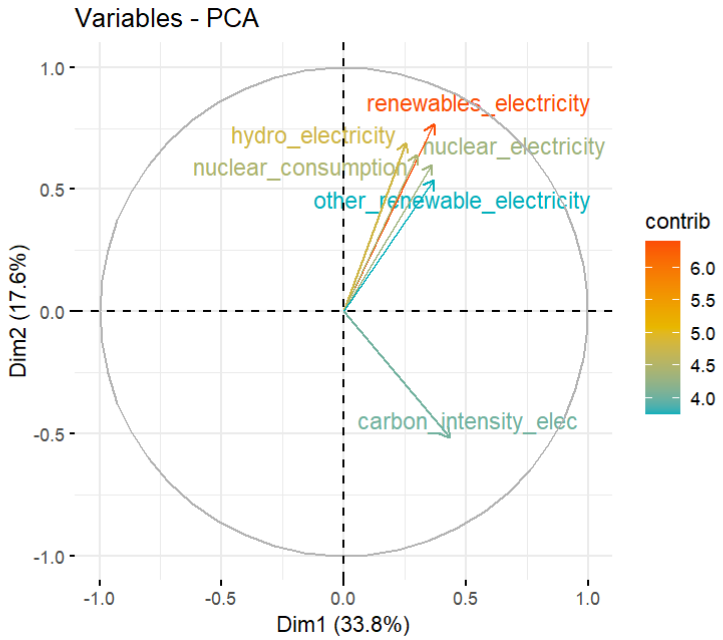
```
##   [6] "Luxembourg"   "New Zealand"  "Norway"         "Paraguay"
"Qatar"
## [11] "Sweden"
```

Vemos que hay países bastante diversos. Intentaremos ver el por qué de estas proyecciones lejos del centro del modelo.

Loading plots

Nos centramos ahora en los gráficos de variables para visualizar la importancia de cada una de las variables en las componentes. No mostraremos aquí todas las variables a la vez porque se satura el gráfico. Tampoco mostraremos todos los loadings plots, solo los más relevantes. El resto estará en el anexo B. Iremos mostrando las variables más importantes en cada componente. Para las variables que más contribuyen a la primera componente.

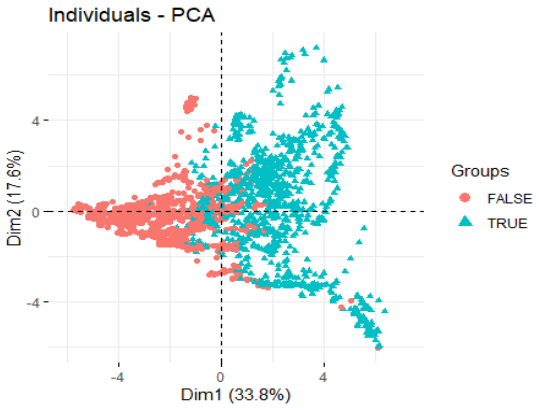


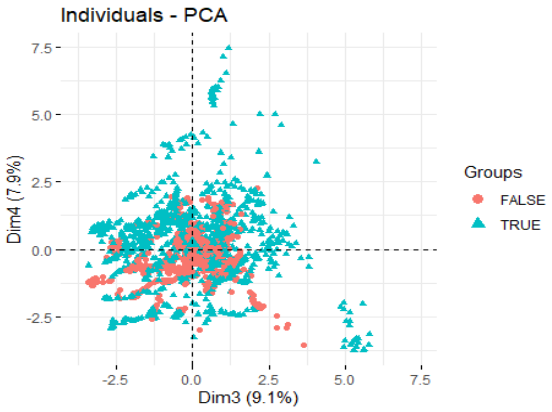


De estos dos gráficos deducimos que la primera componente representa el consumo total del país, que está muy relacionado con el desarrollo del pais ("gdp") y las emisiones que este emite. Podemos intuir que los países más desarrollados serán también aquellos que más consuman. La segunda componente representa el consumo de energía verde, ya que contribuyen a esta componente "renewables_energy", "hydro_electricity" y la energía nuclear, que comienza a considerarse verde también.

Los loading plot de las componentes 3 y 4 se encuentran en el anexo B. Resaltamos que la 4 componente diferencia los países importadores y exportadores de electricidad, siendo estos últimos los que en promedio más producción de gas, crudo y carbón tienen. De aquí podemos concluir que aquellos países con reservas ya sean vegetales o fósiles, tenderán a abastecerse ellos mismos de energía.

Gráficos Scores

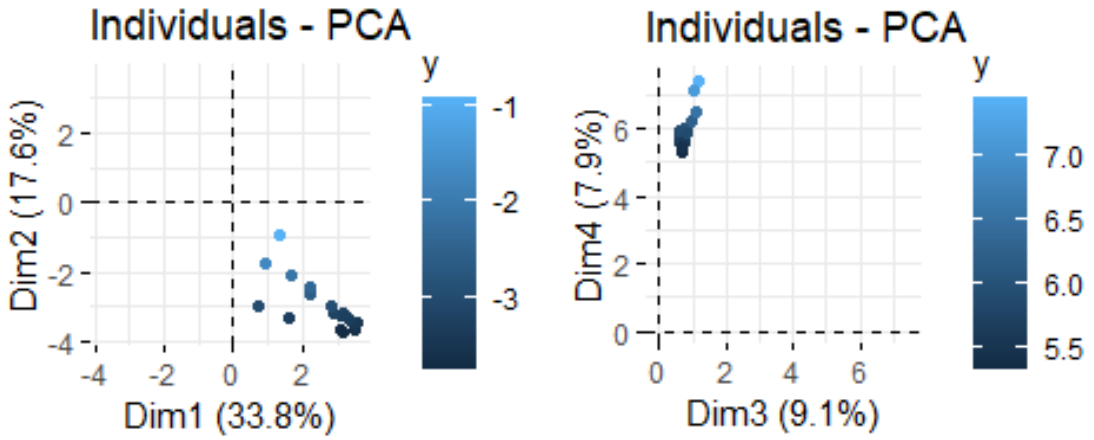




Observamos claramente lo que habíamos supuesto en los loading plots. A la derecha de la primera componente caen los países con un PIB per cápita superior a la media. Estos países más desarrollados, muestran una mayor variabilidad respecto a la segunda componente. Esto significa que entre los desarrollados, hay países que han optado por adoptar energías verdes y otros que las rechazan. Mientras que en el grupo de la izquierda, es más homogéneo respecto al eje Y. Esto nos dice que, los países menos desarrollados no suelen adoptar energías renovables. Esto puede ser a su costosa inversión inicial o porque prefieren fuentes de energía más estables.

Respecto al gráfico de la tercera y cuarta componente, podemos observar que los países que en promedio tendrán mayores importaciones de electricidad son aquellos más desarrollados mientras que los menos ricos tenderán a producir la energía ellos mismos ya que el consumo total que tienen es menor.

En el anexo B podemos ver que el escalado por población hace que se formen dos grupos de países muy poblados en la primera componente. Los que están en vías de desarrollo y los más importantes debido a su población y gran desarrollo. A pesar de su separación en las dos primeras componentes, entre la tercera y la cuarta no se diferencian tanto, por lo que tenderán a tener pocas importaciones y una alta producción de carbón en promedio. Es decir, los países poblados se suelen comportar de una manera parecida, usando los desarrollados en promedio más energías verdes por sus scores más alto en la segunda componente.



Resulta curioso ver, a la vista de este gráfico donde se han seleccionado los individuos con mayor score en la cuarta componente (relacionada con las importaciones), que los países con mayores importaciones de electricidad en promedio son aquellos con menor poco consumo de energías verdes y un alto consumo de combustibles fósiles.

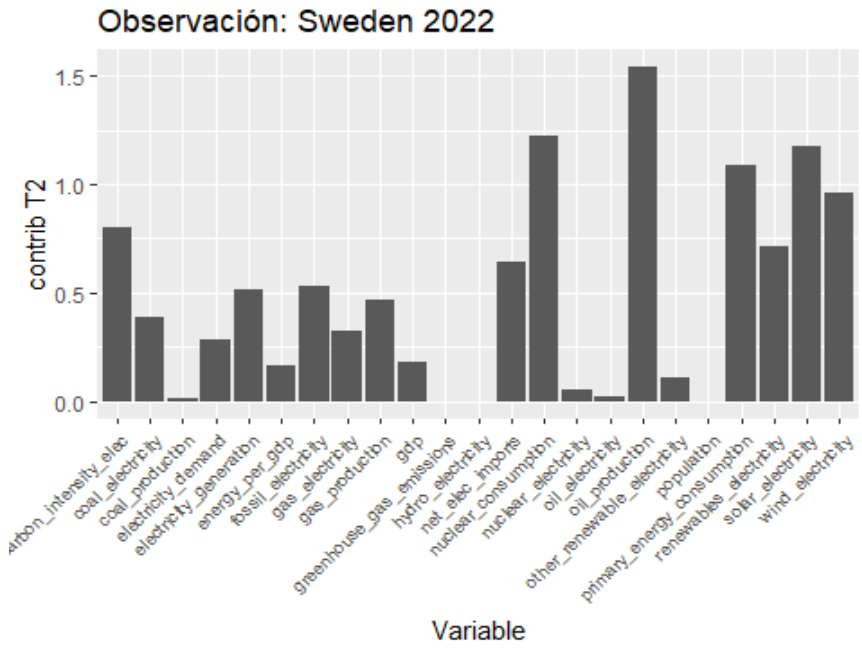
Los países con mayores emisiones y consumo de energías fósiles por millón de habitantes son:

```
## [1] "Bahrain" "Kuwait"  "Qatar"
```

Los países con mayor uso de energías verdes:

```
## [1] "Finland"  "Norway"   "Paraguay" "Sweden"
```

Para poder entender la naturaleza de algunos de las observaciones anómalas vemos el siguiente gráfico.

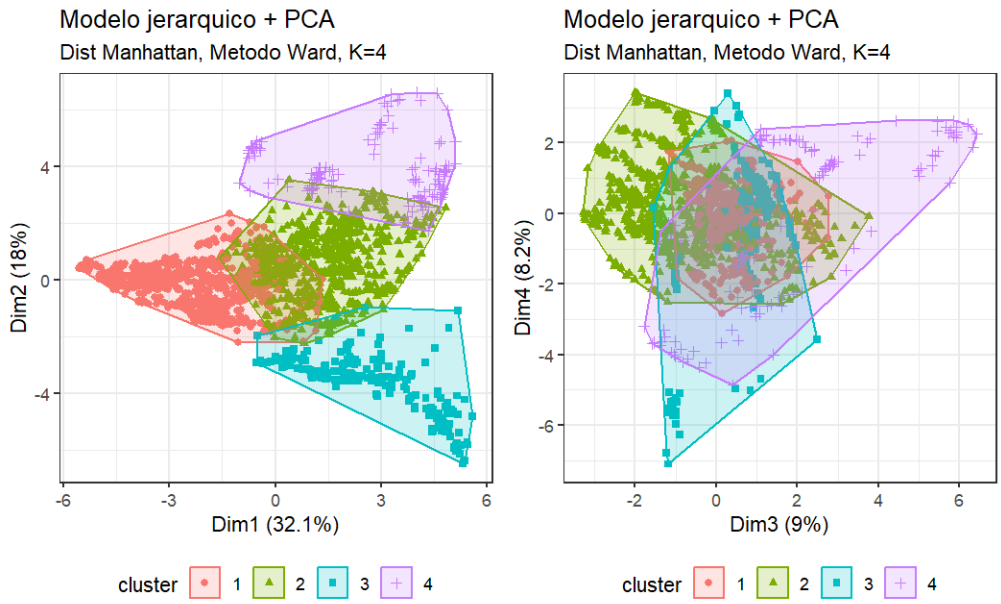


Aquí vemos el gráfico de contribución a la T2 de la observación con más T2. Se trata de Suecia en el año 2022 que registró valores altos en consumo de energías verdes. Esta observación ha aportado información al estudio ya que demuestra que sí hay países que están optando por un mix energético más limpio.

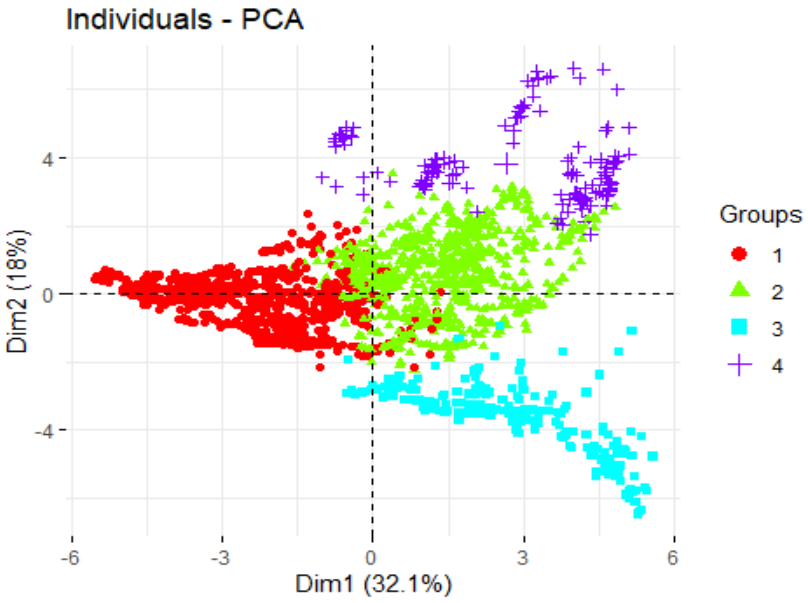
Análisis clustering

Vamos a ver como se agrupan los países en función de las energías consumidas y su relación con las emisiones de efecto invernadero, por lo que esa variable la excluiremos.

Tal y como se explica en el anexo C, utilizaremos la distancia de Manhattan por ser más robusta a valores extremos y que tiene más tendencia de agrupamiento en los datos en el índice de Hopkins. Tras explorar los modelos disponibles, usaremos un modelo jerárquico con el método de Ward con 4 clústers. Los resultados que apoyan nuestra elección están en el anexo C. Aquí podemos ver la distribución en una proyección PCA.

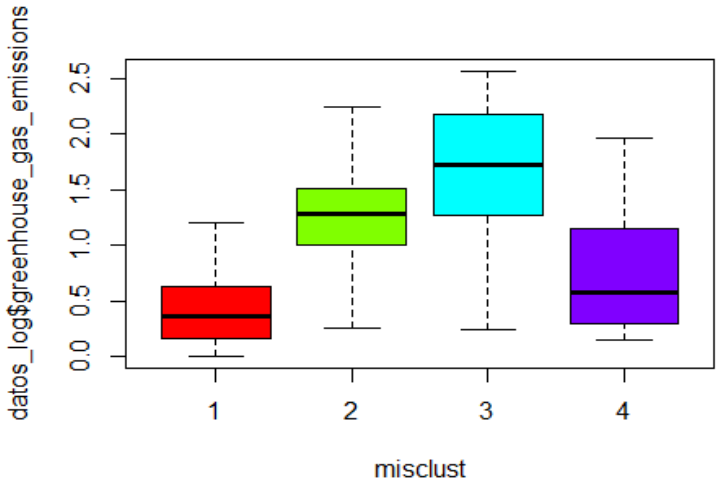


Vemos que la primera componente es la que diferencia los clusters 1 de los demás, mientras que la segunda componente diferencia entre 2, 3 y 4. Podemos interpretar, usando la información obtenida también en PCA, que el cluster 1 engloba a los países menos desarrollados con menos consumo. Los otros clusters representan los países más desarrollados divididos en 3 clusters según su consumo de energías verdes (segunda componente).



Para facilitar la interpretación de este score plot se puede consultar el gráfico de loadings del anexo C o los resultados del PCA. El cluster 3 estará representado por los países con gran consumo de energías fósiles y emisiones por cada millón de habitantes mientras que el 4 estará representado por los países con políticas energéticas más limpias. Cabe destacar que hay un grupo de observaciones del cluster

4 a la izquierda de la primera componente, por lo que podemos afirmar que algunos países menos desarrollados también están adoptando formas de energía limpias.



Este boxplot confirma el resultado esperable de lo que habíamos observado. El cluster que más emisiones tiene es el 3. Sin embargo, resulta que el cluster que menos contamina no es el 4 que es el que representa los países desarrollados con políticas energéticas verdes, sino el 1. Esto nos lleva a pensar que a pesar de invertir en energías limpias, al ser países con mucho consumo total, también necesitan de energías más contaminantes para abastecerse. Los intervalos de Tukey están en el anexo C.

Análisis PLS

Comenzamos el análisis PLS. Tendremos una única variable respuesta, "greenhouse_gas_emissions" que trataremos de predecir a partir del resto de variables de consumo energético. Como tenemos muchas observaciones vamos a dividir en datos Test y datos Train. Para ajustar el número de componentes en el modelo PLS en los datos de entrenamiento usamos la validación cruzada representada en el gráfico que combina R2 y Q2 del anexo D. Escogeremos 4 componentes ya que a pesar que R2 y Q2 aumenten, no lo hacen significativamente.

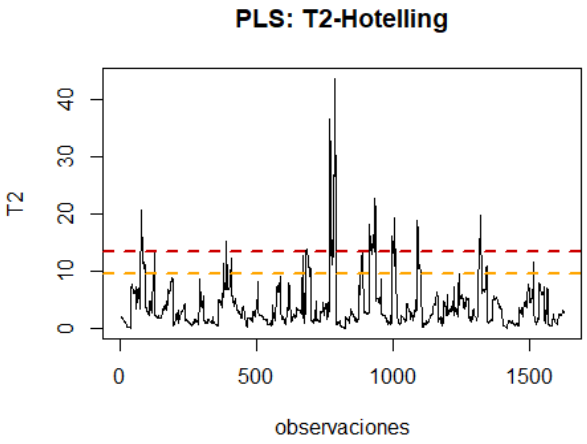
```
##                                                                    PLS
##     1623     samples     x     21     variables     and     1     response
##      standard      scaling      of      predictors      and      response(s)
##            R2X(cum)  R2Y(cum)  Q2(cum)    RMSEE  pre  ort     pR2Y      pQ2
## Total    0.625    0.986    0.985 0.0736   4    0 0.0333 0.0333
```

La cantidad de valores anómalos es:

```
## [1] 71
```

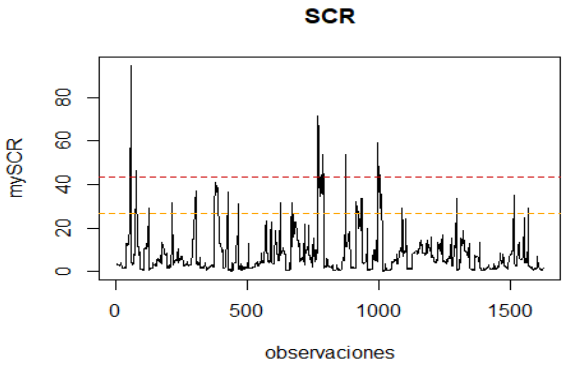
Hay 4 veces más anómalos de los esperados pero tratándose de observaciones de países muy dispares, vamos a proseguir con el análisis para tratar de averiguar el por qué.

```
##       Año       Pais       T2
## 1000 2020 Luxembourg 43.68424
## 980  2000 Luxembourg 36.71925
## 999  2019 Luxembourg 34.61145
## 997  2017 Luxembourg 34.38614
## 1001 2021 Luxembourg 33.68588
## 981  2001 Luxembourg 29.27288
```



Vemos que la mayoría de observaciones anómalas provienen de países pequeños, como Luxemburgo. En PCA vimos algo parecido, por lo que mantendremos estas observaciones para poder entender su comportamiento. Continuamos con los residuos.

```
##       Año       Pais      SCR
## 68   2022  Australia 94.53784
## 67   2021  Australia 73.67163
## 980  2000 Luxembourg 71.34250
## 981  2001 Luxembourg 62.84058
## 1253 2000   Paraguay 59.37785
## 998  2018 Luxembourg 53.74881
```

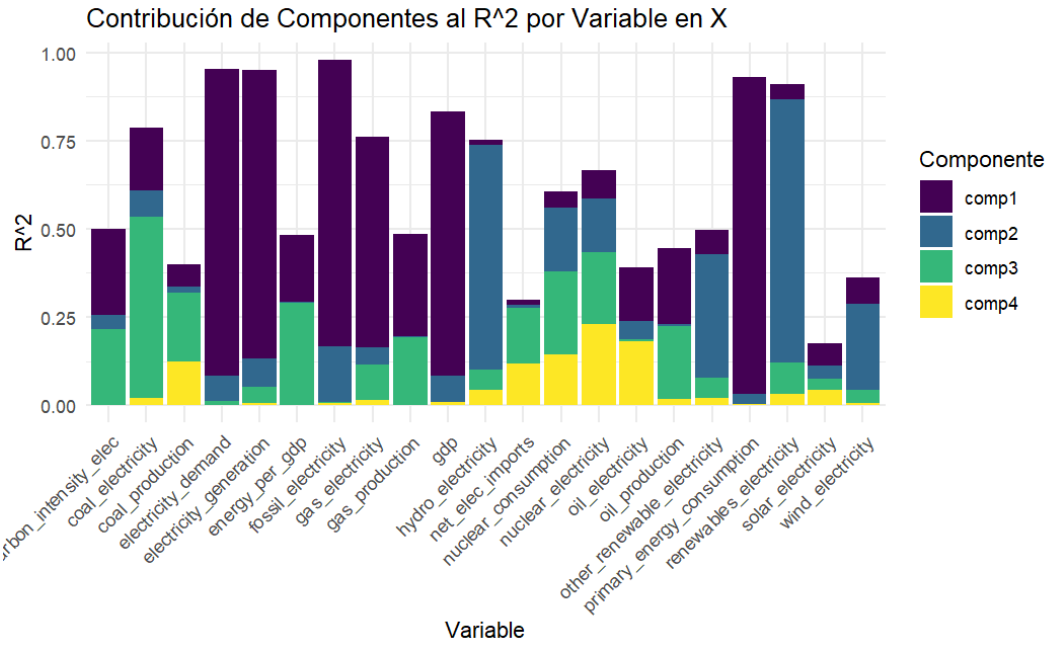


También nos encontramos con observaciones con muchos residuos. Las dos que más tienen pertenecen a años cercanos de Australia. Mantendremos las observaciones en el estudio.

La linealidad de los scores se puede ver en el anexo D.

```
## PLS
## 1623 samples x 21 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total    0.311    0.834   0.833 0.251   1   0 0.05 0.05
## PLS
## 1623 samples x 21 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total    0.448    0.953   0.952 0.133   2   0 0.05 0.05
## PLS
## 1623 samples x 21 variables and 1 response
## standard scaling of predictors and response(s)
```

```
##       R2X(cum) R2Y(cum) Q2(cum)  RMSEE pre ort pR2Y  pQ2
## Total    0.575    0.976   0.976 0.0948   3   0 0.05 0.05
## PLS
## 1623 samples x 21 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum)  RMSEE pre ort pR2Y  pQ2
## Total    0.625    0.986   0.985 0.0736   4   0 0.05 0.05
```



El ajuste por componente del espacio Y está en el anexo D.

Vemos que la primera componente es la que más explica en las variables más importantes en el consumo de un país, mientras que la segunda ajusta mucho mejor las variables relacionadas con energías verdes. Situación muy parecida a la que encontramos en el análisis PCA. El segundo gráfico no aporta mucha información porque solo tenemos una variable de respuesta. Con esto hemos visto el comportamiento de la R2, vamos a ver Q2. Ya conocemos la Q2 en el conjunto de entrenamiento vamos a ver en el de test.

Vamos a calcular la Q2 en los datos Test.

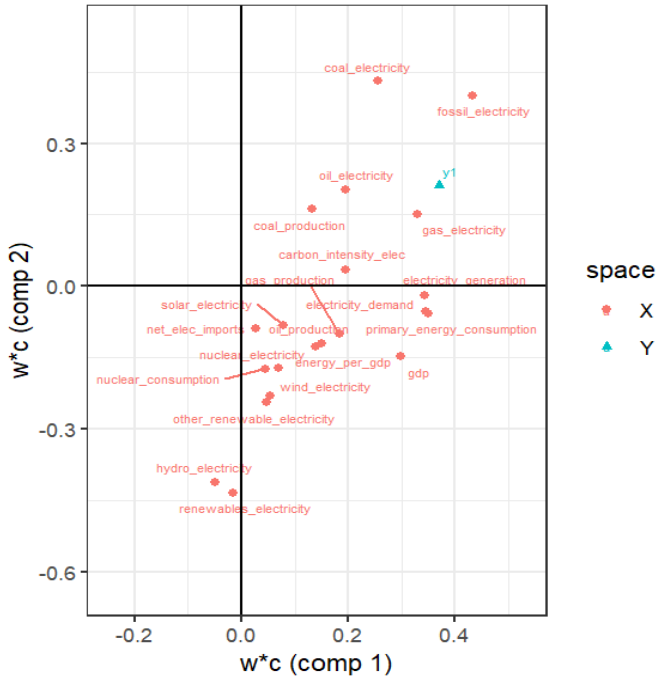
```
##          pred      Real
## 4  -0.8937432 -0.8776744
## 6  -0.8486029 -0.8269303
## 12 -0.6969396 -0.6444226
## 14 -0.6362534 -0.5653262
## 24 -0.7868755 -0.6819657
## 26 -0.7653609 -0.6503655
```

Podemos ver que tiene una alta capacidad predictora en el conjunto Test. Cabe destacar que las emisiones no pueden ser negativas, esto ocurre porque se ha centrado y escalado.

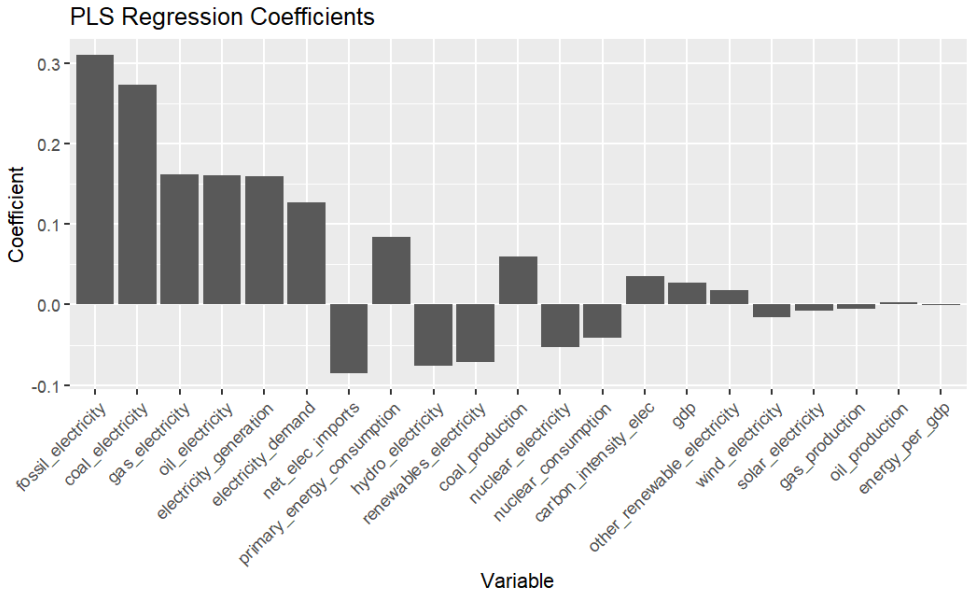
Lo corrobora la Q2.

```
##           [,1]
## [1,] 0.9957112
```

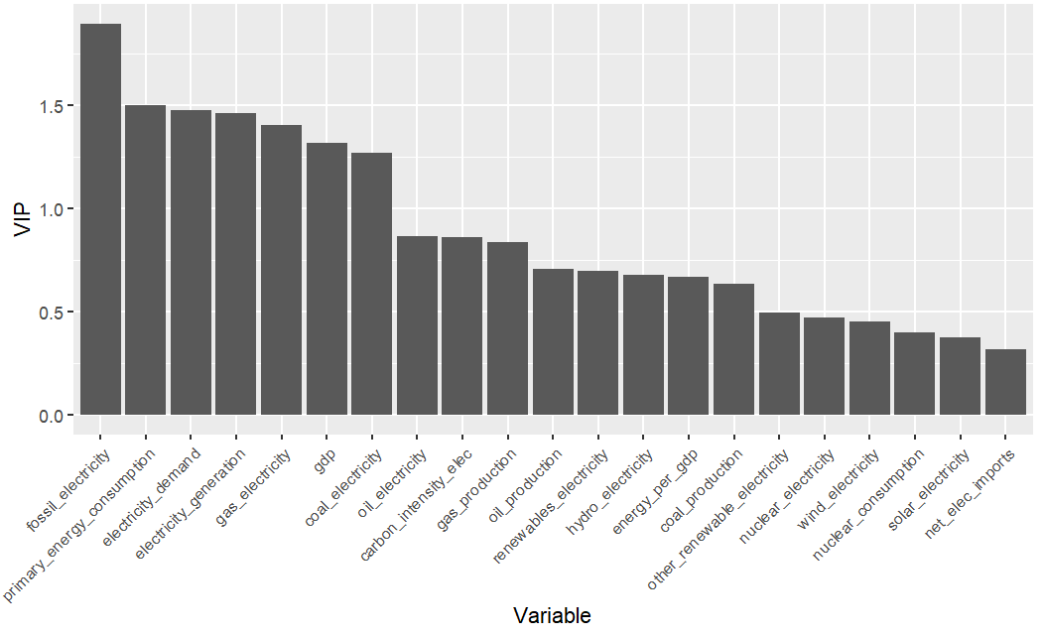
Vamos a ver la relación entre las variables predictoras y la variable de respuesta.



Las variables que más influyen en las emisiones de efecto invernadero son “fossil_electricity” y “coal_electricity” mientras que las que más las disminuyen son “renewables_electricity” e “hydro_electricity”. De las energías verdes, la que más ayuda a reducir las emisiones según el modelo es la hidráulica.



Fijándonos en los coeficientes de regresión PLS observamos que la variable con mayor coeficiente es fossil_electricity, por lo que podemos confirmar que se trata de una variable importante para las emisiones de un país, seguida de la electricidad producida a partir de carbón. Este gráfico lo podemos usar para teorizar que la energía nuclear debería considerarse por los países que buscan reducir sus emisiones, ya que con un coeficiente negativo tenderá a reducirlas.



En este gráfico se aprecian las variables más importantes para el modelo son las que las que caracterizan el consumo total de un país y la electricidad fósil.

Conclusiones

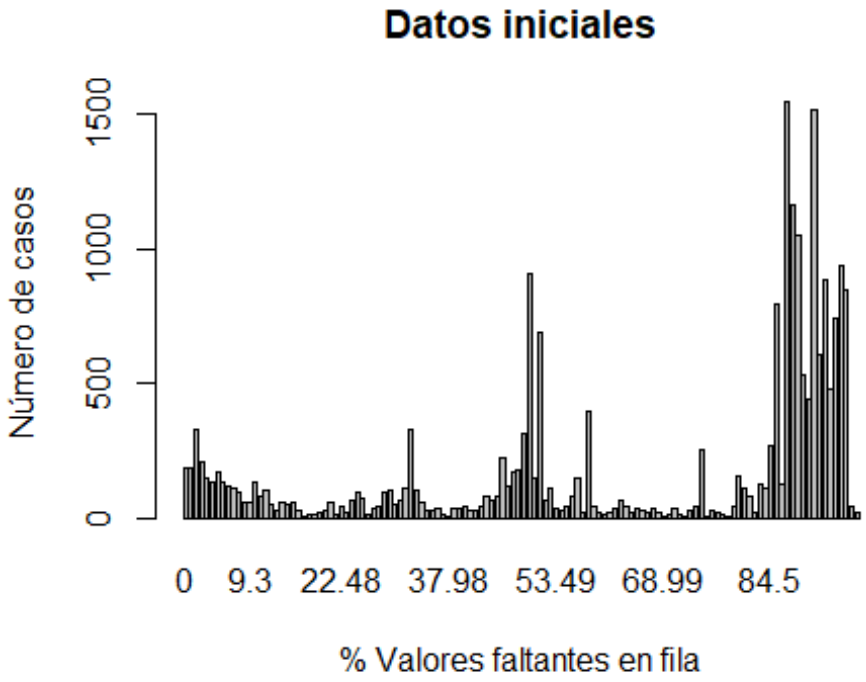
Tras los análisis realizados hemos observado que en los últimos 24 años las formas de energía más utilizadas son las no renovables. Sin embargo, hay países que están optando por un mix energético más limpio, pero hemos encontrado que en general suelen ser los países desarrollados los que eligen esta opción, y no tanto los que se encuentran en vías de desarrollo. También hemos visto que las energías renovables y la nuclear contribuyen a reducir las emisiones de efecto invernadero, por lo que si esto es lo que se busca de cara al futuro, habrá que buscar formas de integrarlas.

Anexos

1. Anexo A: Tratamiento y limpieza de datos
2. Anexo B: PCA
3. Anexo C: Clustering
4. Anexo D: Pls

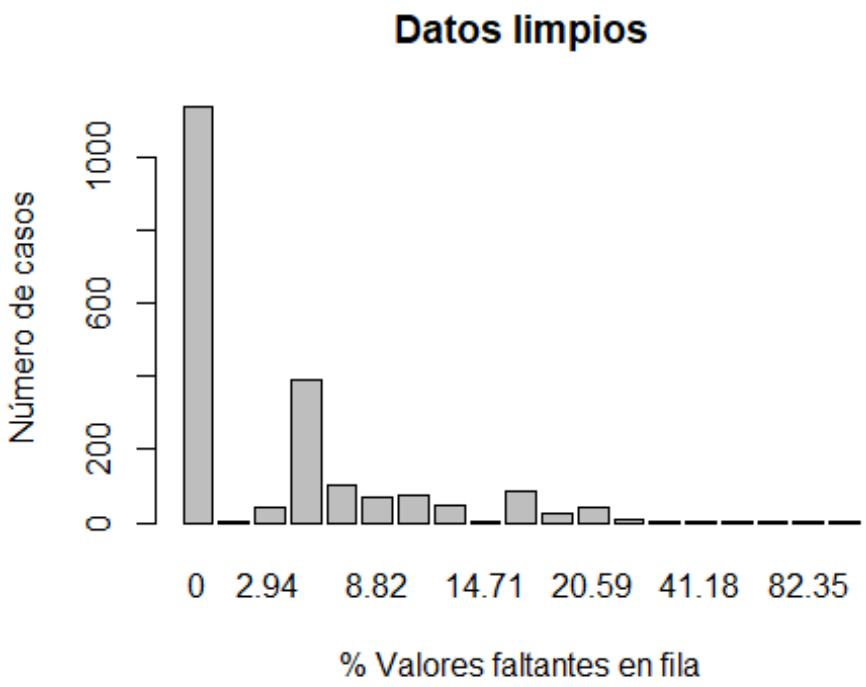
Anexo A

Este es el porcentaje de datos faltantes por fila al comenzar. Debemos eliminar las observaciones con muchos faltantes ya que no aportan información.



Se nos queda un dataframe con **2039** observaciones y **68** variables.

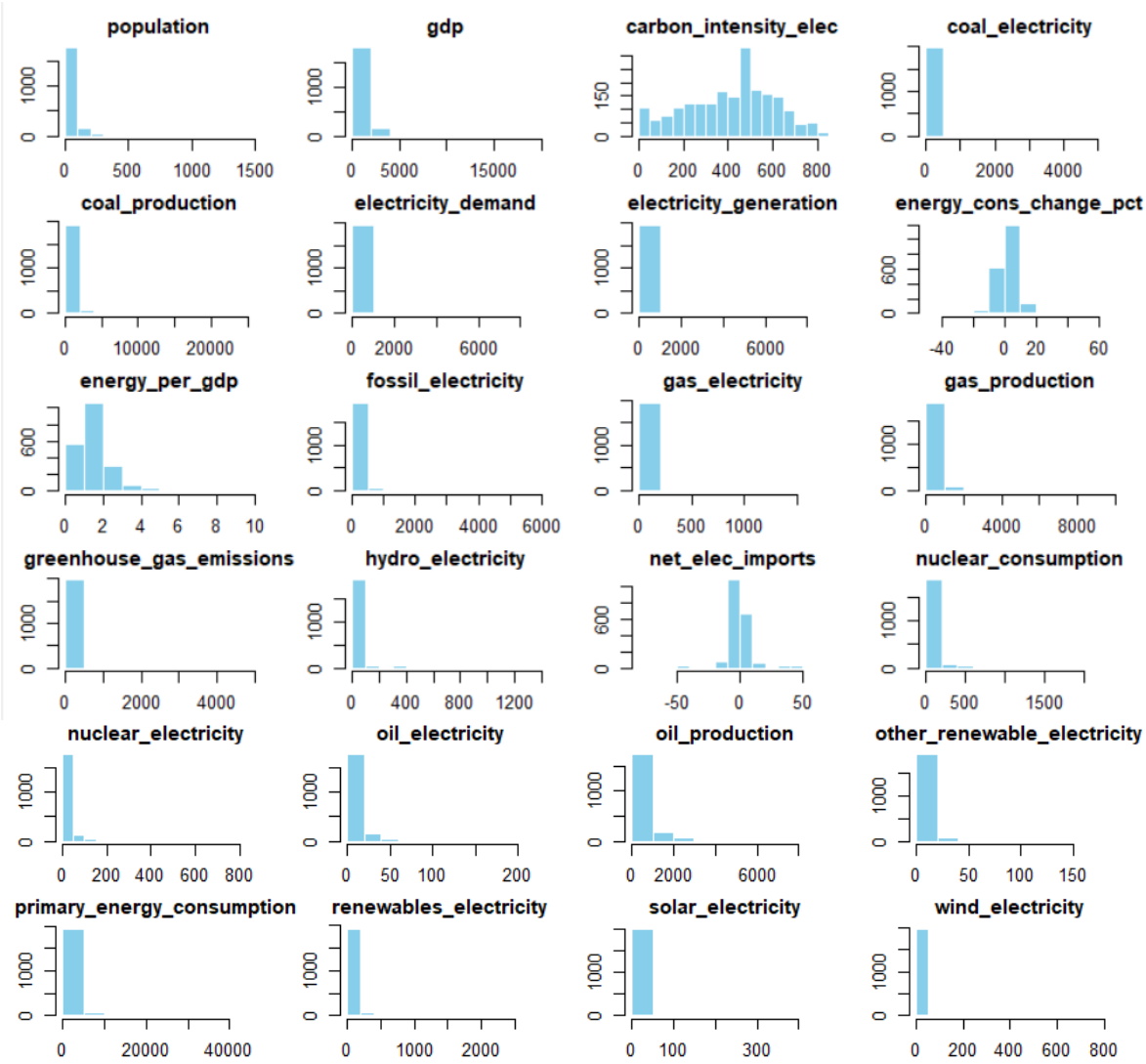
Veamos ahora que hay bastantes menos valores faltantes.



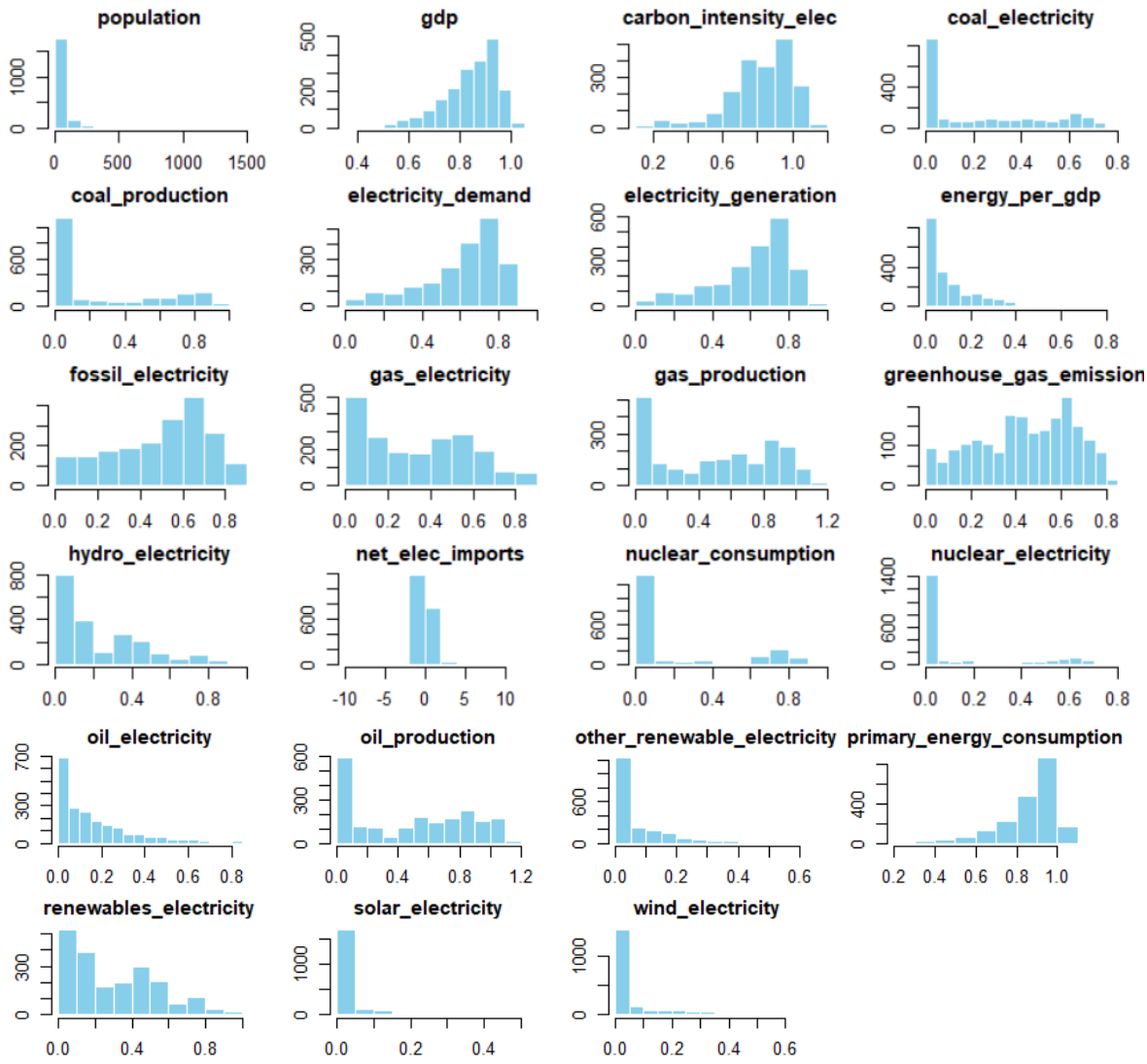
Tras imputar los datos y aplicar las dos transformaciones, vamos a ver la distribución de los datos.

Las transformaciones realizadas consisten en un escalado de todas las variables en función a la población del país y en una conversión logarítmica log(1+x)(el la distribución que viendo el histograma aparentaba corregir la mayoría de variables) se ha excluido net electricity imports de esta transformación logarítmica ya que tomaba valores negativos.

Observamos los histogramas pretratamiento para todas las variables:



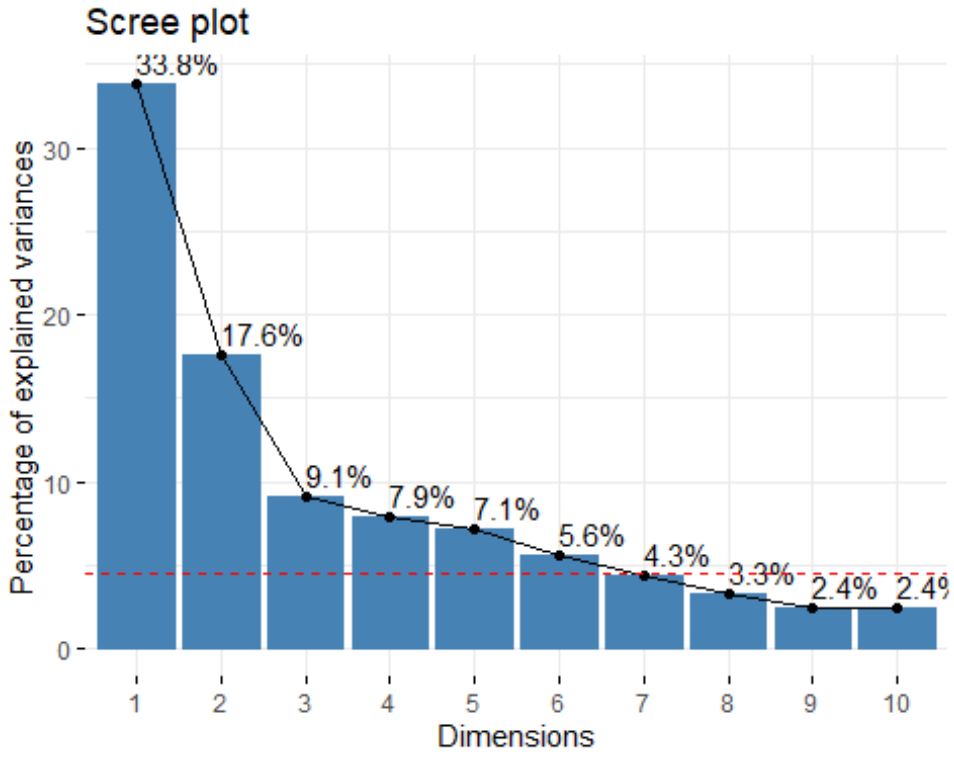
Seguidamente los histogramas post tratamiento:



No en todas las variables se ha solucionado el problema pero en la mayoría de ellas se aprecia una corrección, es por esto que, nos quedaremos con esta transformación final.

Anexo B

Realizamos un scree plot para poder elegir el número de componentes adecuado para nuestro modelo, a la vista del siguiente gráfico escogemos 4 componentes.

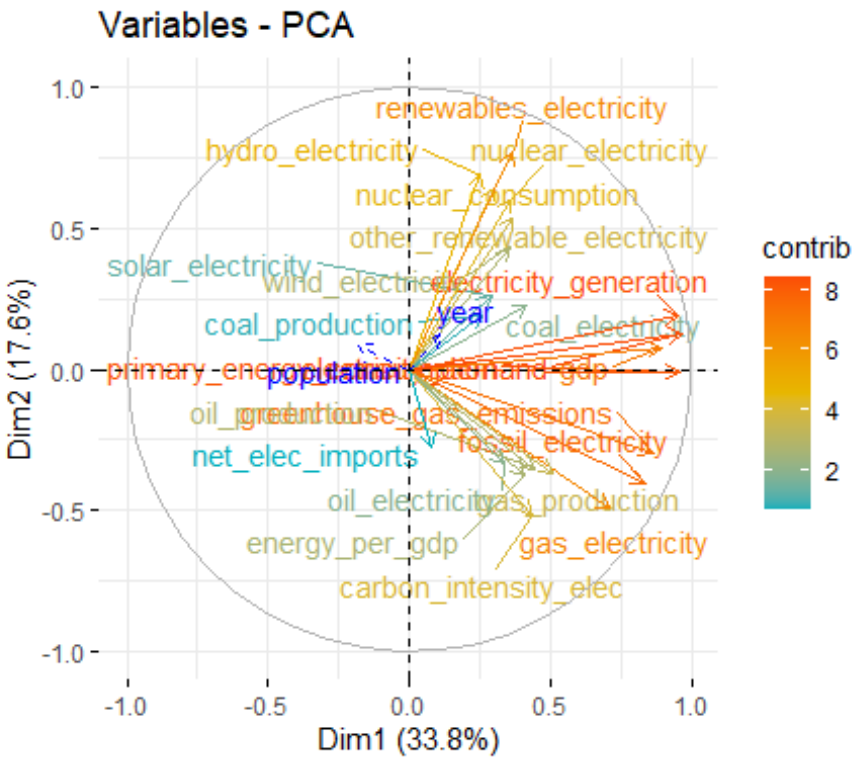


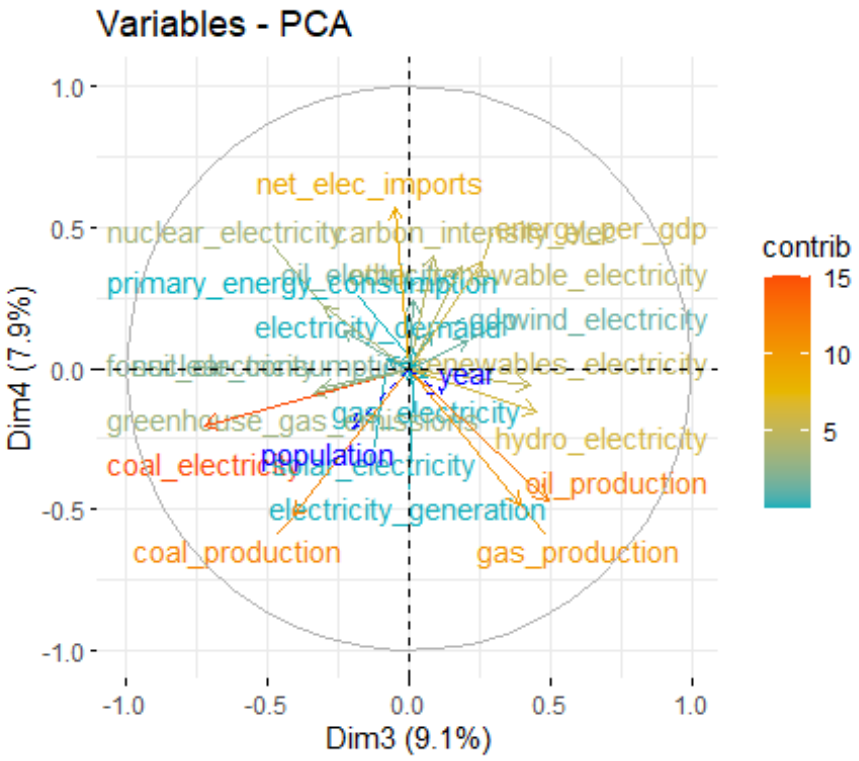
Elegimos 4 componentes principales. Observamos que con estas 3 componentes solo se explica un 68.427818 de la variabilidad total. También observamos que tampoco hemos escogido todas las componentes que explican más de lo que explicaría cada una de ellas si todas explicasen lo mismo.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	7.4404223	33.8201016	33.82010
Dim.2	3.8746408	17.6120035	51.43211
Dim.3	2.0090981	9.1322641	60.56437
Dim.4	1.7299587	7.8634488	68.42782
Dim.5	1.5675699	7.1253179	75.55314
Dim.6	1.2242226	5.5646482	81.11778
Dim.7	0.9536067	4.3345759	85.45236
Dim.8	0.7321611	3.3280051	88.78037
Dim.9	0.5379736	2.4453344	91.22570
Dim.10	0.5262903	2.3922285	93.61793
Dim.11	0.3401370	1.5460773	95.16401
Dim.12	0.3035637	1.3798352	96.54384
Dim.13	0.2474796	1.1249073	97.66875
Dim.14	0.1579067	0.7177579	98.38651
Dim.15	0.1245573	0.5661695	98.95268

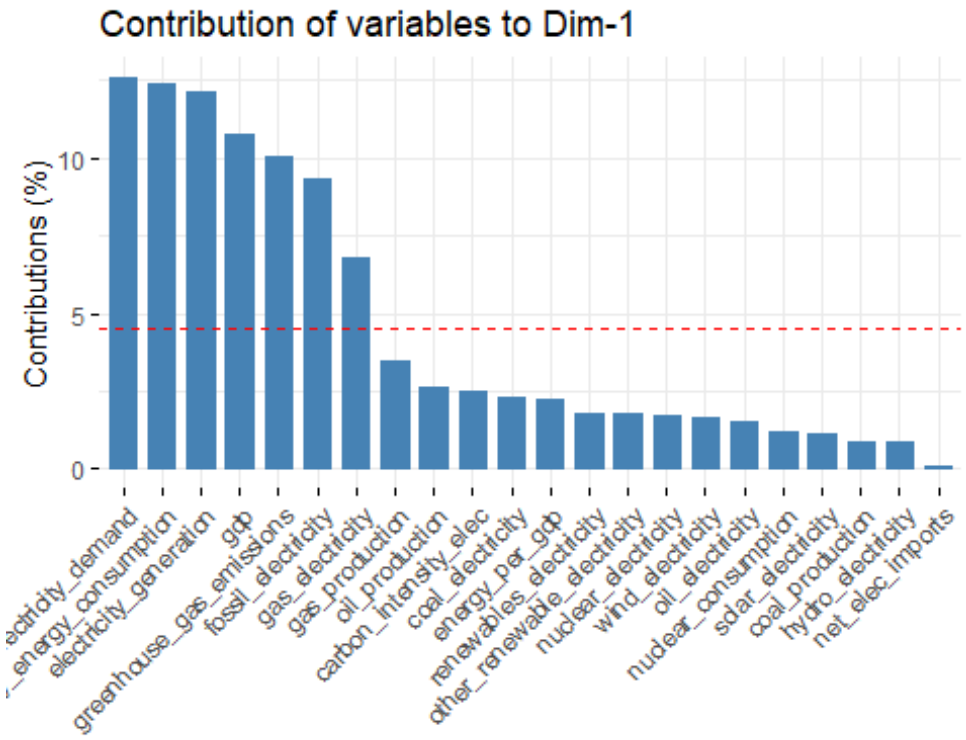
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.16	0.1023904	0.4654111	99.41809
Dim.17	0.0657627	0.2989215	99.71701
Dim.18	0.0352153	0.1600695	99.87708
Dim.19	0.0153772	0.0698965	99.94697
Dim.20	0.0072421	0.0329185	99.97989
Dim.21	0.0024475	0.0111252	99.99102
Dim.22	0.0019761	0.0089824	100.00000

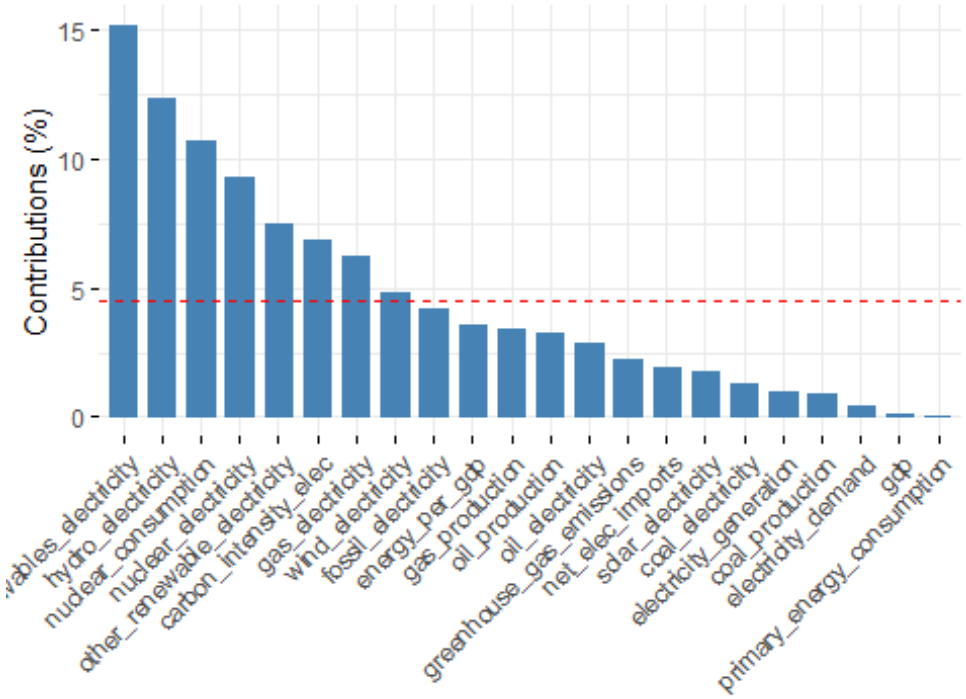
Vemos ahora los loading plots de todas las variables.

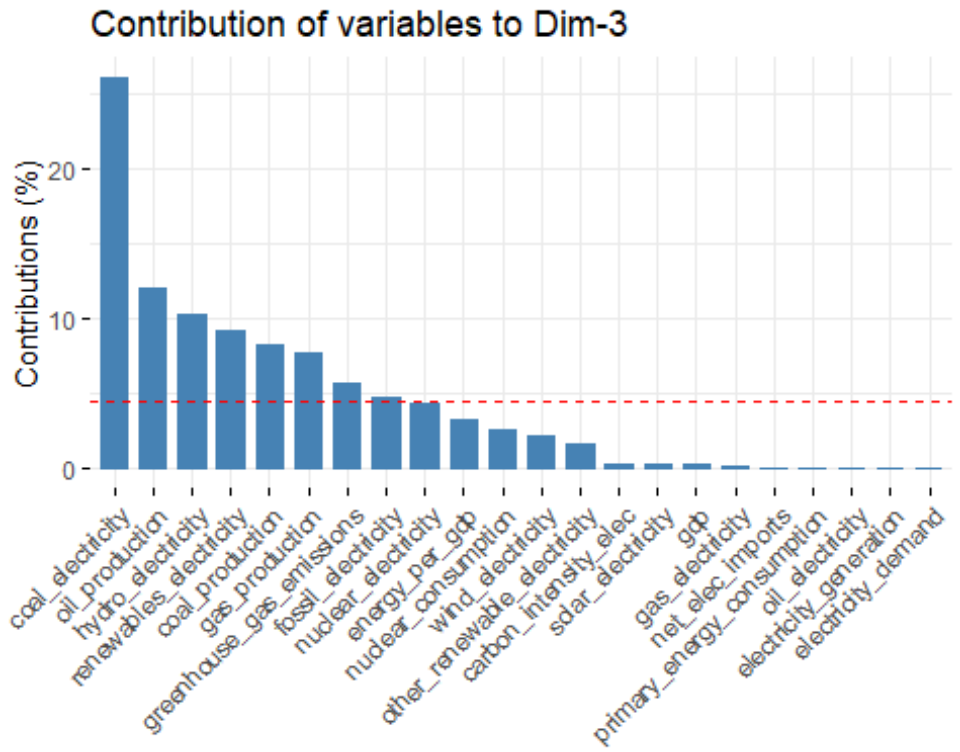


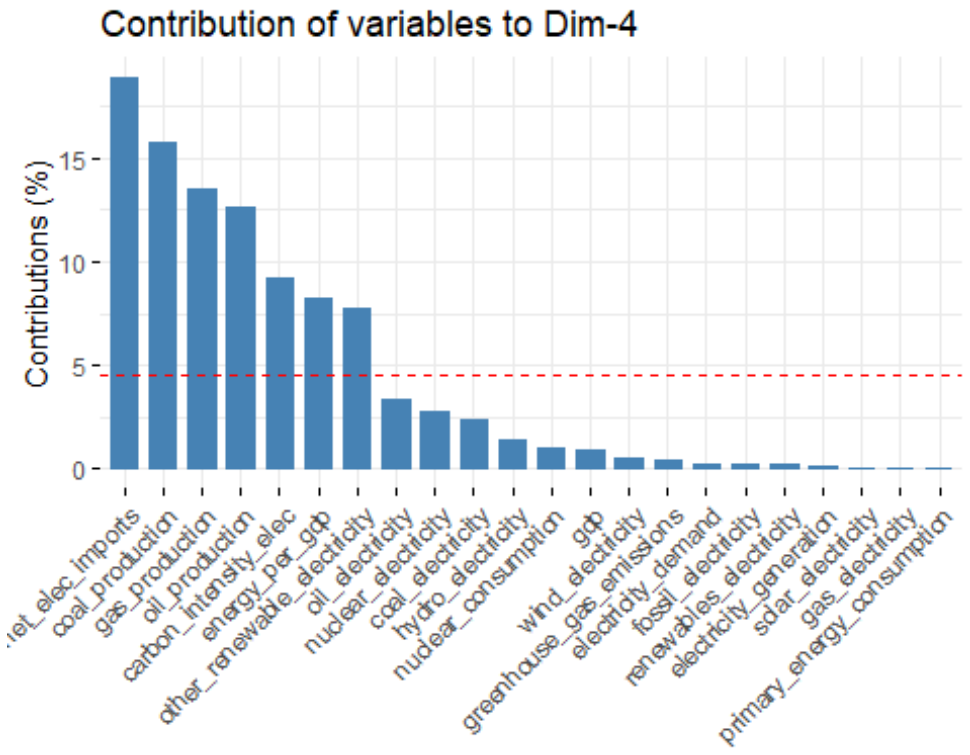


Obtenemos los gráficos de contribuciones para obtener las variables más importantes en cada componente y hacer los loadings plots por componentes, ya que estos salen muy cargados y son poco comprensibles.

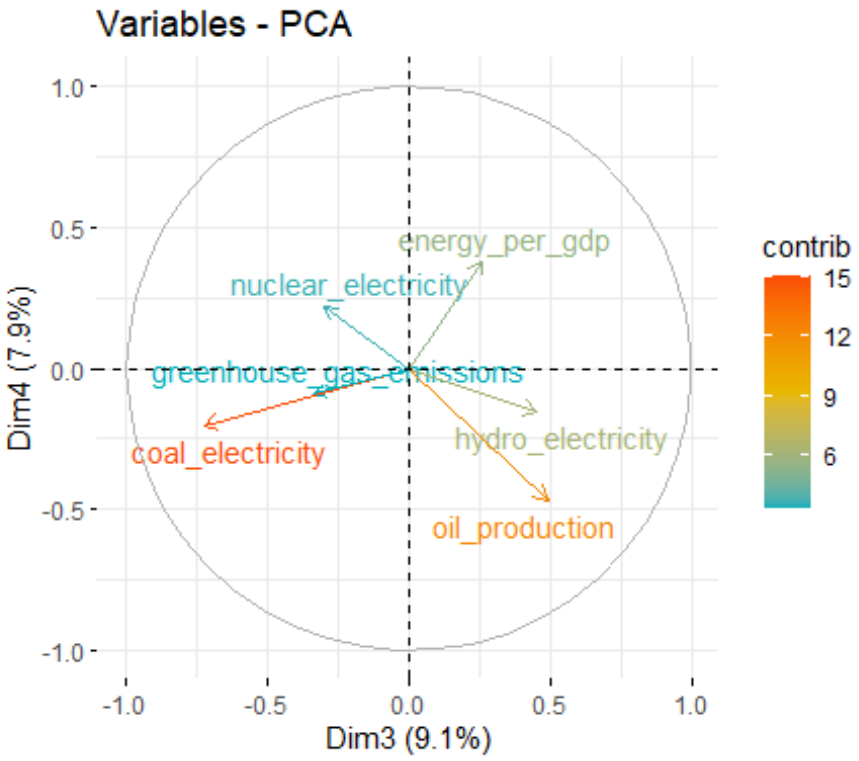

Contribution of variables to Dim-1
Contributions (%)
0
5
10
ectricity_demand
_energy_consumption
electricity_generation
gdp
greenhouse_gas_emissions
fossil_electricity
gas_electricity
gas_production
oil_production
carbon_intensity_elec
coal_electricity
energy_per_gdp
renewables_electricity
other_renewable_electricity
nuclear_electricity
wind_electricity
oil_electricity
nuclear_consumption
solar_electricity
coal_production
hydro_electricity
net_elec_imports

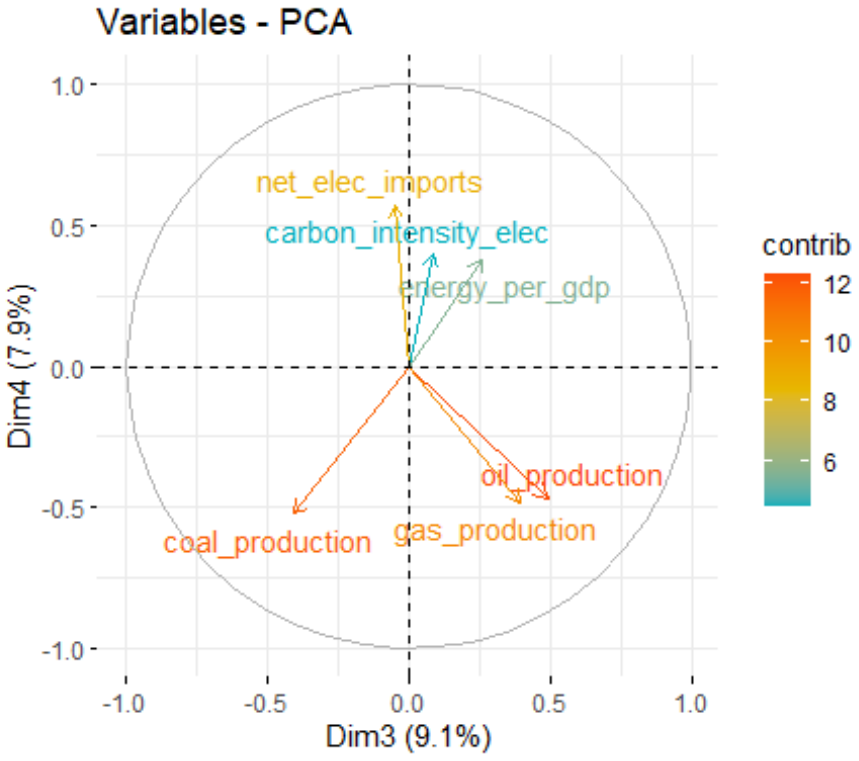

Contribution of variables to Dim-2
Contributions (%)
0
5
10
15
ables_electricity
hydro_electricity
nuclear_consumption
nuclear_electricity
other_renewable_electricity
carbon_intensity_elec
gas_electricity
wind_electricity
fossil_electricity
energy_per_gdp
gas_production
oil_production
oil_electricity
greenhouse_gas_emissions
net_elec_imports
solar_electricity
coal_electricity
electricity_generation
coal_production
electricity_demand
gdp
primary_energy_consumption





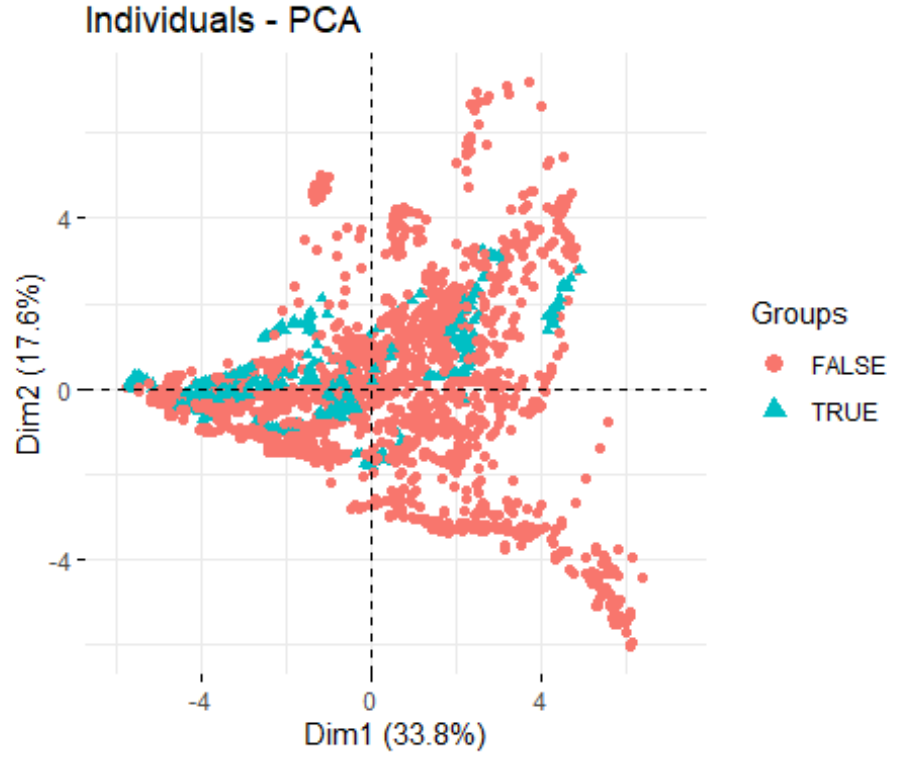
Los loading plots de las componentes 3 y 4 no mostradas en el apartado del PCA.

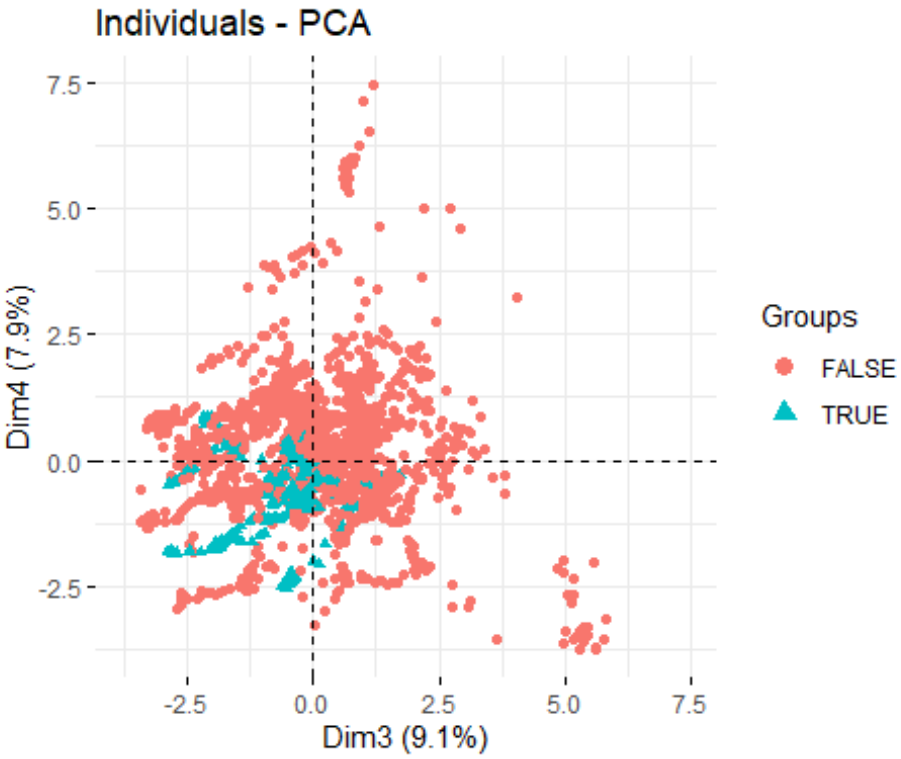

Variables - PCA
energy_per_gdp
nuclear_electricity
greenhouse_gas_emissions
coal_electricity
hydro_electricity
oil_production
contrib
15
12
9
6
Dim4 (7.9%)
Dim3 (9.1%)
1.0
0.5
0.0
-0.5
-1.0


Variables - PCA
net_elec_imports
carbon_intensity_elec
energy_per_gdp
oil_production
coal_production
gas_production
contrib
12
10
8
6
Dim4 (7.9%)
Dim3 (9.1%)
1.0
0.5
0.0
-0.5
-1.0

Observamos que la 4 componente diferencia los países importadores y exportadores de electricidad, siendo estos últimos los que en promedio más producción de gas, crudo y carbón tienen. De aquí podemos concluir que aquellos países con reservas ya sean vegetales o fósiles, tenderán a abastecerse ellos mismos de energía.

Vemos los dos grupos de países poblados en la primera componente.





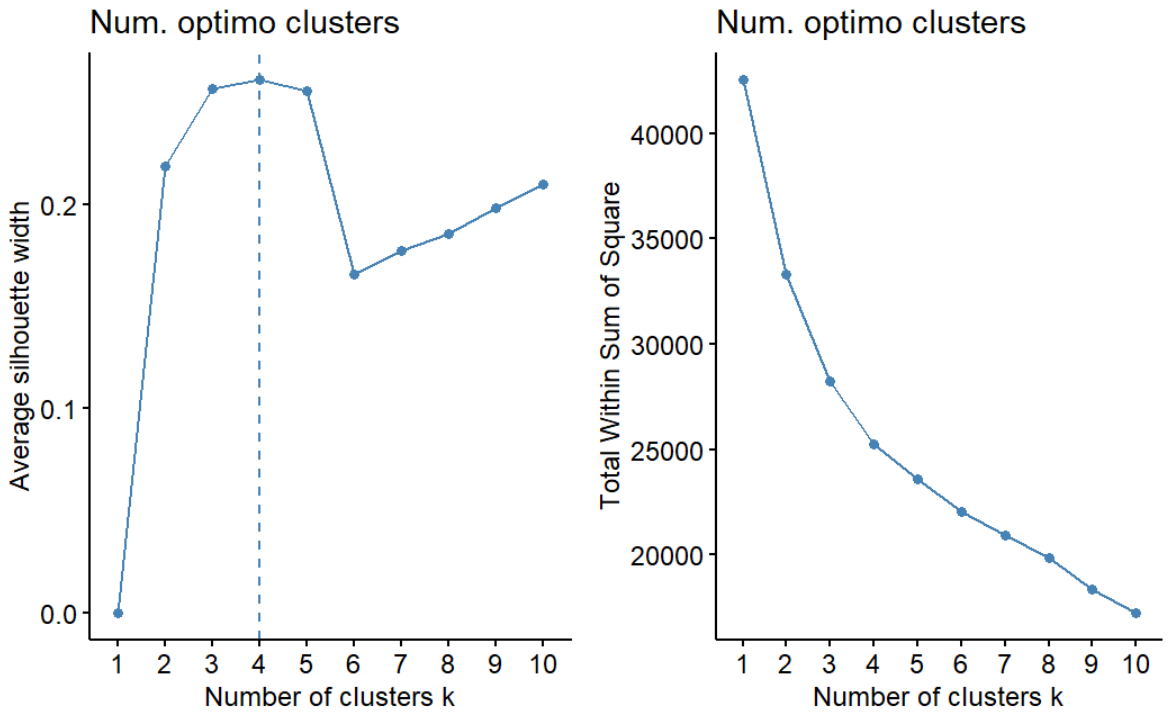
Anexo C

Estos son los coeficientes de Hopkins con las dos distancias Euclídea y Manhattan respectivamente.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9509  0.9525  0.9533  0.9559  0.9600  0.9620

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9576  0.9647  0.9681  0.9672  0.9708  0.9752
```

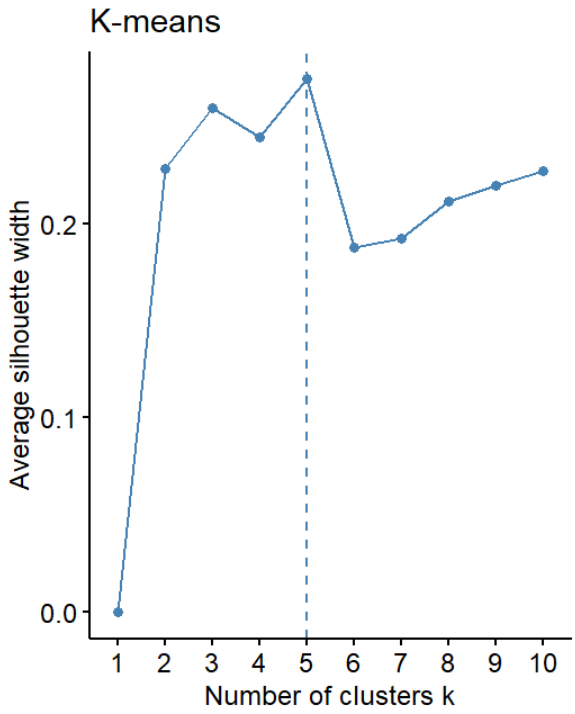
Como la media del estadístico de Hopkins para Euclídea da 0.9559 y para Manhattan da 0.9672, podemos concluir que hay más tendencia agrupamiento con la distancia de Manhattan así que usaremos esta para los modelos jerárquicos.

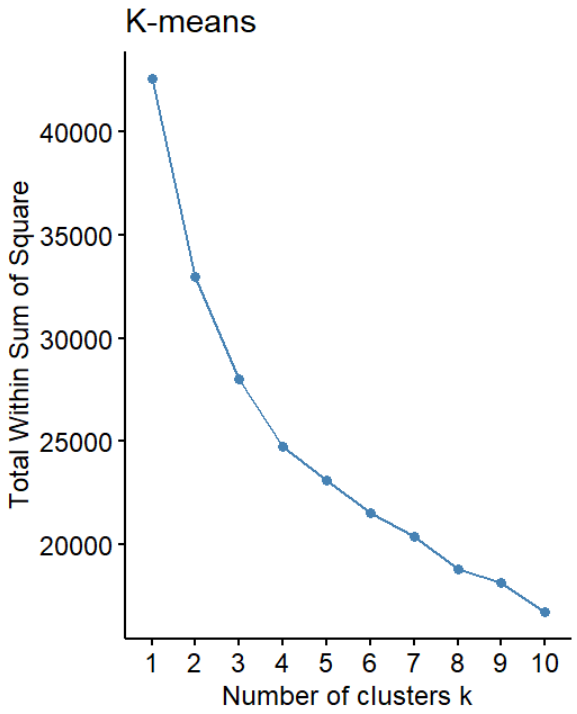
Modelo jerárquico



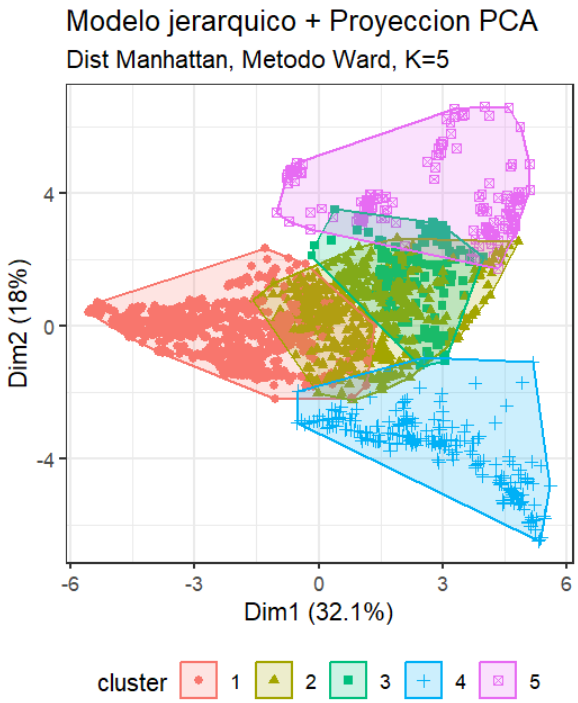
Los resultados para el coeficiente Silhouette indican que el número óptimo de clústers es 4 y además la suma de cuadrados intra-clúster es baja para k=4 y parece ser el punto en el que se crea el codo.

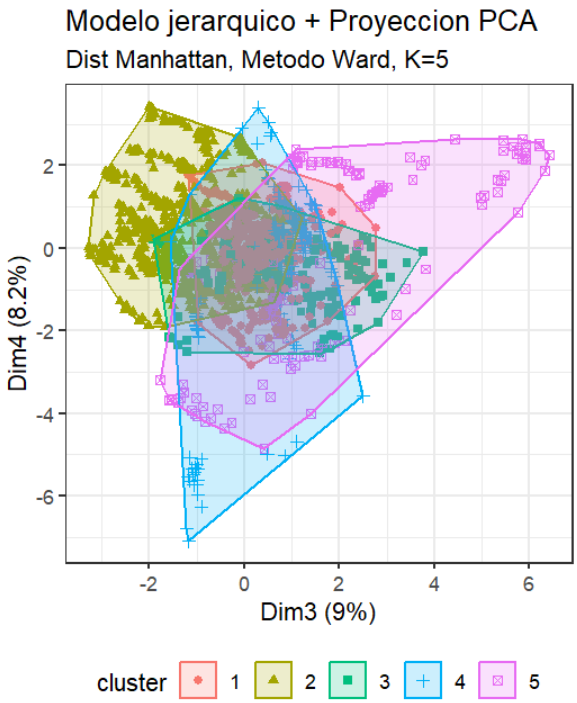
Modelos de partición



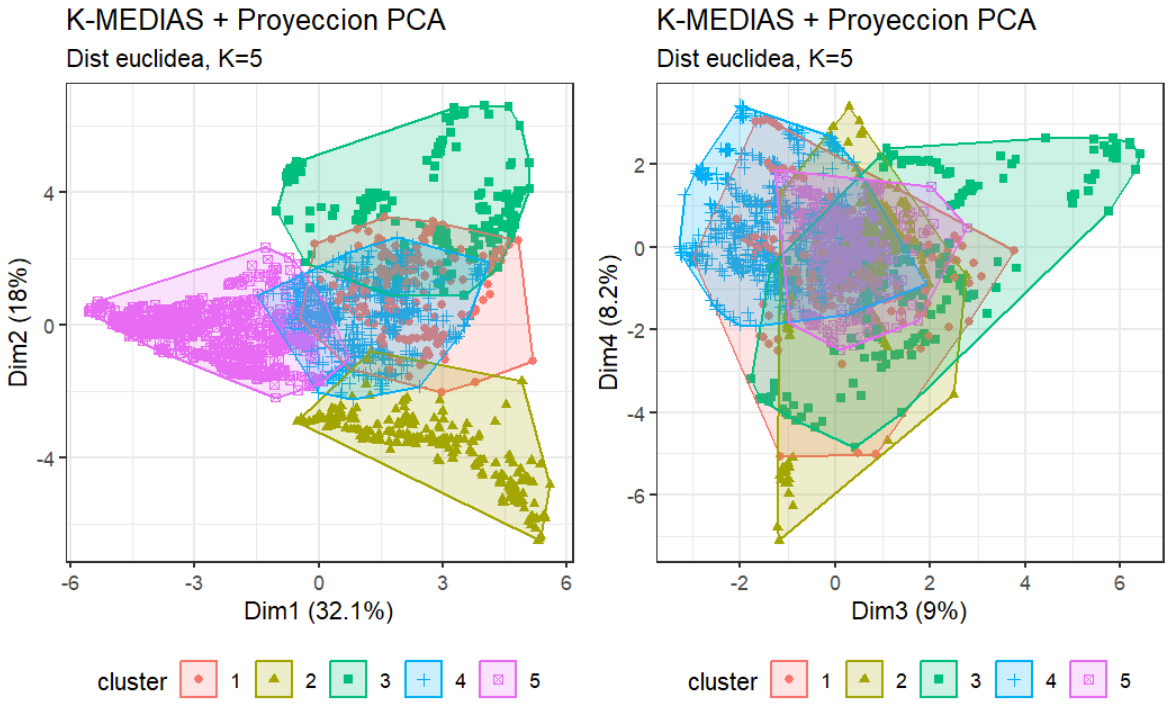


Los resultados para el coeficiente Silhouette indican que el número óptimo de clústers es 5 y además la suma de cuadrados intra-clúster es baja para k=5 y parece ser el punto en el que se crea el codo.

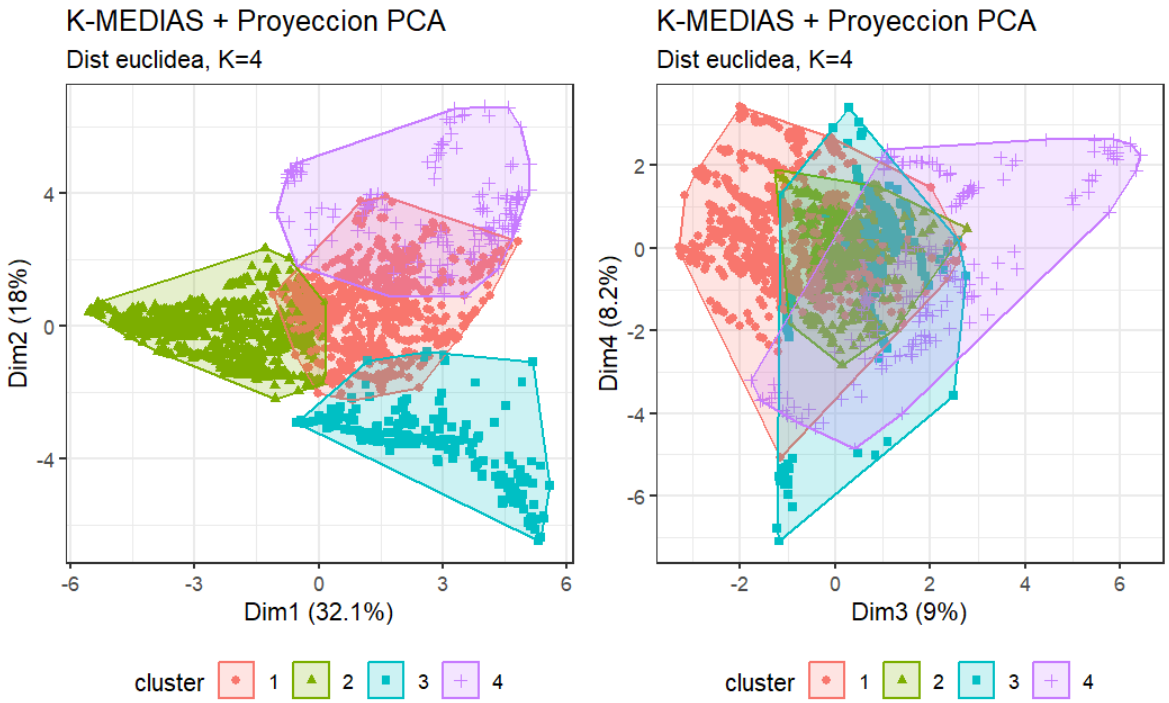




El análisis gráfico con PCA muestra que con las dos primeras componentes principales se solapan principalmente los clústers 1, 2 y 3. En la segunda componente se pueden diferenciar los clústers 4 y 5. Mientas que para la tercera y cuarta componente, se solapan los 5 clústers.



Ahora lo hemos hecho para K-medias, tanto para k=5 como k=4. Para k=5 podemos apreciar que en las dos primeras componentes se solapan todos los clústers, menos el clúster nº2. También podemos apreciar que la segunda componente es la que separa los clústers 2 y 3. Mientras que para las componentes tres y cuatro, se solapan todos los clústers, al igual como pasaba en el modelo jerárquico.

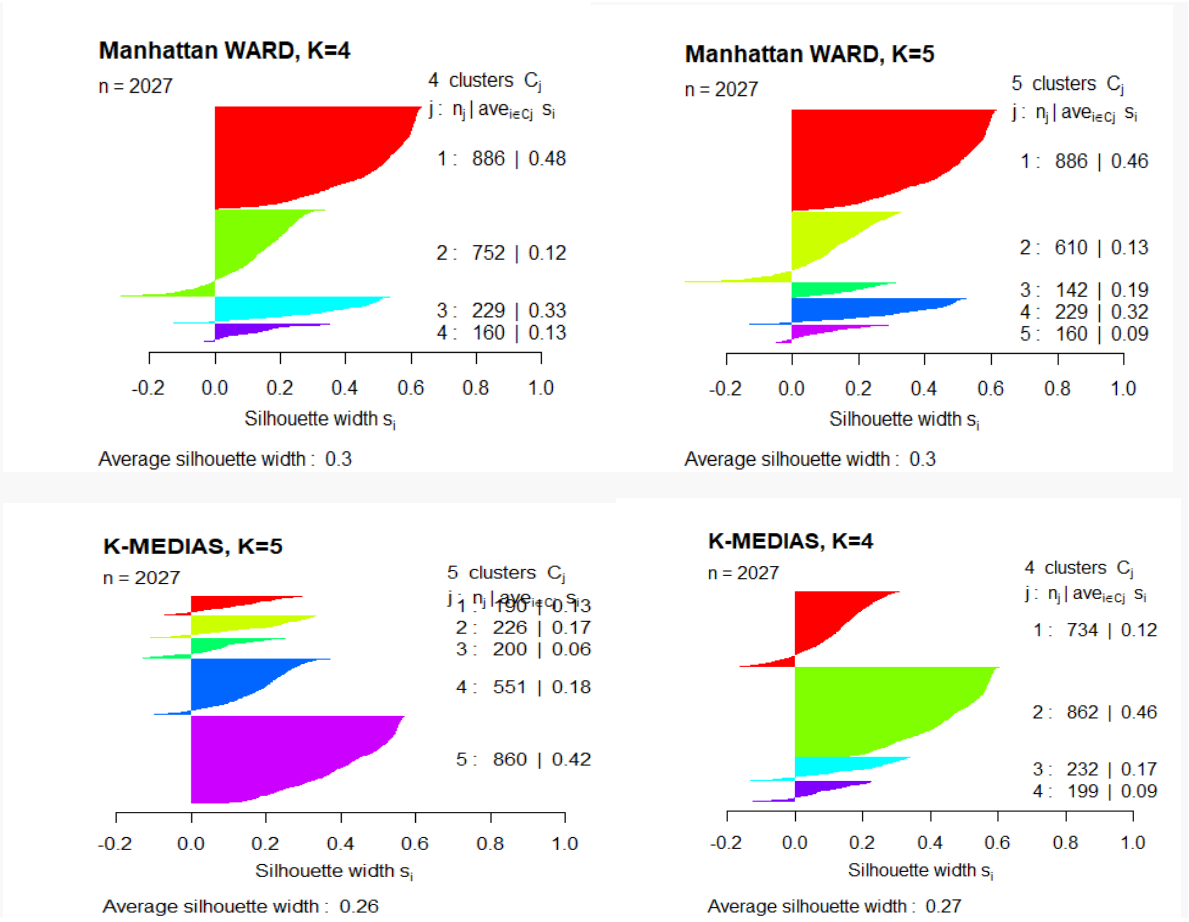


Para k=4 podemos apreciar la primera componente separa a los clústers 2 y 3. Mientras que la segunda componente separa a los clústers 3 y 4. En la tercera y cuarta componente, al igual que anteriormente, se solapan todos los clústers

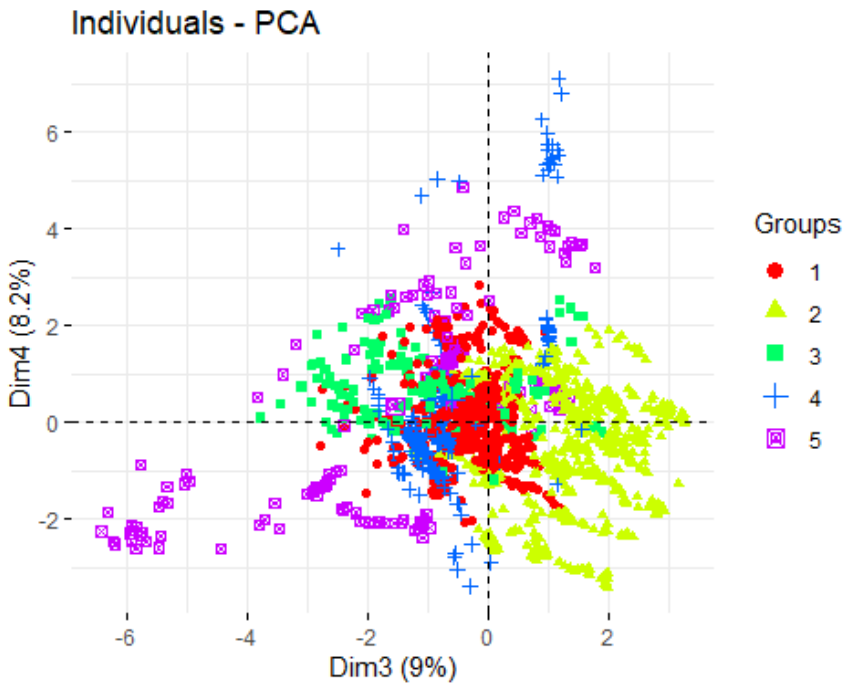
Índices de Dunn

A continuación, hemos calculado el índice de Dunn. Como se puede apreciar, el índice de Dunn es más alto para el método jerárquico de Ward utilizando la distancia Manhattan. Por consiguiente, emplearemos este enfoque para nuestro análisis.

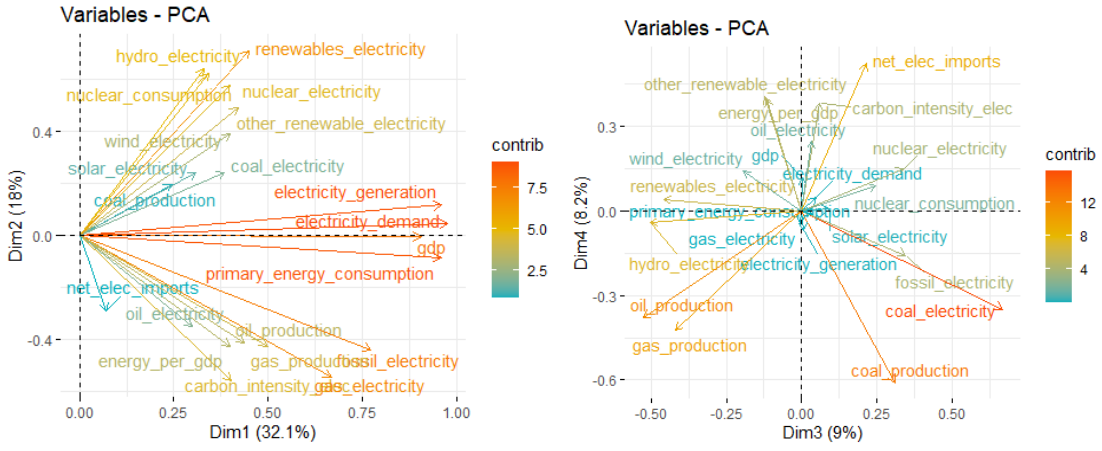
```
## [1] "KMeans"

## $dunn
## [1] 0.01907349

## [1] "Jerárquico Ward con dist Manhattan, K = 5"

## $dunn
## [1] 0.0197391

## [1] "Jerárquico Ward con dist Manhattan, K = 4"

## $dunn
## [1] 0.0197391
```



Aquí vemos los clústers en la 3ª y 4ª componente en una proyección PCA.

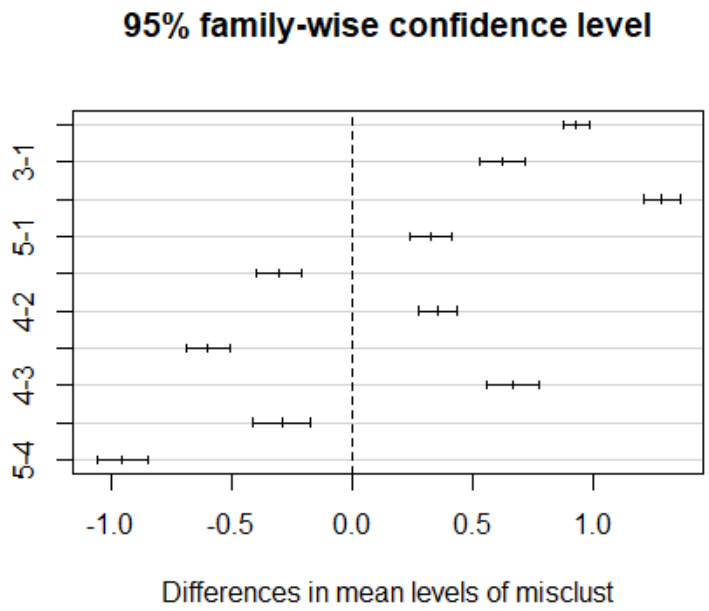


Los loading plots para las interpretaciones de los clústers.



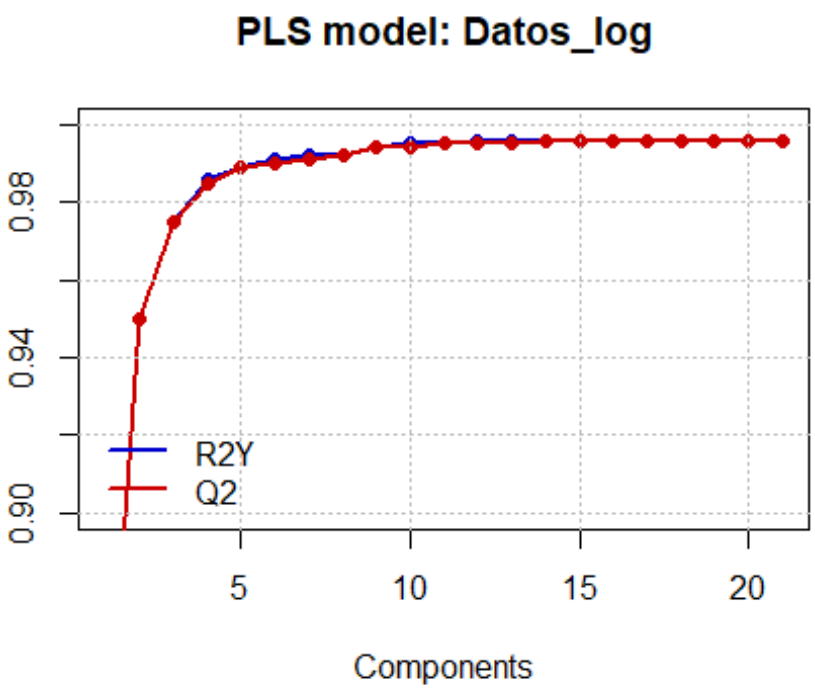
Estos son resultados del test Anova usando la clasificación por clústers como factor y las emisiones como respuesta, así como los intervalos de Tukey.

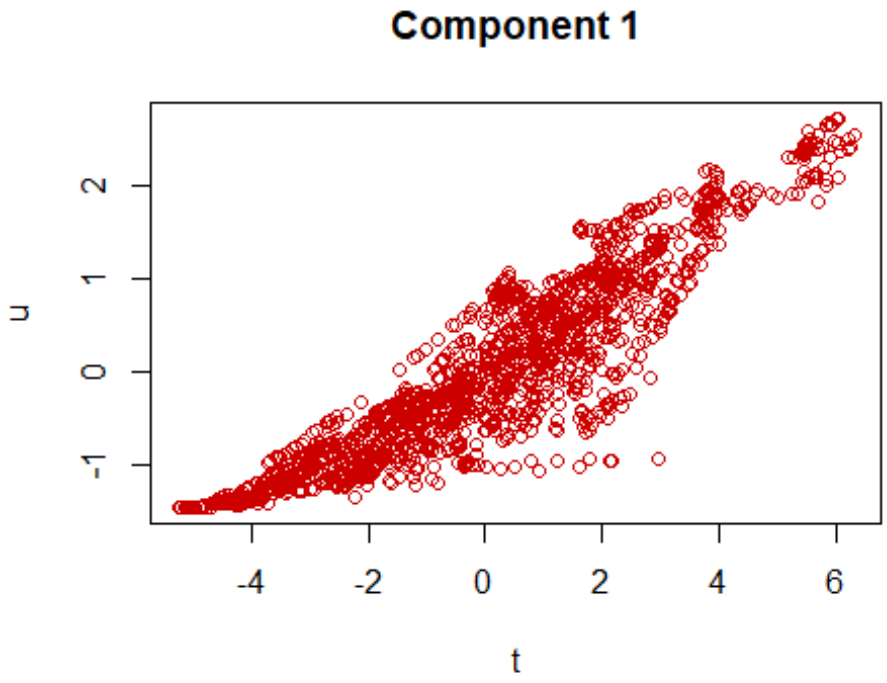
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## misclust       4  481.1  120.28   852.9 <2e-16 ***
## Residuals   2022  285.2    0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

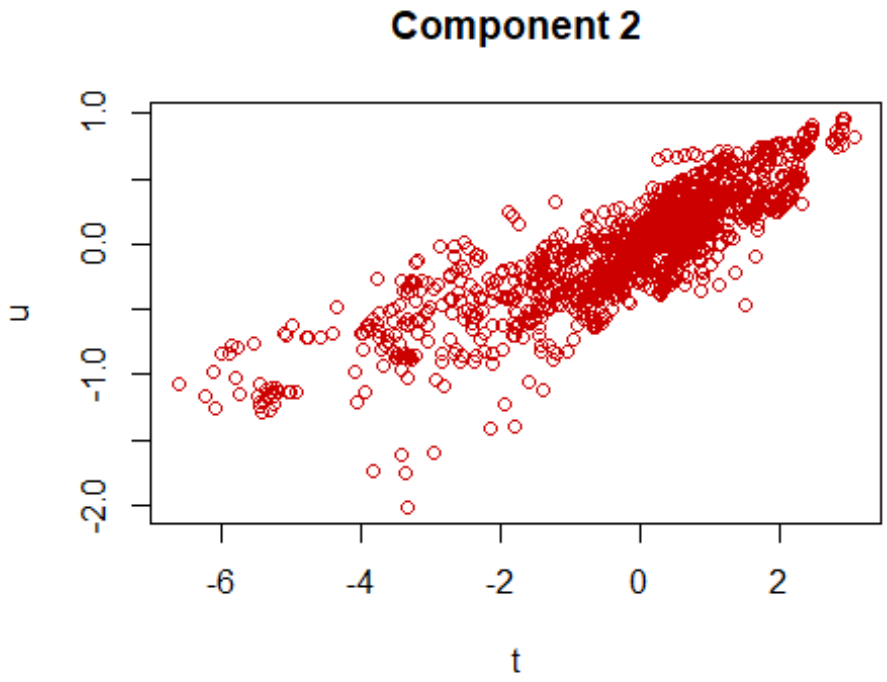


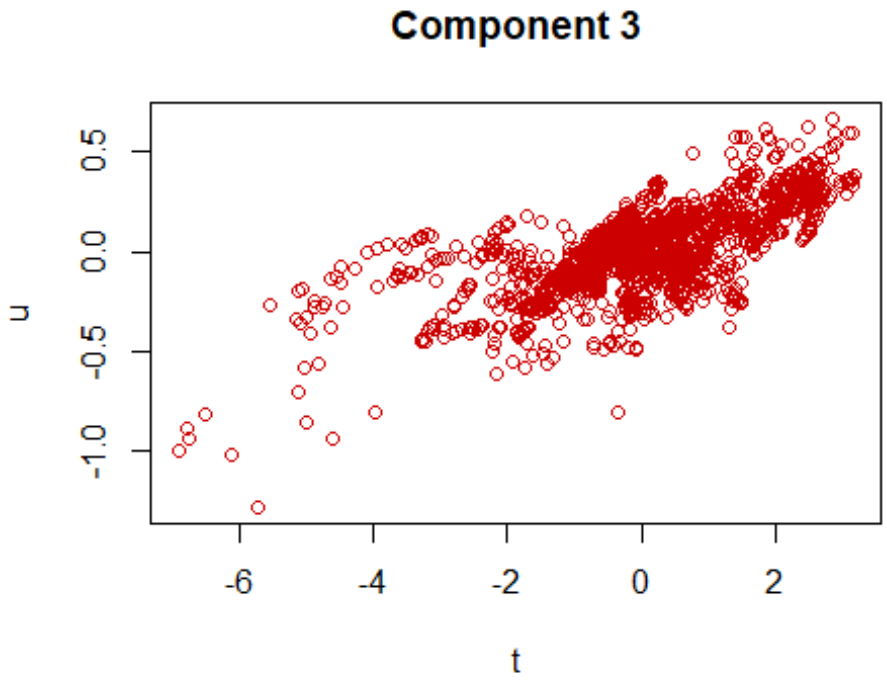
Anexo D

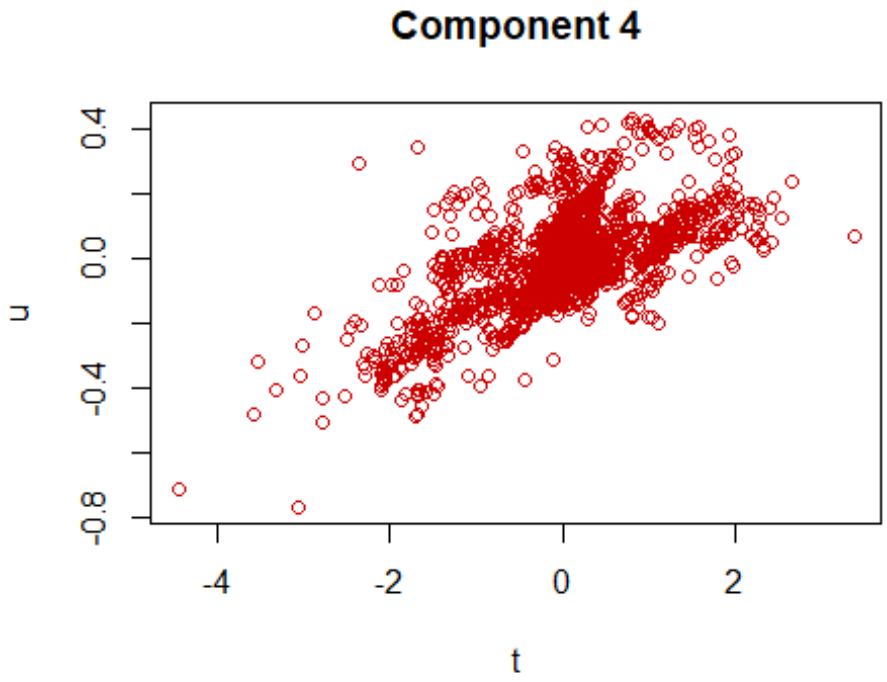
```
## PLS
## 1623 samples x 21 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum)  RMSEE pre ort pR2Y  pQ2
## Total        1    0.996   0.996 0.0367  21   0 0.05 0.05
```


PLS model: Datos_log
0.90
0.94
0.98
R2Y
Q2
5
10
15
20
Components


Component 1
u
t
-4
-2
0
2
4
6
-1
0
1
2

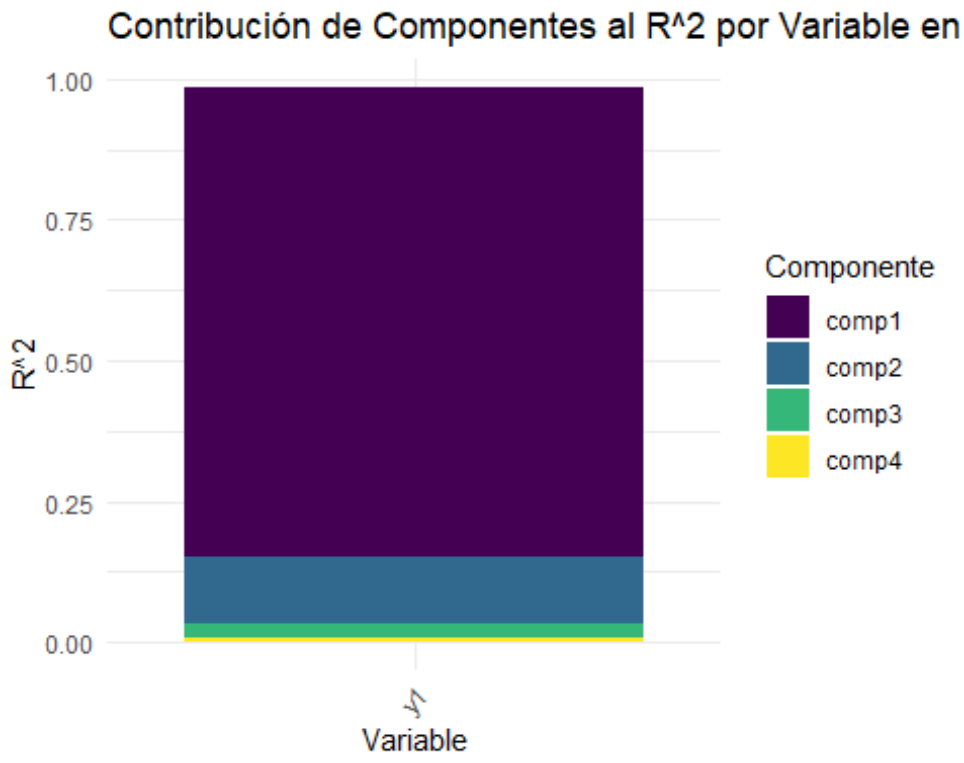

Component 2
u
t
-6
-4
-2
0
2
-2.0
-1.0
0.0
1.0



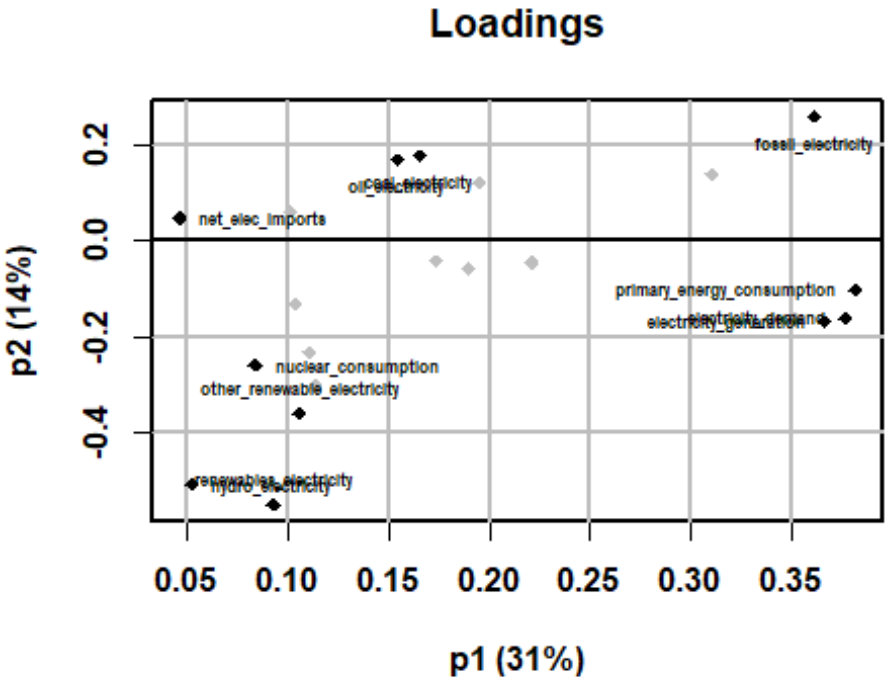


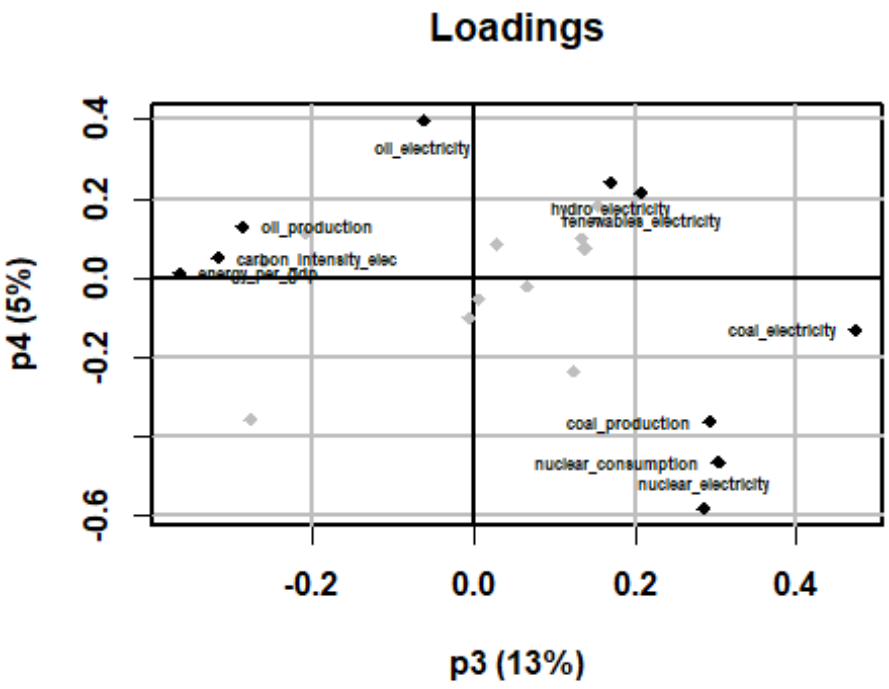

```
##        p1        p2        p3        p4
## 0.9130064 0.8481670 0.7026875 0.6311769
```

Podemos asumir la linealidad entre los scores t y u.

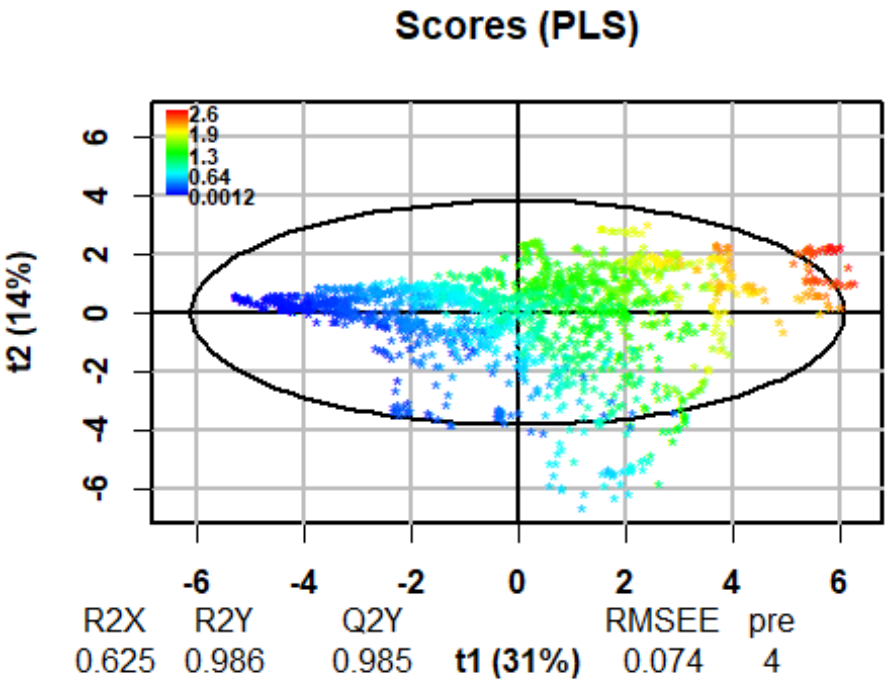


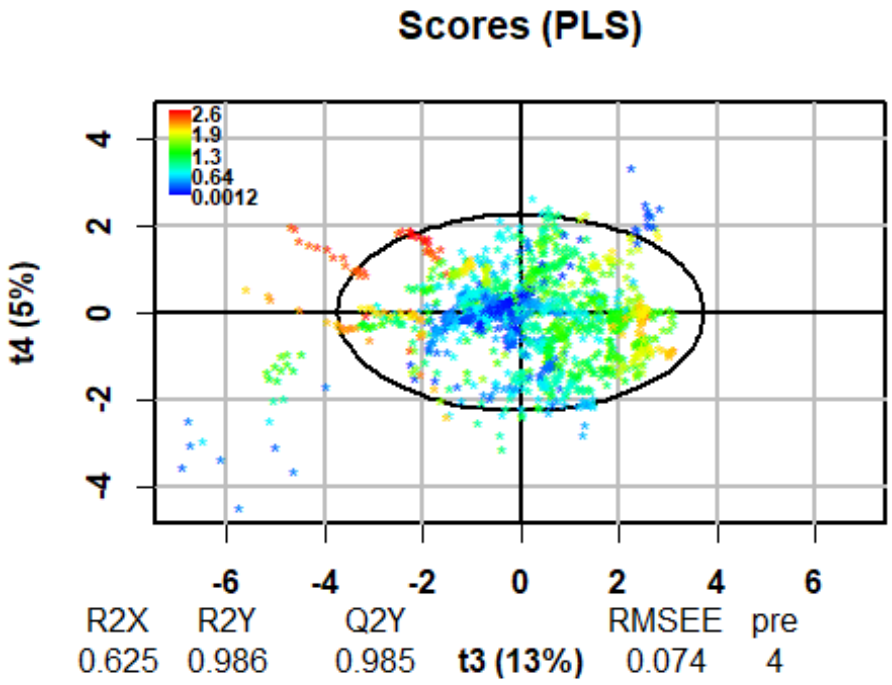
Loading plots.





Y los score plots.





Coloreados por emisiones.