Análisis de Datos

Biomarcadores prematuros de la enfermedad de Parkinson

Autores:

Pau Vives López, Harold Cruz Lorenzo,
Samuel de Paúl Smith

Universitat
de les Illes Balears

Contextualización, objetivos y descripción de los datos

La enfermedad de Parkinson es un transtorno neurodegenerativo crónico caracterizado por los temblores, rigidez y disminución de la mobilidad. Esta enfermedad se debe a un déficit en la secreción de dopamina, hormona liberada por las terminaciones nerviosas de la sustancia negra. A veces comienza con un temblor apenas perceptible en una sola mano. En las etapas iniciales de la enfermedad de Parkinson, el rostro puede tener una expresión leve o nula. Es posible que los brazos no se balanceen al caminar. El habla puede volverse suave o incomprensible. Los síntomas de la enfermedad de Parkinson se agravan a medida que la enfermedad progresa con el tiempo.

A pesar de que la enfermedad de Parkinson no tiene cura, los medicamentos pueden reducir o atenuar notablemente los síntomas. En ocasiones, el médico puede sugerir realizar una cirugía para regular determinadas zonas del cerebro.

Esta enfermedad representa el segundo trastorno neurodegenerativo por su frecuencia, sólo por detrás del Alzheimer. Está extendida por todo el mundo y puede desarrollarse en ambos sexos, afectando a entre un 1% a un 2% de la población mayor de 60 años.

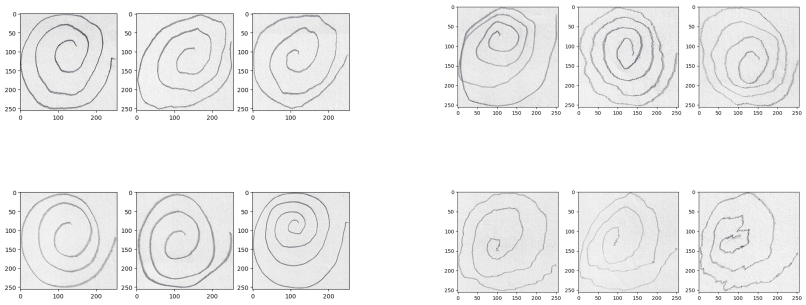


Figure 1: Test de la espiral en pacientes sanos (izquierda) y pacientes con Parkinson (derecha)

Hemos obtenido el dataset con el que vamos a trabajar de la web Kaggle, bajo el título “Early Biomarkers of Parkinson’s Disease”, que podría traducirse como “Biomarcadores prematuros de la enfermedad de Parkinson” Los datos provienen originalmente de un paper publicado a principios de 2017 en nature.com, que puede consultarse en el enlace https://www.nature.com/articles/s41598-017-00047-5.pdf

Hemos elegido este tema porque nos parece que la investigación científica y médica es una área que se beneficia mucho del análisis de datos ya que abunda la información puesto que se realizan todo tipo de pruebas y diagnósticos a los pacientes. Además, estudios de este tipo están enfocados en mejorar la calidad de vida de los pacientes afectados por la enfermedad, y claramente nos parece que merece la pena trabajar de cara a éste objetivo.

El conjunto de datos incluye 30 pacientes con enfermedad de Parkinson (EP) temprana no tratada, 50 pacientes con trastorno de conducta del sueño REM (RBD), que tienen un alto riesgo de desarrollar la enfermedad de Parkinson, y 50 controles sanos (HC).

Todos los pacientes fueron evaluados clínicamente por un neurólogo profesional con experiencia en trastornos del movimiento. Todos los sujetos fueron examinados durante una sola sesión con un especialista del habla. Éstos realizaron la lectura de un texto estandarizado, fonéticamente equilibrado de 80 palabras y monólogos sobre sus intereses, trabajo, familia o actividades actuales durante aproximadamente 90 segundos. Las características del habla fueron analizadas automáticamente por Jan Hlavnička et al.

Con el análisis de éste conjunto de datos se pretende:

- Hallar biomarcadores de la enfermedad de Parkinson en los distintos pacientes estudiados
- Clasificar a los pacientes en grupos de riesgo
- Distinguir qué rasgos están más estrechamente relacionados con el desarrollo del Parkinson

Descripción de los datos

En el dataset original las variables tienen nombres muy largos y engorrosos por lo que primero las renombraremos, para poder trabajar con ellas con mas facilidad, y a continuación detallaremos la información que aporta cada una.

El dataset consta de las siguientes variables para cada una de las observaciones:

code: (carácter) Contiene un código de identificación para cada paciente.
age: (numérica) Edad de cada paciente en años.
gender: (categórica con 2 niveles) Género del paciente.
history: (categórica de 2 niveles) Variable que indica si el paciente tiene familiares con Parkinson.
onset: (numérica) Edad del paciente al inicio de la enfermedad, en años.
duration: (numérica) Duración de la enfermedad desde los primeros síntomas, en años.
antidepr: (categórica con 10 niveles): Terapia con Antidepresivos del paciente, en caso afirmativo se especifica cuál.
antipark: (categórica con 1 nivel): Medicación antiparkinsoniana del paciente.
antipsych: (categórica con 1 nivel): Medicación antipsicótica del paciente.
benzodiazepine: (Categórica con 4 niveles): Medicación con benzodiazepinas del paciente, en caso afirmativo se especifica cuál.
levodopa: (numérica) Cantidad en miligramos del consumo de Levodopa diario del paciente.
clonazepam: (numérica) Cantidad en miligramos del consumo de Clonazepam diario del paciente.
HY_scale: (categórica con 7 niveles) Mide el estado de la discapacidad funcional asociada a la enfermedad de Parkinson en la escala Hoehn-Yahr
UPDRSIII: (numérica) Puntuación total en la "UNIFIED PARKINSON'S DISEASE RATING SCALE"

Evaluación Motora y Cognitiva:

Las siguientes variables son los resultados de unas pruebas realizadas exclusivamente a los sujetos que no están completamente sanos, relacionadas con esta enfermedad (por ejemplo, mediciones sobre los temblores en un brazo)

Hay variables en este grupo que contienen resultados de una misma prueba, pero para direntes regiones del cuerpo. Por tanto, para describir todas las variables de una manera resumida definiremos las pruebas realizadas y el significado de las abreviaturas que acompañan a la variable (por ejemplo, la variable "X19-LRU" es el resultado de la prueba "X19" sobre la zona "LRU" del cuerpo)

NOTA: Todas estas variables son de tipo categórico, tomando como valores números enteros a partir del 0 que miden las complicaciones que experimenta el paciente para realizar dicha prueba.

- *Pruebas:*

Éstas variables miden diferentes pruebas de la escala UPDRS III, que introducida en 1987, es la escala más ámpliamente utilizada para el análisis de la situación clínica de un paciente parkinsoniano.

X18: Pronunciación
X19: Expresión facial
X20: Temblores en reposo
X21: Temblor de acción o postural
X22: Rigidez
X23: Golpes con la punta de los dedos
X24: Movimientos de la mano
X25: Movimientos rápidos alternados
X26: Agilidad en piernas
X27: Levantarse de una silla
X28: Postura
X29: Forma de andar

X30: Estabilidad postural
X31: Bradicinesia y hipocinesia corporal

- *Zonas del cuerpo*

RUE: ("Right Upper Extremity") Hace referencia a la extremidad superior derecha.
LUE:("Left Upper Extremity") Hace referencia a la extremidad superior izquierda.
RLE:("Right Lower Extremity") Hace referencia a la extremidad inferior derecha.
LLE:("Left Lower Extremity") Hace referencia a la extremidad inferior izquierda.
head: Cabeza
neck: Cuello

Mediciones acústicas:

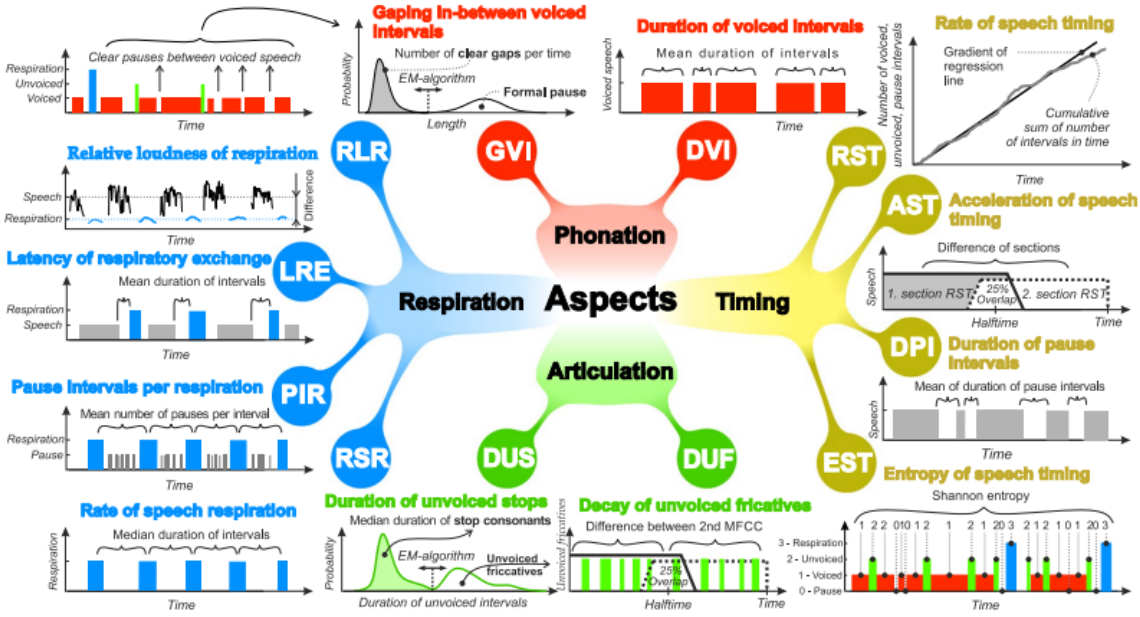


Figure 2: Resumen de las variables utilizadas

(En el dataset, éstas variables se repiten, ya que se han realizado dos mediciones, una durante la lectura de un texto fonéticamente equilibrado y otra durante un monólogo de unos 90 segundos. Las variables correspondientes al monólogo tienen la terminación "-M")

EST: (numérica) Mide la heterogeneidad en el habla en términos de la ocurrencia de intérvalos de sonoridad, insonoridad, pausas y respiraciraciones. (este valor se haciendo uso de la entropia de Shannon)
RST: (numérica) Ratio de la sincronización del habla.
AST: (numérica) Mide la aceleración o deceleración en la velocidad del habla.
DPI: (numérica) Mide la media de la duración de los intérvalos de pausas de cada paciente.
DVI: (numérica) Mide la duración media de los intérvalos de voz.
GVI: (numérica) Separación media entre intérvalos de voz.
DUS: (numérica) Duración media de las consonantes oclusivas.
DUF: (numérica) Mide la diferencia entre el 2º Coeficiente Cepstral en las Frecuencias de Mel.
RLR: (numérica) Sonoridad relativa de la respiración
PIR: (numérica) Intervalos de pausa por respiración
RSR: (numérica) Tasa de respiración del habla.
LRE: (numérica)Latencia del intercambio respiratorio.

Dataset Secundario para Series Temporales

Por completitud respecto al temario de la asignatura, tomamos un segundo dataset con variables evaluadas a lo largo del tiempo para trabajar Series Temporales.

Éste segundo dataset tiene como título "Health Indicator - Parkinson's Disease - Age-Standardized Incidence Rate" y lo hemos obtenido a través de la web del Ministerio de Sanidad de Alberta, Canadá. Puede consultarse en el siguiente enlace: http://www.ahw.gov.ab.ca/IHDA_Retrieval/selectSubCategory.do.

En el dataset figuran diversas variables poblacionales relacionadas con la enfermedad de Parkinson tomadas a la población de Alberta entre los años 2004 y 2020. Cambiamos el nombre de las variables porque son largos y engorrosos. Además, el dataset está dividido en zonas, y subzonas y regiones más pequeñas sucesivamente; sin embargo, a nosotros simplemente nos interesará estudiar las zonas de: Alberta, en general, y las subzonas: South Zone, Calgary Zone, Central Zone, Edmonton Zone y North Zone.

Finalmente, las variables de las que consta el dataset son:

year: (numérica) Año en que se realiza la medición.
region: (categórica con 8 niveles) Región en la que se mide.
sex: (categórica con 3 niveles) sexo del habitante.
incidence_rate: (numérica) Tasa de incidencia (estandarizada por edad)
incident_cases: (numérica) Casos de Parkinson (para todas las edades)
pop_at_risk: (numérica) Población en riesgo de padecer Parkinson.
SE: (numérica): "standard error", error estándar.
SS: (numérica): "standard score", puntuación estándar.
ABr: (numérica): "Alberta Rate".

Limpieza de datos y valores perdidos:

En el dataset principal, tenemos una gran cantidad de variables categóricas, para trabajar con ellas de manera óptima primero las transformaremos a factores con sus respectivos niveles.

Una vez hecho esto, notamos que tenemos muchas variables en que aparecen guiones "-" por varias razones, y lo que haremos será transformar todos estos guiones en *NA's*, y seguidamente explicar el motivo para que éstos datos sean faltantes.

Las variables *history*, *onset* y *duration* tienen *NA's* a partir de la observación 81 ya que los pacientes 81-130 son los controles sanos y por tanto no tiene sentido tomar mediciones acerca de su enfermedad si no tienen ninguna. Similarmente, la variable *HY.scale* toma el valor *NA* a partir de la observación 31, ya que los pacientes 1-30 son los que padecen la enfermedad de Parkinson y la escala H&Y se utiliza para medir el avance de los síntomas de dicha enfermedad.

Finalmente, hay una gran concentración de *NA*'s en las columnas correspondientes a las pruebas X18 - X30 de la escala UPDRS III, y esto se debe a que las pruebas sólo se han realizado a los pacientes que padecen la enfermedad de Parkinson o trastorno de conducta del sueño REM, es decir, en las primeras 80 observaciones.

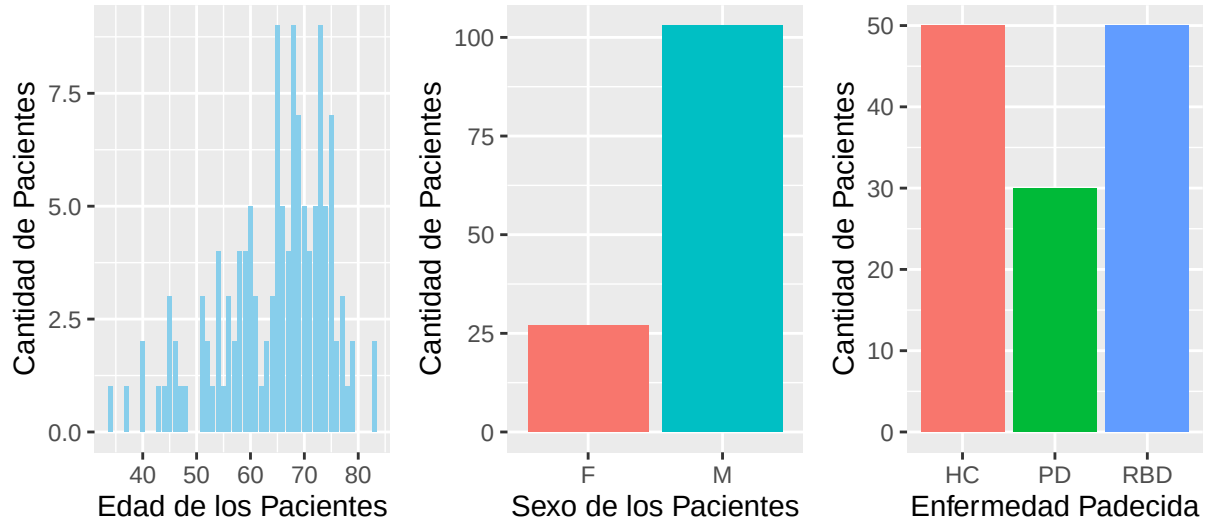
Con el objetivo de limpiar los datos lo máximo posible, eliminaremos algunas variables que no aportan información en absoluto. Durante la recogida de datos se tomaron datos de los pacientes que han resultado no ser útiles para la investigación. Las variables *antiparkinsonian*, *antipsychotic* y *levodopa*, que miden si el paciente se medica con ciertos medicamentos, tienen el mismo valor para las 130 observaciones (No, No y 0mg/día respectivamente). De modo que las eliminamos del conjunto de datos por ser supérfluas.

Otra cosa que nos facilitará trabajar con los datos será tener una variable que nos indique qué enfermedad padece un paciente (si es que padece alguna). Ésta información ya se encuentra en la variable *code*, que nos da el código del paciente. Sin embargo, los códigos incluyen también un número indentificativo para el paciente por lo que no podemos trabajar con ésta variable. Lo que haremos será utilizar la función *separate()* para dividir la variable *code* en dos variables nuevas, una que conservará el mismo nombre y únicamente

guardará el número asignado a cada paciente, y otra a la que llamaremos *disease* que indicará qué enfermedad padece el paciente; enfermedad de Parkinson (EP), trastorno de conducta del sueño REM (RBD) o si son controles sanos (HC).

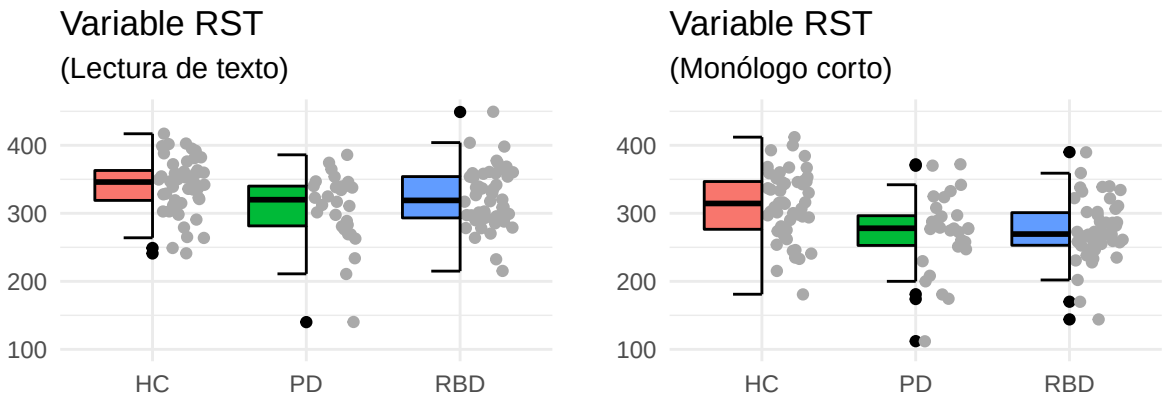
Fase Descriptiva: Visualización y Resúmen de variables.

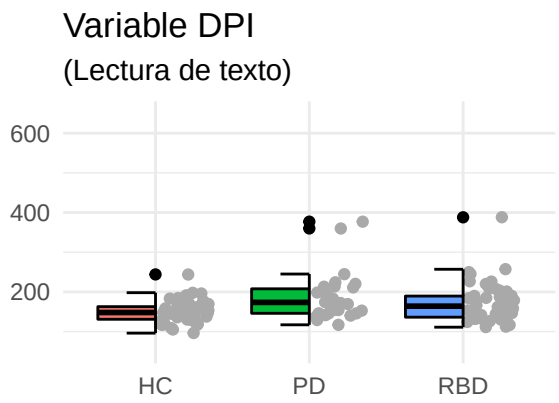
Vamos a representar gráficamente las tres variables *age, gender, disease* para poner en perspectiva algunas de las cualidades más genéricas de los pacientes evaluados:

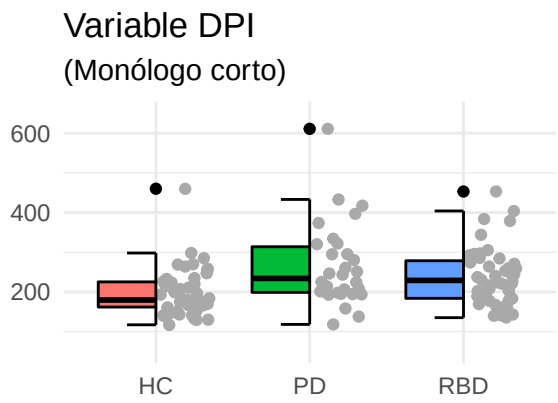


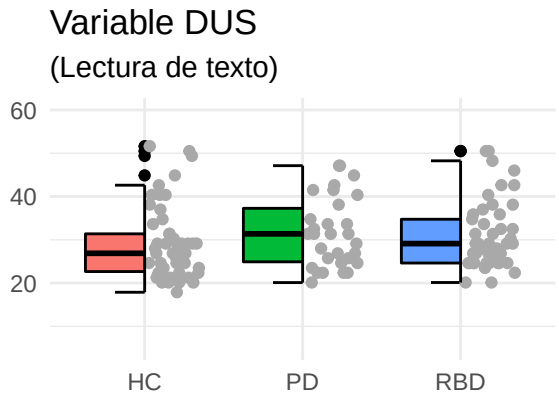
Por tanto vemos claramente que las observaciones se tratan de pacientes de edades comprendidas mayormente entre 50 y 80 años, donde predominan los hombres (aproximadamente una mujer por cada 4 hombres). Además confirmamos la distribución por enfermedad padecida mencionada en la contextualización del trabajo; 50 controles sanos (HC), 30 pacientes que padecen enfermedad de Parkinson (PD) y 50 que padecen trastorno de la conducta del sueño REM (RBD).

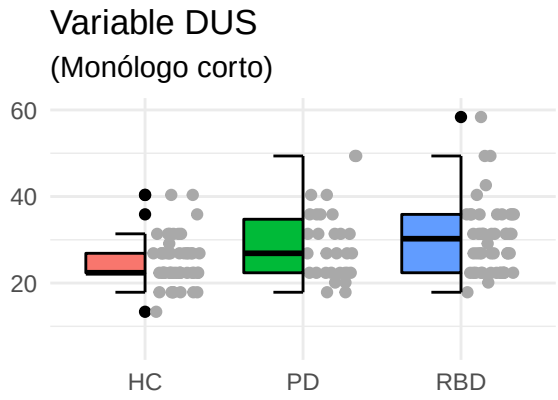
Ahora, lo que queremos hacer es, para cada una de las mediciones realizadas a los pacientes durante la lectura de un texto fonéticamente equilibrado y durante la exposición de un monólogo de 90 segundos, comparar los valores obtenidos en función de la variable *disease*, es decir, comparar según si los pacientes son enfermos de Parkinson, de trastorno de conducta del sueño REM o controles sanos. (Representamos gráficamente aquellas variables en que se se aprecie una diferencia fácil de observar).









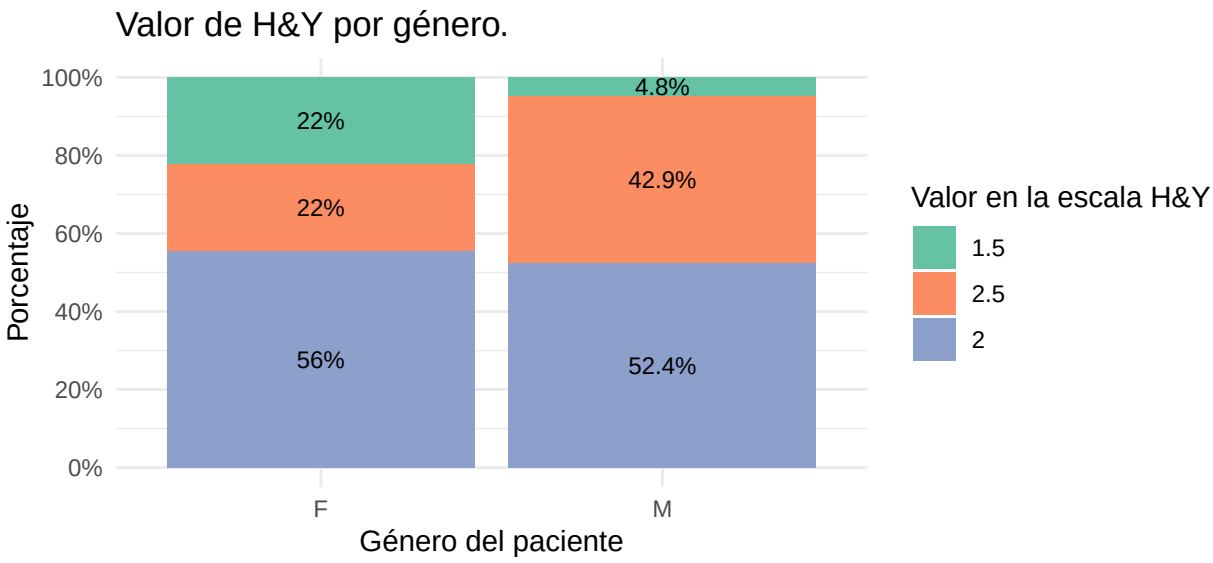


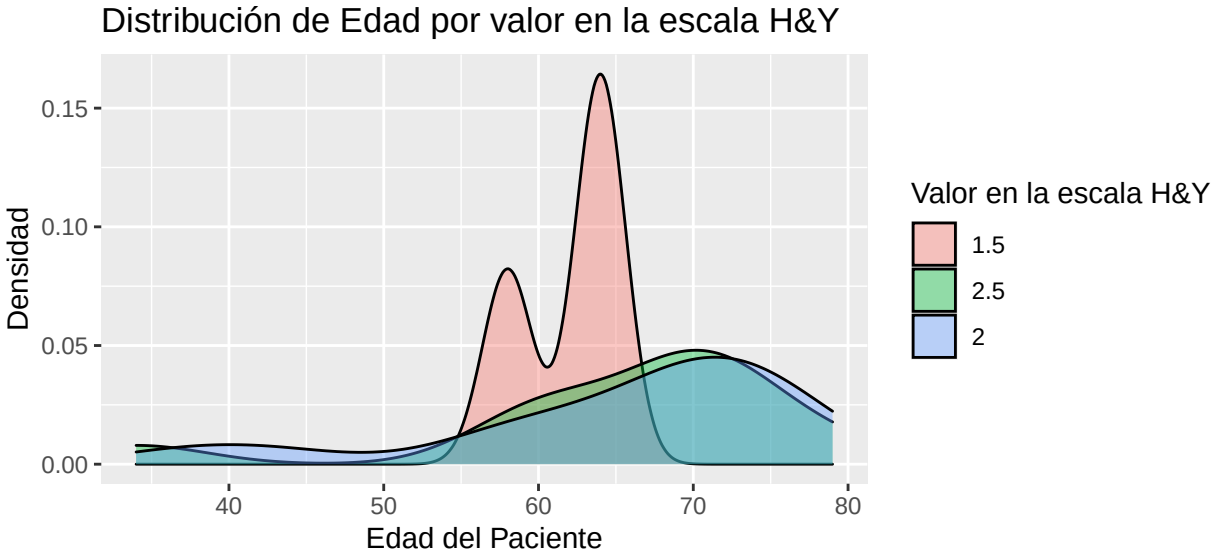
En los boxplots anteriores podemos ver que en una primera inspección no muy detallada de las variables, hay 3 en que se observa una diferencia entre los grupos. Empezamos analizando la variable RST. En el diagrama de cajas podemos ver que el valor medio de ésta variable es mayor para los pacientes sanos que para los enfermos (tanto de Párkinson como de Trastorno de la conducta del sueño REM). Esto ocurre tanto en las mediciones realizadas durante la lectura de un texto fonéticamente estandarizado y las realizadas durante un monólogo corto. Además, los outliers que se observan en los grupos PD y RBD se sitúan prácticamente en su totalidad, por debajo de la línea que marca $Q1 - 1.5 * IRQ$

En el la segunda figura analizamos la variable DPI. En éste caso destaca el hecho de que la media para el grupo de los pacientes sanos sea menor que la de los otros dos grupos de pacientes. Además, las observaciones de los grupos PD y RBD están más dispersas, cosa que aumenta el rango inter-cuartílico. Esto se aplica tanto a las mediciones de lectura como a las de monólogo.

Finalmente, en la tercera figura se representa la variable DUS. De nuevo, observamos que, en la lectura como en el monólogo, el valor medio de ésta variable es menor para los pacientes sanos que para los pacientes que padecen Párkinson o Trastorno de la conducta del sueño REM. No observamos ninguna diferencia muy notable en cuanto a la dispersión de los datos.

La variable *HY.scale* nos indica el grado de avance de los síntomas del Parkinson. Vamos a observar si los síntomas avanzan más rápido dependiendo de la edad o sexo del paciente:





Por lo tanto, en el primer gráfico vemos que, en lo que respecta al avance de los síntomas, el porcentaje de hombres con una escala de 1.5 en H&Y (síntomas leves) es menor que el de mujeres. A su vez, el porcentaje de hombres en escala 2.5 (síntomas más graves) es mayor. De ésta manera observamos que los síntomas de la enfermedad de Parkinson afectan de forma más grave a los hombres que a las mujeres.

Por otra parte, en el segundo gráfico vemos, para cada valor en la escala H&Y, la distribución de las edades de los pacientes afectados. Se aprecia que los pacientes que están en los valores 2 y 2.5 de la escala H&Y tienen una distribución de edades muy parecida. Sin embargo, observamos claramente que los pacientes que están en el valor 1.5, es decir, los de síntomas más leves, tienen una edad significativamente más baja que los otros dos grupos. Así, constatamos que los pacientes más jóvenes presentan una sintomatología más leve.

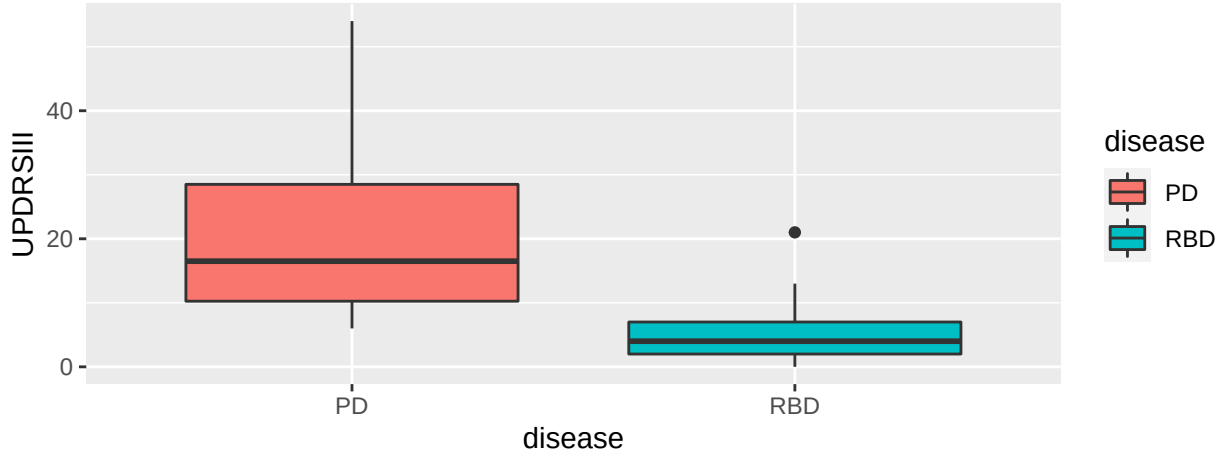
Tenemos un conjunto de variables $(X18, \ldots, X31)$ que se corresponden con las diferentes pruebas, la suma de cuyos resultados dará lugar a la variable genérica $UPDRSIII$. La puntuación de cada paciente en ésta escala consiste de la suma de las puntuaciones del paciente en cada una de las pruebas $X18, \ldots, X31$. Queremos ver, a grandes rasgos, el aspecto que tienen éstas variables. Con éste fin mostramos el gráfico siguiente:

Vemos que en su mayoría las puntuaciones de los pacientes en cada prueba se distribuyen decrecientemente, en general, el mayor número de pacientes es valorado con un 0 y el número de pacientes decrece a medida que aumenta la puntuación. Esta distribución se debe a que, como el objetivo del estudio es determinar biomarcadores prematuros de la enfermedad del Parkinson, los pacientes analizados se caracterizan por no estar en un estado avanzado del desarrollo de los síntomas de la enfermedad.

Los resultados de las pruebas son similares entre ellos por lo que únicamente analizaremos en profundidad la

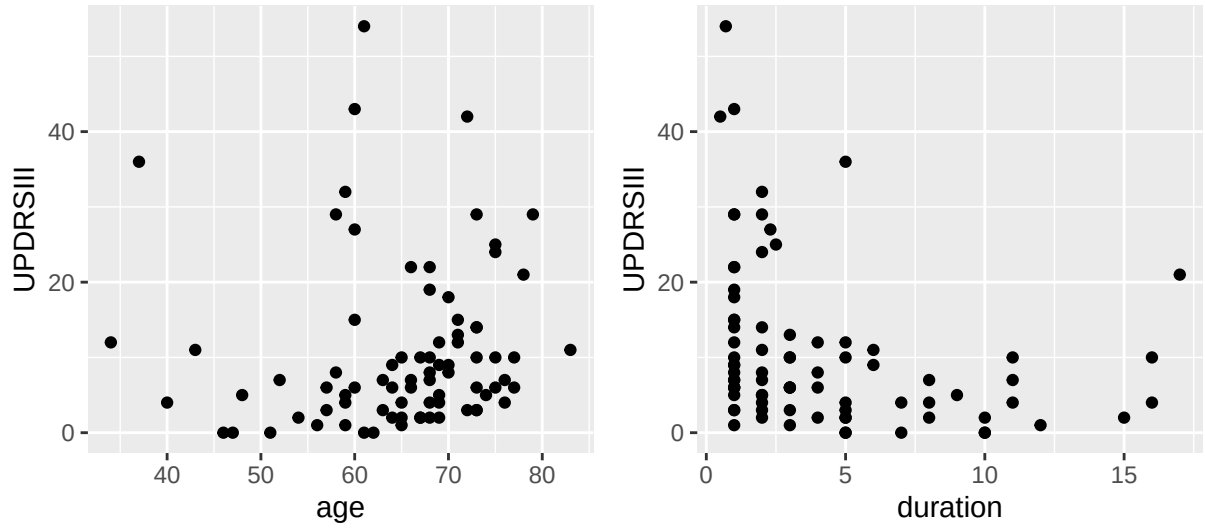
variable (????????), que consideramos de interés.

Queremos comprobar también si las personas que padecen Parkinson, tienen una puntuación más alta en la escala de UPDRSIII que aquellas que padecen trastorno de la conducta del sueño REM.



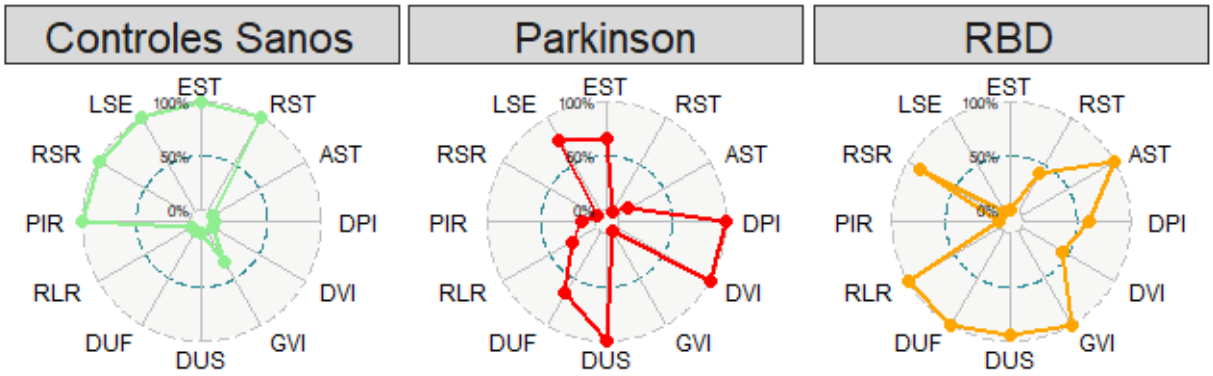
Y como intuíamos, los pacientes con Parkinson tienen valores más altos en este test que los pacientes que padecen transtorno de sueño (basta fijarse en que la media de los primeros queda por encima del valor límite (?) del otro grupo). Además, también podemos fijarnos en que los pacientes con Párkinson tienen una dispersión considerable, mucho mayor que en el grupo de pacientes con transtorno de sueño.

Podríamos pensar que la causa de la alta dispersión de los resultados del test en pacientes de Parkinson se deba a una diferencia muy marcada en la edad de los pacientes seleccionados. Por tanto podemos comprobar cómo se han distribuido los resultados en función de la edad de los pacientes, para ver si hay una tendencia marcada.



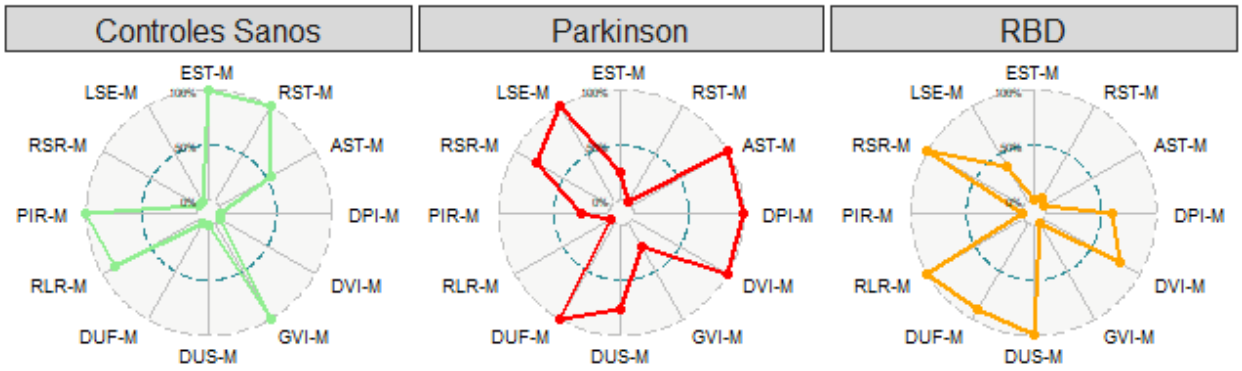
Pero como podemos ver en el gráfico de la izquierda, no parece que exista tal tendencia. Además, en el gráfico de la derecha hemos mirado a ver si el resultado del test se ve afectado por el tiempo total que llevan los pacientes con dicha enfermedad y tampoco parece existir ninguna tendencia ("no se si deberíamos explicar de onde sale la variable Dif pero la había creado solo para esto").

A continuación, realizaremos unos diagramas “de telaraña” o “de radar” que nos permitirán ver de forma clara el perfil de cada grupo de pacientes respecto a las pruebas del habla a las que se han sometido. En primer lugar tomaremos los datos recogidos durante la lectura de un texto fonéticamente estandarizado y siguidamente lo haremos para los datos recogidos durante un monólogo breve.



En la figura anterior podemos observar que, en lo que respecta a ésta prueba, los Controles Sanos se caracterizan por valores altos en las variables PIR y RST, y un valor particularmente bajo de DUS. Por otro lado, los pacientes que padecen la enfermedad de Parkinson se distinguen del resto por un valor alto en la variable DVI y destacan también por un valor más bajo que el del resto de grupos de RSR. Finalmente, los pacientes que padecen trastorno de la conducta del sueño REM, tienen valores notablemente altos de AST, GVI y RLR, mientras que, a diferencia de los otros dos grupos, el valor de EST y LSE se mantiene muy bajo.

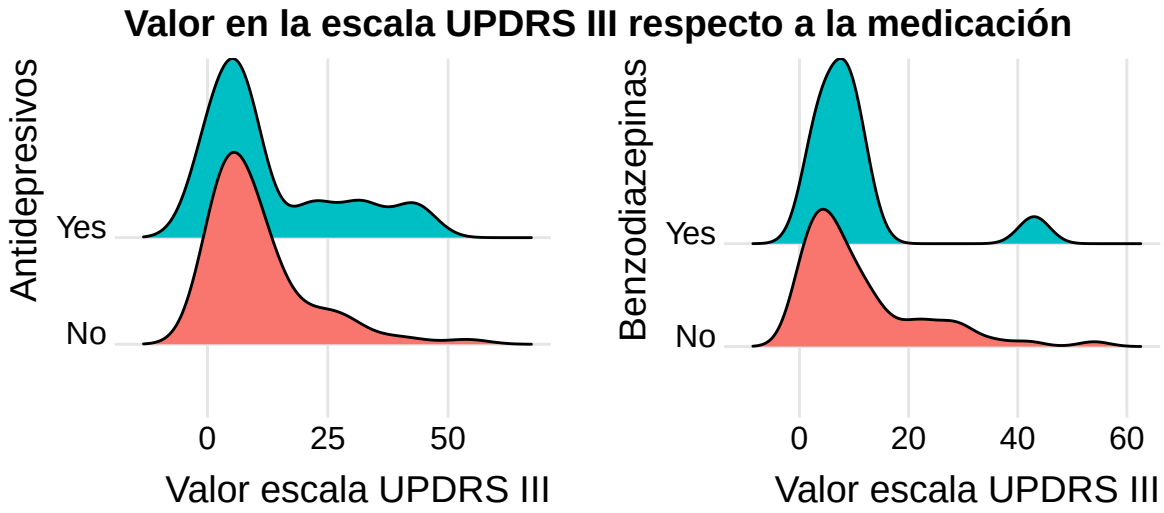
A continuación tomamos los datos recogidos durante el monólogo breve:



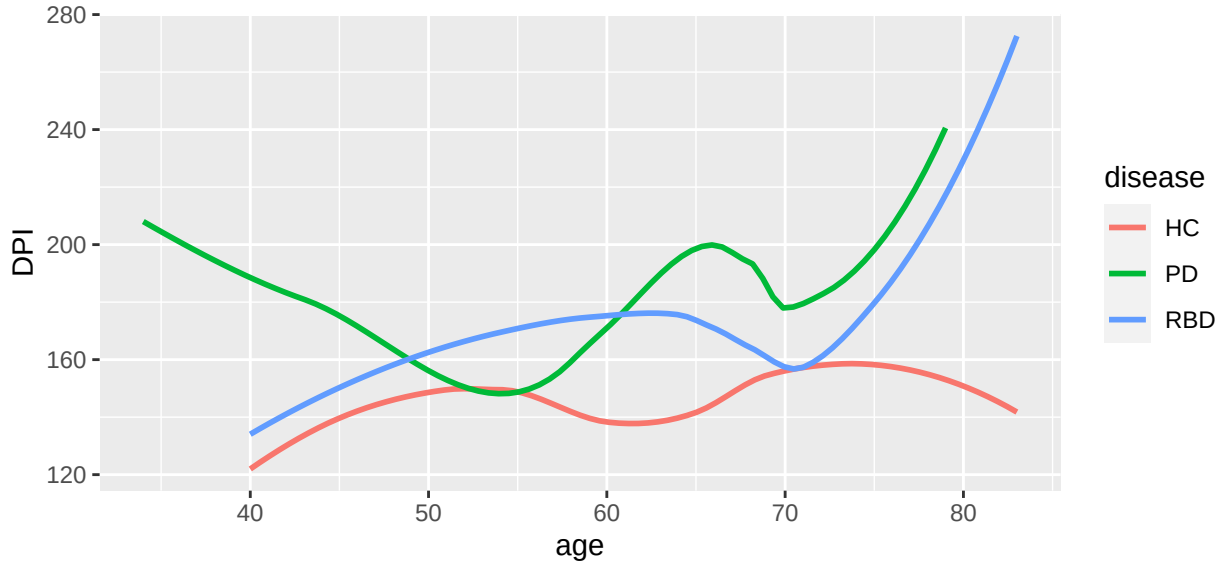
Interpretemos los gráficos en este caso: Los pacientes de control sanos tienen, al igual que en el gráfico anterior (de la lectura de un texto) valores notablemente altos de EST, RST y PIR. Además, observamos valor alto de la variable GVI así como un valor bajo de DUF y DVI. Por otro lado, los pacientes de Parkinson se caracteriazan porque el valor de LSE resulta especialmente alto y el de RLR especialmente bajo. Finalmente, los pacientes con trastorno de la conducta del sueño REM destacan en ésta ocasión por su valor de AST, que en comparación con el de los otros dos grupos es muy bajo.

De ésta forma hemos encontrado cómo las variables recogidas en éstas dos pruebas pueden caracterizar a los diferentes grupos de pacientes y éstas relaciones podrían tenerse en cuenta a la hora de evaluar el riesgo de un paciente de padecer Parkinson.

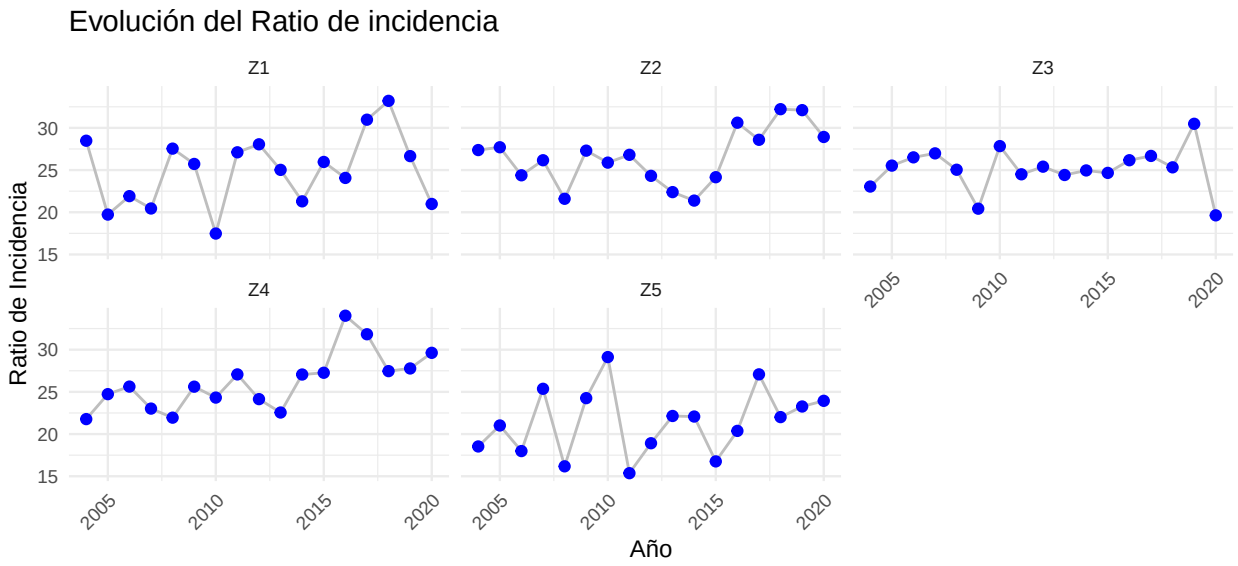
Estudiaremos también si aquellos pacientes toman algún tipo de medicamento presentan un menor valor en la escala UPDRS III, es decir [. . .]

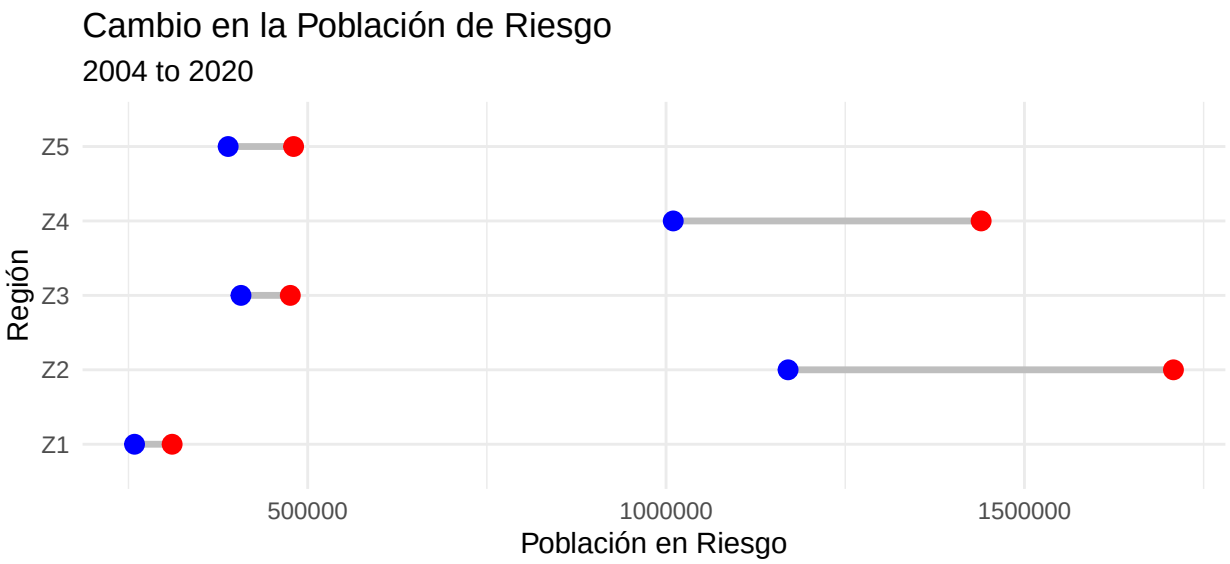


[...] (no sé si sirve de algo la verdad)



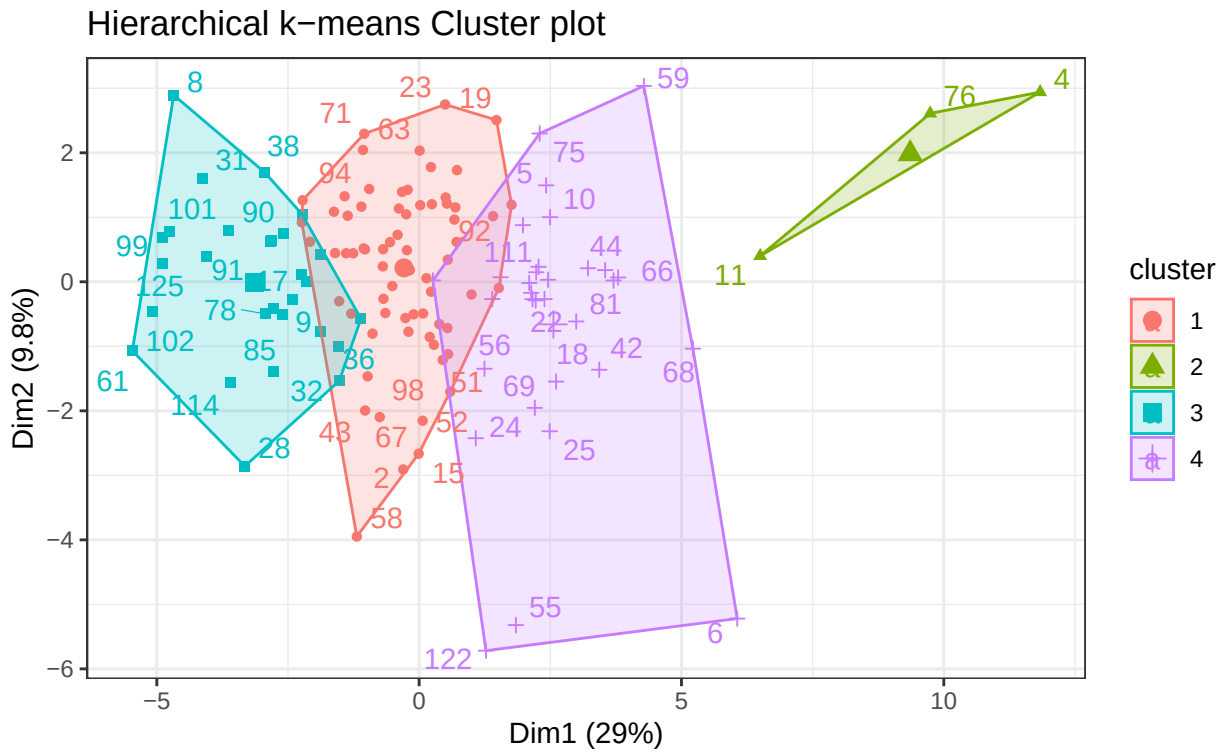
No se que quiere decir pero parece que solo sube pa los enfermitos

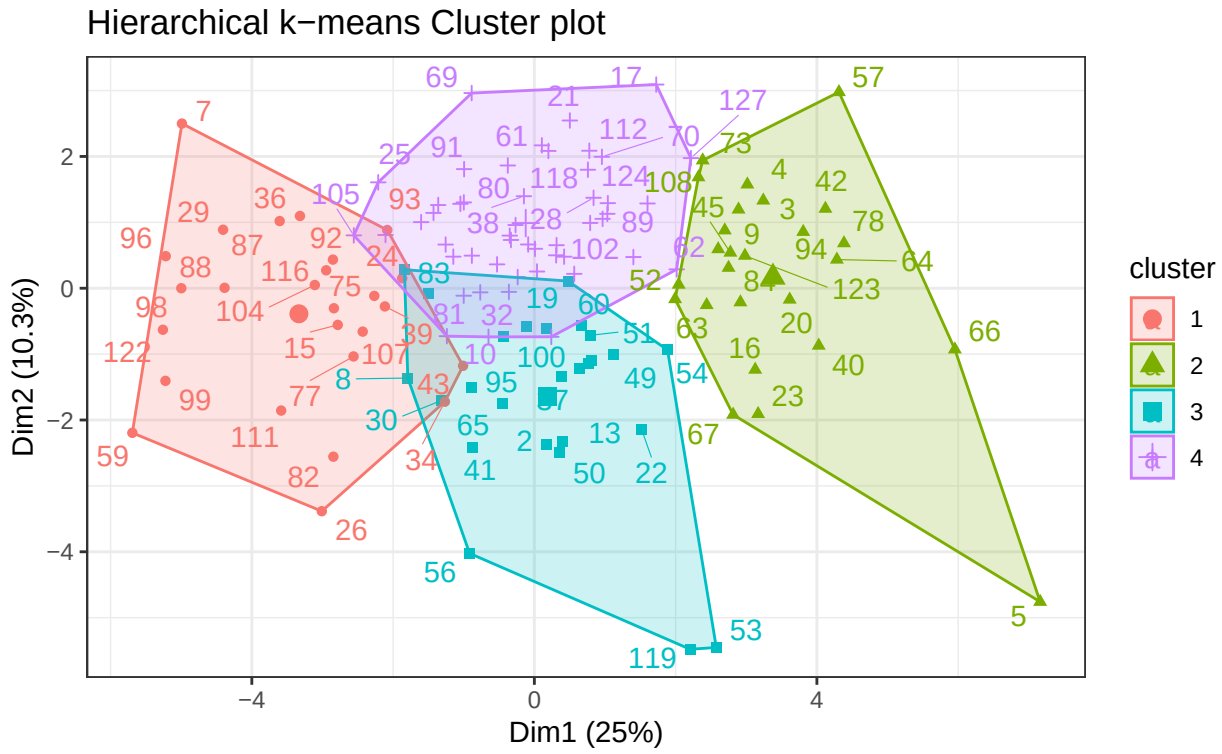

Evolución del Ratio de incidencia
Z1
Z2
Z3
Z4
Z5
Ratio de Incidencia
30
25
20
15
2005
2010
2015
2020
Año

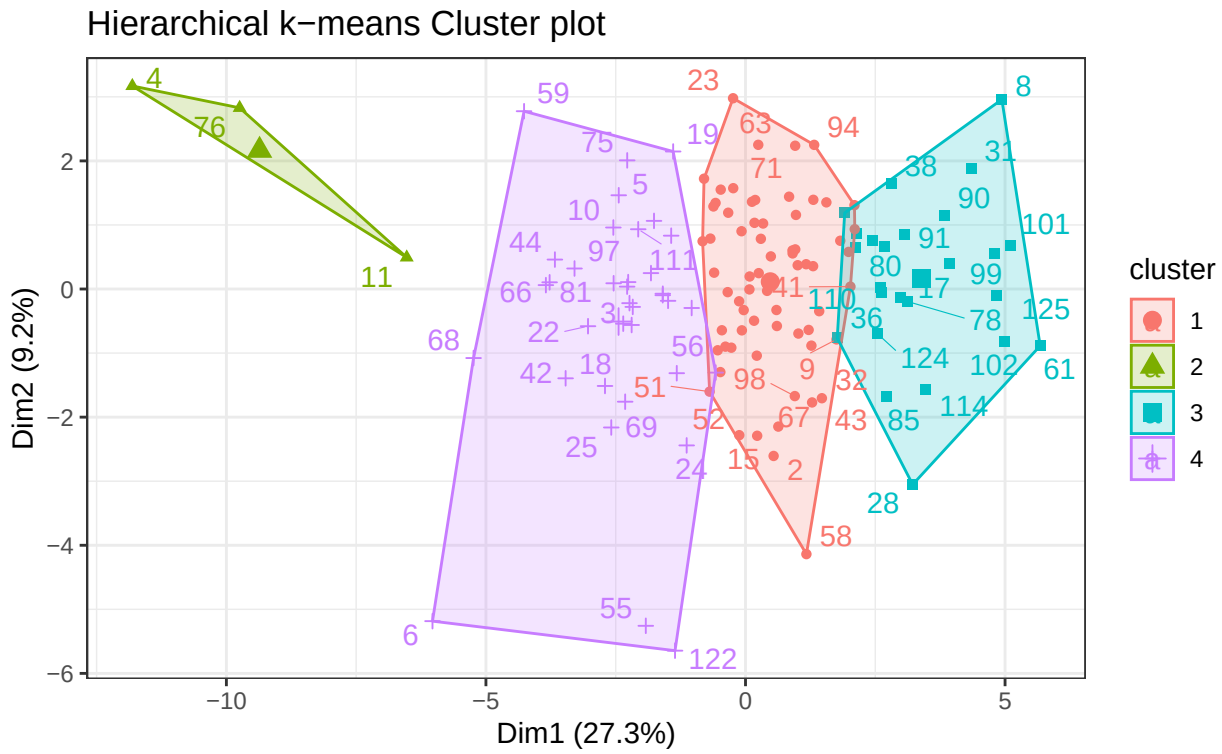

Cambio en la Población de Riesgo
2004 to 2020
Región
Z5
Z4
Z3
Z2
Z1
500000
1000000
1500000
Población en Riesgo

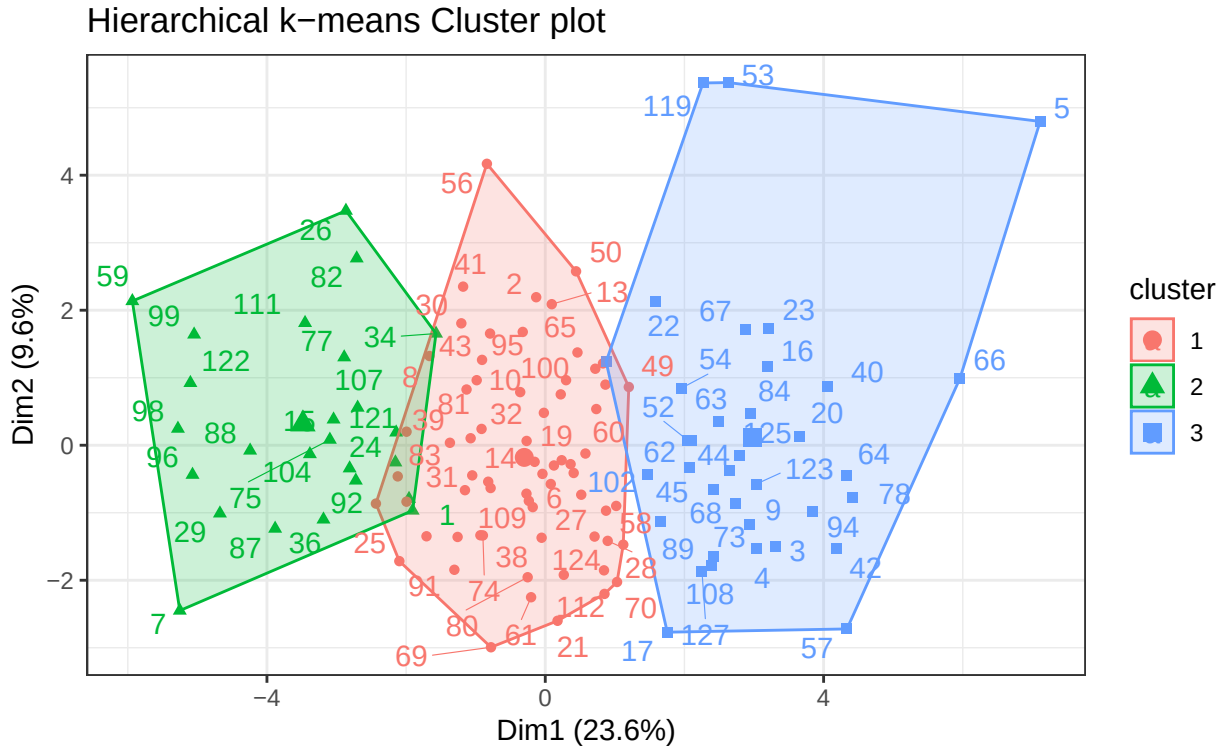
Clustering en "grupos de riesgo":

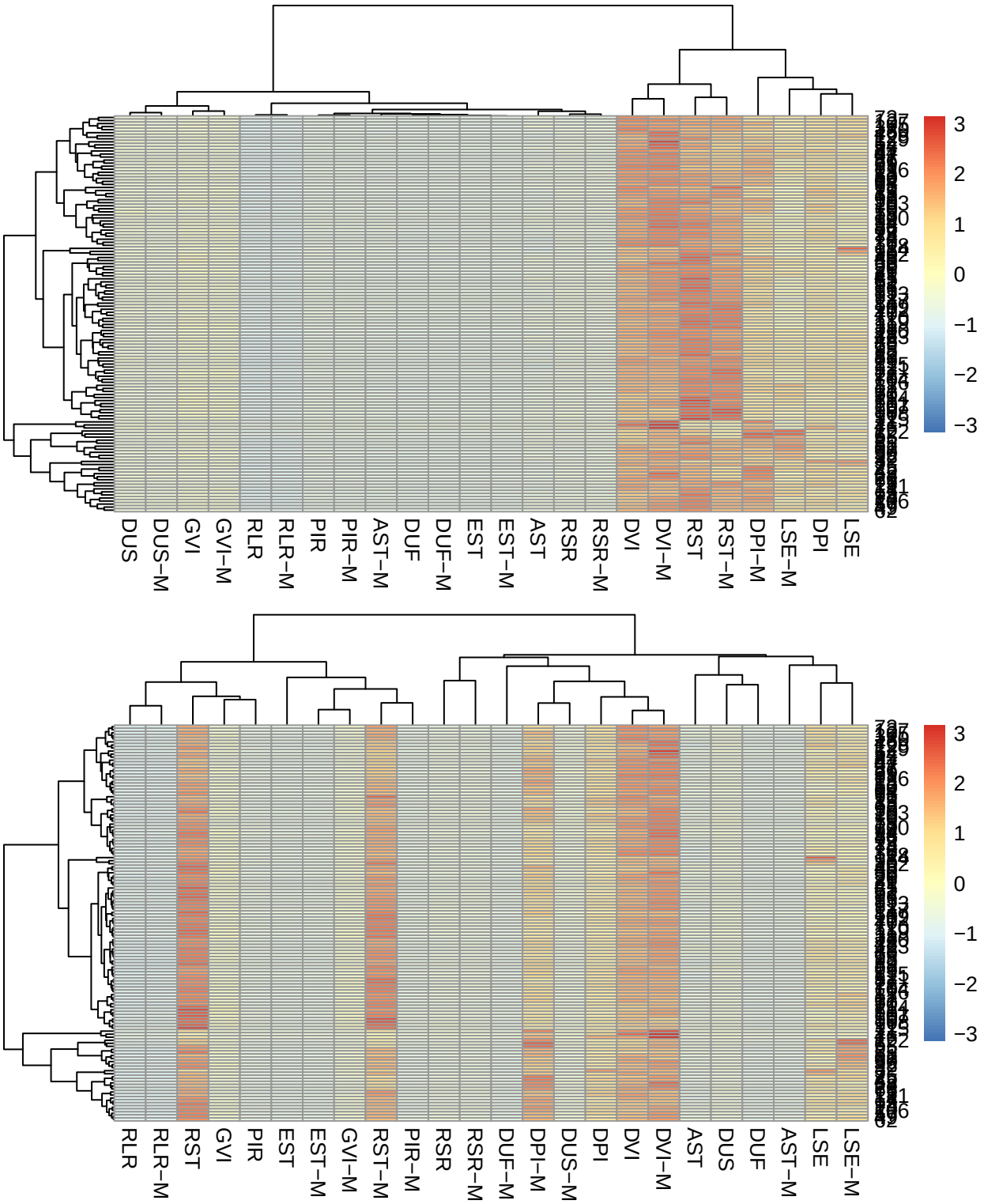
[NO HAY MANERA DE QUE SALGA ALGO DECENTE]

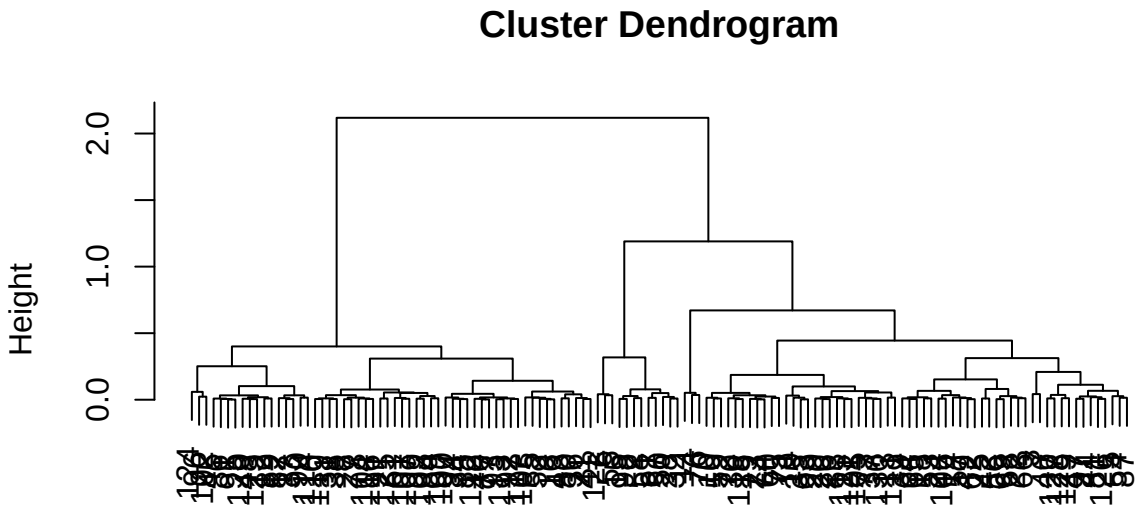


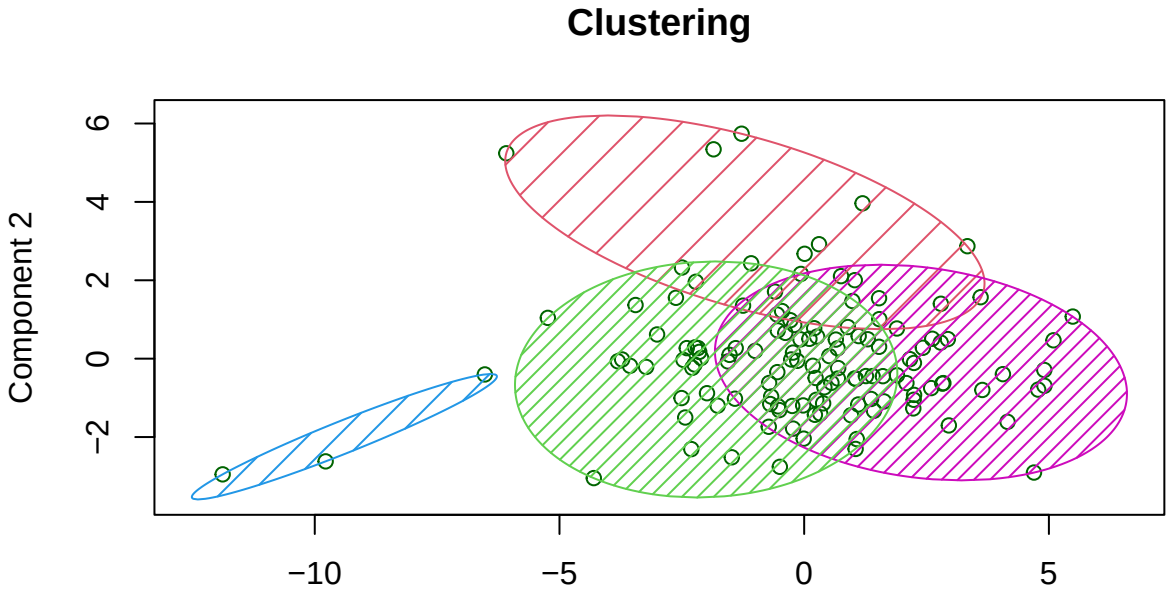



Hierarchical k−means Cluster plot
Dim2 (9.2%)
Dim1 (27.3%)
cluster
1
2
3
4


Hierarchical k−means Cluster plot
Dim2 (9.6%)
Dim1 (23.6%)
cluster
1
2
3


DUS
DUS−M
GVI
GVI−M
RLR
RLR−M
PIR
PIR−M
AST−M
DUF
DUF−M
EST
EST−M
AST
RSR
RSR−M
DVI
DVI−M
RST
RST−M
DPI−M
LSE−M
DPI
LSE
3
2
1
0
−1
−2
−3
RLR
RLR−M
RST
GVI
PIR
EST
EST−M
GVI−M
RST−M
PIR−M
RSR
RSR−M
DUF−M
DPI−M
DUS−M
DPI
DVI
DVI−M
AST
DUS
DUF
AST−M
LSE
LSE−M
3
2
1
0
−1
−2
−3


Cluster Dendrogram
Height
0.0
1.0
2.0
dist2(as_tibble(data_speech))
hclust (*, "ward.D")


Clustering
Component 2
−2
0
2
4
6
−10
−5
0
5
Component 1
These two components explain 38.83 % of the point variability.