Clase 2 - Practico 2

Felix Rojo Lapalma

16 de Mayo de 2018

Practico 2: Entregar un Rmd donde se:

- Elija un dataset clasificado de su preferencia y area (domain expertise), aplique un metodo de clustering y/o mixtura de Gaussianas en el mismo.
- Investigue los resultados en el meta parametro K numero de cumulos e investigue posibles procesos de seleccion del mismo.
- Elabore un resumen, y selecione un mejor valor segun el/los criterios aplicados, discuta el significado de los cumulos encontrados.
- Comente la influencia de la normalizacion de los datos en los resultados del clustering.

Empecemos

Sobre el data-set

El mismo comprende una serie de mediciones sobre ciertos instrumentos pertenecientes a una plataforma satelital. En las mismas se mezclan voltajes, temperaturas,corrientes y demas variables de interes para la plataforma. Las mismas corresponden a un valor medio diario comenzando en el Doy (day of year) 321 del año 2014. En el mismo no existen variables categoricas, por lo que en virtud del practico las generaremos. Adicionalmente crearemos algunos features adicionales. Las variables que consideraremos de interes particular para el analisis son “AXX0044” y “AXX0045”.

Cargamos el data set

Algunos comentarios respecto a las medidas adicionales generadas (no seran utilizadas todas necesariamente):

-box_data[‘year_doy’]: corresponde a la fecha en formato ####.### donde la parte decimal corresponde a la fraccion del año (se considero que un año tiene 366 dias lo cual no es correcto -salvo bisiestos- pero es una aproximación que utilizamos por simplicidad).

-box_data[‘TM_diff’]: corresponde al diff de una dada variable donde se incluye el punto inicial repetido para tener un vector del mismo tamaño de entrada, a la salida.

-box_data[‘season_south’]: correponde a la variable categorica season (hemisferio sur - basado en el doy de entrada).

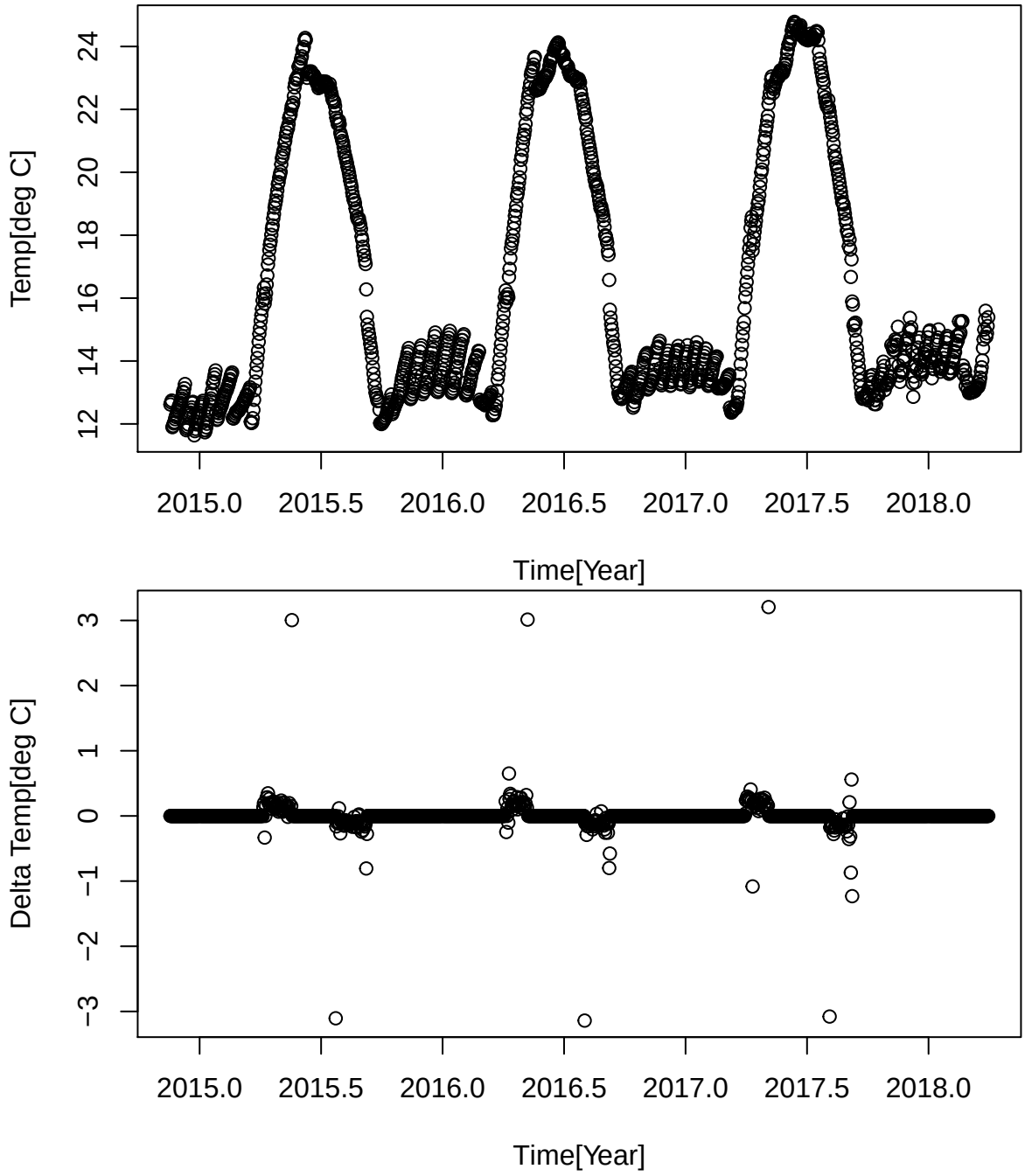
-box_data[‘season_north’]: correponde a la variable categorica season (hemisferio norte - basado en el doy de entrada).

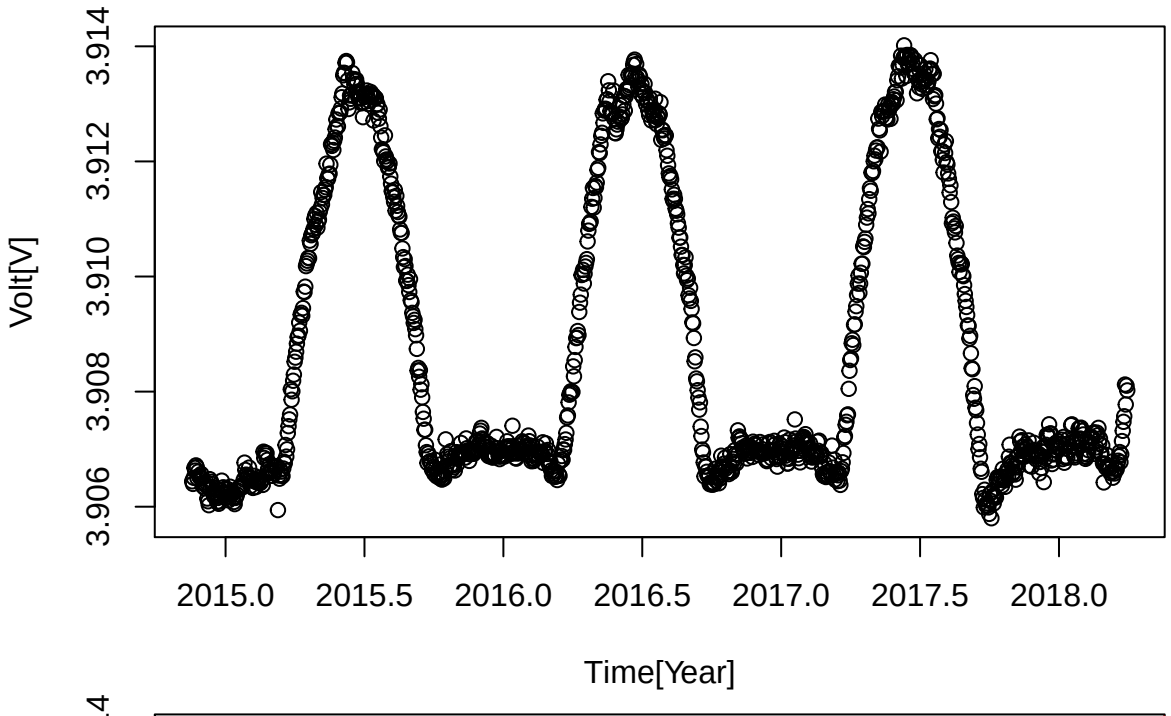
-box_data[‘season_south_num’]: correponde a la variable numerica season (hemisferio sur - basado en el doy de entrada).

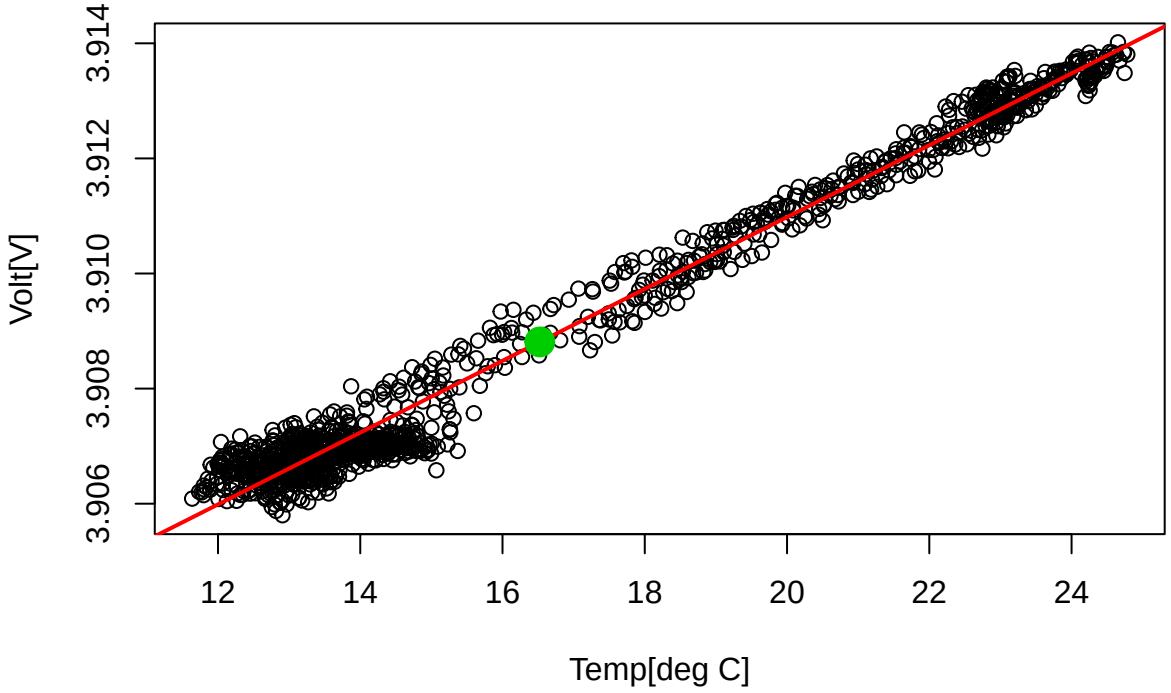
-box_data[‘season_north_num’] correponde a la variable numerica season (hemisferio norte - basado en el doy de entrada).

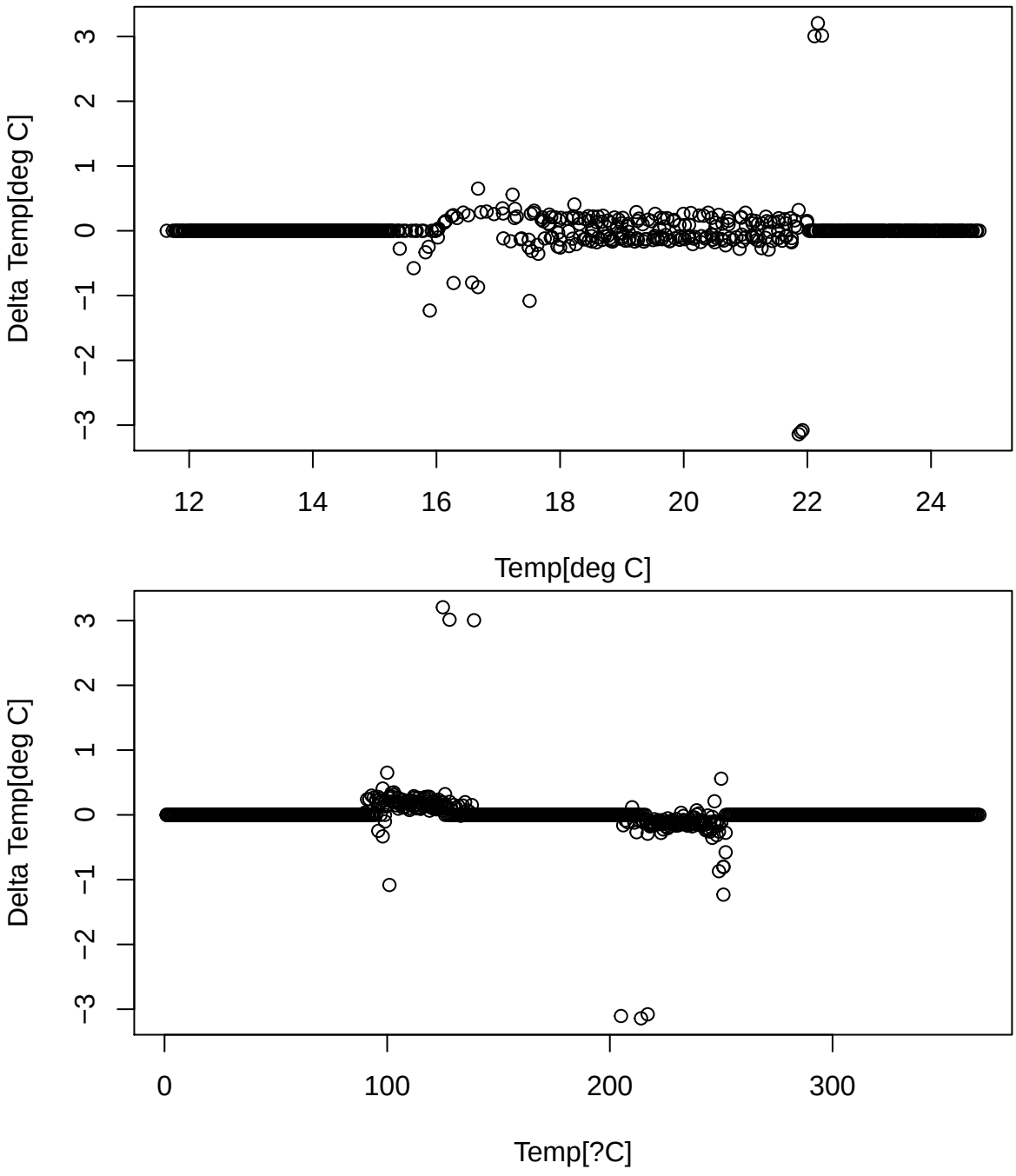
Inspección preliminar

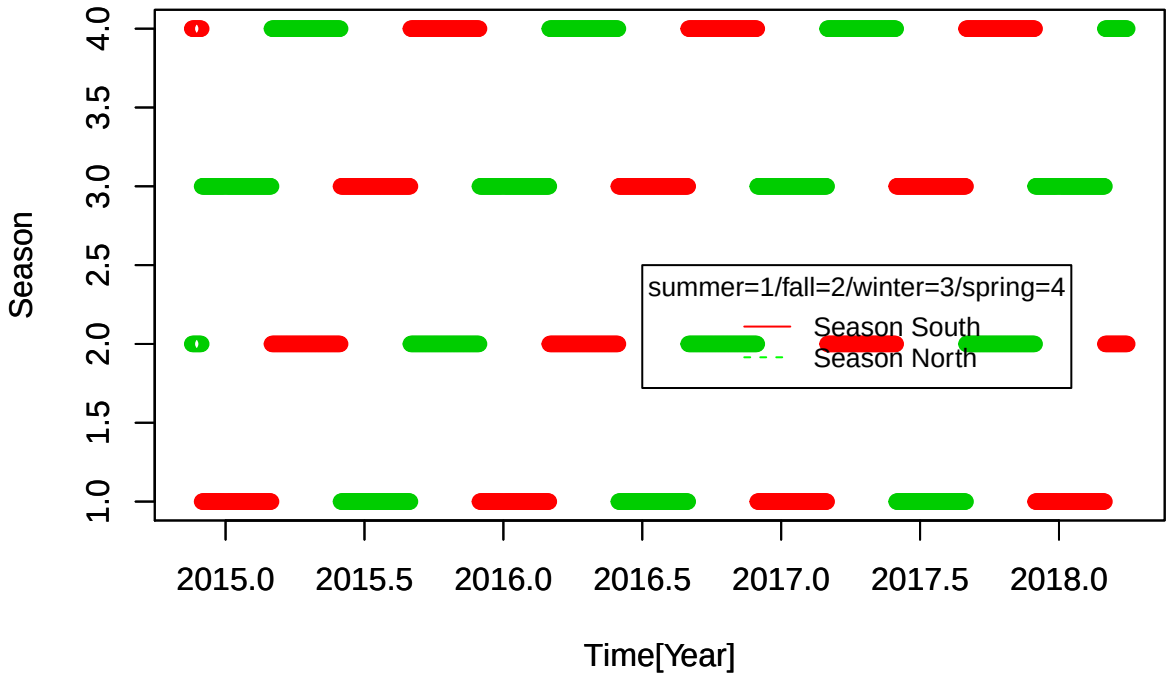
Hagamos algunos plots temporales para ver el comportamiento de las variables en cuestion.




Volt[V]
3.906
3.908
3.910
3.912
3.914
2015.0
2015.5
2016.0
2016.5
2017.0
2017.5
2018.0
Time[Year]


Volt[V]
3.906
3.908
3.910
3.912
3.914
12
14
16
18
20
22
24
Temp[deg C]


Delta Temp[deg C]
3
2
1
0
−1
−2
−3
12
14
16
18
20
22
24
Temp[deg C]
Delta Temp[deg C]
3
2
1
0
−1
−2
−3
0
100
200
300
Temp[?C]



Por la distribución de los valores parece ser que el mencionado instrumento "vive con las estaciones del hemisferio norte" (basicamente es su ubicacion sobre la plataforma). Asimismo observamos que se replica la forma funcional de la temperatura con la correspondiente al voltaje. Inclusive la relación entre ellas es quasi-lineal (basados en el grafico 4).

Enfoque K-means

El enfoque en terminos de kmeans y su parametro k tendra que ver con lo que querramos analizar. Particularmente estamos interesados en ver que se puede decir de los valores alcanzados por la temperatura, el voltage asociado y su estacionalidad. Para ellos construimos clusters de tamaño 2, 3 y 4.

```
set.seed(20)
AXX0045AXX0044_cl_k2 <- kmeans(box_data_n[, c('AXX0045','AXX0044')], 2, nstart = 20)
AXX0045AXX0044_cl_k3 <- kmeans(box_data_n[, c('AXX0045','AXX0044')], 3, nstart = 20)
AXX0045AXX0044_cl_k4 <- kmeans(box_data_n[, c('AXX0045','AXX0044')], 4, nstart = 20)
```

```
# Not Scaled
AXX0045AXX0044_cl_k2_Nnorm <- kmeans(box_data[, c('AXX0045','AXX0044')], 2, nstart = 20)
AXX0045AXX0044_cl_k3_Nnorm <- kmeans(box_data[, c('AXX0045','AXX0044')], 3, nstart = 20)
AXX0045AXX0044_cl_k4_Nnorm <- kmeans(box_data[, c('AXX0045','AXX0044')], 4, nstart = 20)
```

```
## [1] "Scaled"
```

```
##
##     fall spring summer winter
##   1  279    153      0    365
##   2    7    150    276      0
```

```
##
##     fall spring summer winter
##   1   23     90     68      0
##   2  263    127      0    365
##   3    0     86    208      0
```

```
##
##     fall spring summer winter
##   1    0     65    182      0
##   2  258    120      0    365
##   3   28     60     26      0
##   4    0     58     68      0
```

```
## [1] "Not Scaled"
```

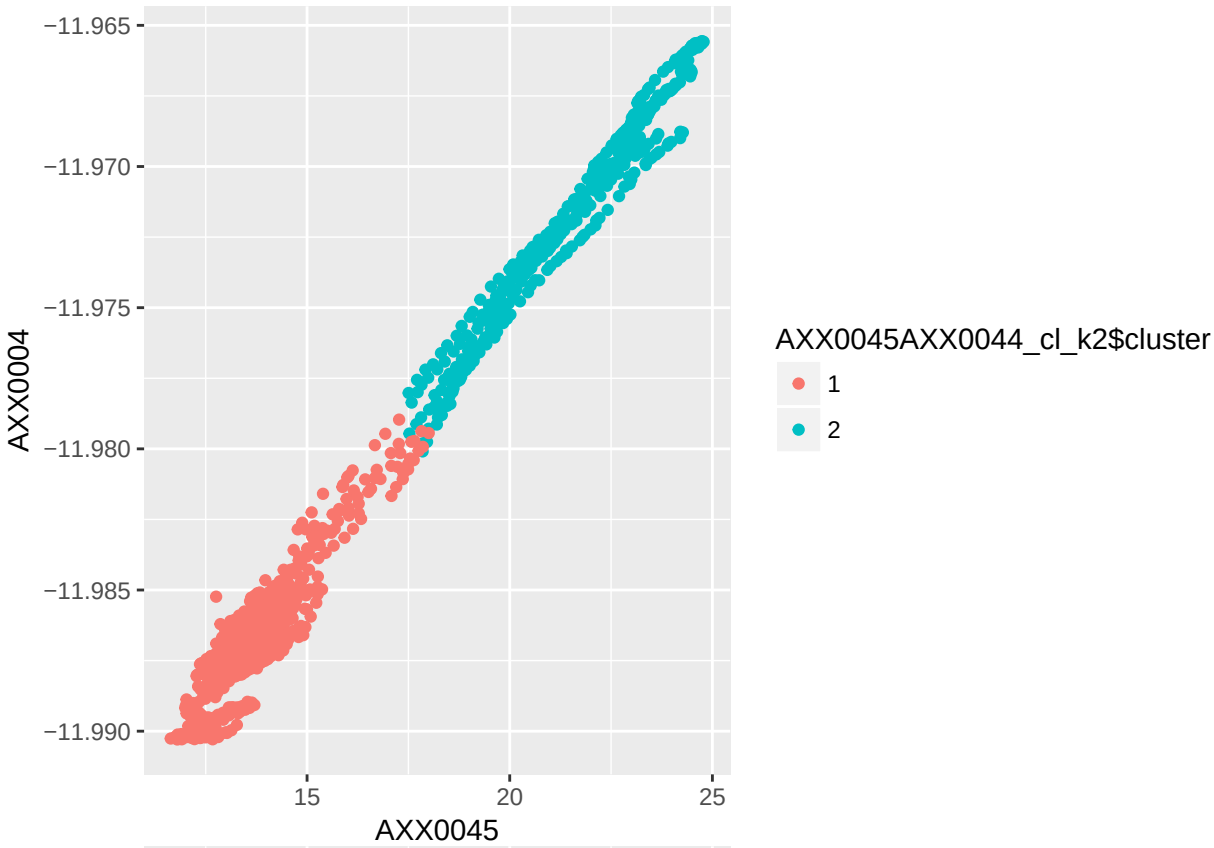
```
##
##     fall spring summer winter
##   1  276    158      0    365
##   2   10    145    276      0
```

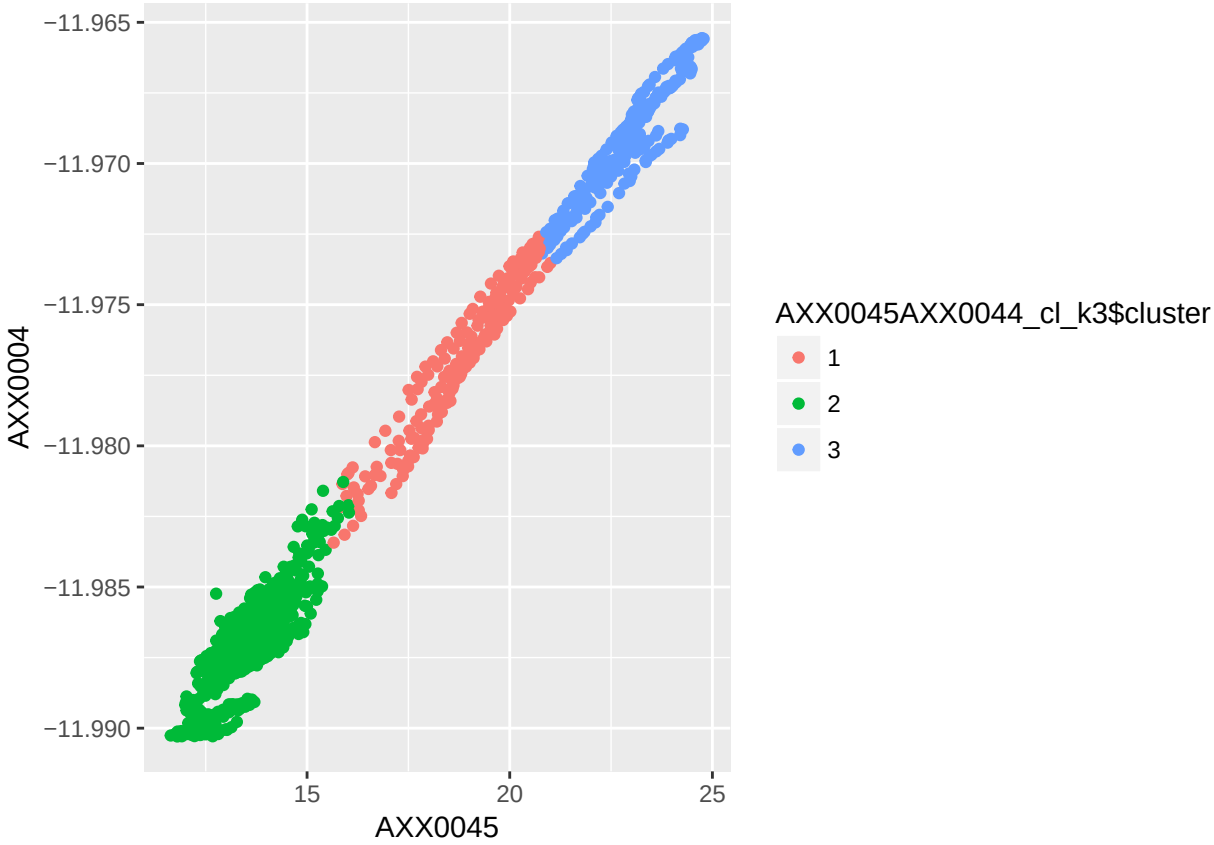
```
##
##     fall spring summer winter
##   1    0     86    203      0
##   2  263    137      0    365
##   3   23     80     73      0
```

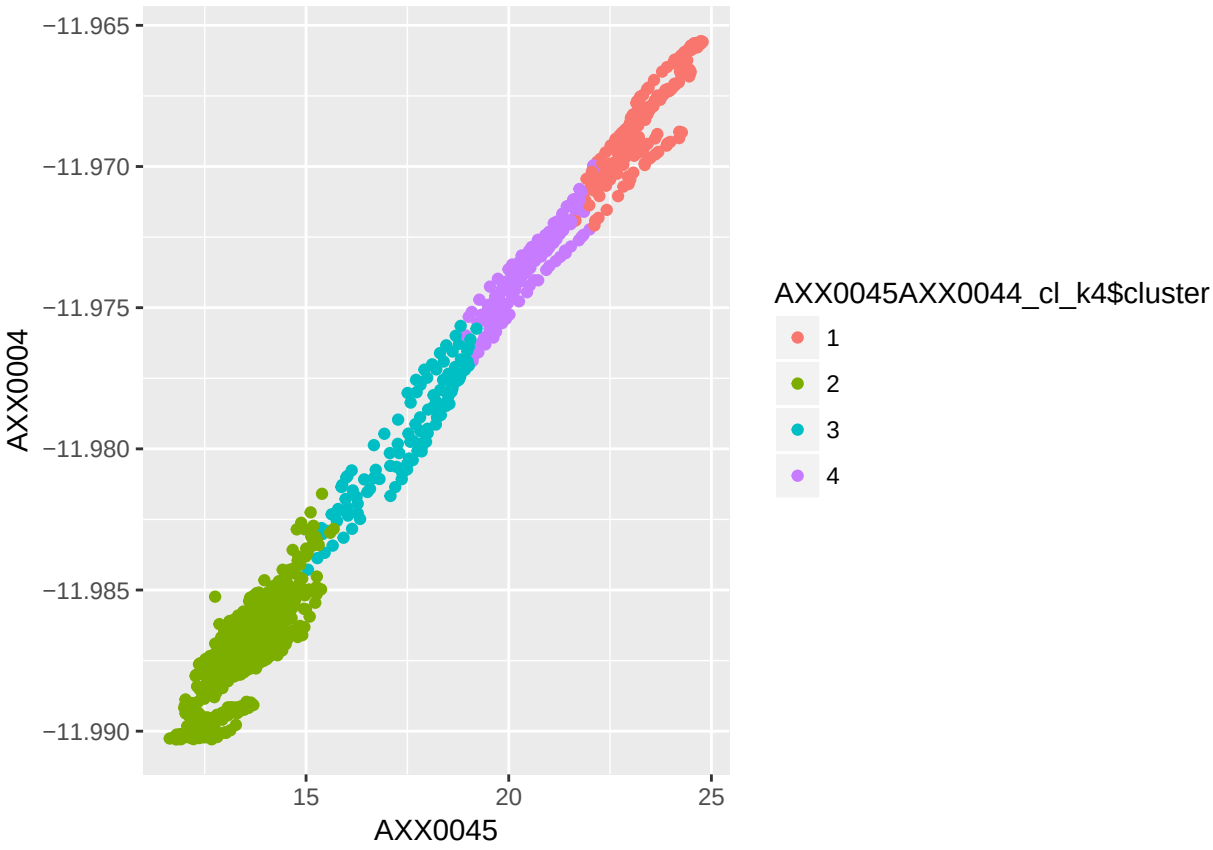
```
##
##     fall spring summer winter
##   1   20     74     74      0
##   2  183     90      0    226
##   3   83     55      0    139
##   4    0     84    202      0
```

De las segmentaciones para k=2,3,4 vemos que esencialmente uno de los clusters se mantiene sin cambios y los restantes empiezan a pasarse los integrantes. A priori hubiera esperado que summer y winter se mantengan esencialmente fijos y la transferencia de informacion se hubiera dado entre fall y spring.

En el caso de usar las variables sin escalarlas, la distribucion alcanzada no es la misma (segun se aprecia numericamente a partir de las tablas - con exepcion de winter).

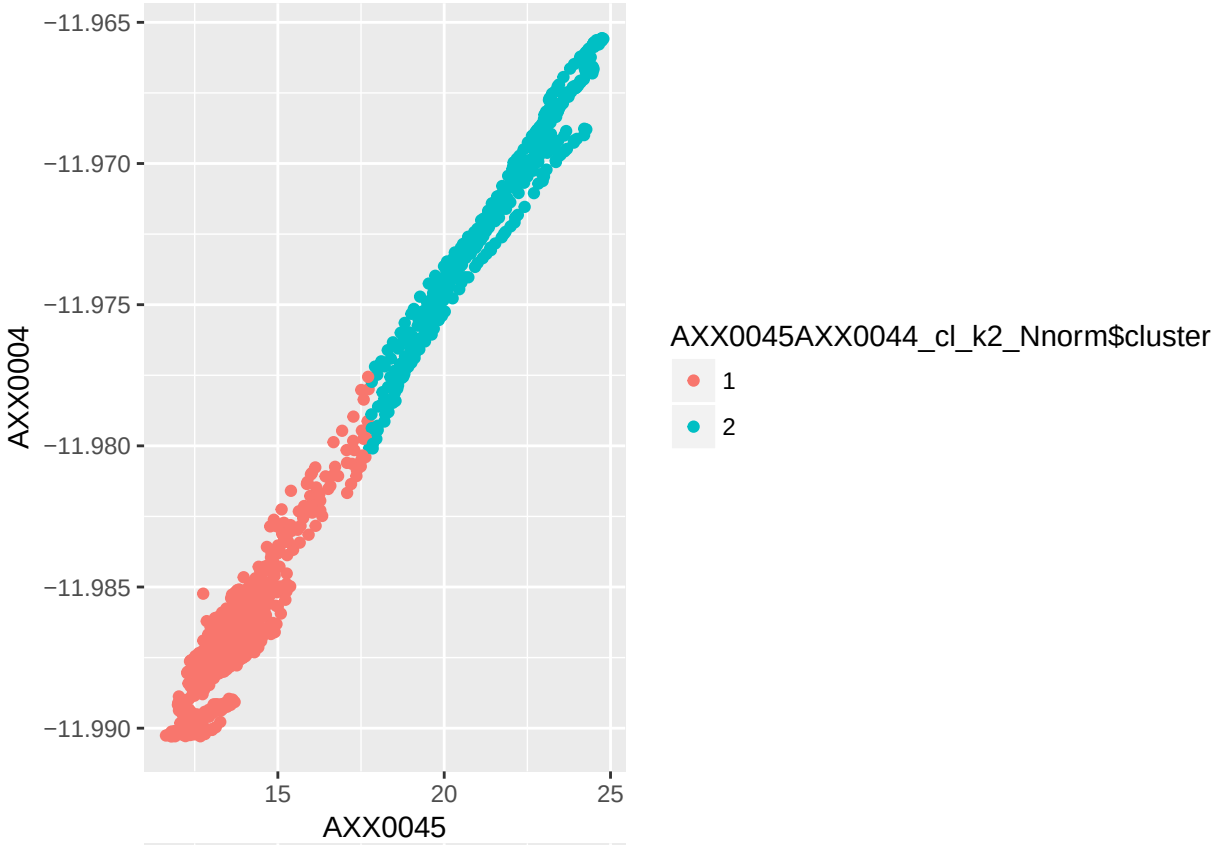

−11.965
−11.970
−11.975
−11.980
−11.985
−11.990
AXX0004
15
20
25
AXX0045
AXX0045AXX0044_cl_k2$cluster
1
2

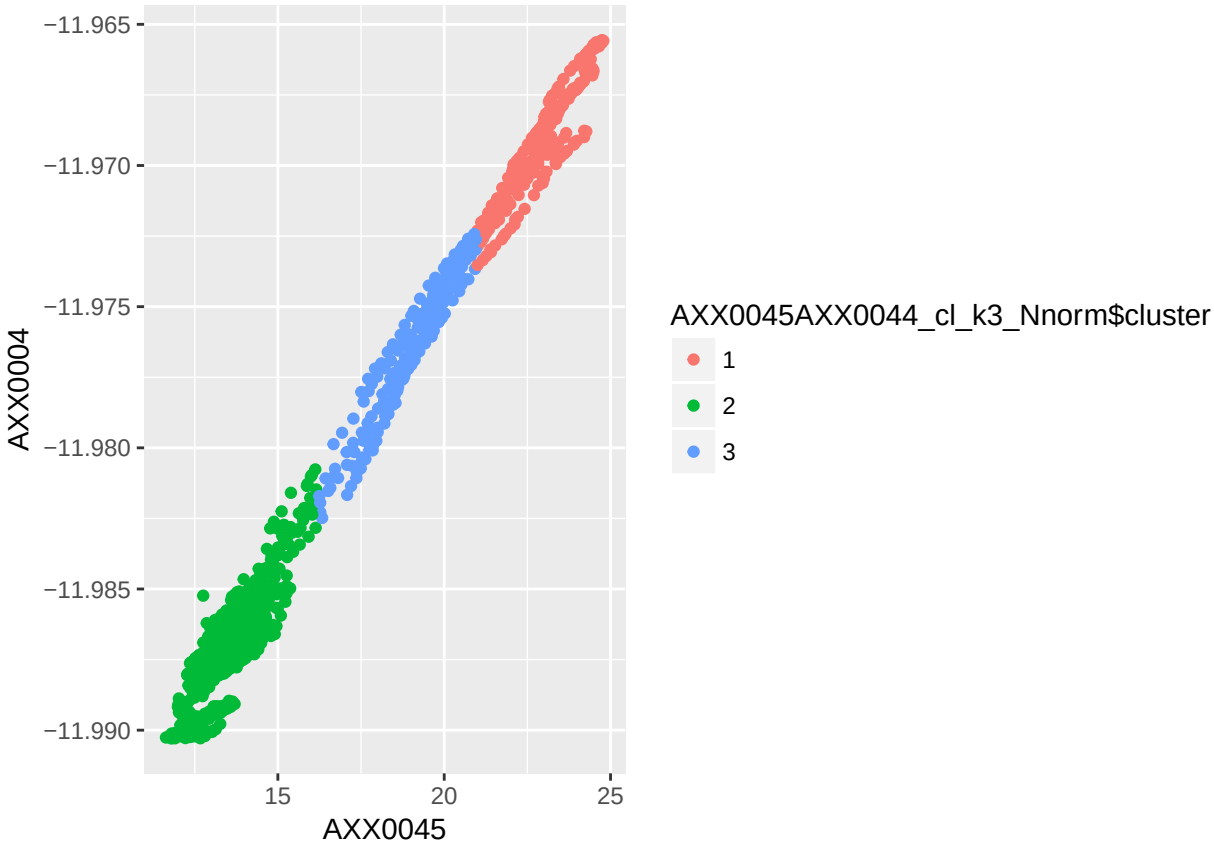

−11.965
−11.970
−11.975
−11.980
−11.985
−11.990
AXX0004
15
20
25
AXX0045
AXX0045AXX0044_cl_k3$cluster
1
2
3

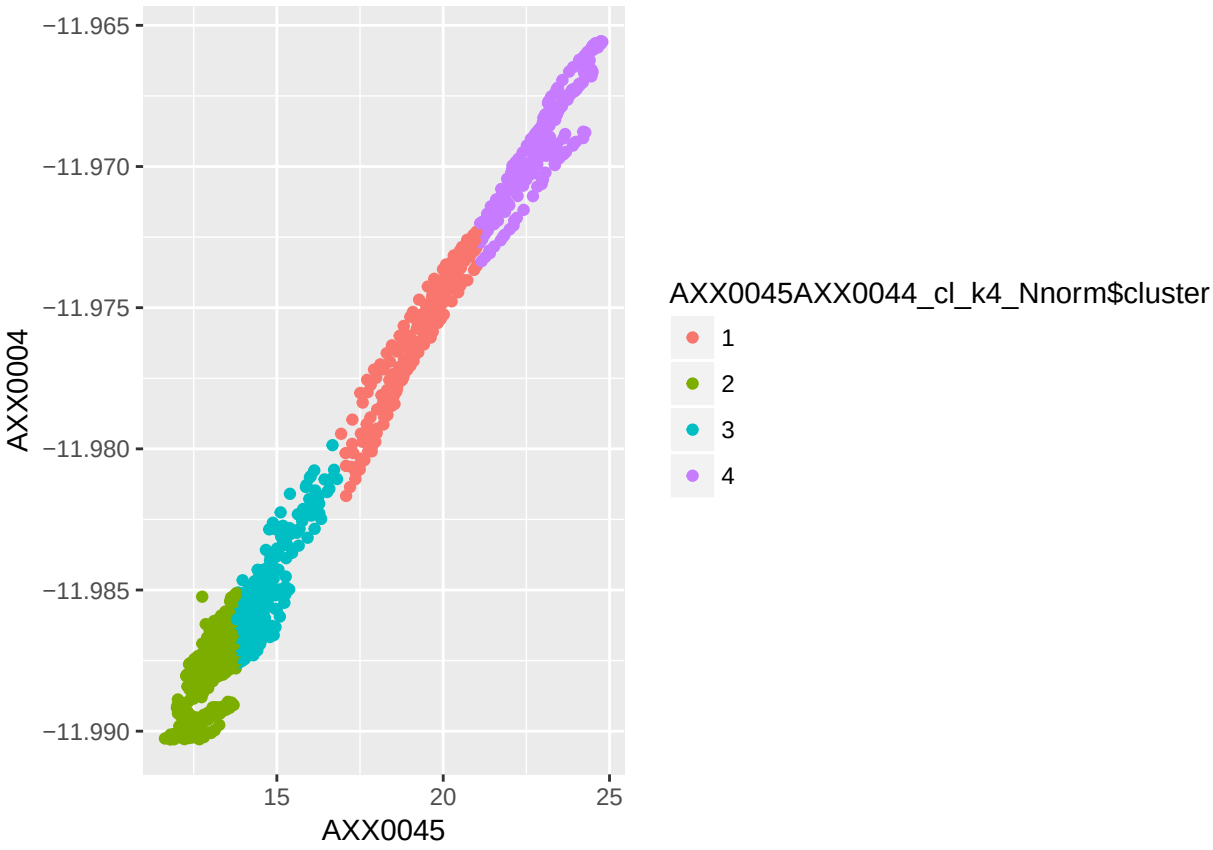


Para la información que queremos recolectar una segmentacion en K>3 no parece brindar informacion extra, en terminos de estacionalidad por ejemplo. Podriamos distinguir en modo global summer y winter y un mix de fall y spring pero no mas. En este caso la selección del parametro k surge de intentar utilizar la menor complejidad para explicar un set de datos (sumado al conocimiento de los mismos, etc). A priori podriamos haber sugerido que el set ideal de cumulos hubiera sido 4, este sale de asociar una dada evolucion a la estacion en curso, pero la seleccion de features utilizado para caracterizar el problema no parece favorecer esta eleccion.

```
## [1] "Not - Scaled"
```


−11.965
−11.970
−11.975
−11.980
−11.985
−11.990
AXX0004
15
20
25
AXX0045
AXX0045AXX0044_cl_k2_Nnorm$cluster
1
2

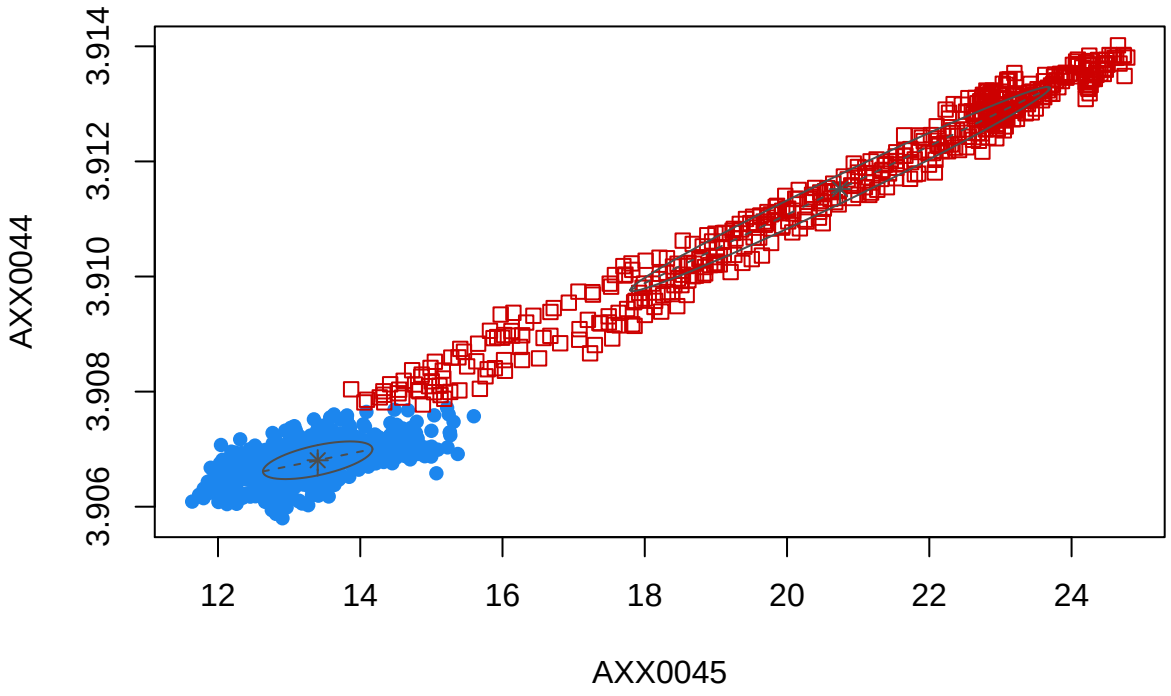

−11.965
−11.970
−11.975
−11.980
−11.985
−11.990
AXX0004
15
20
25
AXX0045
AXX0045AXX0044_cl_k3_Nnorm$cluster
1
2
3



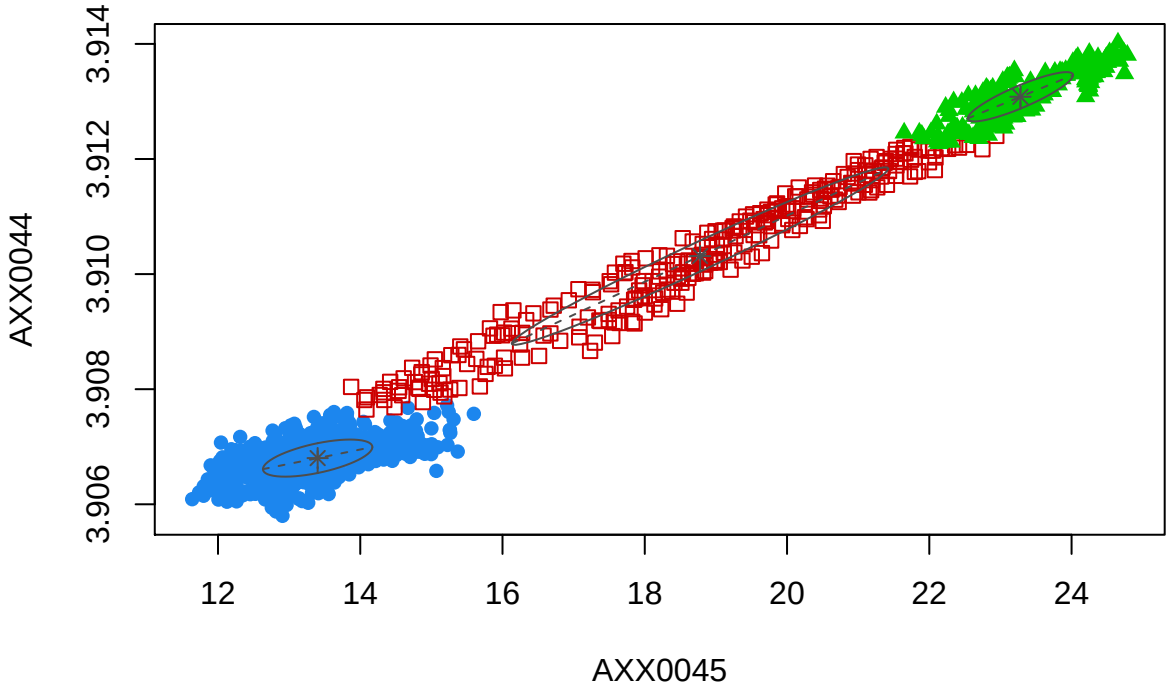
Para el caso no escalado vemos que la distribucion es similar pero cambia la composicion de los cumulos (en sintonia con los valores numericos encontrados con antelacion).

Hagamos unas pruebas con **mclust**. Veamos que diferencias encontramos con kmeans.

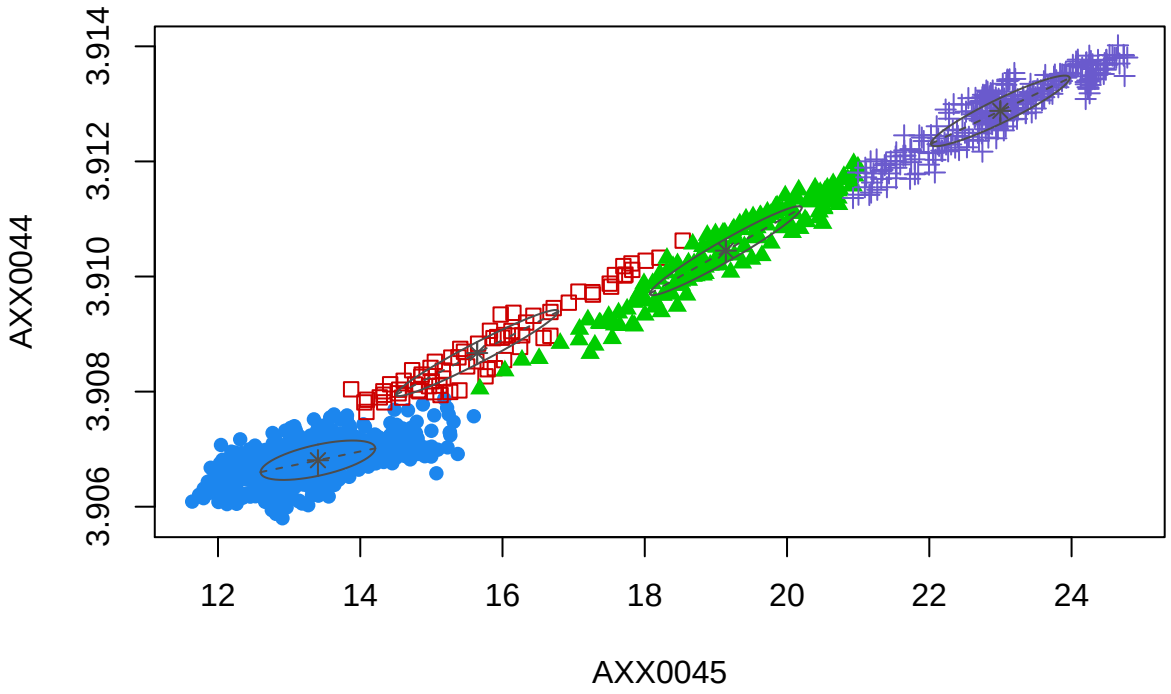
```
## Package 'mclust' version 5.4
## Type 'citation("mclust")' for citing this R package in publications.
```



A diferencia de kmeans para k=2 la distribucion de cumulos no parece simetrica.



Para k=3 parece similar a kmeans. Veamos k=4,



En esta caso se diferencia nuevamente en cuanto una de las elipses parece desplazarse respecto a la linea de tendencia central.

En cualquiera de los casos no logramos diferenciar la estacionalidad o marcarla en forma efectiva. Esto no es un defecto del algoritmo, sino que los features pueden no haber sido los adecuados.

Intentamos ver que pasa si seleccionamos el Doy y la temperatura (no deberia observarse una diferencia con el voltaje dada la relacion ya establecida). Esto a priori puede resultar trivial puesto que estamos justamente caracterizando una estacion usando fechas, pero igualmente veamos como responden los algoritmos, usamos kMeans para probar.

```
set.seed(20)
DOYAXX0045_cl_k2 <- kmeans(box_data_n[, c('Doy','AXX0045')], 2, nstart = 20)
DOYAXX0045_cl_k3 <- kmeans(box_data_n[, c('Doy','AXX0045')], 3, nstart = 20)
DOYAXX0045_cl_k4 <- kmeans(box_data_n[, c('Doy','AXX0045')], 4, nstart = 20)
DOYAXX0045_cl_k5 <- kmeans(box_data_n[, c('Doy','AXX0045')], 5, nstart = 20)
```

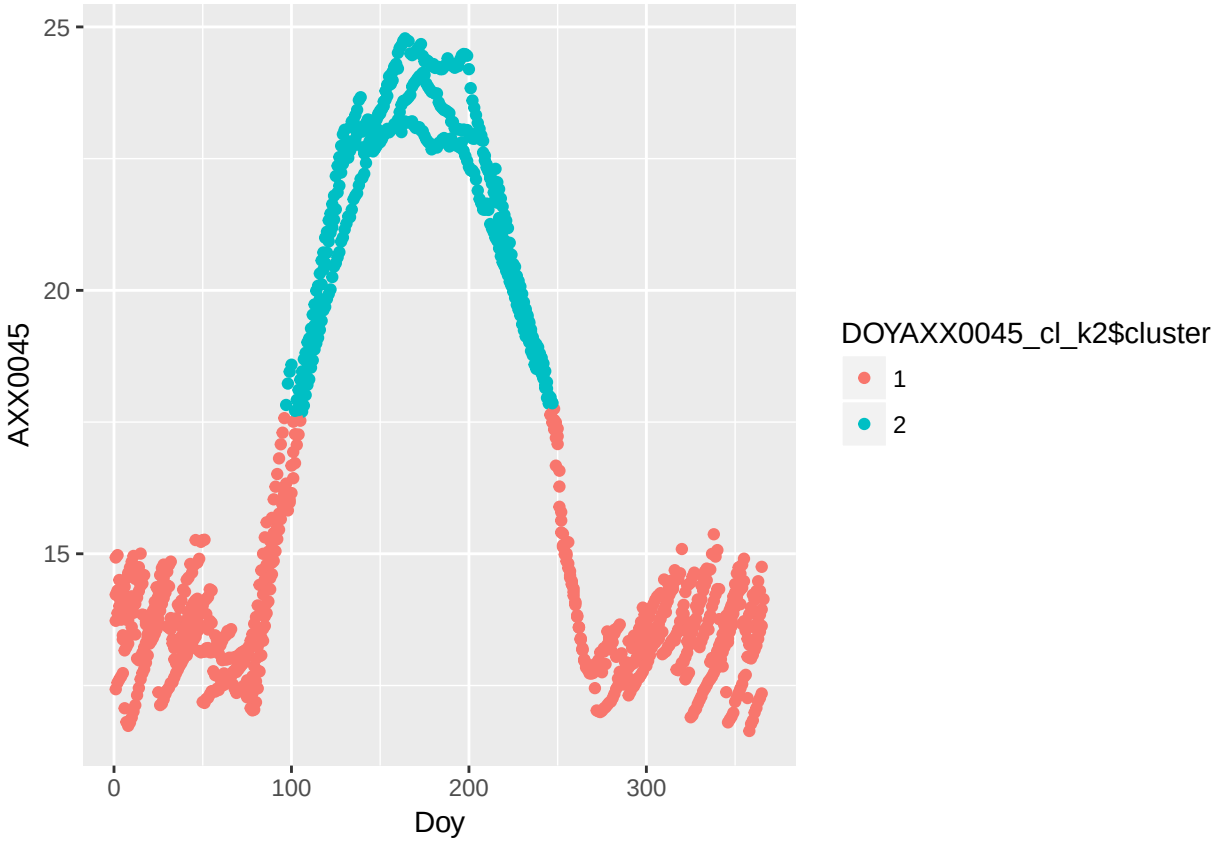
```
## [1] "Scaled Season North"
```

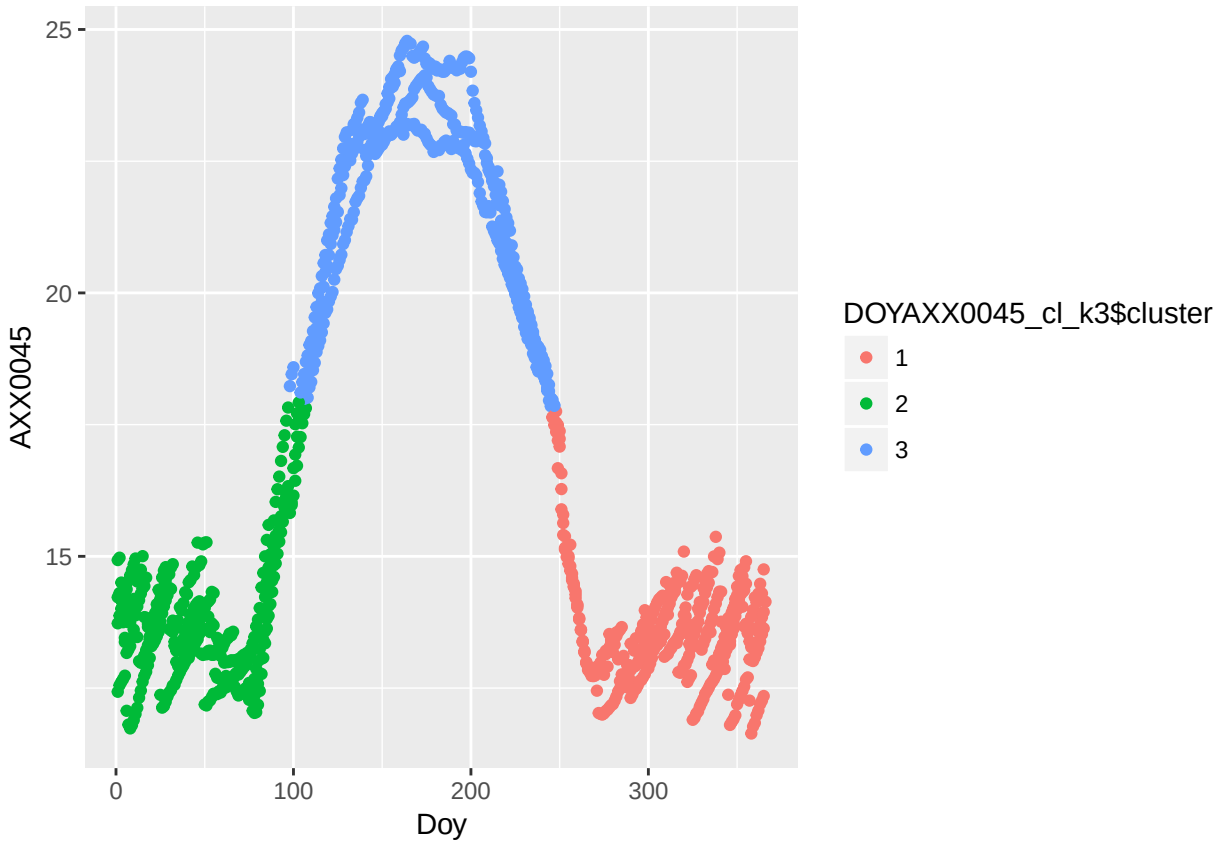
```
##
##     fall spring summer winter
##   1  277    155      0    365
##   2    9    148    276      0
```

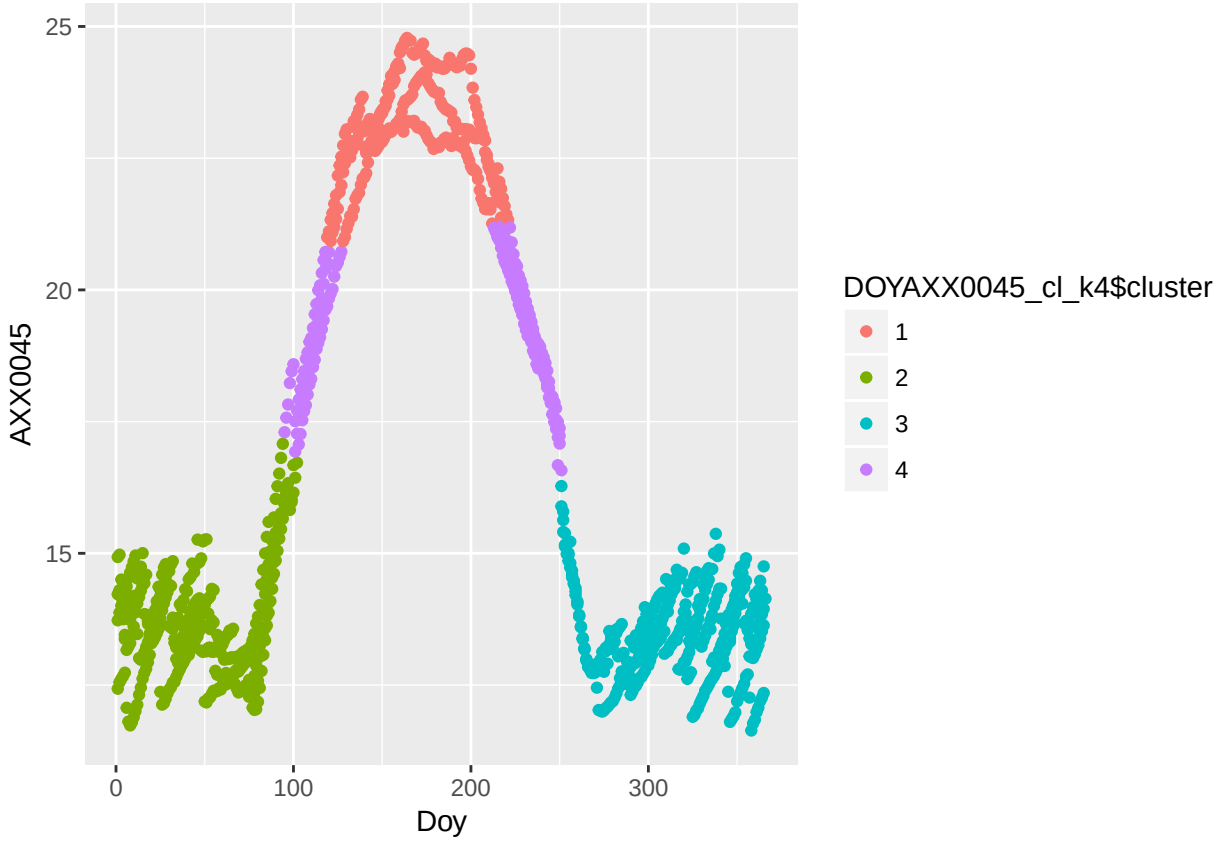
```
##
##     fall spring summer winter
##   1  277      0      0    125
##   2    0    162      0    240
##   3    9    141    276      0
```

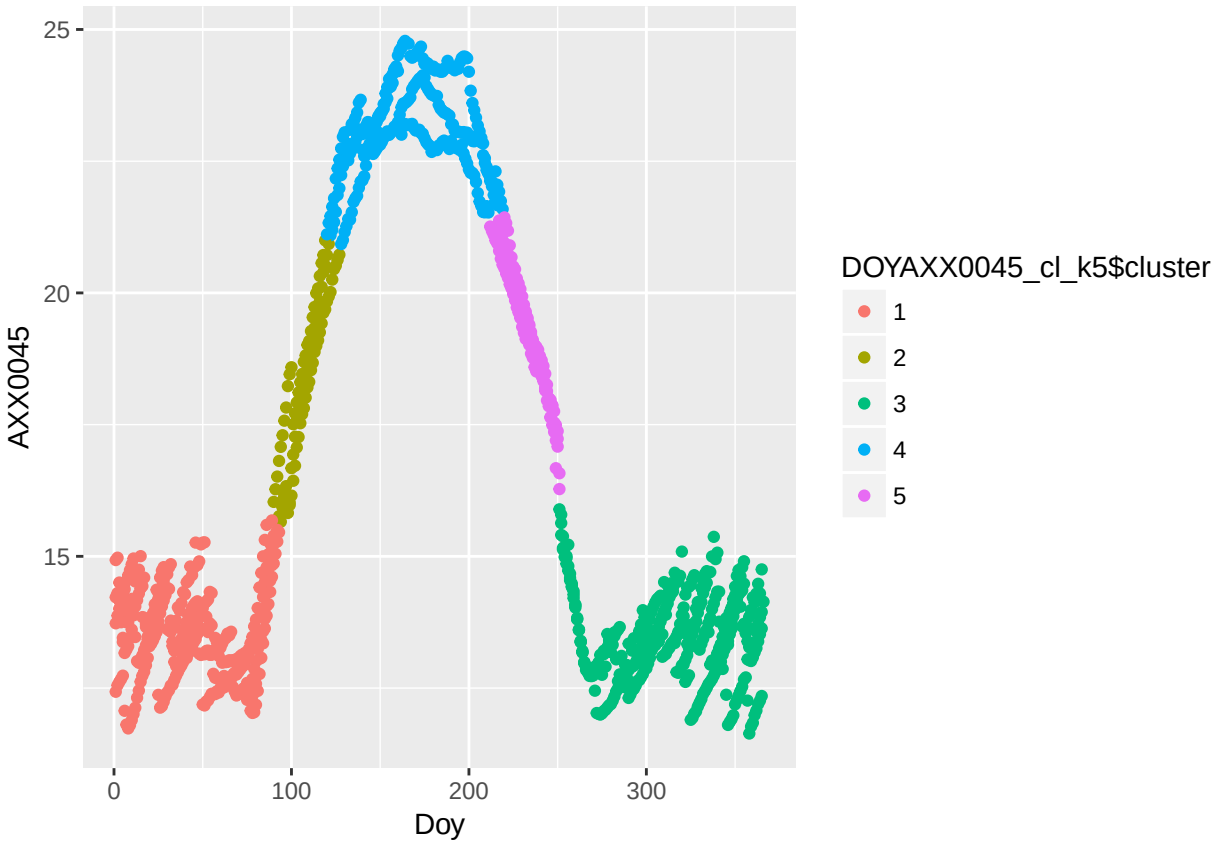
```
##
##     fall spring summer winter
##   1    0     88    197      0
##   2    0    146      0    240
##   3  264      0      0    125
##   4   22     69     79      0
```

```
##
##      fall spring summer winter
##   1     0    124      0    240
##   2     0     93      0      0
##   3   263      0      0    125
##   4     0     86    193      0
##   5    23      0     83      0
```

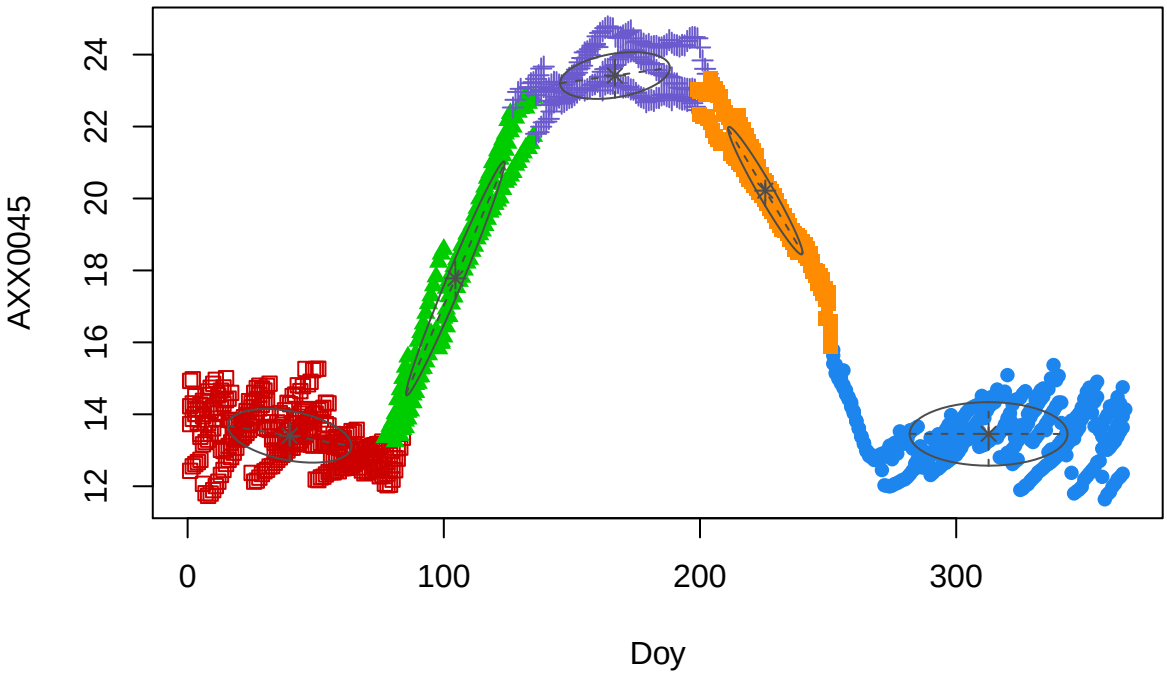



25
20
15
AXX0045
0
100
200
300
Doy
DOYAXX0045_cl_k3$cluster
1
2
3


25
20
15
AXX0045
0
100
200
300
Doy
DOYAXX0045_cl_k4$cluster
1
2
3
4



Utilizando estos features vemos que la clasificacion parece encaminarse a partir de k=5 (k=2/3/4 no aporta nada nuevo en si-respecto a estacionalidad). Mientras que k=5 ya esta marcando spring y fall. Veamos que pasa con **mclust**



Mclust predice incluso mejor un cambio de temporada (summer-fall - hemisferio sur) pero a ambos les cuesta la salida spring-summer. Se observa que la clasificacion sigue teniendo puntos faltantes (se observan mixturas sobre todo en los cambios de temporada). Igualmente por la clase de variables involucradas las fronteras de

definicion (para lo que quiere clasificar) entre clusters es difusa y no se pretende un score del 100% en la identificacion.