

Clusters

Construcción de clusters con las observaciones de 74 estaciones en CA

Existen muchas opciones para agrupar los caudales de las estaciones, así que se elige trabajar con 3:

- Distancia univariada de las variancias de cada estación.
- k-means usando los primeros componentes de PCA.
- Cluster para las series de tiempo utilizando <https://www.jstatsoft.org/article/view/v062i01/v62i01.pdf>

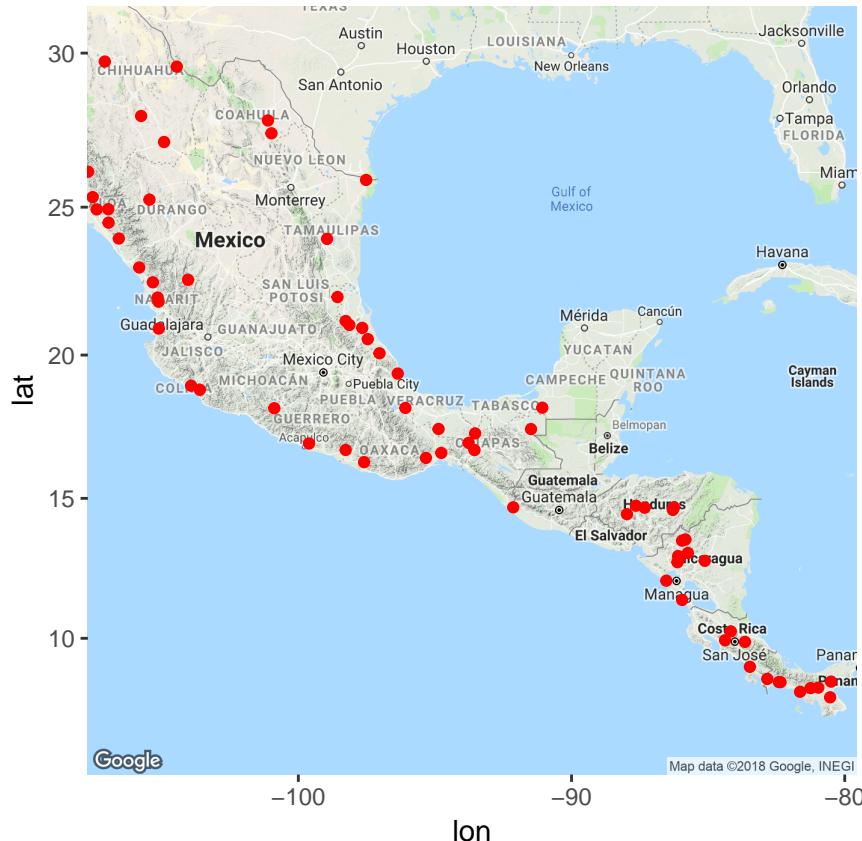
Los datos

Los datos consisten en observaciones mensuales de cuadal en 74 estaciones: las fechas de observación van desde enero de 1969 hasta diciembre de 1979. Cada estación representa una cuenca hidrológica en Centro América, y en total se tienen observaciones de 12 meses en cada uno de los 11 años para 74 locaciones. También, se tienen datos de climatología mensual para cada una de las estaciones.

```
## [1] 132 74  
## [1] 12 74
```

Las locaciones de las estaciones se pueden apreciar en el siguiente mapa:

```
## [1] 74 4  
##      left    bottom     right      top  
## -112.1223   5.6804 -75.2067  31.9196  
## converting bounding box to center/zoom specification. (experimental)  
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=18.8,-93.6645&zoom=5&size=640x640
```



Clustering

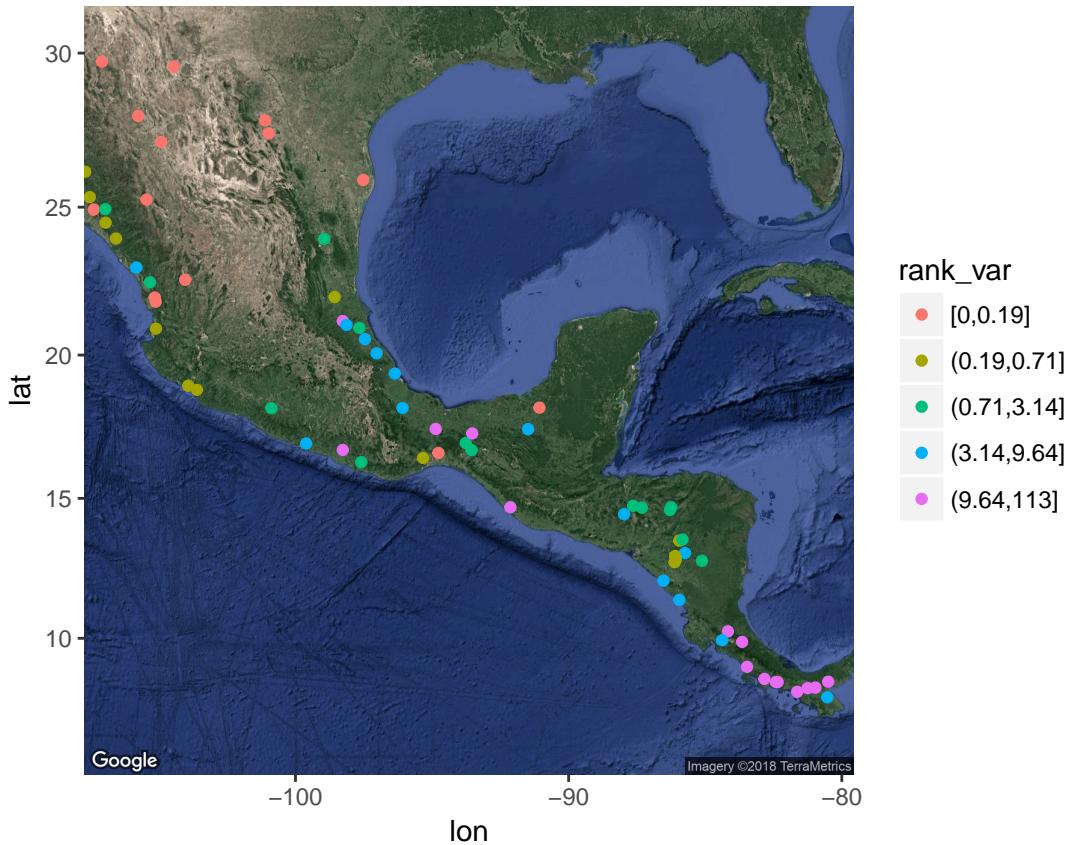
- OPCIÓN 1: Utilizar las variancias y luego agrupar por magnitud.

```

aa<-round(c(apply(caudal[,-c(1:3)],2,var)),4)
loca$rank_var <- cut_number(aa,5)
sq_map <- get_map(location = sbbox, maptype = "satellite", source = "google")

## converting bounding box to center/zoom specification. (experimental)
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=18.8,-93.6645&zoom=5&size=640x640
ggmap(sq_map) + geom_point(data = loca, mapping = aes(x = Longitud, y = Latitud, colour=rank_var))

```



Descripción de los clusters:

```

medianas <- apply(caudal[,-c(1:3)], 2, median)
data <- as_tibble(cbind(loca, medianas))
names(data) <- c("cod", "lat", "lon", "area", "rank_var", "Mediana_de_cluster")
## Área promedio de cada cluster:
round(tapply(data$area, data$rank_var, mean), 4)

##      [0,0.19]  (0.19,0.71]  (0.71,3.14]  (3.14,9.64]  (9.64,113]
##      54370.667     9323.790     4055.396     6172.333 80268425.533

## Mediana del Caudal de cada cluster:
round(tapply(data$Mediana_de_cluster, data$rank_var, median), 4)

##      [0,0.19]  (0.19,0.71]  (0.71,3.14]  (3.14,9.64]  (9.64,113]
##      0.0215       0.1631       0.4938       1.2625       4.7279

```

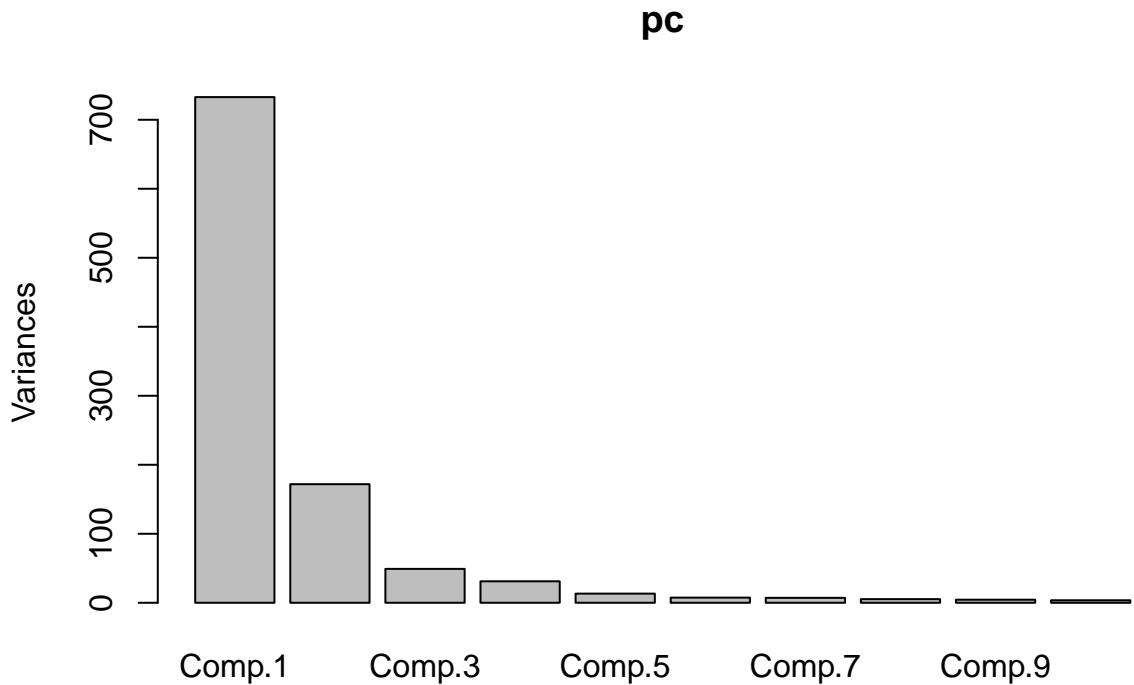
- OPCIÓN 2: k-means usando los componentes de PCA.

Primero, se deben calcular los PCA utilizando las anomalías en lugar de las observaciones. Como se tiene la climatología mensual para cada locación, el cálculo consiste en restar la climatología mensual a cada observación, según el mes correspondiente. Luego, se procede a hacer el PCA y por último agrupar las estaciones utilizando k-means de los primeros 10 componentes.

```

clima2 <- as.matrix(clima[,-1]) %x% rep(1, 11)
anomalies <- caudal[,-c(1:3)] - clima2
pc <- princomp(anomalies)
plot(pc)

```



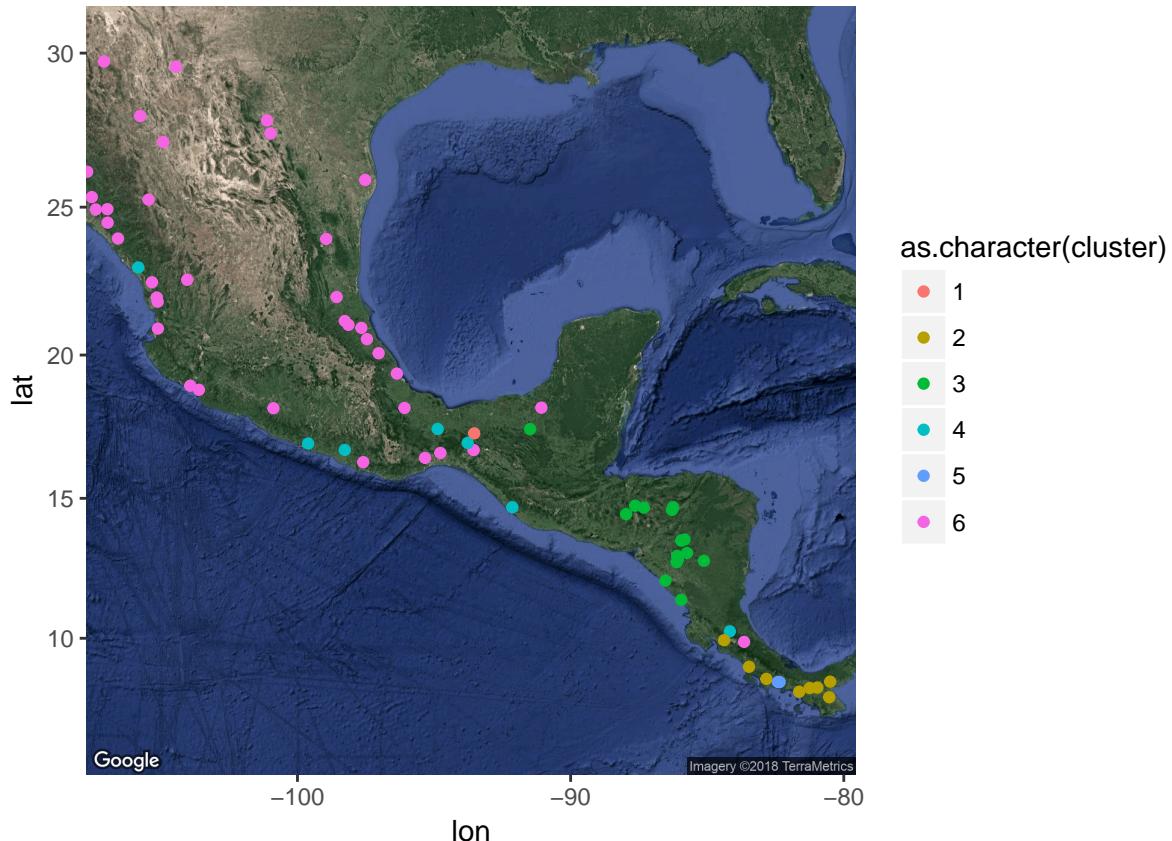
```

mydata <- (pc$loadings[,1:10])
fit <- kmeans(mydata, 6)
aggregate(mydata, by=list(fit$cluster), FUN=mean)

##   Group.1      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5
## 1    1 -0.20806905  0.013471286  0.173478112 -0.8315766147  0.21738611
## 2    2 -0.20322770  0.005580261  0.119450380  0.0172772003 -0.05210568
## 3    3 -0.04963763  0.035985692 -0.004445025  0.0273076181  0.03555346
## 4    4 -0.13415781 -0.049581883 -0.206325913 -0.0543633774 -0.09606152
## 5    5 -0.36826368  0.043582973  0.138011304  0.1755702158  0.02146485
## 6    6 -0.02694113  0.019459538 -0.065333470 -0.0001611807  0.01709005
##          Comp.6      Comp.7      Comp.8      Comp.9      Comp.10
## 1  0.26814490  0.26725038 -0.091022408 -0.06814561 -0.018697883
## 2 -0.08444466 -0.07915992 -0.071527432 -0.09314950  0.113872998
## 3  0.15548519 -0.10944314  0.057859389 -0.03167675 -0.011490048
## 4  0.03670901 -0.07091767 -0.087813852  0.12959367  0.032119010
## 5 -0.03624032  0.26417697  0.211635288  0.22267955 -0.242823127
## 6 -0.02035411  0.01312154 -0.008220804 -0.01979545  0.001146667

loca <- data.frame(loca, cluster=fit$cluster)
ggmap(sq_map) + geom_point(data = loca, mapping = aes(x = Longitud, y = Latitud, colour=as.character(clu

```



Descripción de los clusters:

```

data <- as_tibble(cbind(loca,medianas))
names(data) <- c("cod","lat","lon","area","rank_var","cluster2","Mediana_de_cluster")
## Area promedio de cada cluster:
round(tapply(data$area,data$cluster2,mean),4)

##          1         2         3         4         5
## 200.00 1406.00 5137.56 3804.00 602000668.50
##          6
## 25331.95

## Medianan del Caudal de cada cluster:
round(tapply(data$Mediana_de_cluster, data$cluster2,median),4)

##          1         2         3         4         5         6
## 4.7279 4.1007 0.5439 1.7374 9.0093 0.1344

```

- OPCIÓN 3: Cluster para las series de tiempo.

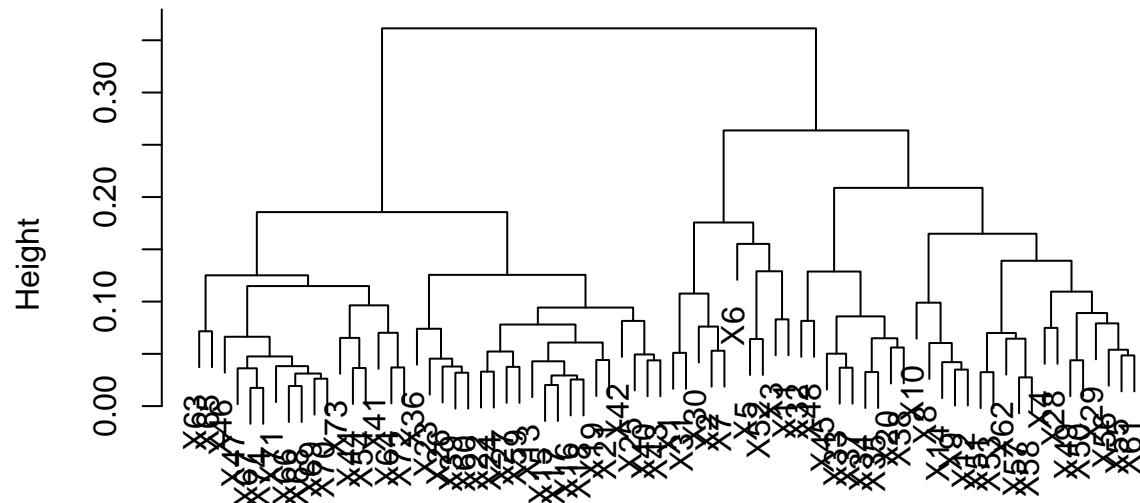
En este caso también se deben calcular las anomalías. Luego, se procede a aplicar el algoritmo de TSclust, que se describe aquí: <https://www.jstatsoft.org/article/view/v062i01/v62i01.pdf>

```

dpred <- diss(anomalies, "ACF", p=0.05)
hc.pred <- hclust(dpred)
plot(hc.pred)

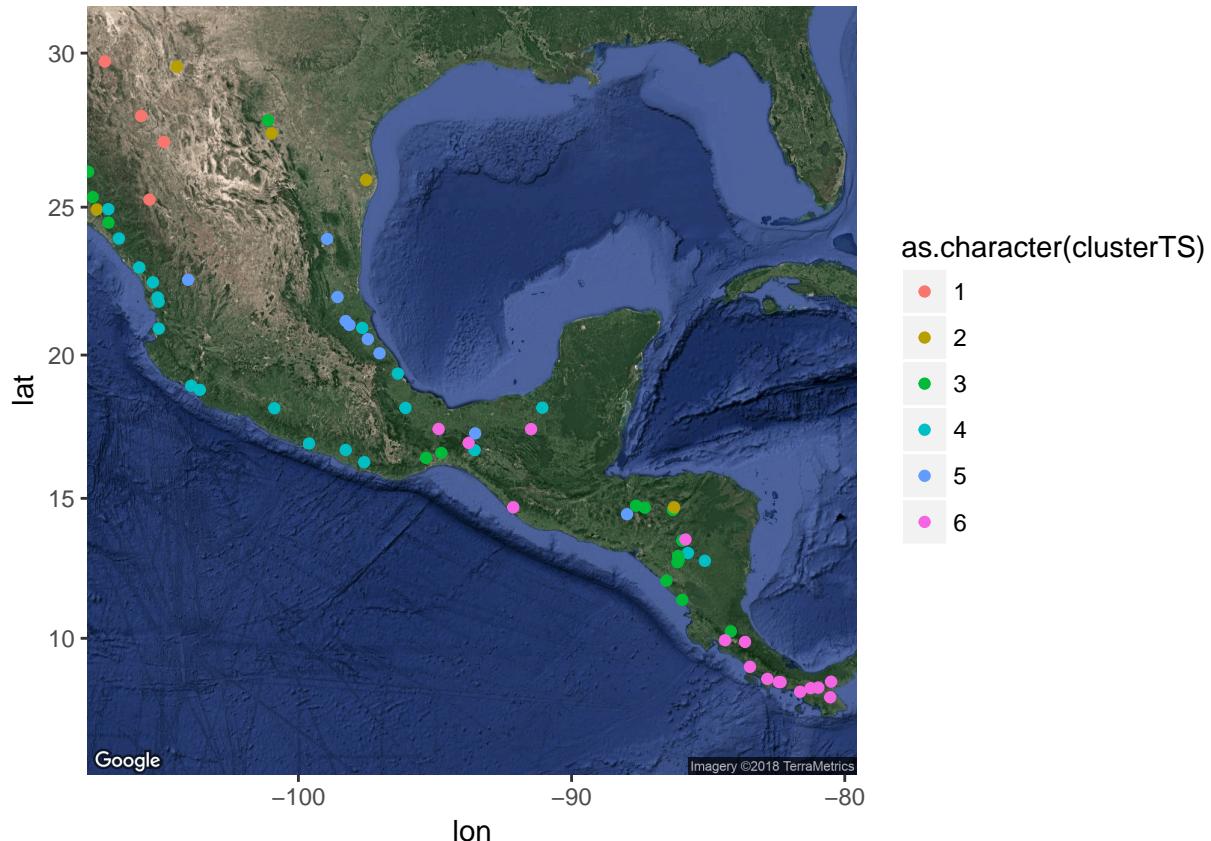
```

Cluster Dendrogram



```
dpred  
hclust (*, "complete")
```

```
aa<-cutree(hc.pred, k = 6)  
loca <- data.frame(loca, clusterTS=aa )  
ggmap(sq_map) + geom_point(data = loca, mapping = aes(x = Longitud, y = Latitud, colour=as.character(clu
```



Descripción de los clusters:

```

data <- as_tibble(cbind(loca,medianas))
names(data) <- c("cod","lat","lon","area","rank_var","cluster2","clusterTS", "Mediana_de_cluster")
## Area promedio de cada cluster:
round(tapply(data$area, data$clusterTS,mean),4)

##           1          2          3          4          5
## 12408.600 111564.200   4427.158  13718.250  8339.222
##           6
## 70828364.091

## Mediana del Caudal de cada cluster:
round(tapply(data$Mediana_de_cluster, data$cluster2,median),4)

##           1          2          3          4          5          6
## 4.7279 4.1007 0.5439 1.7374 9.0093 0.1344

```