spikedesafie

Alexis Suárez

13-10-2019

2.-Analisis general de los datos

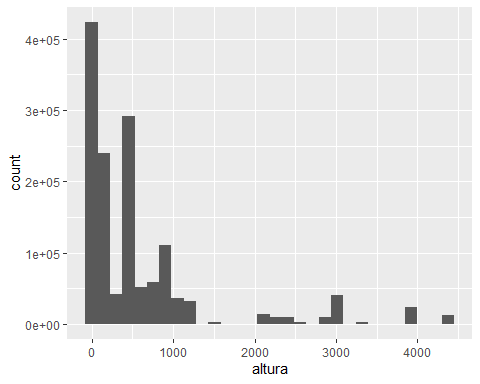
En un comienzo, es dificil hacer una análisis general de la base de datos sin haberla trabajado un poco, es por esto que este análisis lo realicé después de haber estado moviendo los datos un por un tiempo. Los datos vienen desordenados y hay varias variables por donde podrían ordenarse, más adelante preferentemente ordeno los datos por estación y por fecha para realizar los modelos. Tienen muchos datos duplicados, como la id de la cuenca o codigo de estación que podrían ser los mismos, el nombre de la estación deja de ser relevante luego de un tiempo ya que todo se lleva a números. En relación a los datos faltantes es porque el dataset es un promedio de sensores y si no hay sensores puede faltar este dato o el sensor falla y se pierde el dato, hay también distinta cantidad de datos por estación unas incluso a llegar a una sola medición. La distribuciones varían caudal y precipitaciones al parecer están relacionados por la forma del histograma y la temperatura tiene una normal centrada en 19 grados aprox.

## Warning: Missing column names filled in: 'X1' [1]

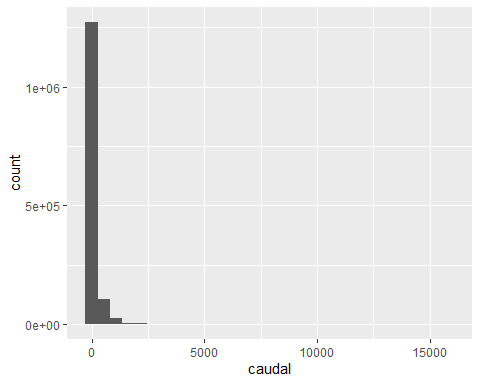
## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## codigo\_estacion = col\_double(),  
## institucion = col\_character(),  
## fuente = col\_character(),  
## nombre = col\_character(),  
## altura = col\_double(),  
## latitud = col\_double(),  
## longitud = col\_double(),  
## codigo\_cuenca = col\_double(),  
## nombre\_sub\_cuenca = col\_character(),  
## cantidad\_observaciones = col\_double(),  
## fecha = col\_datetime(format = ""),  
## caudal = col\_double(),  
## gauge\_id = col\_double(),  
## gauge\_name = col\_character(),  
## precip\_promedio = col\_double(),  
## temp\_max\_promedio = col\_double()  
## )

## X1 codigo\_estacion institucion fuente   
## Min. : 0 5410002: 20706 Length:1411180 Length:1411180   
## 1st Qu.: 352795 4308001: 20590 Class :character Class :character   
## Median : 705590 4320001: 20514 Mode :character Mode :character   
## Mean : 705590 4513001: 20477   
## 3rd Qu.:1058384 9129002: 20355   
## Max. :1411179 3430001: 20211   
## (Other):1288327   
## nombre altura latitud longitud   
## Length:1411180 Min. : 0.0 Min. :-54.04 Min. :-73.28   
## Class :character 1st Qu.: 45.0 1st Qu.:-37.21 1st Qu.:-72.07   
## Mode :character Median : 395.0 Median :-33.59 Median :-71.25   
## Mean : 585.6 Mean :-33.68 Mean :-71.22   
## 3rd Qu.: 750.0 3rd Qu.:-29.95 3rd Qu.:-70.53   
## Max. :4370.0 Max. :-18.23 Max. :-68.14   
##   
## codigo\_cuenca nombre\_sub\_cuenca cantidad\_observaciones  
## Min. : 10.00 Length:1411180 Min. : 802   
## 1st Qu.: 43.00 Class :character 1st Qu.:10255   
## Median : 57.00 Mode :character Median :14889   
## Mean : 62.56 Mean :15740   
## 3rd Qu.: 83.00 3rd Qu.:20611   
## Max. :128.00 Max. :36667   
##   
## fecha caudal gauge\_id   
## Min. :1960-01-02 00:00:00 Min. : 0.00 Min. : 1020003   
## 1st Qu.:1982-12-15 00:00:00 1st Qu.: 1.25 1st Qu.: 4320001   
## Median :1997-01-13 00:00:00 Median : 8.95 Median : 5734001   
## Mean :1994-08-26 11:49:20 Mean : 95.52 Mean : 6279953   
## 3rd Qu.:2007-12-04 00:00:00 3rd Qu.: 69.70 3rd Qu.: 8308000   
## Max. :2018-03-09 00:00:00 Max. :15805.00 Max. :12876004   
##   
## gauge\_name precip\_promedio temp\_max\_promedio  
## Length:1411180 Min. : 0.000 Min. :-19.05   
## Class :character 1st Qu.: 0.000 1st Qu.: 13.70   
## Mode :character Median : 0.000 Median : 19.08   
## Mean : 1.921 Mean : 19.21   
## 3rd Qu.: 0.000 3rd Qu.: 25.20   
## Max. :258.600 Max. : 41.50   
## NA's :27767 NA's :151563

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

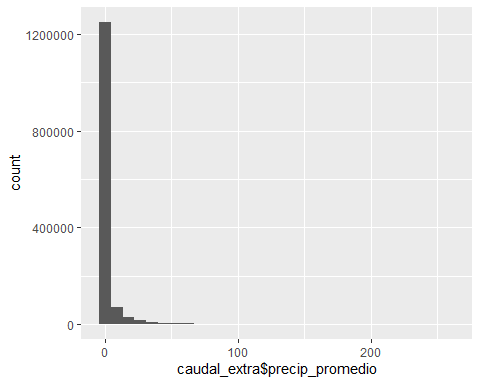


## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



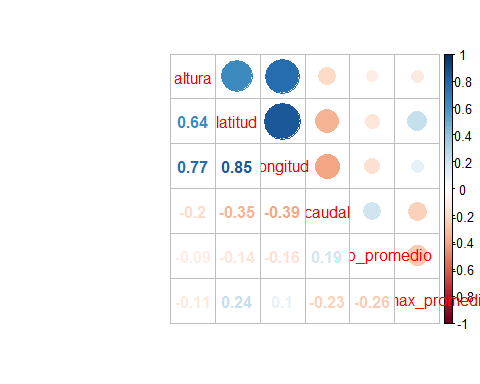
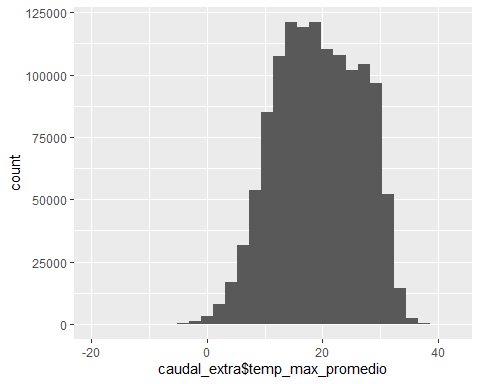
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

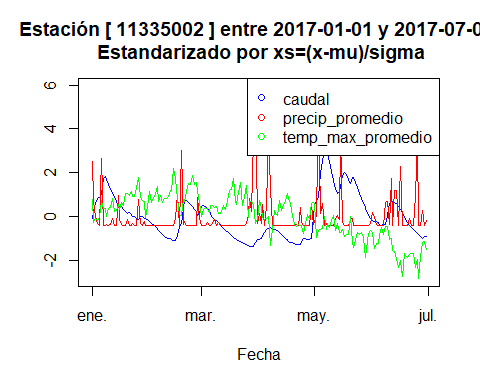
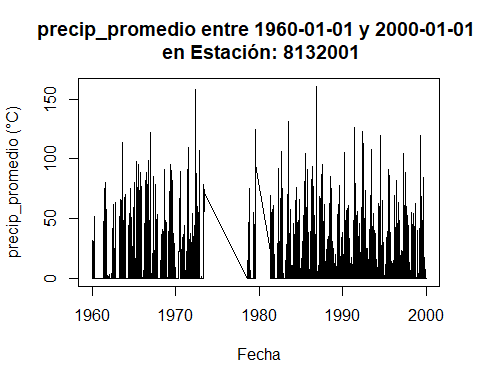
## Warning: Removed 27767 rows containing non-finite values (stat\_bin).



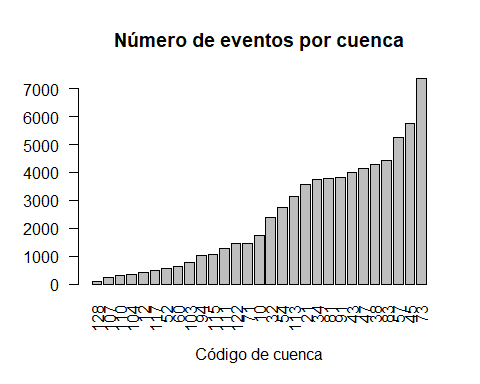
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 151563 rows containing non-finite values (stat\_bin).



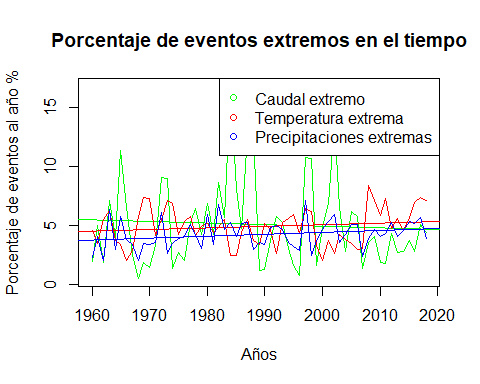


## num [1:1411180] 4540001 4540001 4540001 4540001 4540001 ...

 4.-Para generar las variables extremas, se realizó primero un analisis por cuenca, tomando cada cuenca y dependiendo de la fecha asignar la estación del año, luego de esto, por cada estación y cuenca obtener los percentiles dichos para cada variable (95) y clasificar en eventos extremos estos o no.

Me parece bastante razonable fijar las variables extremas de esta manera, pero igual podría ser un comportamiento extremo hacia abajo, por ejemplo, que la temperatura fuera muy baja también sería un extremo y eso quizá estaría dentro del percentil 5 y se podría catalogar como extremo. 5.-Si, las cuencas son diferentes y unas incluso en el histórico presentando 7 veces la cantidad de eventos extremos que otras. La cuenca con código 73, buscándola en maps se ve que es cercano a Talca cerca de la cordillera, quizá debido a sus condiciones es mucho más probable un evento extremo a una altura de 500 mts aproximadamente, por otro lado la cuenca 128 es a nivel del mar en la isla grande de tierra del fuego. 6.- Para realizar este gráfico se sacó el porcentaje de eventos extremos en relación al total de mediciones de manera anual. Luego, se realizó un gráfico mostrando el como se comportaba a lo largo del tiempo esta variable

caudal\_extra$year=as.factor(substr(caudal\_extra$fecha,1,4))  
j=1  
excau=c()  
extemp=c()  
exprecip=c()  
years=c()  
porcentajetotal=c()  
for (i in levels(caudal\_extra$year)) {  
 aux=caudal\_extra[caudal\_extra$year==i,]  
 excau[j]=sum(aux$caudal\_extremo,na.rm = T)/length(na.omit(aux$caudal\_extremo))\*100  
 extemp[j]=sum(aux$temp\_extremo,na.rm = T)/length(na.omit(aux$temp\_extremo))\*100  
 exprecip[j]=sum(aux$precip\_extremo,na.rm = T)/length(na.omit(aux$precip\_extremo))\*100  
 porcentajetotal[j]=(sum(aux$caudal\_extremo,na.rm = T)+sum(aux$temp\_extremo,na.rm = T)+sum(aux$precip\_extremo,na.rm = T))/(length(na.omit(aux$caudal\_extremo))+length(na.omit(aux$temp\_extremo))+length(na.omit(aux$precip\_extremo)))  
 years[j]=i  
 j=j+1  
 #print(i)  
}  
plot(x =years,y=excau,type="l" ,col="green",ylab = "Porcentaje de eventos al año %",xlab="Años",main="Porcentaje de eventos extremos en el tiempo")  
abline(lm(excau~as.double(years)),col="green")  
lines(x=years,y=extemp,col="red")  
abline(lm(extemp~as.double(years)),col="red")  
lines(x=years,y=exprecip,col="blue")  
abline(lm(exprecip~as.double(years)),col="blue")  
legend("topright",legend=c("Caudal extremo","Temperatura extrema","Precipitaciones extremas"),col = c("green", "red", "blue"),pch=1)

 Realizando una regresión lineal, se puede apreciar como las temperaturas extremas y precipitaciones han ido en aumento a lo largo del tiempo, pero no ocurre con el caudal extremo, puede que ante este aumento de temperatura se genere problemas de sequía que no aumenta el nivel del caudal en los rios. 7.- Para este modelo se generó un modelo de clasificación como output la variable binaria caudal\_extremo, de input se calculó el promedio de los 5 días anteriores del caudal, temperatura y precipitaciones(separando por cada estación y ordenando por la fecha). Además, utilizando las variables de estación de clima y estación de medición como dummy variables. El modelo de clasificación de árbol elegido por tema de clasificación y el tipo de variables, dio como resultado un 68,68% de precisión, lo cual se validó con un 25-fold cross-validation por la cantidad de datos que se entrenó el modelo 120.000 aprox. Para generar esta base de entrenamiento, estaba el problema de un desbalance en el output, ya que exisitían mucho más eventos no extremos que extremos, lo cual podría sesgar nuestro modelo a predecir que todo estaba bien predicho siendo que el modelo que fue entrenado con 90% de ceros como output. Para este desbalance, se tomaron todos los eventos considerados extremos anteriormente y se le agregaron la misma cantidad(aleatorios) de datos considerados como no extremos, así no crear un sesgo en el modelo, además para entrenarlo se mezclaron de forma aleatoria los datos antes de entrar al modelo. Se eligieron 5 días antes para la predicción porque me apreció razonable tomar más de un día, y viendo los datos, me percaté que al haber un aumento en precipitaciones en días anteriores, más de un día anterior, se generaba un evento extremo en el caudal, por lo que podía ser relevante a la hora del modelo. Por lo que para usar este modelo, tendríamos que tomar las mediciones de 5 días anteriores (sin contar el día actual), la estación climática (del día a predecir) y la estación de medición y el modelo tendría un 68% de precisión para predecir o clasificar si habrá un evento extremo o no.

cntrl<-trainControl(method = "cv",number = 25)  
caudal\_extra$verano=(ifelse(caudal\_extra$estacionclima=="Verano",1,0))  
caudal\_extra$otono=(ifelse(caudal\_extra$estacionclima=="Otoño",1,0))  
caudal\_extra$invierno=(ifelse(caudal\_extra$estacionclima=="Invierno",1,0))  
caudal\_extra$primavera=(ifelse(caudal\_extra$estacionclima=="Primavera",1,0))  
datamodeling<-na.omit((caudal\_extra[,c(12,20,13,16,17,21,24:28)]))  
datamodeling$indxcuenca=as.double(datamodeling$indxcuenca)  
aux2=NULL  
for (i in 1:max(datamodeling$indxcuenca)) {  
 aux=datamodeling[datamodeling$indxcuenca==i,]  
 aux=aux[order(aux$fecha),]  
  
 if(nrow(aux)>5) {  
 for (col in 3:5) {  
   
 for (j in 6:nrow(aux)) {  
 aux[j,paste(colnames(aux)[col],"\_5dias",sep = "")]=mean(aux[c((j-5):(j-1)),col],na.rm = T)  
   
 }  
 }  
 aux=na.omit(aux)  
 aux$caudal\_extremo=as.factor(aux$caudal\_extremo)  
 aux2=rbind(aux2,aux)  
 }  
 #print(i)  
}  
for (i in 1:max(datamodeling$indxcuenca)) {  
 aux2[,paste("dummy\_estación\_",i,sep = "")]=ifelse(aux2$indxcuenca==i, 1,0)  
 # print(i)  
}  
dataformodel=aux2[,c(6,8:ncol(aux2))]  
  
dataformodel0=dataformodel[dataformodel$caudal\_extremo==0,]  
  
set.seed(123)  
numberofdata=nrow(dataformodel[dataformodel$caudal\_extremo==1,])  
train\_ind <- sample(seq\_len(nrow(dataformodel0)), size = numberofdata)  
dataformodeltrain=rbind(dataformodel[dataformodel$caudal\_extremo==1,],dataformodel0[train\_ind,])  
  
#dataformodeltrain=dataformodel[train\_ind,]  
#dataformodeltrain[,c(2:5,9:ncol(dataformodeltrain))]=as.factor(dataformodeltrain[,c(2:5,9:ncol(dataformodeltrain))])  
dataformodeltrain=dataformodeltrain[sample(nrow(dataformodeltrain)),]  
model=train(caudal\_extremo~.,data = dataformodeltrain,method = "rpart", trControl = cntrl )  
model

## CART   
##   
## 126302 samples  
## 140 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (25 fold)   
## Summary of sample sizes: 121250, 121250, 121249, 121250, 121249, 121250, ...   
## Resampling results across tuning parameters:  
##   
## cp Accuracy Kappa   
## 0.006360417 0.6868459 0.3736916  
## 0.054021314 0.6605201 0.3210422  
## 0.261136007 0.5653118 0.1306279  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was cp = 0.006360417.

pred=predict(model, dataformodeltrain[,-1])  
confusionMatrix(pred,dataformodeltrain[,1])

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 51897 28583  
## 1 11254 34568  
##   
## Accuracy : 0.6846   
## 95% CI : (0.682, 0.6872)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3692   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.8218   
## Specificity : 0.5474   
## Pos Pred Value : 0.6448   
## Neg Pred Value : 0.7544   
## Prevalence : 0.5000   
## Detection Rate : 0.4109   
## Detection Prevalence : 0.6372   
## Balanced Accuracy : 0.6846   
##   
## 'Positive' Class : 0   
##

varImp(model)

## rpart variable importance  
##   
## only 20 most important variables shown (out of 140)  
##   
## Overall  
## caudal\_5dias 100.0000  
## precip\_promedio\_5dias 83.0914  
## temp\_max\_promedio\_5dias 25.2587  
## verano 4.3817  
## dummy\_estación\_22 4.1658  
## dummy\_estación\_27 3.0372  
## dummy\_estación\_75 1.4317  
## dummy\_estación\_13 1.4195  
## dummy\_estación\_12 0.1573  
## dummy\_estación\_58 0.0000  
## dummy\_estación\_100 0.0000  
## dummy\_estación\_64 0.0000  
## dummy\_estación\_82 0.0000  
## dummy\_estación\_101 0.0000  
## dummy\_estación\_98 0.0000  
## dummy\_estación\_45 0.0000  
## dummy\_estación\_107 0.0000  
## dummy\_estación\_106 0.0000  
## dummy\_estación\_84 0.0000  
## dummy\_estación\_51 0.0000

numberofdata=nrow(dataformodel[dataformodel$caudal\_extremo==1,])\*3/7  
train\_ind <- sample(seq\_len(nrow(dataformodel0)), size = numberofdata)  
dataformodeltrain=rbind(dataformodel[dataformodel$caudal\_extremo==1,],dataformodel0[train\_ind,])  
model=train(caudal\_extremo~.,data = dataformodeltrain,method = "rpart", trControl = cntrl )  
model

## CART   
##   
## 90215 samples  
## 140 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (25 fold)   
## Summary of sample sizes: 86606, 86607, 86607, 86606, 86606, 86607, ...   
## Resampling results across tuning parameters:  
##   
## cp Accuracy Kappa   
## 0.01877032 0.7577122 0.3011795  
## 0.04932752 0.7321847 0.2915721  
## 0.08649867 0.7157676 0.2059015  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was cp = 0.01877032.

8.-a) En relación a lo anterior las métricas son de precisión, es decir la cantidad de verdaderos positivos dividido por la suma de los falsos positivos y verdaderos positivos. Las variables más importantes son el caudal los 5 días previos, luego la precipitación los 5 días previos y la temperatura de esos días.Luego en menor importancia la dummy variable del verano y las dummy de las estaciones de medición. b) Mi modelo es capaz de capturar 68% lo cual es bastante cercano, quizá haciendo un refinado más fino de las variables de input o de la cantidad de días previos a tomar mejoraría bastante en predicción. También podría alterar el balance de la base de datos, para que prediga más los 1 que los 0 y así mover un poco la balanza. Al cambiar esto la precisión aumentó a un 75% es bastante útil pero como que se fuerza un poco al modelo a predecir los eventos extremos.