



International Center for Tropical Agriculture  
*Since 1967 Science to cultivate change*

## Procesamiento de datos faltantes

4 julio 2018

Juan Camilo Rivera

[j.c.rivera@cgiar.org](mailto:j.c.rivera@cgiar.org)

Hugo Dorado

[h.a.dorado@cgiar.org](mailto:h.a.dorado@cgiar.org)



CIAT is a CGIAR Research Center



# Fuentes de información

## WorldClim - Global Climate Data

*Free climate data for ecological modeling and GIS*



**IDEAM**

Instituto de Hidrología,  
Meteorología y  
Estudios Ambientales

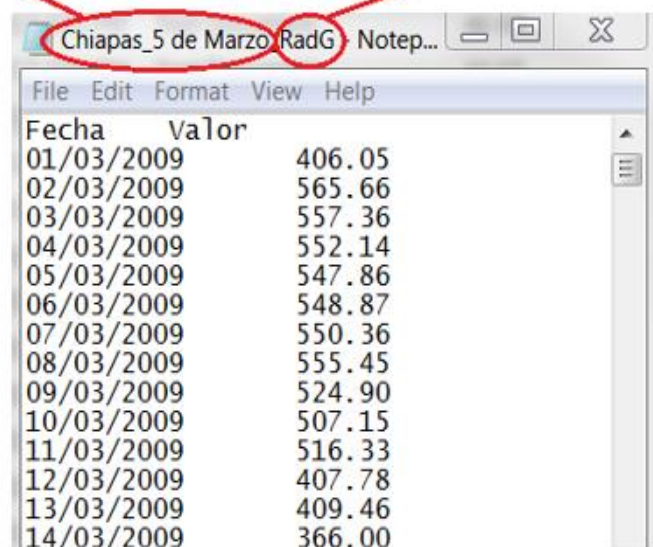


# Archivos planos

## Tipos

Nombre estación

Variable climatologica



Fecha	Valor
01/03/2009	406.05
02/03/2009	565.66
03/03/2009	557.36
04/03/2009	552.14
05/03/2009	547.86
06/03/2009	548.87
07/03/2009	550.36
08/03/2009	555.45
09/03/2009	524.90
10/03/2009	507.15
11/03/2009	516.33
12/03/2009	407.78
13/03/2009	409.46
14/03/2009	366.00



	A	B	C	D	E	F
1	DATE	ESOL	RAIN	RHUM	TMAX	TMIM
.557	4/5/2009	412.8747	0	70.99139	36	24.3016
.558	4/6/2009	513.9043	0	75.20833	34.8	24.9
.559	4/7/2009	396.5338	0	73.85714	34.1	25.6
.560	4/8/2009	397.8491	0	74.09524	33.9	25.4
.561	4/9/2009	448.4498	0	76.82609	34.6	24.9
.562	4/10/2009	481.8188	0	66.20671	39	24.8
.563	4/11/2009	448.1053	0	72.66386	35.9	25.4

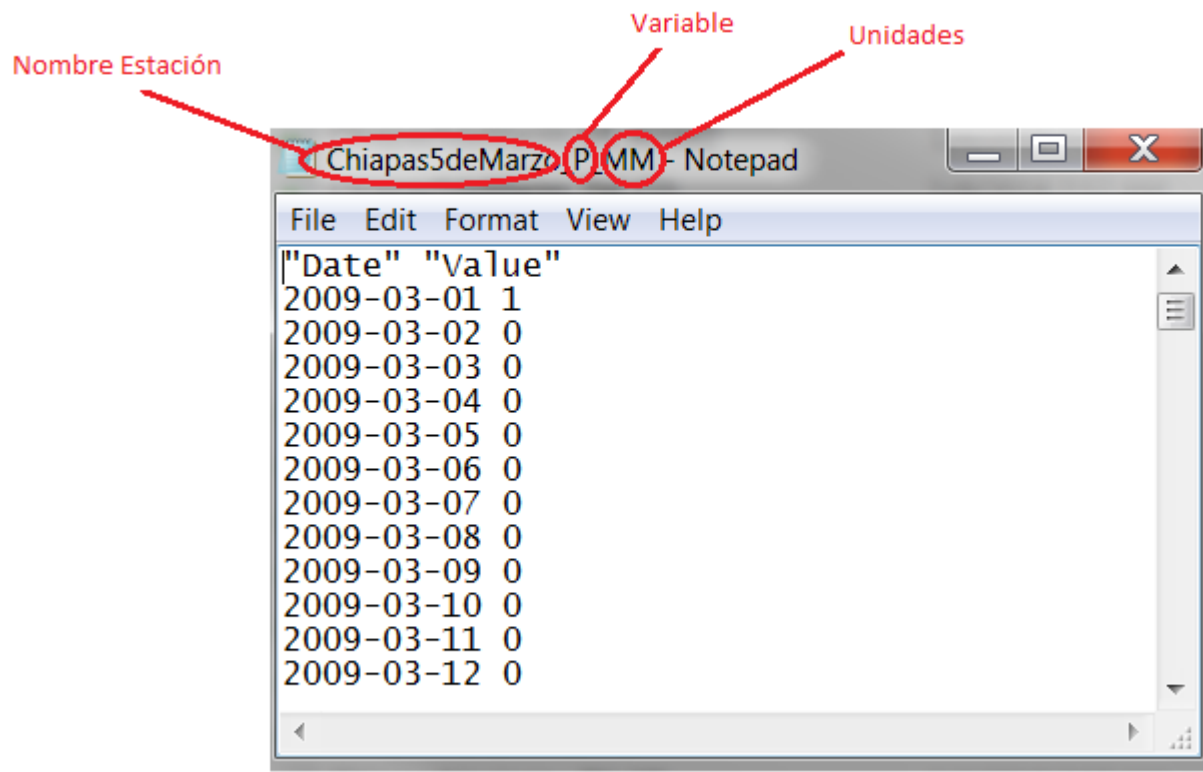
# Variables y unidades

Abreviación	Significado (Ingles)	Significado (español)
TX	Maximum temperature	Temperatura máxima
TM	Minimum temperature	Temperatura mínima
P	Precipitation	Precipitación
RH	Relative humidity	Humedad relative
SR	Solar radiation	Radiación solar

# Unidades

Abreviacion	Unidad de Medida
CD	Grados Celisus
FD	Grados Fahrenheit
MM	Mililitros
NE	Número entre 0 y 100
CCM2	Calorias por centimetro cuadrado
MJM2	Mega Julio por metro cuadrado
WAM2	Watts por metro cuadrado

# Formato único



# Llenado de faltantes

Date	value
19800101	NA
19800102	NA
19800103	NA
19800104	NA
19800105	NA
19800106	NA
19800107	NA
19800108	NA
19800109	NA
19800110	NA
19800111	35.2
19800112	NA
19800113	NA
19800114	36.2
19800115	35.2
19800116	NA

## Vector Autoregresivo Regresión (VAR)

$$x_t = A_1 \cdot x_{t-1} + \dots + A_p \cdot x_{t-p} + u_t$$

$x_t$  = Vector de dimension K, conjunto de variables de clima.

$A_i$  = Es el coeficiente de la matriz K x K

$u_t$  = Es un proceso estocastico de dimension K

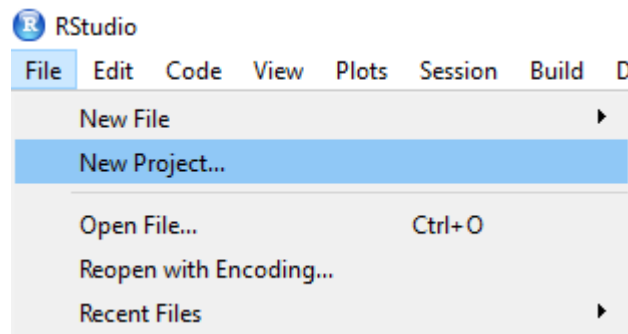


## Pasos para el ejercicio

1. Abrir el programa R Studio.

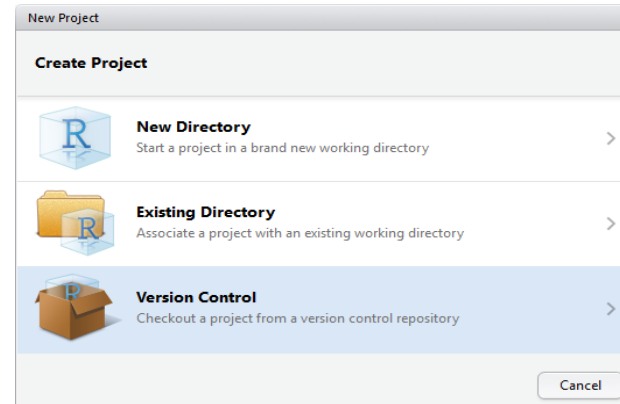


2. Crear un nuevo proyecto.

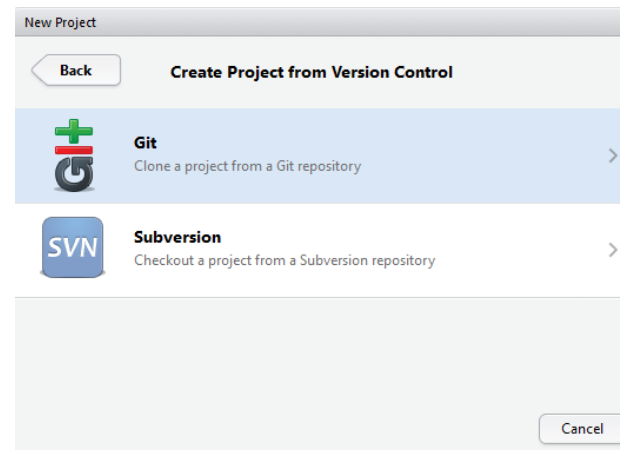




3. Click version control

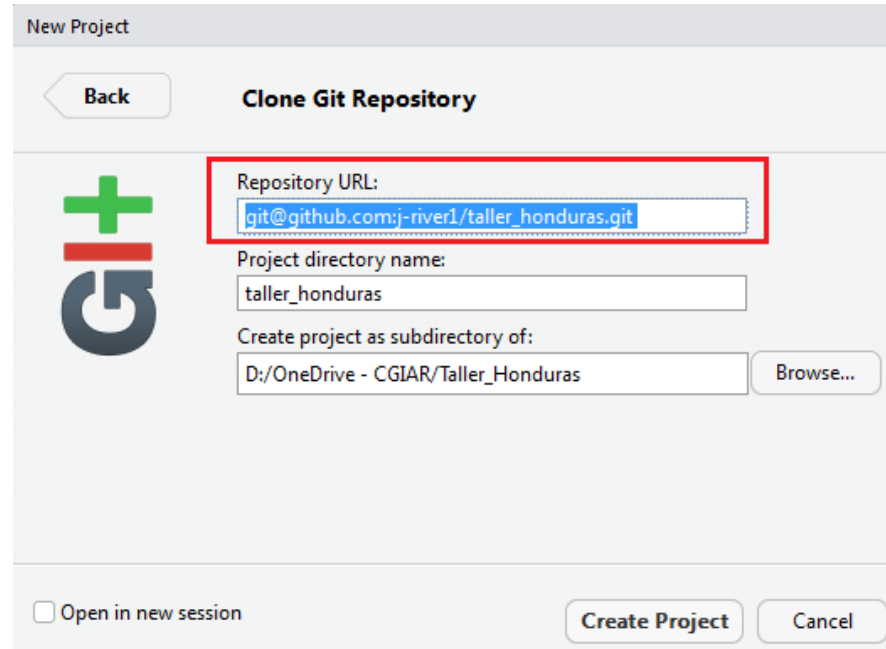


4. Click en Git



5. Ingresar esta dirección en la casilla repository URL y dar click en create project:

git@github.com:j-river1/taller\_honduras.git



New Project

Back Clone Git Repository

Repository URL:  
git@github.com:j-river1/taller\_honduras.git

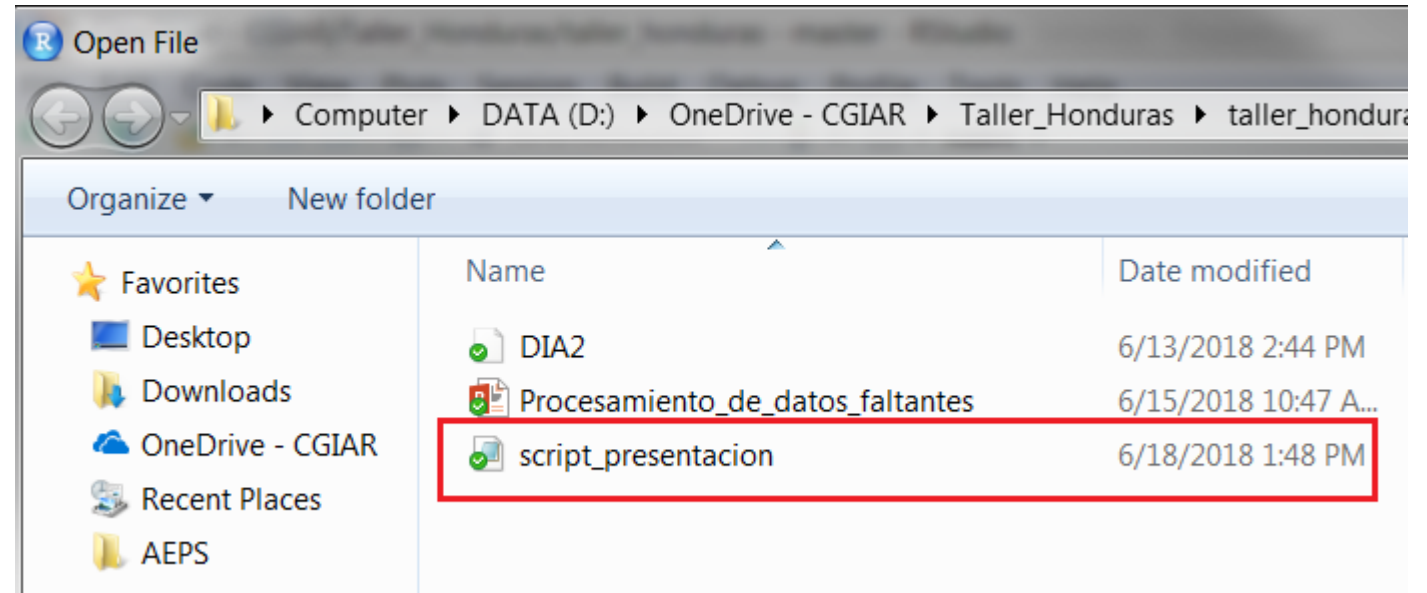
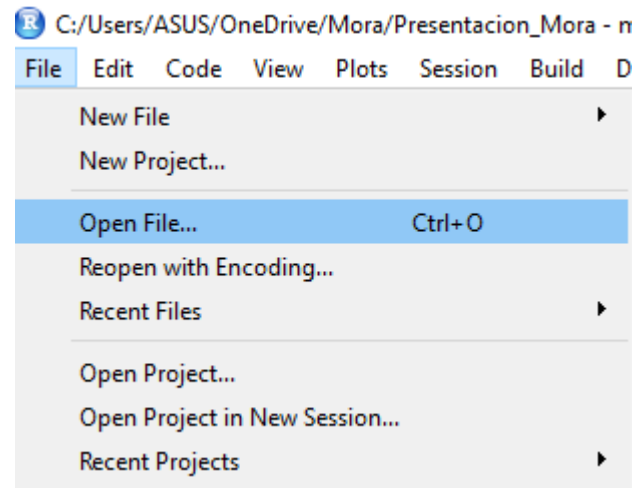
Project directory name:  
taller\_honduras

Create project as subdirectory of:  
D:/OneDrive - CGIAR/Taller\_Honduras Browse...

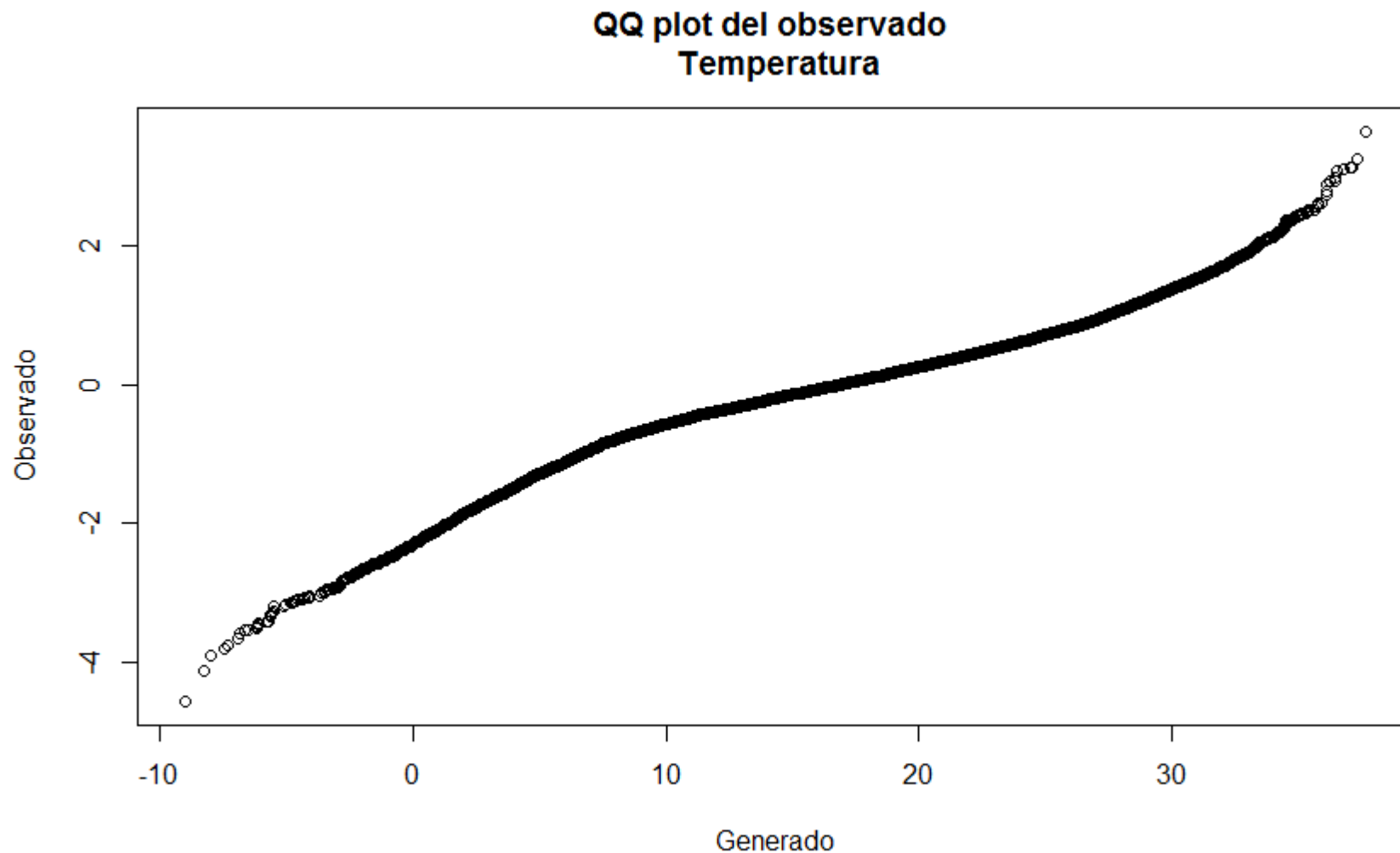
☐ Open in new session

Create Project Cancel

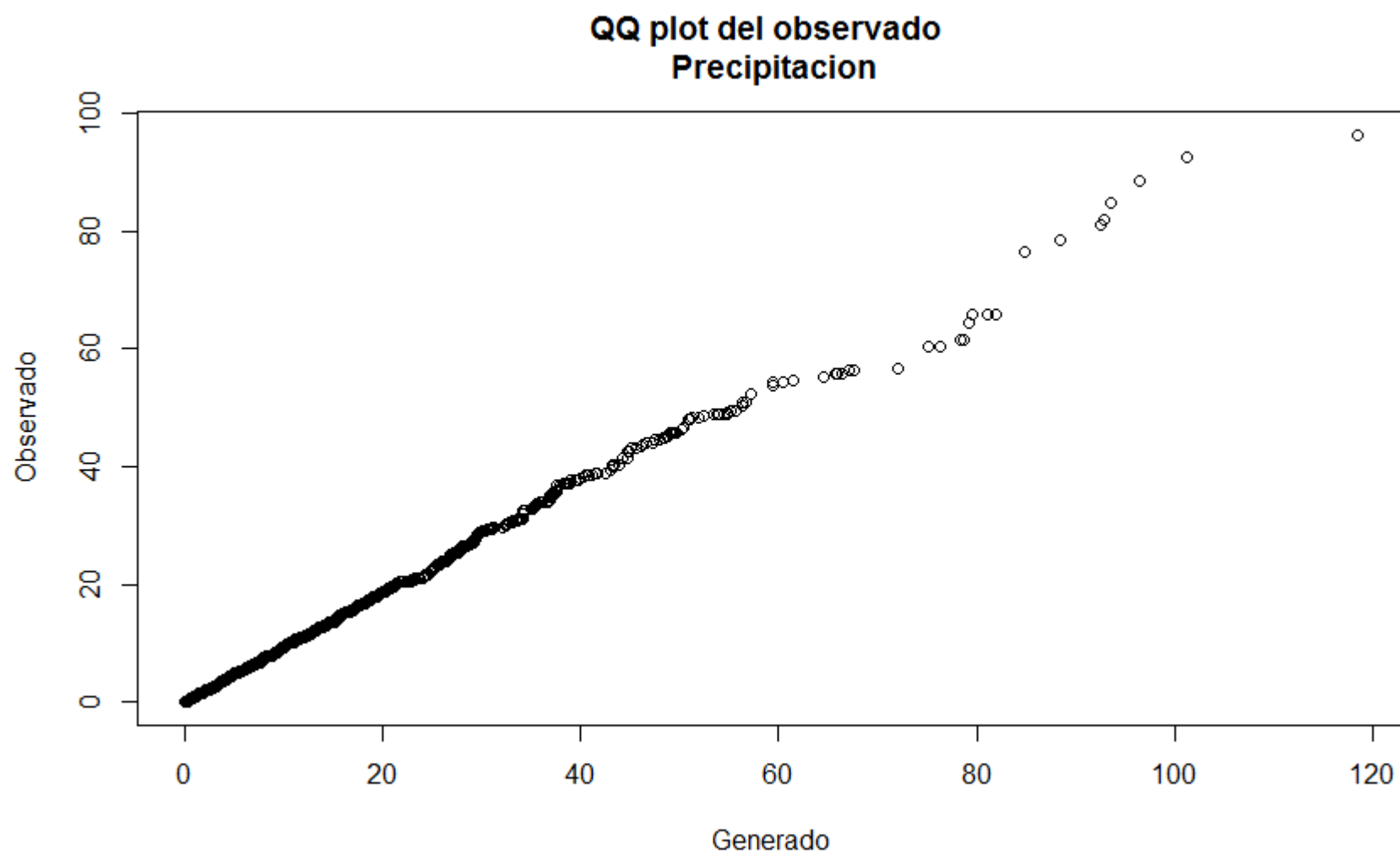
- 6. En R studio ir al menú File > Open File  
> DIA\_2 > script\_presentacion



# Ejemplo de Rmwagen





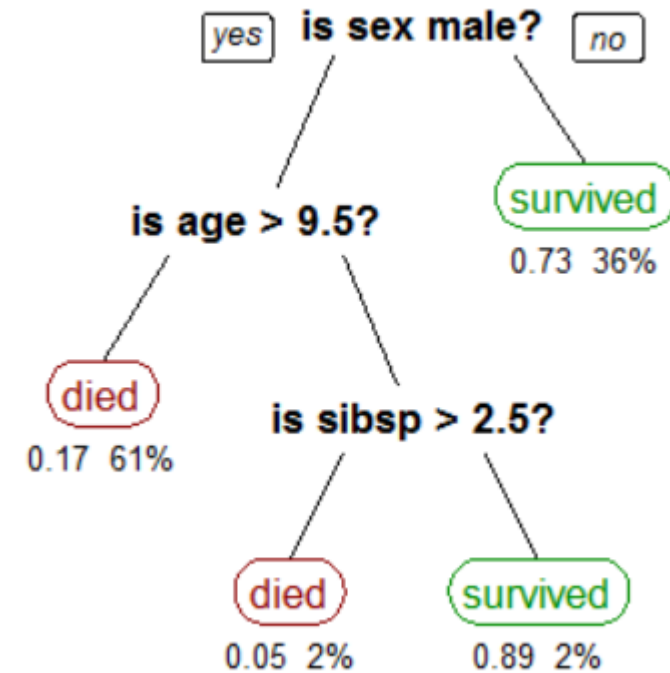


# Random Forest

Método de **regresión** y **clasificación**.

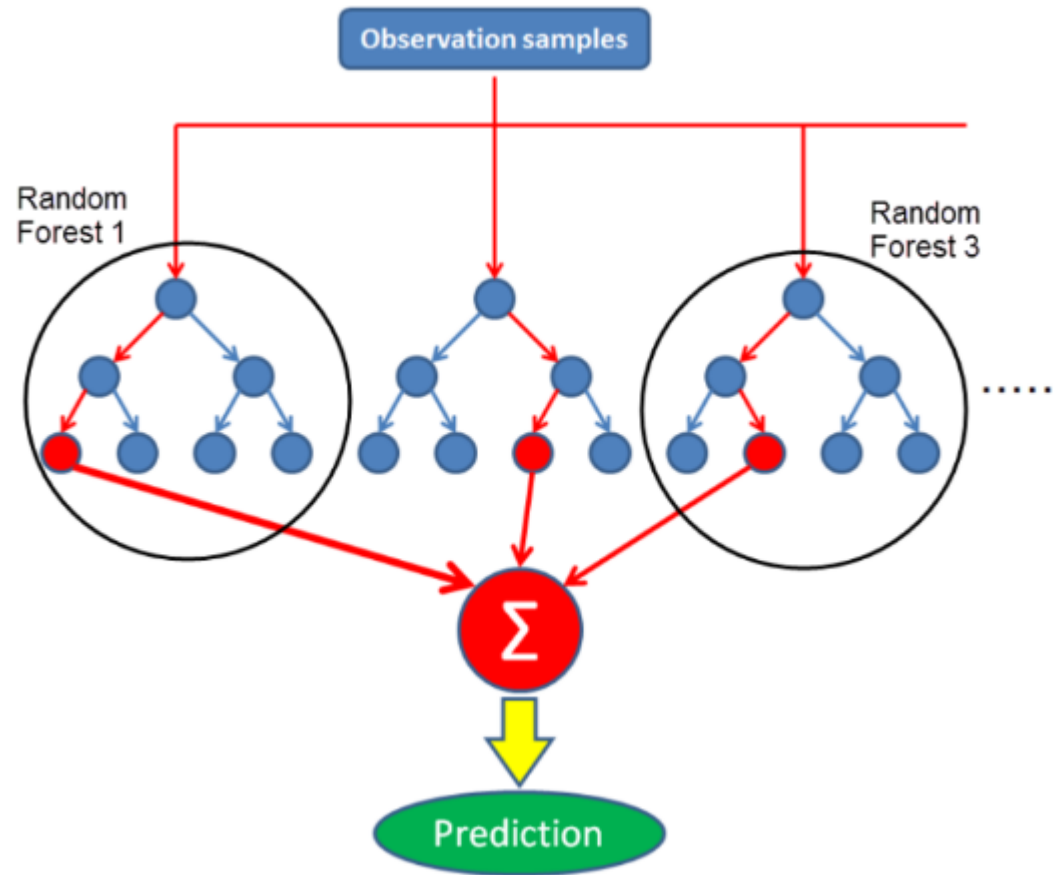
Esta basado en **CART** (Árboles de decision).

- Cada nodo corresponde a una variable de entrada.
- Ramas son los posibles valores que puede tomar la variable.
- Las ramas inferiores muestran la variable a predecir.



# Estructura Random Forest

- Cadena de Árboles aleatorios
- No correlacionados
- Combinados usando nodo optimización
- El resultado más votado será el ganador



```
Call:
randomForest(formula = Species ~ ., data = training, ntree = 100,
              = TRUE)
              mtry = 2, importance
```

Type of random forest: classification

Number of trees: 100

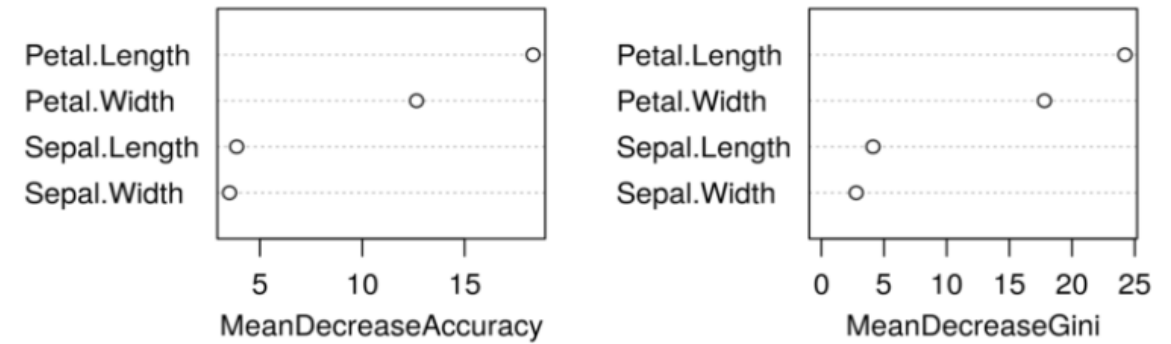
No. of variables tried at each split: 2

OOB estimate of error rate: 5.33%

Confusion matrix:

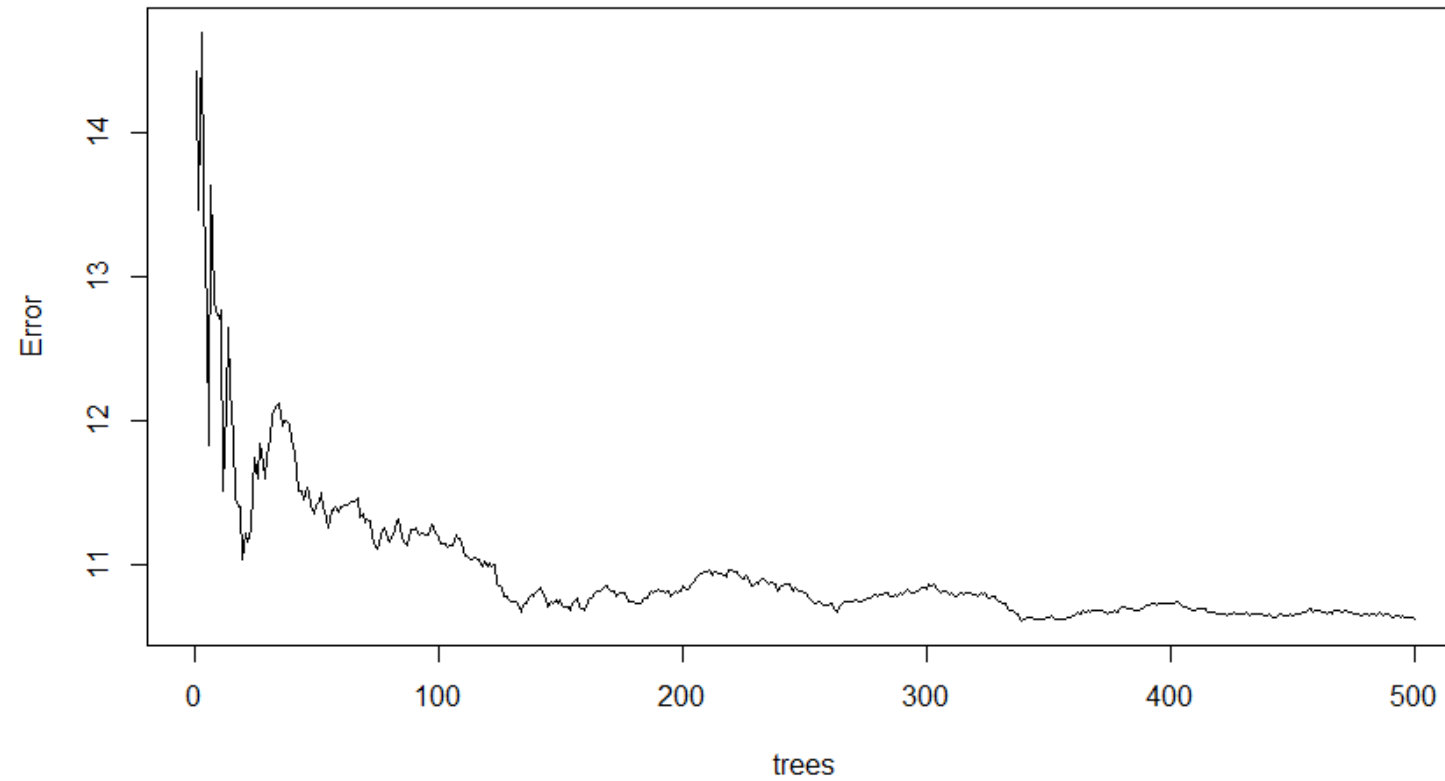
	setosa	versicolor	virginica	class.error
setosa	21	0	0	0.00000000
versicolor	0	25	2	0.07407407
virginica	0	2	25	0.07407407

rf\_classifier

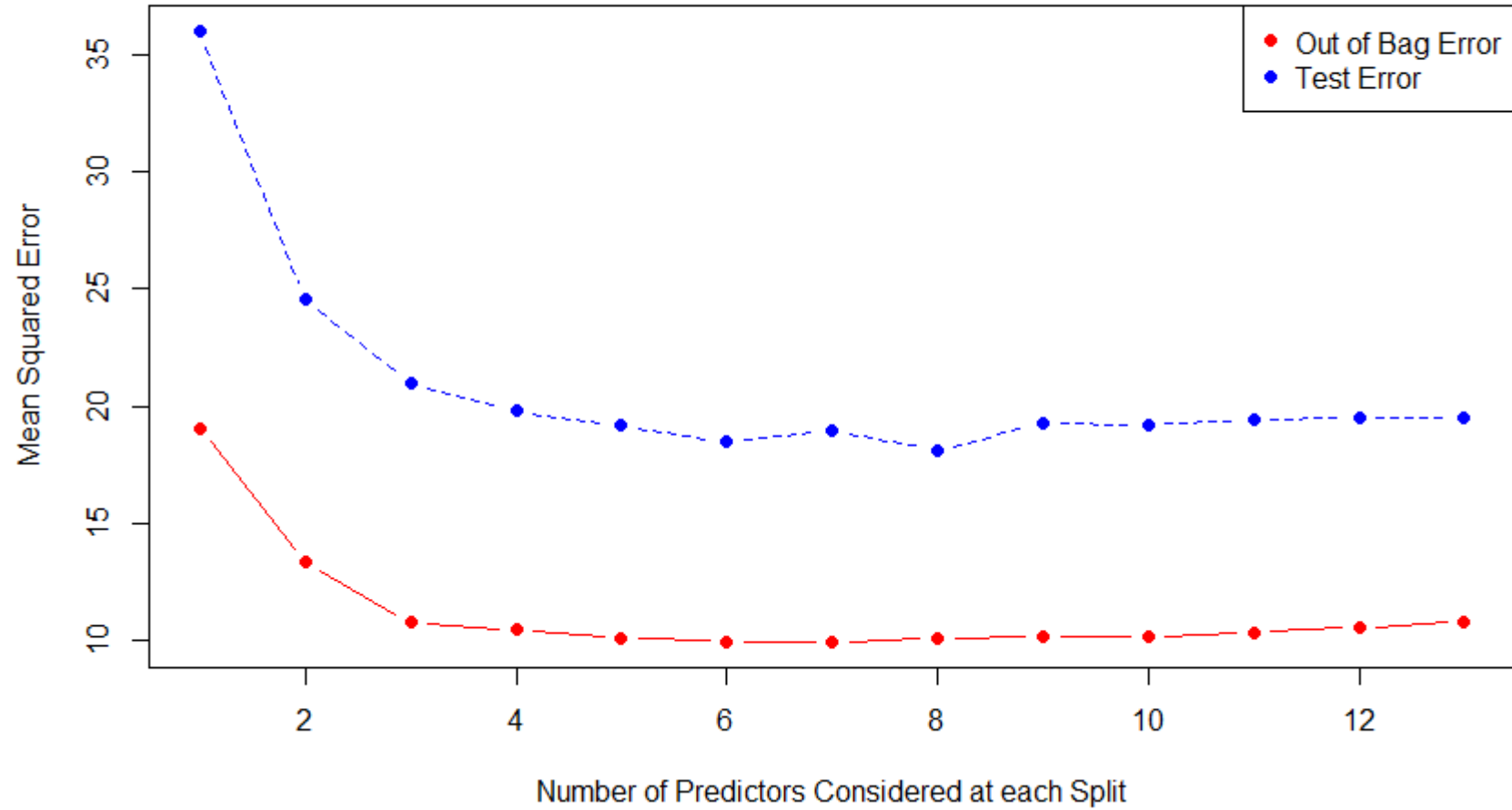




### Cantidad Optima de árboles



## Cantidad Óptima de variables



## Control de calidad

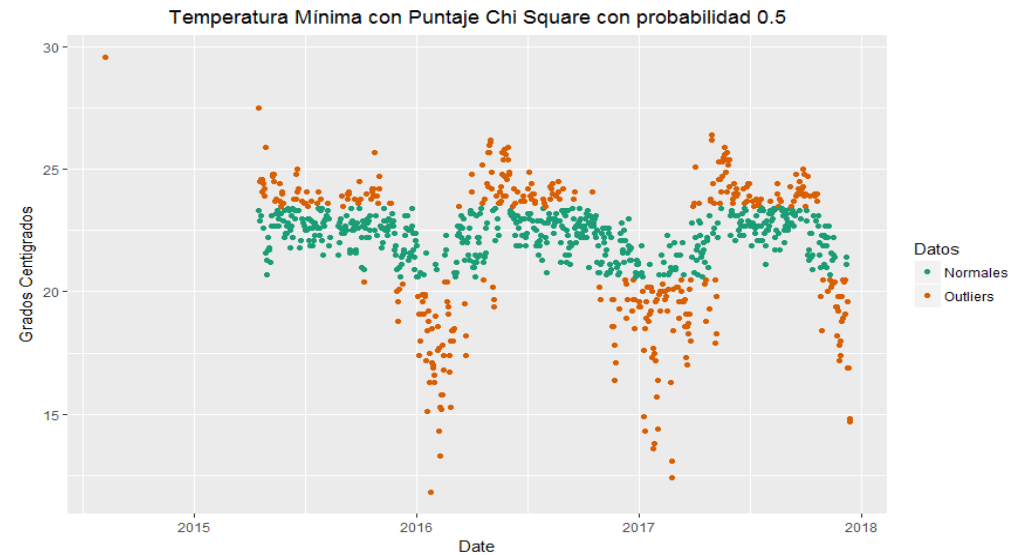
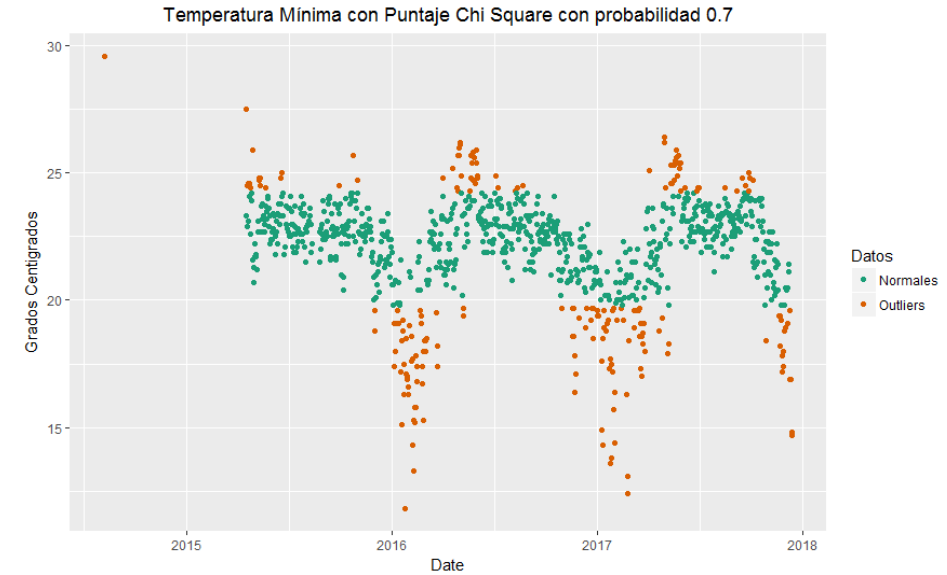
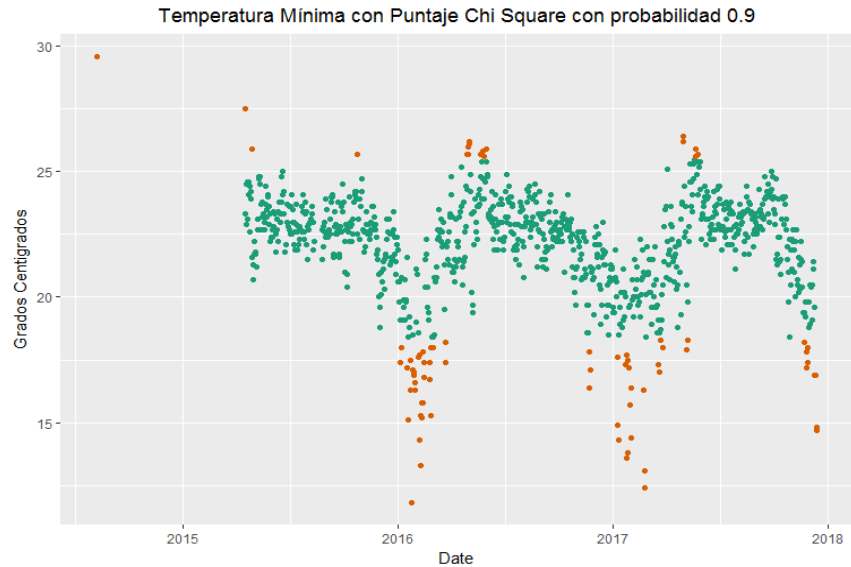
- Valores maximos y minimos para TX, TM, RH y SR .
- Para el valores maximo de SR se configuran de acuerdo a la WMO (World Metereroligal Organization) que es  $1600 \text{ WAM}^2$
- Latitud y longitud promedio de la zona de estudio (para posteriores calculos de horas de amanecer y anocheecer)
- La zona horaria de la zona de estudio

# Scores

## Chi Squared

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$O$  = observed score  
 $E$  = expected score





# Scores

z

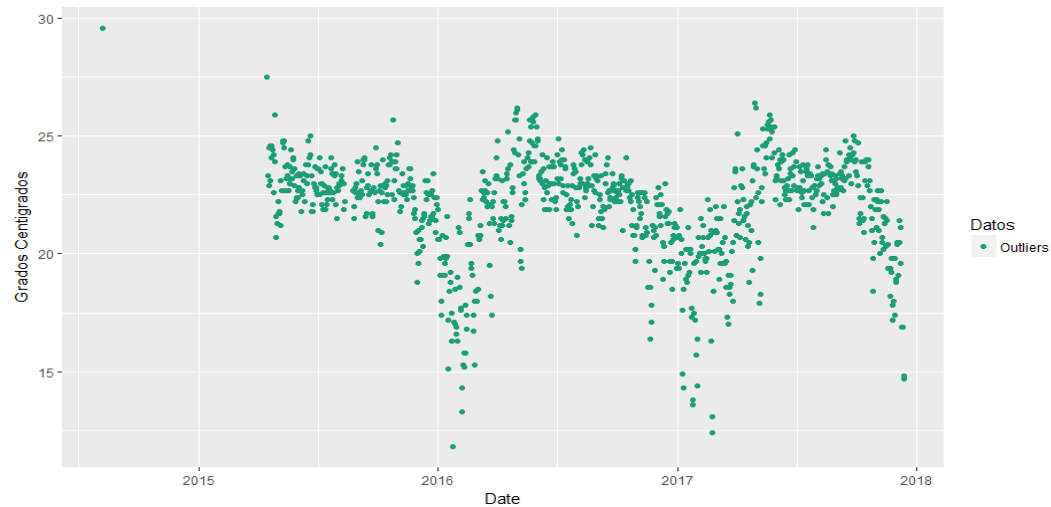
$$z = \frac{x - \mu_x}{\sigma_x}$$

score  $\rightarrow x - \mu_x$

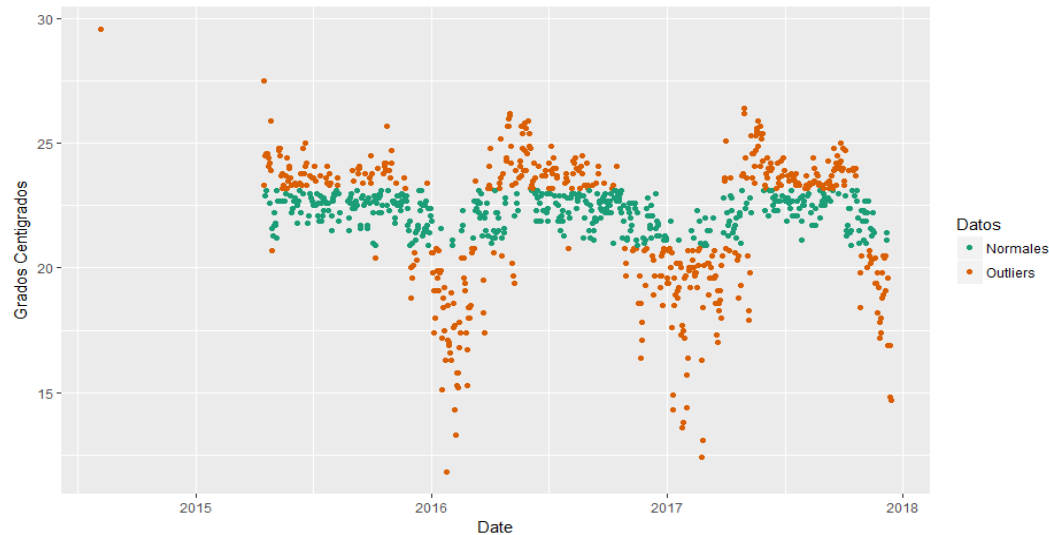
Population mean  $\rightarrow \mu_x$

Population standard deviation  $\rightarrow \sigma_x$

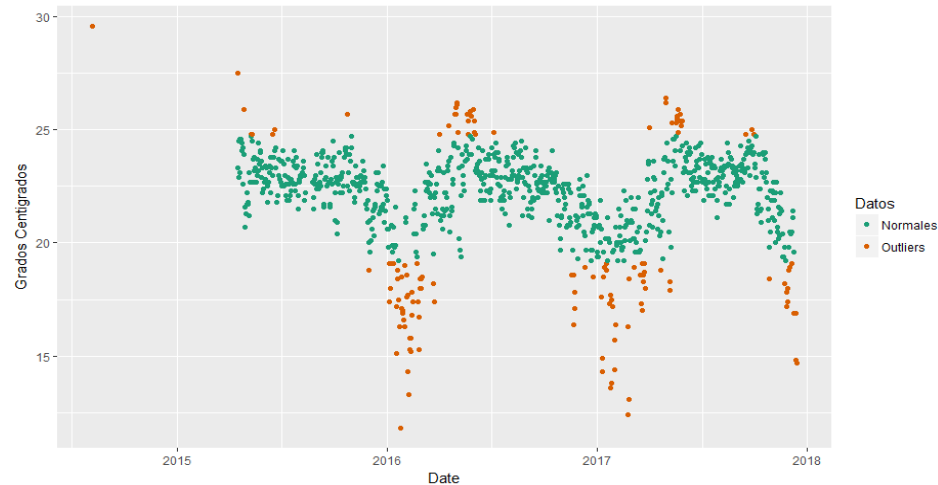
Temperatura Mínima z scored con probabilidad 0.5



Temperatura Mínima z scored con probabilidad 0.7



Temperatura Mínima z scored con probabilidad 0.9



# Análisis Multivariado

- Modelo de regresión.
- Distancia de Cook. Entre las variables más importantes.

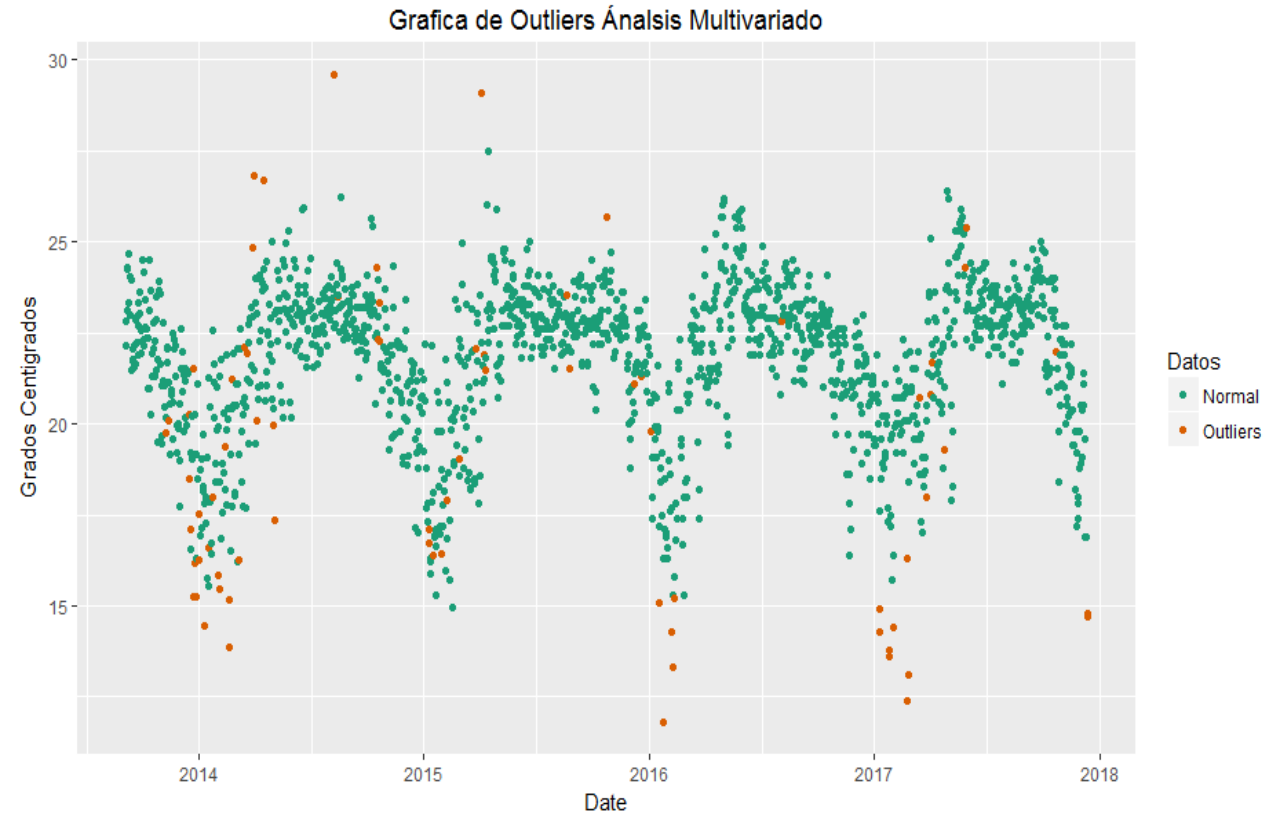
$$D_i = \frac{\sum_{j=1}^n \left( \hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{p \times MSE}$$

$\hat{Y}_j$  El valor en la posición  $j$  dado por el modelo cuando todas las observaciones son incluidas.

$\hat{Y}_{j(i)}$  El valor en la posición  $j$  dado por el modelo cuando no se incluye la posición  $i$ .

$p$  El número de coeficientes en el modelo de regresión.

$MSE$  Error cuadrático medio



# Thank you!



WE'RE PROUD TO  
HAVE CELEBRATED 50 YEARS  
OF AGRICULTURAL RESEARCH  
FOR DEVELOPMENT

## International Center for Tropical Agriculture - CIAT

Headquarters and Regional Office  
for South America and the Caribbean

+57 2 445 0000

Km 17 Recta Cali-Palmira  
A.A. 6713, Cali, Colombia

✉ [ciat.cgiar.org](mailto:ciat.cgiar.org)

🌐 [ciat.cgiar.org](http://ciat.cgiar.org)



CIAT is a CGIAR Research Center