Formateo de Datos

Formato de Entrada

- Los datos deben tener el siguiente formato

year	month	day	precip	tmax	tmin
1980	1	1		35.95	22.15
1980	1	2		35.3	22.8
1980	1	3		35.39	22.19
1980	1	4		35.54	22.54
1980	1	5		35.98	23.98
1980	1	6		36	22.6
1980	1	7		35.6	23
1980	1	8		34.41	21.81
1980	1	9		34.71	20.91
1980	1	10		35.4	22.7
1980	1	11		34.86	23.26
1980	1	12		35.47	22.07
1980	1	13		35.21	21.81

El archivo DEBE ser un CSV.

El número de archivos depende exclusivamente de la cantidad de estaciones a utilizar.

(n estaciones = n archivos)

NOTA

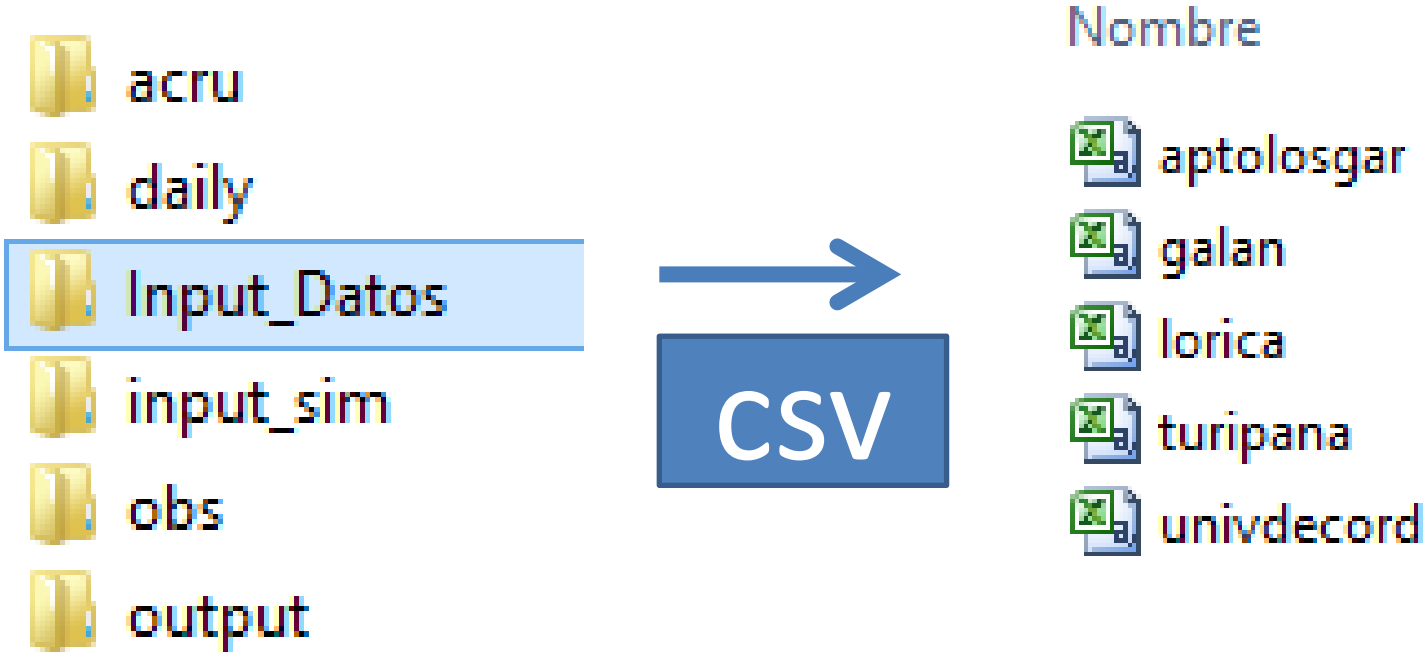
- ¿Puede una estación carecer de datos en ciertos períodos?

- **SÍ**

- **La metodología se encargará de rellenar la serie (de manera estadística)**

Pegado de Archivos

- Pegar los datos POR ESTACIÓN, en la carpeta Input_Datos (dentro del proyecto simgen)



Uso de script «Formateo_Datos_Pais»

```
#Esta primera parte es útil para determinar qué estaciones se van
# a utilizar, haciendo un análisis subjetivo y visual.
#-------------------------------------------------------

#archivos_pais en directorio
archivos_pais <- list.files("Input_Datos", pattern = ".csv")

#Indicar fecha donde aparece el primer dato observado (de T y Pp)
#-------------------------------------------------------------

#crear un Data Frame que almacene la información de cada estación (fecha en que apare
st_datainfo <- data.frame(matrix(nrow = length(archivos_pais), ncol = 7))
colnames(st_datainfo) <- c("station", "precip_first", "precip_last", "tmin_first", "t
st_datainfo$station <- archivos_pais
```

Se crea un data frame que almacena las fechas de inicio/término de las variables por estación.

Análisis de datos

- Ejecutadas las ~80 líneas, se obtiene lo siguiente: tabla (st_datainfo)

station	precip_first	precip_last	tmin_first	tmin_last	tmax_first	tmax_last
aptolosgar.csv	1980-10-1	2012-3-31	1980-1-1	2013-11-28	1980-1-1	2013-12-31
galan.csv	1980-1-1	2013-12-31	Completo	Completo	Completo	Completo
lorica.csv	1983-10-1	2013-12-31	Completo	Completo	Completo	Completo
turipana.csv	1981-1-1	2012-10-31	Completo	Completo	Completo	Completo
univdecord.csv	1980-6-20	2013-12-31	Completo	Completo	Completo	Completo

1er dato de Pp

último dato de Pp

1er dato de Tmin

Último dato de Tmin

1er dato de Tmáx

Último dato de Tmáx

Análisis de datos

- Una vez observada la tabla anterior, el usuario debe determinar el período común para TODAS LAS ESTACIONES.

- Tomar en cuenta que el set de datos debe comenzar el 1° de enero del año más «viejo» y finalizar el 31 de diciembre del año más «joven». (ej: 01/01/1960 – 31/12/2014)

- Si hay una estación con un período muy corto, se puede descartar y trabajar con las otras.

- Tomada la decisión, se debe editar cada uno de los archivos para que se establezcan dentro del mismo período.

Análisis de datos

- Corregido los archivos, se hace funcionar nuevamente el script. (filas 1 - ~80)

- Recomendable revisar la tabla st_datainfo nuevamente para asegurarse que esté todo dentro de lo pensado.
- Importante: las fechas no necesariamente van a coincidir, dado que pueden existir períodos iniciales sin datos.

Relleno de datos

- Correr filas ~100 a ~145.

```
100  #data frame donde se almacena la info de NA por estacion y variable
101  GapsEst <- data.frame(matrix(nrow = length(archivos_pais), ncol = 4))
102  GapsEst[,1] <- archivos_pais
103  colnames(GapsEst) <- c("Station", "Pp", "Tmax", "Tmin")
104
105
106  #EstFiltroImp <- list()
107  for (j in 1:length(archivos_pais)){
108
109    print(paste("procesando", archivos_pais[j]))
110
111    stat_for <- read.delim(paste(getwd(), "/Input_Datos/", archivos_pais[j], sep
112
113    #Porcentaje de Gaps
114    #------------------
115    GapsEst[j, 2:4]<- apply(stat_for[4:6], 2, function(x) sum(is.na(x)/length(x)
116
117
118    #impute
119    #------
120    stat_imp <- missForest(stat_for[,4:6])
121    stat_imp <- stat_imp$ximp
122
```

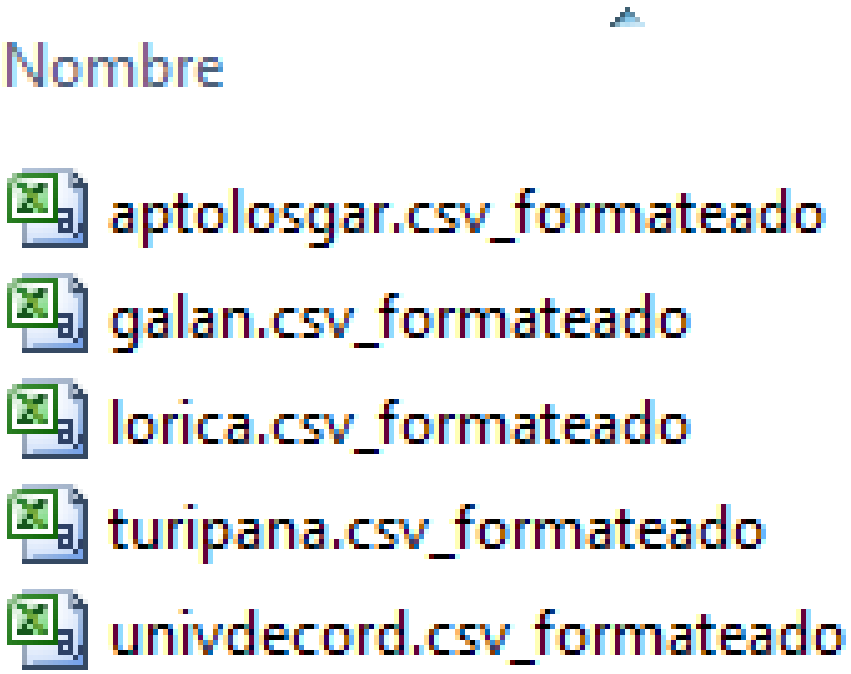
Relleno de datos

- Se van a generar 2 productos:
- Primero: GapsEst, que es un data frame en donde se puede ver el porcentaje de datos rellenados (gaps/(longitud de registro)).

	Station	Pp	Tmax	Tmin
1	aptolosgar.csv	8.897657	39.69724	41.78275
2	galan.csv	16.249295	0.00000	0.00000
3	lorica.csv	12.690233	0.00000	0.00000
4	turipana.csv	9.010387	0.00000	0.00000
5	univdecord.csv	13.084789	0.00000	0.00000

Relleno de Datos

- Segundo: almacenamiento de datos en carpeta daily.



- Fin
list
03_

year	yday	t_min	t_max	precip	solar	date
1971	121	10.0	15.3	0.0	4.7	1971-05-01
1971	122	4.4	16.7	0.0	4.6	1971-05-02
1971	123	2.8	14.3	0.0	4.6	1971-05-03
1971	124	3.7	13.4	0.0	4.5	1971-05-04
1971	125	3.7	13.4	0.0	4.5	1971-05-05
1971	126	7.3	16.0	0.0	4.4	1971-05-06
1971	127	10.4	14.3	0.0	4.4	1971-05-07
1971	128	8.8	17.8	0.0	4.4	1971-05-08
1971	129	8.0	17.3	0.0	4.3	1971-05-09
1971	130	7.3	20.2	0.0	4.3	1971-05-10
1971	131	5.0	21.2	0.0	4.3	1971-05-11
1971	132	6.3	21.3	0.0	4.3	1971-05-12
1971	133	8.3	18.3	0.0	4.2	1971-05-13
1971	134	3.0	18.3	0.0	4.2	1971-05-14

- 01_cmip5_corr_and_param
- 02_Formateo_Datos_Pais
- 03_prep_ntcc_Pais