

Proyecto Final-Temperaturas

Luraghi-Rodriguez

June 18, 2018

BORRADOR

INTRODUCCION

OBJETIVO:

Relizar un análisis exploratorio de los datos utilizados para un proyecto cuyo objetivo general es la modelización y predicción de las temperaturas mínimas extremas en Uruguay, utilizando el enfoque de la teoría de valores extremos.

La modelización de los eventos extremos climáticos resulta de particular interés en la actualidad, debido al gran impacto que estos fenómenos producen tanto en la población como en los sectores productivos. Estos últimos, que resultan especialmente sensibles a la variabilidad climática, adolecen en general de un conocimiento confiable sobre la ocurrencia de sucesos extremos y les es imprescindible asignarles ciertas probabilidades de ocurrencia.

Uno de los objetivos específicos del proyecto sobre valores extremos consiste en comparar dos metodologías para la modelización, obteniendo predicciones de los niveles de retorno mediante el Método de Valores Extremos por Bloques y el Método del Umbral. Además, se busca contrastar los resultados obtenidos mediante el uso de los diversos paquetes del software estadístico R actualmente disponibles, por lo tanto realizaremos también la exploración de las bases de datos de valores extremos utilizadas para cada uno de los dos métodos mencionados.

DATOS

La base de datos original está compuesta por registros diarios de temperaturas mínimas de 26 estaciones meteorológicas de Uruguay para el período 2002-2014. Los datos están comprendidos entre el 1o de enero de 2002 y el 31 de diciembre de 2014, lo cual implica un total de 4.526 observaciones por estación (118664 observaciones en total).

La base incluye las siguientes variables:

- nroEstacion: Número de Estación
- lon: Longitud en la cual se encuentra ubicada la Estación
- lat: Latitud en la cual se encuentra ubicada la Estación
- altitud: Altura de la estación respecto al nivel del mar
- anio: Año en el cual se registró la temperatura
- mes: Mes en el cual se registró la temperatura
- dia: Día en el cual se registró la temperatura
- tmin: valor en grados celsius de la temperatura mínima del día.
- modis1: Temperatura registrada por satélite
- modis2: Temperatura registrada por satélite

Sin embargo para la realización del proyecto las variables a utilizar son:

- Número de estación
- Fecha: para tener un orden cronológico, creamos esta nueva variable a partir de las variables año, mes y día
- Temperatura mínima registrada

- Coordenadas geográficas para un análisis espacial.
- Departamento: asignada de acuerdo a la ubicación geográfica (latitud, longitud) de cada estación.
- Creamos una variable con el nombre de cada estación.

ANALISIS EXPLORATORIO DE DATOS

En este apartado se presentan los resultados del análisis exploratorio realizado para comprender los datos a efectos de poder aplicar teoría de extremos.

El primer punto a tener en cuenta es la existencia de estacionalidad. Sabemos que la temperatura tiene un comportamiento estacional, es decir todos los años la temperatura se comporta de la misma manera (temperaturas altas en verano, bajas en invierno). Lo corroboramos con el siguiente gráfico:

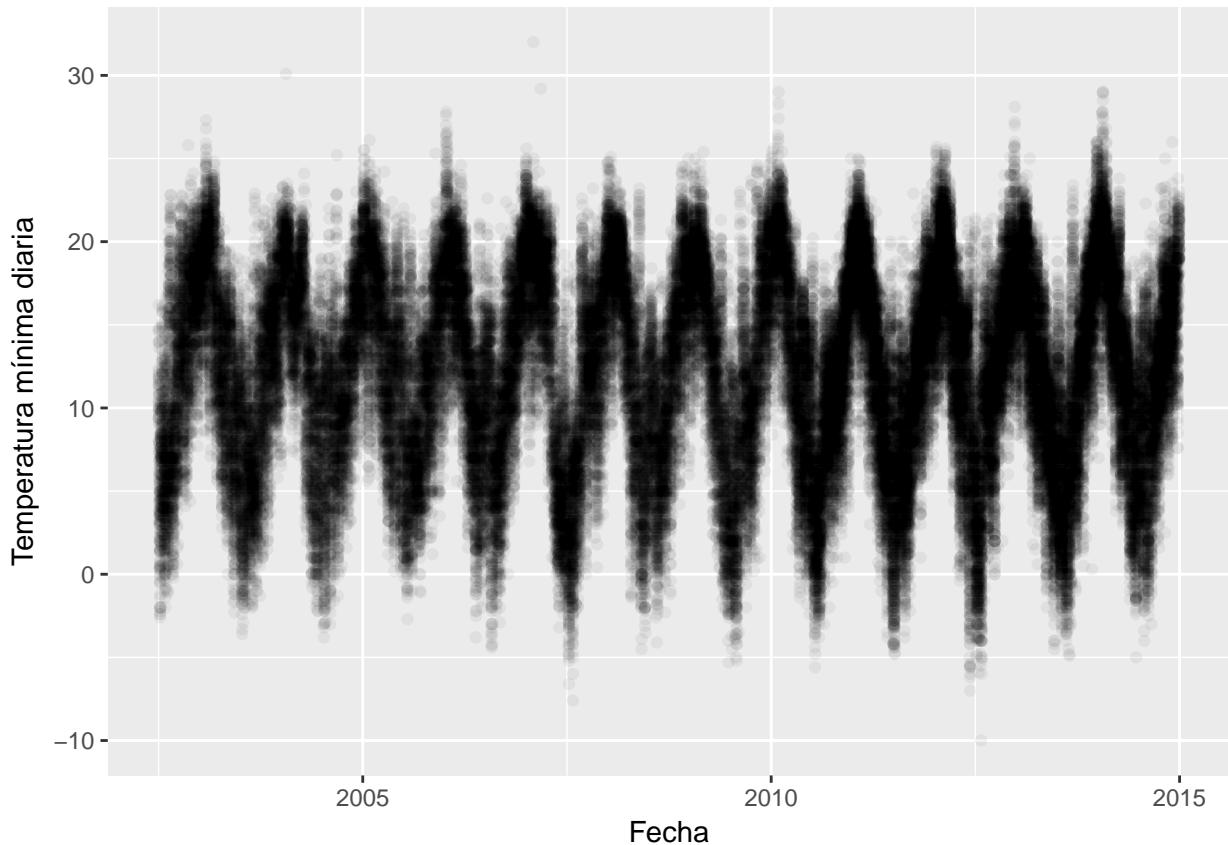


Figure 1: Graficamos los registros de temperatura de las 26 estaciones. Como es esperado, se observa una gran presencia de estacionalidad. Claramente en cada invierno las temperaturas mínimas son las más bajas, en verano las más altas y en primavera y otoño nos ubicamos en la franja media.

De acuerdo a la Figura @ref(Figure 1), parecería que la media es constante. Calculemos la media de las temperaturas mínimas para cada año:

	Año	Media
1	2002	11.89
2	2003	12.05
3	2004	12.39
4	2005	12.41
5	2006	12.42
6	2007	11.95
7	2008	12.45
8	2009	12.15
9	2010	11.65
10	2011	11.88
11	2012	12.68
12	2013	11.73
13	2014	12.79

Como se puede observar en la tabla, la temperatura mínima media año a año no sufre mucha variación, por lo que podríamos afirmar que es constante

Analizaremos como es el comportamiento de la temperatura dentro del año para una estación en particular. Elegimos la estación número 2, la cual corresponde a la estación de Melilla, Canelones.

De acuerdo al gráfico 3 no se observan indicios de existencia de una tendencia, es decir que al pasar de los años haya un corrimiento de las funciones (calentamiento o enfriamiento global)

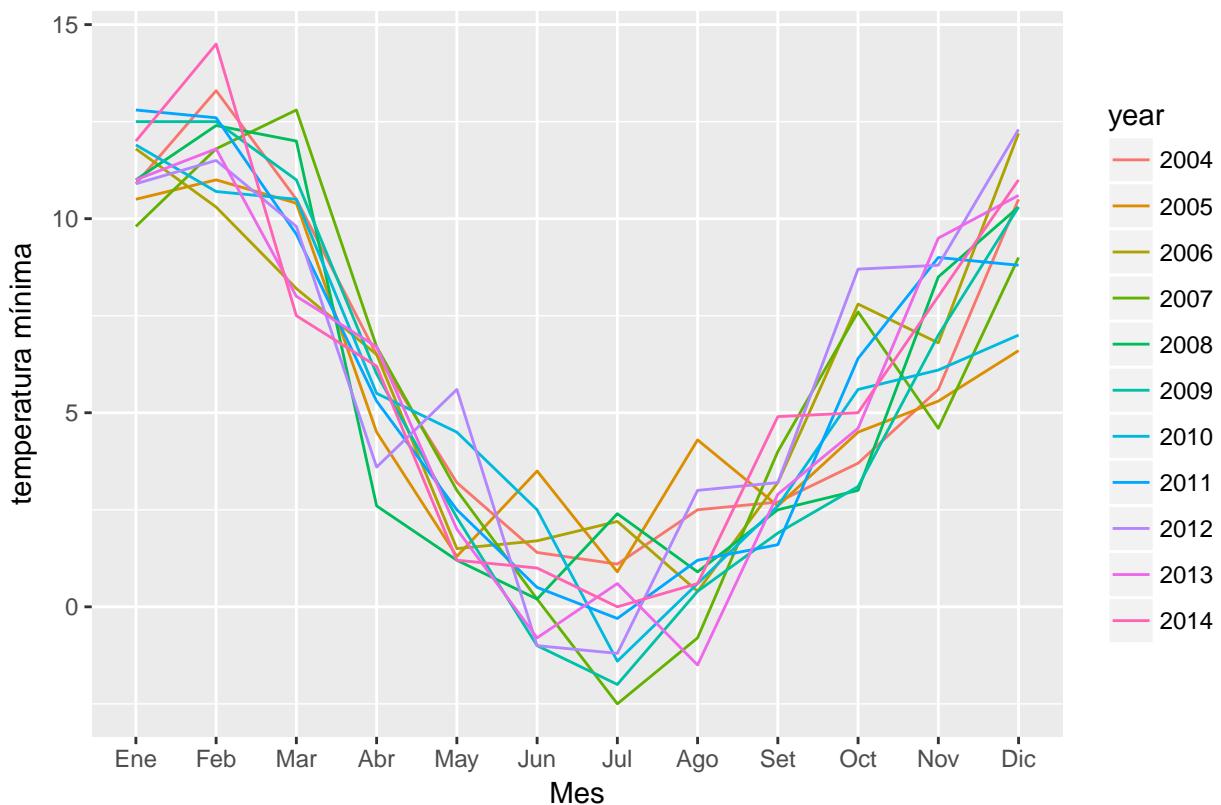


Figure 2: Comportamiento de las temperaturas mínimas para la estación Melilla, Canelones en el período 2004-2014

Análisis de Outliers

Nuestro interés son las temperaturas extremas, por lo cual veremos si tenemos valores atípicos entre los meses de mayo y setiembre.

De la Figura 4 surge que tenemos outliers en varias estaciones, pero como lo que nos interesa es estudiar las temperaturas mínimas extremas sólo miraremos a las estaciones que presentan outliers en los mínimos, es decir las estaciones Florida, Prado y Punta del Este.

Outlier de la estación Florida:

	nroEstacion	tmin	fecha
1	7	-10.00	2012-07-28

La comparamos con la estación Durazno que se encuentre geográficamente cerca

nroEstacion	tmin	fecha	Nombre
1	6	-3.00	2012-07-28 Santa Bernardina,Durazno

Outlier para la estación Prado:

Lo comparamos con la estación Melilla de Canelones

Outliers de la estacion Punta del Este

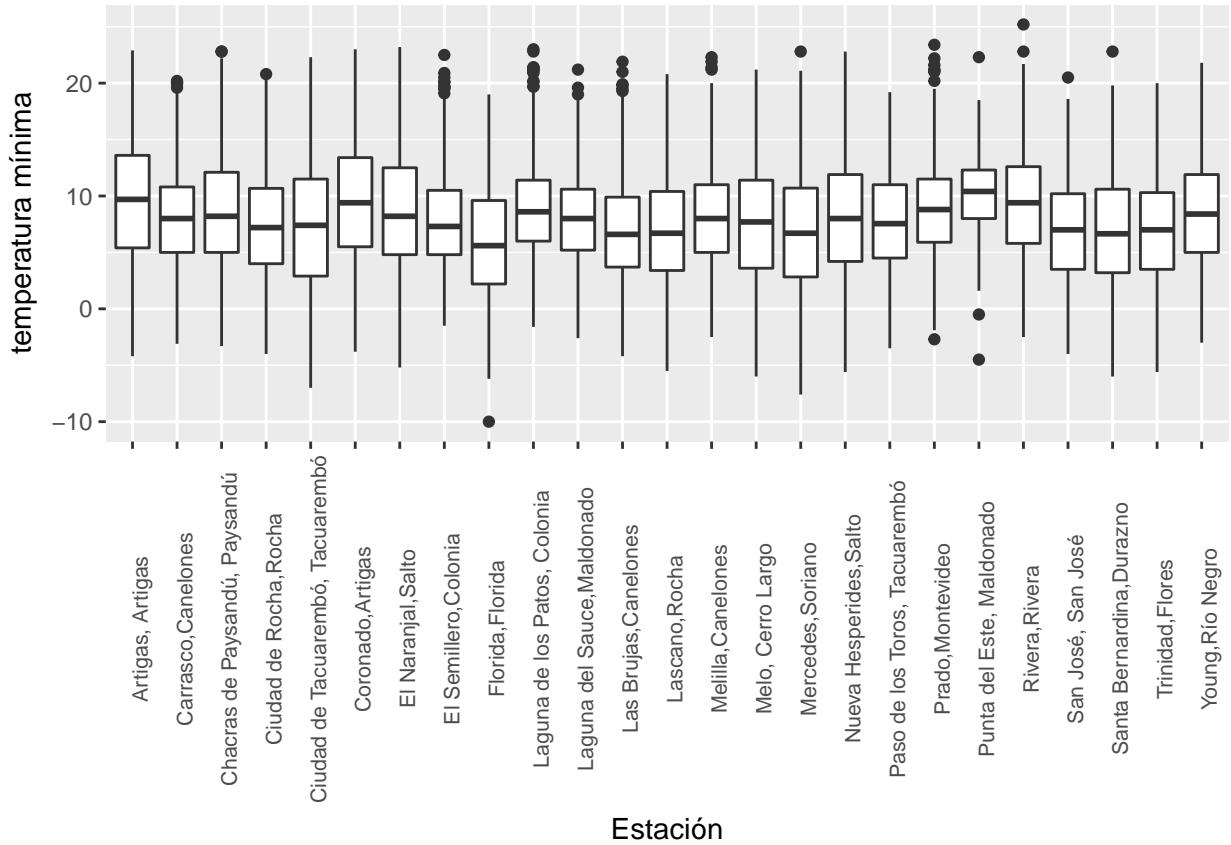


Figure 3: Boxplot de temperaturas mínimas para cada estación

nroEstacion	tmin	fecha
1	-2.70	2009-07-31

Comparamos con la otra estación Laguna del Sauce,Maldonado

Temperaturas mínimas registradas para cada estación -ubicación geográfica(sur-centro-norte). Se agrega una variable mas

detallar que departamentos son parte de cada grupo.

hay correlación entre las temperaturas mínimas y la altitud?

Plot de temperatura mínima registrada en cada año, (puede ser un plotly) coloreado por zona de estación. La idea es explorar si los mínimos siempre se dan en la misma zona o no. si hay una correlación.

Exploración espacial:

Se presentan dos mapas, uno en el cual se ubican las estaciones seleccionadas y otro que muestra las temperaturas mínimas por mes en cada departamento coloreado por las temperaturas.

Metodología.

El interés de la Teoría de valores extremos está centrado en modelar el comportamiento de $M_n = \max\{X_1, X_2, \dots, X_n\}$ siendo X_1, X_2, \dots, X_n una secuencia de variables aleatorias independientes con

nroEstacion	tmin	fecha	Nombre
1	2	-2.00 2009-07-31	Melilla,Canelones

nroEstacion	tmin	fecha
1	14	-0.50 2013-07-22
2	14	-4.50 2013-08-23

distribución F, y M_n representa el máximo del proceso sobre n unidades de tiempos de observación.

En Teoría la distribución de M_n podría calcularse de manera exacta a partir de la función de distribución F:

$$P(M_n \leq z) = P(X_1 \leq z, \dots, X_n \leq z) = P(X_1 \leq z) \times \dots \times P(X_n \leq z) = [F(z)]^n$$

Sin embargo, al sustituir F (desconocida) por su estimación y elevar a la n, pueden generarse grandes discrepancias.

Un camino alternativo a este problema, es considerar F desconocida y aproximar directamente la distribución de F^n utilizando un subconjunto de la base, los valores extremos.

Esta solución, análoga al Teorema Central del Límite, consiste en la búsqueda de secuencias constantes $\{b_n; n \geq 1\}$ y $\{a_n; n \geq 1\}$ tales que la distribución de $M_n^* = \frac{M_n - b_n}{a_n}$ converge a una distribución no degenerada cuando $n \rightarrow +\infty$.

Teorema de Valores extremos:

Si existen sucesiones constantes $a_n > 0$ y b_n tales que:

$$P(M_n^* = \frac{M_n - b_n}{a_n} \leq z) = F^n(a_n z + b_n) \rightarrow G(z)$$

entonces G pertenece a una de las siguientes familias:

El Teorema (1.1) afirma que M_n^* converge en distribución a una de las tres familias de distribuciones mencionadas en el teorema las que, en conjunto, se denominan distribuciones de valores extremos (DVE). En particular, (a) es la distribución de Gumbel, (b) es la distribución de Fréchet y (c) es la distribución de Weibull. Cada familia tiene un parámetro de ubicación y escala, μ y σ respectivamente, y un parámetro de forma α en el caso de las familias Fréchet y Weibull.

Si bien el Teorema (1.1) no garantiza la existencia de un límite no degenerado para M_n , ni nos dice cuál es el límite cuando existe, la distribución límite de la variable normalizada M_n^* tiene que ser alguna de las distribuciones incluidas en el teorema, cualquiera sea la distribución poblacional F. \

Los tres tipos de distribución del teorema pueden ser combinados en una sola distribución con una parametrización común, propuesta por Von Mises (1954) y Jenkinson (1955), que se conoce como la **Distribución de Valores Extremos Generalizada** (GEV, por sus siglas en inglés). La forma de esta distribución es:

$$G_{\xi, \mu, \sigma}(x) = \exp \left\{ - \left(1 + \xi \left(\frac{x - \mu}{\sigma} \right)_+^{-1/\xi} \right) \right\}$$

En los siguientes métodos asumiremos que la función F (distribución de los datos) es desconocida, y estimaremos directamente la distribución extrema, para esto intentamos captar en una nueva base aquellos valores extremos sobre los cuales nos interesa trabajar.

Método del umbral: Consiste en seleccionar un valor u llamado umbral, y considerar aquellas observaciones que superen este valor. La elección de un umbral muy bajo aumentará el sesgo del modelo, mientras que la elección de un umbral demasiado alto aumentará la varianza.

nroEstacion	tmin	fecha	Nombre
1	16	4.20	2013-07-22 Ciudad de Rocha,Rocha
2	16	3.20	2013-08-23 Ciudad de Rocha,Rocha
	Nombre	tmin	fecha
1	Artigas, Artigas	-4.20	2011-07-04
2	Carrasco,Canelones	-3.10	2007-07-29
3	Chacras de Paysandú, Paysandú	-3.30	2012-06-07
4	Ciudad de Rocha,Rocha	-4.00	2007-07-29
5	Ciudad de Tacuarembó, Tacuarembó	-7.00	2012-06-09
6	Coronado,Artigas	-3.80	2012-06-08
7	El Naranjal,Salto	-5.20	2007-07-12
8	El Semillero,Colonia	-1.50	2012-07-11
9	Florida,Florida	-10.00	2012-07-28
10	Laguna de los Patos, Colonia	-1.60	2007-07-13
11	Laguna del Sauce,Maldonado	-2.60	2007-07-29
12	Las Brujas,Canelones	-4.20	2009-07-31
13	Lascano,Rocha	-5.50	2012-06-09
14	Melilla,Canelones	-2.50	2007-07-11
15	Melo, Cerro Largo	-6.00	2012-06-09
16	Mercedes,Soriano	-7.60	2007-07-29
17	Nueva Hesperides,Salto	-5.60	2012-06-09
18	Paso de los Toros, Tacuarembó	-3.50	2012-07-30
19	Prado,Montevideo	-2.70	2009-07-31
20	Punta del Este, Maldonado	-4.50	2013-08-23
21	Rivera,Rivera	-2.50	2007-07-12
22	San José, San José	-4.00	2012-07-30
23	Santa Bernardina,Durazno	-6.00	2007-07-29
24	Trinidad,Flores	-5.60	2010-07-16
25	Young,Río Negro	-3.00	2012-06-07

Método “Block Maxima”: Útil cuando los datos pueden ser divididos en m bloques de tamaño m, el método consiste en seleccionar la temperatura extrema por bloque, asumiendo que las observaciones de los máximos por bloque son independientes. La elección del tamaño del bloque genera un trade-off entre sesgo y varianza, tomar bloques pequeños generan un mayor sesgo mientras que bloques de mayor tamaño aumentan la varianza en la estimación de los parámetros de la distribución.

Descripción de la aplicación shiny

A través de la aplicación Shiny, mostraremos el análisis exploratorio de los datos, aparecerán gráficos de la serie de temperaturas, pudiendo filtrar por número de estación, y pudiendo ubicar geográficamente las mismas. Para todo el territorio uruguayo se podrá observar por mes las temperaturas mínimas registradas en cada departamento.

Comentarios finales

Referencias

-cole -R -librerías