Temperaturas mínimas del Uruguay

Lorena Luraghi & Carolina Rodriguez

June 25, 2018

1 Introduction

OBJETIVO:

Relaizar 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. A modo de ejemplo, golpes de calor o frios extremos pueden afectar negativamente a las personas en situación de calle, o por otro lado arruinar cosechas enteras, motores de aviones, etc. A pesar de la gran importancia de estos eventos, se adolece 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. Realizaremos también la exploración de las bases de datos de valores extremos utilizadas para cada uno de los dos métodos mencionados.

2 Metodología

El interés de la Teoría de valores extremos está centrado en modelar el comportamiento de

$$M_n = max\{X_1, X_2, \ldots, X_n\} \tag{1}$$

siendo $X_1, X_2, \ldots, X_n$ una secuencia de variables aleatorias independientes con distribución F, y M_n representa el maximo 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:

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

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, para esto intentamos captar en una nueva base aquellos valores extremos sobre los cuales nos interesa trabajar.

Hay dos posibles métodos populares para la selección de la base mencionada:

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 eleccion de un umbral demasiado alto aumentará la varianza.

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 estimacion de los parametros de la distribución.

2.1 Teorema de Valores extremos para el método del umbral

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

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

entonces G pertenece a una de las siguientes tres familias de distribuciones: Weibull, Gumbel o Fréchet.

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). 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:

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

2.2 Teorema de valores extremos para el método de Block Maxima

De forma análoga que para el método del umbral, existe un teorema para el método Block Maxima, sin embargo la diferencia con el teorema anterior, es que ahora el ajuste de los datos se realizará mediante una distribución de Pareto Generalizada.

3 Base de 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

Nuevas variables creadas:

- Departamento: se crea esta variable a partir de las coordenadas gegráficas de cada estación
- Nombre: le asignamos a cada estación un nombre.
- Zona: Clasificamos a cada estación en tres zonas(Centro, Norte, Sur) de acuerdo al departamento al que pertenece.
- Fecha:para tener un orden cronológico, creamos esta nueva variable a partir de las variables año, mes y día

Para la realización del proyecto las variables a utilizar son:

- Número de estación
- Fecha

- Temperatura mínima registrada
- Departamento
- Nombre
- Zona

4 Análisis exploratorio de los 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:

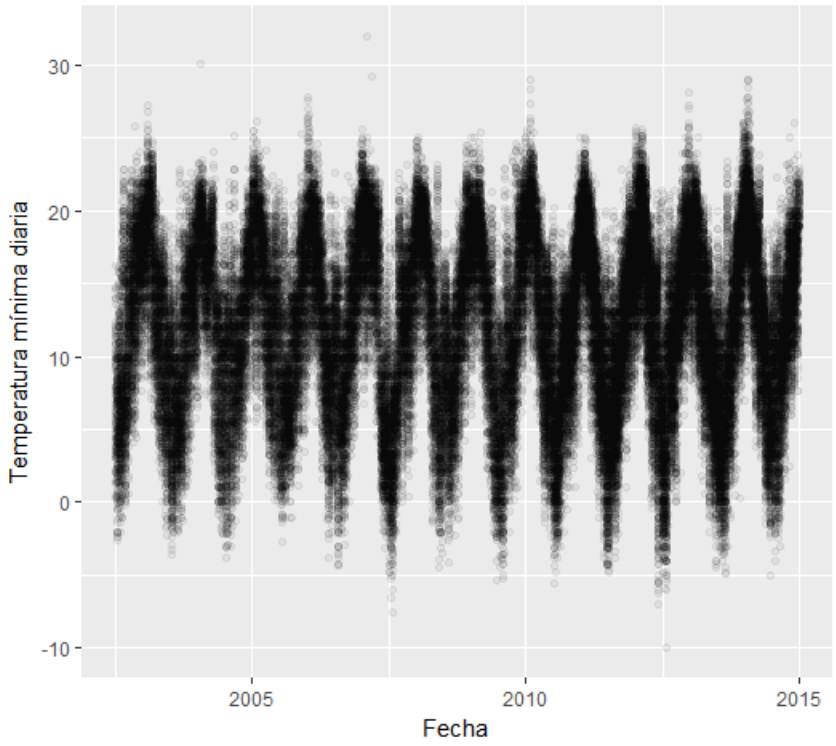


Figure 1: Graficamos los registros de temperatura de las 26 estaciones. Como es esperado, se observa una gran presencia de estacionalidad. Claramente en cada invierno las temperaturas mínimas son las más bajas, en verano las más altas y en primavera y otoño nos ubicamos en la franja media.

De acuerdo a la Figura 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

Analizamos los datos faltantes:

	N°Estacion	porcentaje NA
1	3	0.10
2	12	0.10
3	15	0.20
4	10	0.30
5	25	0.30
6	9	0.40
7	16	0.40
8	17	0.60
9	2	0.90
10	4	3.60
11	5	5.60
12	1	6.90
13	13	16.50
14	6	22.60
15	8	38.20
16	20	39.60
17	22	60.40
18	11	62.70
19	19	63.40
20	18	63.50
21	21	63.70
22	7	63.90
23	14	66.70

Las estaciones que presentan menos datos faltantes son la estación 3 (Artigas) y la 12 (Chacras de Paysandu).

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.

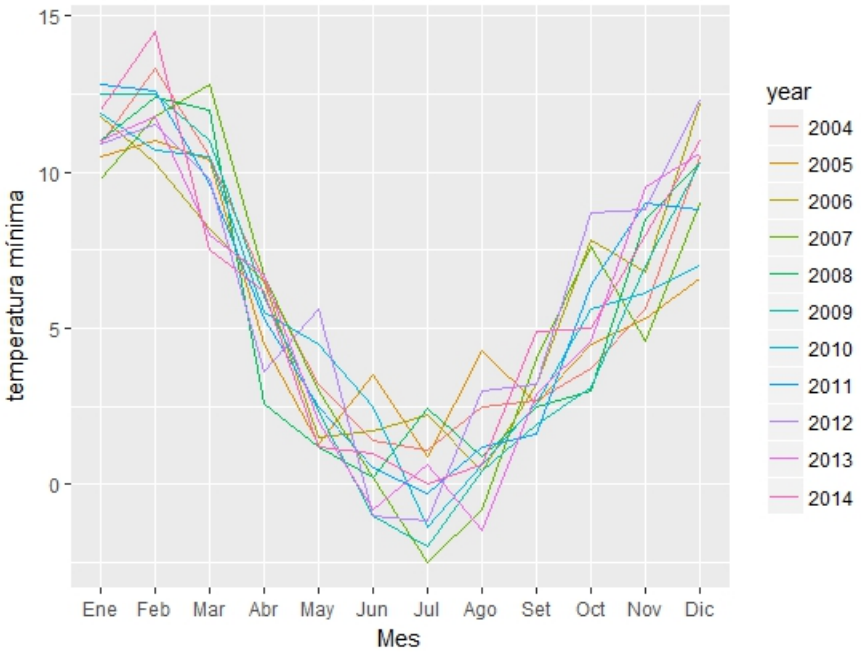


Figure 2: Comportamiento de las temparaturas mínimas para la estación Melilla, Canelones en el período 2004-2014

De acuerdo a la Figura 2 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)

4.1 Análisis de outliers:

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

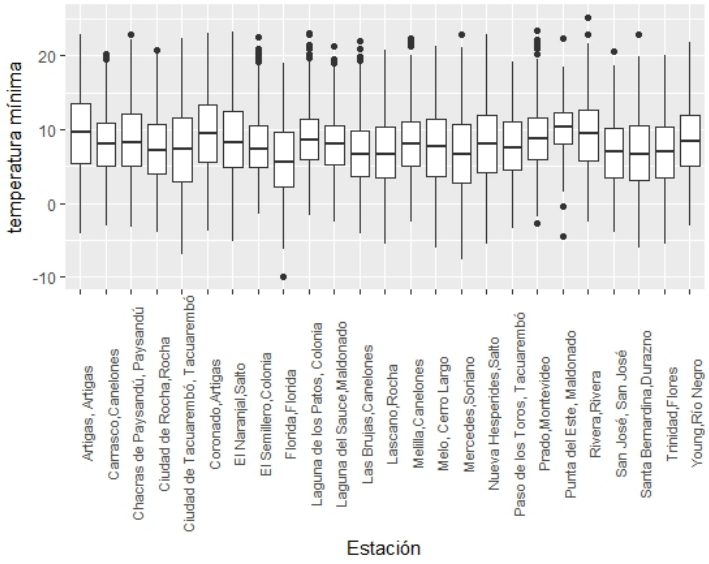


Figure 3: Boxplot de temperaturas mínimas para cada estación.

De la Figura 3 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 Floria, 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:

	nroEstacion	tmin	fecha
1	13	-2.70	2009-07-31

Lo comparamos con la estación Melilla de Canelones:

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

Outliers de la estación Punta del Este:

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

Comparamos con la otra estación Laguna del Sauce,Maldonado:

	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

Queremos ver si hay algún año en particular en el que se cumpla que todas las estaciones registraron su mínimo del periodo:

	Nombre	tmin	fecha	nroEstacion
1	El Semillero,Colonia	-1.50	2012-07-11	24
2	Laguna de los Patos, Colonia	-1.60	2007-07-13	5
3	Melilla,Canelones	-2.50	2007-07-11	2
4	Rivera,Rivera	-2.50	2007-07-12	15
5	Laguna del Sauce,Maldonado	-2.60	2007-07-29	8
6	Nueva Hesperides,Salto	-2.60	2002-07-10	20
7	Prado,Montevideo	-2.70	2009-07-31	13
8	Young,Río Negro	-3.00	2012-06-07	22
9	Carrasco,Canelones	-3.10	2007-07-29	1
10	Chacras de Paysandú, Paysandú	-3.30	2012-06-07	12
11	Paso de los Toros, Tacuarembó	-3.50	2012-07-30	11
12	Coronado,Artigas	-3.80	2012-06-08	4
13	Ciudad de Rocha,Rocha	-4.00	2007-07-29	16
14	San José, San José	-4.00	2012-07-30	18
15	Artigas, Artigas	-4.20	2011-07-04	3
16	Las Brujas,Canelones	-4.20	2009-07-31	25
17	Punta del Este, Maldonado	-4.50	2013-08-23	14
18	El Naranjal,Salto	-5.20	2007-07-12	26
19	Lascano,Rocha	-5.50	2012-06-09	23
20	Nueva Hesperides,Salto	-5.60	2012-06-09	17
21	Trinidad,Flores	-5.60	2010-07-16	21
22	Santa Bernardina,Durazno	-6.00	2007-07-29	6
23	Melo, Cerro Largo	-6.00	2012-06-09	9
24	Ciudad de Tacuarembó, Tacuarembó	-7.00	2012-06-09	19
25	Mercedes,Soriano	-7.60	2007-07-29	10
26	Florida,Florida	-10.00	2012-07-28	7

Constatamos que los años mas fríos fueron 2012 y 2007.

¿Hay alguna relación entre la zona y los mínimos registrados?:

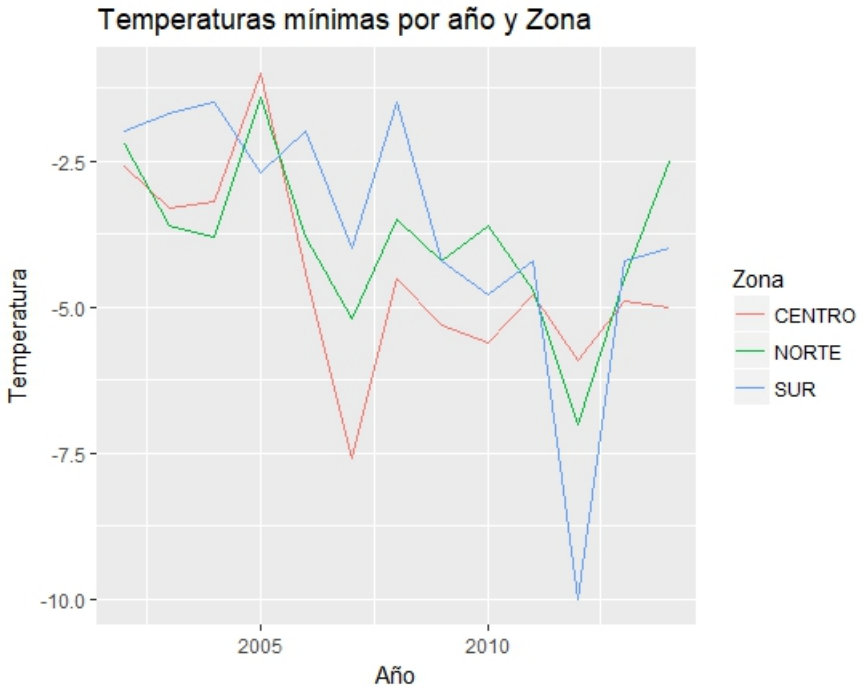


Figure 4: Se observan las temperaturas mínimas registradas por año en cada zona. En la mayoria de los años la tempreatura mínima del Sur son más altas que el Norte y el Centro

	anio	Zona	tmin
1	2002	CENTRO	-2.60
2	2002	NORTE	-2.20
3	2002	SUR	-2.00
4	2003	CENTRO	-3.30
5	2003	NORTE	-3.60
6	2003	SUR	-1.70
7	2004	CENTRO	-3.20
8	2004	NORTE	-3.80
9	2004	SUR	-1.50
10	2005	CENTRO	-1.00
11	2005	NORTE	-1.40
12	2005	SUR	-2.70
13	2006	CENTRO	-4.40
14	2006	NORTE	-3.80
15	2006	SUR	-2.00
16	2007	CENTRO	-7.60
17	2007	NORTE	-5.20
18	2007	SUR	-4.00
19	2008	CENTRO	-4.50
20	2008	NORTE	-3.50
21	2008	SUR	-1.50
22	2009	CENTRO	-5.30
23	2009	NORTE	-4.20
24	2009	SUR	-4.20
25	2010	CENTRO	-5.60
26	2010	NORTE	-3.60
27	2010	SUR	-4.80
28	2011	CENTRO	-4.80
29	2011	NORTE	-4.70
30	2011	SUR	-4.20
31	2012	CENTRO	-5.90
32	2012	NORTE	-7.00
33	2012	SUR	-10.00
34	2013	CENTRO	-4.90
35	2013	NORTE	-4.50
36	2013	SUR	-4.20
37	2014	CENTRO	-5.00
38	2014	NORTE	-2.50
39	2014	SUR	-4.00

5 Descripción de la aplicación shiny

A través de la aplicación Shiny, mostraremos el análisis exploratorio interactivo.

Se estructura en 4 pesatañas:

5.1 Base de datos

En esta pesataña el usuario puede realizar múltiples filtraciones para una primer exploración de la base de datos. Además se incluyó un nuevo data frame que muestra el porcentaje de datos faltantes por estación, de esta forma el usuario puede seleccionar bases que esten mas completas que otras para trabajar.

5.2 Visualización

Aquí introducimos la exploración espacial, donde se pueden seleccionar las estaciones y a través del paquete Leaflet de R , se despliega un mapa marcando geograficamente el punto donde se encuentra la estación. Por otro lado se ve también un mapa de la serie de temperaturas para las estaciones seleccionadas anteriomente, con la opción de poder elegir el periodo. En esta pesataña encontraremos un mapa del territorio uruguayo, el cual se encuentra coloreado por la temperatura mínima regsitrada en el mes para cada departamento. La escala de colores utilizada ayuda a visualizar que departamentos registran temperaturas por encima de la media (rojos) y por debajo (azules) y con color blanco los que están sobre la media.

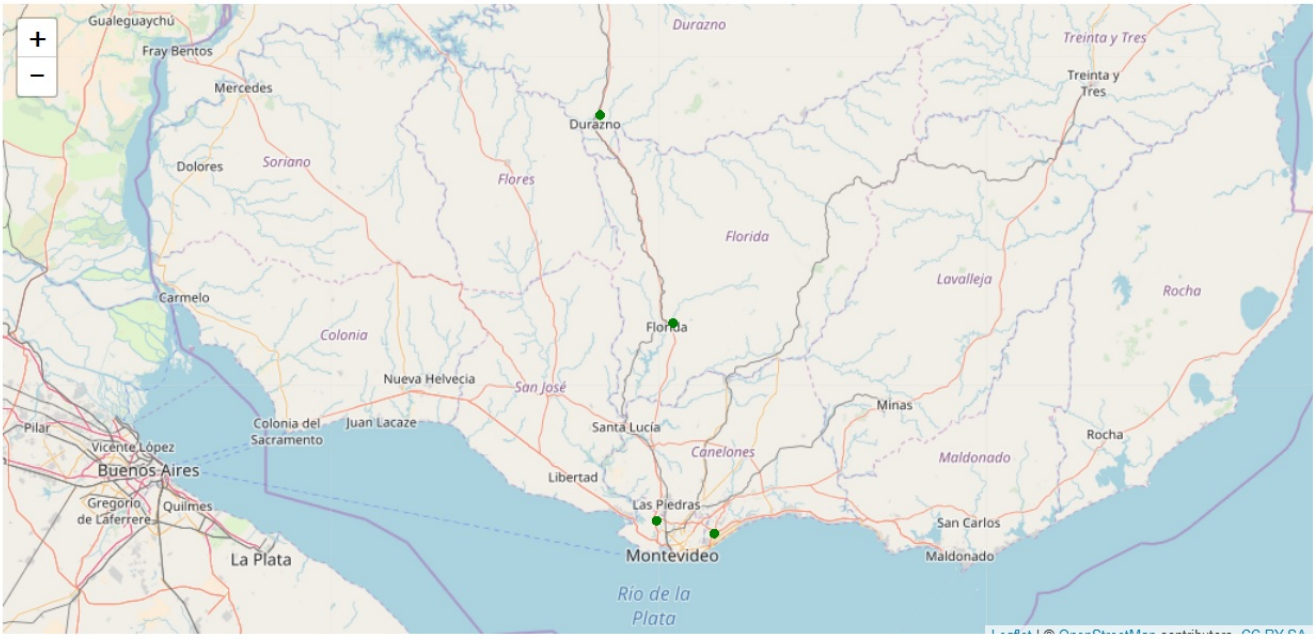


Figure 5: El mapa nos permite un análisis sobre la ubicación espacial de las distintas estaciones.

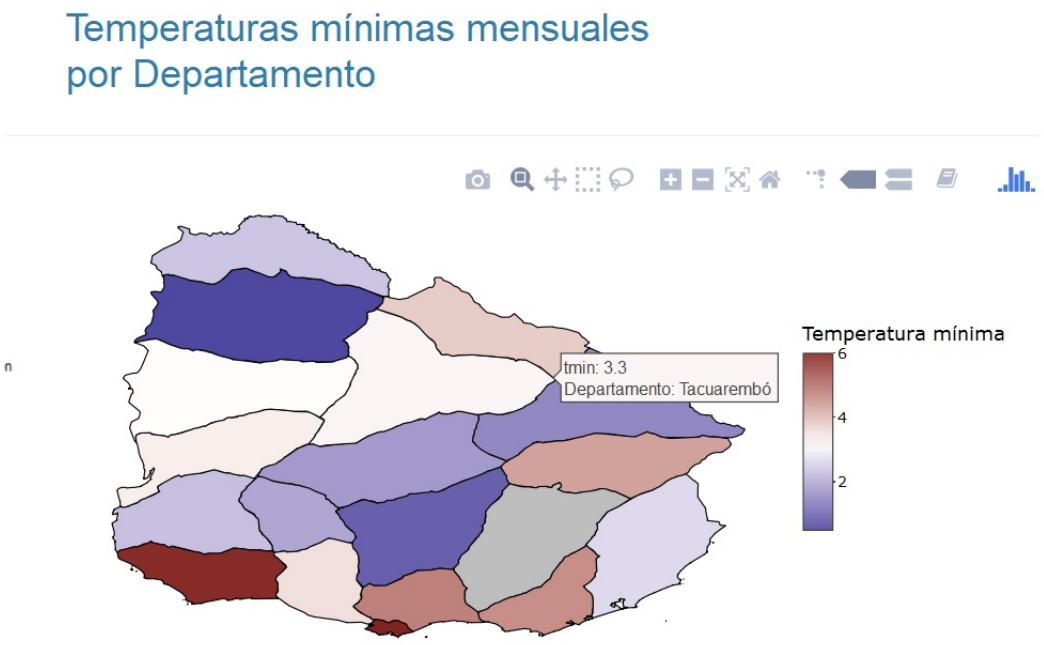


Figure 6: Visualización interactiva que muestra espacialmente las temperaturas mínimas mensuales registradas.

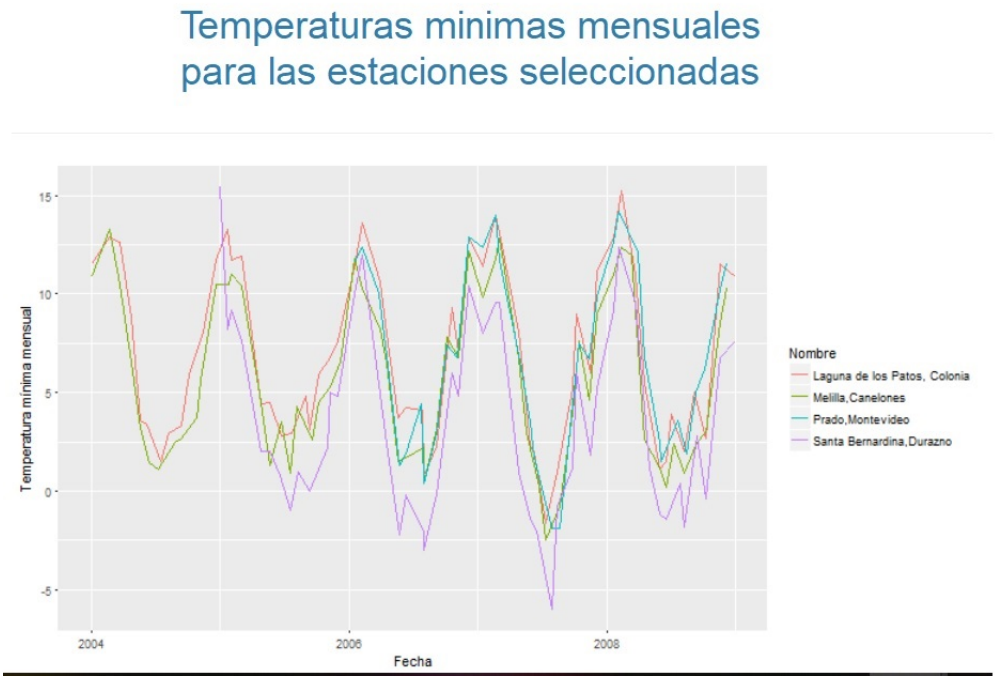


Figure 7: Gráfico de las series de temperaturas para las estaciones seleccionadas

5.3 Video

Se presenta para el mismo mapa anteriormente creado, una animación donde se puede observar a lo largo de un año, como van cambiando las temperaturas para los distintos meses en cada departamento.

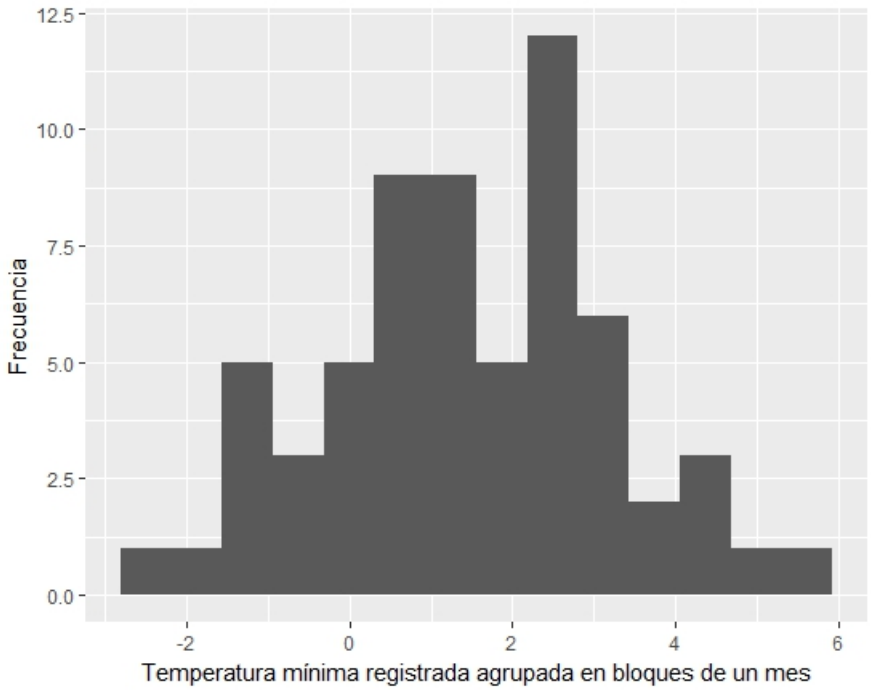
5.4 Base de datos de valores extremos:

Para llevar a cabo ambos métodos mencionados en la metodología (Block Maxima y Umbral) nos generamos una base donde intentamos captar los valores extremos. Incorporamos una nueva página en la Shiny, la cual nos da la opcion de elegir el método. Una vez seleccionado, si deseamos trabajar con Block Maxima, podemos elegir el tamaño de los bloques entre dos opciones: que cada año sea un bloque, o un bloque por mes. Por otro lado si decidimos utilizar el Método del Umbral, la app nos permite seleccionar el valor entero que deseemos como valor u

Análisis exploratorio:

Realizamos un pequeño análisis exploratorio de las nuevas bases de datos, en primer instancia se puede observar la base completa, la cual no contendrá muchas observaciones ya que se está trabajando solo con los datos extremos. Por otro lado, la Shiny nos muestra un summary para conocer un poco la distribución de los datos, asi como también un histograma para ayudar la visualización.

A modo de ejemplo, veamos el summary y el histograma para el método de Block Maxima, de la estación número 2, utilizando como tamaño del bloque un mes.




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -2.500   0.400   1.500   1.475   2.550   5.600
```

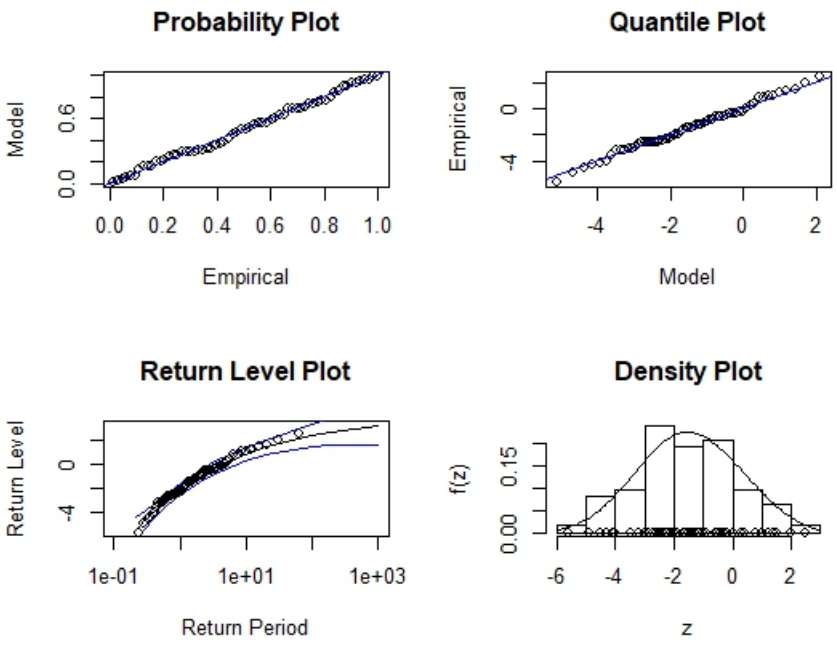
Ajuste de los datos

Luego de realizar la pequeña exploracion de los datos, pasamos al ajuste.

Para el Método del umbral, utilizamos la función "fevd", del paquete "extRemes", la cuál realiza el ajuste de la distribución para datos de valores extremos. Como argumento, la función nos exige el tipo de familia por el cual vamos a modelizar, y siguiendo el teorema de valores extremos para el método del Umbral, la familia que seleccionamos es la de Pareto Generalizada.

Por otro lado, para el Método Block Máxima, realizamos el ajuste a través de los paquetes "evd" e "ismev", además de obtener los parámetros de la familia de distribución GEV, a través de los cuales podemos identificar si el ajuste se hace a través de una Frechet, Gumbell o Weibull, realizamos cuatro plots con el objetivo de evaluar el ajuste.

En los dos plots superiores, comparamos las probabilidades y los cuantiles empíricos con los del modelo. En el plot inferior izquierdo podemos observar en el gráfico el nivel de retorno, el cual se interpreta de la siguiente forma: se espera que el nivel z_p sea excedido en promedio una vez cada $1/p$ unidades de tiempo, es decir, z_p será excedido en una unidad de tiempo con probabilidad p. Por último, en el extremo inferior derecho comparamos el histograma empírico con la función de distribución.




```
$conv
[1] 0

$nllh
[1] 123.3983

$mle
[1] -2.0813145  1.7175915 -0.2833244

$se
[1] 0.23907019 0.16891332 0.08303867
```

6 Comentarios finales

- Se observa una fuerte estacionalidad en la serie de las temperaturas, y media constante, lo cual indica que no hay una tendencia clara.
- Observamos la presencia de tres outliers de temperaturas mínimas, que son aquellos de interés para la futura investigación.
- Los años más frios fueron el 2007 y el 2012, mientras que espacialmente se observa una clara diferencia de temperaturas entre sur con centro y norte,
- Las estaciones que tienen menos NAs son: Chacras de Paysandú y Artigas.
- Las amplitudes térmicas son más grandes en el norte que en el sur. Esto es debido a que en el sur, el agua del Río de la Plata y del Océano Atlántico actúan como reguladores térmicos.

References

[Allaire et al., 2018] Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., and Chang, W. (2018). *rmarkdown: Dynamic Documents for R*. R package version 1.9.

[Chang et al., 2017] Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R*. R package version 1.0.5.

[Cheng et al., 2017] Cheng, J., Karambelkar, B., and Xie, Y. (2017). *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 1.1.0.

[Coles, 2001] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

[functions written by Janet E. Heffernan with R port and documentation provided by Alec G. Stephenson., 2018] functions written by Janet E. Heffernan with R port, O. S. and documentation provided by Alec G. Stephenson., R. (2018). *ismev: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.42.

[Grolemund and Wickham, 2011] Grolemund, G. and Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25.

[Hijmans, 2017] Hijmans, R. J. (2017). *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7.

[R Core Team, 2016] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Sievert et al., 2017] Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy, P. (2017). *plotly: Create Interactive Web Graphics via 'plotly.js'*. R package version 4.7.1.

[Stephenson, 2002] Stephenson, A. G. (2002). evd: Extreme value distributions. *R News*, 2(2):0.

[Wickham, 2009] Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[Wickham and Grolemund, 2016] Wickham, H. and Grolemund, G. (December 2016). *R for Data Science*. O'Reilly Media.