Universidad Zaragoza
1542

Trabajo Fin de Máster

máster en modelización matemática,
estadística y computación

Desarrollo de modelos estocásticos para la proyección de la ocurrencia de extremos: Aplicación a extremos de temperatura

Daniel Antón Galindo

Directora:
Ana Carmen Cebrián

FACULTAD DE CIENCIAS
DEPARTAMENTO DE MÉTODOS ESTADÍSTICOS
25 de noviembre de 2019

Índice general

Resumen

Presentamos la modelización del proceso de ocurrencias de extremos de temperatura en las series de máximas y mínimas diarias del Balneario de Panticosa. Se propone un proceso CPSP, que equivale a la estimación de tres procesos de Poisson No Homogéneos independientes. Estudiamos su evolución en el tiempo y dependencia con algunos predictores de temperatura. Se establecen diferentes herramientas que validan nuestro modelo como una aproximación correcta a los datos disponibles. Este modelo se presenta finalmente como una herramienta útil para la reducción de escala temporal de proyecciones de temperatura. Se obtienen estimaciones de la intensidad del proceso para el siglo XXI a partir de las trayectorias de modelos climáticos globales para diferentes escenarios de cambio climático, observando un incremento del número medio de extremos bajo todos los escenarios, más acentuado en la serie de temperaturas máximas.

We show the modelization of the occurence process of temperatures extremes in the maxima and minima temperatures series of Balneario de Panticosa. We suggest a CPSP process, that is equivalent to the estimation of three Non Homogeneous independient processes. Their temporal evolution and dependence with some temperature predictors is studied. Different tools proposed validate the model as a correct aproximation to the avalaible data., Finally, the model is shown as a usefull tool for de reduction of the temporal scale of temperature proyections. We obtain intensity estimations for the process in the XXI century using differents scenarios of climatic change, observing an increase of the mean number of extremes under all the scenarios, specially in the temperature maxima serie.

Capítulo 1

Introducción y objetivos

El estudio de los extremos de temperatura es, en climatología, un campo de gran interés por su impacto en el ecosistema, la salud e incluso la economía [5]. La definición de extremo depende del enfoque empleado, pero de manera general se considera extrema una observación inusualmente alta dentro de una serie temporal [9]. En el caso de las temperaturas, numerosos estudios hablan de la importancia de su análisis tanto en la serie de máximas diarias como en las mínimas [26]. Por otro lado, a pesar de ser un campo estudiado mayoritariamente en los periodos estivales, los extremos cálidos en los meses de invierno presentan también gran interés por su efecto en la variación del manto nivoso, o en el mantenimiento de los glaciares en zonas montañosas.

Un aumento de la ocurrencia de extremos climáticos ha sido observado en los últimos años, como se ha recalcado en numerosos estudios [19, 3]. Esta situación se enmarca dentro del contexto general de un cambio climático global del que no parece haber dudas ya, y que puede ser uno de los desafíos más importantes de la humanidad para el próximo siglo [16]. Esta situación, ampliamente estudiada a nivel global, requiere sin embargo de resultados a escala local, imprescindibles para la prevención de los impactos del clima en situaciones concretas [24]. Por esta razón, nos proponemos el análisis de la ocurrencia de extremos de temperatura para un caso local particular.

Por ejemplo, es interesante el estudio de zonas donde un posible cambio climático pueda tener efectos negativos sobre el medio ambiente. En concreto, en el entorno del Pirineo se ha observado en los últimos años una disminución del manto nivoso y de la superficie de los glaciares [22], fenómeno posiblemente conectado con el aumento de las temperaturas, concretamente en el periodo invernal. Esta situación puede tener numerosas consecuencias, no sólo en el valioso ecosistema pirenaico, sino también en otras regiones interconectadas como la cuenca hidrográfica del Ebro, cuyo caudal depende fuertemente del nivel de precipitación y deshielo producido en la cordillera pirenaica. De entre las diferentes localidades de esta zona, centraremo nuestro estudio en la estación del Balneario de Panticosa, por el interés de su ubicación (una zona turística dependiente de la nieve y cercana a picos de alta montaña) y por la calidad de los datos disponibles.

Nos proponemos el desarrollo de un modelo para la caracterización de los extremos cálidos en invierno de las series de máximas y mínimas de Panticosa, con el objetivo de realizar su proyección sobre escenarios futuros. El marco elegido para la modelización es el proceso CPSP (*common Poisson shock process*), que nos permitirá caracterizar la ocurrencia de los extremos de ambas series mediante procesos de Poisson independientes. Para la obtención de proyecciones, se utilizarán resultados de modelos GCM (*Global Circulation Model*) bajo diferentes escenarios de cambio climático. Estos modelos de simulación numérica aportan trayectorias para la serie de temperatura diaria a mediana escala, por lo que es necesaria una reducción de escala espacial para su aplicación a nivel local. Sin embargo, las proyecciones obtenidas no pueden considerarse fiables a escala diaria [3], por lo que deben tomarse variables agregadas. De esta manera, el

modelo de ocurrencia de extremos anteriormente construido, se plantea como una herramienta de reducción de la escala temporal, permitiendo la obtención de respuestas diarias a partir de las variables agregadas proyectadas.

El primer capítulo presenta los resultados teóricos disponibles que justifican el comportamiento asintótico Poisson de un proceso de ocurrencia de extremos cuando el umbral definido cumple determinadas características. El segundo capítulo describe la metodología empleada para la reducción de series bivariantes, dependientes y no estacionarias a un conjunto de tres procesos de Poisson No Homogéneos, así como las técnias de ajuste y validación de los modelos construidos. Presentamos en el tercer capítulo la aplicación de estos resultados al caso concreto de la estación de Panticosa, obteniendo un modelo CPSP para el proceso de ocurrencia de extremos en invierno. Por último, se desarolla la proyección de estos modelos para el siglo XXI en el capítulo cuarto, comentando los resultados obtenidos.

Capítulo 2

Teoría de la modelización de valores extremos

La modelización de valores extremos tiene como objetivo la caracterización del **comportamiento estadístico de los valores inusuales** (ya sean inusualmente altos o inusualmente bajos) de un proceso estocástico. Este análisis tiene diferentes enfoques: la caracterización de la distribución del valor máximo de una serie, la caracterización del proceso de ocurrencia de valores por encima de un determinado umbral... Comenzaremos presentando una visión general de la teoría clásica de valores extremos, que se centra en el análisis de los máximos de una serie. Una vez vistas las limitaciones de esta caracterización, presentaremos el modelo de "Excesos Sobre Umbral" (EOT, por sus siglas en inglés) para series de variables *i.i.d.* (independientes e idénticamente distribuidas). Por último, analizaremos el caso particular de series no *i.i.d.*.

2.1. Teoría clásica de extremos: estudio de máximos

Suponemos que tenemos una serie de *variables aleatorias independientes e idénticamente distribuidas* (*v.a. i.i.d.*) $X_1, ..., X_n$ con distribución marginal F. Mostramos un ejemplo de serie en la Figura 2.1.

Queremos establecer el comportamiento estadístico de la siguiente variable "máximo":

$$M_n = max\{X_1, ..., X_n\} \tag{2.1}$$

Como las variables X_i son independientes entre ellas, podríamos establecer la distribución de M_n haciendo una estimación de la distribución F y calculando $Pr(M_n \leq z) = [F(z)]^n$. Sin embargo F puede ser desconocido, y además cualquier pequeño error en la estimación de F crecería exponencialmente al elevarlo a n, por lo que debemos buscar un método más general.

Por otro lado, si el rango de distribución de la función F tiene un máximo finito z_+, tendremos que:

$$\lim_{n\to\infty}[F(z)]^n = \begin{cases} 0 & , z < z_+ \\ 1 & , z \geq z_+ \end{cases} \tag{2.2}$$

Y entonces la distribución de M_n converge a una función degenerada en z_+ para $n \to \infty$. Este problema se soluciona renormalizando la variable M_n de la siguiente manera:

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{2.3}$$

De esta manera, deberemos buscar dos series $\{a_n\}$ y $\{b_n\}$ que aseguren la no-degeneración de la función de distribución de M_n^* para $n \to \infty$.

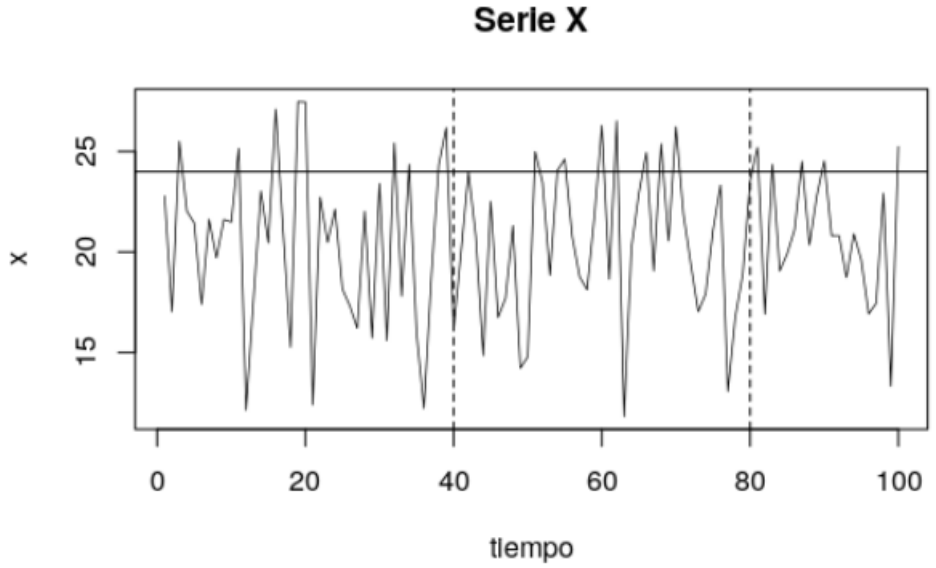


Figura 2.1: Serie ejemplo de *v.a. i.i.d.*. En este caso, se trata de una serie temporal donde cada variable representa la observación de una cierta magnitud (por ejemplo, temperatura ambiente) en cada instante de tiempo.

2.1.1. Distribución asintótica del valor máximo

La distribución asintótica del máximo de una serie de *v.a. i.i.d.*, sea cual sea su distribución marginal F, queda determinada (salvo el valor de los parámetros) por el siguiente teorema (Coles 2001 [9]):

Teorema de distribución del valor máximo: *Sea M_n el máximo de una serie de v.a. i.i.d. $X_1, ..., X_n$, y M_n^* su versión normalizada definida por (2.3). Tenemos que, si existe la distribución límite de M_n^*, es decir si*

$$\lim_{n\to\infty} Pr(M_n^* \leq z) = G(z)$$

*con $G(z)$ una función de distribución no degenarada, entonces $G(z)$ pertenece a la familia de funciones **GEV** (valor extremo generalizado), que tiene la siguiente expresión:*

$$G(z) = exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \text{ con: } \begin{cases} \sigma > 0\text{: parámetro de escala} \\ \mu\text{: parámetro de posición} \\ \xi\text{: parámetro de forma} \end{cases} \tag{2.4}$$

Este teorema es de gran utilidad ya que reduce la búsqueda de la distribución de M_n a la estimación de tres parámetros: σ, μ y ξ, determinados por las características de la serie que estemos considerando. La condición para su aplicación es la existencia de dos series de constantes $\{a_n\}$ y $\{b_n\}$ que garanticen la convergencia de la distribución asintótica de M_n^*. Así pues, siempre que esta distribución exista, estará determinada por (2.4). La demostración de este resultado excede los propósitos de este trabajo, pero puede ser consultada en Leadbetter et al. (1983) [17]. También se puede encontrar una versión informal de ésta en Coles (2001), p. 49-51 [9].

Además, es importante indicar que el **tratamiento de mínimos** puede tratarse de manera equivalente bajo un simple cambio de signo de las variables de la serie y un posterior análisis del máximo de la nueva serie obtenida.

2.1.2. Limitaciones del modelo de máximos

Una limitación del modelo clásico de extremos es la suposición de independencia de las variables. Esta es una condición muy restrictiva que raramente se observa en situaciones reales, sobre todo en el campo de los problemas medio-ambientales, en el que se enmarca este trabajo. En la sección 2.3 revisaremos los resultados de la teoría de extremos para el caso de **series dependientes**, solventando en gran parte este problema.

Pero el incoveniente principal de esta aproximación es el procedimiento de inferencia mediante el cual se obtienen los parámetros de la distribución límite. Para realizar esta inferencia, sólo disponemos de una observación por cada periodo de longitud n definido, utilizando por tanto una fracción $1/n$ de todos los datos y perdiendo de esta manera mucha información. Para evitar este problema, se han propuesto diferentes alternativas que buscan todas ellas utilizar un mayor número de datos incluyendo aquellos que, aún sin ser máximos de ningún periodo, tienen un carácter que podría considerarse "extremo". Una de ellas es la generalización de los **valores k-extremos**, que considera en cada periodo los k estadísticos ordenados superiores, y fue propuesta por Pickands (1975) y Weissman (1978), véase [21] y [29].

Otra aproximación que ha gozado de gran aceptación en el campo de las ciencias medio-ambientales, sobre todo por su interpretación sencilla, son los métodos de **Excesos Sobre Umbral (EOT)**, que define los extremos como todas las variables con un valor superior a un cierto umbral. Fueron desarrollados formalmente por Davison y Smith en 1990 [11], aunque ya gozaban entonces de un uso extendido en otros campos [25]. Es el método que emplearemos en nuestro análisis, por lo que dedicaremos el resto del capítulo a la presentación y discusión de sus resultados principales.

2.2. Modelo de Excesos Sobre Umbral (EOT)

En el modelo EOT, una observación X se denomina **extremo** o **valor extremo** si $X > u$. Además, denominamos **exceso** de X sobre el umbral u al valor positivo $Y = X - u | X > u$. Nuestro objetivo es, en primer lugar, obtener una expresión para la distribución de la variable Y y, en segundo lugar, una caracterización de la ocurrencia de extremos (es decir, la distribución del número de extremos). Una vez obtenidas las dos distribuciones, presentaremos el modelo bajo un punto de vista de los procesos puntuales, lo que nos permitirá establecer una relación entre todos los resultados obtenidos.

2.2.1. Distribución de los excesos

Sea $X_1, ..., X_n$ una serie de *v.a. i.i.d* con distribución marginal F. A priori, podríamos tratar de encontrar la distribución de los excesos Y de esta serie a partir de la siguiente expresión:

$$Pr(X > u + y | X > u) = \frac{\bar{F}(u+y)}{\bar{F}(u)} \qquad \forall\, y > 0 \tag{2.5}$$

donde $\bar{F}(u) = 1 - F(u)$.

Pero, como en el caso de los máximos, la distribución F puede ser desconocida, por lo que trataremos de obtener la distribución de Y de una manera más general. Efectivamente, como mostramos en el siguiente teorema (Coles 2001 [9]), bajo las mismas condiciones en las que M_n tiene la distribución GEV dada por (2.4), se obtiene la distribución asintótica de Y:

Teorema de distribución de los excesos: *Dada una serie $X_1, ..., X_n$ de v.a. i.i.d que cumple las condiciones en las que el máximo M_n converge a una distribución $GEV(z; \mu, \sigma, \xi)$, entonces la distribución de Y converge, bajo $n \to \infty$ y $u \to \infty$, a la distribución:*

$$\begin{cases} \text{con } \xi \neq 0: & H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} & \text{para} & \{y : y > 0 \ \wedge \ (1 + \xi y/\tilde{\sigma}) > 0\} \\ \text{con } \xi = 0: & 1 - e^{-y/\tilde{\sigma}} & \text{para} & y > 0 \end{cases} \tag{2.6}$$

donde la relación con los parámetros de la GEV viene dada por:

$$\tilde{\sigma} = \sigma + \xi(u - \mu) \tag{2.7}$$

Esta distribución se denomina ***distribución Pareto generalizada (PG)****.*

Presentamos a continuación una prueba informal del teorema anterior, propuesta en Coles (2001) [9]. Para una demostración más rigurosa, consultar Leadbetter et al. (1983) [17].

Prueba: Consideramos la *v.a.* X que sigue una distribución F tal que, para n suficientemente grande, podemos aproximar el resultado 2.4:

$$F^n(z) \approx GEV(z; \mu, \sigma, \xi) = exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

Tomando logartimos en cada lado, obtenemos:

$$n \log F(z) \approx -\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi} \tag{2.8}$$

Además, para z suficientemente grande, podemos aproximar el logaritmo a su serie de Taylor:

$$\log F(u) \approx -[1 - F(u)]$$

que, sustituyendo en 2.8 y reordenando, nos da:

$$1 - F(u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u - \mu}{\sigma}\right)\right]^{-1/\xi} \tag{2.9}$$

De manera similar, podemos obtener, para $y > 0$ y u suficientemente grande, la siguiente expresión:

$$1 - F(u + y) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u + y - \mu}{\sigma}\right)\right]^{-1/\xi}$$

Finalmente, obtenemos la expresión para la distribución de Y utilizando las dos expresiones anteriores:

$$\begin{aligned} P(Y > y) &= P(X > u + y | X > u) \\ &\approx \frac{1/n[1+\xi(u+y-\mu)/\sigma]^{-1/\xi}}{1/n[1+\xi(u-\mu)/\sigma]^{-1/\xi}} \\ &= \left[\frac{1+\xi(u-\mu)/\sigma+\xi y/\sigma}{1+\xi(u-\mu)/\sigma}\right]^{-1/\xi} \\ &= \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi} \end{aligned} \tag{2.10}$$

donde hemos definido $\tilde{\sigma} = \sigma + \xi(u - \mu)$. $\square$

El resultado anterior es muy importante ya que, como en el caso de los máximos, obtenemos la distribución asintótica de Y bajo $n \to \infty$ y $u \to \infty$, que es la misma sea cual sea la distribución F de X, siempre que exista tal convergencia. La inferencia sobre Y se basa de nuevo pues, en la estimación de los parámetros de la función, en este caso ξ y $\tilde{\sigma}$.

2.2.2. Modelización de la ocurrencia de extremos

Más que la distribución de los excesos definidos sobre un umbral, muchas veces queremos conocer la distribución de la ocurrencia de estos excesos, es decir, la probabilidad con la que una variable X va a superar nuestro umbral u en cada momento o, lo que es equivalente, la distribución del número de extremos en un intervalo dado. Como veremos más adelante, se puede demostrar que, bajo ciertas condiciones, la **ocurrencia de estos extremos** se comporta asintóticamente como un **proceso de Poisson**. Por lo tanto, antes de enunciar el teorema correspondiente, recordaremos qué son los procesos de Poisson, empezando por la idea general de los procesos puntuales.

Proceso puntual: Un proceso puntual definido en un conjunto A es una norma estocástica para la ocurrencia y posición de eventos puntuales en este conjunto. Llamamos $N(a)$ a la *v. a.* que indica el *número de eventos* en un subconjunto $a \subset A$. Así mismo, llamamos *intensidad* $\Lambda(a)$ a la esperanza de ésta función en el subconjunto a y *densidad de intensidad* o *tasa* a la función:

$$\lambda(a) = \frac{\partial \Lambda(a)}{\partial x_1 \cdots \partial x_k} \qquad \text{, con } a \subset \mathbb{R}^k \tag{2.11}$$

En concreto, decimos que un proceso puntual tiene *incrementos independientes* si $N(a)$ es independiente de $N(b)$ para todo $a, b \subset A$ con $a \cap b = \emptyset$. También, decimos que un proceso puntual tiene *incrementos estacionarios* si $N(a) = N(b)$ para todo $a, b \subset A$ con $|a| = |b|$ (es decir, si el número de eventos sólo depende del tamaño del intervalo).

Proceso de Poisson: Un proceso de Poisson en $\mathbb{R}^+$ (tiempo t generalmente) de tasa $\lambda(t)$, $PP(\lambda(t))$, es un proceso puntual para la caracterización de la ocurrencia de eventos en $\mathbb{R}^+$. Se define como un proceso puntual que cumple las siguientes condiciones (llamando $N(a_1, a_2)$ al valor $N((a_1, a_2]))$:

$$\begin{array}{l} \text{a) El proceso tiene incrementos independientes} \\ \text{b) } P(N(t, t+\delta) = 1) = \lambda(t)\delta + o(\delta) \\ \text{c) } P(N(t, t+\delta) > 1) = o(\delta) \\ \text{por lo que:} \\ \Rightarrow P(N(t, t+\delta) = 0) = 1 - \lambda(t)\delta + o(\delta) \end{array} \tag{2.12}$$

Si el proceso tiene además incrementos estacionarios, la función $\lambda(t)$ será constate, y diremos que se trata de un proceso de *Poisson Homogéneo* (PPH), si no, diremos que es un proceso de *Poisson No Homogéneo* (PPNH).

Las consecuencias principales de esta definición, y las que realmente caracterizan un PP en la práctica, son las siguientes[1]:

1. **Número de ocurrencias**: El número $N(a)$ de ocurrencias en cualquier intervalo temporal a sigue una distribución de Poisson de parámetro $\Lambda(a)$. Además, los números de ocurrencias en subconjuntos disjuntos de A son mutuamente independientes. Es decir, una definición equivalente para el Proceso de Poisson, es la de todo aquel proceso puntual que cumpla:

$$\begin{array}{ll} \text{a)} & P(N(a) = n) = \frac{e^{-\Lambda(a)} \cdot \Lambda(a)^n}{n!} \quad \forall\, a \subset A \\ \text{b)} & N(a) \text{ y } N(b) \text{ son } v.a. \text{ independientes} \quad \forall\, a, b \subset A : a \cap b = \emptyset \end{array} \tag{2.13}$$

   En el caso de un PP homogéneo, por (2.11), tendremos que $\Lambda(a) = \lambda \cdot |a|$, donde $|a|$ es el tamaño del subconjunto a.

[1] Las demostraciones de estas dos consecuencias pueden consultarse por ejemplo en Ross (1996) [23] (páginas 79 y 65 respectivamente).

2. **Tiempos de recurrencia**: Si llamamos T_i al *instante de ocurrencia* del evento i, definimos el *tiempo de recurrencia* del evento i como $R_i = T_i - T_{i-1}\ \forall i > 1$, siendo $R_1 = T_1$. Así pues, en un PP homogéneo, los tiempos de recurrencia son *v.a. i.i.d* con distribución exponencial de parámetro λ y, como consecuencia, cada instante de ocurrencia T_i tiene distribución $\Gamma(i, \lambda)$. Es decir:

   a) $P(R_i < t) = 1 - e^{-\lambda t}$

   b) $P(T_i < t') = \int_{t'}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^{i-1}}{(i-1)!} dt\ , \quad t' \geq 0$

Ahora que hemos definido el proceso de Poisson y sus características, podemos enunciar el siguiente teorema, propuesto por Leadbetter et al. (1983) [17] y que establece que la ocurrencia de extremos tiene un comportamiento Poisson.

Teorema del comportamiento Poisson de los extremos: *Sea $X_1, ..., X_n$ una muestra de v.a. i.i.d con distribución marginal F. Definimos una sucesión de umbrales u_n, y llamamos r_n al número de extremos presentes en la muestra:*

$$r_n := \sharp\{X_i : X_i > u_n\}_{i=1,...,n}$$

Entonces, si tenemos que la sucesión u_n verifica:

$$\lim_{n\to\infty} n\bar{F}(u_n) = \tau \tag{2.14}$$

Tenemos que, para $k = 0, 1, 2, ...$:

$$\lim_{n\to\infty} P(r_n \leq k) = e^{-\tau} \sum_{s=0}^{k} \frac{\tau^s}{s!} \tag{2.15}$$

Es decir: r_n tiene una distribución asintótica Poisson de intensidad τ bajo $n \to \infty$

Prueba: En primer lugar, podemos ver que la *v.a.* r_n hace un recuento del "número de éxitos" en un proceso de n ensayos (que son las n *v.a. i.i.d* X_i), siendo la probabilidad de "éxito" $p_n = 1 - P(X < u_n) = \bar{F}(u_n)$. Por lo tanto, r_n tiene una distribución Binomial $Bin(n, \bar{F}(u_n))$. Además, bajo la condición (2.14) y para $n \to \infty$, es bien conocido que esta distribución tiende a una distribución Poisson de parámetro τ: $\lim_{n\to\infty} Bin(n, \bar{F}(u_n)) = Poi(\tau)$, de ahí la expresión (2.15). $\square$

Con este teorema, hemos establecido que la distribución asintótica de r_n es Poisson cuando la sucesión de umbrales u_n cumplen la condición (2.14). Para interpretar esta condición debemos fijarnos en que, para que el límite (2.14) sea finito, $\bar{F}(u_n)$ debe tender a 0 cuando $n \to \infty$, o lo que es lo mismo, u_n debe tender a z_+ (valor máximo de la distribución de X, que puede ser infinito). Por otro lado, $n\bar{F}(u_n)$ corresponde al número medio de extremos, por lo que realmente el teorema anterior nos dice que, **cuando u_n se elige tal que el número medio de extremos $\bar{r}_n$ es finito bajo $n \to \infty$**, la distribución asintótica de r_n es una distribución Poisson. Además, teniendo en cuenta que nuestras variables son independientes, podemos afirmar que, bajo las condiciones anteriores, **el proceso de ocurrencia de extremos se comporta asintóticamente como un proceso de Poisson** (se cumple la definición (2.13) de proceso de Poisson).

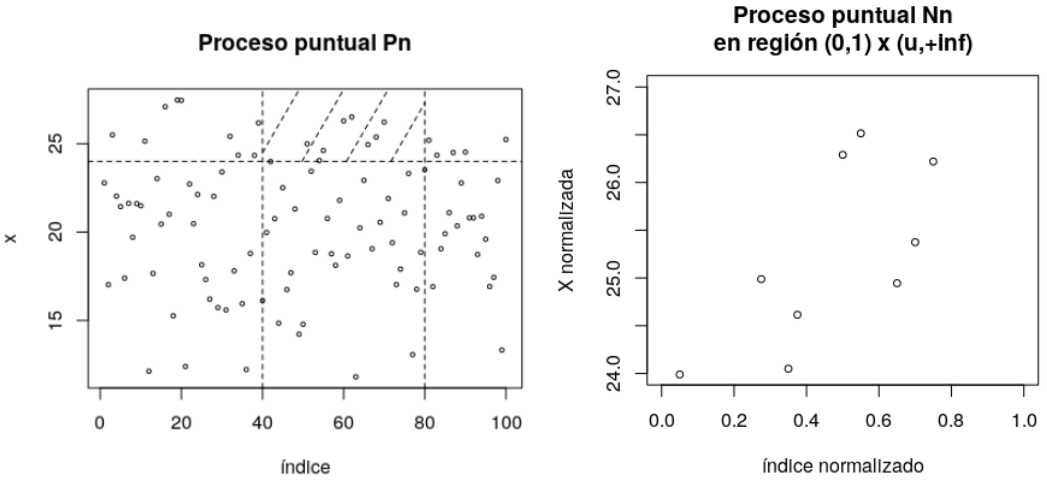
2.2.3. Modelo de Procesos Puntuales

Presentamos ahora un marco formal más general que permite agrupar todos los modelos de extremos enunciados hasta ahora (el modelo de los k estadísticos superiores, el caso concreto de los máximos y el modelo EOT presentado en la sección anterior). A pesar de no aportar resultados nuevos para el estudio de extremos, permite hacer una interpretación global que unifica los resultados y establece relaciones entre ellos.

La idea principal del modelo es la siguiente. Sea una serie $\{X_i\}_{i=1,...,n}$ de *v.a. i.i.d.* con distribución marginal F que cumple las condiciones de convergencia del máximo. Podemos caracterizar esta serie como un proceso puntual bidimensional definido de la siguiente manera:

$$P_n = \{(i, X_i) : i = 1, ..., n\} \tag{2.16}$$

Como en el estudio de extremos únicamente nos interesa la región correspondeinte a la cola superior de la distribución X, estudiaremos este proceso puntual en regiones del tipo $[t_1, t_2] \times (u_n, +\infty)$, donde $[t_1, t_2]$ es el intervalo temporal que nos interesa y u_n es una sucesión de umbrales no decreciente. De manera ilustrativa, podemos ver en la Figura 2.2.(a) el proceso puntual asociado a la serie que habíamos mostrado en la Figura 2.1.



(a) (b)

Figura 2.2: Procesos puntuales P_n y N_n asociados a la serie de *v.a. i.i.d.* presentada en la Figura 2.1. La región rayada de la Figura (a), corresponde a la región de interés de un posible estudio de extremos: valores superiores a u (24 en este caso) en el intervalo temporal $[t_1, t_2]$ ([40,80] en este caso).

Para caracterizar el comportamiento asintótico de un proceso de este tipo, definimos un nuevo proceso puntual escalado[2]

$$N_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) : i = 1, ..., n \right\} \tag{2.17}$$

donde a_n y b_n son constantes normalizadoras que garantizan la estabilidad del comportamiento para $n \to \infty$. La normalización de las abcisas, por su parte, reescala el conjunto de n puntos en el intervalo $(0, 1)$, garantizando así que el número esperado de puntos por unidad de tiempo sea constante cuando $n \to \infty$ (la proporción de excesos cada vez más pequeña debida al aumento de u_n se compensa con la mayor densidad de puntos debida al aumento de n). Esto garantiza que se cumple la condición (2.14), y por lo tanto, aplicando el teorema anterior, se puede demostrar que el comportamiento de N_n en regiones $[t_1, t_2] \times (u_n, +\infty)$ es asintóticamente Poisson.

[2]En la Figura 2.2, podemos ver ambos procesos, P_n y N_n, para una misma serie.

La caracterización que acabamos de hacer tiene la ventaja de establecer el estudio de los máximos M_n o de los excesos Y como casos particulares de este proceso puntual N_n. Por ejemplo, el evento $M_n < z$ es equivalente al evento $N_n(a_z) = 0$, donde $a_z = (0,1] \times (z, +\infty)$. Esto permite establecer una relación entre los parámetros de ambas distribuciones, tal y como ilustra el siguiente teorema (Coles 2001 [9]):

Teorema de relación *GEV-PPH*: *Sea $X_1, ..., X_n$ una serie de v.a. i.i.d. y M_n su valor máximo. Supongamos que se cumplen las condiciones de convergencia del máximo, es decir, suponemos que existen dos series de constantes normalizadoras, a_n y b_n, tales que:*

$$\lim_{n\to\infty} Pr\{(M_n - b_n)/a_n \leq z\} = G(z) = exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

Entonces, si llamamos z_- y z_+ a los valores mínimo y máximo de $G(z)$ respectivamente, tenemos que, el proceso puntual N_n definido en (2.17) converge (bajo $n \to \infty$) a un proceso de Poisson de intensidad:

$$\Lambda(a) = (t_2 - t_1)\left[1 + \xi\left(\frac{u_n - \mu}{\sigma}\right)\right]^{-1/\xi}$$

en cada región $a = [t_1, t_2] \times [u_n, z_+)$, siempre que u_n cumpla la condición (2.14).

Prueba: El carácter asintótico Poisson de N_n en a viene dado por la aplicación directa del teorema de la sección 2.2.2. Este teorema indica que la intensidad del proceso Poisson obtenido vale $\tau = \lim_{n\to\infty} np$, con $p = Pr(X > u_n)$, que no es otra cosa que la probabilidad de que un punto de N_n caiga en la región a. Si además asumimos que M_n tiene una distribución *GEV*, entonces podemos aproximar la función p a (2.9) para n y u_n suficientemente altos, obteniendo entonces:

$$\tau = \lim_{n\to\infty} np = n \cdot \frac{1}{n}\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}$$

Que, añadido a las condiciones de homogeneidad supuestas en nuestro proceso puntual, encontramos la expresión de la intensidad para cualquier intervalo (t_1, t_2):

$$\Lambda(t_1, t_2) = (t_2 - t_1)\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}$$

□

Más allá de las ventajas interpretativas, el modelo de procesos puntuales permite generalizar los resultados que hemos obtenido para series que presentan dependencia entre las variables [17], situación que trataremos en la próxima sección. Además, a nivel práctico, permite establecer la relación entre los diferentes modelos de extremos como ya hemos comentado. Aunque no hemos establecido una demostración rigurosa de la equivalencia de estos modelos (puede consularse en [17] p. 32), podemos establecer la siguiente conclusión. En general tenemos que, **bajo las condiciones en las que los excesos de una serie *i.i.d.* sobre un umbral u presentan una distribución asintótica PG, el proceso de ocurrencia de estos extremos presentará un comportamiento asintótico Poisson**.

2.3. Extremos en series no *i.i.d.*

Todos los resultados obtenidos hasta ahora parten del supuesto de que nuestra serie $\{X\}_{i=1,...,n}$ es *i.i.d.*. Sin embargo, en el caso de series medio-ambientales, es muy frecuente encontrar dependencias entre las variables, por lo que no son aplicables los resultados anteriores. Presentamos

en esta sección algunas generalizaciones que se han desarrollado para los resultados anteriores que permiten su aplicación para series no *i.i.d.*.

En primer lugar, consideraremos una serie estacionaria no independiente. Recordamos la definición de serie estacionaria:

Definición de serie estacionaria: *Un proceso se dice estrictamente* ***estacionario*** *si la distribución d de todo vector de dimensión finita es invariante bajo desplazamiento temporal:*

$$d(X_{t_1}, ..., X_{t_m}) = d(X_{t_1+h}, ..., X_{t_m+h})$$

$$\forall\, h \in \mathbb{Z},\ t_1 < ... < t_m$$

Una consecuencia de la definición de proceso estacionario es que la distribución marginal de cada *v.a.* es idéntica. Sin embargo, estas *v.a.* pueden ser dependientes, siempre que esta dependencia sea homogénea respecto al tiempo. Es decir, la dependencia entre X_1 y X_5 debe ser la misma que entre X_{11} y X_{15}, o entre X_{12} y X_{16}. Para series de este tipo se podrán generlizar los resultados obtenidos en cuanto a los extremos, siempre que se cumplan ciertas características en la dependencia. En concreto, definimos:

Definición de condición $D(u_n)$: *Diremos que una serie X_n verifica la condición $D(u_n)$ si para enteros cualesquiera,*

$$1 \leq i_1 < ... < i_p < j_1 < ... < j_q \leq n,$$

y, definiendo $A_1 = \{i_1, ..., i_p\}$, $A_2 = \{j_1, ..., j_q\}$, se verifica:

$$\left| Pr\left(\max_{i \in A_1 \cup A_2} X_i \leq u_n \right) - Pr\left(\max_{i \in A_1} X_i \leq u_n \right) Pr\left(\max_{i \in A_2} X_i \leq u_n \right) \right| \leq \alpha_{n,l_n} \tag{2.18}$$

siempre que $j_1 - i_p > l_n$, y $\lim_{n\to\infty} \alpha_{n,l_n} = 0$ para alguna sucesión $l_n = o(n)$.

Es fácil ver que, en el caso de un proceso de *v.a.* independientes, la diferencia en (2.18) vale 0. En cambio, cuando dos conjuntos disjuntos cualesquiera de variables (A_1 y A_2 en la definición, separados por al menos l_n variables) presenten algún tipo de dependencia, la diferencia (2.18) será diferente de 0. Lo que exigimos en la condición $D(u_n)$ es que, para un cierto valor u_n, esa diferencia sea pequeña, en concreto, que tienda a 0 cuando $n \to \infty$. Esta condición corresponde con la de una serie que presenta una **dependencia débil a largo plazo** (ya que exigimos una cierta separación entre los conjuntos). Lo que ahora vamos a presentar es que, cuando la condición se cumple para una cierta secuencia de valores u_n, crecientes con n, entonces los resultados límite para los máximos de la serie siguen siendo válidos (Coles 2001, [9]):

Teorema: *Sea M_n la sucesión de máximos de un proceso estacionario. Si existen constantes a_n y b_n tales que la condición $D(u_n)$ se verifica para $u_n = a_n x + b_n\ \forall\, x \in \mathbb{R}$ y*

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \right) = G(x)$$

Entonces la distribución $G(x)$ es GEV.

Se ha demostrado pues que en una serie estacionaria con dependencia, siempre que ésta sea débil a largo plazo (condición D), también se observa el comportamiento límite *GEV* para el máximo. Sin embargo, en este caso no se puede establecer la relación entre las condiciones de convergencia del máximo y el carácter Poisson de la ocurrencia de extremos [17] (p.58). Esta relación se podrá establecer solamente si se establecen nuevas restricciones, más estrictas que la

condición D. En concreto, deberemos exigir una limitación en la posible **dependencia a corto plazo**, sobre la cual no se han puesto restricciones. Esta dependencia a corto plazo se traduce en la tendencia a la formación de "clusters" de variables. Por ejemplo, en una serie con dependencia a corto plazo será común observar que los extremos ocurran en agrupaciones de varias variables consecutivas. Se puede por lo tanto definir una condición de dependencia débil a corto plazo, llamada $D'(u_n)$ o condición "anti-cluster", bajo la cual se puede demostrar el carácter *Poisson* del proceso de ocurrencia de los extremos [17] (p.59).

Capítulo 3

Metodología para la modelización de extremos

En el capítulo anterior hemos establecido el comportamiento asintóticos de los extremos de una serie de variables aleatorias *i.i.d.*, generalizable a series con dependencia bajo ciertas condiciones, D y D'. Necesitamos establecer una metodología que nos permita aplicar estos resultados a las situaciones reales a las cuales nos vamos a enfrentar. En concreto, consideramos la situación de una **serie bivariante con dependencia y no estacionaria**, como puede ser el caso de una serie de temperaturas máximas y mínimas. Esta metodología no es única y pueden existir diversos enfoques que nos permitan todos ellos resolver un mismo problema, por lo que en este capítulo presentaremos simplemente una propuesta que consideramos la más adecuada para el problema que intentaremos resolver en el próximo capítulo.

En la primera sección, presentaremos el modelo CPSP que nos permitirá reducir el proceso bivariante a tres procesos Poisson independientes. A continuación, enunciaremos las aproximaciones necesarias para la caracterización de procesos de extremos univariantes, continuando con la descripción detallada de las etapas de modelización. En primer lugar, estableceremos un método para la correcta elección de los umbrales u, comentando en segundo lugar el proceso de estimación de los procesos de Poisson No Homogéneos independientes. Por último, presentaremos algunas de las diferentes herramientas de validación disponibles para evaluar el modelo propuesto.

3.1. Modelo CPSP

Suponemos que disponemos de dos series de variables aleatorias que presentan dependencia entre ellas y cuyos procesos de ocurrencia de extremos queremos caracterizar. Dado que las series no son independientes, debemos considerar ambos procesos de ocurrencia de extremos como parte de un mismo *proceso bivariante*. El modelo CPSP tiene como objetivo la caracterización de este proceso bivariante, permitiendo la modelización de la dependencia entre los dos procesos subyacentes y el comportamiento marginal de cada uno de ellos. La idea principal es la simplificación del problema mediante la definición de tres procesos independientes equivalentes al proceso inicial, representando uno de ellos la dependencia entre los procesos y los otros dos los procesos marginales independientes.

Más generalmente, un CPSP es un proceso puntual multivariante que asume que existe un proceso Poisson de "choques" subyacente que puede producir d tipos de eventos, de tal manera que el proceso puntual del evento j es el proceso N_j, para $j = 1, ..., d$. Definimos el vector de variables binarias $\mathbf{I}_r = (I_{1,r}, ..., I_{d,r})$ que indica qué evento ocurre en cada punto r del proceso N_0. El modelo CPSP asume que la serie de $\mathbf{I}_r$ es *i.i.d.* multivariante Bernoulli. De esta manera,

cada subproceso N_j puede expresarse como

$$N_j(t) = \sum_{r=1}^{N_0(t)} I_{j,r}$$

que, dado que corresponde a una selección independiente de un proceso de Poisson compuesto, es un proceso de Poisson él mismo.

Volviendo al caso bivariante ($d = 2$), los eventos pueden dividirse en tres tipos: (1,0),(0,1) y (1,1) dependiendo de las componentes observadas en cada uno de ellos. Es decir, el proceso puede descomponerse en tres procesos marginales que corresponden a la ocurrencia de los eventos en los que sólo ocurre el evento de tipo 1, $N_{(1)}$, los eventos en los que sólo ocurre el tipo 2, $N_{(2)}$ y los eventos en los que ocurren ambos; $N_{(12)}$. Ilustramos esta situación en la Figura 3.1. De esta manera, la intensidad de ocurrencia del evento 1 es $\lambda_1 = \lambda(1) + \lambda_{(12)}$ y la intensidad del evento 2 es $\lambda_2 = \lambda(3) + \lambda_{(12)}$. El modelo CPSP supone por lo tanto, que la dependencia entre ambos procesos originales se caracteriza totalmente por el proceso de eventos simultáneos, $N_{(12)}$. De esta manera, hemos establecido el mecanismo necesario para reducir nuestro problema bivariante a una modelización de tres procesos de Poisson univariantes.

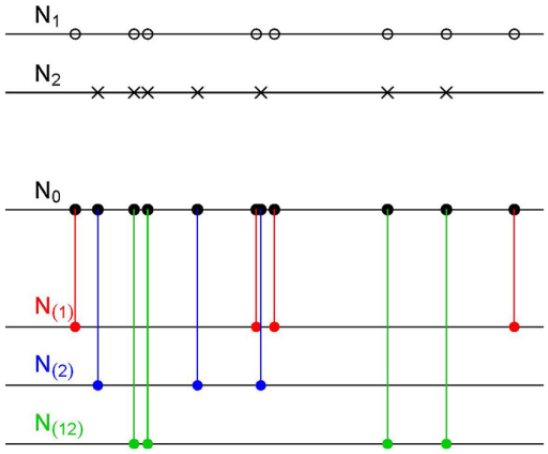


Figura 3.1: Descomposición del proceso CPSP con subprocesos N_1 y N_2 en tres procesos marginales independientes Poisson, $N_{(1)}$, $N_{(2)}$ y $N_{(12)}$. Fuente: [3].

3.2. Aproximaciones para la modelización de extremos univariantes

Una vez hecha la descomposición del proceso original en tres procesos independientes, queremos caracterizar cada uno de éstos. Presentamos a continuación la metodología general y las aproximaciones que nos permiten asumir que cada uno de estos procesos de ocurrencia pueden modelizarse mediante un proceso de Poisson No Homogéneo. Como es habitual en modelización, asumiremos que estas aproximaciones son correctas y estableceremos herramientas de validación (sección 3.5) que nos permitan evaluar a posteriori la calidad del modelo construido. Si nuestro modelo pasa las herramientas de validación, asumiremos que nuestros datos satisfacen todas las hipótesis planteadas en el inicio.

3.2.1. Series finitas

En primer lugar, debemos tomar series con tamaños n y umbrales u lo suficientemente grandes como para poder asumir que la distribución de los extremos se aproxima a su distribución

asintótica, presentada en el capítulo anterior. Para poder hacer esta hipótesis, estableceremos en 3.3, las herramientas que nos permitirán elegir el valor de u de manera razonable.

3.2.2. Series con dependencia: proceso POT para la identificación de extremos

Por otro lado, debemos garantizar que la dependencia entre las variables de nuestra serie es débil a corto y a largo plazo (condiciones D y D') para poder asumir que el proceso de ocurrencia de extremos tiene un comportamiento Poisson. Sin embargo, la condición D' raramente se cumple en series medio-ambientales [9], ya que éstas suelen tener fuerte dependencia a corto plazo (es más probable por ejemplo, que si un día ha llovido, llueva al día siguiente). Para la modelización de este tipo de series se propone una definición alternativa de extremo que nos permita asumir que la dependencia entre la ocurrencia de éstos no presenta dependencia a corto plazo (condición D').

Como decíamos en el capítulo 2, la dependencia a corto plazo se caracteriza por la tendencia a la formación de *clusters*. Por lo tanto, debemos definir los extremos de tal manera que éstos no presenten una tendencia a la agrupación. La solución es aportada por el método de **Picos Sobre Umbral** (**POT**, por sus siglas en inglés) que consiste en lo siguiente:

1. Consideramos una serie de variables aleatorias X con dependencia débil a largo plazo (cumple condición D).

2. Definimos una *ola de extremos* como un conjunto de variables sucesivas que superan el umbral u definido.

3. Definimos un *pico* o *extremo* como la variable X perteneciente a una ola de extremos y que toma el valor máximo de éste. El resto de variables pertenecientes a la ola de extremos no son identificadas como extremos.

De esta manera, definiendo únicamente un extremo en cada conjunto de variables sucesivas que superan el umbral, eliminamos la posibilidad de observación de dos extremos sucesivos. De esta manera, hemos eliminado a la fuerza la dependencia a corto plazo del proceso de extremos asociado, por lo que podemos considerar que se cumple la condición D'. Por lo tanto, disponemos de los resultados teóricos asintóticos que indican que el proceso de ocurrencia de extremos podrá aproximarse a un proceso de Poisson.

3.2.3. Series no *i.d.*: procesos No Homogéneos

Consideramos una serie de variables aleatorias con dependencia débil pero con distribuciones diferentes, lo que corresponde a algún tipo de dependencia temporal, ya sea por un comportamiento estacional o por la presencia de una tendencia. Dado que si la serie fuera *i.d.* el proceso de ocurrencia de extremos podría considerarse Poisson Homogéneo de parámetro λ, consideramos la siguiente generalización. Suponemos que el proceso de extremos de nuestra serie sigue un proceso Poisson No Homogéneo de intensidad $\lambda(t)$.

Pese a que no existe ninguna base teórica que justifique esta última suposición, se trata sin embargo de una práctica muy extendida por su buen funcionamiento [13] y [20].

3.3. Selección de los umbrales

Como hemos comentado, la elección del umbral u no es trivial. Hay tres aspectos a tener en cuenta: que el umbral tenga interés para la aplicación considerada, que sea lo suficientemente grande para que la aproximación POT sea válida y que el tamaño de la muestra de extremos obtenida sea lo suficientemente grande para realizar el procedimiento de inferencia con precisión.

Los últimos dos puntos son un compromiso entre sesgo y varianza que debe solucionarse seleccionando el **menor valor de u dentro del rango** en el que consideremos que la aproximación de nuestro modelo es razonable. Presentamos a continuación un método para la obtención de ese rango de u en el que la aproximación POT es razonable.

El método que hemos elegido para la obtención de un rango de umbrales "válidos" es el propuesto por Coles en [9]. Se basa en el carácter Pareto de los excesos Y que debemos observar en nuestra serie para poder suponer que la serie de ocurrencias sigue un proceso de Poisson. Una manera de comprobar que la distribución de Y es efectivamente la que buscamos es fijarse en la media de estos excesos, que en una Pareto debería valer:

$$E(Y) = \frac{\tilde{\sigma}}{1-\xi} = \frac{\sigma + \xi(u-\mu)}{1-\xi} \tag{3.1}$$

En la expresión anterior, los parámetros μ, σ y ξ son independientes del umbral u y serán teóricamente los mismos para cualquier serie que cumpla las condiciones de convergencia. Además tenemos que, si un determinado umbral u_0 es lo suficientemente grande como para observarse una distribución Pareto en los excesos obtenidos, entonces cualquier umbral $u > u_0$ será también lo suficientemente grande para que se observe el mismo comportamiento. Por lo tanto, si se cumple (3.1) para u_0, tendremos que, para todo $u > u_0$:

$$E(Y) = \frac{\sigma + \xi(u-\mu)}{1-\xi} = \frac{\tilde{\sigma}_{u_0} + \xi(u-u_0)}{1-\xi} \tag{3.2}$$

donde $\tilde{\sigma}_{u_0} = \sigma + \xi(u_0 - \mu)$.

Por lo tanto, en el rango de valores de u en el que la aproximación Pareto es correcta para los excesos definidos de esta manera, se debería cumplir (3.2) y por lo tanto deberíamos observar una dependencia lineal entre la esperanza de los excesos y el umbral escogido. Como estimador de la esperanza de nuestra serie de observaciones, podemos tomar la media muestral que nos permite obtener intervalos de confianza asumiendo que presenta normalidad. Podemos entonces obtener el rango de umbrales válidos de manera gráfica, representanto la media muestral de los excesos frente a u. Es lo que conocemos como *gráfica de media de vida residual* (*MRL* por sus siglas en inglés), que formalmente se representa por:

$$\left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} y : u < x_+ \right)$$

donde x_+ es el valor máximo observado de X_n y n_u el número de extremos observados sobre el umbral u. Este tipo de gráficas puede obtenerse numéricamente mediante el paquete de R *POT*.

En la práctica, observar la linealidad de la media muestral no siempre es fácil. Hay que tener en cuenta que cuanto más alto sea el valor de u, menos excesos tendremos para calcular la media muestral y peor calidad tendrá ésta como estimador de la esperanza. Por lo tanto, deberemos con frecuencia obviar de nuestra gráfica el rango superior de u en el que el número de observaciones extremas sea muy pequeño, y buscar la región lineal inmediatamente anterio. En la Figura 3.2 podemos ver un ejemplo ilustrativo de esta situación. En este caso, la media de los excesos presenta dos regiones que podrían considerarse lineales: [30,60] y [60,85], pero la segunda representa medias realizadas con un número muy pequeño de observaciones, por lo que no se considera fiable. Por lo tanto, en este caso se tomaría [30,60] como rango de umbrales válidos.

3.4. Estimación de los procesos Poisson No Homogéneos

La estimación de un Proceso de Poisson requiere principalmente la estimación de la intensidad. En el caso que nos interesa, el de una serie de variables no independientes ni estacionarias,

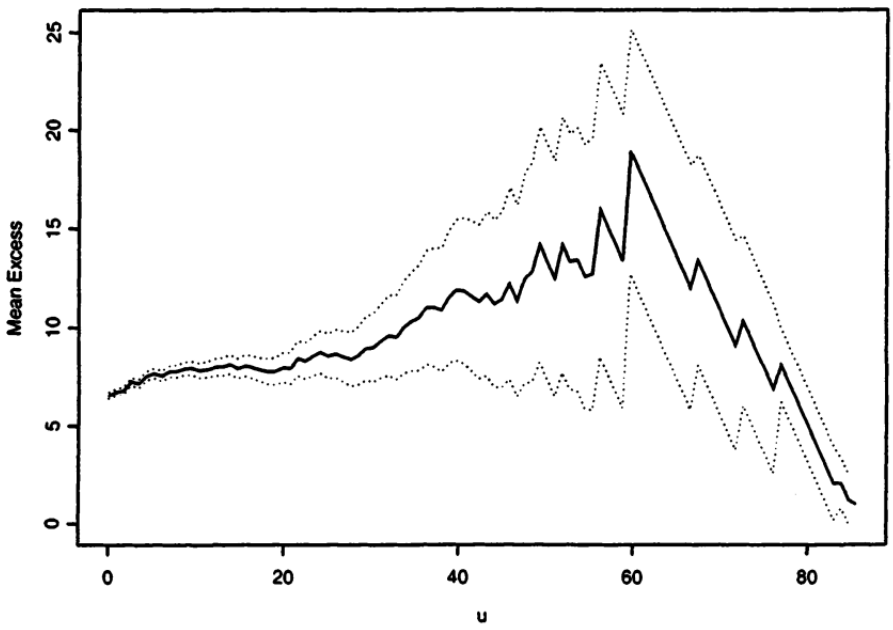


Figura 3.2: Ejemplo de *Mean Residual Life Plot*, tomado de [9]. Consideramos que las medias muestrales calculadas a partir de $u = 60$ no son fiables, por lo que tomamos [30,60] como rango de posibles umbrales.

el proceso de Poisson será No Homogéneo y la intensidad será por lo tanto dependiente del tiempo, $\lambda(\mathbf{t})$. Una posible modelización es plantear la intensidad como una función de m covariables dependientes del tiempo:

$$\lambda(t) = \exp(\mathbf{X}(\mathbf{t})^T \beta) \tag{3.3}$$

donde se ha tomado la función exponencial para garantizar que $\lambda(t) > 0 \; \forall t$.

- $\mathbf{X}(\mathbf{t}) = (1, X_1(t), ..., X_m(t))$ es el vector de covariables en el instante t, de dimensión $(1 \times m + 1)$.
- $\beta = (\beta_0, \beta_1, ..., \beta_m)$ es el vector de coeficientes de nuestro ajuste paramétrico, de dimensión $(1 \times m + 1)$.

La elección de las covariables $\mathbf{X}(\mathbf{t})$ que introduciremos en el modelo es seguramente la tarea más complicada del ajuste pues no tiene una respuesta única, ya que puede haber varios modelos igualmente buenos construidos a partir de diferentes covariables. En todo caso, comentaremos qué criterios seguir en el proceso de elección de variables en 3.4.2, para centrarnos primero en el proceso de estimación para un vector $\mathbf{X}(\mathbf{t})$ dado.

3.4.1. Estimación de la intensidad

La estimación de la intensidad se reduce en nuestro modelo a la estimación de los parámetros β. De entre los diferentes métodos de estimación paramétrica que existen, emplearemos el de **máxima verosimilitud (ML)**, que como principal ventaja nos permite obtener facilmente intervalos de confianza para los estimadores, ya que éstos se comportan con una distribución normal. La función de verosimilitud condicionada a las covariables vale [6]:

$$L\left(\beta; (t_i)_{i=1}^n\right) = \exp\left[-\int_A \lambda(t;\beta)dt\right] \prod_{i=1}^n \lambda(t_i;\beta)$$

donde $\lambda(t;\beta) = \exp(\mathbf{X}(\mathbf{t})^T \beta)$ y A es el espacio en el que se define el proceso puntual. Asumiendo que $\lambda(t_i;\beta)$ es constante en cada unidad de tiempo, hacemos el paso del continuo al discreto y

obtenemos la siguiente expresión para el logaritmo de L:

$$LL(\beta; (t_i)_{i=1}^n) = -\sum_{t=1}^{T} \lambda(t;\beta) + \sum_{i=1}^{n} \log \lambda(t_i;\beta) \tag{3.4}$$

donde T es la longitud del periodo de observación y n el número de observaciones disponibles.

En el caso de una serie de extremos obtenidos mediante la aproximación POT, es necesario modificar la función de verosimilitud. Esto se debe a que hemos considerado únicamente un evento extremo por cada conjunto de variables consecutivas que superan el umbral, correspondiente al máximo de este conjunto. Sin embargo, las covariables tomarán un "carácter extremo" en todos los puntos del evento y no únicamente en el máximo. Por lo tanto, la identificación del resto de puntos como no-ocurrencia (respuesta de nuestro modelo) podría llevar a una mala caracterización de la dependencia con los predictores del modelo. Para evitar esta situación, se propone eliminar todos aquellos puntos en los que se supere el umbral pero no se considere un extremo [6]. Una vez redefinida LL, debemos encontrar los parámetros β que la maximizan mediante alguna herramienta de optimización numérica.

Como comentábamos anteriormente, este método nos permite obtener un intervalo de confianza basado en la distribución asintótica de los estimadores ML. En efecto, la estimación de la varianza de $\hat{\beta}$ se puede calcular como:

$$\hat{V}(\hat{\beta}) = \left[-\left. \frac{\partial^2 LL(\beta)}{\partial\beta_l \partial\beta_k} \right|_{\beta=\hat{\beta}} \right]^{-1}$$

donde:

$$\frac{\partial^2 LL(\beta)}{\partial\beta_l \partial\beta_k} = -\sum_{t=1}^{T} \lambda(t;\beta) X_l(t) X_k(t)$$

Además, podemos obtener el intervalo de confianza para el predictor lineal $\nu(t) = \mathbf{X(t)}^T\beta$ en función de la varianza de $\hat{\beta}$ obtenida:

$$\hat{\nu}(t) \pm z_{1-\alpha/2} s.e.(\hat{\nu}(t)) \quad \text{con} \quad s.e.(\hat{\nu}(t)) = \sqrt{\mathbf{X(t)}^T \hat{V}(\hat{\beta}) \mathbf{X(t)}} \tag{3.5}$$

Así pues, dado que $\lambda(t) = \exp(\nu(t))$, basta con hacer una transformación exponencial del intervalo de confianza (3.5) para obtener el **intervalo de confianza para la intensidad** $\lambda(t)$:

$$\exp\left[\log(\hat{\lambda}(t)) \pm z_{1-\alpha/2} \cdot s.e.(\hat{\nu}(t))\right]$$

3.4.2. Selección de covariables

Consideramos que tenemos un conjunto de variables $\mathbf{X_T(t)} = X_1(t), X_2(t), ...$ que toman valores en cada instante de tiempo t dentro del intervalo temporal en el que consideramos nuestro proceso puntual. Queremos establecer un proceso que nos permita seleccionar el subvector $\mathbf{X(t)} \subset \mathbf{X_T(t)}$ que introduciremos en nuestro modelo (3.3). La idea principal es partir de un modelo "vacío" (sin covariables) y probar qué covariables pueden introducirse para mejorar el modelo. Una vez elegida una, probar de nuevo cuál puede mejorar el modelo para volver a añadirla y continuar el proceso, una a una, hasta que ninguna variable mejore el modelo construido. Es lo que se conoce como proceso *paso por paso hacia delante* (*stepwise-forward*.

Así pues, en cada uno de los pasos de selección, tenemos un modelo M_0 construido con l variables y queremos probar si otro modelo M_1, construido con $l+1$ variables, es preferible. Para ello se utiliza el **test de razón de verosimilitud** (**RL**, por sus siglas en inglés), que está fundamentado en el siguiente resultado ([9] p.35). Si consideramos la función de desviación

$$D = 2[LL(M_1) - LL(M_0)]$$

tenemos que, para ciertas condiciones de regularidad, ésta se comporta como un χ^2_k. Por lo tanto, podemos rechazar M_0 en favor de M_1 con un nivel de significación α si $D > c_\alpha$, donde c_α es el $(1-\alpha)$ cuantil de χ^2_1. En la práctica, definiremos el **p-valor** del test RL como el cuantil de χ^2_1 al que corresponde el valor D. Por lo tanto, con un nivel de significancia del 95 %, aceptaremos la nueva variable (introducida en el modelo M_1), siempre que el p-valor del test sea inferior a 0.05.

3.5. Validación del modelo CPSP

Una vez establecido el modelo, debemos evaluar su validez para representar los datos de los que disponemos. Principalmente deberemos comprobar dos cosas: que cada uno de los tres procesos se comporta aproximadamente como un NHPP y que los tres son indenpendientes entre sí.

El primer punto requiere de la siguiente transformación temporal:

$$t^* = \int_0^t \lambda(u)du$$

que transforma nuestro proceso de Poisson No Homogéneo en un proceso de Poisson Homogéneo de intensidad unidad, simplificando su análisis. Así pues, una vez hecha esta transformación, deberemos comprobar el carácter Poisson de la serie obtenida. Basaremos esta comprobación en dos aspectos: el carácter exponencial de los tiempos de espera y el buen ajuste de la intensidad estimada.

Definimos en primer lugar las *distancias inter-eventos* $d_i = t^*_i - t^*_{i-1}$, donde t^*_i es el instante de ocurrencia del evento i. Bajo la hipótesis nula, estos eventos deberían ser una muestra *i.i.d.* exponencial o, equivalentemente, los *residuos* $r_i = \exp(-d_i)$ deberían ser una muestra *i.i.d* uniforme. Para comprobar que se cumple esta última característica, comprobaremos por un lado la incorrelación entre los residuos y a continuación la distribución uniforme de éstos.

3.5.1. Incorrelación de los residuos

En primer lugar podemos estudiar la correlación de primer orden, es decir, la correlación que tiene cada uno de los n residuos con el anterior. Para estudiar esta correlación, podremos representar la ***gráfica de correlación de primer orden***, que viene dada por los puntos:

$$(r_{i-1}, r_i : i = 2, ...n)$$

Podremos acompañarla de un ajuste lineal para descartar posibles dependencias.

Esta correlación entre los residuos puede cuantificarse mediante el cálculo del coeficiente de correlación de Pearson, que para dos variables X e Y cualesquiera vale $Cor(X,Y) = Cov(X,Y)/\sqrt{V(X)\cdot V(Y)}$, donde Cov representa la covarianza y V la varianza. Así pues, la correlación de primer orden de nuestra serie puede estimarse mediante la siguiente expresión:

$$Cor(r_i, r_{i-1}) = \frac{n\sum_{i=2}^n r_i r_{i-1} - \sum_{i=2}^n r_i \sum_{i=2}^n r_{i-1}}{\sqrt{n\sum_{i=2}^n r_i^2 - (\sum_{i=2}^n r_i)^2}\sqrt{n\sum_{i=2}^n r_{i-1}^2 - (\sum_{i=2}^n r_{i-1})^2}}$$

Además, este estimador presenta una distribución t con $n-3$ grados de libertad, lo que nos permite definir el **test de incorrelación**. Si el p-valor de este test vale más de 0.05, consideramos que no hay indicios para rechazar la incorrelación.

Para evaluar correlaciones de órdenes superiores, una herramienta muy útil es la gráfica ***ACF***, que representa la correlación entre residuos separados por h instantes de tiempo $\rho(h) = Cor(r_i, r_{i-h})$, para diferentes valores de h:

$$(h, \hat{\rho}(h) : h = 1, 2, ...)$$

Además, el estimador $\hat{\rho}(h)$ tiene un comportamiento normal insesgado por lo que podemos obtener una banda de confianza dentro de la cual cualquier estimación de $\rho(h)$ puede considerarse nula al nivel de confianza α establecido. De esta manera, si hay una fracción $(1-\alpha)$ o menos de los $\hat{\rho}(h)$ estimados fuera del intervalo de confianza, podemos asumir que no existen correlaciones de órdenes superiores.

3.5.2. Distribución uniforme de los residuos

Además de la incorrelación, debemos comprobar que los residuos obtenidos presentan una distribución uniforme. Para ello la mejor herramienta es un ***qqplot***, es decir, una gráfica que represente los sucesivos cuantiles de la muestra frente a los cuantiles teóricos de una serie de *v.a.* uniformes. Si la muestra es uniforme, los puntos deberían ajustarse a la recta unidad, por lo que podemos añadir ésta en nuestra gráfica junto con un intervalo de confianza construido gracias a la distribución beta de los cuantiles.

Para cuantificar la similitud de nuestra muestra con una muestra uniforme realizamos el **test Kolmogorov-Smirnov**. Éste nos devuelve un p-valor que, para valores inferiores a 0.05, nos indicará el rechazo de la hipótesis de uniformidad bajo un nivel de confianza del 95 %.

3.5.3. Media nula de los *residuos brutos*

Además del análisis de los residuos uniformes, podemos realizar un análisis basado en los denominados *residuos brutos*. Si los anteriores residuos correspondían a la distribución exponencial de las distancias inter-evento, éstos corresponden al error cometido en la estimación del número de eventos en un intervalo. Se definen de la siguiente manera (en su versión escalada) para cada intervalo (l_1, l_2):

$$r(l_1, l_2) = \frac{1}{l_2 - l_1} \left(\sum_{t_i \in (l_1, l_2)} h(t_i) I_{t_i} - \int_{l_i}^{l_2} h(t_i) \hat{\lambda}(u) du \right)$$

donde I_{t_i} vale 1 en el instante t_i y $h(t_i)$ es la función de peso, que vale $h(u) = 1/\sqrt{\hat{\lambda}(u)}$ en el caso de los **residuos de Pearson**, los que emplearemos. Estos residuos permiten comprobar que la estimación de la intensidad es correcta, ya que bajo un modelo correcto $r(l_1, l_2) \approx 0$. Para comprobar esta situación, representamos los residuos obtenidos **frente al tiempo** acompañados de un ajuste y un intervalo de confianza dado por la aproximación de Baddeley et al. (2005) [4].

3.5.4. Independencia de los procesos POT

Este punto será simplemente comprobado mediante el test de independencia propuesto en Abaurrea et al. (2013) [1] y disponible en la librería de R *IndTestPP*.

Capítulo 4

Aplicación: estudio de la serie de temperaturas de Panticosa

En estes capítulo pondremos en práctica los métodos hasta ahora presentados para la modelización de ocurrencia de extremos en una serie bivariante de temperatura. Tomaremos para ello la serie de temperaturas máximas y mínimas diarias de la estación meteorológica del Balneario de Panticosa (Huesca). Como mencionábamos en la introducción, el estudio climático en el entorno del Pirineo tiene gran interés por su especial vulnerabilidad al cambio climático y por su valioso patrimonio natural. De todas las estaciones meteorológicas pertenecientes a la AEMET (Agencia Estatal de Meteorología) que se encuentran en la parte aragonesa de esta cordillera (la zona con más altitud y con más presencia de glaciares), la estación elegida tiene gran interés por disponer de observaciones en un amplio rango temporal. El periodo que elegimos para el estudio de los extremos es el de *invierno* o *DEF* (diciembre, enero y febrero), por su importancia en la localidad, tanto a nivel ecológico como turístico.

Realizaremos por tanto el análisis de extremos con los datos disponibles de esta serie, cuyas características principales presentaremos en la primera sección. Continuaremos después con todo el proceso de modelización de la ocurrencia de extremos para las dos series de temperatura. Presentaremos en primer lugar el análisis pertinente para la selección de umbrales suficientemente grandes, para continuar con la modelización de los procesos Poisson No Homogéneos, cuya intensidad será modelizada mediante la dependencia con diferentes covariables. Finalmente, comprobaremos la validez de los modelos obtenidos mediante diferentes técnicas propuestas.

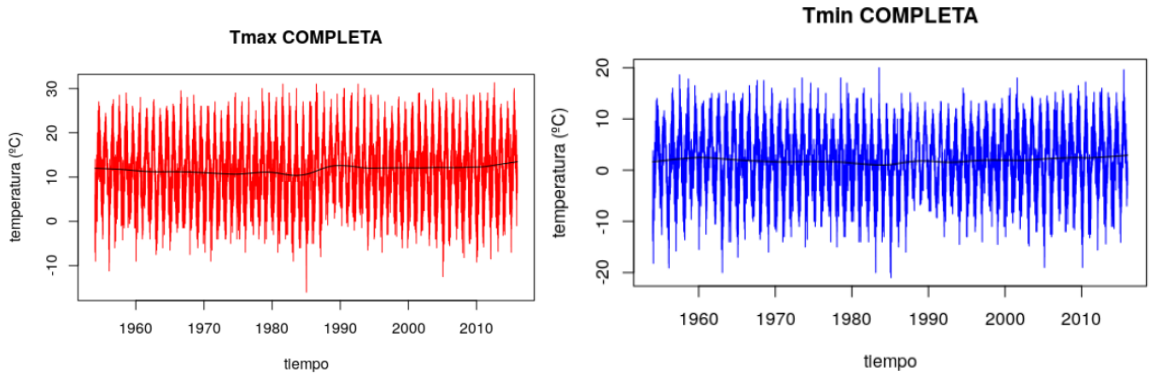
Todos los pasos aquí presentados han sido implementados mediante los paquetes *NHPoisson*, *POT* y *IndTestPP* del lenguaje de programación estadística R.

4.1. Presentación de los datos y análisis preliminar

Los datos han sido proporcionados por AEMET. Corresponden a las medidas hechas en la estación meteorológica con identificador 9451, coordenadas $42{,}763\,^{o}\,N - 0{,}234\,^{o}\,O$ y altitud 1660 m. Se tratan de las observaciones diarias realizadas desde el 1 de enero de 1954 al 31 de diciembre de 2015. En cada fecha tenemos una medición de la temperatura máxima diaria y una medición de la temperatura mínima, ambas en grado Celsius. Esta serie ha sido construida a través de la unión de dos series, la correspondiente a la estación "Balneario" (con datos entre 1941 y 1993) y la estación "Casa de Piedra" (con datos entre 1993 y 2018), ambas situadas en el complejo turístico del Balneario de Panticosa. Entre los años 1987 y 1993, se han tenido que imputar los datos, pues éstos faltaban. El método empleado para ésto, ha sido de tipo regresión, basada en los cuatro puntos de la red SPAIN02 que rodean la estación: los correspondientes con las coordenadas $42{,}8\,^{o}\,N - 0{,}3\,^{o}\,O$, $42{,}8\,^{o}\,N - 0{,}2\,^{o}\,O$, $42{,}7\,^{o}\,N - 0{,}3\,^{o}\,O$ y $42{,}7\,^{o}\,N - 0{,}2\,^{o}\,O$. Para saber más sobre este tipo de métodos de interpolación, consultar [15].

Analizamos en primer lugar la evolución general de la temperatura en Panticosa que nos

muestran los datos. Presentamos en la Figura 4.1 las dos series de temperatura junto con un suavizado polinomial que nos da la tendencia general (tomando como ventana de tiempo una década). Podemos observar una **tendencia ligeramente creciente en ambas series**, más marcada en la serie de máximos. Efectivamente, el valor máximo del suavizado se encuentra para ambas series en el último año (2015), observándose un aumento total en el periodo 1954-2015 de 1.5ºC para T_{max} y 1.3ºC para T_{min}. Esta observación está en concordancia con lo observado en promedio en España, que es un aumento de aproximadamente 1ºC en la temperatura media durante la segunda mitad del siglo XX [8]. Se observa también una pequeña reducción de la variabilidad en el periodo 87-93, que corresponde a los datos interpolados.



(a) (b)

Figura 4.1: Series de temperaturas máximas (a) y mínimas (b) diarias y suavizado polinomial con ventana de 10 años para cada una.

Al contrario que en la serie completa (con todos los meses del año), **en la serie de *invierno* (DEF) no se observa una tendencia clara**. Esto da indicios de una evolución climática desigual en las estaciones del año. Para evaluar estas diferencias, realizamos un análisis de la evolución de estas series en cada uno de los meses del año. Presentamos en la Figura 4.2 la evolución general de las temperaturas máximas y mínimas en cada mes del año, agrupadas según estaciones. En general podemos observar comportamientos muy diversos en los diferentes meses y tendencias con bastante variabilidad. Como comportamientos más destacables, podemos decir que ha habido un claro **aumento de la temperatura en verano**, que rondaría los 2ºC. Los meses de invierno, que utilizaremos para hacer el análisis, presentan en general una tendencia especialmente constante en comparación con el resto, a excepción de la ligera disminución en el mes de febrero, bastante acentuada en las mínimas[1]. Esto da como resultado una tendencia constante o incluso ligeramente decreciente en el promedio de los tres meses *DEF*.

4.2. Selección de los umbrales

Para obtener una serie de extremos, debemos en primer lugar definir el umbral a partir del cual definiremos una observación como "extrema". Como hemos comentado en la sección 3.3, la mejor solución es el método gráfico a través de un "Mean residual life plot" (gráfico de vida media residual), que llamaremos *MRLP* de ahora en adelante. Éste nos permite obtener un rango en el cual la aproximación *POT* es válida: el rango en el que el valor $E(X > u|X > u_0)$ es lineal respecto de u.

[1]Hay que tener en cuenta, sin embargo, que el suavizado polinomial es menos fiable en las colas.

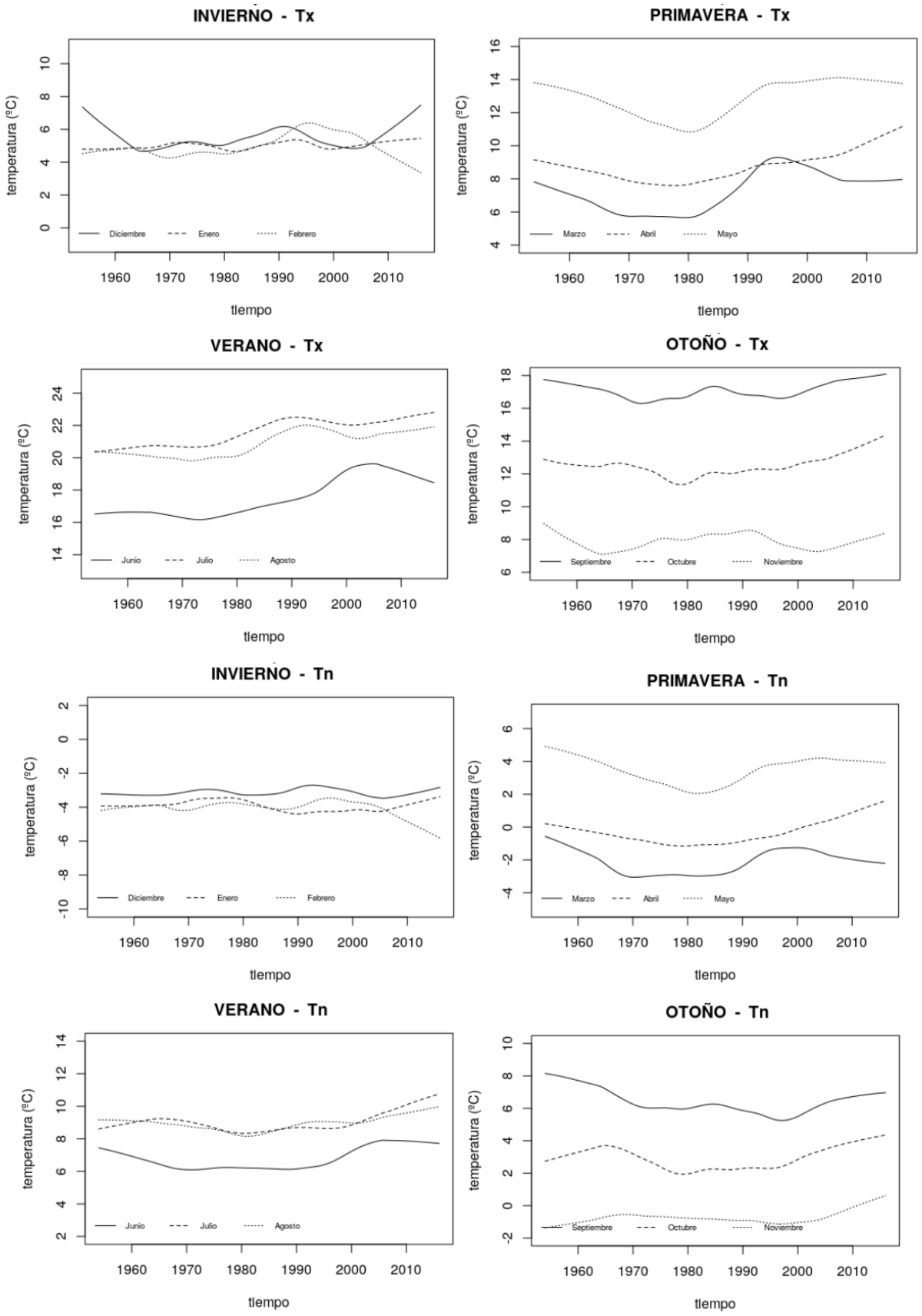


Figura 4.2: Evolución de las temperaturas máximas (T_x) y mínimas (T_n) diarias de Panticosa (suavizado polinomial con ventana de 20 años) en cada mes, agrupadas por estaciones. El rango de la escala en el eje Y es de 12ºC en todas las gráficas para facilitar su comparación.

En la Figura 4.3 podemos ver las gráficas correspondientes a la serie de mínimos y a la de máximos, además de un ajuste a una recta que define nuestro rango de posibles umbrales. En la *MRLP* de las mínimas, observamos la presencia de un pico final con una gran incertidumbre en los valores superiores a u =5ºC, que no debe considerarse fiable pues se trata de medias hechas con muy pocos valores (sólo hay un 15 valores superiores a 5). Podemos considerar por lo tanto que el rango en el que la media de los excesos presenta un comportamiento lineal es la inmediatamente anterior a u =5ºC, que según la recta trazada en nuestra gráfica es el rango [0,5]. Este rango nos permite elegir el **valor 0ºC como umbral de nuestra serie de mínimos**, valor con especial significado físico, ya que entonces un "día extremo en la mínima" será aquel en el que no hiele durante la noche. Teniendo en cuenta que la nieve funde a temperaturas superiores a 0ºC, la presencia temperaturas negativas por la noche es imprescindible para mantener el manto nivoso y compensar la fusión que se produce durante el día. En nuestra serie de temperaturas, el 0 representa el **percentil 10.8**. Un hipotético aumento de este porcentaje representaría por lo tanto una disminución de los días con nieve en Panticosa.

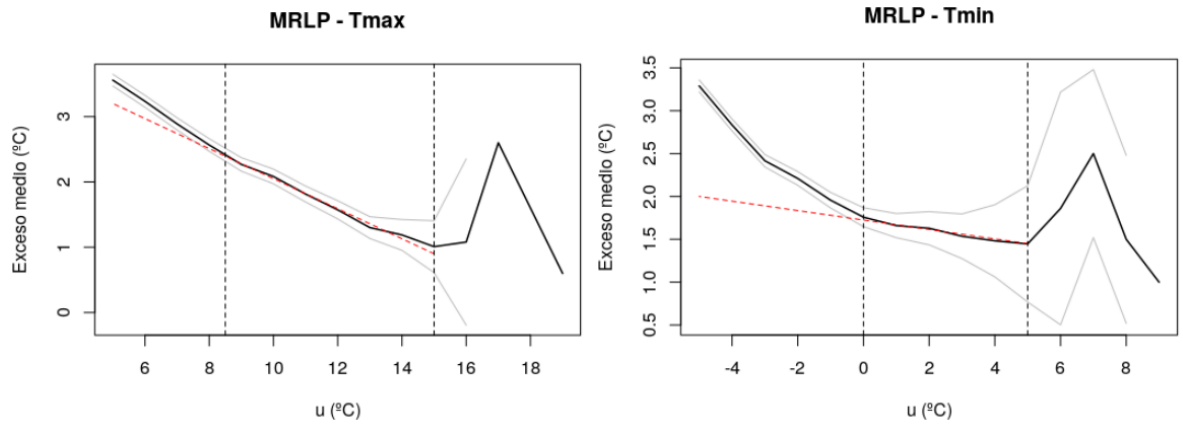


Figura 4.3: Gráfico de vida media residual para las series de temperaturas mínimas y máximas. La recta trazada indica la región con comportamiento lineal, que nos da el rango de umbrales (región entre las líneas discontinuas) para los cuales la aproximación *POT* es válida.

Analizamos a continuación la *MRLP* correspondiente a la serie de máximos. Se observa un gran aumento de la incertidumbre para valores superiores a u =15ºC, por lo que tomamos este valor como límite máximo. Trazando una recta, podemos definir una región aproximadamente lineal que empezaría en u =8.5ºC o incluso antes. En cualquier caso, elegimos un umbral que represente un percentil similar al que hemos hallado en las mínimas (10.8), por lo que escogemos el **valor 10ºC como umbral de nuestra serie de mínimos**, que representa el percentil **11.2** (el siguiente valor entero, 11, representa el percentil 7.3, bastante menor).

4.3. Ajuste del modelo CPSP

4.3.1. Obtención de los procesos independientes

Presentamos ahora el proceso de modelización para la ocurrencia de extremos en nuestras series de temperatura. En el marco del modelo CPSP, esto requiere la definición de tres series de extremos: T_n (extremos únicamente en la serie de mínimas), T_x (extremos únicamente en las máximas) y T_{nx} (extremos en ambas) que suponemos independientes. Cada uno de estos procesos de ocurrencia obtenidos, será modelizado después por un Proceso de Poisson No Homogéneo tal y como hemos comentado en el capítulo 3. Recordemos que definimos la ocurrencia de un extremo como el punto de ocurrencia del máximo de cada cluster de observaciones consecutivas que superan el umbral. Además, si dos clusters de T_{max} y T_{min} se solapan en un cierto rango de

instantes de tiempo, ese rango se considera un cluster diferente cuyo máximo es definido como un extremo de T_{nx} (considerando la suma de T_{max} y T_{min} en el cálculo del máximo). Ilustramos este proceso para el rango de un año en la Figura 4.4.

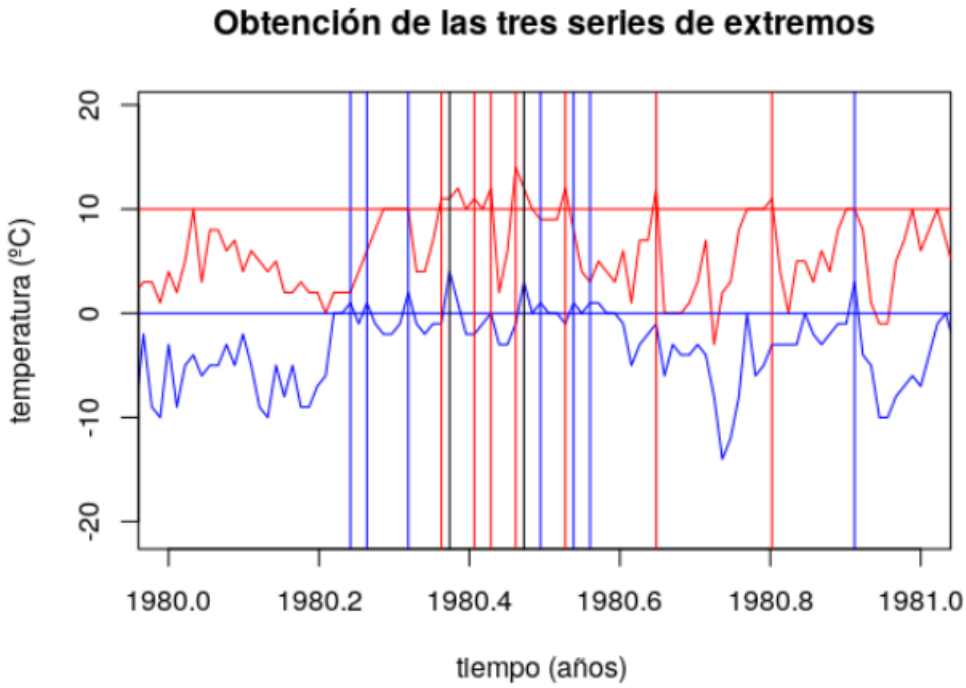


Figura 4.4: Obtención de las tres series de ocurrencias de extremos a partir de las dos series de temperaturas diarias en el año 1980. La serie T_n corresponde a las líneas azules, la serie T_x a las líneas rojas y T_{nx} a las negras.

Antes de comenzar el ajuste, analizamos el comportamiento general de estas tres series obtenidas. Podemos observar en la Figura 4.5 la evolución del "ratio empírico de ocurrencia" anual para cada serie, es decir la evolución del porcentaje de extremos ocurridos en cada año. En general, podemos observar que el ratio anual de extremos presenta gran variabilidad pero se mantiene bastante constante en promedio a lo largo del tiempo, salvo en la serie de máximos, donde se observa un ligero aumento a partir de 1990. Esto cuadra con las observaciones hechas en la sección 4.1, donde recalcábamos la ausencia de una tendencia clara en las temperaturas de los meses de invierno y una variabilidad constante.

4.3.2. Covariables potenciales en series de temperatura

Una vez definidas las series de extremos, debemos construir las covariables que intentaremos introducir en nuestro ajuste a un PPNH. En nuestro caso, no debemos olvidar que el fin último de nuestra modelización es obtener proyecciones de nuestro proceso sobre el futuro, por lo que debemos considerar covariables cuyos valores futuros (en el siglo XXI) puedan ser obtenidos. En el caso de la temperatura, existen modelos planetarios que pueden ser reescalados a escala local para obtener las proyecciones de una serie de temperatura (ver capítulo 5). Sin embargo estas proyecciones no son fiables a escala diaria [28], por lo que consideraremos siempre variables agregadas que promedien varias observaciones y representen la tendencia de nuestras temperaturas, además de otros términos que den cuenta del carácter estacional de la serie. Por lo tanto, consideraremos a priori los siguientes predictores:

1. **Armónicos**: Representan la componente puramente estacional de las series, y vienen dados

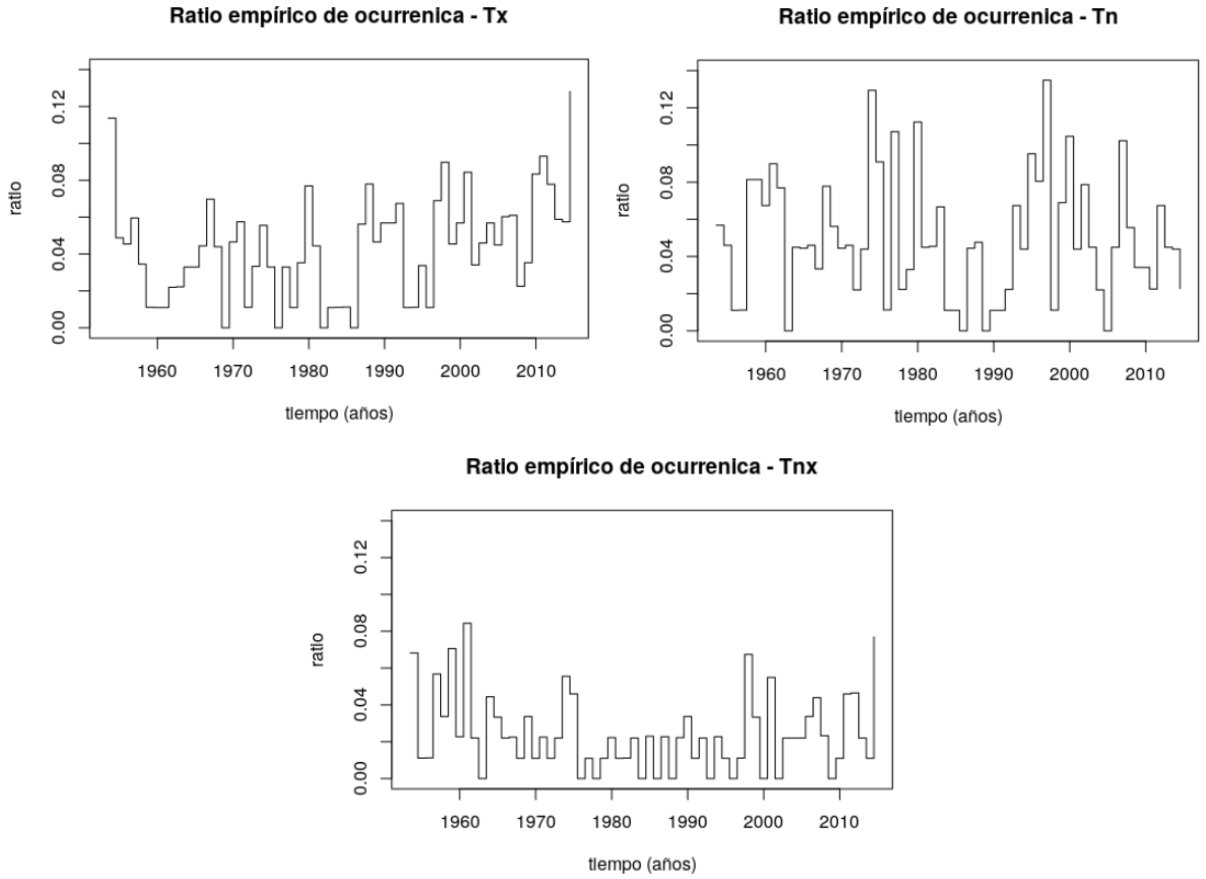


Figura 4.5: Evolución del ratio empírico de ocurrencia anual de extremos para las tres series.

por las expresiones

$$\cos\left(2\pi n\frac{d_i}{T}\right)\ ,\ \sin\left(2\pi n\frac{d_i}{T}\right)$$

donde T es el periodo estacional (el más grande observable en la serie) y d_i es el índice de cada observación dentro de ese periodo. En nuestro caso tenemos que T vale 365 y d_i corresponde al número del día i en su año. Podemos definir infinitos armónicos según el valor que tome n (1, 2, ...), que corresponden a tomar los periodos T, $T/2$, $T/3$, etc.

2. **Tendencias a corto plazo**: Representan la tendencia de la serie en las últimas observaciones. Definiremos dos: la media móvil realizada con los 15 días anteriores (T_{x15} y T_{n15} para las máximas y mínimas respectivamente) y la media móvil realizada con los 31 días anteriores[2] (T_{x31} y T_{n31}). Los valores de estas medias en las primeras observaciones corresponderán sin embargo a la media de los únicos valores anteriores disponibles, hasta llegar al día 15 o 31, según el caso. Es decir, para T_{x15}, por ejemplo, el primer valor será el de la primera observación de la serie, el segundo será la media entre el primero y el segundo, y así sucesivamente hasta el día 16, a partir del cual ya se tomarán solamente los 15 días anteriores para hacer la media.

3. **Términos cuadráticos**: Consideraremos la posibilidad de introducir los términos cuadráticos de las covariables que incluyamos en el modelo, por si una dependencia lineal fuera insuficiente.

[2]El hecho de realizar la media con los días anteriores y no de manera centrada (la mitad de los datos tomados del pasado y la mitad del futuro) nos permitirá más adelante justificar más razonablmente el uso de este modelo para la proyección de datos futuros.

4. **Términos de interacción**: Si algún armónico se introduce en el modelo, consideraremos la posibilidad de introducir su interacción con el resto de covariables incluídas (el producto de ambas covariables).

No se ha considerado la introducción de tendencias a largo plazo como predictores de la intensidad, ya que suelen presentar situaciones de extrapolación en la proyección, lo que impide la correcta estimación del proceso para situaciones futuras. La ilustración de este problema para nuestra serie de temperaturas se puede consultar en el anexo A.

Estas serán por lo tanto el conjunto de variables $\mathbf{X_T(t)}$ cuya posible introducción en el modelo iremos evaluando una a una mediante el test *RL*. En el proceso que hemos descrito, es importante el orden en el que evaluemos el test, ya que la inclusión de una variable puede ser positiva respecto a un modelo M_1 y sin embargo no serlo respecto a otro modelo M_2 (generalmente si éste contiene más variables que M_1). Por esta razón, establecemos unos ciertos criterios en cuanto al orden a llevar a cabo. Priorizaremos en primer lugar los armónicos sobre el resto de predictores ya que, si existe una fuerte componente estacional, ésta podría ser modelizada parcialmente por alguna variable a corto plazo (sugiriendo el test *RL* su inclusión), pero siempre será más precisa su modelización mediante armónicos. Por otro lado, priorizaremos las tendencias con menor periodo de agregación, por su mayor variación que puede aportar más información sobre posibles cambios temporales. En resumen, consideraremos el siguiente orden:

- Armónicos (empezando por los de primer orden).
- Tendencias a corto plazo de la serie correspondiente (T_{max} en la modelización de T_x y T_{min} en la de T_n). Se evalúa el promedio de 15 días y depués el de 31.
- Tendencias a corto plazo de la otra serie.
- Términos cuadráticos de los predictores introducidos.
- Interacción de las tendencias introducidas con los armónicos, si es que han entrado en el modelo.

4.3.3. Resultados

Proceso $\mathbf{T_x}$

1. Comenzamos creando un modelo sin predictores para nuestra serie de extremos, que comparamos con un modelo en el que introducimos el primer armónico. El test *RL* nos devuelve un p-valor de 0.20, por lo que se rechaza la introducción del armónico. Probamos a continuación el test entre el modelo sin predictores y uno que incluya T_{x15}, y obtenemos un p-valor 0, por lo que tomamos $\mathbf{T_{x15}}$ como primera variable del modelo.

2. Comparamos el modelo obtenido con modelos que incluyan además otras variables. Continuamos con la tendencia de T_{max}, obteniendo un p-valor de 0.055 en el test para la introducción de T_{x31}, por lo que rechazamos esta covariable. Obtenemos sin embargo 0 en el caso de T_{n15}, por lo que tomamos $\mathbf{T_{n15}}$ como segunda variable del modelo.

3. La covariable T_{n31} devuelve un p-valor de 0.28, por lo que es rechazada. Como no se ha introducido ningún armónico, sólamente nos queda probar la posible introducción de los términos cuadráticos de las dos variables introducidas. El test nos devuelve para ambas un p-valor de 1, por lo que rechazamos su inclusión, terminando el proceso de exploración.

El modelo obtenido es el siguiente:

$$M_x: \quad \log(\hat{\lambda}_x(t)) = -6{,}09 + 0{,}40T_{x15} - 0{,}15T_{n15} \tag{4.1}$$

Proceso $\mathbf{T_n}$

1. El p-valor del test RL vale 0.14 para la inclusión de los armónicos como primer predictor, por lo que rechazamos su introducción. La inclusión de $\mathbf{T_{n15}}$ devuelve 0 sin embargo, por lo que la introducimos en el modelo.

2. La otra variable de tendencia de T_{min}, T_{n31}, presenta un p-valor de 0.57, por lo que pasamos a evaluar la tendencia de T_{max}. Obtenemos 0 como resultado del test RL para $\mathbf{T_{x15}}$, por lo que introducimos esta variable.

3. La posible introducción de T_{x31} se rechaza con un p-valor de 0.60. De nuevo, no hay posibles términos de interacción y los términos cuadráticos son rechazados también con p-valores 1, por lo que terminamos nuestro proceso.

El modelo obtenido es el siguiente:

$$M_n: \quad \log(\hat{\lambda}_n(t)) = -0{,}71 + 0{,}41T_{n15} - 0{,}17T_{x15} \tag{4.2}$$

Proceso $\mathbf{T_{nx}}$

1. Rechazamos la inclusión de los armónicos con p =0.78. Exploramos ahora las variables de tendencia sin priorizar las de la de una serie sobre la otra. Como obtenemos p-valores nulos tanto para T_{n15} como para T_{x15}, nos fijamos en el valor del ratio devuelto por el test RL, que vale 53 en el primer caso y 68 en el segundo. Esto quiere decir que la función de verosimilitud del modelo con T_{n15} vale 53 veces lo que la función de verosimilitud del modelo sin predictores, y 68 veces en el caso del modelo con T_{x15}. Por esta razón, introducimos $\mathbf{T_{x15}}$ como primera covariable del modelo.

2. La posterior inclusión de T_{n15} sigue dando un p-valor favorable, 0.016, por lo que introducimos esta variable.

3. Las variables T_{n31} y T_{x31} no parecen en cambio aportar gran cosa pues devuelven p-valores de 0.35 y 0.42 respectivamente. Finalmente, la introducción de términos cuadráticos es rechazada con p-valores 1.

El modelo obtenido es el siguiente:

$$M_{nx}: \quad \log(\hat{\lambda}_{nx}(t)) = -4{,}57 + 0{,}22T_{x15} + 0{,}15T_{n15} \tag{4.3}$$

Presentamos a modo de resumen la Tabla 4.6 con los valores estimados para los parámetros de los tres modelos, que comentamos a continuación. Como podemos observar, el proceso T_x presenta un coeficiente positivo respecto a la covariable T_{x15} y uno negativo respecto a T_{n15}, más pequeño en valor absoluto. Este primer coeficiente significa que la intensidad del proceso de ocurrencia de un extremo en la serie T_x está positivamente correlacionada con la temperatura media en las máximas de los 15 días anteriores, como era de esperar. Sin embargo, el coeficiente negativo, no tiene por qué decir que la superación del umbral en las máximas esté negativamente correlacionada con las 15 observaciones anteriores de la serie de temperaturas mínimas. Hay que tener en cuenta que la serie T_x para la que estamos estimando la intensidad corresponde a las temperaturas que superan el umbral únicamente en T_{max}, y no aquellas que lo superan en T_{max} y T_{min}. Por lo tanto, cuando la temperatura mínima aumenta (T_{n15}), también aumenta la probabilidad de superar el umbral en T_{min}, lo que puede provocar un extremo en T_{nx}, que es incompatible con un extremo en T_x. Nos encontramos la situación inversa en el caso de T_n: un coeficiente positivo respecto a T_{n15} -el más relevante-, que puede interpretarse como una correlación directa respecto al comportamiento de las mínimas en los días anteriores, y un

	Térm. indep.	T_{x15}	T_{n15}
T_x	-6.09	0.40	-0.15
T_n	-0.71	-0.17	0.41
T_{nx}	-4.57	0.22	0.15

Figura 4.6: Parámetros β para la estimación de la intensidad de cada proceso.

coeficiente negativo respecto a T_{x15}, que puede estar relacionado con la ocurrencia de extremos en T_nx. En el caso de esta última serie, ambos coeficientes son positivos.

En la Figura 4.7, podemos ver la comparación de las intensidades empírica y ajustada para cada modelo, calculadas en intervalos de 4 años para facilitar la visualización. Observamos ajustes bastante buenos, que captan el orden de magnitud de las intensidades así como la tendencia general de éstas.

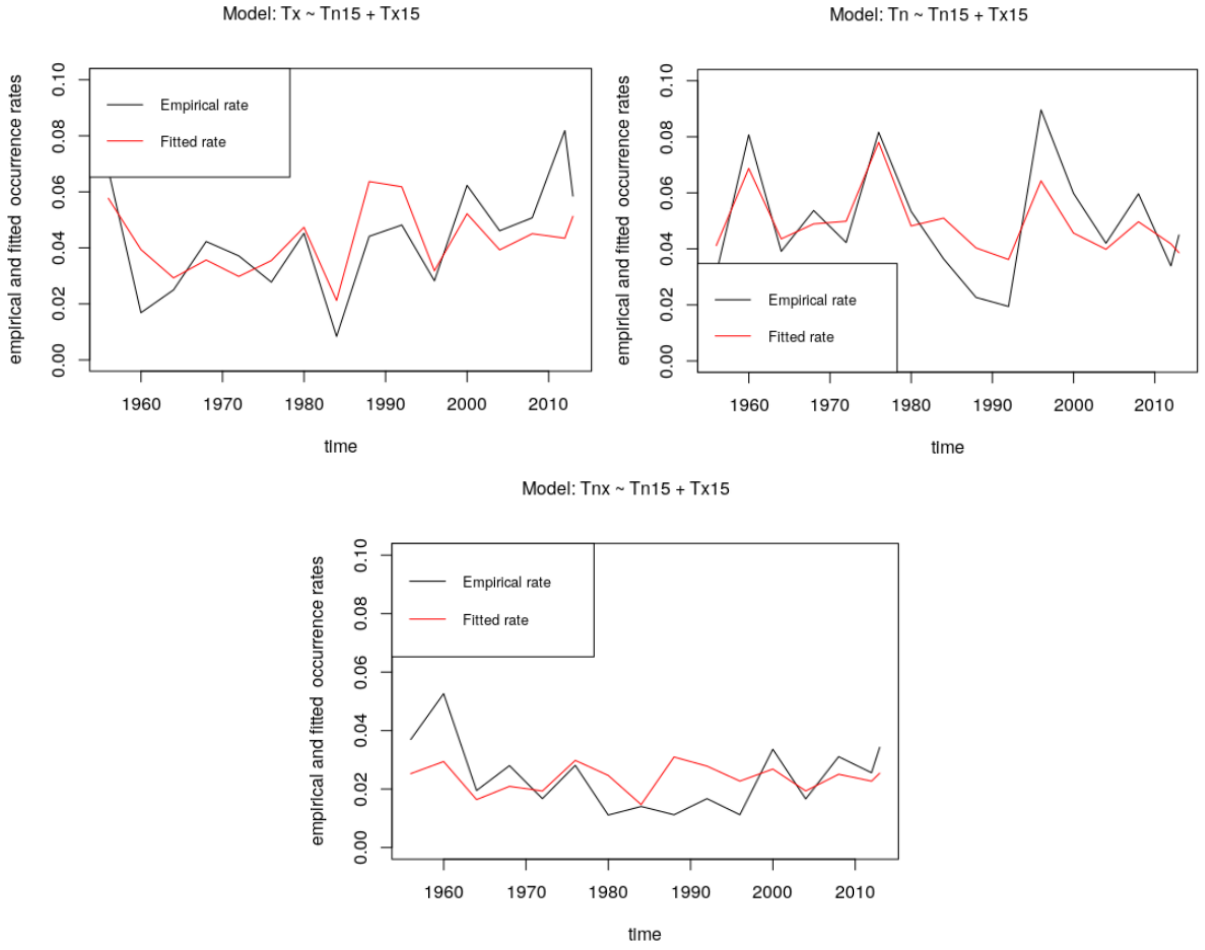


Figura 4.7: Ratio de ocurrencia de extremos estimado (*fitted rate*) y observado (*empirical rate*) para cada serie. Los ratios están calculados para intervalos de 361 días (inviernos de 4 años).

4.4. Validación del modelo

Analizaremos a continuación la validez del ajuste realizado con cada modelo al proceso de ocurrencia de extremos asociado. Por un lado, analizaremos el comportamiento de los residuos $\{r_i\}$ del modelo (ver definición en 3.5), que deben ser incorrelados entre ellos y presentar una distribución uniforme. Por otro lado, analizaremos los residuos brutos que deben presentar una

media nula. Una vez realizada la validación de cada modelo, comprobaremos que los tres procesos son independientes entre sí.

Modelo para el proceso $\mathbf{T_x}$ Obtenemos los residuos de nuestro proceso T_x y analizamos su correlación, obteniendo un coeficiente de Pearson de 0.01 y un p-valor 0.85 para el test de incorrelación, por lo que suponemos que no existe correlación de primer orden. Presentamos en la primera gráfica de la Figura 4.8 el *gráfico de correlación de primer orden* con su debido ajuste lineal, que muestra una nube de puntos sin aparente dependencia, confirmando la hipótesis de incorrelación. En la segunda gráfica, observamos el ACF de nuestros residuos, que presenta dos valores fuera del intervalo de confianza. Sin embargo, éstos valores son pequeños y representan un porcentaje aceptable dentro del nivel de confianza establecido (95 %), por lo que parece razonable asumir incorrelación también en los órdenes superiores.

Para verificar la distribución uniforme de nuestros residuos, representamos el *qqplot* correspondiente (tercera gráfica de 4.8). Observamos un buen ajuste a la recta unidad salvo en los cuantiles superiores, donde los correspondientes a los residuos parecen tomar valores inferiores a los teóricos. Sin embargo, los puntos que salen fuera de los intervalos de confianza representan un porcentaje pequeño de la muestra y nuestro test Kolmogorov-Smirnov arroja un p-valor 0.19, por lo que no rechazamos la hipótesis de uniformidad.

Por último, analizamos los resiudos brutos de nuestro ajuste M_x para intervalos disjuntos de 89 días (entero más cercano a 90, el periodo de un invierno no bisiesto, que garantiza intervalos de aproximadamente igual tamaño en todo el periodo). Representamos estos residuos en la cuarta gráfica de la Figura 4.8, donde podemos ver un intervalo de incertidumbre bastante constante entre -0.2 y 0.2, totalmente compatible con la media nula que esperábamos. Por otro lado, únicamente cinco puntos se encuentran fuera del intervalo de confianza, que para un total de 62 puntos (número de intervalos de 89 días) corresponde a un 8 % de los residuos. A pesar de ser superior al 5 % (porcentaje esperado en nuestro nivel de confianza), no es excesivamente grande por lo que no debe preocuparnos.

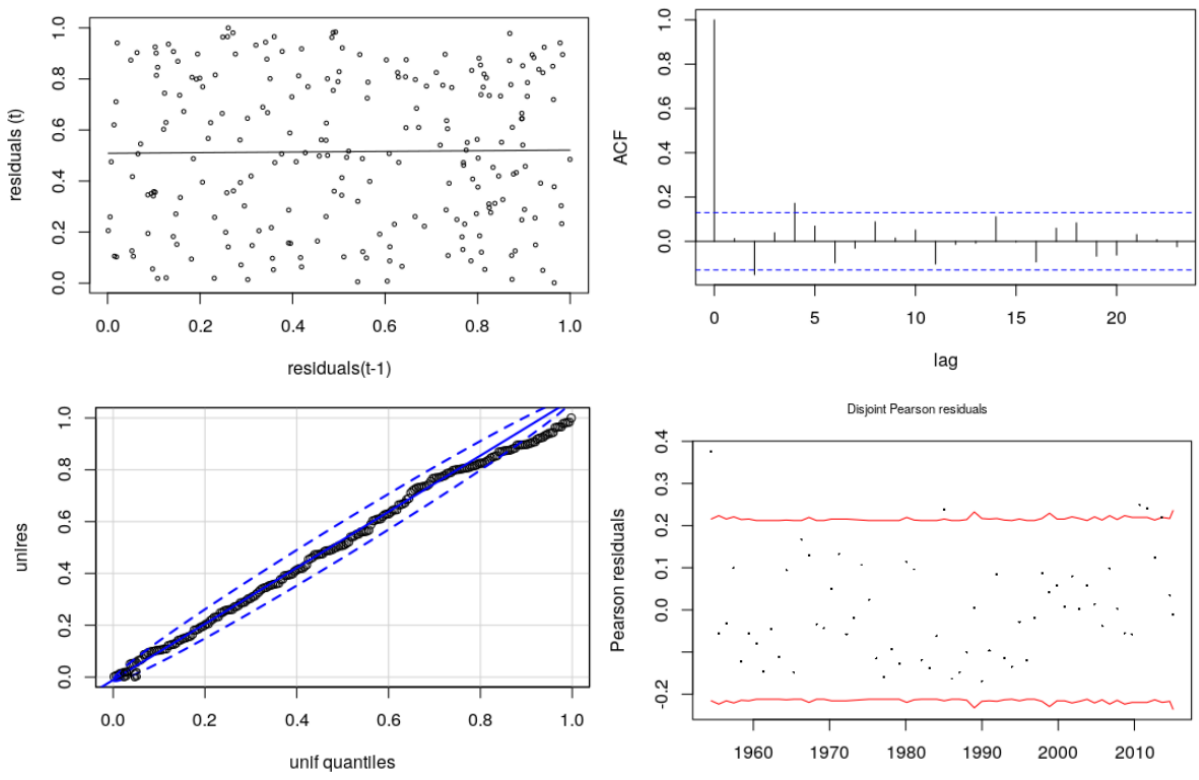


Figura 4.8: Validación gráfica del modelo M_x (4.1) para la serie T_x.

Modelo para el proceso $\mathbf{T_n}$ Obtenemos los residuos del proceso T_n y analizamos su correlación, obteniendo un coeficiente de Pearson de -0.04 y un p-valor 0.49 para el test de incorrelación, por lo que suponemos que no existe correlación de primer orden. Además, podemos ver el *gráfico de correlación de primer orden* en la primera gráfica de la Figura 4.9 que presentamos junto con un ajuste lineal. Podemos ver de nuevo una nube de puntos sin aparente dependencia, confirmando la hipótesis de incorrelación. En el ACF de nuestros residuos (segunda gráfica de la Figura), todos los valores se encuentran dentro del intervalo por lo que tampoco rechazamos la incorrelación en órdenes superiores.

Por otro lado, el *qqplot* parece en concordancia con la posible distribución uniforme de nuestros residuos.(tercera gráfica de 4.9). El ajuste a la recta unidad es de nuevo bastante bueno salvo para los valores más pequeños de los cuantiles, siendo sin embargo un buen ajuste en conjunto. El test Kolmogorov-Smirnov devuelve un p-valor 0.09 que nos permite mantener la hipótesis de uniformidad.

Terminamos de nuevo con los resiudos brutos de M_n, que calculamos para intervalos disjuntos de 89 días. Representamos estos residuos en la cuarta gráfica de la Figura 4.8, donde podemos ver de nuevo un intervalo de incertidumbre constante y centrado en el cero que es compatible con una buena estimación de la intensidad. Además, hay 3 puntos fuera del intervalo (5 %).

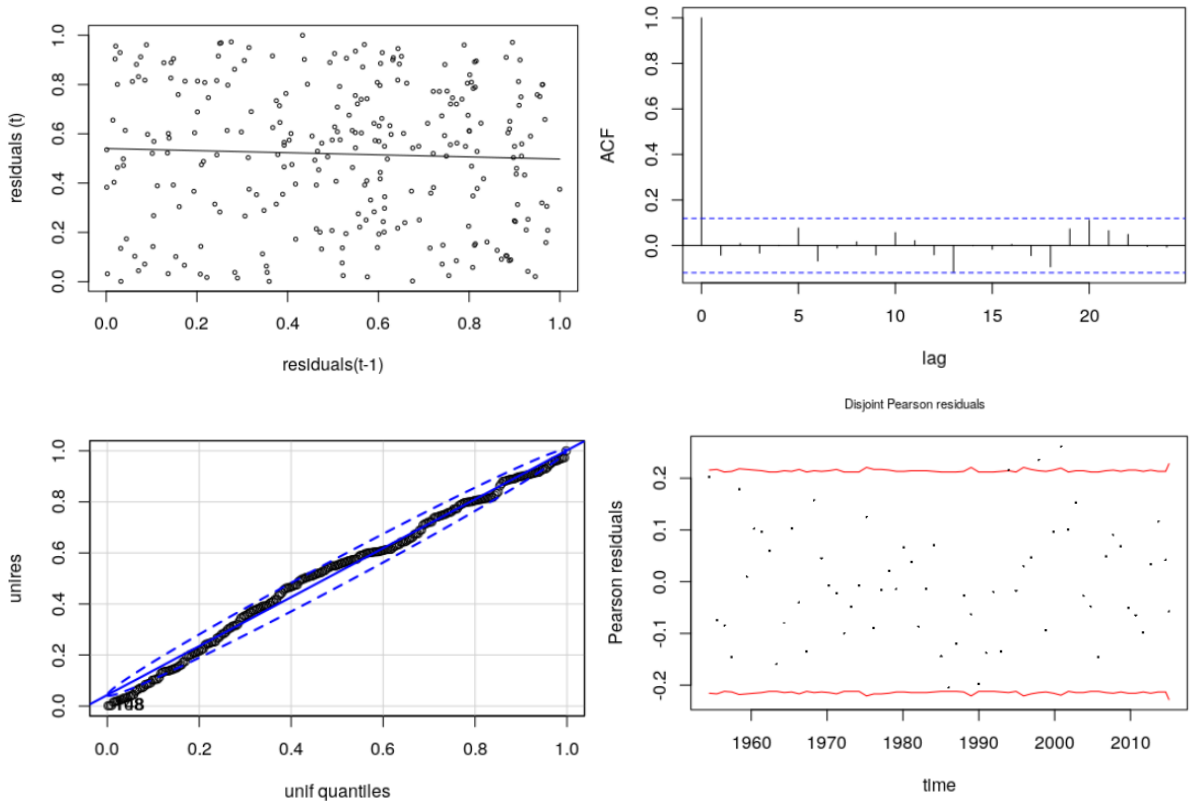


Figura 4.9: Validación gráfica del modelo M_n (4.2) para la serie T_n.

Modelo para el proceso $\mathbf{T_{nx}}$ Realizamos el mismo proceso para la última serie de ocurrencias, empezando por los residuos de T_{nx} y analizamos su correlación. Encontramos un coeficiente de Pearson de -0.09 y un p-valor 0.30 para el test de incorrelación, además de no observar tendencia en el *gráfico de correlación de primer orden* (primera gráfica de la Figura 4.10). Además, el ACF sólo muestra un valor fuera del intervalo, por lo que suponemos que no hay correlación de ningún tipo en los residuos.

El *qqplot* (tercera gráfica de 4.10), de nuevo, presenta una ligera desviación respecto a la recta unidad en los valores de cuantiles altos, pero el test Kolmogorov-Smirnov devuelve un

p-valor 0.22 por lo que suponemos uniformidad en los residuos.

Por último, en la cuarta gráfica de 4.10 podemos ver la evolución de los residuos brutos de M_{nx}, que parecen mantener una media nula. Observamos sólamente cuatro puntos fuera del intervalo.

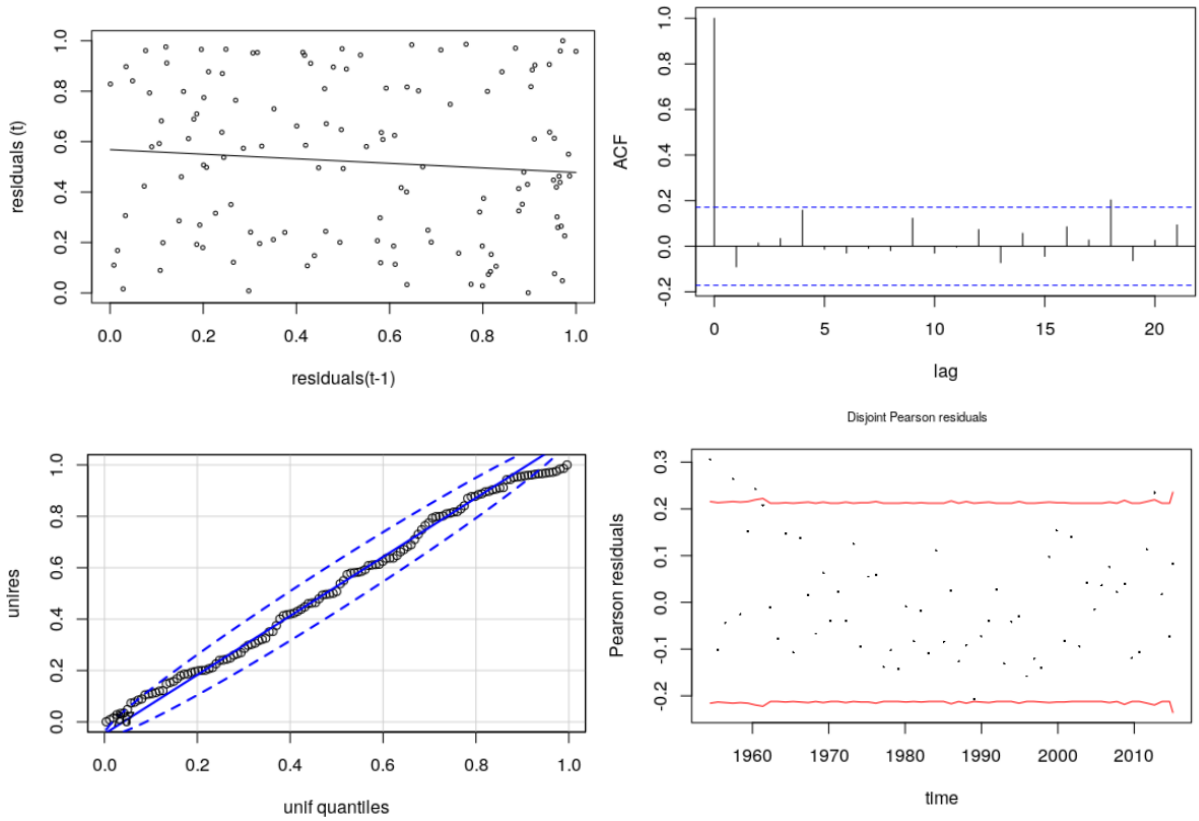


Figura 4.10: Validación gráfica del modelo M_{nx} (4.3) para la serie T_{nx}.

Validación del modelo conjunto Una vez validados los tres modelos, debemos validar la independencia de los tres procesos modelizados. Realizamos la evaluación mediante el test mencionado en 3.5 que nos devuelve un p-valor de 0.94, por lo que no rechazamos nuestra hipótesis nula.

Capítulo 5

Proyección de la ocurrencia de extremos bajo condiciones de cambio climático

El modelo construido en el capítulo anterior es un modelo de reducción de escala que nos permite caracterizar la ocurrencia de extremos mediante su dependencia con las tendencias a corto plazo de las series de temperatura máxima y mínima. Así pues, es una herramienta muy útil para la proyección de intensidades de ocurrencia dadas por las covariables de entrada del modelo. En concreto, buscaremos obtener una caracterización del proceso de ocurrencia de extremos bajo diferentes posibles escenarios de cambio climático en el siglo XXI.

Presentaremos la metodología utilizada en este proceso, que consistirá en primer lugar en la obtención de una serie regional de temperaturas máxima y mínima a través de los llamados modelos GCM. Con este conjunto de series obtenidas para diferentes escenarios de cambio climático, obtendremos una serie local de temperaturas para nuestra estación de Panticosa, mediante el proceso de reducción de escala espacial que describiremos. A partir de éste, podremos construir las covariables de entrada de nuestros modelos y así obtener la estimación de la intensidad para cada una de las series de proyección. Antes de presentar los resultados obtenidos mediante este proceso para nuestra serie, comentaremos ciertos datos disponibles de proyección en la media escala cercana a Panticosa. Terminaremos comentando las proyecciones obtenidos en nuestro modelo local.

5.1. Metodología utilizada

5.1.1. Modelos GCM

Los modelos GCM (General Circulation Models) son modelos numéricos que caracterizan los procesos físicos que se desarrollan en la atmósfera, el océano, la criosfera y la superficie terrestre. Son actualmente la mejor herramienta para la modelización de la respuesta climática al aumento de emisiones de gases de efecto invernadero[1]. Establecen una respuesta climática en una rejilla tridimiensional con resolución horizontal de entre 250 y 600 km de lado y de entre 10 y 20 subdivisiones verticales en la atmósfera y alrededor de 30 en el océano. Permiten por lo tanto, establecer una respuesta climática regional ante diferentes situaciones introducidas en el modelo.

Usaremos en nuestra proyección los GCM propuestos en el último informe del IPCC (Comité Intergubernamental para el Cambio Climático de la ONU), llamado AR5 y publicado en 2014 [16]. Este informe define un total de 33 regiones en las que agrupar las divisiones de la rejilla,

[1] Ver descripción detallada en http://www.ipcc-data.org/guidelines/pages/gcm_guide.html

estando Panticosa en la región MED (Mediterráneo). De todos los modelos aceptados por el IPCC, tomaremos tres:

- El MPI-ESM-LR, que abreviaremos MPI. Ha sido elaborado por el Max-Planck-Institut (MPI) for Meteorology, Alemania [14].
- El IPSL-CM5A-MR, que aberviaremos IPSL. Ha sido elaborado por el Institut Pierre-Simon Laplace (IPSL), Francia [12]. Como característica a destacar, este modelo no considera los días bisiestos.
- El MRI-CGCM3, que aberviaremos MRI. Ha sido elaborado por el Meteorological Research Institute, Japón: Yukimoto et al. (2012) [30].

Como novedad respecto a los anteriores informes de la IPCC, la variable de entrada de estos modelos es el nivel de emisión de gases de efecto invernadero, el llamado *forzamiento radiativo*, medido en W/m^2. Así pues, se definen cuatro posibles escenarios de emisión para el siglo XXI, las llamadas trayectorias RCP (Representative Concentration Pathways). En la Figura 5.1 podemos ver la evolución de la emisión global considerada en cada uno de estos cuatro escenarios. Como vemos, el número final en el nombre de cada escenario corresponde al umbral máximo de forzamiento radiativo que se esperaría no superar en el año 2100. Todos los escenarios presentan una monotonía creciente en las emisiones, salvo el escenario RCP3PD/RCP2.6, que establece un pico de emisión a mitad del siglo XXI y una posterior disminución suave hasta alcanzar menos de 2.6 W/m^2 en 2100. En los escenarios RCP4.5 el crecimiento monótono alcanza una estabilización a finales de siglo, mientras que los escenarios RCP6 y RCP8.5 presentan un crecimiento aproximadamente lineal. Hay que tener en cuenta que los escenarios se han construido en base a modelos de emisión y no tienen por qué tener una relación directa con una tesitura socio-económica concreta, ya que éstas pueden ser muy variadas y llevar a situaciones de emisión muy diferentes. Aun así, está claro que el modelo RCP3PD/RCP2.6 corresponde al escenario de emisiones más optimista y el escenario RCP8.5 al más pesimista.

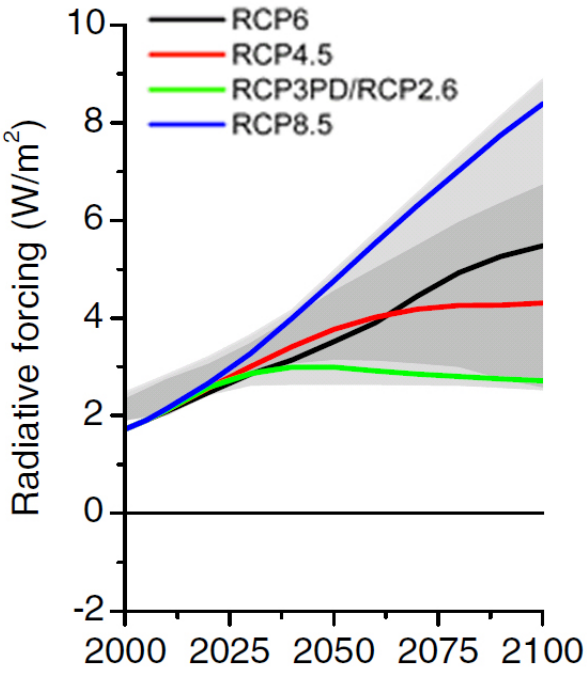


Figura 5.1: Evolución del forzamiento radiativo global según cada uno de los cuatro escenarios RCP [27].

Usaremos pues los datos de temperatura proporcionados en la región de Panticosa por cada uno de los tres modelos elegidos en tres de los cuatro escenarios: RCP4.5, RCP6, RCP8.5, salvo

para el modelo MPI que no dispone de respuesta al escenario RCP6. Estas ocho series de datos pueden descargarse de la web del Data Distribution Center (DDC)[2].

5.1.2. Reducción de escala espacial

Como hemos comentado, los modelos GCM nos devuelven proyecciones de temperatura a escala regional, por lo que es necesaria una reducción de escala (*downscalling*) que nos permita obtener proyecciones para nuestra serie de temperaturas locales. Esta reducción de escala espacial, o *regionalización*, puede hacerse mediante técnicas dinámicas que utilizan modelos regionales de clima, o mediante técnicas estadísticas, ya sea por el método de análogos o por el de regresión lineal (SDSM), el que usaremos en nuestra proyección. La técnica SDSM establece relaciones entre las variables atmosféricas a gran escala que ofrecen los modelos GCM (predictores) y las variables locales que pretendemos obtener (respuesta), que en nuestro caso son las temperaturas mínima y máxima diarias. El proceso de reducción de escala consiste entonces en ajustar mediante una regresión lineal múltiple la respuesta de temperatura observada en el periodo histórico disponible con las covariables aportadas por el modelo en ese periodo. Así pues, suponiendo que la relación entre predictores y respuestas es invariante frente al cambio del clima, obtenemos la serie de temperaturas diarias de nuestra localidad en el escenario modelizado por el GCM elegido.

Sin embargo, la serie obtenida no puede interpretarse a escala diaria, pues sólo es fiable en promedios de varios días [3]. Esto no es un problema para la proyección de nuestro modelo de extremos, pues las covariables introducidas en él son medias móviles de varios días. Así pues, una vez obtenida las series T_{max} y T_{min} para cada modelo y escenario, calcularemos las covariables correspondientes, T_{x15} y T_{n15} para el periodo de proyección. Dadas estas covariables, obtendremos, a través de nuestros modelos M_x, M_n y M_{nx}, las respuestas en intensidad de ocurrencia de los procesos T_x, T_n y T_{nx}.

5.2. Proyecciones regionales en el entorno de Panticosa

Antes de realizar la proyección sobre nuestras series de temperaturas en Panticosa, presentamos los resultados disponibles en cuanto a la proyección regional de la serie de temperaturas en el entorno de nuestra localidad. Todos los resultados que vamos a exponer a continuación han sido obtenidos por AEMET mediante el método de reducción de escala estadístico SDSM, y son de acceso público en su página web[3]. Para cada variable (T_{max} o T_{min}) y cada escenario, se utilizan un gran conjunto de modelos, cuyo resultado regionalizado (serie diaria de temperatura) se promedia otorgando igual peso a cada modelo.

Mostramos en la Tabla 5.2 los resultados obtenidos para la evolución de la temperatura en el conjunto de la España peninsular. Los datos presentan el incremento medio de la temperatura (ºC) respecto al nivel medio del periodo de referencia (1961-1990). Se obtienen datos para la serie de máximos y de mínimos, así como para la temperatura media anual y la temperatura media en invierno, periodo que nos interesa especialmente en nuestro estudio. En general, observamos un incremento de ambas series bajo todos los escenarios y en los dos periodos, siendo más fuerte en las máximas respecto a las mínimas y más leve en el invierno respecto al promedio anual. Así pues, para un mismo escenario (RCP8.5) obtenemos un incremento de 4ºC en las mínimas de invierno y de 7ºC en las máximas anuales.

Presentamos en la Figura 5.3 varios mapas que muestran los incrementos esperados en cada estación meteorológica para el periodo 2081-2100 en los meses de invierno, para las dos series de temperatura y bajo dos escenarios, el RCP4.5 y el RCP8.5. Cabe remarcar la gran diferencia espacial de los incrementos esperados en la península. Podemos ver cómo, en la zona de Panticosa

[2] https://www.iiasa.ac.at/web-apps/tnt/RcpDb/dsd?Action=htmlpage&page=welcome
[3] http://www.aemet.es/es/serviciosclimaticos/cambio_climat/result_graficos

	T_{min} (RCP4.5, RCP6, RCP8.5)	T_{max} (RCP4.5, RCP6, RCP8.5)
Anual	2.5, 3, 5.5	3, 4, 7
Invierno	2, 2.5, 4	2, 3, 5

Figura 5.2: Incremento medio de temperatura (°C) en la España peninsular respecto al nivel de referencia (61-90) para las series de temperaturas máximas y mínimas, para el periodo anual y de invierno y para los tres escenarios de emisiones disponibles. La proyección se construye promediando todos los modelos disponibles. Fuente: AEMET.

(frontera norte de Aragón) se espera un incremento más significativo que en la mayoría de la península, siendo una de las zonas con más incremento de la península, sobre todo en las máximas.

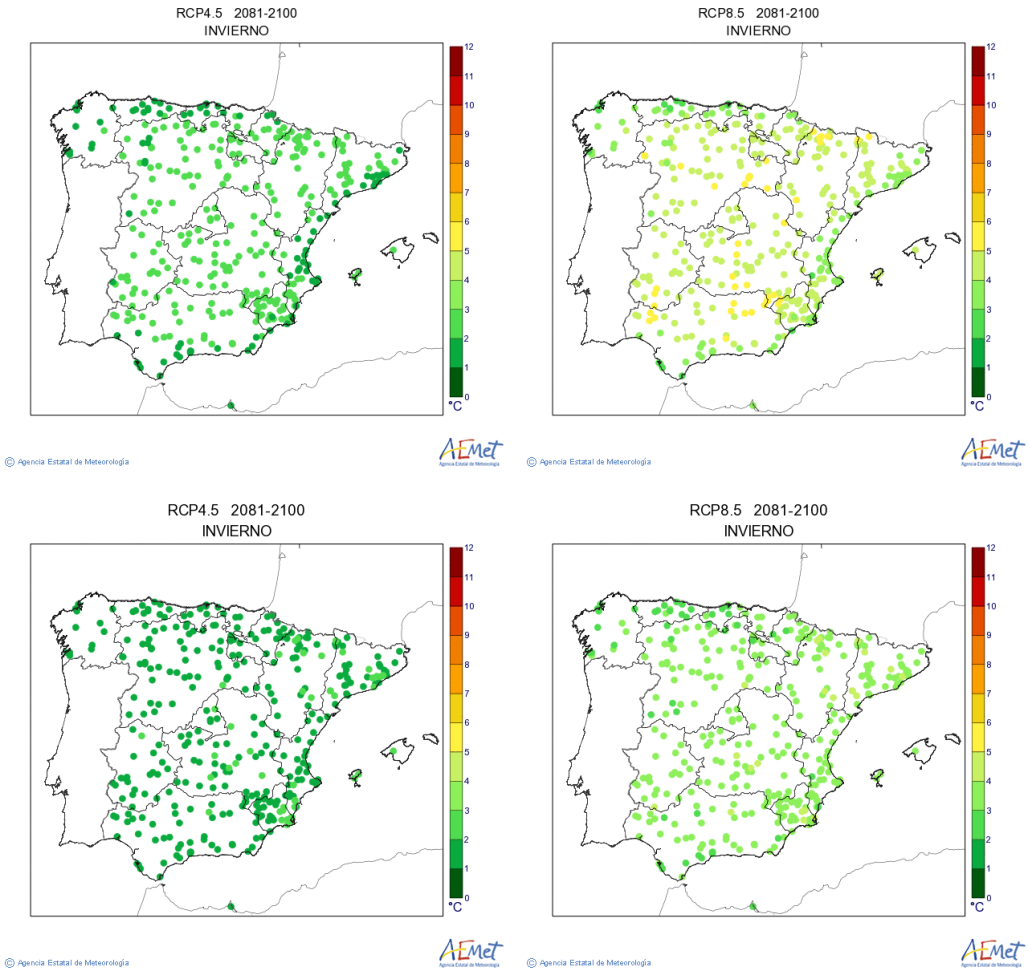


Figura 5.3: Incrementos esperados en cada estación meteorológica de la España peninsular para el periodo 2081-2100 en los meses de invierno, para las dos series de temperatura y bajo dos escenarios, el RCP4.5 y el RCP8.5. Fuente: AEMET.

Por último, podemos evaluar las proyecciones que existen para la provincia de Huesca, donde se encuentra Panticosa. En la Figura 5.4, podemos observar un mayor incremento en todas las series respecto a lo proyectado para el conjunto de la España peninsular. Además, están disponibles datos sogre el aumento de días cálidos en el año (número de días por encima del percentil 10 de las observaciones hechas en 1961-1990). En este sentido, se espera que aumente hasta un 30, 40 y 60 % de los días totales del año, bajo los escenarios RCP4.5, RCP6 y RCP8.5

respectivamente.

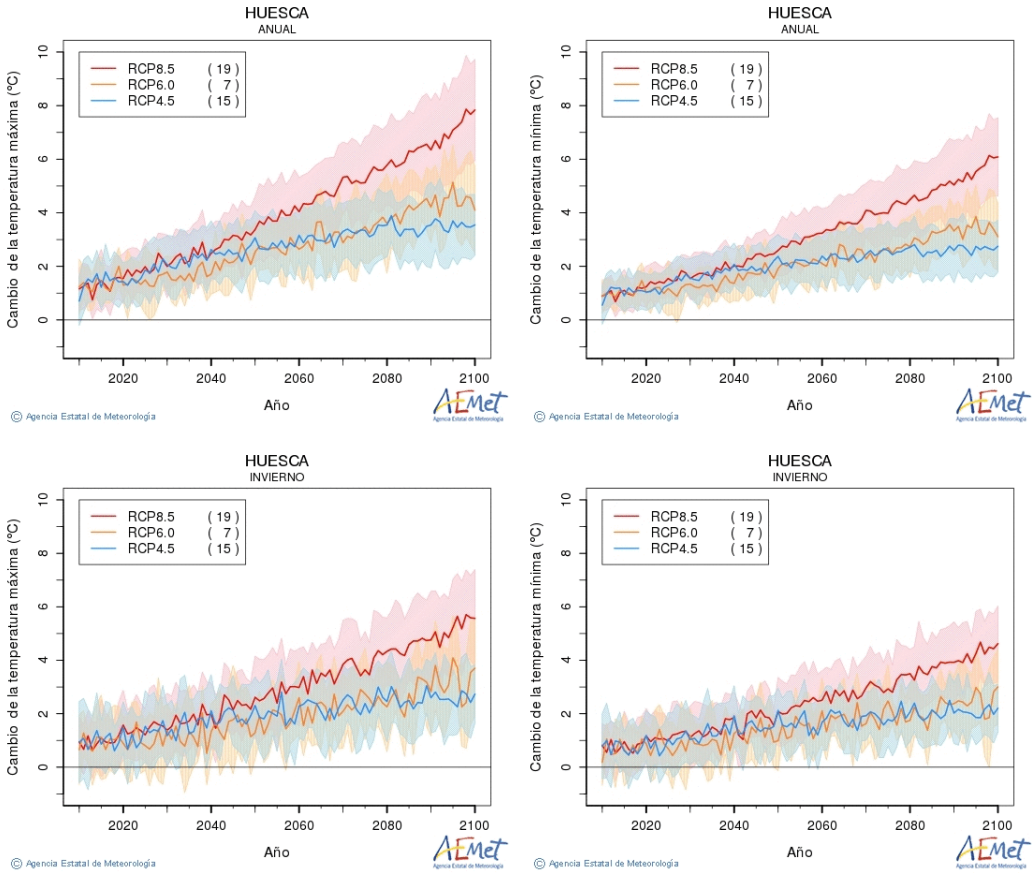


Figura 5.4: Evolución de la temperatura media (ºC) de la provincia de Huesca respecto al periodo 1961-1990, para las temperaturas máximas y mínimas y para las medias anuales y las medias invernales. Fuente: AEMET.

5.3. Proyección local en la estación de Panticosa

5.3.1. Obtención de los datos y tratamiento preliminar

El procedimiento de regionalización a escala local de los modelos GCM para Panticosa ha sido aplicado por AEMET, y también está disponible en su página web[4], junto con el de muchas más estaciones meteorológicas del estado. Para obtener una descripción detallada del proceso de reducción de escala empleado, consultar [2]. La respuesta del modelo regionalizado consiste en series de temperatura diarias bajo diferentes escenarios, con las cuales construiremos las covariables de entrada de nuestros modelos de ocurrencia de extremos.

Además de presentar diferentes escenarios de cambio climático en el periodo 2006-2100, los modelos GCM proporcionan la respuesta del modelo ante las situaciones ya observadas en el periodo 1961-2000. Son las trayectorias *históricas* (HIST), y sirven para validar el modelo y la reducción de escala local. En nuestro caso, al comparar las series HIST con las de temperatura observadas, observamos la existencia de un error de nivel (Figura 5.5, gráfica de la izquierda). Es decir, la serie histórica presenta una media diferente a la serie observada. Esto se debe a que, durante el proceso de unión de las dos series que componen las observaciones, se utilizó la media de la serie Casa de Piedra como referencia para igualar los niveles de las dos series. Sin embargo,

[4]http://www.aemet.es/es/serviciosclimaticos/cambio_climat/datos_diarios

durante la reducción de escala espacial, se ha utilizado la media de la serie Balneario. En todo caso, la solución es sencilla y consiste en ajustar los niveles medios de las series mediante una estandarización de la media o *reescalado*. Es decir, en cada uno de los modelos y variables (T_{max} y T_{min}) debemos calcular la media m_h de la serie histórica y m_o de la serie observada, en el periodo en el que se solapan (1961-2000), y restar $m_h - m_o$ a todas las series del modelo (la serie HIST y las de los tres escenarios). En la Figura 5.5 ilustramos este proceso para MRI-T_{min}, en el que mostramos los suavizados de la serie de temperaturas en invierno (DEF) observada, histórica y proyectada bajo los tres escenarios para el caso sin reescalar y tras el reescalado.

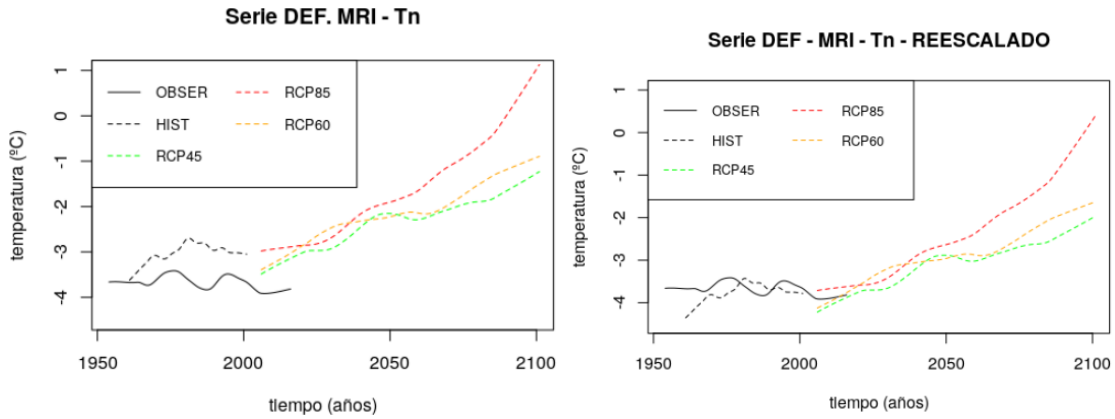


Figura 5.5: Suavizado polinomial con ventana del 30 % de la serie de temperaturas máximas en invierno (DEF) observada, histórica y proyectada mediante el modelo MRI bajo los tres escenarios, para el caso sin reescalar y tras el reescalado.

5.3.2. Proyección de nuestro modelo de extremos para el siglo XXI. Resultados.

Procedemos a analizar la respuesta de nuestros modelos a las covariables construidas mediante las trayectorias de los modelos GCM. Presentamos en la Figura 5.6 la intensidad de ocurrencia devuelta por los modelos M_x, M_n y M_{nx}, tanto para la serie observada como para las series proyectadas mediante los tres modelos y en los tres escenarios. Se observa **en todos los casos un aumento de la ocurrencia de extremos durante el siglo XXI**, lo que contrasta con la ausencia de tendencia observada en la segunda mitad del siglo XX para los meses de invierno, pero que está en consonancia con el aumento general de la temperatura previsto por estos modelos incluso en los meses DEF. Este incremento es bastante mayor en el proceso T_x que en el resto. Además, obtenemos las intensidades asociadas a los procesos de ocurrencia originales (extremos en T_{max} y extremos en T_{min}) que presentamos en la misma figura. Podemos ver con claridad como el **incremento en las máximas es más importante**, ya que la intensidad de su proceso es menor a la de los mínimos al inicio de siglo XXI, llegando a superarla en la segunda mitad de siglo. En efecto, a principios de siglo observamos un 6 % de observaciones extremas en T_{max} en promedio (5 días extremos por invierno), alcanzando a final de siglo entre 9 y 18 % según el escenario (entre 8 y 16 días extremos). En el caso de las mínimas, pasamos de 7 % (6 días) a un 9-16 % (8-14 días).

Para analizar el comportamiento estacional del modelo, representamos en la Figura 5.7 la intensidad media (truncada al 25 %) calculada en cada uno de los tres meses de nuestro periodo (diciembre, enero y febrero) y para cada una de las trayectorias disponibles: la observada y todas las proyectadas. Mostramos el resultado para dos periodos representativos del siglo XXI: 1921-1950 y 1971-2100. El patrón mensual de la serie de observaciones es el siguiente. El mes de diciembre presenta la mayor intensidad de extremos para las tres series. Por otro lado, en las series T_x y T_{nx}, enero y febrero presentan intensidades medias similares mientras que en

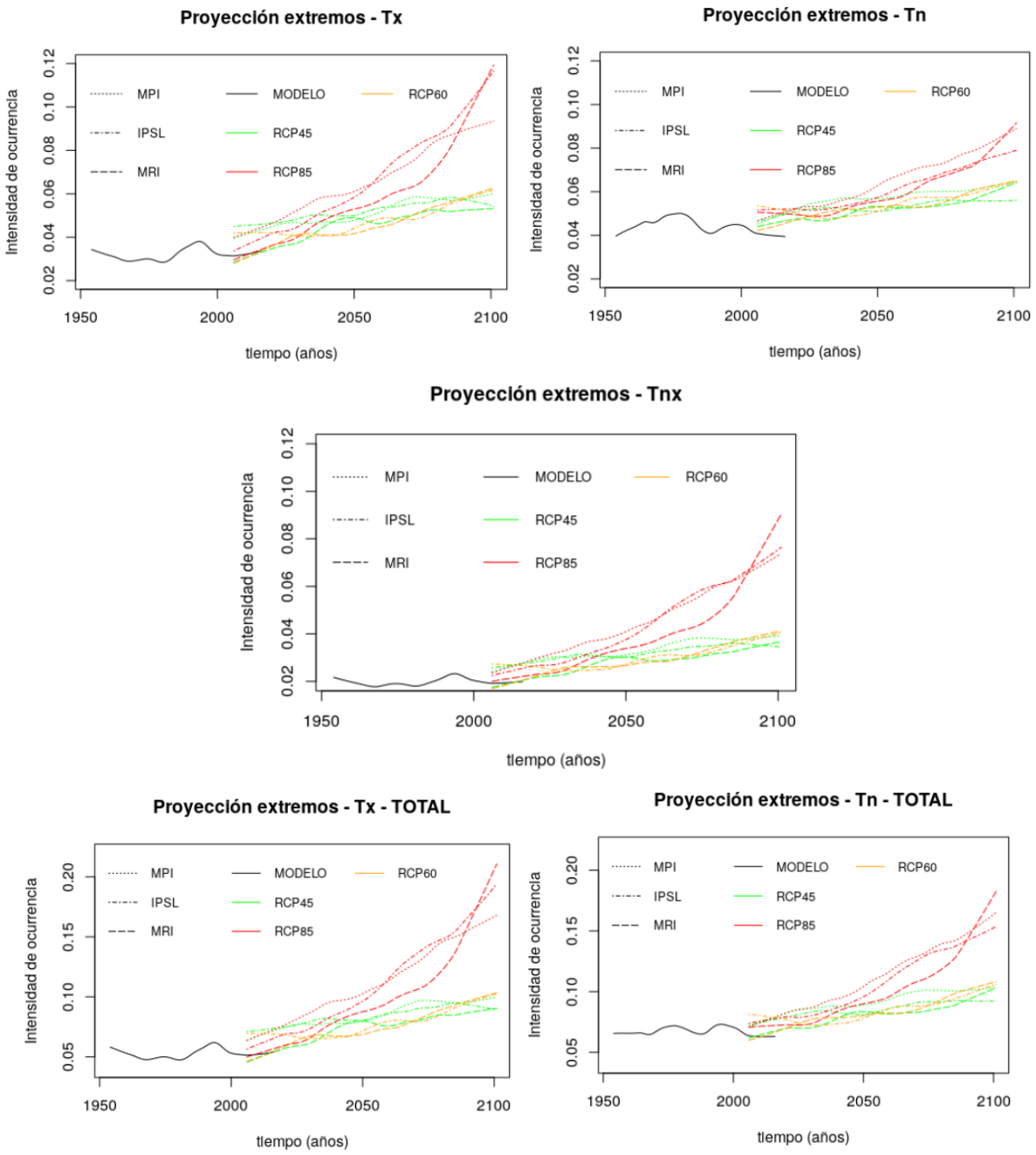


Figura 5.6: Intensidad de ocurrencia (suavizado con ventana del 30 %) de extremos en los diferentes procesos, T_x, T_n y T_{nx} devuelta por los modelos correspondientes para el periodo de observación y para el periodo de proyección, bajo los tres escenarios y los tres modelos GCM considerados. Presentaos también el resultado para los procesos originales de ocurrencia en T_{max} y en T_{min}.

T_n la intensidad en febrereo es claramente menor. En cuanto a las proyecciones, es curioso cómo los modelos predicen una acentuación de este patrón, aumentando la convexidad de la intensidad en T_x (sobre todo en las trayectorias MRI) y manteniéndose la tendencia decreciente diciembre-febrero en T_n. Podría decirse que la proyección prevee un **aumento de la diferencia inter-mensual**. La diferencia del modelo MRI sobre el resto es remarcable en este aspecto, ya que la intensidad media de la trayectoria RCP8.5 se sitúa en el mes de enero por debajo de la correspondiente a la RCP4.5 en el primer periodo. En el segundo periodo, como es lógico, los diferentes escenarios aumentan la distancia entre ellos.

Por último, cabe destacar el mayor incremento observado para los extremos del proceso T_x.

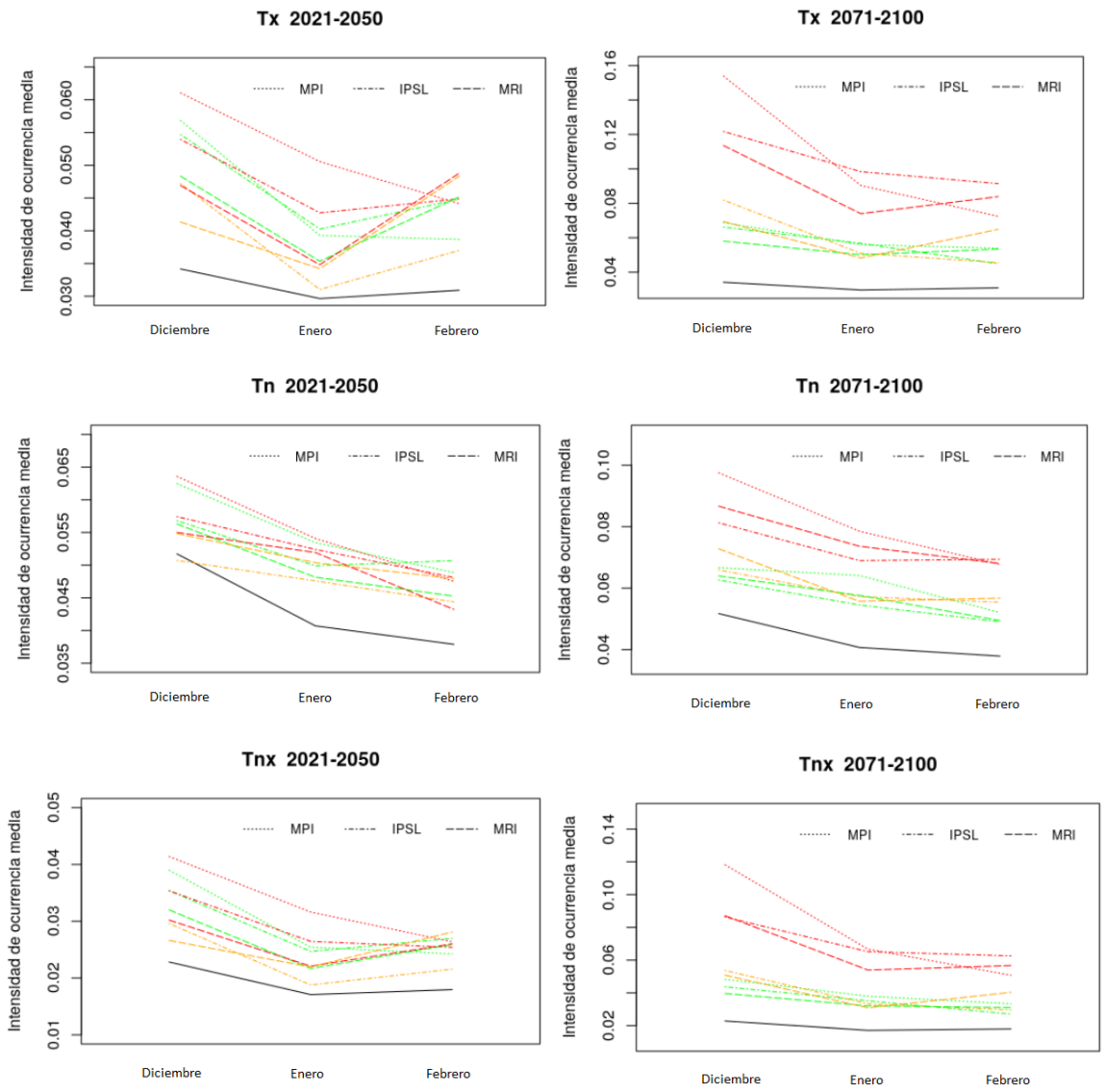


Figura 5.7: Intensidad media mensual (truncada al 25 %) de ocurrencia de extremos en los diferentes procesos, T_x, T_n y T_{nx} devuelta por los modelos correspondientes para el periodo de observación y para dos periodos de proyección: 2021-2050 y 2071-2100. Consideramos los tres modelos GCM indicados en la leyenda y los tres esceneraios indicados por el color de las líneas: RCP4.5 en verde, RCP6.0 en naranja y RCP8.5 en rojo.

Efectivamente, en ese proceso se proyecta una intensidad de hasta 0.08 en diciembre, predicha por el modelo MPI bajo el escenario RCP8.5, que corresponde a doblar la intensidad de la serie de observaciones actual. Un cambio más leve se da sin embargo en los extremos de temperaturas mínimas, cuyo mayor incremento es únicamente de 0.025 y corresponde nuevamente al MPI-RCP8.5 en diciembre. Si sumamos las intensidades de los procesos T_x y T_{nx} obtenemos la intensidad total de los extremos en temperaturas T_{max}, resultando aproximadamente 0.057 para la serie observada en diciembre y 0.135 para la trayectoria MPI-RCP8.5 (la más alta) en el mismo mes, lo que corresponde a un incremento del 140 % en el mes de diciembre. En el caso de los extremos de temperaturas mínimas, sumando las intensidades de T_n y T_{nx} en el mismo mes y para las mismas trayectorias obtenemos 0.075 y 0.13 respectivamente, lo que supone un incremento del 70 %.

Capítulo 6

Conclusiones

6.1. Proceso y resultados

En este trabajo, hemos mostrado la modelizaciónde de la ocurrencia de extremos en series de temperaturas máximas y mínimas mediante el modelo CPSP. Este modelo se ha presentado como una herramienta útil de caracterización de ambos procesos mediante su simplificación en tres NHPP independientes. El carácter Poisson de estos modelos marginales, justificado teóricamente como comportamiento asintótico, ha sido verficado en situaciones reales de series finitas.

Los resultados obtenidos para este proceso de modelización son un buen ajuste del modelo a las observaciones disponibles, pasando éste las herramientas de validación presentadas. Se utilizan como predictores las covariables de tendencia a corto plazo, recalcando el riesgo de extrapolación bajo proyecciones que presentan otras tendencias de mayor orden.

El modelo descrito ha demostrado ser una herramienta útil de reducción de escala temporal, para la proyección de eventos a escala diaria, ya que en esta escala los modelos dinámicos de clima no alcanzan a dar resultados. Se insiste en la necesidad de una reducción de escala espacial (regionalización) para la construcción de los predictores. Dado que los modelos regionalizados no reproducen con mucha precisión las observaciones disponibles, es necesario una estandarización de la medi en todos los modelos.

Finalmente, se han presentado los resultados de la proyección para la estación del Balneario de Panticosa. Se observa una tendencia creciente bajo todos los escenarios, mucho más acentuada en el caso de los máximos. Se observa también una acentuación de la diferencia inter-mensual, siendo enero el mínimo para el proceso de máximas y febrero para las mínimas. En concreto, se estiman para final de siglo entre 8 y 16 días por invierno de extremos en las máximas y entre 8 y 14 días en las mínimas.

6.2. Perspectivas para futuros trabajos

Como posibles vías de investigación de este trabajo, se propone la exploración de otras covariables atmosféricas que no presenten situaciones de extrapolación y permitan un ajuste más preciso de nuestro proceso. Por otro lado, el estudio aquí presentado podría realizarse durante otro periodo del año, como por ejemplo el verano, para evaluar posibles diferencias estacionales que parecen darse lugar, y su evolución bajo proyección.

6.3. Agradecimientos

Agradecemos a la *Agencia Estatal de Meteorología* (AEMET) la proporción de los datos de observaciones diarias empleadas en este trabajo. Así mismo, agradecemos a Jesús Asín la preparación y adecuación de dichos datos.

Bibliografía

[1] Abaurrea J., Asín J., Cebrián A. C. (2014) A bootstrap test of independence between three temporal nonhomogeneous poisson processes and its application to heat wave modeling. *Environ Ecol Stat* **pp** 1–18, http://dx.doi.org/10.1007/s10651-014-0288-1.

[2] Amblar P. et al. (2017) GUÍA DE ESCENARIOS REGIONALIZADOS DE CAMBIO CLIMÁTICO SOBRE ESPAÑA A PARTIR DE LOS RESULTADOS DEL IPCC-AR5. *Agencia Estatal de Meteorología*.

[3] Abaurrea J., Asín J., Cebrián A. C. (2015) *Stoch Environ Res Risk Assess* **29**: 309. https://doi.org/10.1007/s00477-014-0953-9

[4] Baddeley A.J., Turner R. (2005) **spatstat**: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, **12** (6), 1–42. URL http://www.jstatsoft.org/v12/i06/

[5] Beniston M., Stephenson D., Christensen O., Ferro C., Frei C., Goyette S., Halsnaes K., Holt T., Jylhä K., Koffi B., Palutikof J., Schöll R., Semmler T., Woth K. (2007) Future extreme events in European climate: an exploration of regional climate model projections. *Clim. Change* **81** (1), 71–95.

[6] Cebrián A. C., Abaurrea J., Asín J. (2015) NHPoisson: An R package for Fitting and Validating Nonhomogeneous Poisson Process. *Journal of Statistical Software* **64**.

[7] Cleveland, W. S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* **35**, 54.

[8] Clima en España: Pasado, presente y futuro. Informe de Evaluación del Cambio Climático Regional (2010). *Clivar-España*.

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

[10] David M. Gay (1990) Usage summary for selected optimization routines. *Computing Science Technical Report* **153**, AT&T Bell Laboratories, Murray Hill.

[11] Davison A. C., Smith R. L. (1990) Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society* **B** (52), 393-442.

[12] Dufresne J. L., Foujols M. A., Denvil S. et al. (2013) Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Climate Dynamics* **40**: 2123. https://doi.org/10.1007/s00382-012-1636-1

[13] Evin G., Favre A .(2013) Further developments of a transient Poisson-cluster model for rainfall. *Stoch Environ Res Risk Assess* **27**:831–47.

[14] Giorgetta M. A., et al. (2013) Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, *J. Adv. Model. Earth Syst.* in press.

[15] Herrera S., Fernández J., Gutiérrez, J. M. (2016) Update of the Spain02 gridded observational dataset for EURO CORDEX evaluation: assessing the effect of the interpolation methodology. *International Journal of Climatology*, **36** (2), 900-908.

[16] IPCC (2014) Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A:Global and Sectoral Aspects. Contribution of Working Group II to the FifthAssessment Report of the Intergovernmental Panel on Climate Change[Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir,M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy,S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]. Cambridge UniversityPress, Cambridge, United Kingdom and New York, NY, USA, 1132 pp.

[17] Leadbetter M. R., Lindgren G., Rootzen H. (1983) Extremes and Related Properties of Random Sequences and Series. *Springer Verlag, New York*.

[18] Lindskog P., McNeil A. (2003) Common poisson shock models: applications to insurance and risk modelling. *ASTIN Bull* **33**:209–238

[19] Meehl G., Washington W.M., Collins W., Arblaster J., Hu A., Buja L., Strand W., Teng H. (2005) How much more global warming and sea level rise? *Science* **307**, 1769–1772.

[20] Ogata Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *J Am Stat Assoc* **83**(401):9–27.

[21] Pickands J. (1975) Statistical inference using extreme order statistics. *Annals of Statistics* **3**, 119-131.

[22] Rico I., Izagirre E., Serrano E., López-Moreno J. I. (2017) Superficie glaciar actual en los Pirineos: Una actualización para 2016. *Pirineos* **172**.

[23] Ross S. M. (1996) Stochastic Processes. *Wiley* (2ª edición).

[24] Sivakumar B., Christakos G. (2011) Climate: patterns, changes and impacts. *Stoch Environ Res Risk Assess* **25**:443–44.

[25] Todorovic P., Zelenhasic E. (1970) A Stochastic Model for Flood Analysis. *Water Resources Research* **6** (6), 1641-1648.

[26] Tryhorn L., Risbey J. (2006). On the distribution of heatwaves over the Australian region. *Aust. Meteorol. Mag.* **55**, 169–182.

[27] Van Vuuren et al. (2011) The Representative Concentration Pathways: An Overview. *Climatic Change*, **109** (1-2), 5-31.

[28] Wang X., Yang T., Shao Q., Acharya K., Wang W., Yu Z. (2012) Statistical downscaling of extremes of precipitation and temperature and construction of their future scenarios in an elevated and cold zone. *Stoch Environ Res Risk Assess* **26** (3), 405–18.

[29] Weissman I. (1978) Estimation of parameters and quantiles based on the k largest observations. *Journal of the American Statistical Association* **73**, 812-815.

[30] Yukimoto S. A. et al. (2012). A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3. Model Description and Basic Performance. *Journal of the Meteorological Society of Japan.* **90A**. 23-64. 10.2151/jmsj.2012-A02.

Anexo A

Situación de extrapolación para las tendencias a largo plazo

Ilustramos a continuación la situación de extrapolación que presentan las tendencias a largo plazo de nuestras series de temperatura bajo proyección mediante modelos GCM.

En primer lugar, describimos lo que entendemos como "situación de extrapolación". Cuando se ajusta un modelo de regresión, se establecen las relaciones entre unas variables, los predictores, y otra variable respuesta. El modelo permite, siempre que se garanticen esas relaciones, obtener una respuesta para diferentes valores de los predictores. Así pues, suponemos que esas relaciones establecidas por el modelo son correctas en el rango de valores tomado por los predictores en la muestra de ajuste. Por lo tanto, cuando queramos establecer la respuesta de nuestro modelo ante un valor de un predictor fuera de ese rango, diremos que nos encontramos en *situación de extrapolación*. Al no poder garantizar que las relaciones entre predictores y respuesta son las mismas, cualquier estimación que hagamos de la respuesta para ese valor de extrapolación será errónea.

Es el caso de las tendencias a largo plazo de nuestras series de temperatura. Definimos una posible tendencia a largo plazo como el suavizado polinomial con ventana del 40 % de la serie de máximas, TT_x, y mínimas, TT_n. Al calcular estas covariables para la serie de observaciones y para las series proyectadas mediante los modelos GCM, encontramos que en la proyección, nos encontramos en situación de extrapolación (ver Figura A.1). Por lo tanto, un posible modelo que tuviera estas covariables como predictores no sería capaz de realizar una proyección fiable bajo estos escenarios. No es el caso sin embargo de las tendencias a corto plazo, tal y como podemos ver en los diagramas de caja.

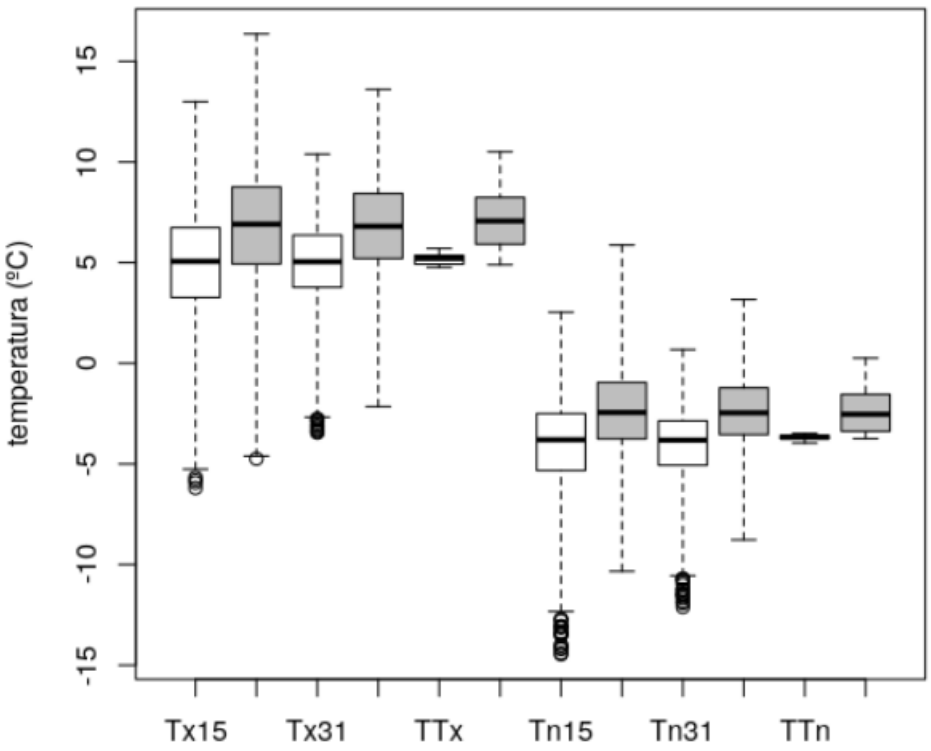


Figura A.1: Diagramas de caja de seis covariables de tendencia definidas, para la serie de observaciones (diagramas blancos) y para la serie proyectada (diagramas oscuros) bajo el modelo MRI escenario RCP8.5.