# Procesos de Poisson para la modelización de sucesos extremos



**Ana Isabel Barbería Sánchez**

Trabajo de fin de grado en Matemáticas

Universidad de Zaragoza

# Summary


**Extreme Value Theory**

This report is based on the classical theory of extreme values and the theory of excesses over threshold .The main objective of the classical theory of extreme values is to characterize the distribution of the maximum in a time period. That is, given a sequence of independent random variables, $X_1,..,X_n$ , with the same distribution function $F$ and observed at regular time intervals in a period, it seeks to characterize the distribution function of the maximum $M_n$. In practice, the behavior of $X_i$ is usually unknown, so the calculation of the exact behavior of $M_n$ is impossible. One possibility is to use standard statistical techniques to estimate F from observed data. Unfortunately, very small discrepancies in the estimate of F can lead to substantial discrepancies for $F^n$.

Therefore, extreme value theory has been developed, in that it is estimated $F^n$ using the information of extreme data. Specifically, the aim of the classical theory of extreme values is to characterize the asymptotic distribution of $M_n$.

Some examples where practical applications of maximum analysis are interesting are the measurements of sea level or daily temperatures.

One of the main theorems, Theorem of Fisher-Tippett, characterizes the distribution of the maximum limit and defines the extreme value distribution. Besides, this theorem allows us to ensure that under fairly general conditions, the asymptotic distribution of the maximum always corresponds to one of the VE0, VE1 or VE2 distributions defined in the theorem, independently of the original distribution. These three types can be parameterized in a single expression

$$G(z) = exp\left\{ -\left[1+\varepsilon\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\varepsilon}} \right\} \tag{1}$$

defined on $\{z : 1+\frac{\varepsilon(z-\mu)}{\sigma} > 0\}$ with $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \varepsilon < \infty$, where $\mu$ is a location parameter, $\sigma$ a scale parameter y $\varepsilon$ a shape parameter.

In the section of maximum domain of attraction are summarized the results that allow to characterize the limit distribution of $M_n$.

There is a problem if we apply these results, because the information from numerous observations, that may also have extreme character, can be lost if only the maximum of observations is considered in a certain time. To overcome this problem there are techniques such as EOT processes (Excess over Threshold).

The EOT methods (Excesses Over Thresholds) consider that an observation of a sequence is extreme if its value exceeds some high threshold. These methods are based on the occurrence of excesses over a threshold in a sequence i.i.d.r.v. have asymptotically, a Poisson behavior, and excesses have a generalized Pareto distribution.

This method presents a problem, due to the fact that this method has as hypothesis that the excesses are independent and in practice, usually, the excesses are dependent. To solve this problem, the POT (Peaks Over Threshold) method is developed. This method analyzes the sequence formed by the peaks of the sequence of excesses. An event is defined as a sequence of consecutive excesses and taking each peak as the value of maximum intensity within each event. On the sequence of peaks, which we assume now independent identically distributed, the EOT method is applied.

If the parent distribution F were known, the distribution of threshold exceedances would also be known. In practical applications, this is not the case, approximations for high values of the threshold are sought.

There is an important theorem which shows that if there is a limit distribution $M_n$, and n is large, the occurrence of excesses over a threshold $u_n$ behaves like a Poisson process of rate $\tau$.

In conclusion, the distribution of the occurrence of excess converges to a Poisson process and also excess distribution converges to a PG. That is, this helps us to analyze the sequence formed by the peaks of the sequence of excesses. In addition, there is an equivalence between the POT procedure and the classic maximum analysis.

There are some conditions, such as the condition D ($u_n$), which ensures the long-term dependence is weak or does not exist, and the condition D '($u_n$) which is a condition anti-cluster.

If both conditions are guaranteed, the sequence $M_n$ behaves as a sequence $\widetilde{M_n}$, so the problem of a maximum is reduced to an independent identically distributed squence. Consequently, under these conditions, convergence to a Poisson process of the process of occurrence of excesses on a threshold is maintained.

**Modeling of extreme events**

To estimate and validate a model of extreme events we will use the POT approach. An extreme event is defined as every observation which is above a certain threshold. To model the behavior of an extreme event, we select a Poisson process (PP), due to the occurrence of excesses on strict thresholds converges to a PP.

There are different methods of estimation, which allow to estimate the parameters that define a theoretical model with certain properties. In this report we will use the maximum likelihood method because it is the one with better properties.

To define the likelihood function of a NHPP based on the POT approach some adjustments are necessary, because only a point of occurrence is assigned to each event. The likelihood function of a PP will be

$$L(\beta; x_1, ..., x_n) = exp\left[-\int_A \lambda(x;\beta)dt\right]\prod_{i=1}^{n}\lambda(x_i;\beta). \tag{2}$$

To select the covariates that will be part of the model, various tools are available. To work with a high number of covariates, there are automatic selection methods, such as Forward, Backward and Stepwise method. The latter is one of the most commonly used methods, which is a combination of the other two. When the number of covariates is not very high, it is preferable to make a selection of covariates manually.

Once the model is selected, it is necessary to validate it, in other words, to check that the data satisfy model hypothesis. There are no specific tools to validate a NHPP, although transforming a no-homogeneous Poisson process (NHPP) into one homogeneous (HPP), we can use the tools for validation a HPP.

In conclusion, analysis validation begins transforming the points of occurrence of a NHPP in a PPH. Later, uniform residuals in a homogeneous process are calculated. Finally, uniform behavior and serial correlation of the uniform residuals is studied.

**Aplication**

In our application we analyze the occurrence of heat extreme events in the summer months. To do this we have a daily temperature $T_X$, for the months of May to September, between 1951 to 2005 in Zaragoza.

For that, we consider temperature signals, which are variables that give information about the evolution of temperature long-term and short-term, and seasonal terms.

Due to the characteristics of extreme events heat (nonmonotonic behavior and seasonal trend) we need to model through a Poisson process with no-homogeneous intensity. The representation of the intensity is a deterministic function of temperature covariates as the terms of short and long term and seasonal terms. Temperature harmonic interaction terms are also considered as possible covariates.

Once the estimation and selection of variables is done, we obtain the final model of Zaragoza. Finally, the model is validated through the uniform residuals and raw resicuals.

# Índice general

**Summary** III

**1. Teoría de Valores Extremos** **1**
1.1. Teoría de Valores Extremos clásica . . . 1
1.1.1. Distribución valor extremo . . . 2
1.1.2. Máximo dominio de atracción . . . 3
1.1.3. Limitaciones de La Teoría de Valores Extremos . . . 6
1.2. Teoría de Excesos sobre umbral . . . 6
1.2.1. Distribución Pareto generalizada . . . 7
1.2.2. Procesos de Poisson . . . 8
1.3. Teoría de valores extremos para series con dependencia . . . 9

**2. Modelización de sucesos extremos** **13**
2.1. Definición de los sucesos basada en POT . . . 13
2.1.1. Selección umbral . . . 13
2.2. Estimación del modelo . . . 13
2.2.1. Función de verosimilitud . . . 14
2.2.2. Estimación . . . 15
2.3. Selección de covariables . . . 16
2.3.1. Test máxima verosimilitud . . . 16
2.4. Validación . . . 16

**3. Aplicación** **19**
3.1. Estudio de los datos . . . 20
3.2. Selección del modelo . . . 20
3.3. Validación . . . 24
3.3.1. Residuos uniformes . . . 24
3.3.2. Residuos brutos . . . 25

**Bibliografía** **27**

# Capítulo 1

# Teoría de Valores Extremos

## 1.1. Teoría de Valores Extremos clásica

El objetivo principal de la teoría de valores extremos clásica es caracterizar la distribución del máximo en un periodo de tiempo. Es decir, dada una serie de variables aleatorias independientes, $X_1,..,X_n$, con la misma función de distribución $F$ y observadas en intervalos de tiempo regulares en un determinado periodo (por ejemplo un año), se busca caracterizar la función de distribución del máximo $M_n$. Si la distribución de probabilidad de $X_i$ es conocida, podemos calcular el comportamiento exacto de $M_n$. Sin embargo, en la practica el comportamiento de $X_i$ suele ser desconocido por lo que calcular el comportamiento exacto de $M_n$ no es posible.

En esta sección se van a ver las definiciones y resultados mas importantes de la teoría de extremos clásica. Nos centraremos en el comportamiento estadístico de los máximos, aunque sabiendo que $minX_i = -max(X_i)$, podremos aplicarlo al análisis de mínimos.

**Definición** (Máximo)**.** Sea $X_1,..,X_n$ una secuencia de variables aleatorias independientes idénticamente distribuidas con función de distribución F. Se define la variable máximo,

$$M_n = max\{X_1,...,X_n\} \tag{1.1}$$

Cuya función de distribución viene dada por

$$P\{M_n \leq x\} = P\{X_1 \leq x,...,X_n \leq x\} = P\{X_1 \leq x\}...P\{X_n \leq x\} = \{F(x)\}^n \tag{1.2}$$

Como se ha mencionado anteriormente, en los problemas reales la función de distribución $F$ suele ser desconocida. Se podrían utilizar técnicas estadísticas para estimar la función de distribucion a través de los valores observados y sustituirla en (1.2), pero el problema de esta aproximación es que aunque las discrepancias que se puedan dar a la hora de estimar $F$ sean pequeñas, cuando elevamos $F$ a la potencia n, estas discrepancias pueden cobrar mucha importancia.

Por ello se ha desarrollado la teoría de valores extremos, en la que se trata de estimar $F^n$, utilizando la información de los datos extremos. En concreto, el objetivo de la teoría de valores extremos clásica es caracterizar la distribución asintótica de $M_n$. Como se verá posteriormente, la distribución asintótica (cuando $n \to \infty$) del máximo no depende de la función $F$, por lo que en la práctica se podrá estimar la distribución de $M_n$ directamente, utilizando sólo la muestra de máximos. Sin embargo, hay que tener en cuenta, que por ser resultados asintóticos, no se pueden utilizar con $n$ pequeño.

Algunos ejemplos de aplicaciones prácticas donde el análisis de máximos es de interés son las medidas del nivel del mar, véase en [6], o las temperaturas diarias.

En esta sección se comienza presentando el Teorema de Fisher-Tippett, el teorema central de la teoría de valores extremos, que caracteriza la distribución límite del máximo y definiendo la distribución de valor extremo. A continuación se introducirá el concepto de máximo dominio de atracción y se terminará indicando las limitaciones que tiene este tipo de análisis.

### 1.1.1. Distribución valor extremo

En primer lugar es necesario recordar que dado que la sucesión de máximos $M_n$ converge trivialmente al extremo superior (o punto final) $x_F = sup\{x \in R | F(x) < 1\}$ y por lo tanto su distribución límite es degenerada, se debe trabajar con la variable máximo reescalada

$$\frac{M_n - d_n}{c_n}$$

con $c_n > 0$ y $d_n$ constantes, a los que llamamos coeficientes normalizadores.

Veamos el teorema que establece las distribuciones limite de $\frac{M_n - d_n}{c_n}$.

**Teorema 1.1.** *(Fisher-Tippett) Si existen { $c_n > 0$} y {$d_n$} de manera que*

$$\lim_{n \longrightarrow \infty} P\left\{\frac{M_n - d_n}{c_n}\right\} = G(z) \tag{1.3}$$

*donde G es una función de distribucion no degenerada, d es el parámetro de localización y c es el parámetro de escala. Entonces G pertenece a una de estas familias:*

*a)*

$$G_0(z) = exp\left\{-exp\left[-\left(\frac{z-d}{c}\right)\right]\right\} \tag{1.4}$$

*b)*

$$G_1(z) = \begin{cases} 0 & si\ z \leq 0 \\ exp\left\{-\left(\frac{z-d}{c}\right)^{-\alpha}\right\} & si\ z > 0 \end{cases} \tag{1.5}$$

*c)*

$$G_2(z) = \begin{cases} exp\left\{-\left(-\frac{z-d}{c}\right)^{-\alpha}\right\} & si\ z \leq d \\ 1 & si\ z > d \end{cases} \tag{1.6}$$

Los detalles de esta demostración son muy técnicos y se pueden encontrar en [9].

Este teorema es de gran importancia ya que nos permite asegurar que, bajo condiciones bastante generales, la distribución asintótica de los máximos siempre corresponde a una de esas distribuciones, independientemente de la distribución original. Es decir que las únicas posibles distribuciones límite de $M_n^*$ son las tres distribuciones anteriores. Todas las familia tienen parámetros de localización y escala d y c. Además las familias Fréchet y Weibull tienen un parámetro de forma $\alpha$.

Es importante señalar que aunque este teorema es muy importante, no garantiza que exista el límite no degenerado de $M_n$, ni especifica cual de las tres distribuciones es la que se da, si el límite existe. Existen otros resultados teóricos, que no se exponen en esta memoria, que permiten garantizar si la distribución límite existe, véase [9] y otros resultados que permiten caracterizar las distribuciones que tienen cómo distribución límite cada uno de los tres tipos de distribución VE.

Para ello, en primer lugar vamos a formalizar la definición de esta distribución límite y a introducir otras definiciones necesarias.

**Definición** (Distribución Valor Extremo)**.** Una función de distribución se dice de Valor Extremo, VE, si es una de las tres distribuciones definidas con anterioridad en el Teorema de Fisher-Tippett. Dentro de la distribución VE se distinguen tres tipos:

1. Tipo 0 (VE0) o distribución Gumbel, $G_0$.
2. Tipo 1 (VE1) o distribución de Fréchet, $G_1$.
3. Tipo 2 (VE2) o distribución de Weibull, $G_2$.

Estos tres tipos se pueden reparametrizar en una única expresión que los contiene como casos particulares,

$$G(z;\varepsilon,\sigma,\mu) = exp\left\{-\left[1+\varepsilon\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\varepsilon}}\right\} \tag{1.7}$$

definida en el conjunto $\{z : 1+\dfrac{\varepsilon(x-\mu)}{\sigma} > 0\}$ con $-\infty < \mu < \infty, \sigma > 0$ y $-\infty < \varepsilon < \infty$, donde $\mu$ es un parámetro de localización, $\sigma$ un parámetro de escala y $\varepsilon$ un parámetro de forma.

La relación entre las dos parametrizaciones cuando $\mu = 0$ y $\sigma = 1$ es:

$$G(z;\varepsilon,\sigma,\mu) = \begin{cases} G_0(z;1,0) & \text{si } \varepsilon = 0 \\ G_1(z;\frac{1}{\varepsilon},\frac{1}{\varepsilon},-\frac{1}{\varepsilon}) & \text{si } \varepsilon > 0 \\ G_2(z;-\frac{1}{\varepsilon},-\frac{1}{\varepsilon},-\frac{1}{\varepsilon}) & \text{si } \varepsilon < 0 \end{cases}$$

**Definición.** La distribucion correspondiente a una variable aleatoria X se dice máximo estable, si existen constantes normalizadoras $d_n$ y $c_n$ tal que

$$\frac{M_n - d_n}{c_n} \xrightarrow[d]{} X. \tag{1.8}$$

El siguiente resultado proporciona una caracterización de la distribución VE.

**Teorema 1.2.** *Una distribución es máximo estable si y sólo si es una VE*

Se puede encontrar una demostración de este teorema en [6].

### 1.1.2. Máximo dominio de atracción

En esta sección se resumen los resultados que permiten caracterizar cuál es la distribución límite de una distribución. Para ello se requieren algunos resultados previos.

**Definición.** Se dice que una función de distribución $F$ pertenece al máximo dominio de atracción (MDA(G)) de una distribución $G$ si existen constantes $c_n > 0$ y $d_n \in \mathbb{R}$ de manera que la distribución límite de la sucesión de máximos normalizados $M_n$ con función de distribución de base $F$, es $G$.

**Teorema 1.3** (Caracterización de MDA(G))**.** *La función de distribución F pertenece al MDA de una distribución G Valor extremo si y sólo si*

$$\lim_{x\longrightarrow\infty} n\overline{F}(c_n x + d_n) = -lnG(x) \in \mathbb{R}.$$

*con* $\overline{F}(x) = 1 - F(x)$

**Definición.** Una distribución $\overline{F}$ es una variación regular con índice - $\alpha$ y se denota con $\overline{F} \in R_{-\alpha}$ si

$$\lim_{x\longrightarrow\infty} \frac{\overline{F}(xt)}{\overline{F}(x)} = t^{-\alpha}, \text{ con } t > 0.$$

Algunas propiedades de las distribuciones de variación regular se encuentran en [3].

El siguiente concepto define una relación de equivalencia en el conjunto de todas las funciones de distribución.

**Definición** (Cola-equivalente)**.** Dos funciones de distribución $F$ y $H$ se dicen cola-equivalentes si $x_F = x_H$ y

$$\lim_{x\longrightarrow x_F} \frac{\overline{F}(x)}{\overline{H}(x)} = c$$

para alguna constante $0 < c < \infty$.

Una propiedad de las distribuciones cola-equivalentes es que $F \in$ MDA(G) si y solo si $H \in$ MDA(G), donde las constantes normalizadoras son las mismas para cada función.

**Máximo dominio de atracción en las funciones de distribución Gumbel**

**Definición** (Función de Von Mises)**.** Una función de distribución F es una función de Von Mises, si existe z < $x_F$ tal que F se puede expresar como

$$\overline{F}(x) = c\ \exp\left[-\int_z^x \frac{1}{a(t)}df\right]; \qquad z < x < x_F \tag{1.9}$$

donde c es una constante positiva y $a(t)$ es una función positiva absolutamente continua de manera que $\lim_{t\longrightarrow x_F} \frac{da(t)}{dt} = 0$.

**Teorema 1.4** (Condición de Von Mises del MDA($\Lambda$))**.** *Si F es una función de Von Mises, entonces F pertenece al MDA($\Lambda$). Además sólo pertenecen al MDA($\Lambda$) las funciones de Von Mises y las que son cola-equivalentes a ellas.*

**El MDA($\Lambda$) se compone de las distribuciones que satisfacen las condiciones Von Mises y sus distribuciones cola-equivalentes.**

**Ejemplos de distribuciones en el MDA de la distribución Gumbel**

Algunas de las distribuciones que pertenecen al MDA($\Lambda$) son:

1. Weibull: $\overline{F} = exp\{-cx^{\tau}\}$, $x \geq 0$, c, $\tau > 0$
2. Erlang: $\overline{F}(x) = e^{-\lambda x}\sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!}$ $x \geq 0$, $\lambda > 0, n \in \mathbb{N}$
3. Normal: $\overline{F}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
4. Exponencial: $F(x) = 1 - e^{-x}$. A continuación se muestra la pertenencia al MDA($\Lambda$) para el caso de exponencial con parámetro $\lambda = 1$.

Si $\{X_1, ..., X_n\}$ es una secuencia de variables independientes que siguen una distribución exponencial con parámetro $\lambda = 1$, $F(x) = 1 - e^{-x}$. En este caso, tenemos las constantes normalizadoras $c_n = 1$ y $d_n = ln(n)$,

$$P\left\{\left(\frac{M_n - d_n}{c_n}\right) \leq x\right\} = F^n(x + ln(n)) =$$

$$= \left[1 - e^{-(x+ln(n))}\right]^n = \left[1 - n^{-1}e^{-x}\right]^n \underset{n\longrightarrow\infty}{\longrightarrow} exp(-e^{-x})$$

para $x \in \mathbb{R}$. Así la distribucion límite de $M_n$ cuando $n \longrightarrow \infty$ es una distribución Gumbel, con $\varepsilon = 0$ en la familia VE.

**Máximo dominio de atracción de la distribución Fréchet**

**Teorema 1.5** (Máximo dominio de atracción de $\Phi_\alpha$)**.** *Una función de distribución F pertenece al máximo dominio de atracción de la función de distribución Fréchet $\Phi_\alpha$ con $\alpha > 0$, si y solo si $\overline{F}(x) = x^{-\alpha}L(X)$ donde L(x) es una función de variación regular.*

$$F \in MDA(\Phi_\alpha) \Leftrightarrow \overline{F} \in R_{-\alpha} \tag{1.10}$$

*Además si $F \in MDA(\Phi_\alpha)$*

$$\frac{M_n}{c_n} \underset{d}{\longrightarrow} \Phi_\alpha$$

con constantes normalizadoras $d_n = 0$ y $c_n = F^{-1}(1 - n^{-1})$.

El siguiente resultado permite caracterizar el MDA($\Phi_\alpha$) de una forma más sencilla, a través de la función de densidad.

**Teorema 1.6** (Condición de Von Mises para el MDA($\Phi_\alpha$))**.** *Sea F una función de distribucion absolutamente continua, de manera que*

$$\lim_{x\longrightarrow\infty}\frac{xf(x)}{\overline{F}}=\alpha>0 \tag{1.11}$$

*entonces* $F\in MDA(\Phi_\alpha)$.

**El MDA($\Phi_\alpha$) se compone de las funciones de distribución que satisfacen las condiciones Von Mises para MDA($\Phi_\alpha$) y sus distribuciones cola equivalentes.**

**Ejemplos de distribuciones en el MDA de la distribución Fréchet**
La distribución Pareto, distribución Fréchet, distribución de Cauchy y distribución de Burr.

**Ejemplo**
Si $\{X_1,...,X_n\}$ es una secuencia de variables independientes que siguen una distribución estándar Fréchet $F(x)=e^{-\frac{1}{x}}$ para $x>0$. Dadas las constantes normalizadoras $c_n=n$ y $d_n=0$,

$$P\left\{\left(\frac{M_n-d_n}{c_n}\right)\leq x\right\}=F^n(nx)=\left[e^{-\frac{1}{nx}}\right]^n\underset{n\longrightarrow\infty}{\longrightarrow}exp(\frac{-1}{x})$$

para $x>0$. Así la distribución límite en este caso es también una distribución Fréchet, con $\varepsilon=1$ en la familia VE.

**Máximo dominio de atracción de las funciones de distribución Weibull** Recordamos que $\Phi_\alpha(x)=\Psi_\alpha(-x^{-1})$. Por lo que esperaremos que sus respectivas MDA también estén relacionadas.

**Teorema 1.7** (Máximo dominio de atracción de $\Psi_\alpha$)**.** *La función de distribución F pertenece al máximo dominio de atracción de la distribución Weibull* $\Psi_\alpha$ *con* $\alpha>0$*, si y solo si* $\overline{F}(x_F-x^{-1})=x^{-\alpha}L(X)$ *donde L(x) es una función de variación regular y su punto final es finito,* $x_F<\infty$.

*Si* $F\in MDA(\Psi_\alpha)$ *entonces*

$$\frac{M_n-x_F}{c_n}\underset{d}{\longrightarrow}\Psi_\alpha$$

con constantes normalizadoras $d_n=x_F$ y $c_n=x_F-F^{-1}(1-n^{-1})$.

El siguiente resultado permite caracterizar el MDA($\Psi_\alpha$) de una forma más sencilla, de nuevo a través de la función de densidad.

**Teorema 1.8** (Condición de Von Mises para MDA($\Psi_\alpha$))**.** *Sea F una función de distribución absolutamente continua con función de densidad positiva, de manera que si en un intervalo finito (z,*$x_F$*).*

$$\lim_{x\longrightarrow\infty}\frac{(x_F-x)f(x)}{\overline{F}}=\alpha>0 \tag{1.12}$$

*entonces* $F\in MDA(\Psi_\alpha)$.

**El MDA($\Psi_\alpha$) se compone de las funciones de distribución que satisfacen las condiciones Von Mises para MDA($\Psi_\alpha$) y sus distribuciones cola equivalentes.**

**Ejemplos de distribuciones en el MDA de la distribución Weibull**

Algunas distribuciones que pertenecen al MDA($\Psi_\alpha$) son:

1. La distribución Beta $f(x) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$.

2. La distribución Uniforme(0,1) y veamos porque esta pertenece a la familia VE2:

Si $\{X_1,...,X_n\}$ es una secuencia de variables independientes que siguen una distribución Uniforme U(0,1), $F(x) = x$ para $0 \leq x \leq 1$. Para algún $x < 0$ fijado y supuesto n> - x y dadas las constantes normalizadoras $c_n = \dfrac{1}{n}$ y $d_n = 1$.

$$P\left\{\left(\frac{M_n - x_F}{c_n}\right) \leq x\right\} = F^n(n^{-1}x+1) = \left(1+\frac{x}{n}\right)^n \underset{n\longrightarrow\infty}{\longrightarrow} e^x$$

Es decir, que la distribucion límite de $M_n$ es una distribución Weibull, con $\varepsilon = -1$ en la familia VE.

### 1.1.3. Limitaciones de La Teoría de Valores Extremos

Un problema de la aplicación de estos resultados en modelización estadística es que los procedimientos de inferencia basados en la distribucion límite del máximo son ineficientes. Al considerar únicamente el máximo de las observaciones en un determinado intervalo de tiempo, se pierde información de numerosas observaciones que también pueden tener carácter extremo. Para subsanar este problema existen técnicas que incluyen en la muestra un mayor número de observaciones extremas. Una de ellas es considerar en cada periodo, en lugar del máximo, los k estadísticos ordenados superiores. Sin embargo esta técnica, si k es pequeño, tampoco garantiza la inclusión de todos los valores extremos y si k es demasiado alto puede incluir valores que no sean extremos. Por ello la aproximación más utilizada en modelización estadística, es la que se describe en la sección siguiente denominada 'Exceso sobre Umbral' o EOT (Excess over Threshold).

Otra restricción de los resultados de la teoría de valores extremos expuestos es que las observaciones deben ser independientes. Sin embargo, en muchos casos, especialmente en las observaciones medio-ambientales, las series de observaciones presentan una estructura de dependencia a corto plazo. Este es un problema importante que también sera tratado en la próxima sección.

## 1.2. Teoría de Excesos sobre umbral

Los métodos EOT (Excesses Over Thresholds) consideran que una observación de una serie es extrema si su valor está por encima de un determinado umbral. Como veremos en esta sección, estos métodos se basan en que, asintóticamente la ocurrencia de los excesos sobre un umbral, en una serie v.a.i.i.d, tienen un comportamiento de Poisson y los excesos siguen una distribución Pareto Generalizada. Aunque este método presenta un problema, puesto que el método EOT tiene como hipótesis que los excesos sean independientes y en la práctica no suelen serlo. Para solucionar este problema, se desarrolla el método POT (Peaks Over Threshold). Este método consiste en analizar la serie formada por los picos de la serie de excesos, definiendo un suceso como una serie de excesos consecutivos y tomando cada pico como el valor de intensidad máxima dentro de cada suceso. Sobre la serie de picos, que suponemos ahora independientes idénticamente distribuidos, se aplica el método EOT.

Sea $X_1, X_2, ...$ una serie de v.a.i.i.d. con distribución F. Consideramos como sucesos extremos los $X_i$ que excedan un umbral alto u. Denotando un termino arbitrario en la serie $X_i$ como X y u como el umbral fijado, el comportamiento estocástico de los sucesos extremos viene dado por la siguiente probabilidad condicional

$$P\{X > u+y | X > u\} = \frac{1-F(u+y)}{1-F(u)} \qquad y \geq 0 \tag{1.13}$$

Si conocemos la distribución de F, también podremos saber cual es la distribución de los excesos sobre un umbral alto. En la mayoría de casos prácticos la distribución F no es conocida, por lo que se buscan aproximaciones que sean ampliamente aplicables para valores altos del umbral. Una búsqueda análoga es la que se llevó a cabo en el capitulo anterior, donde vimos la distribución VE como una aproximación de la distribución de $M_n$, independientemente de la distribución original.

Antes de justificar los métodos introduciremos dos conceptos importantes, la distribución Pareto Generalizada y el Proceso de Poisson.

### 1.2.1. Distribución Pareto generalizada

Para justificar la definición de la distribución Pareto generalizada enunciaremos el siguiente teorema.

**Teorema 1.9.** *Sean $X_1, X_2, \ldots$ una serie de variables aleatorias independientes idénticamente distribuidas, con distribución F y $M_n = max\{X_1, \ldots, X_n\}$. Suponiendo que se satisface que, para un n grande, $P\{M_n \leq z\} \approx G(z)$ donde G(z) es la distribución VE (vista en la sección 1.1.1):*

$$G(z;\varepsilon,\sigma,\mu) = exp\left\{-\left[1+\varepsilon\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\varepsilon}}\right\} \tag{1.14}$$

*con $\mu \in \mathbb{R}, \sigma > 0$, se tiene que para un u suficientemente grande, la función de distribución de $(X-u)$, condicionada a que $X > u$, es aproximadamente*

$$H(x;\varepsilon,\sigma,\mu) = 1-\left(1+\frac{\varepsilon(x-\tilde{\mu})}{\tilde{\sigma}}\right)^{-\frac{1}{\varepsilon}} \qquad si\ \varepsilon \neq 0 \tag{1.15}$$

*para $x \geq \tilde{\mu}$, cuando $\varepsilon \geq 0$ y $x \leq \tilde{\mu} - \frac{\tilde{\sigma}}{\varepsilon}$ cuando $\varepsilon < 0$ donde $\tilde{\mu} \in \mathbb{R}$.*

Una demostración de este teorema se encuentra en [15].

Llamaremos distribución **Pareto generalizada**, PG, a la distribución H(x), cuya versión estandarizada, siendo $\tilde{\mu} = 0$ y $\tilde{\sigma} = 1$ es:

$$H(x) = 1-(1+\varepsilon x)^{-\frac{1}{\varepsilon}} \tag{1.16}$$

con

$$\begin{array}{ll} x \geq 0 & \text{si } \varepsilon \geq 0 \\ 0 \leq x \leq -\dfrac{1}{\varepsilon} & \text{si } \varepsilon < 0 \end{array}$$

La familia PG contiene tres familias de distribuciones: Exponencial o PG0, Pareto o PG1 y Beta (con parámetros ($\alpha$, 1))o PG2. Siendo $\alpha > 0$, las tres familias de distribuciones en su versión estandarizada son:

$$PG0 : H_0(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1-exp(-x) & \text{si } x \geq 0 \end{cases}$$

$$PG1 : H_1(x) = \begin{cases} 0 & \text{si } x < 1 \\ 1-x^{-\alpha} & \text{si } x \geq 1 \end{cases}$$

$$PG2 : H_2(x) = \begin{cases} 0 & \text{si } x < -1 \\ 1-(-x)^{-\alpha} & \text{si } -1 \leq x \leq 0 \\ 1 & \text{si } x \geq 0 \end{cases}$$

La relación entre las parametrizaciones es:

$$H(x;\varepsilon,\sigma,\mu) = \begin{cases} H_0(x;1,0) & \text{si } \varepsilon = 1 \\ H_1(x;1/\varepsilon,1/\varepsilon,-1/\varepsilon) & \text{si } \varepsilon > 0 \\ H_2(x;-1/\varepsilon,-1/\varepsilon,-1/\varepsilon) & \text{si } \varepsilon < 0 \end{cases}$$

Como hemos visto en el teorema 1.9, existe una relación entre las aproximaciones que llevan a definir las distribuciones PG (H(x)) y VE (G(x)), la cual viene dada por la siguiente relación funcional:

$$H(x) = 1 + ln[G(x)]$$

con $ln[G(x)] > -1$.

En particular cada distribución PG0 , PG1 y PG2 se corresponden respectivamente con las distribuciones de valor extremo VE0 (Gumbel), VE1 (Fréchet) y VE2 (Weibull).

La siguiente propiedad nos permite caracterizar de forma unívoca la distribución PG.

**Definición** (POT-Estable)**.** La distribución F de una variable X se dice POT-estable si existen $c_n$ y $d_n$ constantes normalizadoras de manera que la distribución del exceso sobre u, $X_u = X - u|(X > u)$, normalizado, tiene la misma distribucion que X, es decir

$$F_{X_u}\left(\frac{x-d_n}{c_n}\right) = F(x). \tag{1.17}$$

**Teorema 1.10.** *La distribución Pareto Generalizada es POT-estable y es la única distribución que presenta esta propiedad; es decir, si* $X \sim PG(\varepsilon, \sigma, \mu)$ *para* $u > \mu$ *tenemos que*

$$X_u \sim PG(\varepsilon, \sigma + \varepsilon(u-\mu), 0). \tag{1.18}$$

### 1.2.2. Procesos de Poisson

**Definición** (Proceso de Poisson)**.** Un proceso de Poisson de tasa $\lambda(t)$, es un proceso puntual donde los puntos $T_1, T_2, ...$ son los instantes de ocurrencia, los cuales ocurren de forma aleatoria. Denotamos $H_t$ a la trayectoria del proceso hasta el instante t, y N(u,v) al número de ocurrencias en el intervalo (u,v].

Un proceso de Poisson cumple que para todo t,

$$P(N(t,t+\delta) = 1|H_t) = \lambda(t)\delta + o(\delta) \tag{1.19}$$

$$P(N(t,t+\delta) > 1|H_t) = o(\delta) \tag{1.20}$$

Por lo que:

$$P(N(t,t+\delta) = 0|H_t) = 1 - \lambda(t)\delta + o(\delta) \tag{1.21}$$

Si la función $\lambda(t)$ es constante, el proceso de Poisson es homogéneo (PPH), en el caso contrario, el proceso es no homogéneo (PPNH).

Esta definición no es fácil de comprobar en la práctica, por lo que se presentan dos propiedades equivalentes más fáciles de verificar.

a) Caracterización del número de ocurrencias en un intervalo. Sea A un conjunto arbitrario sobre el eje temporal y N(A) el número de puntos del proceso en él. En un PPH las v.a $N(A_1), N(A_2), ...$, con $A_1, A_2, ...$ intervalos temporales disjuntos, son independientes y tienen distribución de Poisson de media $\lambda|A_1|, \lambda|A_2|, ...$, con $|A|$ longitud del conjunto.

b) Caracterización de los tiempos de recurrencia. Los tiempo de recurrencia de un proceso puntual $T_{r1} = T_1, T_{r2} = T_2 - T_1$, se definen como los intervalos transcurridos entre dos ocurrencias consecutivas. Los tiempos de recurrencia de un PPH son variables aleatorias independientes idénticamente distribuidas con distribución exponencial de parámetro $\lambda$.

**Ocurrencia de los excesos como Proceso de Poisson**

La ocurrencia de los excesos sobre el umbral de una serie i.i.d, se comporta asintóticamente, cuando el umbral es muy estricto, como un Proceso de Poisson con intensidad

$$\Lambda(t_1,t_2) = (t_2 - t_1)\left(1+\gamma\frac{u-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}. \tag{1.22}$$

Veamos a continuación uno de los resultados fundamentales de los métodos EOT, el cual permite justificar el carácter Poisson del proceso de ocurrencia de los excesos. Para justificar este teorema, habrá que tener en cuenta que, dada una muestra de tamaño n, el número de excesos $r_n$ sobre un umbral fijo $u_n$ es aleatorio y tiene una distribución Binomial de parámetros n y $p_n = \overline{F}(u_n)$ y utilizar que una distribución Binomial se aproxima a una distribución Poisson, $P(\tau)$ cuando $np_n \underset{n\longrightarrow\infty}{\longrightarrow} \tau$.

**Teorema 1.11.** *Sea $(X_n)$ una serie v.a.i.i.d con distribución F y $r_n$ el número de excesos de la serie sobre el umbral tomado $u_n$. Si la sucesión de umbrales $(u_n)$ verifica,*

$$\lim_{n\longrightarrow\infty} n\overline{F}(u_n) = \tau \tag{1.23}$$

*entonces para k= 0, 1, 2...*

$$\lim_{n\longrightarrow\infty} P(r_n \leq k) = e^{-\tau}\sum_{s=0}^{k}\frac{\tau^s}{s!} \tag{1.24}$$

*Inversamente, si esta propiedad se cumple para un valor de k, entonces se verifica la condición 1.23, y por lo tanto la propiedad se cumple para todo k.*

Como hemos dicho con anterioridad, este teorema justifica el carácter de Poisson del proceso de ocurrencia de los excesos. En efecto, escalando el rango temporal con un factor n, $r_n$ corresponde al número de excesos en el intervalo (0,1] y su distribución es, en las condiciones del teorema anterior, aproximadamente Poisson. Análogamente se prueba que el número de excesos en cualquier intervalo acotado tiene una distribución límite $P(\tau)$. Además, por hipótesis, el número de excesos en intervalos disjuntos son v.a independientes.

Los teoremas vistos en esta sección nos aseguran que la distribución de la tasa de excesos converge a una Poisson y además la distribución de los excesos converge a una distribución PG. Es decir, nos ayudan a analizar la serie formada por los picos de la serie de excesos, lo cual se pretende analizar con el método POT. Además existe un teorema, el cual se puede encontrar en [15], que nos asegura que existe una equivalencia entre el procedimiento POT y el análisis clásico de máximos.

**Teorema 1.12.** *La convergencia de $M_n$ a una distribución VE, es equivalente a que la distribución de la tasa de excesos converja a una distribucion Poisson, y la distribucion de los excesos a una PG.*

## 1.3. Teoría de valores extremos para series con dependencia

Hasta ahora hemos estado tratando con resultados que se basaban en la hipótesis de independencia de la serie $(X_n)$. Sin embargo, en los problemas reales nos encontramos con series que no cumplen con la hipótesis de independencia entre las observaciones. Esto es debido a que algunas características frecuentes de series reales son la dependencia a corto y largo plazo, así como la estacionalidad. A continuación se definen unas condiciones de dependencia bajo las cuales los resultados de la teoría de extremos siguen siendo aplicables.

**Definición** (Proceso estacionario)**.** Un Proceso aleatorio $X_1, X_2, ...$ se dice proceso estacionario si dados un conjunto de enteros $i_1, ..., i_k$ y un número entero m, tenemos que $(X_{i_1}, ..., X_{i_k}) = (X_{i_1+m}, ..., X_{i_k+m})$.

**Definición** (Proceso asociado)**.** Un proceso asociado a un proceso estacionario $(X_n)$ es un proceso $(\widetilde{X}_n)$ de v.a.i.i.d con la misma distribución que las variables del proceso $(X_n)$.

En esta sección vamos a dar unas condiciones sobre la serie estacionaria $(X_n)$ con distribución F, que aseguran que el máximo muestral de esta serie $(M_n)$ y el de la serie correspondiente asociada $(\widetilde{M}_n)$, tienen el mismo comportamiento asintótico.

Una de las condiciones necesarias para aplicar la teoría de valores extremos en series dependientes es la condición D$(u_n)$.

**Definición** (Condición D$(u_n)$)**.** Una serie $X_n$ verifica la condición $D(u_n)$ si para cualquier p , q y n enteros tal que $1 \leq i_1 < ... < i_p < j_1 < ... < j_q < n$ de manera que $j_1 - i_p \geq l_n$ tenemos

$$\left| P\left( \max_{i \in A1 \cup A2} X_i \leq u_n \right) - P\left( \max_{i \in A1} X_i \leq u_n \right) P\left( \max_{i \in A2} X_i \leq u_n \right) \right| \leq \alpha_{n,l_n} \tag{1.25}$$

donde $A_1 = \{i_1, ..., i_p\}$ , $A_2 = \{j_1, ..., j_p\}$ y $\alpha_{n,l_n} \underset{n \longrightarrow \infty}{\longrightarrow} 0$, para alguna sucesión $l_n = o(n)$

La condición D$(u_n)$ es una propiedad que garantiza que la dependencia a largo plazo es débil, o no existe.

El teorema que se enuncia a continuación garantiza que si se satisface la condición $D(u_n)$, la distribución asintótica de $M_n$ es una distribución VE.

**Teorema 1.13** (Condición Distribución VE)**.** *Sea $M_n$ la sucesión de máximos de $(X_n)$ y constantes $c_n > 0$ y $d_n \in \mathbb{R}$. Si se verifica la condición D$(u_n)$ con $u_n = c_n x + d_n$ y*

$$\frac{(M_n - d_n)}{c_n} \underset{d}{\longrightarrow} G \tag{1.26}$$

*entonces la distribución limite G, es una distribución VE*

Este resultado no nos proporciona información sobre el tipo de distribución VE, ni sobre su relación con el límite de la serie asociada. Para obtener esta información, se imponen condiciones más fuertes.

**Definición** (Condición D'$(u_n)$)**.** Diremos que una serie $X_n$ verifica la condición $D'(u_n)$ si

$$\lim_{k \longrightarrow \infty} \lim_{n \longrightarrow \infty} \sup n \sum_{j=2}^{[n/k]} P(X_1 > u_n, X_j > u_n) = 0 \tag{1.27}$$

D'$(u_n)$ es una condición anti-cluster, es decir, verifica que no se producen agrupaciones de las ocurrencias. Así, esta condición limita la probabilidad de que se produzca más de un exceso en $X_1, ..., X_{[\frac{n}{k}]}$.

**Teorema 1.14.** *Sea $(X_n)$ una serie estacionara con función de distribución $F \in MDA(VE)$ y $(\widetilde{X}_n)$ el proceso asociado y con $c_n$ y $d_n$ las constantes normalizadoras de $(X_n)$. Si además $(X_n)$ cumple las condiciones D$(u_n)$ y D'$(u_n)$ donde $u_n = c_n x + d_n$ con $x \in \mathbb{R}$ entonces tenemos*

$$\frac{(M_n - d_n)}{c_n} \underset{d}{\longrightarrow} G \tag{1.28}$$

$$\frac{(\widetilde{M}_n - d_n)}{c_n} \underset{d}{\longrightarrow} G \tag{1.29}$$

*donde G es una distribución VE.*

Este teorema nos garantiza que la sucesión $M_n$ se comporta como la sucesión $\widetilde{M_n}$, por lo que el problema de la distribución del máximo de una serie estacionaria se reduce al de una serie i.i.d.

En la realidad, es plausible que se cumpla la condición D$(u_n)$, puesto que esta no es muy restrictiva, mientras que la condición D'$(u_n)$ es incompatible con que exista dependencia a corto plazo, por lo que muchas series de datos reales no la verifican.

**Convergencia del máximo de series con dependencia a corto plazo**

Para comprender el teorema que enunciaremos con posterioridad debemos definir índice extremal

**Definición** (Indice extremal)**.** Sea $(X_n)$ un proceso estacionario. Si para todo $\tau > 0$ existe $u_n$ tal que

$$\lim_{n\longrightarrow\infty} \widetilde{F}(u_n) = \tau \tag{1.30}$$

entonces llamamos índice extremal de la serie a un valor $\theta$ (con $\theta \geq 0$ ) tal que

$$\lim_{n\longrightarrow\infty} P(M_n \leq u_n) = e^{-\theta\tau} \tag{1.31}$$

Otra caracterización posible, suponiendo de nuevo que la sucesión P($M_n \leq u_n$) converge, es

$$\theta = \lim_{n\longrightarrow\infty} P(max(X_2,...,X_{p_n}) \leq u_n)|X_1 \geq u_n),$$

donde $p_n$ = o(n).

El índice extremal mide el grado de agrupación, de hecho, se puede interpretar como el recíproco del tamaño medio asintótico de los clusters en niveles extremos. Esta interpretación se basa en los resultados de [13]. Los valores de $\theta$ están comprendidos entre 0 y 1. En un proceso independiente, el indice extremal es 1, aunque esto no implica que si el índice es 1, el proceso sea independiente.

En el siguiente resultado veremos el efecto de la dependencia en el comportamiento asintótico de los máximos.

**Teorema 1.15.** *Sea $(X_n)$ un proceso estacionario que satisface la condición D($u_n$) con $u_n = c_n x + d_n$ e índice extremal $\theta$. Entonces tenemos que*

$$\frac{(M_n - d_n)}{c_n} \xrightarrow[d]{} G \tag{1.32}$$

*si y solo si*

$$\frac{(\widetilde{M}_n - d_n)}{c_n} \xrightarrow[d]{} G_* \tag{1.33}$$

*con G y $G_*$ distribuciones no degeneradas de manera que G(x)= $G_*^{\theta}(x)$*

Este teorema asegura que la distribución asintótica del máximo de un proceso con dependencia a corto plazo es una distribución VE. Además, el tipo de la distribución, equivalentemente el parámetro de forma, es el mismo que si el proceso estuviera formado por variables independientes. Su único efecto es elevar a la potencia $\theta$ la distribución límite del máximo la serie independiente $(\tilde{X}_n)$.

En conclusión, los resultados asintóticos para procesos independientes son aplicables a procesos con dependencia a corto plazo si se cumple la condición D($u_n$).

# Capítulo 2

# Modelización de sucesos extremos

En este capítulo veremos la estimación de un modelo de sucesos extremos y su posterior validación basándose en la aproximación POT. Uno de los resultados teóricos proporcionados por este método nos permite asegurar que considerando umbrales suficientemente estrictos, el proceso de ocurrencia de los picos sobre un umbral estricto se comportará como un PP.

## 2.1. Definición de los sucesos basada en POT

Dada una serie de observaciones, definimos un suceso extremo como toda observación la cual está por encima de un determinado umbral. En los problemas reales, las ocurrencias tienen una duración, por lo que éstas no corresponden exactamente a un punto. Sin embargo, si la duración de los sucesos extremos es pequeña, en relación al tiempo que transcurre entre las ocurrencias, la utilización de un PP como modelo de ocurrencia es una buena aproximación (método visto en la sección 1.2.2). Por lo que para estudiar estos sucesos determinaremos el instante al cual se le asocia la ocurrencia de cada valor extremo, como hemos dicho antes en la sección 1.2, este instante será el punto de máxima intensidad dentro de cada suceso.

Para modelar el comportamiento de un suceso extremo, elegimos un proceso de Poisson homogéneo (HPP), ya que estos han sido ampliamente utilizados para modelar la ocurrencia de eventos extremos, debido a que se establece que la ocurrencia de excesos sobre umbrales muy estrictos converge a un PP. En el caso que la tasa de ocurrencia no fuese constante, tomaríamos un PPNH, con su posterior transformación a un HPP.

### 2.1.1. Selección umbral

En la elección de un umbral adecuado nos encontramos con dos objetivos contradictorios, el primero es que necesitamos definir un umbral suficientemente estricto para que sea posible la utilización de la teoría de valores extremos, y por otro lado, si tomamos un umbral muy estricto se tendría un número muy pequeño de observaciones extremas, por lo que se tendría muy pocos datos con los que trabajar. Además, hay que tener en cuenta que el umbral tomado debe adaptarse a la definición del tipo de suceso analizado.

## 2.2. Estimación del modelo

Existen distintos métodos de estimación, como el método de los momentos o el método de máxima verosimilitud. Estos métodos permiten estimar los parámetros que definen un modelo teórico con ciertas propiedades. En esta memoria vamos a utilizar el método de máxima verosimilitudes debido a que es el que presenta mejores propiedades.

Este método encuentra el valor de los parámetros $\beta$ que maximizan la función de verosimilitud. Para entender este método, primero recordaremos lo que es una función de verosimilitud, la cual viene definida por la función de densidad o de probabilidad conjunta de las variables independientes.

### 2.2.1. Función de verosimilitud

**Definición** (Función de verosimilitud)**.** Sea $X_1,...,X_n$ una muestra aleatoria de una población X con función de probabilidad conjunta $P_\theta$ (o con función de densidad $f_\theta$). Para cada muestra particular $(x_1,...,x_n)$, la función de verosimilitud se define como la función de probabilidad (o de densidad) conjunta de $(X_1,...,X_n)$ evaluada en $(x_1,...,x_n)$.

$$L(\theta) = L(x_1,...,x_n;\theta) = P_\theta(X_1 = x_1,...,X_n = x_n)$$

si $X$ es discreta,

$$L(\theta) = L(x_1,...,x_n;\theta) = f_\theta(x_1,...,x_n)$$

si $X$ es continua.

Para definir la función de probabilidad de un NHPP basado en el enfoque POT son necesarios algunos ajustes, debido a que sólo se asigna un punto de ocurrencia a cada evento, y éste es el punto de máxima intensidad. Así la probabilidad resultante es la que habríamos obtenido si solo el punto de máxima intensidad hubiese sido observado.Veamos a continuación como obtener la función de verosimilitud de un PP.

Sea $x_1,...,x_n$ una muestra aleatoria observada en una región $\mathbf{A} \subset \mathbb{R}$ la cual sigue un proceso de Poisson en $\mathbf{A}$ con función de intensidad $\lambda(x;\beta)$, para algún valor de $\beta$ y dado $I_i = [x_i, x_i + \delta_i]$, para i=1,...,n un conjunto de intervalos pequeños, tomaremos $L = \mathbf{A} / \bigcup_{i=1}^{n} I_i$. Teniendo en cuenta las propiedades de un proceso de Poisson

$$P\{N(I_i) = 1\} = exp\{-\Lambda(I_i;\beta)\}\Lambda(I_i;\beta) \tag{2.1}$$

donde

$$\Lambda(I_i;\beta) = \int_{x_i}^{x_i+\delta_i} \lambda(u)du \approx \lambda(x_i)\delta_i. \tag{2.2}$$

Sustituyendo 2.2 en 2.1 tenemos que

$$P\{N(I_i) = 1\} = exp\{-\lambda(x_i)\delta_i\}\lambda(x_i)\delta_i \approx \lambda(x_i)\delta_i, \tag{2.3}$$

donde hemos usado que $exp\{-\lambda(x_i)\delta_i\} \approx 1$ cuando $\delta_i$ es pequeño. Además,

$$P\{N(L) = 0\} = exp\{-\Lambda(L)\} \approx exp\{-\Lambda(\mathbf{A})\}, \tag{2.4}$$

si $\delta_i$ es pequeño. Así la verosimilitud será

$$\begin{aligned} L(\beta;x_1,...,x_n) &= P\{N(L) = 0, N(I_1) = 1, N(I_2) = 1,...,N(I_n) = 1\} \\ &= P\{N(L) = 0\}\prod_{i=1}^{n} P\{N(I_i) = 1\} \\ &\approx exp\{-\Lambda(\mathbf{A};\beta)\}\prod_{i=1}^{n}\lambda(x_i;\beta)\delta_i. \end{aligned}$$

Dividiendo para $\delta_i$ se obtiene la densidad que nos conduce a la siguiente función de verosimilitud

$$L(\beta;x_1,...,x_n) = exp\left[-\int_A \lambda(x;\beta)dx\right]\prod_{i=1}^{n}\lambda(x_i;\beta), \tag{2.5}$$

siendo

$$\Lambda(\mathbf{A};\beta) = \int_{\mathbf{A}} \lambda(x;\beta)dx, \tag{2.6}$$

Definiendo T como la longitud del periodo observado, la función de log-verosimilitud vendrá definida por la siguiente ecuación

$$LL(\beta,(t_i)_{i=1}^{n}) = -\sum_{t=1}^{T}\lambda(t;\beta) + \sum_{i=1}^{n} log\lambda(t_i,\beta) \tag{2.7}$$

Para dar flexibilidad al modelo, definimos la función $\lambda(t;\beta)$ en función de las covariables que seleccionaremos. Puesto que la intensidad debe ser positiva, tomamos la función exponencial para asegurarnos de que $\lambda(t;\beta)$ cumple esta condición, por lo que definimos $\lambda(t;\beta) = exp(\mathbf{X}^T(t)\beta)$ con $\lambda(t,\beta)$ constante para cada t, siendo $\mathbf{X}^T(t)$ el vector de covariables en el tiempo t y $\beta$ el vector de los parámetros a estimar.

### 2.2.2. Estimación

Se define $\hat{\beta}$ como el valor del vector de parámetros el cual hace que L($\beta$, **x**) alcance su valor máximo. Dada la muestra **x**, el método propone como estimador de $\beta$ el valor que maximiza la función de verosimilitud, dada una muestra finita de datos. Recordemos el método de máxima verosimilitud.

**Definición** (Método de máxima verosimilitud)**.** Sea $x_1,...,x_n$ una muestra aleatoria de una población X con función de verosimilitud L($\beta$). El valor $\hat{\beta}$= $\hat{\beta}(x_1,...,x_n)$ el cual cumple

$$L(x_1,...,x_n;\hat{\beta}) = \underset{\beta\in\mathfrak{B}}{max} L(x_1,...,x_n;\beta) \tag{2.8}$$

recibe el nombre de estimador máximo-verosímil de $\beta$.

El estimador de máxima verosimilitud de $\beta$ será aquel que haga que la derivada de la función de log-verosimilitud sea cero. Para maximizar la función de verosimilitud se puede recurrir a métodos numéricos.

Recordemos que el método de máxima verosimilitud proporciona estimadores con buenas propiedades debido a que son insesgados, consistentes, suficientes, la distribución del estimador máximo verosímil es asintóticamente Normal con esperanza igual al valor del parámetro estimado y además son asintóticamente eficientes, es decir, entre todos los estimadores consistentes de un parámetro $\beta$, los de máxima verosimilitud son los de varianza mínima.

A través de éste método podemos obtener la matriz de covarianza de los estimadores. La estimación de la matriz de covarianza de los MLE $\hat{\beta}$ se basa en la distribución asintótica de los estimadores de máxima verosimilitud y se calcula como la inversa del hessiano

$$\hat{V}(\hat{\beta}) = \left[ -\frac{\delta^2 LL(\beta)}{\delta\beta_l \delta\beta_k}\bigg|_{\beta=\hat{\beta}} \right]^{-1} \tag{2.9}$$

donde la segunda derivada de la verosimilitud es

$$\frac{\delta^2 LL(\beta)}{\delta\beta_l \delta\beta_k} = -\sum_{t=1}^{T} \lambda(\beta;t) X_l(t) X_k(t) \tag{2.10}$$

A través de esta estimación de la varianza se pueden calcular los intervalos de confianza. Esto es lo que veremos a continuación.

#### Intervalos de confianza

Dada la matriz de covarianza de $\hat{\beta}$, los intervalos de confianza para $\lambda(t)$ se pueden obtener realizando dos aproximaciones, el método delta o una transformación de los intervalos de confianza para $v(t) = \mathbf{X}(t)^T\beta$.

**Intervalos de confianza basados en los métodos delta.** Los métodos delta estiman la varianza de g($\mathbf{Z};\theta$), una función de un vector de C covariables, **Z**, en términos de la matriz de covarianza

$$V[g(\mathbf{Z};\theta)] \approx \sum_{l=1}^{C}\sum_{k=1}^{C} \frac{\delta g}{\delta\theta_l}\frac{\delta g}{\delta\theta_k} Cov[Z_l, Z_k], \tag{2.11}$$

con $\frac{\delta g}{\delta\theta_j} = \frac{\delta g(\mathbf{Z})}{\delta \mathbf{Z}_j}\Big|_{\mathbf{Z}=\theta}$ y $\theta$ un vector de parámetros los cuales, asintóticamente cumplen que $E[\mathbf{Z}] = \theta$. Aplicando esta aproximación a $exp(\mathbf{X}^T\hat{\beta})$, la varianza estimada de la tasa ajustada será

$$\hat{V}[\hat{\lambda}(t)] = \sum_{l=1}^{C}\sum_{k=1}^{C} \hat{\lambda}(t) X_l(t) \hat{\lambda}(t) X_k(t) \hat{V}[\hat{\beta}_l, \hat{\beta}_k]. \tag{2.12}$$

Sabemos que $\hat{\lambda}(t)$ es el estimador máximo verosímil de $\lambda(t)$ por lo que es asintóticamente Normal.

Consecuentemente, un intervalo de confianza para $\lambda(t)$ será

$$(\hat{\lambda} - z_{\alpha/2} s.e(\hat{\lambda}), \hat{\lambda} + z_{\alpha/2} s.e(\hat{\lambda})), \tag{2.13}$$

donde s.e ($\hat{\lambda}$) es la raíz cuadrada de $\hat{V}[\hat{\lambda}(t)]$.

El intervalo podría incluir valores fuera del rango [0,∞), así el valor más bajo del intervalo se define como el máximo entre 0 y $\hat{\lambda} - z_{\alpha/2} s.e(\hat{\lambda})$. Si el máximo es 0, esta aproximación no garantiza el nivel de confianza del intervalo, y por consiguiente debe ser considerada la siguiente aproximación.

**Transformación de intervalos de confianza** La segunda aproximación para calcular un intervalo de confianza de $\lambda(t)$ consiste en aplicar una transformación exponencial, debido a que ésta garantiza que los valores sean positivos, a un intervalo de confianza para el predictor lineal $\hat{v}(t) = \mathbf{X}^T(t)\hat{\beta}$. El cálculo de este intervalo se basa de nuevo en las propiedades asintóticas del estimador máximo verosímil $\hat{v}(t)$, y $\mathrm{V}(\hat{v}(t))$ se estima de la siguiente manera:

$$\hat{V}(\hat{v}(t)) = \mathbf{X}^T(t)\hat{V}(\hat{\beta})\mathbf{X}(t).$$

## 2.3. Selección de covariables

Para seleccionar las covariables que formarán parte del modelo, se disponen de diversas herramientas. Para trabajar con un número de covariables alto, existen métodos de selección automática, como el método Forward, Backward y Stepwise. Este último es uno de los métodos más empleados, el cual es una combinación de los dos anteriores.

Para entender el método Stepwise, debemos conocer en que consisten los métodos Forward y Backward. El método Forward comienza por el modelo mas simple, y a través de éste, realizando algún test, se elige la covariable que más información aporta al modelo según el test realizado. En los pasos sucesivos se selecciona la segunda covariable que aporta más información al modelo y así sucesivamente para estudiar todas las covariables e interacciones posibles. El algoritmo finaliza cuando las variables no seleccionadas son las que no aportan información relevante al modelo. El método Backward (o de eliminación hacia detrás) actúa de forma inversa. Se comienza seleccionando todas las variables. El algoritmo finaliza cuando las variables seleccionadas son las que aportan información relevante al modelo.

Así el método Stepwise consiste en introducir o eliminar en cada paso una covariable dependiendo de si aporta, o no, información relevante al modelo. Además este modelo permite la posibilidad de modificar decisiones tomadas en pasos anteriores, bien sea eliminando del conjunto seleccionado la variable introducida en un paso anterior del algoritmo, bien sea seleccionando una variable previamente eliminada.

Cuando el número de covariables no es muy alto, es preferible hacer una selección de las covariables de forma manual. En este caso, se propone un procedimiento paso a paso, basado en el método Stepwise, en el que partiendo del modelo más sencillo, se introduce una variable en cada paso de acuerdo al test de razón de máxima verosimilitudes que veremos a continuación.

### 2.3.1. Test máxima verosimilitud

Este test compara dos modelos de manera que el primero (modelo reducido $\mathscr{M}_0$) es un caso particular del segundo modelo (el modelo alternativo o general $\mathscr{M}_1$). El test analiza si los datos son más verosímiles bajo el modelo general que bajo el modelo reducido, y consecuentemente si el modelo general debería ser modelizado o no.

**Definición** (Test de razón de verosimilitud)**.** Supongamos que el modelo $\mathscr{M}_0$ con parámetro $\theta^{(2)}$ es el submodelo de $\mathscr{M}_1$ con parámetro $\theta_0 = (\theta^{(1)}, \theta^{(2)})$ bajo la restricción $\theta^{(1)} = 0$. Sean $L_0(\mathscr{M}_0)$ y $L_1(\mathscr{M}_1)$ los valores máximos de la función log-verosimilitud de los modelos $M_0$ y $M_1$ respectivamente. El test de contraste de $M_0$ frente a $M_1$, con un nivel de significación $\alpha$, consiste en rechazar $M_0$ si $D = 2\{L_1(M_1)) - L_0(M_0)\} > c_\alpha$, siendo $c_\alpha$ el cuantil (1-$\alpha$) de la distribución $\chi^2_k$.

## 2.4. Validación

Una vez seleccionado el modelo, es necesario validarlo, es decir comprobar que los datos satisfacen las hipótesis necesarias en las que se basa el modelo. No existen herramientas específicas para validar un PPNH, aunque utilizando una técnica habitual para transformar un proceso de Poisson no homogéneo (PPNH) en uno homogéneo (PPH), podemos utilizar las herramientas para la validación de un PPH.

**Transformación de un PPNH a un PPH**

Un PPNH puede ser homogeneizado a través de la siguiente transformación de escala de tiempo, donde $t_i^{NH}$ son los puntos de la ocurrencia del proceso original y $t_i^H$ serán los puntos de ocurrencias del PPH.

$$t_i^H = \int_0^{t_i^{NH}} \lambda(t)dt. \tag{2.14}$$

**Validación de un PPH**

Veamos a continuación dos tipos de residuos a través de los cuales podemos validar el modelo.

**Residuos exponenciales o uniformes**

Una vez transformado el PPNH en un PPH, la validación de análisis consiste en analizar los residuos exponenciales, es decir, la distancia entre los eventos ($d_i^*$= $t_i^H$- $t_{i-1}^H$), que bajo la hipótesis nula deben ser independientes idénticamente distribuidas con distribución exponencial. Equivalentemente, la validación puede basarse en los residuos uniformes, es decir, las distancias transformadas $\exp(-d_i^*)$, que deben ser independientes idénticamente distribuidas con distribución uniforme.

En conclusión, el análisis de validación empieza transformando los puntos de ocurrencia de un PPNH a un PPH. Después se calculan los residuos uniformes del proceso homogéneo. Finalmente se estudia el comportamiento uniforme y la correlación serial de los residuos uniformes.

Para estudiar el comportamiento uniforme, se realiza un test Kolmogorov-Smirnov y una qqplot uniforme con un nivel de confianza de un 95 % y para analizar la correlación, se calcula el coeficiente de Pearson y una gráfica de correlación serial.

Recordemos que la prueba de Kolmogorov-Smirnov para una muestra es un procedimiento de "bondad de ajuste", que permite medir el grado de concordancia existente entre la distribución de un conjunto de datos y una distribución teórica específica. Su objetivo es señalar si los datos provienen de una población que tiene la distribución teórica especificada.

Recordemos también, que el coeficiente de Pearson es un valor entre -1 y 1. Si éste valor es 0 o muy próximo a 0 podremos considerar que los residuos son incorrelados.

**Residuos brutos**

Una aproximación complementaria para validar un PPNH es el análisis de los residuos brutos. El proceso de error es $\varepsilon(t) = N(t) - \int_0^t \lambda(u)du$.

El proceso de resido bruto correspondiente será:

$$R(l) = \sum_{t\in(0,l)} I_{t_i} - \int_0^l \hat{\lambda}(u)du.$$

donde $I_{t_i}$ es un indicador de variable, el cual es 1 en el instante $t_i$. En esta memoria los residuos se definen considerando los incrementos del proceso bruto en los intervalos $(l_1, l_2)$, que pueden ser disjuntos o solapados. Para obtener los valores instantáneos, los residuos se dividen por la longitud del intervalo $l_2 - l_1$.

$$r(l_1,l_2) = \frac{1}{l_2-l_1}\left(\sum_{t_i\in(l_1,l_2)} I_{t_i} - \int_{l_1}^{l_2} \hat{\lambda}(u)du\right). \tag{2.15}$$

Estos residuos pueden interpretarse como los valores residuales observados menos ajustados y estos son muy útiles para validar la intensidad ajustada del PP. Si el modelo es adecuado, $r(l_1,l_2) \approx 0$, es decir, el valor de los residuos debe ser es aproximadamente 0.

Para calcular los residuos brutos, se calculan los residuos en intervalos solapados y disjuntos.

-Los residuos $r_s(l_1,l_2)$ se basan en los intervalos solapados de una longitud dada, centrados en cada instante t. Debido a su estructura de solapamiento no serán independientes.

-Los residuos $r_d(l_1,l_2)$ se calculan en intervalos disjuntos de igual longitud, y se asigna al punto medio del intervalo.

Para calcular estos residuos se usan las intensidades empíricas y las intensidades medias ajustadas utilizando las siguientes funciones:

$$\frac{\sum_{t_i\in(l_1,l_2)} I_{t_i}}{(l_2-l_1)} \text{ y } \frac{\int_{l_1}^{l_2}\hat{\lambda}du}{(l_2-l_1)}.$$

.

La versión escalada de los residuos

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

se puede calcular utilizando una función de peso no negativa $h(u)$. La función de peso mas utilizada es $h(u) = \frac{1}{\sqrt{\hat{\lambda}}}$.

# Capítulo 3

# Aplicación

Una ola de calor se define como un intervalo de días consecutivos, con observaciones de temperatura máxima diaria, $T_x$, por encima de un determinado umbral extremo. En este modelo tomaremos el umbral determinado por la Agencia Española de Meteorología (AEMET) que definen el umbral de ola de calor como el percentil 10 de la serie de temperaturas en los meses de verano en un periodo de referencia. En Zaragoza el umbral será de 37.0ºC.

Para definir los valores de calor extremos utilizaremos la aproximación POT, como hemos justificado anteriormente. Para realizar el ajuste utilizaremos el paquete NHPoisson.

En este estudio, analizaremos la ocurrencia de olas de calor en los meses de verano. Para ello tenemos los datos de la temperatura diaria $T_x$, medidos en décimas de grado, correspondientes a los meses de mayo a septiembre comprendidos entre los años 1951 y 2005 en Zaragoza.

Los datos proporcionados se reflejan en la figura 3.1, donde sese visualiza el umbral tomado, y los datos que lo sobrepasan.

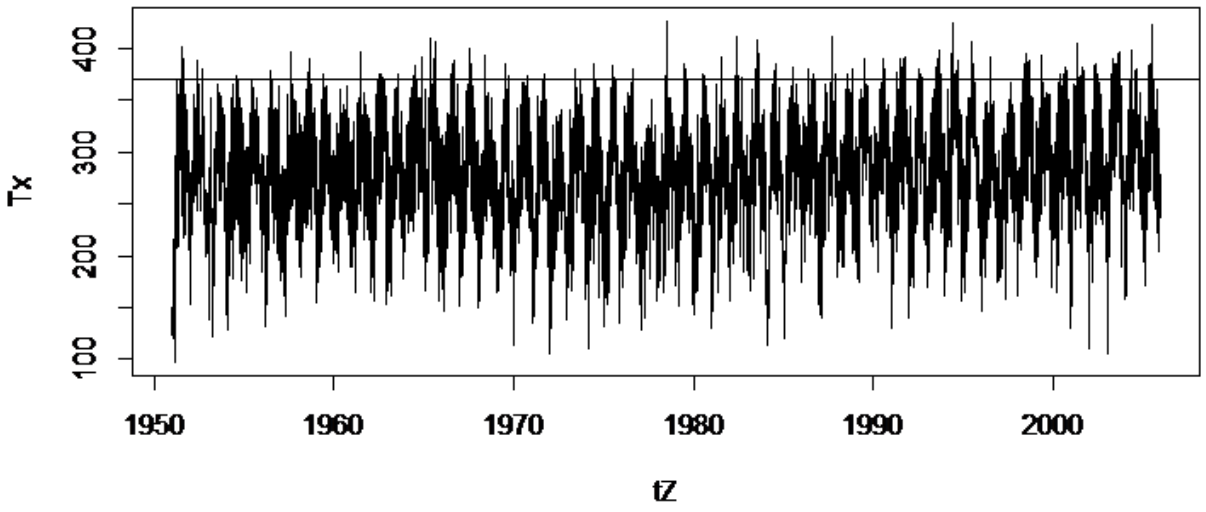


Figura 3.1: Temperatura máxima diaria.

## 3.1. Estudio de los datos

Para el siguiente estudio consideramos las siguientes variables como potenciales covariables:

1. Términos de temperatura:
   - TTx y TTn son variables que nos dan información de la evolución de la temperatura a largo plazo, definidas usando un suavizado tipo lowess de la temperatura máxima diaria, con una ventana del 30 %.
   - Txm31 y Tnm31 son variables que nos dan información de la evolución de la temperatura a corto plazo, definidas como un promedio móvil de 31 días centrado en cada día, la cual proporciona información sobre el estado local de la temperatura.
2. Términos estacionales: Consideramos como covariables los armónicos de primer orden, segundo, etc. Los términos armónicos ayudan a representar las variaciones estacionales, como la suma de una serie de funciones de coseno y seno.

Los términos de temperatura cuadráticos y las interacciones de la temperatura con los armónicos son covariables a tener en cuenta. Así, la función intensidad tendrá la siguiente estructura teniendo en cuenta el primer armónico:

$$\begin{aligned} log(\lambda(t)) = \beta_0 + \beta_1 * cos(2\pi * dia/365) + \beta_2 * sen(2\pi * dia/365) \\ + f_1(\ variables\ de\ termpertatura; \beta_j) \\ + f_2(cos(2\pi * dia/365)\ variables\ de\ temperatura; \beta_K) \\ + f_3(sin(2\pi * dia/365)\ variables\ de\ temperatura; \beta_l). \end{aligned}$$

Debido a que las olas de calor son fenómenos anómalos, la duración de estos sucesos será mucho menor que la duración de los periodos donde no se dan olas de calor.

Para empezar definamos los sucesos extremos, teniendo en cuenta que estamos realizando una aproximación POT donde sólo los puntos de máxima intensidad en cada suceso son considerados.

```
dateZ <- cbind(ano,mes, diames)

ZarEv <-POTevents.fun(T=Tx,thres= 370,date= dateZ)

 Number of events:  136
 Number of excesses over threshold 370 : 251
```

Por lo que tenemos 251 valores por encima del umbral, y como, a través del enfoque POT hemos dicho que, sólo asignamos un punto de ocurrencia a cada suceso, tendremos 136 sucesos. Los puntos de máxima intensidad en esos sucesos son los instantes de ocurrencia.

Los datos que estudiamos están tomados en tiempo discreto. Estos datos son tomados diariamente durante un periodo de tiempo de 8416 días. Por consiguiente, debido a que la unidad de tiempo tomada es muy pequeña comparada con la longitud de tiempo, podemos considerar que estamos ante un modelo de carácter continuo.

## 3.2. Selección del modelo

Debido a las características de los sucesos extremos de calor (comportamiento estacional y tendencia no monótona) tenemos que modelar a través de un proceso de Poisson con una intensidad no homogénea. La representación de la intensidad es una función determinista de covariables como los términos de temperatura a corto y largo plazo y términos estacionales, definidos con anterioridad. Las interacciones de los términos de temperatura con los términos estacionales serán estudiados posteriormente como posibles covariables.

La selección stepwise, en la cual nos basamos, consiste en un procedimiento paso a paso basada en el test de razón de verosimilitudes, visto en la sección 2.3.

En primer lugar, se ajusta el modelo más sencillo posible.

```
modZ.1<- fitPP.fun(covariates = NULL, posE = ZarEv$Px, inddat = ZarEv$inddat,
 tit = "ZARAGOZA Tx; Intercept", start = list(b0 = 1))
  Number of observations  not used in the estimation process:  115
  Total number of time observations:  8415
```

```
  Number of events:  136

  Convergence code:  0
  Convergence attained
  Loglikelihood:  -695.144

  Estimated coefficients:
  b0
  -4.111
  Full coefficients:
  b0
  -4.111
  attr(,"TypeCoeff")
  [1] "Fixed: No  fixed parameters"
```

A continuación veremos si el armónico de primer orden debe ser incluido en el modelo.

```
> covZ <- cbind(cos(2 * pi * dia / 365), sin(2 * pi * dia / 365))
> modZ.2 <- fitPP.fun(covariates = covZ, posE = Px, inddat = inddat,
  + tit = "ZARAGOZA Tx; Cos, Sin", start = list(b0 = -100, b1 = 1,  b2 = 1),
  + modSim = TRUE, dplot = FALSE, modCI = FALSE)
> aux <- testlik.fun(ModG = modZ.2, ModR = modZ.1)
       General Model (hypothesis H1):  ZARAGOZA Tx; Cos, Sin
       Reduced Model (hypothesis H0):  ZARAGOZA Tx; Intercept
       ML ratio test statistic:  185.83
       P-value: 0
```

Se contrasta la hipótesis nula $H_0 : \beta_{cos} = \beta_{sen} = 0$. Tomando un nivel de significación 0,05, rechazaremos la hipótesis nula, puesto que el p-valor es 0 (menor que el nivel de significación), es decir introduciremos en el modelo el primer armónico.

Veamos si introducimos ahora el armónico de segundo orden.

```
> covZ3 <- cbind(cos(2*pi * dia /365), sin(2*pi*dia /365), cos(4*pi* dia /365),
  + sin(4 * pi * dia /365))
> modZ.3 <- fitPP.fun(covariates = covZ3, posE = Px, inddat = inddat,
  + tit = "ZARAGOZA Tx; Cos, Sin, Cos2, Sin2", start = list(b0 = -100,  b1 = 1,
  + b2 = 1, b3 = 1, b4 = 1), modSim = TRUE, dplot = FALSE, modCI = FALSE)
> aux <- testlik.fun(ModG = modZ.3, ModR = modZ.2)
       General Model (hypothesis H1):  ZARAGOZA Tx; Cos, Sin, Cos2, Sin2
       Reduced Model (hypothesis H0):  ZARAGOZA Tx; Cos, Sin
       ML ratio test statistic:  1.21
       P-value: 0.545
```

El p-valor es mayor que el nivel de significación ($\alpha = 0{,}05$), por lo que no rechazamos la hipótesis nula $H_0 : \beta_{cos2} = \beta_{sen2} = 0$, es decir no introducimos en el modelo el segundo armónico.

Veamos ahora si introducimos la covariable Txm31.

```
> covZ4 <- cbind(cos(2 * pi * dia /365),  sin(2 * pi * dia /365), Txm31)
> modZ.4 <- fitPP.fun(covariates = covZ4, posE = Px, inddat = inddat,
  + tit = "ZARAGOZA Tx; Cos, Sin, Txm31", start = list(b0 = -100,  b1 = 1,
  + b2 = 1, b3 = 0), modSim = TRUE, dplot = FALSE, modCI = FALSE)
> aux <- testlik.fun(ModG = modZ.4, ModR = modZ.2)
       General Model (hypothesis H1):  ZARAGOZA Tx; Cos, Sin, Txm31
       Reduced Model (hypothesis H0):  ZARAGOZA Tx; Cos, Sin
       ML ratio test statistic:  95.77
       P-value: 0
```

Se contrasta la hipótesis nula: $H_0 : \beta_{Txm31} = 0$. Tomando un nivel de significación $\alpha$ = 0.05, rechazamos la hipótesis nula, puesto que el p-valor es 0 y éste es menor que $\alpha$. Debido a que la variable contiene información relevante para el modelo, ésta será introducida en el modelo.

Consecutivamente vemos si introducimos la covariable Tnm31.

```
> covZ5 <- cbind(cos(2 * pi * dia /365),sin(2 * pi * dia /365),Txm31,Tnm31)
> modZ.5 <- fitPP.fun(covariates = covZ5, posE = Px, inddat = inddat,  tit =
"ZARAGOZA Tx; Cos, Sin, Txm31, Tnm31", start = list(b0 = -100,  b1 = 1, b2 = 1,
 b3 = 0, b4=0), modSim = TRUE, dplot = FALSE, modCI = FALSE)
> aux <- testlik.fun(ModG = modZ.5, ModR = modZ.4)
      General Model (hypothesis H1):  ZARAGOZA Tx; Cos, Sin, Txm31, Tnm31
      Reduced Model (hypothesis H0):  ZARAGOZA Tx; Cos, Sin, Txm31
      ML ratio test statistic:  4.24
      P-value: 0.039
```

El p-valor es menor que 0.05, por lo que rechazamos la hipótesis nula $H_0 : \beta_{Tnm31} = 0$ e introduciremos en el modelo la variable Tnm31.

Posteriormente veamos porque no introducimos las covariables TTx y TTn en el modelo.

Contrastando la hipótesis nula: $H_0 : \beta_{TTx} = 0$, el p-valor resultante es 0.205 (mayor que $\alpha$=0.05), por lo que no se rechaza la hipótesis nula. Así la variable TTx no entra en el modelo. Así mismo, contrastando la hipótesis nula: $H_0 : \beta_{TTn} = 0$ vemos que el p-valor resultante es 0.148 (mayor que $\alpha$=0.05), por lo que no se rechaza la hipótesis nula. Así la variable TTn tampoco entrará en el modelo.

Además se han estudiado las siguientes interacciones:

$cos(2\pi * dia/365) * Txm31$, $sin(2\pi * dia/365) * Txm31$.

$cos(2\pi * dia/365) * Tnm31$, $sin(2\pi * dia/365) * Tnm31$.

Cuyos p-valores fueron 0.018 en ambos casos. Haciendo el estudio de validación, ambos son válidos y ademas presentan los mismos resultados, por lo que se pueden considerar ambos por igual. En esta memoria introduciremos en el modelo $cos(2\pi * dia/365) * Txm31, sin(2\pi * dia/365) * Txm31$.

Veamos algunos datos del modelo tomado, como el máximo valor de la log-verosimilitud, y los coeficientes estimados.

```
Number of observations  not used in the estimation process:  115
Total number of time observations:  8415
Number of events:  136
Convergence code:  0
Convergence attained
Loglikelihood:  -548.209
Estimated coefficients:
    b0      b1      b2      b3      b4      b5      b6
-77.378 -54.947 -24.570   0.242  -0.027   0.167   0.075
Full coefficients:
    b0      b1      b2      b3      b4      b5      b6
-77.378 -54.947 -24.570   0.242  -0.027   0.167   0.075
```

Por último utilicemos el criterio de información de Akaike (AIC), el cual es una medida de ajuste, que evalúa el grado de calidad del ajuste según el subconjunto de covariables.

Veamos el AIC para el modelo ajustado es cual es 1110.418.

```
> extractAIC(modZ.final)
[1]    7.000 1110.418
```

Por lo que concluimos que el modelo final de Zaragoza es:

$$log(\lambda(t)) = -77{,}37 - 54{,}94 * cos(2\pi * dia/365) - 24{,}570 * sen(2\pi * dia/365) + 0{,}242 * Txm31$$
$$- 0{,}027 * Tnm31 + 0{,}167 * cos(2\pi * dia/365) * Txm31 + 0{,}075 * sin(2\pi * dia/365) * Txm31. \qquad (3.1)$$

A continuación, veremos para cada covariable en el modelo (excepto el intercepto) el p-valor de un test de razón de verosimilitudes, que estudia si cada covariable aporta información relevante al modelo que contiene a las restantes.

Veamos que el p-valor de todas las covariables es menor que un nivel de significación 0.05, Así se rechazarán las hipótesis nulas asociadas a los contrastes realizados, es decir, todas las covariables aportaran información relevante al modelo.

```
> aux <- LRTpv.fun(modZ.final)
  The p-values of the LRT comparing the initial model and the model without the covariate
  p-values
  Cos          0.004
  Sin          0.005
  Txm31        0.000
  Tnm31        0.037
  cos*Tnm31    0.000
  sen*Tnm31    0.008
```

Veamos los intervalos de confianza al 95 % de los parámetros $\beta$, basados en la aproximación normal asintótica de los estimadores máximo verosímil de $\beta$.

```
> confintAsin.fun(modZ.final)
        2.5 %        97.5 %
b0 -121.03108606 -33.724366890
b1  -97.62488808 -12.269135809
b2  -43.94494135  -5.194564735
b3    0.10218779   0.381561325
b4   -0.05080150  -0.002472863
b5    0.03183852   0.301729930
b6    0.01419234   0.135937333
```

La figura 3.2 muestra la intensidad ajustada y los intervalos de confianza transformados. Este último es un método de obtención de los intervalos de confianza visto en la sección 2.3.

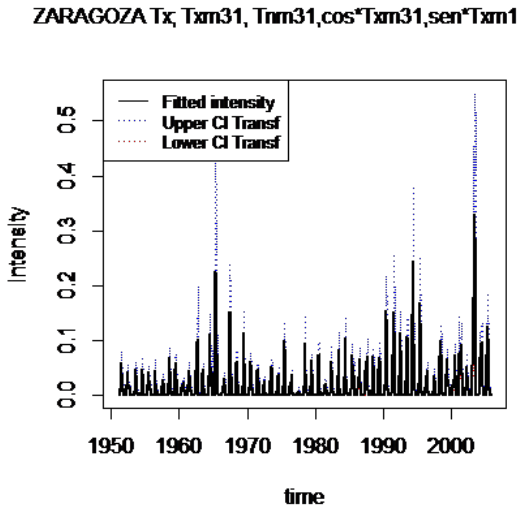


Figura 3.2: Intensidad ajustada e intervalos de confianza transformados.

## 3.3. Validación

Como hemos explicado en el capítulo anterior, podemos realizar la validación a través de los residuos uniformes y residuos brutos.

### 3.3.1. Residuos uniformes

#### Transformación del PPNH a un PPH

Para validar el modelo, a través de los residuos uniformes, necesitamos transformar el PPNH ajustado a un PPH (de la forma explicada en el capítulo anterior (sección 2.4)).

```
>  posEHZ <- transfH.fun(modZ.final)$posEH
```

#### Validación del PPH

A continuación calculamos los residuos uniformes y posteriormente estudiaremos el comportamiento uniforme y la correlación serial de los residuos como se ha explicado en la sección 2.4.

```
> resZ <- unifres.fun(posEHZ)
> graphresU.fun(unires = resZ$unires, posE = modZ.final@posE, Xvariables = cbind(covZ.final,
 dia) , namXv = c("cos", "sin", "Txm31","Txm31","cos*Txm31","sen*Txm31", "summer day index"),
  tit =
  "ZARAGOZA;  cos, sin, Txm31, Tnm31, cos*Txm31, sen*Txm31", addlow = FALSE)
```

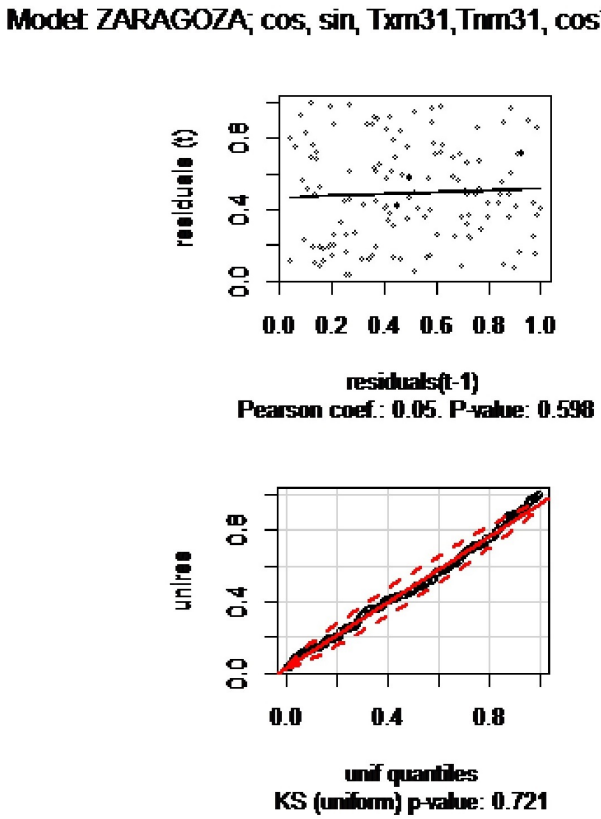


Figura 3.3: Gráficas de validación modelo final.

**Correlación entre los residuos**

Para analizar la correlación, se calcula el coeficiente de Pearson, como vemos en la parte inferior de la primera gráfica de la figura 3.3, éste coeficiente es 0.05. Contrastando la hipótesis $H_0 : \rho = 0$, el p-valor es de 0.598, el cual es mayor que 0.05, por lo que no rechazaremos la hipótesis de que los residuos sean incorrelados. Esto también se puede visualizar a través de la gráfica de dispersión, la cual debido a la dispersión de los puntos a través de toda la gráfica y la recta horizontal muestra que no hay una correlación evidente entre los residuos.

**Comportamiento uniforme**

Para analizar el comportamiento uniforme, se realiza un test Kolmogorov-Smirnov. Se obtiene un p-valor de 0.721, no se rechaza la hipótesis de que los residuos tienen un comportamiento uniforme. Además en la segunda gráfica de la figura 3.3 podemos observar que los puntos están dentro (o muy próximos) de la banda dibujada, por lo que los datos pueden razonadamente proceder de una distribución uniforme.

### 3.3.2. Residuos brutos

Finalmente, veamos en figura 3.4 una gráfica donde se compara la tasa empírica y la tasa ajustada.

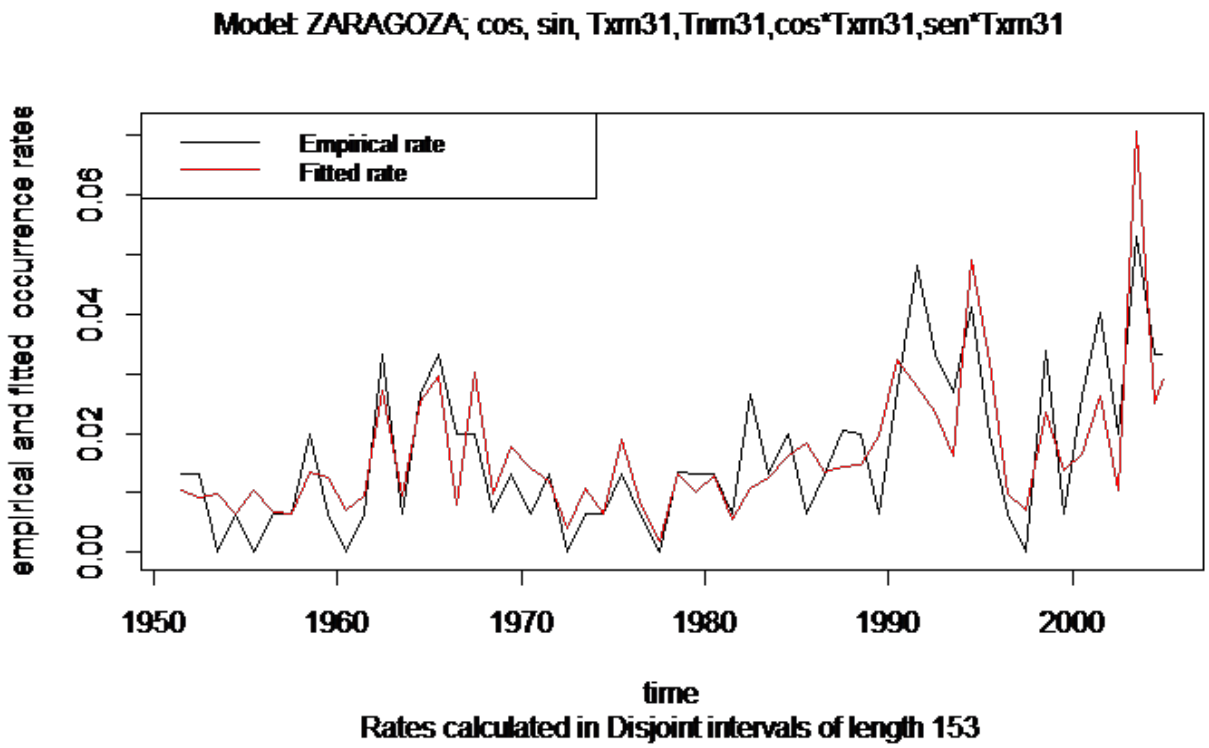


Figura 3.4: Intensidades ajustadas y empíricas de los intervalos disjuntos.

Esta gráfica nos muestra que la tasa ajustada y la tasa empírica se semejan bastante, por lo que el residuo bruto será bastante pequeño.

En conclusión el modelo tomado de Zaragoza es un modelo válido. Además se puede concluir que la ocurrencia de valores extremos cálidos muestra un comportamiento estacional. Así mismo, vemos que la tasa tiene un comportamiento creciente conforme al tiempo.

# Bibliografía


[1] CEBRIÁN, A., ABAURREA, J. AND ASÍN, J. (2015). , *NHPoisson: An R Package for Fitting and Validating Nonhomogeneous Poisson Processes.* Journal of Statistical Software, 64(6). disponible en

`https://www.jstatsoft.org/article/view/v064i06/v64i06.pdf`.

[2] ABAURREA, J. (2006), *Modeling and forecasting extreme hot events in the central Ebro valley, a continental-Mediterranean area*. Glob.Planet. Change.

[3] BINGHAM, N., GOLDIE, C. AND TEUGELS, J. (1987). , *Regular variation.* Cambridge u.a.: Cambridge Univ. Pr.

[4] CASTILLO, E. (1988), *Extreme value theory in Engineering.* Boston: Academic Press.

[5] CEBRIAN, A.C, *Análisis, Modelización y predicción de episodios de sequía*, Doctorada en Ciencias (Matemáticas). Departamento de métodos estadísticos. Universidad de Zaragoza, disponible en

`http://metodosestadisticos.unizar.es/personales/acebrian/tesis.pdf`.

[6] COLES, S. (2001), *An Introduction to Statistical Modeling on Extreme Values*. Bristol: Springer-Verlag, pp.20-249.

[7] COLLETT, D. (1994)., *Modelling survival data in medical research.*Chapman and Hall.

[8] COX, D. AND ISHAM, V. (1980)., *Point processes.*Chapman and Hall.

[9] EMBRECHTS, P., KLÜPPELBERG, C. AND MIKOSCH, T. (1999)., *Modelling extremal events*. Springer.

[10] FACULTAD DE CIENCIAS DE LA UNIVERSIDAD DE ZARAGOZA, *Directrices propias para la elaboración del trabajo fin de grado en Matemáticas*, disponible en

`https://ciencias.unizar.es/trabajo-fin-de-grado-en-matematicas`.

[11] GARCÍA, A. (2004)., *La teoría del valor extremo: una aplicación al sector asegurador.*,1st ed. [ebook] Alcalá de Henares (28802). Madrid., pp.30-36. disponible en

`http://www.actuarios.org/espa/web-nueva/publicaciones/anales/2004/art%2027-53.pdf`

[12] HÜSLER, J(1989). *Multivariate extreme values in stationary random sequences.* Departamento de estadística matemática. Universidad de Berna. Suiza. disponible en

`http://www.sciencedirect.com/science/article/pii/030441499090125C`.

[13] HSING, LEADBETTER AND HU?SLER, 1987 *On the exceedance point process for a stationary sequence.* Leuven: Kath. Univ.

[14] IBAÑEZ, A. (2011) *Análisis estadístico de valores extremos y aplicaciones.*Departamento de Estadística e Investigación Operativa.Universidad de Granada. disponible en

`http://masteres.ugr.es/moea/pages/tfm1011/analisisestadisticodevalore sextremosyaplicaciones/!`.

[15] LEADBETTER, M., LINDGREN, G. AND ROOTZE?N, H. (1983)., *Extremes and related properties of random sequences and processes.* New York: Springer-Verlag.

[16] MURILLO GÓMEZ, JUAN GUILLERMO. (2009). *La teoría de valor extremo y el riesgo operacional: una aplicación en una entidad financiera*. Revista Ingenierías Universidad de Medellín, 8(15, Suppl. 1), 59-70., disponible en

`http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-33242009000300007`.

[17] REAL ACADEMIA ESPAÑOLA, `http://www.rae.es/`.