Análisis, modelización y predicción de episodios de sequía

Memoria presentada por
Ana Carmen Cebrián Guajardo
para optar al grado de Doctor
en Ciencias (Matemáticas)

DEPARTAMENTO DE METODOS ESTADISTICOS
Universidad de Zaragoza

Índice General

Capítulo 1

Planteamiento del problema

El objetivo de esta memoria es el desarrollo de modelos estadísticos para analizar la ocurrencia y características de los episodios de sequía, y su aplicación en la predicción de magnitudes de interés en la planificación y gestión de recursos hídricos. La sequía es un fenómeno difícil de definir debido a su gran complejidad. Por una parte tiene diversas facetas -climática, hidrológica, edáfica, etc.- que pueden no coexistir temporal o espacialmente; por otra, su ocurrencia no depende sólo del comportamiento de la naturaleza sino que en ella influyen de forma decisiva factores humanos y socio-económicos: es la sociedad la que, con su organización y gestión de recursos, delimita su umbral de sensibilidad a la sequía.

De modo genérico, la sequía puede definirse como un periodo de tiempo en el que se observa un déficit hídrico inusual, que altera sensiblemente el desarrollo normal de la vida colectiva de una región. Obviamente, este concepto varía dependiendo de las características climáticas y de las necesidades de cada región, por lo que es imposible establecer una definición precisa de carácter universal. Algunas de las definiciones más generales son las propuestas por Baldwin (1941), que define un periodo de sequía como un intervalo de más de tres meses consecutivos con precipitación inferior al 50% de la normal, y por Gibbs & Maher (1967) que consideran que un año es seco, si la precipitación acumulada es inferior al percentil décimo calculado a partir de un periodo de referencia.

En esta memoria sólo se analiza el aspecto climático del fenómeno, aspecto que se considera fundamental por dos razones. En primer lugar, es un análisis básico por

ser el factor desencadenante: la escasez de precipitación -sequía climática- provoca la carencia de recursos hídricos -sequía hidrológica- que ocasiona daños económicos graves en el sector agrícola, que pueden trasmitirse a todo el conjunto de actividades socio-económicas. Por otra parte, los modelos y métodos desarrollados para este estudio pueden ser aplicados al tratamiento de otros aspectos del fenómeno, dada la común estructura formal en la que se pueden formular este tipo de problemas.

El estudio de las sequías es de gran interés, en particular en áreas como la Península Ibérica donde se producen de forma recurrente. En muchas regiones de la Península el volumen de lluvia de los meses de verano es escaso y la mayor parte de la precipitación anual se concentra en las épocas de primavera y otoño. Esta patrón climático resulta muy sensible a la sequía en zonas donde los recursos hídricos cubren la demanda de forma muy ajustada, ya que cualquier alteración de ese patrón, por ausencia de precipitación en una de las estaciones lluviosas, desemboca en un periodo prolongado de escasez hídrica.

El objetivo de esta memoria es el desarrollo de modelos que permitan caracterizar la ocurrencia y magnitud de las sequías, dando respuesta a cuestiones que se plantean, o deberían plantearse, en la planificación de recursos; por ejemplo, la duración y el déficit esperados de las sequías en una región, el tiempo medio de recurrencia entre ellas o las características del episodio seco más grave que cabe esperar en un intervalo de tiempo de longitud dada.

Organización de la memoria

En este capítulo se plantea el problema analizado, estableciendo la definición y caracterización de episodio de sequía que se utiliza. Se revisan las diferentes aproximaciones ensayadas en estudios anteriores y se describen las series históricas de precipitación en las que se aplicarán los modelos propuestos.

El capítulo segundo está dedicado a revisar la teoría de extremos necesaria para el desarrollo de los modelos propuestos en capítulos posteriores. Los resultados se presentan organizados en tres bloques, la teoría de extremos clásica, el método de 'excesos sobre umbral', y la aproximación basada en procesos puntuales, que incluye como caso particular a las dos anteriores.

En el tercer y cuarto capítulo se desarrollan modelos para la ocurrencia de periodos secos y sequías respectivamente. Una vez justificado teóricamente el carácter Poisson del proceso de ocurrencia, se estudia la validez de un proceso de Poisson compuesto para modelizar el proceso de los periodos secos, y de un proceso de Poisson cluster compuesto para las sequías. La caracterización se completa en el capítulo quinto con el ajuste de distribuciones adecuadas a las tres magnitudes fundamentales asociadas a una sequía, duración, déficit e intensidad máxima, y la determinación, para cada una de ellas, de la distribución del máximo en un periodo de tiempo determinado.

Finalmente, en el capítulo sexto se analiza al problema de la predicción en tiempo real. Se desarrollan modelos con estructura GLM para predecir la duración media restante de la sequía dada la información observada sobre ella hasta ese momento, y modelos para el riesgo de fallo con covariables dependientes del tiempo, con los que se predice la finalización o continuidad de un episodio en un instante.

Los modelos se contrastan utilizando series históricas españolas; en los capítulos tercero y cuarto se analizan las series de precipitación de seis observatorios -Burgos, Daroca, Huesca, Madrid, Murcia y San Fernando- con el fin de comprobar que las hipótesis requeridas por el modelo propuesto se verifican con carácter general en la vertiente mediterránea de la Península, la más propensa a la sequía, que de acuerdo con Benito, Orellana & Zurita (1994) corresponde aproximadamente a la mitad oriental de la misma. En el resto de los capítulos sólo se presentan, a modo de ejemplo, los resultados correspondientes a la serie de Huesca.

La realización de todos los análisis se ha programado con el paquete estadístico S-Plus. Al final de cada capítulo se incluye un apartado donde se describe la aplicación y características de las funciones construidas.

1.1 Aproximaciones al análisis de sequías

La investigación sobre fenómenos meteorológicos e hidrológicos extremos se ha centrado más en el análisis de excesos, tormentas extraordinarias e inundaciones, que en el de episodios de sequía. La aproximación más frecuente para estudiar este

fenómeno ha sido la teoría de rachas, aunque recientemente se han desarrollado otras alternativas. A continuación se exponen las diferentes aproximaciones utilizadas, y algunos de los trabajos más importantes de cada una de ellas.

Teoría de rachas Un periodo seco se puede definir como una racha o secuencia ininterumpida de observaciones menores o iguales que un valor de referencia, precedida y seguida por, al menos, una observación mayor que el nivel de truncación. El planteamiento de modelos probabilísticos para la duración de estas rachas permite obtener resultados de interés sobre algunas propiedades de las sequías. Los primeros modelos propuestos eran sencillos; Yevjevich (1967) fue el primero en desarrollar un modelo de este tipo basado en la distribución Geométrica, y en incorporar técnicas del análisis de series temporales para predecir la ocurrencia de sequías (Saldariaga & Yevjevich 1970).

Sen (1976) utiliza la teoría de rachas para analizar los periodos secos y no secos de series anuales de flujos fluviales. Posteriormente, se plantea el interés de encontrar modelos específicos para los valores extremos y sus trabajos se dirigen a obtener la función de distribución de la máxima longitud de racha en una serie de longitud finita. Obtiene esta distribución para series de variables i.i.d. (Sen 1991*b*), y deduce la función de distribución exacta del número de cruces en cadenas de Markov estacionarias de primer y segundo orden, utilizando técnicas de enumeración (Sen 1990, Sen 1991*a*).

Moyé, Kapadia & Cech (1988), tomando como base el modelo de Yevjevich y utilizando ecuaciones en diferencias, desarrollan una distribución de probabilidad que permite estimar el número esperado de sequías en un periodo de tiempo dado y la duración media de las mismas; aplican el modelo a series anuales de precipitación de Texas. En un trabajo posterior, Moyé & Kapadia (1995), analizan los episodios de sequía de las series anteriores utilizando resultados basados en estadísticos ordenados.

Un trabajo reciente en este campo es el de Lall, Rajagopalan & Tarboton (1996), que desarrollan un modelo no paramétrico para describir las rachas de días secos y no secos basado en estimadores de la densidad tipo núcleo, con el objeto de generar series sintéticas de precipitación. Aplican el modelo a series diarias de lluvia en

Utah.

Uno de los mayores inconvenientes de esta aproximación es que permite caracterizar la duración de los episodios secos, pero no otras características de interés como su déficit e intensidad. Intentando resolver esta carencia, Griffiths (1990), en un análisis de la sequía en Nueva Zelanda, propone un modelo probabilístico que no sólo describe las rachas secas de una serie de precipitación mensual, sino que caracteriza además el déficit asociado a las rachas ajustando una distribución Gamma; bajo estas hipótesis, estima las propiedades de la mayor longitud de racha y el máximo déficit esperados en un periodo de tiempo determinado.

Métodos de excesos sobre umbrales Estos métodos se basan en la hipótesis de que el proceso de ocurrencia de un fenómeno extremo tiene carácter Poisson. La aplicación de estas técnicas es frecuente en el análisis de fenómenos medio-ambientales: huracanes (Castro & Pérez-Abreu 1994), polución atmosférica (Smith 1989), etc. En el campo de las sequías se han aplicado al estudio de series fluviales. Por ejemplo, Zelenhasic & Salvai (1987) analizan la serie de flujo diario del río Sava, antigua Yugoslavia, y describen las sequías a partir de su duración, déficit e instante de ocurrencia, con el objetivo de caracterizar la mayor sequía durante un determinado periodo de tiempo. Madsen, Rosbjerg & Harremoes (1994) proponen un modelo POT para analizar episodios de sequía fluvial en dos ríos daneses, bajo la hipótesis de que la duración y el déficit de las mismas tienen una distribución Exponencial generalizada. Posteriormente, Madsen & Rosbjerg (1998), mejoran este modelo puntual y, a partir de él, desarrollan procedimientos de tipo bayesiano para realizar estimaciones en puntos donde no existen medidas.

Modelos de renovación alternante Una aproximación más general, que permite el estudio y caracterización tanto de los periodos secos como de los no secos, son los modelos de renovación alternante. Utilizando estos modelos, Kendall & Dracup (1992) realizan un estudio de los sucesos de sequía en series de flujos fluviales; su objetivo es resolver uno de los problemas de los modelos para generar series sintéticas de flujo fluvial: la incapacidad para reproducir consistentemente la frecuencia de ocurrencia de los sucesos de sequía más severos observada en los registros históricos,

persistencia conocida como efecto Hurst.

En los últimos años, el interés por el comportamiento futuro de los elementos climáticos -aumento de las temperaturas, alteración de la ocurrencia de fenómenos extremos, etc.- ha crecido significativamente debido a las predicciones sobre un posible cambio climático, provocado por la acumulación de gases de efecto invernadero en la atmósfera. Este interés ha promovido la celebración de reuniones científicas interdisciplinares sobre la problemática general asociada a la sequía. Algunas publicaciones importantes, resultado de este tipo de congresos son:

- *'Coping with droughts'* (1983): publicación que recoge las ponencias presentadas en una reunión del *NATO Advanced Study Institute* dedicada al estudio de las sequías, celebrada en Lisboa en Junio de 1980.

- *'Stochastic Hidrology and Hydraulics'* (1991): esta revista dedica el número 4 del volumen 5 a presentar los trabajos expuestos en el *Workshop on Drought* celebrado en Washington en 1990 organizado por la *National Science Foundation* de EEUU.

- *'Extreme values: floods and droughts'* (1994): recoge los trabajos sobre este tema presentados en la conferencia internacional *Stochastic and Statistical Methods in Hidrology and Environmental Engineering* que tuvo lugar en Waterloo, Canadá, en 1993.

A pesar de la proximidad del problema, los estudios realizados en España sobre la sequía no son numerosos; desde la obra centenaria 'Memoria sobre las causas de la sequía de las provincias de Almería y Murcia y los medios de atenuar sus efectos' de J. Echegaray, los trabajos sobre el análisis de la sequía han sido principalmente de carácter descriptivo. Entre ellos, cabe destacar los siguientes.

- 'Las sequías en España' es una publicación del mismo tipo que las monografías citadas, que recoge las comunicaciones presentadas en las 'Jornadas sobre la sequía en España', organizadas en 1990 por la Real Academia de Ciencias Exactas, Físicas y Naturales en Madrid, tras la sequía de los años 1988-90 que tuvo gran repercusión social.

- Algunos trabajos se centran en el análisis de un episodio concreto grave. Raso et al. (1981) y Sales, Jambrino & Juste (1982) analizan la sequía que se produjo en los años 1978-81; la de los años 1988-89 es estudiada por Capel (1989), en toda España, y por Beser & Tico (1993) en el ámbito del País Vasco.

- Ascaso & Casals (1981) y Ortigosa (1987) describen los periodos de sequía observados en distintos observatorios de la depresión central del Ebro y La Rioja, respectivamente.

- Olcina & Rico (1994) y (1995) se centran en el análisis de las causas y efectos de las sequías en el sureste de la Península Ibérica.

- Pérez & Escrivá (1982), Pérez (1988) y Pita (1995) establecen una definición de la sequía y proponen una metodología para la descripción del fenómeno.

- Finalmente, señalaremos los trabajos de modelización de rachas de días secos utilizando cadenas de Markov de distinto orden realizados por Pérez et al. (1984) y Martín, Conesa & Moreno (1992).

1.2 Caracterización de la sequía

El procedimiento habitual en Hidrología para determinar la ocurrencia de una sequía consiste en definir una señal, $s(t)$, relacionada con la magnitud de interés -la precipitación, como en este caso, el caudal de un río, el nivel de un pantano, el grado de humedad en el suelo, etc.- y una curva de referencia, $Qr(t)$, que expresa, para cada periodo, un umbral crítico relativo a las necesidades de dicha magnitud. Diremos que nos encontramos en un periodo seco cuando la trayectoria del proceso estocástico $s(t)$ se encuentra bajo la curva de referencia $Qr(t)$.

Las magnitudes utilizadas para caracterizar un episodio seco, que se ilustran en la figura 1.1, son la **duración** o **longitud** $L = ts - te$, es decir, el intervalo de tiempo transcurrido entre los instantes de entrada y salida de la señal bajo $Qr(t)$, el **instante de ocurrencia** asociado al episodio seco, que generalmente se define como el punto medio del mismo $(ts + te)/2$, el **déficit** o **severidad** de la sequía, D, que es el área comprendida entre $Qr(t)$ y $s(t)$ mientras dura el episodio, y la **intensidad máxima**, IM, la máxima diferencia entre $Qr(t)$ y $s(t)$ durante el episodio seco.

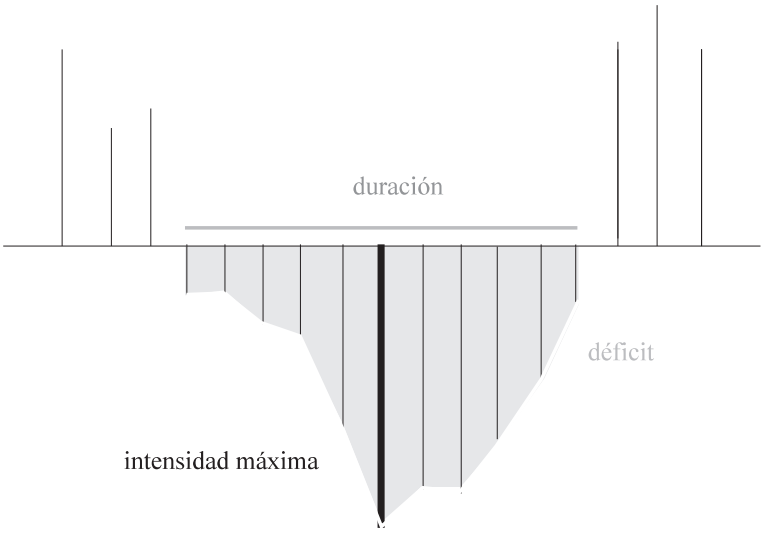

Figura 1.1: Caracterización de un episodio de sequía.

Planteado el problema de esta forma, el primer punto a determinar es la señal $s(t)$ que se va a utilizar, ya que los modelos y metodología posteriores dependerán, en parte, de esa elección. Dadas las características de la precipitación en la Península Ibérica, consideramos que el periodo temporal unitario más adecuado para el estudio es el mes. Partiendo de la serie cronológica de la precipitación acumulada por meses, figura 1.2, se define como señal $s(t)$ la serie de la lluvia anual acumulada móvil. Esta señal resulta adecuada porque, al tener una movilidad mensual, permite establecer de forma precisa el periodo de ocurrencia de un episodio; por otra parte, en zonas áridas, como gran parte de la Península, la sequía requiere un periodo temporal amplio para manifestarse y con la señal definida una observación por debajo del umbral implica que la precipitación del año anterior ha sido extrema. Una ventaja adicional es que la ausencia de componente estacional en $s(t)$ permite definir curvas de referencia constantes.

1.3 Series de precipitación analizadas

Las series de precipitación analizadas se seleccionaron entre las estudiadas por Almarza, López & Flores (1996), publicación en la que se revisan las 50 series de lluvia más largas de nuestro país. Además de presentar los valores medidos, detectan y

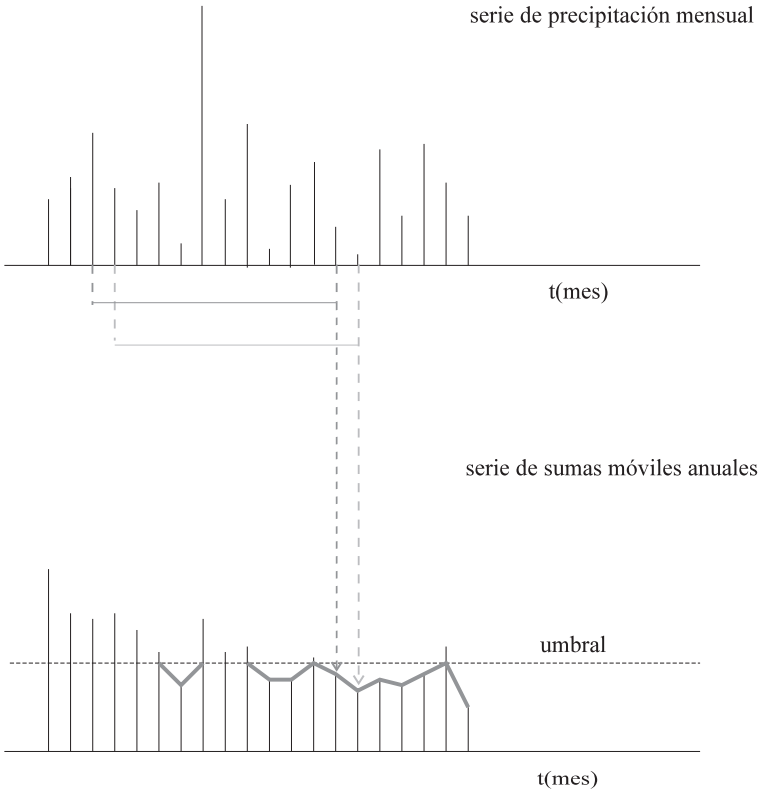

Figura 1.2: Construcción de la serie $s(t)$.

corrigen las inhomogeneidades observadas con el fin de obtener series coherentes. Entre estas series, se han seleccionado las que presentaban menor número de observaciones perdidas, mayor calidad y longitud, y una adecuada distribución espacial: Burgos, Huesca, Madrid, San Fernando -consideradas por Almarza series de referencia debido a su calidad- y Murcia. Los autores señalan que la serie de Murcia no se puede considerar homogénea, incluso después de la corrección aplicada, debido a un cambio en el aparato de medición en 1901; por esta razón, no se ha utilizado la serie registrada entre 1865 y 1900. Se han analizado además las series de Huesca y Daroca, proporcionadas por el Centro Territorial de Aragón, Navarra y Rioja del Instituto Nacional de Meteorología.

En las series seleccionadas existe un pequeño número de observaciones perdidas. Como la ausencia de una observación en la serie original de lluvia mensual implica doce observaciones perdidas en la serie $s(t)$, se han completado los huecos de cada serie con valores estimados. Se consideró la posibilidad de estimarlos mediante

correlación con series de observatorios próximos, pero la dificultad de obtener series de calidad y las bajas correlaciones existentes entre las disponibles llevó a buscar otras alternativas. Finalmente, las observaciones perdidas se estimaron a partir de la propia serie utilizando el siguiente procedimiento.

- La serie temporal de precipitación acumulada mensual se separa en doce series según el mes.

- Se calcula el orden p del percentil al que corresponden cada uno de los seis meses anteriores y los seis posteriores a la observación perdida en la serie mensual correspondiente.

- Con los valores de p calculados -doce, si todas las observaciones adyacentes están completas, o los disponibles en otro caso- se calcula una media ponderada con pesos, 1, 2, 3, 4, 5, 6, 6, 5, 4, 3, 2, 1 y se redondea al entero más próximo, pm. Se exige que el número de observaciones con el que se calcula la media ponderada sea al menos seis; en la práctica el número mínimo ha sido nueve.

- El valor de la observación perdida se estima con el percentil de orden pm de la serie mensual correspondiente.

A continuación se describen brevemente las características de los registros de los seis observatorios utilizados.

- **Burgos**. Serie registrada de 1-1862 al 12-1994 con 4 observaciones perdidas correspondientes a los meses: 11-1869, 7-1880, 4-1946, 7-1978. Ha tenido tres localizaciones: Instituto (1862 a 1943), Villafría (1944 a 10-1947), y Aeródromo de Villafría (11-1947 a 1994).

- **Daroca**. Serie registrada de 10-1909 al 5-1998 con 5 observaciones perdidas correspondientes a los meses: 7 a 10-1937, 5-1994. Ha tenido una única localización: Observatorio.

- **Huesca**. Serie registrada de 1-1862 al 12-1998 sin ninguna observación perdida. Ha tenido dos localizaciones: Instituto (1862 a 1948) y Monflorite (1949 a 10-1998).

- **Madrid**. Serie registrada de 1-1859 al 12-1994 con 2 observaciones perdidas correspondientes a los meses: 3 y 4-1939. Ha tenido dos localizaciones: Observatorio Astronómico (1859 a 1900) e Instituto Meteorológico (1901 a 1994).

- **Murcia**. Serie registrada de 1-1901 al 12-1994 con 5 observaciones perdidas correspondientes a los meses: 11 a 12-1936, 5-1942, 5-1988 y 7-1988. Ha tenido dos localizaciones: Instituto (1901 a 1944), y Aeródromo (1945 a 1994).

- **San Fernando**. Serie registrada de 1-1853 al 12-1988 con 1 observación perdida correspondientes al mes: 3-1986. Esta serie fue reconstruida por Almarza et al. (1996) uniendo información de dos observatorios próximos: la serie Hnos. Urrutia registrada en Cádiz de 1853 a 1881 y la de 1882 a 1988 del Observatorio del Instituto de la Marina en San Fernando; la serie final verifica todos los controles de homogeneidad.

En la tabla 1.1 se presenta un breve resumen descriptivo de las series de precipitación anual móvil de los observatorios analizados, que se completa con los correspondientes histogramas, figura 1.3, y representaciones de la series temporales, figura 1.4. No se aprecian indicios claros de existencia de tendencia o comportamiento cíclico en ninguna serie.

Benito et al. (1994) en su análisis de los patrones de precipitación en la Península Ibérica diferencian dos zonas en cuanto a la estructura temporal de la precipitación; por un lado, la vertiente atlántica en la que predomina la periodicidad anual, y por otro, la mediterránea en la que predomina el ciclo semestral. Comparando las series de precipitación mensual de los distintos observatorios estudiados, tabla 1.2 y figuras 1.5 y 1.6, se observa que San Fernando es un observatorio del primer tipo, mientras

Serie	Burgos	Daroca	Huesca	Madrid	Murcia	S.Fernando
N^o obs.	1596	1064	1644	1632	1128	1632
Media	5416.4	4284.0	5427.8	4284.3	2859.3	5628.5
Dt	1207.8	1004.2	1505.6	1023.2	1015.9	1626.5
Mín.	3031	1690	2484	2402	953	2309
Máx.	8902	6927	11350	7464	5747	9990
p50	5230	4213	5166	4078	2813	5246

Tabla 1.1: Análisis descriptivo de las series de precipitación anual móvil, medidas en dl.

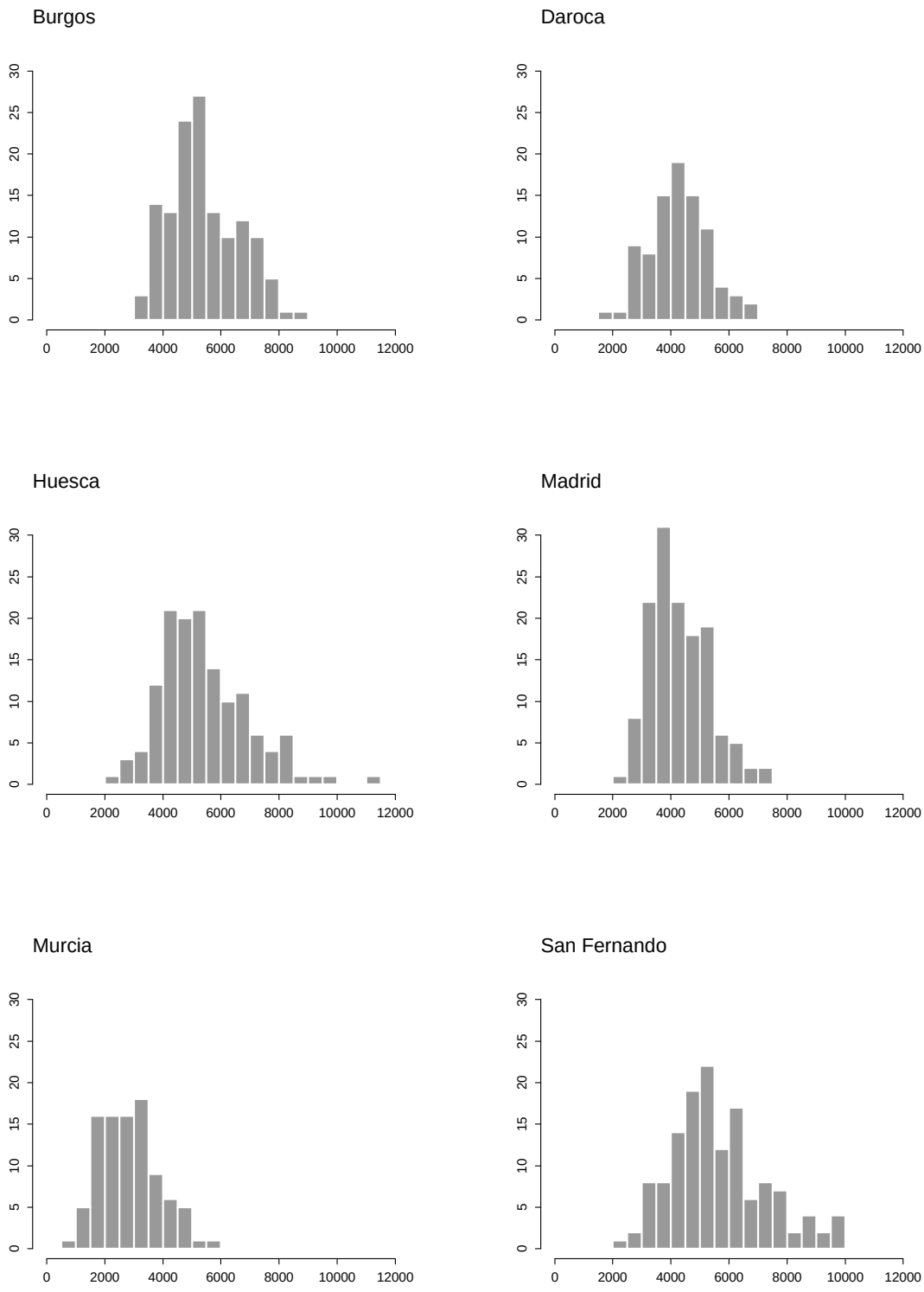

Figura 1.3: Histogramas de la precipitación anual.

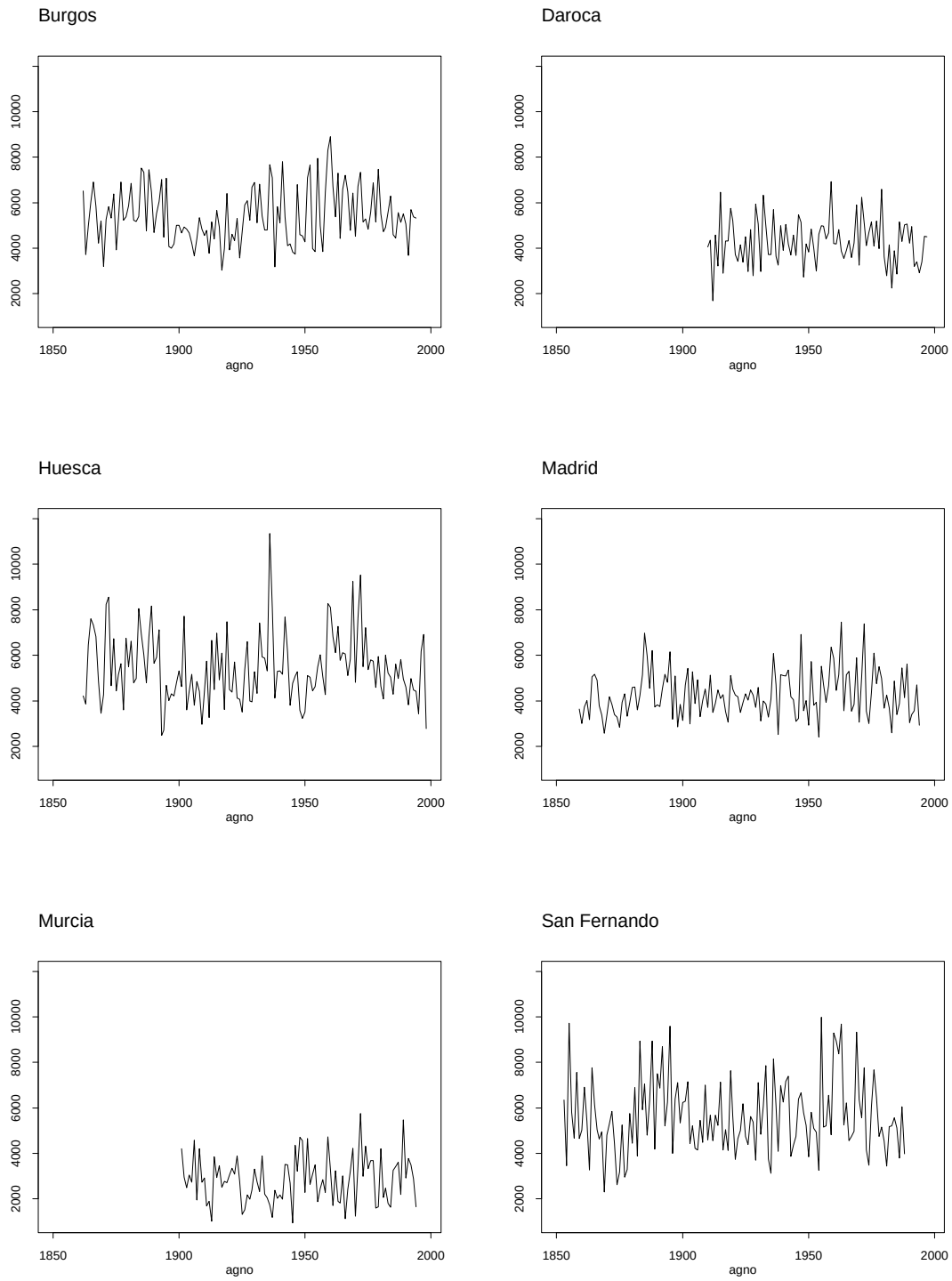

Figura 1.4: Series temporales de la precipitación anual.

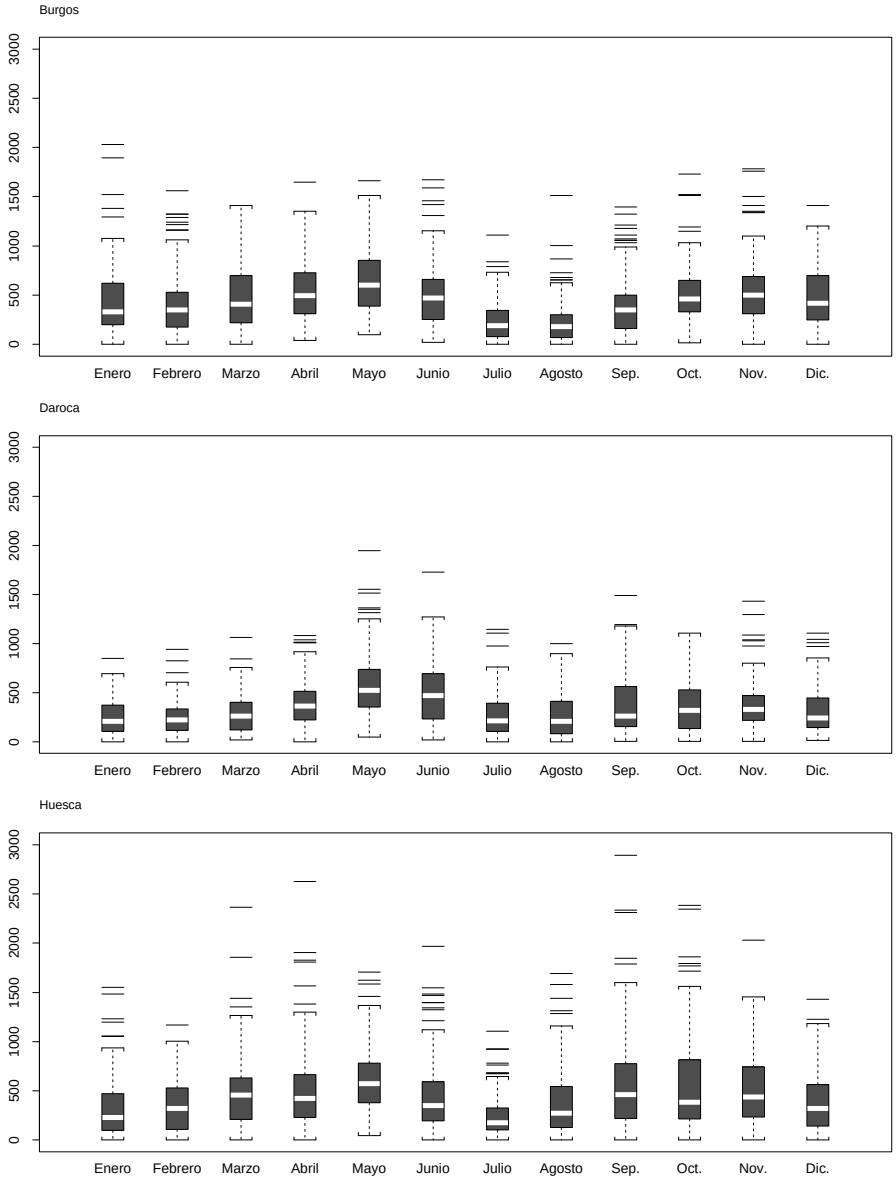

Figura 1.5: Gráficos de caja de las precipitación mensual.

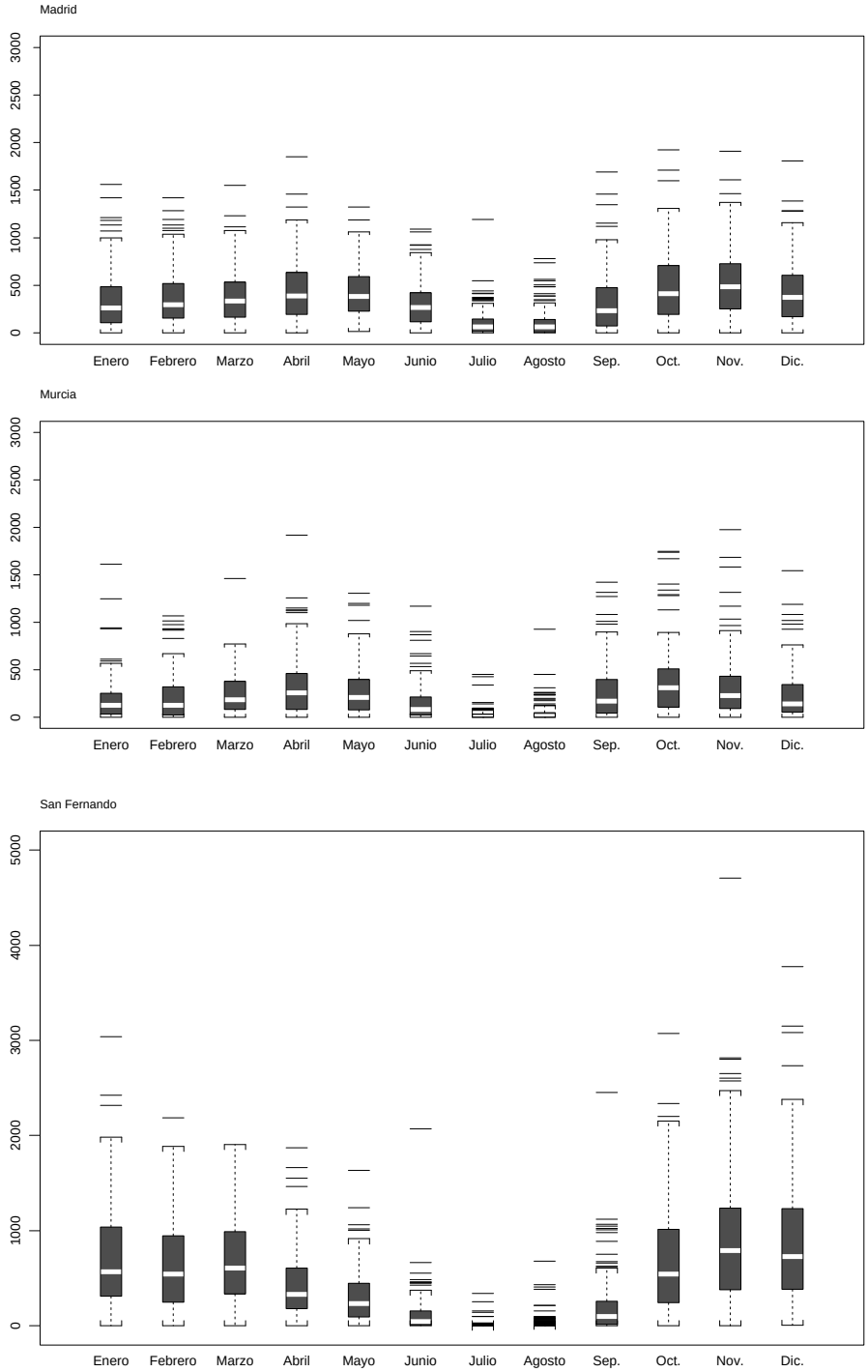

Figura 1.6: Gráficos de caja de la precipitación mensual.

Serie	Burgos	Daroca	Huesca	Madrid	Murcia	San Fernando
Enero	329	205	231	262	126	570
	197-618	104-370	100-473	108-482	33-248	312-1036
Febrero	351	220	322	295	123	545
	173-530	115-334	110-527	154-515	25-316	248-939
Marzo	405	261	458	337	184	609
	219-699	121 -399	210-633	166-531	80-375	337-987
Abril	494	363	421	387	224	332
	311-728	220-513	227-663	198-633	79-452	180-594
Mayo	600	522	571	382	208	233
	389-851	353-736	378-783	230-591	80-397	93-447
Junio	472	468	352	265	81	47
	252-659	233-690	193-594	119-415	25-211	10-158
Julio	189	211	176	61	0	0
	80-344	108-388	101-324	17-143	0-30	0-2
Agosto	180	207	271	62	8	0
	66-299	81-410	129-543	11-139	0-48	0-7
Septiembre	351	262	461	234	170	100
	161-499	155-563	220-775	75-476	40-391	16-252
Octubre	460	319	386	410	309	542
	330-650	134-529	213-813	197-707	107-508	248-1011
Noviembre	498	327	436	483	226	789
	309-690	216-471	234-740	252-725	90-425	384-1236
Diciembre	415	239	319	373	137	731
	247-697	144-446	142-562	170-607	52-342	387-1224

Tabla 1.2: Mediana y rango intercuartílico de las series de precipitación mensual, en dl.

Murcia es el representante más claro del segundo. El resto combinan en diferente medida ambos caracteres.

1.4 Implementación en S-Plus

- **Función: formatda.fun** (formatda.txt). Esta función cambia el formato de las series de Almarza et al. (1996). El resultado es una base de datos que contiene tres columnas correspondientes al año, el índice del mes y la precipitación mensual. Si está disponible la precipitación anual, verifica que no hay errores comprobando que la suma de los meses correspondientes coincide con el valor anual.

  Argumentos: No tiene argumentos; una vez ejecutado se debe introducir el nombre del fichero con los datos originales.

- **Función: relleno.fun** (relleno.txt). Esta función localiza, estima y completa las observaciones perdidas de una serie. El resultado es una base de datos con las series completas.

  Argumentos: No tiene argumentos; una vez ejecutado pregunta el nombre del fichero que contiene la serie a completar.

- **Función: mcsmo.fun** (mcsmo.txt). Esta función construye la serie de precipitación anual móvil. El resultado es una base de datos con tres columnas, el año, el mes y la correspondiente suma móvil. La primera observación corresponde a la observación número doce de la serie de precipitación mensual.

  Argumentos:

  - dataf: base de datos que contiene tres columnas, el año, el mes y la correspondiente precipitación mensual.

- **Función: descrip.fun** (descrip.txt). Esta función realiza un breve análisis descriptivo de la serie de precipitación mensual: medidas descriptivas, gráficos de caja, histogramas, representación temporal, correlogramas y correlogramas parciales por meses. También calcula la serie anual y hace un pequeño análisis descriptivo, histograma, gráfico de caja, gráfico probabilístico de normalidad y correlogramas.

  Argumentos:

  - dataf: base de datos de la serie de precipitación mensual.

Capítulo 2

Teoría de extremos

En este capítulo se presentan las definiciones y resultados fundamentales de la teoría de extremos que serán la base de las propiedades y modelos desarrollados en capítulos posteriores. Se pretende dar una visión general de su evolución hasta su estado actual, prestando especial atención a los resultados de interés para el análisis de sequías.

El objetivo central de la teoría de extremos es desarrollar procedimientos, estadísticamente justificables, para estimar la cola de una distribución desconocida, F, a partir de una muestra de datos. Este objetivo surge para dar respuesta a problemas que se plantean en multitud de áreas de aplicación, en particular en estudios de carácter medio-ambiental. Por ejemplo, la construcción de una estructura que debe soportar las condiciones generadas por un cierto proceso aleatorio, y que fallará en condiciones extremas del proceso. Obviamente, interesa diseñar la estructura de forma que su probabilidad de fallo sea muy pequeña, y para calcular esa probabilidad es necesario un estimador del comportamiento del proceso en niveles extremos.

Los principales problemas con los que hay que enfrentarse para realizar esa estimación son:

- La escasez de datos en la cola de la distribución; los datos extremos son por definición anómalos, por lo que el número de observaciones disponibles es, generalmente, muy pequeño.

- La necesidad de extrapolar la inferencia a situaciones no observadas, ya que

las estimaciones de interés corresponden a valores que sobrepasan los máximos de la muestra.

- La imposibilidad de usar la mayor parte de las técnicas de estimación estándar, que suelen producir estimadores de las colas muy sesgados.

Las bases de la teoría de extremos en el marco de muestras aleatorias fueron establecidas por Fisher y Tippett en la década de los 20, y unificadas y extendidas por Gnedenko en los años 40. La aplicación estadística de los modelos probabilísticos fue estudiada y formalizada por Gumbel en los años 50. En la década de los 70, Pickands generalizó las leyes límite clásicas, lo que permitió la mejora de los procedimientos de modelización. A partir de los años 80, se ha analizado el comportamiento de valores extremos en procesos estocásticos con estructura más general.

La teoría de extremos se ha desarrollado habitualmente para el análisis de máximos, por lo que la notación estándar hace referencia a este tipo de datos. La mayor parte de los resultados y técnicas son aplicables al análisis de mínimos, sin más que tener en cuenta la relación,

$$\min_{i\leq n} X_i = -\max_{i\leq n}(-X_i).$$

En este capítulo se revisa el método tradicional, basado en los resultados de convergencia de la sucesión de máximos, y otras aproximaciones alternativas más recientes.

2.1 Teoría clásica: análisis de máximos

Gumbel (1958) es la referencia clásica para estos métodos cuyo objetivo es caracterizar el comportamiento del máximo de muestras i.i.d. Dada una serie $X_1, X_2, \ldots, X_n$ de v.a. i.i.d. con distribución F, el comportamiento del máximo

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

es, en principio, trivial ya que,

$$P(M_n \leq x) = F^n(x).$$

En la práctica, F puede ser desconocida o dar lugar a una expresión de la f.d., función de distribución, del máximo compleja. Este inconveniente lleva a plantear una aproximación basada en argumentos asintóticos: el objetivo es caracterizar el comportamiento asintótico de M_n, en particular determinar la distribución límite cuando $n \to \infty$, para utilizarla como aproximación cuando n sea finito pero grande. En el caso de que la distribución límite fuera independiente de F se podría estimar la distribución de M_n directamente.

2.1.1 Leyes límite

La convergencia de M_n es un problema trivial y degenerado, ya que con probabilidad 1 la sucesión convergerá a $x_F = \sup\{x \in R \mid F(x) < 1\}$, el extremo superior, o punto final, de F. Alternativamente, como en el teorema central del límite, se puede considerar la distribución límite de la variable reescalada $(M_n - b_n)/a_n$ donde b_n y a_n son sucesiones de coeficientes normalizadores. Antes de presentar el teorema que resuelve la cuestión es necesario definir la siguiente relación de equivalencia en las distribuciones.

Definición 2.1. *Se dice que dos distribuciones F y F^* son del* **mismo tipo** *si existen constantes a y b tales que $F^*(ax + b) = F(x)$ para todo x.*

Teorema 2.1. *(Fisher-Tippett). Si existen sucesiones de constantes, a_n y b_n, tales que M_n tiene una distribución límite no degenerada, con función de distribución G,*

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

entonces G es del mismo tipo que una de las distribuciones siguientes, con $\alpha > 0$:

$$VE0: \quad G_0(x) = \exp[-\exp(-x)] \qquad si\ -\infty < x < \infty$$

$$VE1: \quad G_1(x) = \begin{cases} 0 & si\ x \leq 0 \\ \exp(-x^{-\alpha}) & si\ x > 0 \end{cases}$$

$$VE2: \quad G_2(x) = \begin{cases} \exp[-(-x)^{\alpha}] & si\ x < 0 \\ 1 & si\ x \geq 0. \end{cases}$$

Inversamente, cada una de estas distribuciones G_i, con $i = 0, 1, 2$, es la distribución límite de una sucesión $(M_n - b_n)/a_n$; en particular, del máximo de una serie de v.a. independientes con distribución G_i.

Es importante resaltar que este teorema no garantiza la existencia de límite no degenerado de la sucesión M_n, ni especifica qué tipo de distribución aparecerá si el límite existe.

2.1.2 Distribución Valor Extremo

El resultado anterior lleva a definir el concepto de distribución Valor extremo.

Definición 2.2. *Una función de distribución se dice de* **Valor extremo**, *VE, si es de uno de los tres tipos especificados en el teorema 2.1, que se denominan: distribución de Valor extremo de tipo 0 (VE0), 1 (VE1) o 2 (VE2), o Gumbel, Fréchet y Weibull*, respectivamente.*

Habitualmente, la distribución Weibull se define como la distribución de una v.a. positiva, con $F_W(x) = 1 - \exp(-x^\alpha)$. En el contexto de la teoría de extremos, con este nombre se denota a una de las posibles distribuciones límite de M_n, que se concentra en $(-\infty, 0)$ y tiene f.d. $G_2(x;\alpha) = 1 - F_W(-x;\alpha)$; para evitar confusiones la denotaremos Weibull*.

Las características de estas distribuciones son muy diferentes, aunque existe una relación funcional entre ellas,

$$X \sim VE1(\alpha) \Leftrightarrow \alpha \ln(X) \sim VE0 \Leftrightarrow -X^{-1} \sim VE2(\alpha).$$

Introduciendo un parámetro de forma, γ, se puede adoptar una parametrización que las comprende a todas,

$$G(x;\gamma) = \exp\left[-(1+\gamma x)^{-1/\gamma}\right] \qquad \text{para } 1+\gamma x > 0.$$

Esta distribución se puede generalizar añadiendo parámetros de localización, μ, y escala, σ:

$$G(x;\gamma,\sigma,\mu) = G\left(\frac{x-\mu}{\sigma};\gamma\right).$$

Los tipos 1 y 2 de la clase VE corresponden, respectivamente, a los casos $\gamma > 0$ y $\gamma < 0$. La distribución Gumbel, o tipo 0, se obtiene como el límite de $G(x;\gamma)$ cuando $\gamma \to 0$. Con esta notación, la relación entre las dos parametrizaciones es:

$$G(x;\gamma) = \begin{cases} G_1(x; 1/\gamma, 1/\gamma, -1/\gamma) & \text{si } \gamma > 0 \\ G_2(x; -1/\gamma, -1/\gamma, -1/\gamma) & \text{si } \gamma < 0 \\ G_0(x; 1, 0) & \text{si } \gamma = 0. \end{cases}$$

Las funciones de densidad de las distribuciones VE son unimodales. Las distribuciones Fréchet y Gumbel tienen asimetría a derecha y la distribución Weibull* es asimétrica negativa si $\alpha < 3.6$, positiva si $\alpha > 3.6$ y simétrica si α toma ese valor. En la distribución Weibull*, el parámetro de localización μ es el punto extremo inferior y en la Fréchet, el punto extremo superior.

Los momentos de orden j son,

$$\begin{aligned} m_{j,G1,\alpha} &= \Gamma(1 - j/\alpha) \qquad \text{si } \alpha > j \\ m_{j,G2,\alpha} &= (-1)^j \Gamma(1 + j/\alpha) \end{aligned}$$

y los cumulantes de la distribución Gumbel,

$$\begin{aligned} \kappa_{1,G_0} &= C_{Eu} \\ \kappa_{j,G_0} &= (-1)^j \frac{d\Upsilon}{dt}^{(j-1)}(1) \qquad \text{para } j \geq 2 \end{aligned}$$

siendo $\Upsilon(t) = \ln[\Gamma(1-t)]$.

La distribución VE se puede caracterizar a través de la clase de distribuciones max-estables.

Definición 2.3. *La distribución correspondiente a una v.a. X se dice* **max-estable***, si existen constantes normalizadoras $a_n > 0$ y b_n tales que,*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} X.$$

Teorema 2.2. *Una distribución es max-estable si y sólo si es una distribución Valor extremo.*

Las constantes normalizadoras de las distribuciones VE son:

- Gumbel: $b_n = \ln(n)$ y $a_n = 1$
- Fréchet: $b_n = 0$ y $a_n = n^{1/\alpha}$
- Weibull*: $b_n = 0$ y $a_n = n^{-1/\alpha}$.

2.1.3 Convergencia débil de los máximos. Dominios de atracción

Un gran esfuerzo en la investigación de teoría de extremos se ha dirigido a la caracterización de condiciones suficientes bajo las que el máximo normalizado converge, al cálculo de las constantes normalizadoras y a la determinación del tipo de distribución límite. A continuación se exponen algunos de los resultados de convergencia más importantes.

Teorema 2.3. *Dada una sucesión u_n de números reales y $\tau \in [0, \infty]$, la condición*

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

es equivalente a,

$$\lim_{n\to\infty} P(M_n \leq u_n) = \exp(-\tau).$$

Teorema 2.4. *Sea τ una constante en $[0, \infty)$ y F una f.d. con $x_F \leq \infty$; entonces, existe una sucesión u_n que verifica,*

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

si y sólo si,

$$\lim_{x\uparrow x_F} \frac{\bar{F}(x)}{\bar{F}(x^-)} = 1,$$

con $\bar{F}(x) = 1 - F(x)$.

Como consecuencia de este resultado, si una distribución F tiene un salto en el punto final de la distribución, no existe una distribución límite de M_n no degenerada. Por ejemplo, la distribución límite del máximo de variables con distribución Poisson, Geométrica o Binomial negativa, es degenerado.

Teorema 2.5. *Si M_n es la sucesión de máximos de una sucesión de v.a. i.i.d. con distribución F absolutamente continua, sus constantes normalizadoras son:*

$$\begin{aligned} a_n &= h^*(b_n) \\ b_n &= F^{-1}(1 - 1/n) \end{aligned}$$

donde h^* es la función de riesgo recíproca, $h^*(x) = \bar{F}(x)/f(x)$. Si la distribución límite de M_n es no degenerada, su parámetro de forma será $\gamma = \lim_{x\to x_F} dh^*(x)/dx$.

En conclusión, el teorema 2.1 asegura que bajo condiciones que garanticen la existencia de la distribución límite de M_n, ésta es VE, y el teorema 2.5 proporciona la forma concreta de esa distribución.

Otro punto de interés es la caracterización de las distribuciones F para las que el máximo normalizado converge a un determinado tipo de distribución VE.

Definición 2.4. *Una función de distribución F pertenece al* **max-dominio de atracción** *de una distribución G, MDA(G), si la distribución límite de la sucesión de máximos normalizados M_n con distribución de base F, es G.*

La caracterización del MDA de cada una de las tres distribuciones posibles requiere algunos resultados previos.

Proposición 2.1. *Una distribución F con constantes de normalización a_n y b_n pertenece al MDA de una distribución G Valor extremo, si y sólo si,*

$$\lim_{n\to\infty} n\bar{F}(a_n x + b_n) = -\ln[G(x)].$$

Definición 2.5. *Una distribución, F, se dice de* **variación regular** *con índice $-\alpha$, y se denota $F \in R_{-\alpha}$, si:*

$$\lim_{x\to\infty} \frac{\bar{F}(xt)}{\bar{F}(x)} = t^{-\alpha} \qquad con\ t > 0.$$

Estas funciones se caracterizan porque la cola del máximo M_n determina la cola de la suma $S_n = \sum_{i=1}^{n} X_i$, ya que las probabilidades $P(M_n > x)$ y $P(S_n > x)$ son del mismo orden.

Definición 2.6. *Dos distribuciones F y H se dicen* **cola-equivalentes**, *si $x_F = x_H$ y,*

$$\lim_{x\to x_F} \frac{\bar{F}(x)}{\bar{H}(x)} = c \qquad con\ 0 < c < \infty.$$

Una propiedad de estas distribuciones es que pertenecen al mismo dominio de atracción y tienen las mismas constantes normalizadoras, propiedad que permite definir los dominios de atracción como clases de equivalencia con constantes de normalización únicas asociadas a cada clase.

MDA de la distribución Fréchet

Teorema 2.6. *Una distribución pertenece al MDA(VE1) si y sólo si su función de distribución es una función de variación regular:*

$$F \in MDA(VE1) \Leftrightarrow F(x) \in R_{-\alpha}.$$

Las constantes normalizadoras son $b_n = 0$ y $a_n = F^{-1}(1 - n^{-1})$.

Teorema 2.7. *Condición de von Mises del MDA(VE1). Si F es una f.d. absolutamente continua tal que,*

$$\lim_{x\to\infty} \frac{xf(x)}{\bar{F}(x)} = \alpha,$$

entonces $F \in MDA(VE1)$. Además, sólo pertenecen al MDA(VE1) las funciones que verifican la condición de von Mises y las que son cola-equivalentes a ellas.

Las distribuciones Pareto y Cauchy pertenecen al MDA(VE1).

MDA de la distribución Weibull*

Teorema 2.8. *Una distribución F pertenece al MDA(VE2) si y sólo si su punto final es finito y $F(x_F - x^{-1})$ es una función de variación regular:*

$$F \in MDA(VE2) \Leftrightarrow \begin{cases} x_F < \infty \\ \bar{F}(x_F - x^{-1}) \in R_{-\alpha}. \end{cases}$$

Las constantes normalizadoras son $b_n = x_F$ y $a_n = x_F - F^{-1}(1 - 1/n)$.

Teorema 2.9. *Condición de von Mises del MDA(VE2). Si F es una f.d. absolutamente continua con densidad positiva en un intervalo finito (z, x_F), y*

$$\lim_{x\to\infty} \frac{(x_F - x)f(x)}{\bar{F}(x)} = \alpha,$$

entonces, $F \in MDA(EV2)$. Además, sólo pertenecen al MDA(VE2) las distribuciones que verifican la condición de von Mises y las cola-equivalentes a ellas.

Las distribuciones Beta y Uniforme pertenecen al MDA(VE2).

MDA de la distribución Gumbel

Definición 2.7. *Una función F es una* **función de Von Mises**, *si existe $z < x_F$ tal que F se puede expresar como,*

$$\bar{F}(x) = c \exp\left[-\int_z^x \frac{1}{a(t)} dt\right] \qquad \text{para } z < x < x_F$$

siendo c una constante positiva, y $a(t)$ una función positiva absolutamente continua tal que $\lim_{t \to x_F} da(t)/dt = 0$. Si existe, una elección posible de $a(t)$ es la función de riesgo recíproca $h^*(t)$.

Teorema 2.10. *Condición de von Mises del MDA(VE0). Si F es una función de von Mises, entonces $F \in$ MDA(VE0). Además, sólo pertenecen al MDA(VE0) las funciones de von Mises y las que son cola-equivalentes a ellas.*

Las distribuciones Exponencial, Weibull, Erlang, Normal, Gamma y Lognormal pertenecen al MDA(VE0).

2.1.4 Limitaciones del análisis de máximos

Una limitación importante de los resultados del análisis clásico de extremos, es que requieren la independencia de las observaciones. Esta condición es muy restrictiva, especialmente en problemas medio-ambientales, donde es frecuente que las series de observaciones presenten estructura de dependencia a corto plazo. En la sección 2.4 se revisarán resultados de la teoría de extremos aplicables en series dependientes.

Otro inconveniente de esta aproximación es que los procedimientos de inferencia basados en la distribución límite del máximo son ineficientes. Al considerar sólo el máximo observado en un determinado periodo de tiempo, generalmente un año, se pierde la información de un gran número de observaciones que también tienen carácter extremo. Para evitar este problema se han desarrollado técnicas que permiten incluir en el análisis un mayor número de datos extremos de la muestra. Por ejemplo, la generalización propuesta por Pickands (1975) y Weissman (1978), y desarrollada por Gomes (1981), Smith (1986) y Tawn (1988), consistente en considerar en cada periodo, un número k de estadísticos ordenados superiores. Otra aproximación, válida en las mismas condiciones que el análisis de máximos, que permite incorporar a la muestra un mayor número de observaciones extremas son los métodos EOT.

2.2 Excesos sobre un umbral

Los métodos de excesos sobre un umbral, EOT (*Excesses Over Thresholds)*, se basan en la hipótesis de que la ocurrencia de excesos sobre un umbral estricto en una serie de v.a. i.i.d. presenta un comportamiento Poisson, y que los excesos tienen una distribución exponencial o, más generalmente, Pareto generalizada, (Davison & Smith 1990).

Este tipo de métodos se desarrollaron en el campo de la Hidrología (Todorovic & Zelenhasic 1970), donde se aplicaban, sin justificación teórica, al análisis de series que presentaban dependencia a corto plazo utilizando el procedimiento denominado de picos sobre un umbral, POT, (*Peaks Over Thresholds*). Este método consiste en analizar la serie formada por los picos de la serie de excesos, definiendo un pico como el máximo valor observado durante una secuencia de excesos consecutivos. Sobre la serie de picos, que se supone i.i.d, se aplica el procedimiento EOT. Posteriormente se han aplicado también al análisis de series medio-ambientales, (Smith 1989, Davison & Smith 1990, Rasmussen et al. 1994).

Antes de justificar la aproximación EOT se revisan algunas propiedades del proceso de Poisson y la distribución Pareto generalizada.

2.2.1 Proceso de Poisson

Definición 2.8. *Un* **proceso de Poisson** *en* $\mathbb{R}^+$ *de tasa* $\lambda(t)$*,* $PP(\lambda(t))$*, es un proceso puntual en el que los puntos,* $T_1, T_2, \ldots$ *ocurren de forma totalmente aleatoria; denotando* H_t *a la trayectoria del proceso hasta el instante* t*, y* $N(u, v)$ *a la variable número de ocurrencias en el intervalo* $(u, v]$*, un proceso de Poisson verifica, para todo* t*,*

$$P(N(t, t+\delta) = 1 \mid H_t) = \lambda(t)\delta + o(\delta)$$
$$P(N(t, t+\delta) > 1 \mid H_t) = o(\delta)$$

de forma que,

$$P(N(t, t+\delta) = 0 \mid H_t) = 1 - \lambda(t)\delta + o(\delta).$$

Si la función $\lambda(t)$ *es constante, el proceso se denomina homogéneo, PPH, y en otro caso, no homogéneo, PPNH.*

Esta definición no es fácil de comprobar en la práctica, pero existen caracterizaciones equivalentes más sencillas de verificar.

i. 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. $N(A_1), N(A_2), \ldots$, con $A_1, A_2, \ldots$ intervalos temporales disjuntos, son independientes y tienen distribución Poisson de media $\lambda|A_1|, \lambda|A_2|, \ldots$, donde $|A|$ es la longitud del conjunto A.

ii. Caracterización de los tiempos de recurrencia. Los tiempos de recurrencia de un proceso puntual en $\mathbb{R}^+$ se definen como los intervalos transcurridos entre dos ocurrencias consecutivas, $Tr_1 = T_1,\ Tr_2 = T_2 - T_1, \ldots$ En un PPH los tiempos de recurrencia son realizaciones independientes de una v.a. Exponencial de parámetro λ. Una consecuencia de esta caracterización es que cada instante de ocurrencia T_i tiene distribución $\Gamma(i, \lambda)$.

La siguiente propiedad, (Leadbetter, Lindgren & Rootzén 1983), es de gran aplicación y se utilizará en análisis posteriores,

Proposición 2.2. *Todo proceso de ocurrencia obtenido como resultado de un proceso de borrado aleatorio -proceso en el que cada punto se elimina o permanece de forma independiente, con una probabilidad determinada p- de un proceso de Poisson, es también un proceso de Poisson.*

2.2.2 Distribución Pareto Generalizada

Definición 2.9. *Una distribución se dice* **Pareto generalizada***, PG, si su función de distribución es:*

$$H(x) = 1 - (1 + \gamma x)^{-1/\gamma},$$

con

$$\begin{aligned} x \geq 0 \qquad & si\ \gamma \geq 0 \\ 0 \leq x \leq -1/\gamma \qquad & si\ \gamma < 0. \end{aligned}$$

La distribución se puede generalizar añadiendo parámetros de localización μ y escala σ,

$$H(x; \gamma, \sigma, \mu) = H\left(\frac{x - \mu}{\sigma}; \gamma\right).$$

Esta distribución contiene a tres familias de distribuciones: Exponencial o PG0, Pareto o PG1, y Beta* o PG2, una subfamilia de las distribuciones Beta con parámetros $(\alpha, 1)$. Para comprobarlo, basta reparametrizar de forma conveniente,

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

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

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

con $\alpha > 0$. La relación entre las dos parametrizaciones es:

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

Entre las funciones de distribución PG y VE existe una relación funcional,

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

con $\ln[G(x)] > -1$; en concreto, las distribuciones PG0, PG1, y PG2 se corresponden con los tres tipos de la distribución Valor extremo, VE0, VE1 y VE2, respectivamente.

Las distribuciones Pareto, Exponencial y Beta* con parámetro $\alpha > 1$ son decrecientes en su dominio; la distribución Beta* con $\alpha < 1$ es creciente y, si $\alpha = 1$, coincide con la distribución Uniforme$[-1, 0]$. Los momentos de orden j correspondientes son,

$$\begin{aligned} m_{j,H1,\alpha} &= \alpha/(\alpha - j) \qquad \text{si } \alpha > j \\ m_{j,H2,\alpha} &= (-1)^j \alpha/(\alpha + j), \end{aligned}$$

y los cumulantes de la distribución Exponencial,

$$\kappa_{j,H0,\alpha} = (j-1)!\alpha^{-j}.$$

La siguiente propiedad es de gran utilidad en los análisis EOT y POT y caracteriza la distribución PG.

Definición 2.10. *La función de distribución F de una variable X se dice* **POT-estable** *si existen constantes a_u y b_u tales que la distribución del exceso sobre u, $X_u = X - u \mid (X > u)$, normalizado, tiene la misma distribución que X,*

$$F_{X_u}\left(\frac{x - b_u}{a_u}\right) = F(x).$$

Proposición 2.3. *La distribución PG es POT-estable y es la única que presenta esta propiedad; en particular, si X es $PG(\gamma, \sigma, \mu)$, para $u > \mu$ se verifica,*

$$X_u(x) \sim PG(\gamma, \sigma + \gamma(u - \mu), 0).$$

La ausencia de memoria de la distribución Exponencial es un ejemplo de la propiedad anterior.

2.2.3 Justificación asintótica

Cuando se definen umbrales muy estrictos, se puede justificar asintóticamente que la ocurrencia de los excesos sobre el umbral de una serie i.i.d. se comporta como un proceso de Poisson de intensidad,

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

En 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 $p_n = \bar{F}(u_n)$. Utilizando la aproximación de una distribución Binomial por una Poisson cuando $np_n \to \tau$, se obtiene el siguiente resultado (Leadbetter et al. 1983).

Teorema 2.11. *Sea (X_n) una serie de v.a. i.i.d. con distribución F, y r_n el número de excesos de la serie sobre el umbral u_n. Si la sucesión de umbrales (u_n) verifica,*

$$\lim_{n\to\infty} n\bar{F}(u_n) = \tau, \tag{2.1}$$

entonces, para $k = 0, 1, 2, \ldots$

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

Inversamente, si esta propiedad es cierta para un valor de k, se verifica la condición 2.1 y, en consecuencia, la propiedad es cierta para todo k.

Este teorema permite justificar el carácter 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 unidad $(0, 1]$, y su distribución es, en las condiciones del teorema anterior, aproximadamente Poisson. De forma análoga 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. En consecuencia, en condiciones que garanticen la existencia de la distribución límite de M_n, y con n grande, la ocurrencia de los excesos sobre u_n se comporta como un proceso de Poisson de tasa τ.

En el método EOT, la inferencia de los excesos se basa en la distribución de la variable exceso sobre u, X_u,

$$F_u(x) = P(X - u \leq x \mid X \geq u) = \frac{F(u+x) - F(u)}{\bar{F}(u)}.$$

La forma del límite de esta distribución cuando $u \to \infty$, se deduce del siguiente teorema obtenido por Pickands (1975) y Balkema & Haan (1974).

Teorema 2.12. *Sea H una función de distribución PG con valor γ dado; entonces,*

$$\lim_{u \uparrow x_F} \inf_{0<\sigma_u<\infty} \sup_{0 \leq x < \infty} |F_u(x\sigma_u + u) - H(x;\gamma)| = 0,$$

si y sólo si,

$$\lim_{n \to \infty} F^n(a_n x + b_n) = G(x;\gamma)$$

para $1 + \gamma x > 0$, y G una función de distribución VE.

Esencialmente, este resultado significa que F_u se aproxima bien por una distribución PG, si y sólo si F pertenece al máximo dominio de atracción de una distribución VE con el mismo parámetro de forma γ.

Los resultados anteriores se resumen en el siguiente teorema sobre la equivalencia del procedimiento EOT y el análisis clásico de máximos.

Teorema 2.13. *La convergencia del máximo muestral M_n a una distribución VE es equivalente a que la distribución de la tasa de excesos sobre umbrales u_n, con $u_n \uparrow \infty$, converja a una distribución Poisson, y la de los correspondientes excesos a una PG.*

2.3 Aproximación utilizando procesos puntuales

El procedimiento EOT y el procedimiento de los k-estadísticos superiores, con el análisis de máximos como caso particular cuando $k = 1$, se pueden incluir dentro de una aproximación más general basada en resultados de procesos puntuales. Esta aproximación permite incorporar a la inferencia todas las observaciones extremas, definiendo como extremos los valores que exceden un umbral estricto.

La idea básica consiste en formar, a partir de una serie (X_i) de v.a. i.i.d. con distribución desconocida F, un proceso puntual bidimensional $\{(i, X_i) : i = 1, \ldots, n\}$, y caracterizar el comportamiento de este proceso en regiones $[t_1, t_2] \times (u, \infty)$; de esta forma, se obtiene una representación del comportamiento de X_i en las colas. Formalmente, el argumento asintótico es el siguiente: dada $F \in MDA(G)$, con constantes normalizadoras a_n y b_n, y una serie de umbrales crecientes u_n, se define el proceso puntual bidimensional escalado,

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

El sentido de escalar el proceso al intervalo $[0, 1)$ es mantener el número esperado de excesos aproximadamente constante, ya que al crecer n, y en consecuencia u_n, los excesos sobre el umbral se hacen cada vez más escasos. Definiendo para cada n un proceso en los puntos $j/(n+1)$, con $j = 1, 2, \ldots, n$, la disminución del número de excesos provocada por el aumento de u_n se equilibra con la mayor densidad de puntos $j/(n+1)$.

Al crecer n, las ordenadas de P_n tienden a agruparse en torno al punto final inferior de la distribución de la variable normalizada, $(X_i - b_n)/a_n$, pero lejos de la frontera inferior se comportan como un proceso de Poisson no homogéneo. En efecto, en una topología de conjuntos que excluyan la frontera inferior, se comprueba la convergencia débil del proceso P_n a un proceso de Poisson y, en consecuencia, en niveles altos, el proceso P_n se puede aproximar por un proceso de Poisson de intensidad,

$$\Lambda\{(t_1, t_2) \times (x, \infty)\} = (t_2 - t_1) \left[1 + \gamma \frac{x - \mu}{\sigma} \right]^{-1/\gamma}.$$

Para obtener esta expresión basta calcular la intensidad asociada a un conjunto

$A = \{(0,1) \times (x, \infty)\}$,

$$\begin{aligned} \Lambda(A) &= -\ln\left[\lim_{n\to\infty} P\left(N_n(A) = 0\right)\right] \\ &= -\ln\left[\lim_{n\to\infty} P\left(M_n(A) < x\right)\right]. \end{aligned}$$

Si la sucesión de máximos es convergente,

$$\lim_{n\to\infty} P\left(M_n(A) < x\right) = \exp\left[-\left(1 + \gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right],$$

y teniendo en cuenta que $\Lambda(\{(t_1,t_2)\times(x,\infty)\}) = (t_2 - t_1)\Lambda(A)$ se obtiene la expresión buscada.

También la distribución PG de la variable exceso se deduce de esta aproximación; si denotamos $X_{i,n}$ a la i-ésima variable de una muestra de tamaño n,

$$\begin{aligned} \lim_{n\to\infty} \bar{F}_u(x; i, n) &= \lim_{n\to\infty} P\left(X_{i,n} > u + x \mid X_{i,n} > u\right) \\ &= \frac{\Lambda\{(0,1) \times (u+x, \infty)\}}{\Lambda\{(0,1) \times (u, \infty)\}} \\ &= \left(1 + \frac{\gamma x}{\sigma + \gamma u - \gamma\mu}\right)^{-1/\gamma}, \end{aligned}$$

que corresponde a una distribución PG con parámetro de escala $\sigma + \gamma u - \gamma\mu$.

Todos los resultados de teoría de extremos mencionados anteriormente pueden obtenerse a partir de esta representación.

2.4 Teoría de extremos para series dependientes

Los resultados expuestos hasta el momento se basan en la hipótesis de independencia de la serie (X_n). Sin embargo, algunas características frecuentes de las series medio-ambientales son la existencia de tendencias, la dependencia a corto plazo, la estacionalidad, la influencia de covariables externas, etc., características que impiden suponer la independencia de las observaciones. En este apartado se caracterizan las condiciones de dependencia bajo las que los resultados de la teoría de extremos siguen siendo aplicables.

Condiciones de convergencia del máximo de procesos estacionarios

Definición 2.11. *Un proceso se dice estrictamente* **estacionario** *si la distribución de todo vector de dimensión finita es invariante bajo desplazamiento temporal,*

$$(X_{t1}, \ldots, X_{tm}) \stackrel{d}{=} (X_{t1+h}, \ldots, x_{tm+h})$$

para cualquier elección de $t_1 < \ldots < t_m$ *y cualquier entero* h.

Definición 2.12. *Dado un proceso estacionario* (X_n)*, se llama* **proceso asociado** *a un proceso* $(\tilde{X}_n)$ *de v.a. i.i.d. con la misma distribución de probabilidad que las variables del proceso* (X_n).

El siguiente objetivo es establecer condiciones que permitan asegurar que el máximo muestral de una serie estacionaria, y el de la correspondiente serie asociada, tienen el mismo comportamiento límite.

Definición 2.13. *Diremos que una serie* X_n *verifica la* **condición** $D(u_n)$ *si para enteros cualesquiera,*

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

tales que $j_1 - i_p > l_n$ *se verifica,*

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

con $A_1 = \{i_1, \ldots, i_p\}$, $A_2 = \{j_1, \ldots, j_q\}$ *y* $\lim_{n\to\infty} \alpha_{n,l_n} = 0$, *para alguna sucesión* $l_n = o(n)$.

Esta condición es más débil que la mayor parte de las formas clásicas de dependencia y se verifica en procesos en los que no existe dependencia a largo plazo, o ésta es débil. Bajo esta hipótesis, se puede enunciar el siguiente resultado sobre la distribución límite del máximo de un proceso estacionario.

Teorema 2.14. *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 con* $u_n = a_n x + b_n$ *para todo* $x \in \mathbb{R}$, *y*

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

la distribución límite de M_n, G, *es una distribución VE.*

Así, la condición D_n se puede interpretar como una propiedad que garantiza que la dependencia a largo plazo de la serie es lo suficientemente débil para no afectar a la distribución límite de M_n. El teorema 2.14, no proporciona información sobre el tipo de la distribución G, ni sobre su relación con el límite de la serie asociada; para obtener esa clase de resultados es necesario imponer condiciones más fuertes.

Definición 2.14. *Diremos que una serie* (X_n) *verifica la* **condición** $D'(u_n)$ *si,*

$$\lim_{k\to\infty} \limsup_{n\to\infty} \; n \sum_{j=2}^{[n/k]} P(X_1 > u_n, X_j > u_n) = 0.$$

$D'(u_n)$ es una condición anti-cluster que limita la probabilidad de que se produzca más de un exceso en las variables $X_1, \ldots, X_{[n/k]}$; en efecto, esta condición implica,

$$E\left(\sum_{1\le i<j\le[n/k]} I_{\{X_i>u_n, X_j>u_n\}}\right) \le [n/k] \sum_{j=2}^{[n/k]} E(I_{\{X_1>u_n, X_j>u_n\}}) \to 0.$$

Teorema 2.15. *Sea* (X_n) *una serie estacionaria con f.d.* $F \in MDA(VE)$, a_n *y* b_n *las correspondientes constantes normalizadoras, y* $(\tilde{X}_n)$ *su proceso asociado. Si* (X_n) *satisface las condiciones* $D(u_n)$ *y* $D'(u_n)$ *con* $u_n = a_n x + b_n$ *y* $x \in \mathbb{R}$*, entonces,*

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

y

$$\lim_{n\to\infty} P\left(\frac{\tilde{M}_n - b_n}{a_n} < x\right) = G(x).$$

con G una función de distribución VE. Por consiguiente, se puede decir que las condiciones $D(u_n)$ y $D'(u_n)$ obligan a la sucesión (M_n) a comportarse como $(\tilde{M}_n)$, de forma que el problema del máximo de una serie estacionaria se reduce al de una serie i.i.d. Como consecuencia, bajo estas condiciones, la convergencia a un proceso de Poisson del proceso de ocurrencia de los excesos sobre un umbral estricto se mantiene Leadbetter et al. (1983).

En la mayor parte de las series medio-ambientales, la condición $D(u_n)$ es plausible; desafortunadamente, no se suele verificar la condición $D'(u_n)$, ya que es frecuente la existencia de dependencia temporal a corto plazo que provoca la formación de clusters de observaciones extremas.

Convergencia del máximo de series con dependencia a corto plazo

Definición 2.15. *Sea (X_n) un proceso estrictamente estacionario; si para todo $\tau > 0$, existe una sucesión u_n tal que,*

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

se llama **índice extremal de la serie,** *θ, a un valor no negativo tal que,*

$$\lim_{n\to\infty} P(M_n \leq u_n) = e^{-\theta\tau}.$$

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

$$\theta = \lim_{n\to\infty} P(\max(X_2, \ldots, X_{p_n}) \leq u_n) \mid X_1 \geq u_n),$$

donde $p_n = o(n)$.

El índice extremal se puede interpretar como el recíproco del tamaño medio asintótico de los clusters en niveles extremos; esta interpretación, más intuitiva, se basa en los resultados de Hsing, Husler & Leadbetter (1988).

El rango de valores posibles de θ es $[0, 1]$; en un proceso independiente el índice extremal es 1 -aunque $\theta = 1$, no implica independencia- y cuanto mayor es la tendencia del proceso a formar clusters en los valores extremos, más se aproxima θ a 0.

El efecto de la dependencia en el comportamiento asintótico de los máximos se establece en el siguiente teorema.

Teorema 2.16. *Sea (X_n) un proceso estacionario con índice extremal θ que satisface la condición $D(u_n)$ para $u_n = a_n x + b_n$; entonces,*

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

si y sólo si,

$$\lim_{n\to\infty} P\left(\frac{\tilde{M}_n - b_n}{a_n} < x\right) = G_*(x),$$

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

Este resultado indica que la distribución asintótica del máximo en un proceso con dependencia a corto plazo es también VE. Además, el tipo de la distribución, equivalentemente el parámetro de forma, no se ve afectado por la existencia de dependencia. Su único efecto es elevar a la potencia θ la distribución límite de la serie independiente ($\tilde{X}_n$) y, por consiguiente, sólo afecta a los parámetros de localización y escala.

En resumen, bajo la condición $D(u_n)$, los resultados asintóticos para procesos independientes son aplicables a procesos con dependencia a corto plazo.

2.5 Inferencia

En la inferencia de problemas de extremos se pueden utilizar distintos procedimientos -técnicas gráficas, método de los momentos, de mínima distancia, etc.- pero el más frecuente es el de máxima verosimilitud. Este procedimiento presenta la ventaja de que la teoría en la que se apoya está bien estudiada; además, se puede generalizar con facilidad, lo que permite su aplicación en situaciones con estructuras complejas.

La aplicación de resultados de inferencia máximo verosímil requiere la comprobación de las condiciones de regularidad necesarias. Este aspecto fue estudiado por Smith (1984), que estableció las siguientes conclusiones respecto a las distribuciones VE y PG:

- Si $\gamma > -0.5$, los MLE son completamente regulares.
- Si $-1 < \gamma < -0.5$, los MLE existen pero no son regulares.
- Si $\gamma < -1$ los MLE no existen.

El valor $\gamma < -0.5$ corresponde a distribuciones con una cola superior muy corta y acotada, características que no son frecuentes en los datos medio-ambientales por lo que, en general, éstos verificarán las condiciones necesarias para que los estimadores sean regulares.

Capítulo 3

Modelización del proceso de periodos secos mediante un Proceso de Poisson

Un indicio de la existencia de un estado de sequía es la entrada de la serie de precipitación por debajo de un determinado umbral; en consecuencia, una posible aproximación para estudiar este fenómeno consiste en analizar la serie de esos periodos, que denominaremos periodos secos. La teoría de extremos de procesos estocásticos estacionarios y los resultados de la aplicación del método POT a otras series medioambientales, apoyan la hipótesis de que. considerando umbrales suficientemente estrictos como los que definen el estado de sequía, el proceso de ocurrencia de periodos por debajo del umbral corresponderá a un PP.

El proceso de ocurrencia no es el único aspecto de interés de la sequía; en su modelización se deben considerar también las magnitudes que permiten caracterizar la importancia y gravedad de los episodios: el déficit, la duración y la intensidad máxima. Teniendo en cuenta los factores expuestos, un modelo que se adapta a las características del fenómeno y permite dar respuesta a los objetivos planteados es el proceso de Poisson compuesto.

3.1 Proceso de Poisson compuesto

Definición 3.1. *Un* **proceso de Poisson compuesto**, *PPC($\lambda, F_1, \ldots, F_k$), es un proceso puntual, en el que a cada ocurrencia se le asocia un vector aleatorio $(X1, \ldots, Xk)$, que verifica las siguientes hipótesis:*

i.- La sucesión de los instantes de ocurrencia: $0 \leq T_1 < T_2 < \ldots$ forma un proceso de Poisson(λ).

ii.- La muestra $(X1_i, \ldots, Xk_i)_{i=1\ldots n}$ del vector de magnitudes es i.i.d. con distribuciones marginales $F_1, \ldots, F_k$.

iii.- Las variables $X1, \ldots, Xk$, son independientes del proceso de ocurrencia.

El modelo más sencillo supone que las variables $(X1, \ldots, Xk)$ son independientes entre si, aunque se pueden formular modelos más complejos que modelicen la dependencia existente entre esas variables.

3.1.1 Generalizaciones del PPC

Una clase importante de procesos puntuales son los procesos de renovación, PR, generalización de los procesos de Poisson que se obtienen al eliminar la hipótesis de exponencialidad; es decir, son procesos puntuales en los que los tiempos de recurrencia son v.a. independientes e idénticamente distribuidas.

En los PR, y en los PP en particular, se supone que las ocurrencias son instantáneas; en la práctica, es frecuente que las ocurrencias no sean exactamente puntuales sino que tengan una duración asociada. Si esta duración es comparable a la de los intervalos de separación, puede ser más interesante utilizar modelos que representen la sucesión de estados alternantes del proceso.

Definición 3.2. *Un* **proceso de renovación alternante**, *PRA, representa un sistema que puede estar en dos estados, 1 y 2. El sistema permanece durante un tiempo $T1$ en el estado inicial 1, posteriormente pasa al estado 2, en el que permanece durante un tiempo $T2$, y así sucesivamente. El modelo requiere que la muestra del vector aleatorio $(T1_i, T2_i)_{i=1,\ldots n}$ sea independiente e idénticamente distribuida.*

Esta definición implica que las series $(T1_i)$ y $(T2_i)$ sean i.i.d., pero $T1_i$ y $T2_i$ pueden ser dependientes; es decir, la permanencia en el estado 1 no depende del pasado, pero la del estado 2 puede depender de lo ocurrido en el estado 1 anterior.

Un PRA compuesto, PRAC, es un modelo que se obtiene de forma análoga al PPC, definiendo en este caso dos vectores aleatorios de magnitudes que caracterizan los periodos en estado 1 y en estado 2, respectivamente. Las series de variables de cada uno de estos dos vectores deben ser i.i.d, pero puede existir dependencia entre las observaciones del mismo instante i, ya sea entre componentes del mismo vector o de los vectores asociados a distintos estados.

3.1.2 Proceso de los periodos secos. Justificación del modelo

En primer lugar hay que señalar que el proceso de los periodos secos observado es de carácter discreto, mientras que el proceso de Poisson es un proceso en tiempo continuo. La aproximación de un proceso estocástico en tiempo discreto tipo POT se apoya en la teoría asintótica de extremos de series de tiempo, revisada en el segundo capítulo y cuyo desarrollo se puede consultar en Leadbetter et al. (1983). Por otra parte, el proceso de los periodos secos no es estrictamente un proceso puntual puesto que las ocurrencias tienen asociada una duración; no obstante, esta aproximación es razonable, ya que la sequía es un fenómeno anómalo y las duraciones de los periodos secos serán mucho menores que las de los periodos no secos. En estas condiciones es necesario determinar a qué instante se asocia la ocurrencia de cada episodio. Se proponen tres alternativas, que permitirán verificar si las propiedades del proceso dependen de la definición utilizada:

- Punto medio: esta definición es la más natural desde el punto de vista físico.

- Punto inicial: esta localización tiene la ventaja, de interés en los modelos de predicción, de que su valor es conocido desde el comienzo del episodio.

- Punto de intensidad máxima: esta definición aporta aleatoriedad a la localización del punto de ocurrencia dentro del episodio, y permite la aplicación de la siguiente propiedad.

Propiedad 3.1. *Si el proceso de ocurrencia asociado a un umbral u es un* $PP(\lambda_u)$*, el proceso de ocurrencia asociado a un umbral más estricto,* $u_x > u$*, será también Poisson con tasa* $\lambda_{u_x} = \lambda_u[1 - F(u_x)]$*, siendo* $F(y) = P\,(\,X \leq y \mid X > u\,)$*.*

En efecto, consideremos el proceso de ocurrencia correspondiente a un umbral dado; tomando como punto de ocurrencia el de intensidad máxima, los procesos correspondientes a umbrales más estrictos se obtienen al eliminar los puntos cuya intensidad es menor que el nuevo umbral. El carácter aleatorio del borrado lo garantizará la hipótesis de independencia entre el vector de magnitudes, en particular la intensidad máxima, y el proceso de ocurrencia. Finalmente, como un proceso de borrado aleatorio conserva el carácter Poisson, propiedad 2.2, se obtiene el resultado.

Respecto al vector de magnitudes asociado a cada episodio, estará formado por las variables que describen la gravedad de una sequía: duración, déficit e intensidad máxima (L, D, IM).

El proceso de periodos secos se puede enmarcar en un modelo más general, un PRA, en el que los estados alternantes correspondan a periodos secos, que denotaremos ps, y periodos no secos, pns, según la serie de precipitación acumulada esté o no, por debajo del umbral de definición. Para caracterizar los ps se considera el vector de magnitudes (L, D, IM) y para los pns se definen tres magnitudes análogas: duración, exceso e intensidad máxima, $(Lns, Ens, IMns)$.

El resto del capítulo se dedica a describir el proceso de formulación y ajuste del modelo. El primer paso es definir el umbral que determinará la región crítica. Posteriormente se comprueba el carácter Poisson del proceso de ocurrencia de los periodos secos obtenidos con el umbral seleccionado y, finalmente, se analizan las series de magnitudes asociadas a esos episodios.

3.2 Selección del umbral

La elección del umbral debe basarse, en primer lugar, en criterios de tipo climático, ya que los periodos con observaciones inferiores a su valor deben corresponder a situaciones anómalas de escasez de precipitación. Como el concepto de periodo seco

no es preciso y existen distintos criterios de definición, la selección basada en ellos no proporcionará un único umbral válido sino un rango de valores posibles.

Por otra parte, se analizan las propiedades estadísticas de las series asociadas a los umbrales en ese rango: el carácter Poisson del proceso de ocurrencia y el ajuste a una distribución con buenas propiedades, como la PG, de los excesos sobre el umbral. En consecuencia, aparecen dos objetivos contradictorios entre los que se debe buscar un equilibrio:

- Definir umbrales lo suficientemente críticos para poder aplicar los resultados de la teoría de extremos de procesos estocásticos.

- Considerar valores del umbral que proporcionen un número razonable de observaciones; la variabilidad de las estimaciones será menor cuanto mayor sea el tamaño de muestra.

Generalmente, la metodología POT se ha aplicado al análisis de variables exceso, $X_u = X - u \mid (X \geq u)$; en este caso, la variable de interés no corresponde a un exceso sino a un déficit, $u - X \mid (X \leq u)$, pero el tratamiento es análogo y, por sencillez, hemos conservado la notación de exceso que es la habitual.

3.2.1 Criterios climáticos

Existe un gran número de criterios para definir el umbral que determina la ocurrencia de un periodo seco, criterios que dependen de las características climáticas de la zona en estudio y de la unidad temporal considerada. A continuación se citan algunos de los más usuales.

1. Porcentajes del valor medio

    - Baldwin (1941)define como sequía, en Australia, un periodo superior a dos meses consecutivos en el que se hayan registrado precipitaciones inferiores al 50% de la media.
    - Bates (1976) en un estudio para los EEUU utiliza el 75% del valor medio de la precipitación en el caso de datos anuales, y el 60% en los mensuales.

- Rossi (1983) en un análisis de las sequías en Sicilia, basado en datos anuales, utiliza el 70% del valor medio de un periodo de referencia, 1921-1978.
- Moyé et al. (1988), siguiendo el criterio propuesto por el *Texas Almanac*, definen el umbral en el 75% de la precipitación anual media asociada a un periodo de referencia de 30 años de duración.

2. Percentiles muestrales: Este tipo de criterios permite establecer de forma sencilla el equilibrio deseado entre los dos objetivos que se plantean -obtener sólo observaciones de carácter extremo y maximizar el tamaño de muestra- ya que el orden del percentil utilizado determina el número de observaciones por encima del umbral.
   - Gibbs & Maher (1967), en un estudio de las sequías en Australia, utilizan como umbral el percentil décimo de la precipitación anual, valor que elimina el 90% de las observaciones de la muestra.
   - Zelenhasic & Salvai (1987), en un análisis de los episodios de sequía de series de flujos fluviales, utilizan los percentiles 10 y 5 y comprueban que el método POT no es aplicable con umbrales menos estrictos, como los percentiles 20, 30 y 40.

   Otra posibilidad, equivalente al uso de percentiles, consiste en determinar directamente el tamaño de muestra, o el número aproximado de observaciones por periodo de tiempo, que se quiere obtener.
   - En un análisis de series diarias de flujos fluviales, y siguiendo las indicaciones del *Flood Studies Report*, Davison & Smith (1990) proponen utilizar umbrales tales que el número de excesos por año sea un valor entre 1 y 5.
   - Coles (1994) ajusta doce modelos mensuales a datos diarios de precipitación, utilizando umbrales tales que, en cada serie, el número de excesos en un periodo de 25 años sea 25 o 50.

3. Factor de frecuencia estándar: este criterio define umbrales de la forma $u = \bar{X} + kS_X$, con k constante, y es adecuado cuando la distribución de la variable es aproximadamente Normal; en otros casos puede proporcionar valores demasiado estrictos.

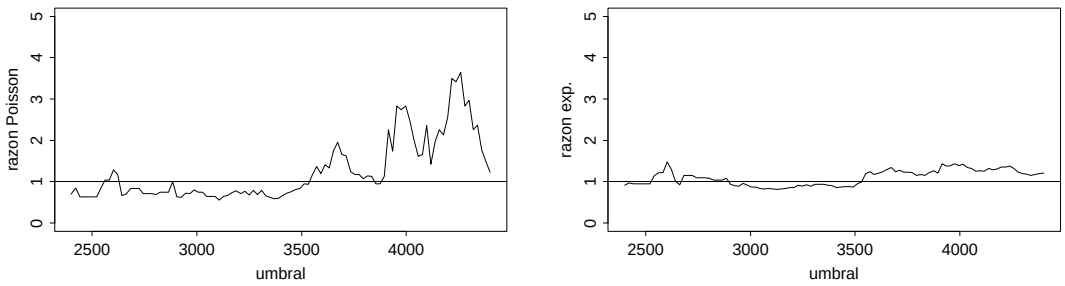

Figura 3.1: Gráficos de la razón Poisson (izda.) y de la razón Exponencial (dcha.); Daroca.

- Rasmussen & Rosbjerg (1991) utilizan como umbral de sequía en series de flujos fluviales un factor de frecuencia definido con $k = 3$, umbral que proporciona un número medio anual de excesos entre 1 y 4.5, según las características climáticas de la región.

3.2.2 Análisis del carácter Poisson en relación con el umbral

En este apartado se proponen distintos gráficos para analizar la evolución del proceso de ocurrencia al disminuir el umbral, con el fin de determinar el valor a partir del cual se puede suponer el carácter Poisson del proceso. Dentro del rango de umbrales climáticos posibles que presenten las propiedades estadísticas deseadas, se seleccionará el umbral menos estricto, con el objeto de obtener el mayor tamaño de muestra posible.

- Gráfico de la razón Poisson. Este procedimiento consiste en dividir el tiempo de observación del proceso en k intervalos iguales disjuntos y representar el cociente entre la media y la varianza muestrales de la variable $N(l)$, el número de ocurrencias en cada intervalo, en función del umbral. En un PPH la media y la varianza de $N(l)$ coinciden, por consiguiente, se seleccionará el umbral tal que, para todos los valores inferiores, el cociente se aproxime a 1, figura 3.1.

- Gráfico de la razón Exponencial. En este gráfico, análogo al anterior, se representa el cociente entre la media y la desviación típica de la muestra de tiempos de recurrencia, figura 3.1, que en un PPH debe tomar valores próximos a 1.

- Gráfico de Castro. Castro & Pérez-Abreu (1994) proponen un gráfico para confirmar el carácter Poisson del proceso asociado a un umbral dado basado

en el funcional generador de probabilidades, fgp, de un proceso puntual N durante un periodo de tiempo Υ, que se define como,

$$\Psi(g) = E\left[\prod_{i=1}^{N(\Upsilon)} g(t_i)\right] = E\left[\exp\left\{\int_{\Upsilon} \ln[g(t)]N(dt)\right\}\right],$$

donde g pertenece a la familia de funciones $G = \{g \mid 0 \leq g(s) \leq 1 \text{ y } g = 1 \text{ fuera de un conjunto compacto}\}$. A partir de $\Psi(g)$ se define el funcional,

$$Y_c(g) = \ln[\Psi(1 - cg)],$$

con g una función definida en Υ, tal que $1 - g \in G$, y c un valor constante $0 < c < 1$.

El fgp de un PP en Υ con función de intensidad $\lambda(t)$ es,

$$\Psi(g) = \exp\left\{-\int_{\Upsilon}[1 - g(t)]\lambda(dt)\right\}$$

y, en consecuencia,

$$Y_c(g) = -c\int_{\Upsilon} g(t)\lambda(dt)$$

es una función lineal de c para toda función g.

Una comprobación del carácter Poisson de un proceso consiste en verificar la linealidad de $Y_c(g)$ para distintas funciones g, figura 3.2. La función $Y_c(g)$ se debe estimar a partir de la muestra; un estimador insesgado y consistente de $\Psi(g)$ es el funcional generador de probabilidades empírico, fgpe,

$$\hat{\Psi}_n(g) = \frac{1}{n}\sum_{j=1}^{n}\prod_{i} g(t_i^j)$$

con $(t_i^j; j = 1, \ldots, n)$ los instantes de ocurrencia en n copias independientes del proceso N. Por consiguiente,

$$\hat{Y}_c(g) = \ln[\hat{\Psi}(1 - cg)].$$

Para realizar la comprobación se utilizarán un conjunto de funciones g representativas de G, por ejemplo:

$$g_1(t) = \frac{t^2}{|T|^2}$$

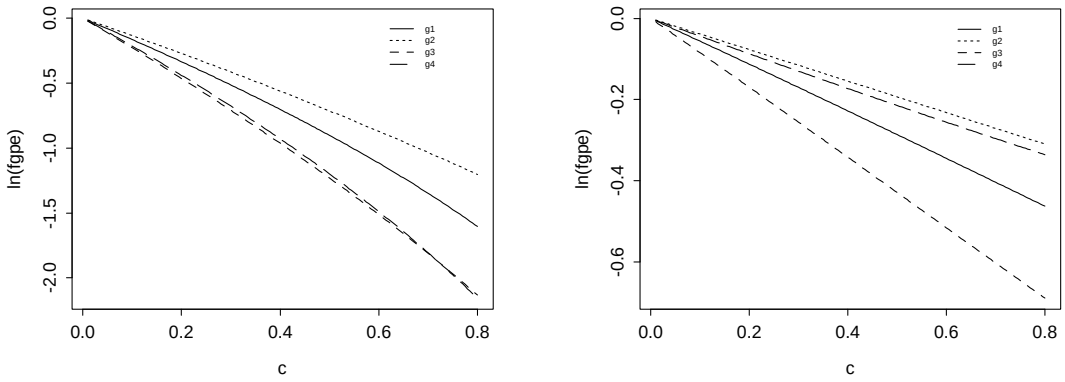

Figura 3.2: Gráficos de Castro para los umbrales 4760 (izda.) y 3310 (dcha.) con 18 intervalos de tiempo; San Fernando.

$$
\begin{aligned}
g_2(t) &= \frac{T^{1/2} - t^{1/2}}{T^{1/2}} \\
g_3(t) &= \begin{cases} 1 - (2/T)t & t < T/2 \\ (2/T)t - 1 & t \geq T/2 \end{cases} \\
g_4(t) &= \begin{cases} (2/T)t & t < T/2 \\ 2 - (2/T)t & t \geq T/2, \end{cases}
\end{aligned}
$$

Estas funciones, definidas en un dominio $\Upsilon = [0, T]$, presentan diferentes características, una es una función creciente, otra decreciente y las otras dos simétricas. Debido a la gran variabilidad de $Y_c(g)$ cuando c se aproxima a 1, el gráfico se limita al intervalo $0 < c < 0.8$. Estudios basados en simulación indican que, aunque el proceso sea Poisson, si el tamaño de muestra no es suficientemente grande, la función g_4 puede presentar un carácter convexo, mientras que el gráfico correspondiente a procesos no Poisson es, generalmente, cóncavo.

Si no se dispone de copias independientes del proceso y éste es homogéneo, se puede separar el periodo de observación en k intervalos disjuntos, y considerar las realizaciones en cada uno de ellos copias independientes.

3.2.3 Análisis de los excesos en función del umbral

Las distribuciones más utilizadas para modelizar excesos sobre umbrales extremos son las distribuciones Exponencial y PG, que contiene a la primera como caso par-

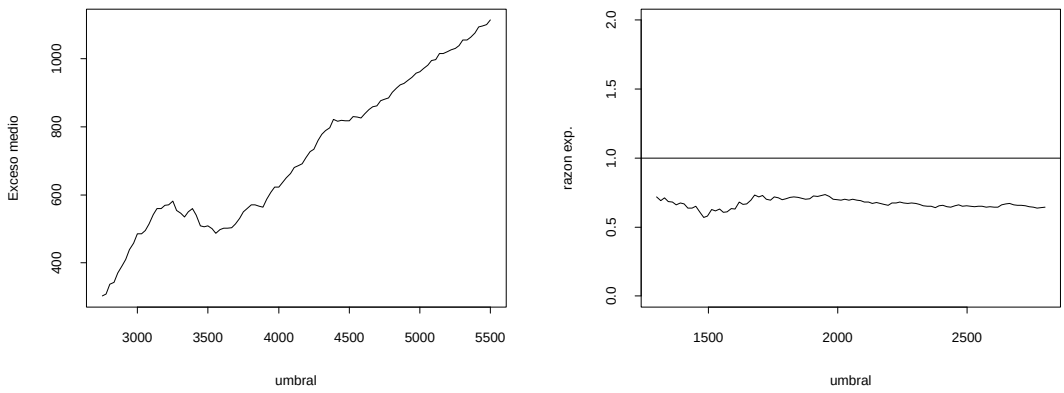

Figura 3.3: Gráficos del exceso medio y de exceso exponencial; San Fernando (izq.) y Murcia (dch.).

ticular. Los excesos de una variable PG también tienen distribución PG, propiedad que permite localizar el valor del umbral a partir del cual éstos se pueden considerar extremos, representando parámetros de la distribución cuyo valor sea constante o una función sencilla del umbral. Recordemos que en el análisis de episodios secos los umbrales más estrictos corresponden a los valores más pequeños.

- Gráfico del exceso medio. Davison & Smith (1990) aplican esta idea al valor esperado del exceso sobre un umbral u que, en el caso de una distribución PG, es una función lineal del umbral.

$$E(X - u \mid X \geq u) = \frac{\sigma - u\gamma}{1 + \gamma}.$$

  En los valores superiores, o inferiores en el caso de periodos secos, al umbral a partir del cual la distribución PG es adecuada, el exceso medio muestral,

$$y(u) = \sum_{Au} \frac{x_i - u}{|Au|}$$

  con $Au = \{i : x_i > u\}$, debe ser aproximadamente una función lineal de u, figura 3.3.

- Gráfico para excesos exponenciales. Se aplica la idea anterior representando el cociente entre la media y la desviación típica de las muestras correspondientes a umbrales crecientes, y teniendo en cuenta que el coeficiente de variación de una distribución Exponencial es siempre 1, figura 3.3. Análogamente se pueden representar estimaciones de la media o del parámetro α que, bajo la hipótesis de exponencialidad, deben ser constantes.

- Test de Gertensgarbe. Gertensgarbe & Werner (1989) proponen un test para distinguir la región de valores extremos de una variable, basado en la hipótesis de que las diferencias en la muestra ordenada, $y_i = x_{(i)} - x_{(i-1)}$, de las observaciones extremas presentan una estructura diferente a las restantes, de forma que el comienzo de la región de extremos corresponderá a un punto de cambio de esa serie. Para detectar este punto se utiliza una versión secuencial del test de Mann-Kendall, TMK. Puesto que el límite de la región de extremos es un valor alto, el test no se aplica a la serie original completa sino a la de los excesos sobre un umbral base, tal que el límite buscado sea un valor más estricto.

  El TMK se aplica en dos direcciones, del comienzo al final de la serie y de forma retrógrada. En el primer caso se define el estadístico,

$$U_i^* = \sum_{k=1}^{i} n_k$$

  para $i = 2, \ldots, n$, donde n_k es el número de elementos de la serie anteriores a y_k, inferiores a ese valor. El correspondiente estadístico estandarizado es,

$$U_i = \frac{U_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(2i+5)}{72}}}$$

  que, en ausencia de tendencia, tiene una distribución $N(0,1)$. El estadístico inverso estandarizado Up_i se obtiene de forma análoga a partir de la serie retrógrada y tiene la misma distribución; finalmente se calcula Up' invirtiendo el orden de la serie Up. El punto inicial de U y el final de Up' son siempre 0 y, si en la serie no hay empates, el valor inicial de Up' coincide con el final de U.

  El punto de intersección entre las series Up' y U, figura 3.4, corresponde a un cambio de tendencia en la serie ordenada, que será significativo si el valor de las curvas en ese punto excede el nivel de significación. Para facilitar la detección del punto de cambio proponemos complementar la información de este gráfico con dos gráficos auxiliares, el de la serie ordenada y el de las correspondientes diferencias y_i, figura 3.5.

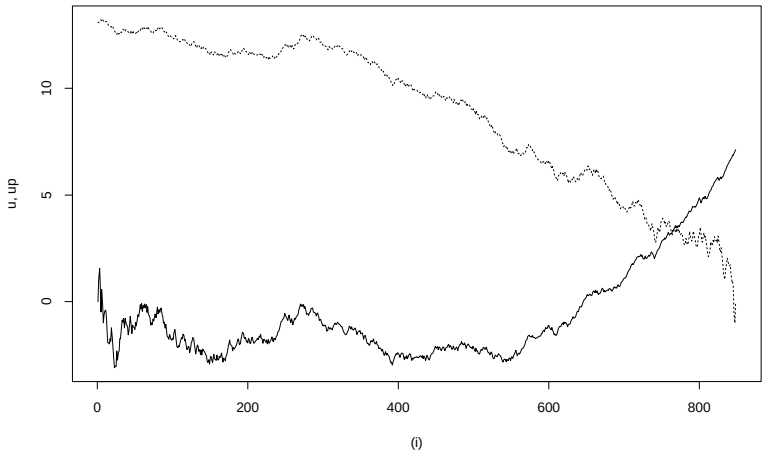

Figura 3.4: Gráfico de Gertensgarbe con un umbral base $u = 5400$; Burgos.

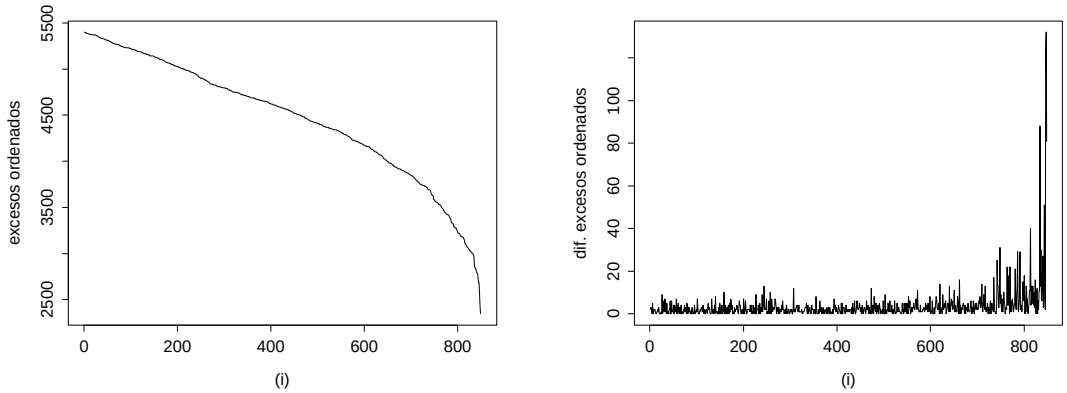

Figura 3.5: Gráficos de los excesos ordenados (izda.) y de las diferencias de los excesos ordenados (dcha.); Burgos.

3.2.4 Análisis de resultados

El primer paso del proceso de selección es calcular un rango de posibles umbrales atendiendo a criterios climáticos, porcentajes 65, 70, 75, 80 y 85 de la precipitación media y percentiles 5, 10, 15, 20 y 25; en la figura 3.6 se representa un fragmento de la serie de precipitación de Huesca utilizando como umbral los distintos percentiles señalados.

Posteriormente se realiza un análisis gráfico, en ese rango de valores, del carácter Poisson del proceso de ocurrencia y del carácter extremo de los excesos, utilizando las herramientas descritas. En los análisis que requieren la división del periodo de

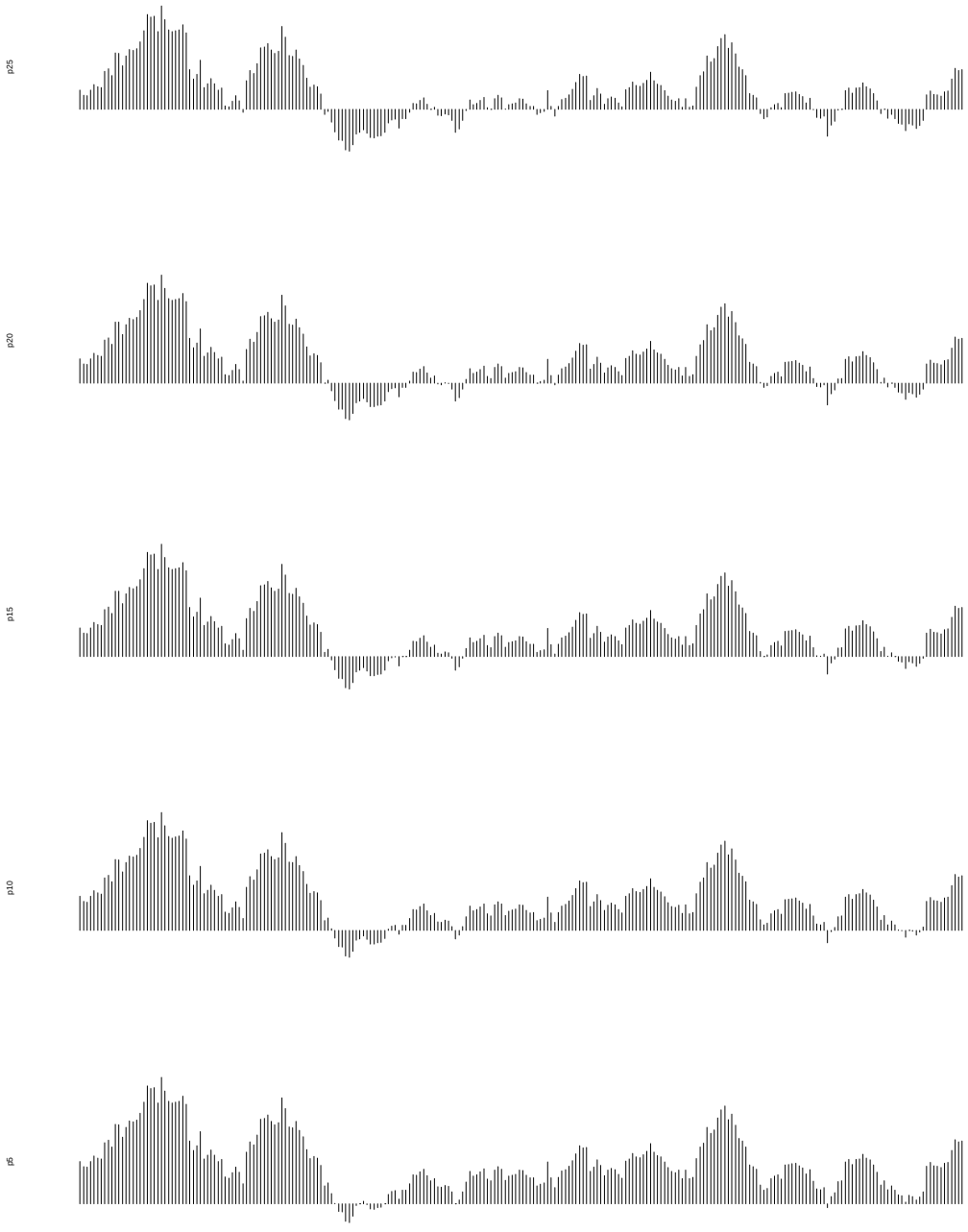

Figura 3.6: Series de precipitación respecto a umbrales definidos con distintos percentiles, correspondientes a las observaciones 300-550; Huesca.

	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Percentil 5	3470	2830	3380	2750	1350	3310
Percentil 10	3880	3060	3740	3010	1630	3630
Percentil 25	4520	3600	4400	3550	2100	4560
Percentil 30	4670	3750	4600	3680	2250	4760
70% media	3790	3000	3810	3010	2000	3940
60% media	3240	2570	3260	2580	1710	3380
C. Poisson	4500	3600	4400	3500	2200	4500
C. extremo	3500	2800	3500	2800	1500	3300
C. PG	3600	2800	3500	2800	1500	3300

Tabla 3.1: Resultados del análisis de selección del umbral.

observación en intervalos se consideran tres particiones, con 12, 18 y 25 intervalos en las cuatro series más largas y con 10, 15 y 20 en Murcia y Daroca. En la tabla 3.1 se resumen los principales resultados de cada observatorio, de los que se deducen las siguientes conclusiones.

- El carácter Poisson del proceso de ocurrencia de los periodos secos se verifica en todos los observatorios a partir de umbrales próximos al percentil 25. El gráfico de Castro confirma este resultado.
- El carácter extremo de las observaciones se manifiesta en umbrales más estrictos, con valores que varían entre los percentiles 10 y 5 de la muestra.
- El carácter PG de la distribución de los excesos requiere umbrales que oscilan, dependiendo de los observatorios, entre los percentiles 5 y 10; estos umbrales coinciden con los de comportamiento extremo.
- No existe ningún umbral cuya serie de excesos presente un comportamiento exponencial.
- El percentil décimo, criterio de Gibbs & Maher (1967), es un valor muy próximo al 70% de la media, criterio de Rossi (1983), excepto en el caso de Murcia y San Fernando.

En conclusión, un umbral aparentemente adecuado para definir los periodos secos, ya que verifica todas las condiciones necesarias, tanto de tipo climático como estadístico, es el percentil 10 de la precipitación anual.

En el aspecto metodológico se señalan las siguiente observaciones.

- Los gráficos de razón exponencial y razón Poisson proporcionan, en general, resultados coincidentes; el gráfico de la razón Poisson es bastante estable al número de intervalos en que se divide el tiempo de observación del proceso.
- El gráfico de Castro permite confirmar la selección del valor del umbral. Este gráfico es más sensible a la partición en intervalos del periodo de observación; cuando el número de intervalos es pequeño, las funciones presentan un aspecto convexo que tiende a desaparecer al considerar un número mayor.
- En ocasiones el gráfico de Gertensgarbe sólo detecta el cambio de tendencia más fuerte de la serie ordenada que, en este tipo de datos, corresponde a un umbral demasiado estricto. Los gráficos auxiliares propuestos muestran la estructura de la serie y permiten localizar otros cambios de tendencia existentes, que también son significativos.
- En la interpretación de los gráficos, en especial en aquéllos en los que se representa una función del umbral, se debe tener en cuenta que el tamaño de muestra correspondiente a los umbrales muy estrictos es pequeño y que, por consiguiente, tienen asociada una gran variabilidad.

3.3 Proceso de ocurrencia. Comprobación de las hipótesis de un PP

De acuerdo con la propiedad 3.1 y los resultados del apartado anterior, los procesos de periodos secos definidos con umbrales inferiores al percentil 25 deben presentar carácter Poisson. Para comprobar esta afirmación se analiza el proceso asociado a ese umbral, que denotaremos ps25, y los correspondientes a cuatro valores, aproximadamente equidistantes, entre el percentil 25 y el 10; el proceso definido con este umbral, ps10, se analiza con mayor detalle. En primer lugar se describen las herramientas y la metodología utilizadas, que se basan en las caracterizaciones de un PP expuestas en el segundo capítulo.

Los detalles relativos a los tests citados en esta sección se pueden consultar en Kendall & Gibbons (1990), KG, Gibbons & Chakraborti (1992), GC, Daniel (1990),

D, y Ansell & Phillips (1994), AP.

3.3.1 Número de ocurrencias en un intervalo

En un $PPH(\lambda)$ el número $N(l)$ de ocurrencias en un intervalo de tiempo de longitud l tiene una distribución $P(\lambda l)$. La comprobación de la distribución de esta variable se realiza con los siguientes controles.

- Valor de la razón entre media y varianza, que bajo la hipótesis de distribución Poisson es igual a 1.

- Ajuste máximo verosímil de una distribución Poisson a $N(l)$. La bondad de este ajuste se comprueba con un análisis gráfico -qqplot y comparación de las distribuciones observada y teórica mediante gráficos de las funciones de probabilidad y distribución- y con los tests χ^2 y Kolmogorov-Smirnov, KS.

3.3.2 Tiempos de recurrencia

1. Carácter Exponencial El carácter exponencial de los tiempos de recurrencia se comprueba con las siguientes herramientas.

- Valor del coeficiente de variación.

- Ajuste máximo verosímil de una distribución Exponencial a los tiempos de recurrencia. La bondad del ajuste se comprueba con las herramientas indicadas para la distribución Poisson; en este caso, la función de densidad se compara con una estimación no paramétrica basada en el núcleo de Epanechnikov.

2. Independencia La mayoría de las comprobaciones de independencia de una serie se basan en el análisis de la hipótesis $\rho = 0$, aunque en series no normales, como las analizadas, estas hipótesis no son exactamente equivalentes.

2.a. Autocorrelación La existencia de autocorrelación se verifica con los siguientes tests y analizando los correlogramas correspondientes.

- Test de Kendall (KG). La versión habitual de este test no admite corrección de continuidad, por lo que se ha utilizado un estadístico equivalente, función de $S = \tau n(n-1)/2$,

$$t_K = \frac{S}{\sqrt{\frac{n(n-1)(2n+5)}{18}}},$$

con la siguiente corrección: sumar, o restar, 1 a S si $\tau < 0$, o $\tau > 0$. Bajo la hipótesis nula, $S = 0$, el estadístico tiene una distribución asintótica $N(0,1)$. Se utiliza una estimación de la varianza específica para muestras en las que pueden existir empates.

- Test de Spearman (KG). La aproximación Normal del estadístico de Spearman usual sólo es válida con muestras de tamaño $n > 35$, por lo que se ha utilizado una modificación,

$$t_S = \frac{\rho_s\sqrt{n-2}}{\sqrt{1-\rho_s^2}},$$

válida con $n > 10$, que bajo la hipótesis $\rho_s = 0$ tiene una distribución asintótica t_{n-2}; este estadístico no permite el cálculo de intervalos de confianza.

2.b. Aleatoriedad Esta hipótesis se analiza con los siguientes tests:

- Test de rachas (GC). Este test se basa en el número de rachas R de la serie de signos, el signo es positivo si la observación es mayor que un cierto valor de referencia y negativo si es menor. El estadístico

$$t_R = \frac{R - (1 + 2n_1n_2/n)}{\frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)}},$$

con $n = n_1 + n_2$ y n_1, n_2 el número de observaciones + y $-$, respectivamente, tiene una distribución asintótica $N(0,1)$, si $n_1 > 12$ y $n_2 > 12$. Los valores de referencia utilizados son la media y la mediana. Se aplica la corrección de continuidad habitual.

- Tests de rachas crecientes y decrecientes (GC). Este test, basado en la misma idea que el test de los puntos de cambio (KG), contrasta el carácter aleatorio de una serie controlando el número de rachas; una racha se define como un conjunto de observaciones consecutivas crecientes o decrecientes. El estadístico,

$$t_{Rcd} = \frac{R - (2n-1)/3}{\sqrt{(16n-29)/90}}$$

tiene una distribución asintótica $N(0,1)$, válida para $n > 25$. Se aplica la corrección de continuidad habitual.

3. Homogeneidad La homogeneidad de un PP requiere que los tiempos de recurrencia sean idénticamente distribuidos, propiedad que implica ausencia de tendencia y de estacionalidad.

3.a. Tendencia

- Tests de tendencia temporal. Se contrasta la existencia de correlación entre los valores de una variable y los tiempos de observación correspondientes utilizando los tests de Kendall y Spearman.

- Test de rangos de von Neumann (GC). Este test se basa en el estadístico,

$$t_{VN} = \frac{\sum_{i=1}^{n-1} [rang(x_i) - rang(x_{i+1})]^2 - 2}{\sum_{i=1}^{n} \left(rang(x_i) - \frac{n+1}{2}\right)^2} \; \frac{5n(n+1)(n-1)^2}{4(n-2)(5n^2-2n-9)},$$

  que es una transformación del coeficiente de correlación de rangos definido por Wald & Wolfowitz (1943), y tiene una distribución asintótica $N(0,1)$.

- Test del signo de la diferencia (KG). Este test evalúa la posible tendencia en términos del número de observaciones crecientes de la muestra; se dice que una observación es creciente si el signo de su diferencia con la observación anterior es positivo.

- Test de Cox-Stuart (D). Este test se basa en el estadístico $t_{CS} = \min(N_+, N_-)$, donde N_+ y N_- son, respectivamente, el número de pares (x_i, x_{i+c}) crecientes y decrecientes, siendo c el punto medio de la serie; la distribución asintótica del estadístico es $N(0,1)$.

- Controles de tendencia específicos para procesos de Poisson y de renovación. Denotaremos (t_0, t_f) al tiempo de observación del proceso y $t_1, \ldots, t_n$ a los instantes de ocurrencia observados en ese periodo.

  - Test del *Military Handbook*, TMH. Contrasta la ausencia de tendencia frente a la existencia de una tendencia de carácter potencial; bajo la

hipótesis nula el estadístico,

$$t_{MH} = -2\sum_{i=1}^{n} \ln\left(\frac{t_i}{t_f}\right),$$

tiene una distribución aproximada χ^2_{2n}.

- Test de Laplace, TL. Plantea como alternativa la existencia de una tendencia de carácter loglineal; si la tendencia es nula, el estadístico,

$$t_L = \frac{\sum_{i=1}^{n} t_i/t_f - n/2}{\sqrt{n/12)}},$$

tiene una distribución aproximada $N(0,1)$.

- Test de Lewis-Robinson, TLR. Los dos tests anteriores pueden sugerir, erróneamente, la existencia de tendencia si el proceso subyacente no es un PP. Este test ajusta el estadístico de Laplace para distribuciones con coeficiente de variación distinto de 1,

$$t_{LR} = t_L \frac{\bar{t}}{s_t},$$

por lo que se puede aplicar a cualquier proceso de renovación; en ausencia de tendencia su distribución asintótica es $N(0,1)$.

- Gráfico del número acumulado de ocurrencias. En un PPH la esperanza del número de ocurrencias en un intervalo $(0,t)$, $E[N(t)]$, debe ser una función lineal del tiempo, y cualquier desviación de la linealidad indicará que la tasa del proceso no es constante.

- Gráfico de los tiempos de recurrencia frente al tiempo. Permite detectar la existencia de ciclos o tendencias en la serie que invalidarían la hipótesis de idéntica distribución requerida por un PRH.

3.b. Estacionalidad La posible falta de homogeneidad de la distribución de las ocurrencias a lo largo del año, provocada por un comportamiento estacional, se comprueba con los siguientes controles.

- Representación gráfica de las frecuencias absolutas y relativas de ocurrencia mensuales.

- Test χ^2 de homogeneidad.

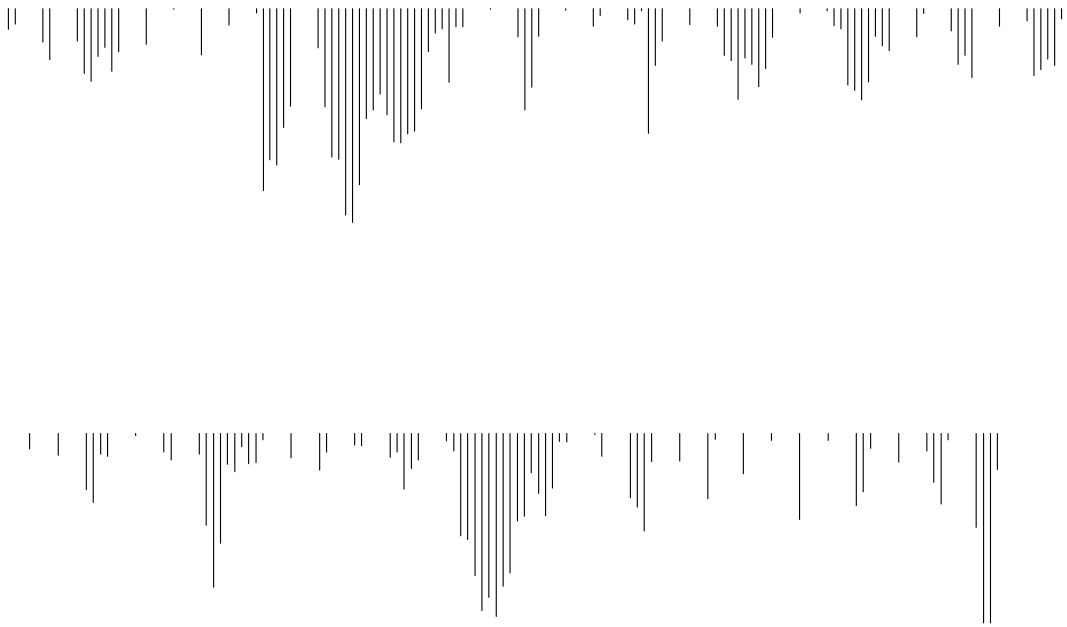

Figura 3.7: Periodos secos definidos con el percentil 10 observados en la serie; Huesca.

3.3.3 Análisis de resultados

A partir de la serie de periodos secos de cada observatorio, como la que se muestra en la figura 3.7, se construyen tres series localizando la ocurrencia en el punto inicial, punto medio, y punto de intensidad máxima respectivamente, y con cada una de ellas se realizan las siguientes comprobaciones:

- Ajuste Poisson del número de ocurrencias en los intervalos obtenidos al dividir el periodo de observación en 12, 18 y 25 intervalos en las series largas y en 10, 15 y 20 en las series de Murcia y Daroca.

- Ajuste Exponencial y controles de independencia y homogeneidad de las series de los tiempos de recurrencia. Se representan los correlogramas hasta orden 12. Los tests de homogeneidad se aplican mensual o bimestralmente, dependiendo del tamaño de muestra.

En las figuras 3.8 y 3.9 se representan las series de los tiempos de recurrencia, con localización en el punto medio, de los periodos secos definidos con el percentil 25 y el percentil 10, respectivamente, de los seis observatorios analizados. Un resumen

	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Umbral	4520	3600	4400	3550	2100	4560
n	77	48	75	93	40	78
Poisson (χ^2)	0.808	0.601	0.076	0.483	0.163	0.213
Exp. (K-S)	0.269	>0.2	0.228	0.173	0.180	>0.2
λ	(0.055)	(0.052)	(0.050)	(0.066)	(0.044)	(0.054)
Indep. T_r	0.624	0.123	0.434	0.635	0.121	0.089
τ Kendall	(0.039)	(-0.157)	(0.063)	(-0.034)	(0.176)	(0.133)
Rachas cre-dec.	0.084	0.882	0.407	0.901	0.309	0.786
Von Neumann	0.609	0.134	0.294	0.800	0.054	0.056
Cox-Stuart	0.188	0.118	0.617	0.883	0.022	0.511
Lewis-Robinson	0.027	0.331	0.273	0.396	0.699	0.459
Homogeneidad	0.239	0.101	0.326	0.536	0.416	0.003

	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Umbral	3880	3060	3740	3010	1630	3630
n	42	37	45	54	26	40
Poisson (χ^2)	0.523	0.132	0.342	0.507	0.163	0.763
Exp. (K-S)	0.301	0.132	0.126	0.129	0.345	0.245
λ	(0.028)	(0.039)	(0.029)	(0.036)	(0.027)	(0.026)
Indep. T_r	0.815	0.010	0.271	0.261	0.214	0.632
τ Kendall	(-0.027)	(-0.306)	(-0.117)	(-0.108)	(-0.185)	(-0.055)
Rachas cre-dec.	0.710	0.034	0.720	0.744	0.750	0.373
Von Neumann	0.851	0.017	0.369	0.240	0.197	0.724
Cox-Stuart	0.502	0.999	0.663	0.845	0.999	0.999
Lewis-Robinson	0.376	0.465	0.105	0.925	0.543	0.418
Homogeneidad	0.634	0.457	0.109	0.999	0.759	0.715

Tabla 3.2: P-valores de los tests del análisis ps25 (sup.) y ps10 (inf.) en la localización punto medio. Número de intervalos en el control Poisson: 15 en Murcia y Daroca, y 18 en el resto.

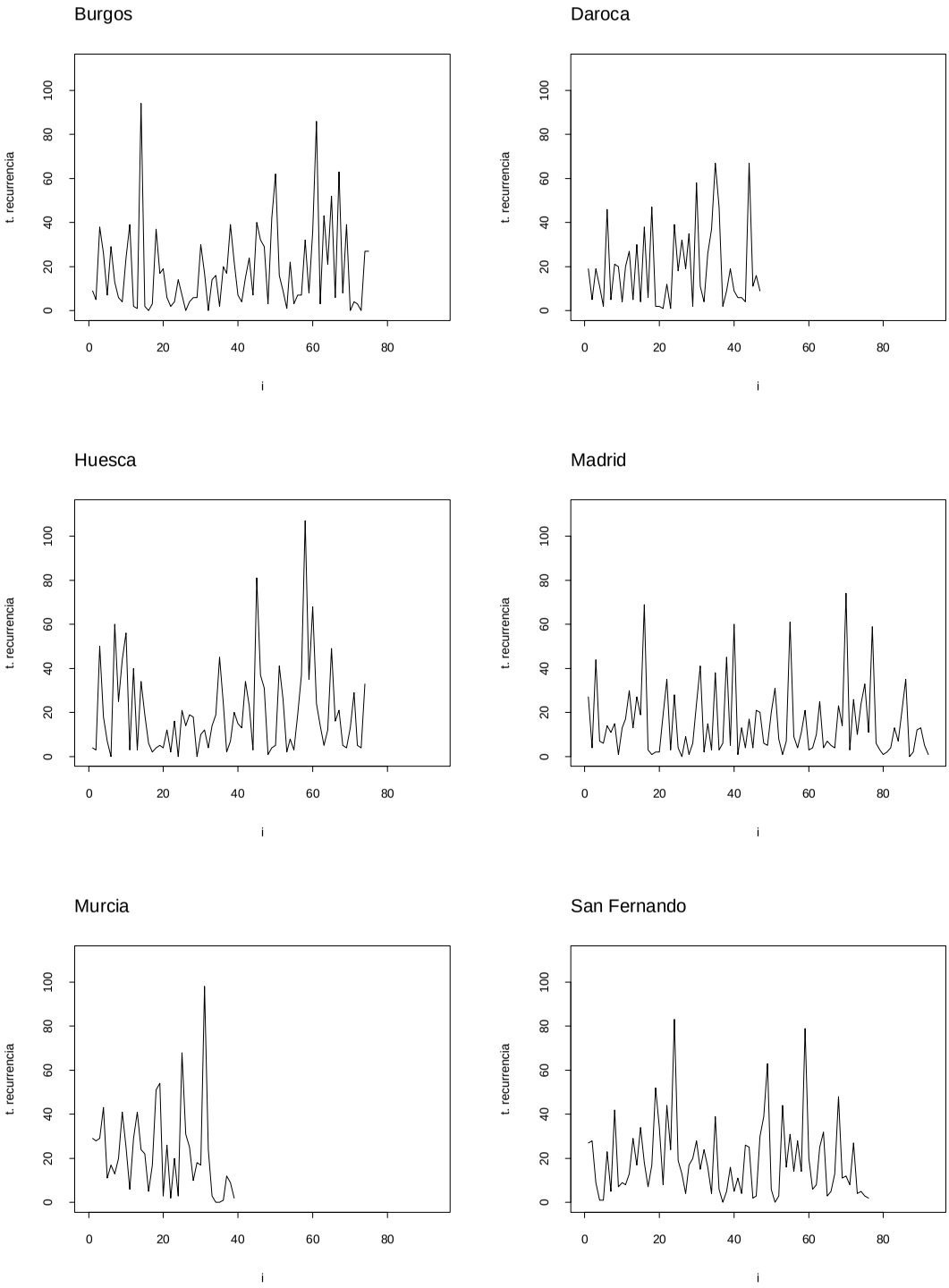

Figura 3.8: Series de tiempos de recurrencia con localización en el punto medio, ps25.

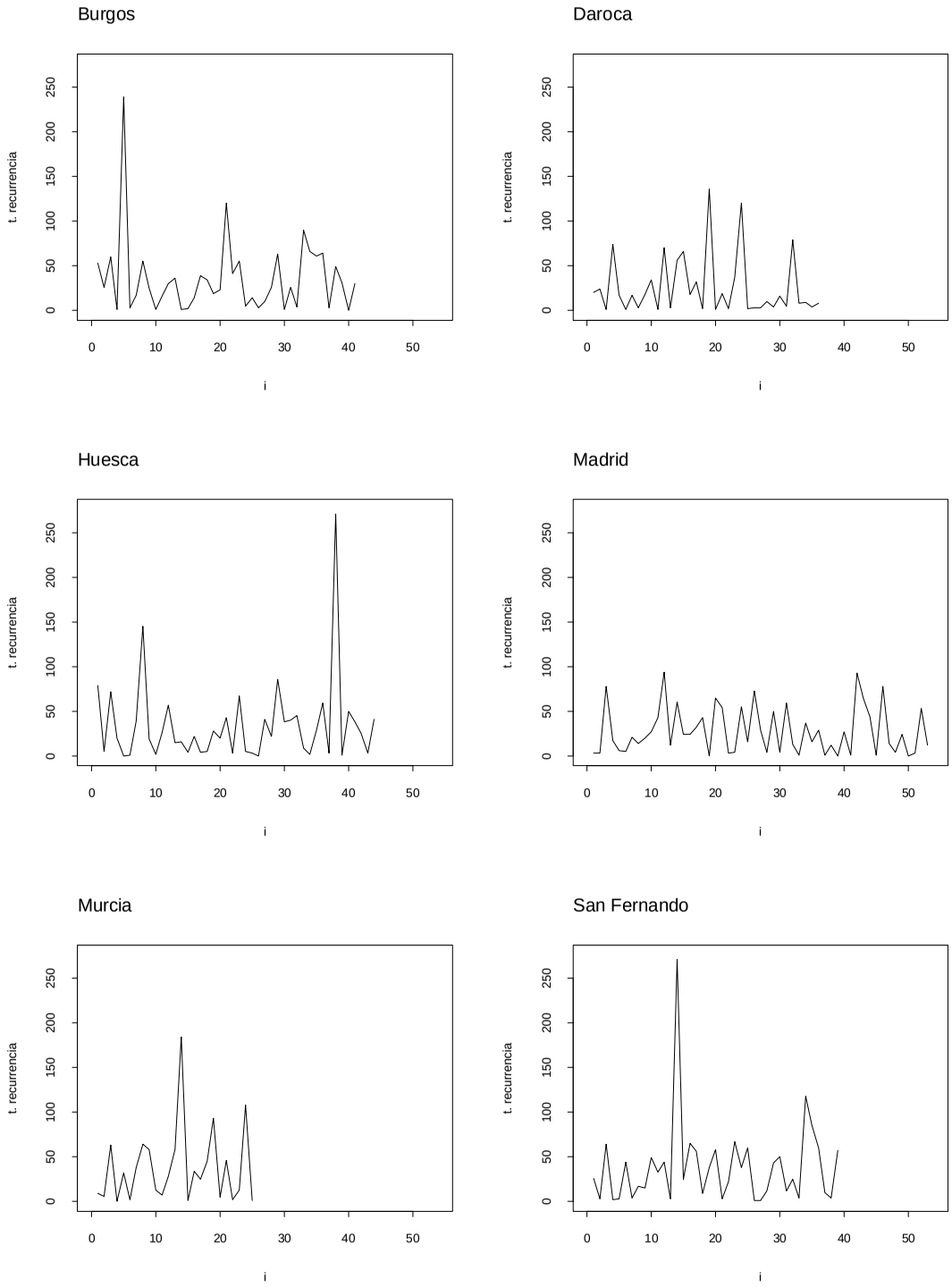

Figura 3.9: Series de tiempos de recurrencia con localización en el punto medio, ps10.

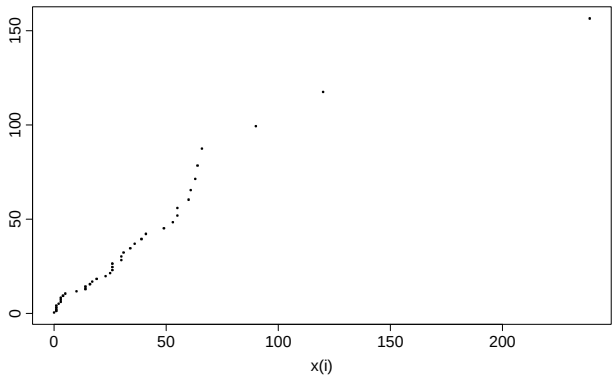

Figura 3.10: QQplot exponencial de los tiempos de ocurrencia de ps10 con localización en el punto medio; Burgos.

de los resultados obtenidos en el análisis de estas series se muestra en la tabla 3.2. A continuación, se señalan los aspectos de interés de cada observatorio.

- **Burgos.** En ps25 hay un aviso aislado de tendencia en el test de Lewis-Robinson que desaparece con umbrales más estrictos. En ps10 se observa un cambio en el comportamiento de la serie de tiempos de recurrencia en torno a la observación número 20, figura 3.9. Este cambio se detecta también en el qqplot, figura 3.10, en el que se observa una desviación de la linealidad en la cola superior de la muestra, en las observaciones que corresponden a los mayores valores posteriores al cambio observado en la muestra.

- **Daroca.** Algunos p-valores de los tests de tendencia en ps25 son bajos pero no significativos. El carácter aleatorio es satisfactorio en todos los umbrales, excepto en ps10. Este análisis presenta, en la serie con localización en el punto medio, p-valores menores que 0.05 en algunos tests de aleatoriedad y en la correlación de orden 1; estos avisos no aparecen en otras localizaciones del punto de ocurrencia ni en umbrales menos estrictos, por lo que no se consideran consistentes.

- **Huesca.** En el análisis ps25 el p-valor del test de rachas crecientes y decrecientes en la localización punto inicial es inferior a 0.05; este efecto es aislado y desaparece en umbrales más estrictos.

- **Madrid.** La serie de Madrid presenta indicios de estacionalidad en la localización punto inicial; este comportamiento es significativo en ps25, pero se debilita

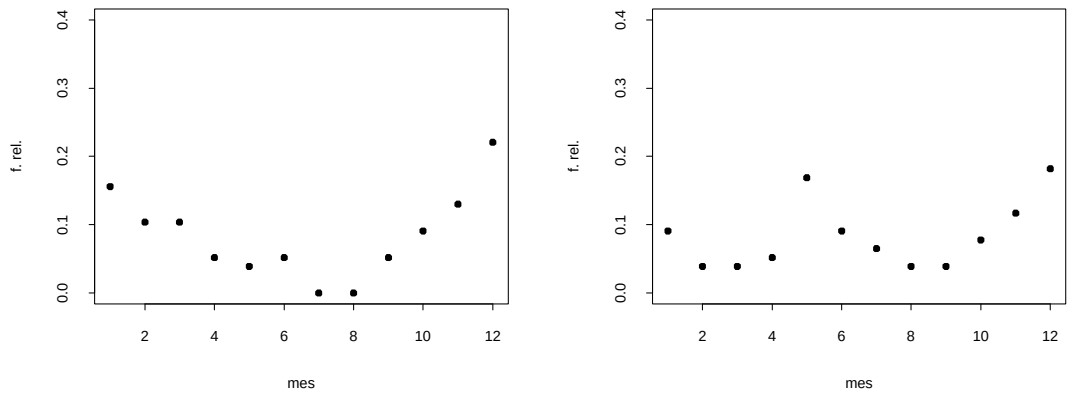

Figura 3.11: Probabilidades mensuales de ocurrencia con localización en el punto inicial (izda.) y en el punto medio (dcha.), ps25; San Fernando.

al considerar umbrales más estrictos, de forma que en ps10 los p-valores de los tests de homogeneidad son bajos pero no graves.

- **Murcia.** En ps25 la serie presenta indicios de una tendencia monótona creciente, que se manifiesta en los tests de tendencia y en una correlación positiva de orden 1 significativa, en algunas localizaciones. Esta tendencia es inapreciable en ps10.

- **San Fernando.** Los tests de correlación de orden 1 y algunos tests de tendencia, rachas y von Neumann, de ps25 son bajos, aunque gráficamente no se aprecia una tendencia monótona, sino la existencia de dos ciclos; este efecto desaparece en los otros umbrales. El principal problema de esta serie es la falta de homogeneidad a lo largo de año. Los tests de homogeneidad mensuales y bimestrales de ps25 son claramente significativos, especialmente los correspondientes a la localización punto inicial; este efecto se aprecia también en las frecuencias de ocurrencia, que son prácticamente nulas en los meses de verano, figura 3.11. De nuevo, el carácter estacional se suaviza en los umbrales más estrictos, de forma que los tests no son significativos en ps10; la mejoría es más lenta en la serie del punto inicial.

  El origen de este carácter estacional se encuentra en la desigualdad de las aportaciones de cada mes a la precipitación anual, que implica que la aparición de un determinado tipo de mes muy seco, puede ser más decisiva en la ocurrencia de una sequía que la de otros, provocando un efecto estacional en

la ocurrencia. En el caso de San Fernando, es frecuente que la precipitación de los meses de verano sea cero, de forma que la aparición de una observación anual muy baja no estará nunca originada por la aportación de esos meses y, en consecuencia, la probabilidad de que un episodio seco comience en esa época es prácticamente nula.

De los análisis realizados se extraen las siguientes conclusiones.

- El carácter exponencial de los tiempos de recurrencia es aceptable en el proceso ps25; este carácter se mantiene con umbrales más estrictos.
- El carácter aleatorio en general, y la ausencia de estructura de correlación en particular, no son plausibles en el proceso ps25, pero mejoran notablemente al considerar umbrales más estrictos; en ps10 no se detectan indicios de dependencia.
- No se detecta la existencia de tendencias monótonas que pudieran ser reflejo de un cambio climático. Sólo en ps25 en Murcia y Daroca existen algunos indicios, pero las series son demasiado cortas para poder realizar una afirmación concluyente, ya que también podrían corresponder a la fase creciente de un ciclo, como los que se han detectado en otros observatorios.

  La existencia de tendencia en una serie puede no trasmitirse a sus datos extremos por lo que, como sucede en Murcia y Daroca, una tendencia apreciable en el proceso correspondiente a un umbral puede desaparecer en umbrales más estrictos.
- El carácter estacional de las series está ligado a las características climáticas de la zona; sólo es significativo en el análisis ps25 de San Fernando y, en menor medida, en el de Madrid. El carácter estacional es más marcado en los procesos definidos con umbrales menos estrictos.

El proceso ps10 se puede suponer un PPH en todos los observatorios. Umbrales menos severos no siempre permiten suponer la homogeneidad del proceso, en Murcia y Daroca debido a la existencia de tendencia y en San Fernando y Madrid debido al carácter estacional.

Respecto a los aspectos metodológicos se concluye:

- Las tres series de localización de la ocurrencia utilizadas -punto medio, punto inicial y de intensidad máxima- proporcionan resultados similares, excepto en los controles de homogeneidad; la localización punto inicial presenta un carácter estacional más marcado.

- Los procedimientos para comprobar el carácter Poisson del número de ocurrencias funcionan mejor cuando el número de intervalos en que se divide el tiempo de observación no es muy pequeño. La comprobación del carácter exponencial de los tiempos de recurrencia presenta la ventaja de que no requiere ningún valor elegido de forma subjetiva.

- El test de rachas para contrastar la aleatoriedad de una serie definido con la media como valor de referencia es muy sensible a la existencia de datos anómalos. En presencia de este tipo de datos, la media toma valores demasiado grandes y el número de rachas correspondiente puede ser pequeño indicando, de forma errónea, la falta de aleatoriedad. Utilizando como referencia la mediana se evitan estos problemas.

- Los tests de von Neumann y los tests de tendencia temporales son los más eficientes para detectar la existencia de tendencias de tipo monótono, mientras que los tests basados en rachas son más útiles para detectar la existencia de ciclos o tendencias no monótonas.

Comparación de los tiempos de recurrencia de los distintos observatorios
En la tabla 3.3 se muestran los resultados de las comparaciones de las series de tiempos de recurrencia de los distintos observatorios, con el objeto de establecer las coincidencias temporales de los mayores valores observados en cada una; una coincidencia significa un periodo no seco común. Se muestran, en orden cronológico, los años inicial y final de algunos de los 10 mayores tiempos de recurrencia observados en cada serie en ps25 y ps10, respectivamente. Entre paréntesis se indica la posición de esa observación en la serie del observatorio correspondiente, ordenada en sentido decreciente; sólo se muestran los periodos que coinciden, aproximadamente, en tres o más observatorios.

Periodo	Huesca	Burgos	Daroca	Murcia	Madrid	S. Fernando
P25(1)	1882-87	1882-90	-	-	1884-90	1883-87
	(5)	(1)	-	-	(2)	(4)
P25(2)	-	-	1918-22	1922-25	1919-23	1919-21
	-	-	(6)	(9)	(4)	(9)
P25(3)	1932-38	-	1935-38	1931-35	-	1936-38
	(2)	-	(9)	(6)	-	(10)
P25(4)	-	1939-43	1939-43	1942-45	1939-44	1939-43
	-	(5)	(5)	(5)	(5)	(3)
P25(5)	1959-67	1958-64	1958-63	1956-61	1958-65	1958-65
	(1)	(2)	(3)	(3)	(1)	(2)
P25(6)	1970-76	1971-75	1971-76	1971-78	-	-
	(3)	(4)	(2)	(2)	-	-
P25(7)	-	1976-81	1976-80	-	1975-80	1976-80
	-	(3)	(4)	-	(3)	(5)
P25(8)	1982-85	1982-85	1986-92	1985-94	-	-
	(9)	(10)	(1)	(1)	-	-

Periodo	Huesca	Burgos	Daroca	Murcia	Madrid	S. Fernando
P10(1)	1882-93	1876-96	-	-	1882-90	1882-905
	(2)	(1)	-	-	(2)	(1)
P10(2)	1919-24	-	1917-23	1914-19	1919-23	-
	(5)	-	(4)	(4)	(10)	-
P10(3)	1931-38	1925-35	1932-38	1931-36	1925-31	1929-35
	(3)	(2)	(5)	(6)	(5)	(5)
P10(4)	1958-81	1958-65	1954-65	1946-61	1957-65	1957-67
	(1)	(3)	(1)	(1)	(1)	(2)
P10(5)	-	1976-81	1971-81	1971-78	1975-81	1975-80
	-	(4)	(2)	(3)	(3)	(7)
P10(6)	-	-	1986-92	1985-94	-	-
	-	-	(3)	(2)	-	-

Tabla 3.3: Comparación de los mayores tiempos de recurrencia, ps25(sup.) y ps10 (inf.).

Los episodios no secos de carácter general detectados en ps25 son: P25(1) -Daroca y Murcia no tienen registro en esa época-, P25(4), y P25(5). Los periodos P25(1) y P25(5) se siguen detectando de forma general en ps10, donde corresponden a P10(1) y P10(4); en este umbral también son comunes a la mayor parte de los observatorios P10(3) y P10(5). Se pueden señalar otras coincidencias más locales:

- Los tres episodios de mayor magnitud de los observatorios de Murcia y Daroca son los mismos en ps25 y ps10.

- Los observatorios de Burgos y Huesca muestran una estructura similar, especialmente en ps25, que se aprecia en los gráficos correspondientes de la figura

3.8.

- En Burgos y Madrid coinciden cuatro de los cinco mayores tiempos de recurrencia en ambas series.

3.4 Análisis de las series de magnitudes

Una vez comprobado que el proceso de ocurrencia definido con el percentil 10 es un PPH, la descripción del proceso de los periodos secos se completa con un análisis de las series del vector de magnitudes asociado a cada episodio. En un PPC la muestra de este vector debe ser independiente e idénticamente distribuida e independiente del proceso de ocurrencia. Se analiza también el vector de magnitudes asociado a los periodos no secos, aunque el modelo propuesto no impone ninguna hipótesis sobre estas variables, y se estudia la relación existente entre los periodos secos y los no secos adyacentes. Estos análisis serían imprescindibles en un PRAC.

Independencia La independencia se comprueba con los controles descritos en el apartado 3.3. Los datos analizados están calculados a partir de sumas móviles anuales, por consiguiente, las observaciones correspondientes a episodios muy próximos, a una distancia menor que doce, comparten información, y es probable que no se pueda suponer la independencia de las series de magnitudes.

Homogeneidad Para analizar la existencia de tendencia se utilizan las mismas herramientas que en el apartado anterior, excepto los tests específicos para procesos de renovación y de Poisson. La comprobación del carácter estacional asociado al instante de ocurrencia en el año, se basa en:

- Análisis descriptivo de la variable en periodos homogéneos, meses o grupos de meses.

- Test de Kruskal-Wallis para contrastar la igualdad de medianas de k grupos. Este test supone que la dispersión es la misma en todos los grupos, hipótesis que no está garantizada en este tipo de datos, por lo que sus resultados sólo son orientativos.

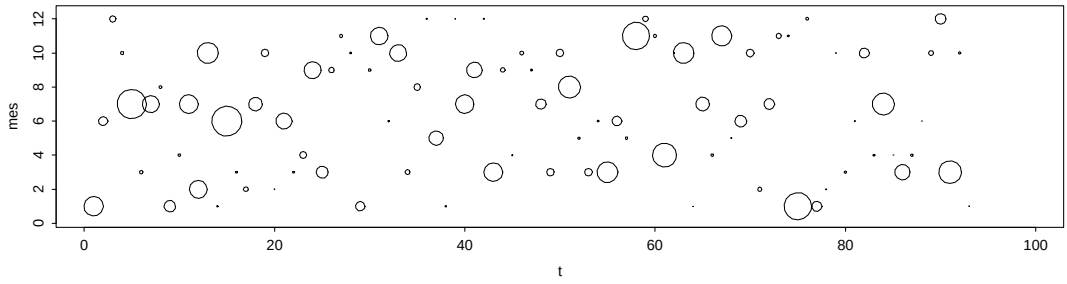

Figura 3.12: Gráfico de burbujas de la intensidad máxima con localización en el punto medio, ps25; Madrid.

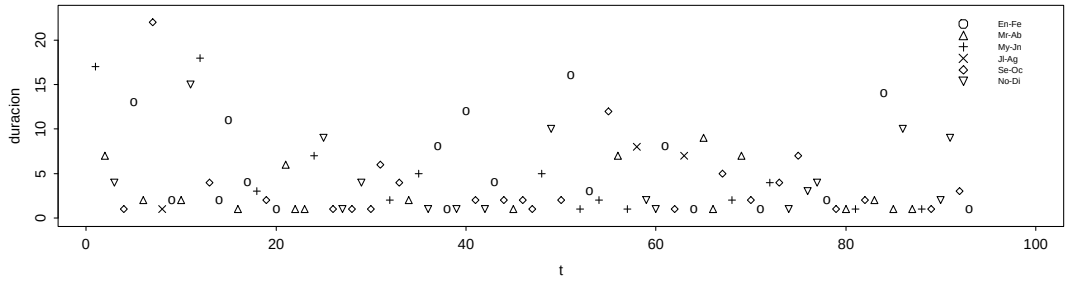

Figura 3.13: Gráfico de signos de la duración con localización en el punto inicial, ps25; Madrid.

- Tabla de contingencia del número de ocurrencias, según la duración del episodio y el mes de ocurrencia.

  Para completar el análisis proponemos dos gráficos que permiten comprobar, de forma sencilla, la existencia de un patrón estacional.

- Gráfico de burbujas. En este gráfico, figura 3.12, se representa la serie de los meses de ocurrencia de los episodios con puntos de tamaño proporcional al valor que toma la variable en esa observación.

- Gráfico de signos. Se representa la serie de una variable utilizando símbolos distintos según el periodo del año en que se produce la observación, figura 3.13.

Independencia del vector de magnitudes y el proceso de ocurrencia Se analiza la correlación entre la serie de cada una de las magnitudes y la serie de los tiempos de recurrencia utilizando las herramientas indicadas en el apartado 3.3.

Observatorio	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Media(L) (mes)	3.8	2.8	3.6	3.0	4.3	4.0
p50(L)	2	2	2	2	3	2
max(L)	15	12	22	12	24	31
p25(L)-p75(L)	1-5	1-3	1-4	1-4	2-5	1-5
Media(D) (l.)	175.7	93.2	179.2	96.1	124.4	204.7
p50(D)	39.2	21.8	38.6	40.4	38.8	40.0
max(D)	1081.9	1049.2	1901.8	529.2	755.8	1840.3
p25(D)-p75(D)	14-192	13-67	18-200	11-123	11-198	16-190
Media(Im) (l)	44.3	30.2	48.6	36.2	32.1	54.5
p50(Im)	25.0	18.9	31.2	31.0	23.2	32.6
max(Im)	153.4	139.5	181.2	116.9	75.1	246.6
p25(Im)-p75(Im)	13-71	9-43	15-62	10-54	9-59	14-91

Tabla 3.4: Descriptiva de las magnitudes asociadas a los periodos secos en ps10.

Dependencia entre periodos secos y no secos La posible dependencia existente entre los periodos secos, ps, y los no secos, pns, adyacentes, se controla mediante el análisis de la correlación entre las series ps-pns y pns-ps de las tres magnitudes asociadas a cada uno de ellos, utilizando los tests de correlación habituales.

3.4.1 Análisis de resultados

Cada una de las series de magnitudes asociadas a los periodos secos y a los no secos se somete a un análisis de independencia y homogeneidad. Los controles de homogeneidad se realizan por bimestres, enero-febrero, marzo-abril, ..., noviembre-diciembre, con cada una de las localizaciones. En la tabla 3.4 se muestran algunas medidas descriptivas de la distribución de las magnitudes asociadas a los periodos secos.

Los aspectos más destacables de los análisis son:

- **Burgos.** Aparecen indicios leves de falta de aleatoriedad en la duración y el déficit. En la representación de la serie y el gráfico de burbujas de estas magnitudes se observa un descenso de nivel de las observaciones de mayor valor, en torno al episodio número 20, que puede ser la causa de esos indicios. El efecto contrario, un aumento de nivel, se aprecia en el exceso y la duración de

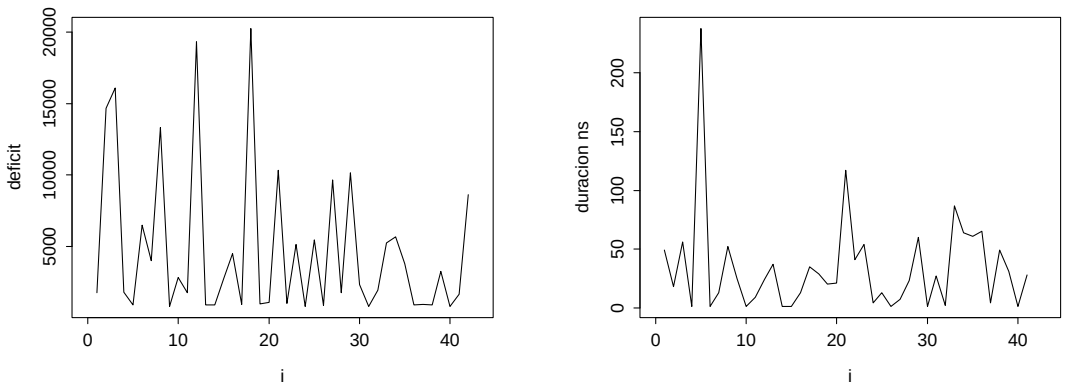

Figura 3.14: Series del déficit y la duración de los periodo no secos, ps10; Burgos.

los periodos no secos en torno a la misma observación, figura 3.14; este punto coincide temporalmente con el cambio de estructura que se ha detectado en el proceso de ocurrencia, figura 3.9. Ese cambio de nivel corresponde, aproximadamente, al año 1924; la causa podría ser una alteración en las condiciones de observación, pero Almarza et al. (1996) no señalan que en esa fecha se produjera ningún cambio en el observatorio.

- **Daroca.** Se detecta correlación y falta de aleatoriedad en la intensidad máxima y el déficit y, en menor grado, en la duración. El test Kruskall-Wallis de homogeneidad de la duración es significativo, aunque gráficamente no hay indicios de estacionalidad; los tamaños de muestra de algunos bimestres son muy pequeños, lo que puede estar influyendo en los resultados del test.

  En los periodos no secos, los p-valores de todos los tests de correlación son pequeños, siendo significativos en la duración; estos resultados son lógicos dados los problemas de correlación detectados en la serie de tiempos de recurrencia, y la relación existente entre éstos y la duración de los periodos no secos.

- **Huesca.** De nuevo aparecen problemas de aleatoriedad y correlación en el déficit y la intensidad máxima y, en menor medida, en la duración. Los periodos no secos verifican todas las hipótesis analizadas.

- **Madrid.** Los tests indican la existencia de una posible tendencia monótona en las magnitudes duración y déficit; gráficamente, figura 3.15, se aprecia un descenso en el nivel de estas magnitudes en torno al epsiodio número 13;

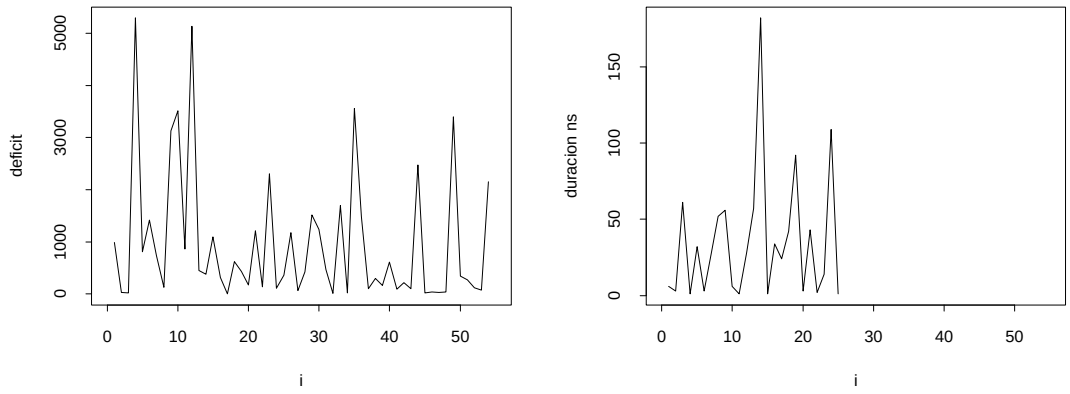

Figura 3.15: Serie del déficit de Madrid y la duración de los periodo no secos de Murcia, ps10.

en esta época, año 1901, se produjo un cambio de ubicación del Observatorio Astronómico al Instituto de Meteorología, lo que puede ser el origen de esta inhomogeneidad. Al analizar la serie correspondiente al Instituto Meteorológico, los tests dejan de ser significativos. El efecto del desplazamiento parece afectar únicamente al nivel de las magnitudes, ya que los resultados relativos al proceso de ocurrencia coinciden con los de la serie completa.

Se detecta estacionalidad en las series del déficit y la intensidad máxima con localización en el punto de intensidad máxima. Como se observa en la figura 3.16, los episodios que tienen su punto de máxima intensidad en invierno o verano son los de menos gravedad; la causa de este hecho es nuevamente el menor peso de las aportaciones de esos meses en la precipitación anual.

- **Murcia.** Las series de magnitudes se pueden considerar homogéneas e independientes.

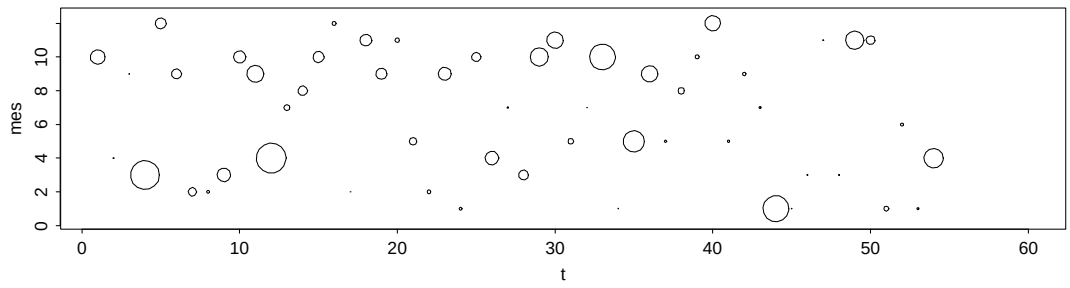

Figura 3.16: Gráfico de burbujas de la intensidad máxima con localización en el punto de intensidad máxima, ps10; Madrid.

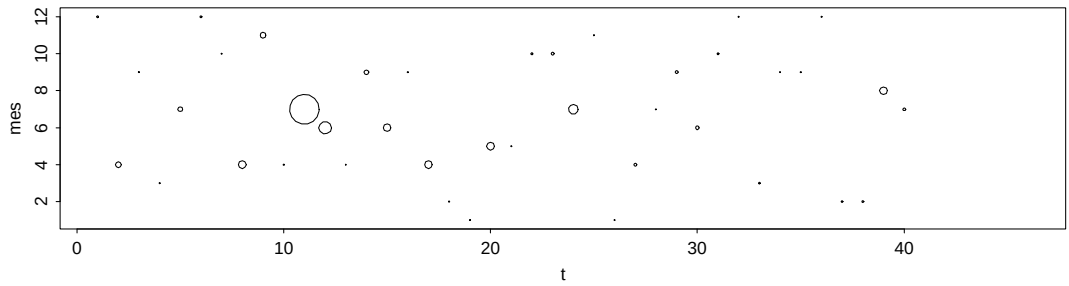

Figura 3.17: Gráfico de burbujas de la duración con localización en el punto medio de ps10; San Fernando.

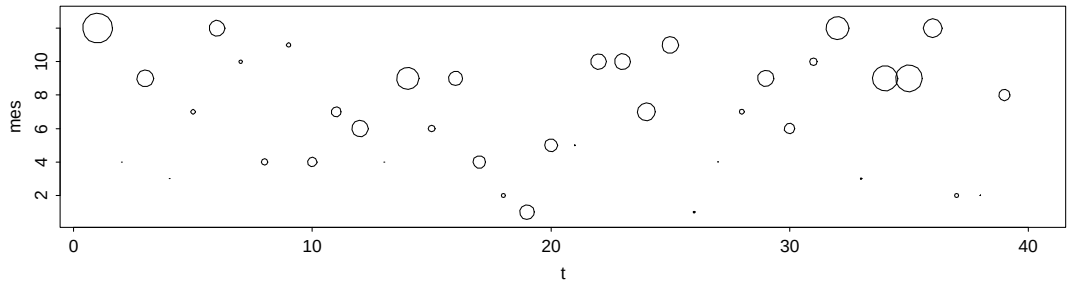

Figura 3.18: Gráfico de burbujas de la intensidad máxima de los periodos no secos con localización en el punto inicial de ps10; San Fernando.

- **San Fernando.** Se observan indicios de carácter estacional en la duración y el déficit de los periodos secos, en las series con localización en el punto medio, figura 3.17.

  En los periodos no secos la estacionalidad es clara en las tres magnitudes, especialmente en la localización punto medio, figura 3.18.

Los análisis realizados se resumen en las siguientes conclusiones.

- Las series de magnitudes, especialmente el déficit, no se pueden suponer independientes ya que presentan problemas de aleatoriedad y, en algún caso, estructura de correlación. La variable duración es la que presenta menos indicios de dependencia.

- No existen indicios significativos de tendencia en ninguna de las series.

- La homogeneidad estacional depende de las características climáticas del observatorio; de las series analizadas sólo las de Madrid y San Fernando son estacionales.

- El carácter estacional de las magnitudes es distinto dependiendo del punto de localización de la ocurrencia del episodio considerado; en general, y a diferencia de los resultados del proceso de ocurrencia, la series con localización en el punto inicial son las menos sensibles al efecto estacional.

- En todos los observatorios la relación de dependencia entre las magnitudes es similar. La dependencia más fuerte se observa entre el déficit y la intensidad máxima, en segundo lugar entre la duración y el déficit, y la más débil, aunque claramente significativa, entre la duración y la intensidad máxima.

- El vector de magnitudes es independiente del proceso de ocurrencia.

- El comportamiento de las series de magnitudes de los episodios no secos es, en general, más aleatorio y homogéneo que el de los episodios secos.

- No existe correlación significativa entre los episodios secos y los no secos adyacentes.

- Respecto a la metodología empleada, cabe destacar la utilidad del gráfico de burbujas para analizar el carácter estacional, tanto de la ocurrencia como de las magnitudes.

En conclusión, la ocurrencia de los periodos secos en los observatorios analizados se puede modelizar con un PP, pero no es posible plantear un modelo PPC debido a la dependencia serial existente en los vectores de magnitudes.

3.5 Implementación en S-Plus

- **Función: dumbr.fun** (dumbr.txt). Esta función realiza todo el proceso de selección del umbral.

  Argumentos:

  - smo: serie de sumas móviles anuales

  Subfunciones:

  - rp.plot: realiza el gráfico de razón Poisson; para obtener la muestra divide el periodo de observación del proceso en k intervalos.

- rpe.plot: representa el gráfico de razón exponencial de los tiempos de recurrencia.
- npo.plot: realiza el gráfico de Castro; para obtener la muestra se divide el periodo de observación del proceso en k intervalos.
- mrl.plot: representa el gráfico de la vida residual media.
- re.plot: construye el gráfico de los excesos exponenciales.
- te.plot: realiza los tres gráficos asociados al test de Gertensgarbe y calcula el p-valor del test en el punto de corte que se le indique; se debe especificar un valor inicial del umbral.
- perc.fun: calcula el percentil al que corresponde un valor en una muestra dada.

- **Función: sseq.fun** (sseq.txt). Esta función construye las tres series de ocurrencia de los periodos secos dado un umbral, y todas las magnitudes asociadas a ese proceso; realiza además un análisis gráfico de la serie y un control, opcional, del carácter Poisson del proceso de ocurrencia.

  Argumentos:

  - datafsmo: base de datos que contiene la precipitación anual móvil.
  - nomb: etiqueta del nombre del observatorio
  - n1, n2, n3: son argumentos para las subfunciones; indican el número de intervalos en que se divide el periodo de observación.

  Subfunciones (fsseq.txt):

  - grafevolper.fun: representa, de forma simultánea, la serie de precipitación anual móvil, respecto a distintos valores del umbral.
  - grafico.fun: representa la serie de precipitación anual móvil correspondiente al umbral seleccionado.
  - contip.fun (contip.txt): realiza todos los controles indicados sobre el carácter Poisson del proceso de ocurrencia. Calcula las series del punto medio, inicial y de intensidad máxima de las ocurrencias, y los correspondientes tiempos de recurrencia; aplica el análisis de comprobación en

cada uno de los tres procesos. En los episodios de duración par, el punto medio se localiza en la posición $L/2$ desde el punto inicial. Utiliza las siguientes subfunciones:

- rp.fun: realiza un control del carácter Poisson del número de ocurrencias en un intervalo; es necesario indicar el número de intervalos en que se divide el tiempo de observación del proceso.
  Esta función llama a poiss.fun (poisson.txt) que realiza el ajuste Poisson de una muestra y los correspondientes controles de bondad de ajuste. Para aplicar el test χ^2 de bondad de ajuste se exige que el valor esperado de cada celda sea mayor que cinco excepto, como máximo, en un 20% de ellas, en las que debe ser mayor que uno. Para calcular los puntos de corte que satisfagan esta condición se utiliza la función nperiopp.fun.
- ajusexp.fun (exponencial.txt): realiza un ajuste exponencial a una muestra y los correspondientes controles de bondad de ajuste. Para aplicar el test χ^2 se exige la misma condición que en el caso Poisson; la función nperiop.fun (fajuste.txt), con argumento dist='exp' calcula los correspondientes puntos de corte.
- tendencia.fun: calcula los tests de tendencia específicos para procesos de Poisson y de renovación.
- acfsk.fun (taleat.txt): contrasta la hipótesis de autocorrelación nula en los retardos indicados aplicando los tests de Kendall y Spearman, y representa el correlograma con los correspondientes intervalos de confianza basados en el estadístico t_K propuesto y en el estadístico de Spearman habitual.
- taleat.fun (taleat.txt): realiza los tests de tendencia generales y los de aleatoriedad; los p-valores se calculan con las aproximaciones y correcciones de continuidad descritas.
- anestacion.fun: realiza el análisis del carácter estacional del proceso.

- **Función: contmag.fun** (contmag.txt). Esta función aplica los controles de independencia, homogeneidad y correlación entre magnitudes, a las seis variables definidas. Los análisis que dependen del instante de ocurrencia se realizan con las tres localizaciones.

Argumentos:

- dataf: base de datos que contiene las series de magnitudes
- datasmo: base de datos que contiene la serie de la precipitación móvil anual
- nomb: etiqueta del nombre del observatorio analizado
- tipod: etiqueta del tipo de episodio analizado, 'p.seco' o 'sequia', y percentil con el que se define el umbral.

Subfunciones (fcontmag.txt): además de acfsk.fun y taleat.fun, ya descritas, se utilizan:

- describe.fun: realiza un análisis descriptivo numérico y gráfico: histograma, gráfico de caja y serie temporal de las observaciones.
- mpacf.fun: calcula y representa la correlación parcial.
- describebi.fun: realiza un análisis descriptivo numérico y un gráfico de caja de la serie por bloques homogéneos, bimestres o trimestres.
- tabladur.fun: calcula la tabla de contingencia según la duración y el instante de ocurrencia.
- grafesta.fun: realiza un análisis gráfico del carácter estacional de una variable y calcula el test de Kruskall-Wallis.

Capítulo 4

Modelización del proceso de sequías mediante un proceso de Poisson Cluster Compuesto

El modelo PPC propuesto inicialmente en el capítulo anterior exige el carácter Poisson del proceso de ocurrencia y la independencia de la muestra del vector de magnitudes; la primera hipótesis se hace verosímil al disminuir el umbral y se acepta en el percentil décimo, pero la segunda no es plausible ni siquiera en ese umbral.

Por otra parte, las características del fenómeno estudiado y la señal elegida justifican la existencia de dependencia entre las magnitudes de los episodios consecutivos. En efecto, en un episodio largo de sequía se pueden alternar periodos secos graves con otros periodos de remisión en los que no se llega a alcanzar la normalidad, pero que sobrepasan ligeramente el umbral crítico durante un corto periodo de tiempo; por consiguiente, la sequía aparece dividida en varios periodos secos, figura 4.1, cuyas magnitudes serán mutuamente dependientes, ya que la pequeña duración del intervalo de separación no elimina el impacto del anterior periodo seco. El objetivo de este capítulo es desarrollar un modelo que permita representar esta estructura, de forma que los periodos secos que sean manifestaciones de una misma sequía se identifiquen como una unidad.

4.1 Proceso de Poisson cluster compuesto

4.1.1 Proceso de ocurrencia Poisson cluster

Un proceso de Poisson cluster, PPCl, es un PP en el que cada punto u ocurrencia tiene asociado un número aleatorio de ocurrencias que forman un proceso subsidiario o cluster. El proceso Poisson cluster corresponde a la superposición de las ocurrencias de todos estos subprocesos, sin identificar los puntos pertenecientes a cada uno de ellos. Los centros o puntos de ocurrencia de cada cluster no tienen que estar necesariamente incluidos en el proceso global, aunque no hay pérdida de generalidad en suponerlo. Se pueden considerar dos tipos de procesos cluster:

- PPCl sin superposición: procesos en los que no se puede producir el solapamiento temporal de dos clusters.

- PPCl con superposición: procesos en los que la probabilidad de solapamiento no es despreciable; en este caso, es necesario especificar no sólo el proceso de ocurrencia de los centros de los clusters, sino también el mecanismo de generación de los puntos que forman cada uno de ellos.

La sequía es intrínsecamente un suceso raro, poco frecuente, por lo que se puede suponer que la posibilidad de que se superpongan dos sequías es nula; en consecuencia, sólo se consideran PPCl sin superposición.

Definición 4.1. *Se llama* **proceso de Poisson cluster**, *PPCl*(λ), *a un proceso que verifica las siguientes propiedades,*

i.- El proceso de ocurrencia de los centros de los clusters es un PP(λ).

ii.- El número de puntos asociado a cada cluster es una variable aleatoria, Np, independiente e idénticamente distribuida. Esta variable también es independiente de la configuración del PP de los centros.

El correspondiente funcional generador de probabilidades es,

$$\Psi(g) = \exp\left[-\lambda \int_{-\infty}^{\infty} (1 - \psi_{Np}[g(x)])dx\right]$$

con $\psi_{Np}(z) = E(z^{Np})$ la función generatriz de probabilidad de Np. Este resultado permite obtener muchas propiedades del proceso, pero resulta más sencillo deducirlas a partir de las propiedades del PP de ocurrencia de los centros y del número de ocurrencias por cluster, teniendo en cuenta que el proceso global, condicionado a las posiciones de los centros, es una superposición de procesos independientes. En particular, se deducen fácilmente las siguientes propiedades:

- El número medio de ocurrencias en un conjunto A es $\lambda|A|E(Np)$.
- El número de ocurrencias en intervalos disjuntos es independiente.

4.1.2 Proceso de Poisson cluster compuesto

Como se comentó en el capítulo anterior, la ocurrencia no es el único aspecto de interés en la sequía; por esta razón se consideraron modelos que complementaban el PP introduciendo un vector aleatorio de variables para representar otras características del fenómeno. El mismo razonamiento se puede aplicar al modelo PPCl.

Definición 4.2. *Se define un* **proceso de Poisson cluster compuesto**, *$PPClC(\lambda, F_1 \ldots, F_k)$, como un PPCl en el que se asocia a cada cluster un vector aleatorio de magnitudes $(X1, \ldots, Xk)$ i.i.d., con distribuciones marginales $F_1, \ldots, F_k$ independientes del proceso de ocurrencia.*

4.1.3 Proceso de las sequías

Al considerar este fenómeno en el marco de un PPClC, cada sequía se identifica con un cluster constituido por un número aleatorio de puntos que corresponden a los periodos secos que la forman. En este modelo las magnitudes asociadas a las diferentes sequías deben ser independientes, pero puede existir dependencia entre las magnitudes correspondientes a periodos secos del mismo cluster.

El vector de magnitudes asociado a las sequías es el mismo que el asociado a los periodos secos en el PPC, (L, D, IM), aunque es necesario adaptar su definición a esta situación, figura 4.1.

- Duración: Se define como la suma de las duraciones de cada uno de los periodos secos que forman el cluster y de los periodos no secos que los separan.

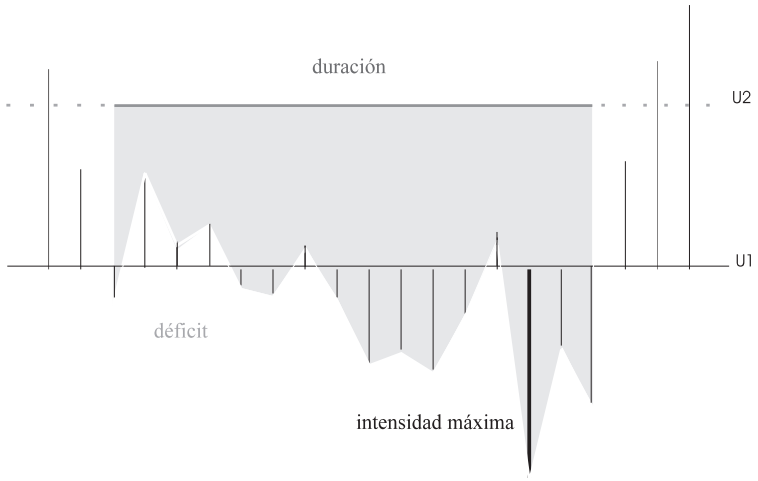

Figura 4.1: Definición de un episodio de sequía y de las magnitudes asociadas.

- Déficit: Definimos esta variable como la suma de las diferencias entre la intensidad y el umbral $U2$ -en este caso el percentil 30- de cada una de las observaciones que forman el cluster. Esta definición presenta algunas ventajas sobre la que se basa en las diferencias al umbral $U1$, que parece la definición natural:
  - Es menos sensible a la elección del umbral que, aunque se base en criterios razonables, no es absolutamente objetiva.
  - Resuelve de forma eficiente el problema relativo a la influencia en el déficit de los excesos asociados a los periodos de remisión. La definición propuesta tiene en cuenta su efecto, pero es menos sensible a su valor que otras definiciones; en particular, evita la aparición de sequías muy largas con déficits anormalmente pequeños, que no reflejan la situación real, provocados por el efecto de los excesos.
  - Garantiza que el déficit tome siempre valores positivos.
- Intensidad máxima: Se define como la intensidad máxima observada en los periodos secos que forman el cluster.

Finalmente señalaremos que si no se verifica la hipótesis de exponencialidad de los tiempos de recurrencia, el análisis de las sequías planteado, se puede enmarcar

dentro de procesos de tipo renovación y renovación alternante, más generales que el PPCl y el PPClC.

4.2 Justificación del modelo

El proceso de los episodios de sequía se obtiene como resultado de una agrupación del proceso de periodos secos cuya ocurrencia se modeliza con un PP; el problema ahora es determinar si el proceso de ocurrencia de los clusters conserva el carácter Poisson y qué condiciones son necesarias para ello. En general, el proceso resultante de la agrupación de puntos en un PP no conserva el carácter Poisson del original, pero los procesos definidos a partir de un umbral presentan la siguiente propiedad.

> **Propiedad 4.1.** *Si el proceso de ocurrencia de periodos por debajo de un umbral u es un PP, el proceso de ocurrencia de agrupaciones de dichos periodos convergerá, al decrecer el umbral, a un PP.*

En efecto, asignando el instante de ocurrencia al punto de intensidad máxima y aplicando la propiedad 3.1, se justificó el carácter Poisson de todo proceso de ocurrencia de episodios definidos con un umbral más estricto que u. Por hipótesis, el proceso de Poisson excluye la acumulación de ocurrencias en intervalos de tiempo pequeños; en consecuencia, al decrecer el valor del umbral las ocurrencias que permanecen, tienden a separarse temporalmente. En estas condiciones, en los correspondientes procesos de clusters, el número de episodios por cluster se aproxima a uno; es decir, al considerar umbrales cada vez más estrictos, el proceso de clusters converge a uno de periodos secos, que se obtiene mediante borrado aleatorio del proceso de periodos secos inicial, por lo que su carácter Poisson está garantizado.

Esta propiedad justifica la hipótesis del carácter Poisson del proceso de ocurrencia de clusters de periodos secos definidos con el percentil 10. En caso de que la hipótesis no fuera plausible en ese umbral, al disminuir éste, se haría más verosímil. En los procesos definidos con umbrales menos estrictos se pueden plantear modelos de renovación alternante compuestos; estos modelos son menos restrictivos pero no permiten la aplicación de algunos de los resultados que se presentarán en el siguiente capítulo.

A continuación se describe el proceso de ajuste del modelo. El primer paso es la determinación de los clusters a partir del proceso de ocurrencia de los periodos secos. Una vez establecido el proceso de centros de los clusters con las tres localizaciones habituales, se comprueba el carácter Poisson de la ocurrencia y se analizan las hipótesis relativas al vector de magnitudes asociado a las sequías y al de los periodos no secos que las separan.

4.3 Identificación de los clusters

Los criterios para identificar clusters pueden desarrollarse a partir de hipótesis paramétricas del modelo o de factores de carácter empírico. En cualquier caso parece razonable utilizar criterios basados en medidas de la magnitud del intervalo que separa dos ocurrencias. Antes de definir los criterios aplicados, se exponen y discuten algunos de los utilizados en problemas afines.

El criterio de unión más natural es suponer que dos ocurrencias consecutivas en los instantes T_{i-1} y T_i, pertenecen al mismo cluster si

$$T_i - T_{i-1} < z*,$$

siendo el problema determinar el valor $z*$. La mayor parte de los autores evitan la formulación de un modelo particular y realizan la selección de forma empírica.

- Tawn (1988) utiliza separaciones de 15 y 30 horas para garantizar la independencia de observaciones del nivel del mar; con los dos criterios obtiene resultados similares y comprueba la estabilidad de los resultados salvo con valores muy alejados de ese rango.

- Smith (1989) supone independientes los registros horarios del nivel de ozono que tienen una una separación mínima de un día. Con este criterio obtiene clusters muy próximos, por lo que repite el análisis considerando una separación de tres días para confirmar la validez de los resultados; éstos presentan pocas diferencias con los obtenidos inicialmente.

Otros autores complementan este tipo de criterios con medidas de la gravedad del periodo de separación.

- Zelenhasic & Salvai (1987) en su estudio sobre las sequías fluviales con medidas diarias, miden el grado de recuperación del proceso basándose en la magnitud de la duración y el exceso asociados al periodo de separación entre dos ocurrencias; consideran que dos observaciones pertenecen al mismo cluster si están separadas menos de seis días y el exceso del periodo de separación es pequeño respecto al déficit del periodo seco posterior. En estas condiciones definen la duración y el déficit del cluster como la suma de las duraciones y de los déficits de los episodios secos que lo forman, sin tener en cuenta los valores correspondientes a los periodos de separación.

  Por otra parte, consideran despreciables todas las ocurrencias con déficits tales que $D_i < p_d D_{max}$, con p_d un factor suficientemente pequeño y D_{max} el mayor déficit observado en la muestra, y las eliminan; utilizan valores de p_d entre 0.005 y 0.01.

- Madsen et al. (1994) plantean un modelo de sequías aplicando el mismo tipo de criterios que Zelenhasic & Salvai (1987), pero modifican la forma de medir la magnitud del exceso del periodo de separación: imponen que la razón entre el volumen del exceso y el déficit precedente debe ser menor que un valor crítico p_c, que definen igual a 0.3.

- Madsen & Rosbjerg (1998) proponen una nueva definición de las variables asociadas al cluster que incorpora información sobre los episodios de separación,

$$\begin{aligned} L &= L_i + L_{i+1} + Lns_i \\ D &= D_i + D_{i+1} - Ens_i \end{aligned}$$

  con Lns_i y Ens_i, la duración y el exceso del periodo entre las ocurrencias i e $(i+1)$-ésima respectivamente. Respecto al criterio de eliminación de episodios despreciables propuesto por Zelenhasic & Salvai (1987) consideran que es muy sensible a la existencia de datos atípicos, por lo que excluyen los periodos con $L_i < r_L E(L)$ o $D_i < r_D E(D)$, donde r_L y r_D son factores prefijados, en este caso con un valor igual a 0.1.

- En su revisión sobre la metodología POT, Rasmussen et al. (1994) subrayan la necesidad de imponer restricciones sobre la duración y el exceso del inter-

valo de separación entre episodios, para garantizar la independencia de las observaciones.

Una aproximación alternativa a este problema consiste en plantear y ajustar un modelo paramétrico para el proceso de ocurrencia de los periodos secos, y calcular el valor de $z*$ bajo esas hipótesis.

- Smith (1984) determina el valor de $z*$ cuando el proceso de ocurrencia corresponde a un PP doblemente estocástico cuyo proceso asociado es una cadena de Markov de dos estados.

- Davison & Smith (1990) sugieren la utilización de los modelos de Neyman-Scott y Bartlett-Lewis.

En general, la aplicación de modelos paramétricos en la determinación de los clusters no mejora los resultados. Davison & Smith (1990), tras analizar los resultados de distintos trabajos, concluyen que los análisis no son demasiado sensibles al valor de $z*$, siempre que se utilice un valor intuitivamente razonable. Finalmente, señalan que en las situaciones en las que los resultados son estables en un amplio rango de valores del umbral, el ajuste paramétrico no presenta ninguna ventaja.

4.3.1 Criterios de separación propuestos

En la elección del criterio de separación se consideran los siguientes factores.

- Existen dos propiedades que borran la memoria de los episodios pasados y que permiten suponer el comienzo de una nueva sequía: el tiempo transcurrido desde el episodio anterior y la importancia del periodo no seco que los separa.

- La utilización de un criterio basado sólo en la distancia presenta algunos inconvenientes: un valor $z*$ grande puede dar lugar a un grado de agrupación excesivo, mientras que un valor pequeño no garantizará la independencia.

- Criterios como el de Madsen et al. (1994), basados en la magnitud del periodo de separación entre dos sucesos, resultan muy sensibles al valor de p_c, de forma que, dependiendo de ese valor, la composición de los clusters varía considerablemente.

Finalmente, se optó por utilizar longitudes de separación $z*$ intermedias y complementar ese criterio con una regla basada en la intensidad máxima del periodo de separación entre los episodios, medida que consideramos menos sensible a los valores fijados a priori, que las propuestas por otros autores.

> **Criterio de separación I:** *Dos periodos secos pertenecen a clusters distintos si entre ellos hay más de seis meses de separación o la intensidad máxima entre dichos periodos es superior al percentil 30 de la precipitación anual.*

La elección de este percentil se basa en la clasificación de Gibbs & Maher (1967) que considera normales las precipitaciones anuales entre el percentil 30 y el 70, y muy inferiores a la media a las que se encuentran entre los percentiles 10 y 20. La elección del valor seis es más subjetiva; además de factores empíricos, como su relación con el ciclo semestral de precipitación en el área mediterránea, se justifica porque en las observaciones separadas por esa distancia la información común es menor que la información nueva. Esencialmente este criterio supone que dos periodos secos -que corresponden a observaciones muy inferiores a la media- pertenecen a diferentes sequías y son independientes, si están suficientemente alejados temporalmente o, si en algún instante entre ellos se alcanza un valor normal de precipitación que determine su finalización.

Para tener una idea relativa de la capacidad separadora de este criterio, se propone otro cuya capacidad de separación es razonable.

> **Criterio de separación II:** *Dos periodos secos pertenecen a clusters distintos si la separación entre ellos es mayor o igual que doce.*

La aplicación de este criterio a sumas anuales móviles implica que las observaciones de episodios separados no comparten información y, en principio, no hay motivos para suponer la existencia de dependencia entre las magnitudes asociadas a esos episodios.

Para comprobar si las observaciones separadas por un criterio son independientes, se aplican los siguientes controles.

- Valor de la correlación calculada emparejando cada observación con la primera posterior a ella que verifique el criterio de separación.

- Independencia de la serie temporal de observaciones separadas, evaluada con las herramientas habituales. La serie de observaciones separadas se construye incluyendo la primera observación de la serie original y aplicando posteriormente el criterio de separación de forma iterativa; es decir, sólo se incorporan a la nueva serie las observaciones que, de acuerdo al criterio, están separadas de la última incluida.

Criterio de eliminación de episodios despreciables Es difícil establecer una regla de carácter general sobre qué episodios se deben considerar despreciables, ya que su importancia está directamente relacionada con el umbral de definición de la sequía, y no sólo con los valores de las magnitudes asociadas. Dado que toda entrada de la suma móvil anual por debajo del umbral implica, al menos, un año de precipitación escasa, se ha considerado que con umbrales estrictos, a partir del percentil 10, todos los episodios secos se deben considerar relevantes. Con otros umbrales se aplica el criterio de Madsen et al. (1994) con $p_c = 0.05$, criterio muy restrictivo que sólo elimina los episodios insignificantes.

Estadísticamente, la inclusión o no de estos sucesos afecta más a la distribución de las magnitudes que a las propiedades del proceso de ocurrencia. En efecto, bajo la hipótesis de independencia entre el vector de magnitudes y el proceso de ocurrencia, la eliminación de los episodios despreciables representa un proceso de borrado aleatorio que no altera el carácter Poisson del proceso original.

4.3.2 Análisis de resultados

Los controles de independencia aplicados a las series de magnitudes de los periodos secos separados con los criterios I y II son satisfactorios en ambos casos aunque, en general, los p-valores correspondientes al criterio II son mayores en el déficit y la intensidad máxima, tabla 4.1; la aplicación de este criterio puede conllevar, sin embargo, una agrupación excesiva. La duración tiene un comportamiento más independiente y no presenta correlación significativa, ni siquiera en la serie original.

Magnitud	Serie	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Duración	pv S.C	0.138	0.333	0.263	0.420	0.771	0.316
	τ	(-0.152)	(-0.103)	(-0.107)	(-0.070)	(0.043)	(-0.104)
	pv C.I	0.362	0.194	0.656	0.810	0.981	0.402
	τ	(-0.094)	(-0.135)	(-0.041)	(-0.020)	(0.003)	(-0.086)
	pv C.II	0.623	0.912	0.559	0.725	0.731	0.252
	τ	(-0.050)	(-0.011)	(-0.055)	(-0.030)	(-0.047)	(-0.117)
Déficit	pv S.C.	0.129	0.186	0.015	0.555	0.944	0.168
	τ	(-0.166)	(-0.156)	(-0.256)	(-0.057)	(0.013)	(-0.155)
	pv C.I	0.238	0.326	0.613	0.933	0.761	0.298
	τ	(-0.128)	(-0.114)	(-0.053)	(-0.008)	(0.043)	(-0.116)
	pv C.II	0.406	0.703	0.863	0.994	0.623	0.081
	τ	(-0.090)	(0.044)	(-0.018)	(0.001)	(0.070)	(-0.194)
Intensidad máxima	pv S.C.	0.334	0.105	0.007	0.394	0.797	0.217
	τ	(-0.106)	(-0.190)	(-0.281)	(-0.081)	(-0.040)	(-0.139)
	pv C.I	0.501	0.182	0.479	0.806	0.981	0.321
	τ	(-0.073)	(-0.156)	(-0.074)	(-0.023)	(0.003)	(-0.111)
	pv C.II	0.613	0.643	0.785	0.896	0.925	0.100
	τ	(-0.055)	(0.054)	(-0.029)	(0.012)	(-0.013)	(-0.186)

Tabla 4.1: Coeficiente de Kendall y p-valor del test de correlación de las series completas, S.C., separadas con el criterio I, C.I, y por el criterio II, C.II.

En conclusión, con umbrales inferiores al percentil décimo, el criterio I asegura la independencia de los episodios, aunque este resultado no se puede generalizar a valores menos severos del umbral.

4.4 Control de las hipótesis de un PPClC

Las hipótesis que debe verificar un proceso PPClC son el carácter Poisson del proceso de ocurrencia de los clusters, la independencia del vector de magnitudes asociado, y la independencia entre éstas variables y el proceso de ocurrencia. Para comprobar estas hipótesis se aplican los procedimientos utilizados en el análisis del proceso de periodos secos.

Una hipótesis adicional del modelo es la independencia e idéntica distribución del número de elementos que forman cada cluster -que se evalúa con las herramientas utilizadas para contrastar la independencia y homogeneidad temporal- y su inde-

pendencia del proceso de ocurrencia.

4.4.1 Análisis de resultados

Se analiza el carácter Poisson del proceso de ocurrencia de los clusters definido con el percentil décimo, que denotaremos seq10. Dado el resultado 4.1 se analiza también un umbral un poco más estricto, correspondiente al valor en el que -de acuerdo con los análisis preliminares de selección del umbral, tabla 3.1- las observaciones presentaban carácter extremo; en torno al percentil 6 o 7, dependiendo del observatorio. El centro de los clusters se define en las localizaciones habituales, punto medio, punto inicial y punto de intensidad máxima.

Proceso de ocurrencia de los clusters En la tabla 4.2 se presenta un resumen del análisis de la serie de ocurrencia de las sequías, y a continuación señalamos los resultados más destacados.

- **Burgos.** En seq10 los tests no rechazan el carácter exponencial, sin embargo el qqplot correspondiente, similar al observado en periodos secos, no es satisfactorio; la causa es, de nuevo, el cambio de comportamiento que se aprecia en la serie, ahora en torno a la observación número 16, como se puede observar en la figura 4.2. Con la disminución del umbral la mejora del gráfico es apreciable, figura 4.3.

- **Daroca.** No hay indicios para rechazar el carácter Poisson del proceso seq10; la hipótesis de exponencialidad se hace más verosímil al disminuir el umbral.

- **Huesca.** Los resultados son similares a los de Burgos: se puede aceptar el carácter Poisson en seq10, y los p-valores correspondientes aumentan con la disminución del umbral.

- **Madrid.** En seq10 se rechaza, gráfica y analíticamente, el carácter Poisson del proceso de ocurrencia. Al disminuir el umbral al percentil 6, los tests dejan de ser significativos y los gráficos de bondad de ajuste experimentan una notable mejoría, figuras 4.4 y 4.5, que permite aceptar el carácter Poisson.

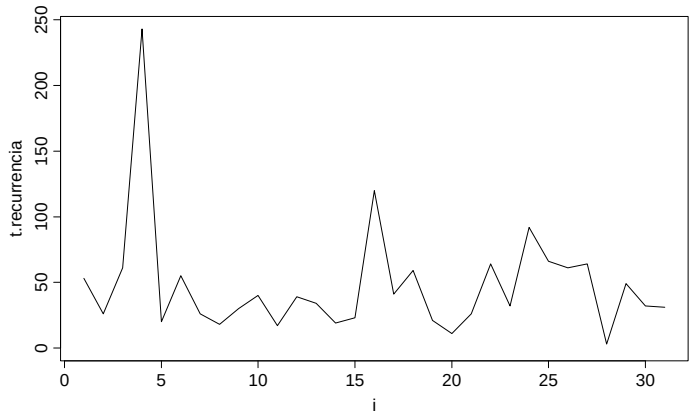

Figura 4.2: Serie de los tiempos de recurrencia con localización en el punto medio, seq10; Burgos.

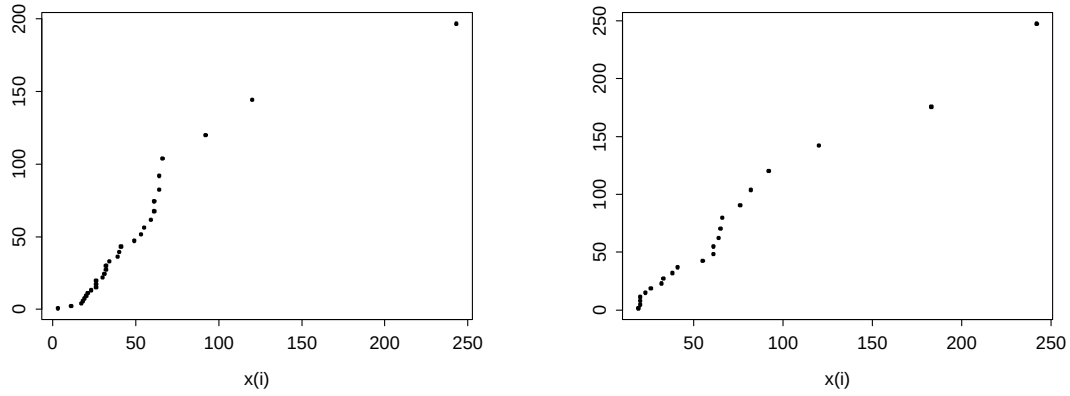

Figura 4.3: QQplot exponencial de los tiempos de recurrencia con localización en el punto medio, seq10 (izda.) y seq6 (dcha.); Burgos.

- **Murcia.** La exponencialidad de Tr, así como el carácter Poisson del número de ocurrencias se rechazan en seq10; los resultados mejoran en seq7, figura 4.6, aunque se debe tener en cuenta el pequeño tamaño de muestra disponible.

- **San Fernando.** El test de exponencialidad indica que no se puede rechazar el carácter Poisson del proceso seq10, aunque los p-valores de algunos de los tests χ^2 son pequeños. Se detecta un comportamiento estacional significativo, especialmente en las series con localización en el punto de intensidad máxima y el punto inicial. Esta inhomogeneidad provoca los bajos p-valores de los tests y la desviación que se aprecia en el qqplot. Con umbrales más estrictos la estacionalidad deja de ser significativa y mejora el carácter Exponencial, figura 4.7.

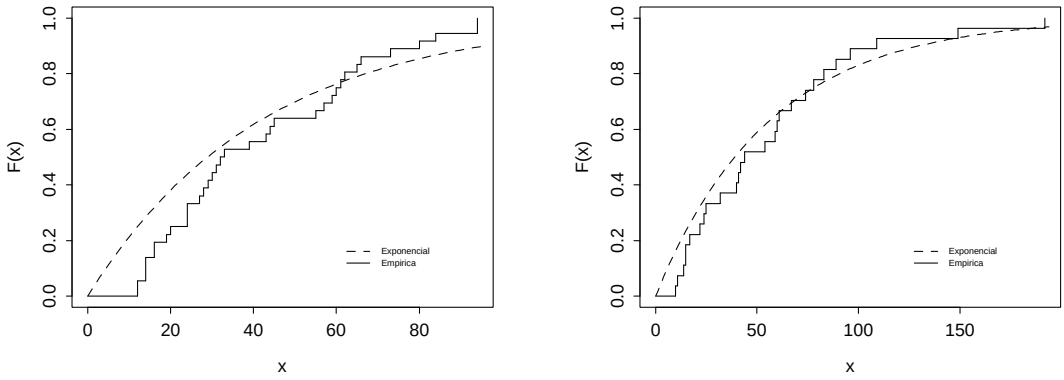

Figura 4.4: Función de distribución del ajuste exponencial de los tiempos de recurrencia con localización en el punto medio, seq10 (izquierda) y seq6 (derecha); Madrid.

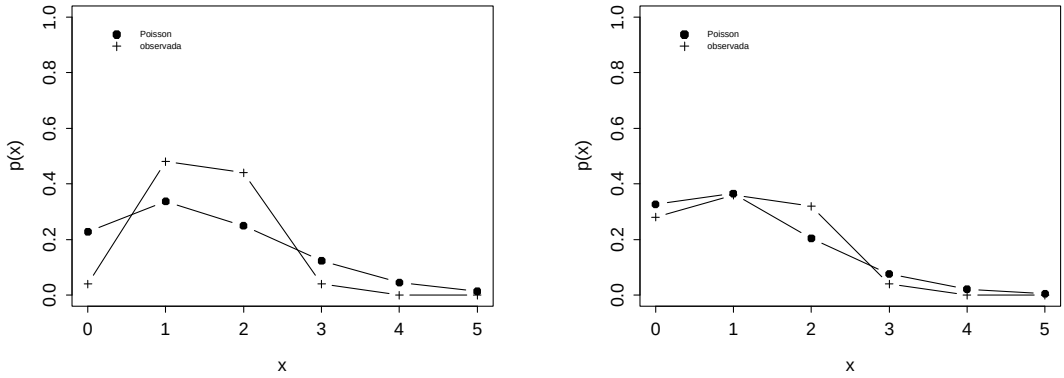

Figura 4.5: Función de probabilidad del ajuste Poisson del número de ocurrencias por intervalo, con localización en el punto medio, n^o intervalos=25, seq10 (izquierda) y seq6 (derecha); Madrid.

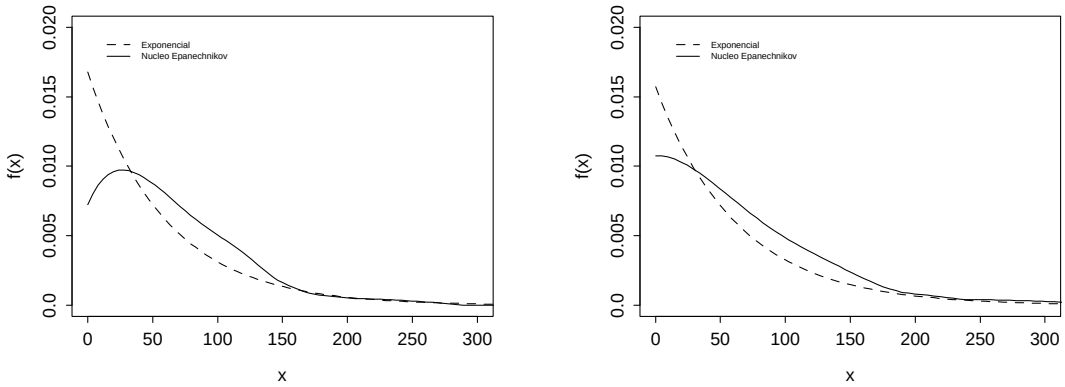

Figura 4.6: Función de densidad del ajuste exponencial de los tiempos de recurrencia con localización en el punto medio, seq10 (izda.) y seq7 (dcha.); Murcia.

	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Proceso	seq10	seq10	seq10	seq10	seq10	seq10
Umbral	3880	3060	3740	3010	1630	3630
n	32	23	28	37	15	31
Poisson (χ^2)	0.239	0.893	0.088	0.136	0.058	0.021
Exp. (K-S)	0.213	0.117	0.144	0.004	0.013	0.167
α	0.021	0.023	0.019	0.024	0.017	0.020
Indep. T_r	0.943	0.650	0.894	0.618	0.951	0.301
τ Kendall	(0.011)	(0.076)	(0.022)	(-0.061)	(-0.026)	(-0.138)
Rachas cre-dec.	0.775	0.497	0.444	0.596	0.199	0.561
Von Neuman	0.806	0.611	0.796	0.892	0.689	0.190
Cox-Stuart	0.121	0.999	0.386	0.814	0.999	0.423
Lewis-Robinson	0.350	0.692	0.510	0.756	0.576	0.641
Homogeneidad	0.741	0.307	0.963	0.882	0.126	0.593
Indep. n^o ps/cluster	0.874	0.882	0.393	0.589	0.943	0.255
τ Kendall	(-0.013)	(-0.022)	(-0.088)	(-0.044)	(-0.022)	(-0.083)

	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Proceso	seq7	seq5	seq6	seq6	seq7	seq6
Umbral	3700	2800	3500	2800	1450	3400
n	23	14	21	28	14	20
Poisson (χ^2)	0.267	0.423	0.397	0.794	0.634	0.372
Exp. (K-S)	0.239	0.300	0.213	0.216	0.206	(> 0.200)
α	0.015	0.014	0.014	0.018	0.016	0.013
Indep. T_r	0.486	0.304	0.916	0.628	0.301	0.325
τ Kendall	(-0.114)	(-0.242)	(-0.023)	(-0.071)	(-0.242)	(-0.176)
Rachas cre-dec.	0.395	0.266	0.593	0.283	0.730	0.010
Von Neuman	0.529	0.467	0.925	0.830	0.321	0.488
Cox-Stuart	0.752	0.221	0.752	0.579	0.683	0.999
Lewis-Robinson	0.998	0.297	0.437	0.074	0.476	0.087
Homogeneidad	0.365	0.335	0.489	0.463	0.294	0.639
Indep. n^o ps/cluster	0.900	0.690	0.260	0.649	0.689	0.751
τ Kendall	(-0.013)	(-0.026)	(-0.116)	(-0.031)	(-0.026)	(-0.041)

Tabla 4.2: P-valores de los tests del análisis con el percentil 10 (sup.) y un umbral inferior (inf.) con localización punto medio. El número de intervalos en el control Poisson es 15 en Murcia y Daroca, y 18 en el resto.

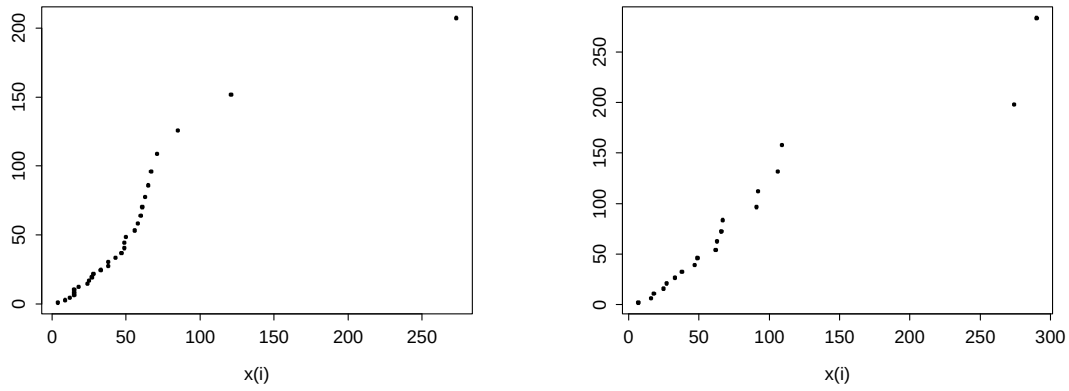

Figura 4.7: QQplot exponencial de los tiempos de recurrencia con localización en el punto medio, seq10 (izda.) y seq6 (dcha.); San Fernando.

Los resultados se resumen en las siguientes conclusiones:

- En general, el proceso seq10 de ocurrencia de los clusters se puede suponer un PP, aunque el carácter Exponencial y Poisson analizados mejora, notablemente en algunos casos, al considerar umbrales más estrictos.

- La hipótesis de independencia del proceso requiere umbrales menos estrictos que las relativas al carácter de la distribución; en el proceso seq10 todos los controles de independencia son satisfactorios.

- En el proceso seq10 no se aprecia tendencia en ningún observatorio y el carácter estacional sólo es significativo en el observatorio de San Fernando.

- En los umbrales próximos al percentil 6, que corresponden a umbrales de carácter extremo, el carácter Poisson del proceso de ocurrencia es completamente satisfactorio. El problema de considerar umbrales tan estrictos es la disminución drástica del tamaño de muestra.

Independencia y homogeneidad del número de elementos por cluster Los controles de independencia en esta serie son satisfactorios; los p-valores de los tests de independencia, tabla 4.2, no detectan ningún patrón de correlación. Tampoco el análisis descriptivo por trimestres muestra indicios de inhomogeneidad.

Observatorio	Burgos	Daroca	Huesca	Madrid	Murcia	S. Fernando
Proceso	seq10	seq10	seq10	seq6	seq7	seq10
Media(L) (mes)	5.4	5.6	6.8	3.9	5.7	5.8
p50(L)	4	4	6	3	5	4
max(L)	15	17	26	9	21	31
p25(L)-p75(L)	1-9	2-8	2-9	1-6	2-6	2-8
Media(D) (l.)	649.4	510.8	858.8	414.8	564.6	903.5
p50(D)	454.6	350.3	582.6	352.9	451.6	525.4
max(D)	2029.9	1865.2	3829.0	922.7	2038.4	5343.3
p25(D)-p75(D)	108-1023	177-704	192-1162	112-688	201-361	193-1171
Media(Im) (l.)	52.8	40.8	66.1	38.3	33.7	65.1
p50(Im)	46.1	32.1	58.5	32.5	40.3	39.5
max(Im)	153.4	139.5	181.2	95.9	57.1	246.6
p25(Im)-p75(Im)	16-84	18-57	29-79	21-51	20-49	16-101
n^{o}ps/cluster	1.3	1.6	1.6	1.3	1.1	1.3

Tabla 4.3: Descriptiva de las magnitudes asociadas a las sequías

Vector de magnitudes Se analizan las series de las magnitudes correspondientes a los procesos en los que se acepta el carácter Poisson del proceso de ocurrencia, seq10 en Burgos, Daroca, Huesca y San Fernando, seq7 en Murcia y seq6 en Madrid. En la tabla 4.3 se muestran algunos medidas resumen de la distribución de las magnitudes asociadas a las sequías.

Los resultados a destacar en cada observatorio son:

- **Burgos.** Se detecta el mismo comportamiento encontrado en el análisis de periodos secos: en la serie del déficit se produce un descenso del nivel medio en torno a la observación número 16. En las magnitudes no secas se observa el efecto contrario, un aumento del exceso a partir de la misma observación, figura 4.8. Una consecuencia de este problema, como se puede observar en la figura 4.9, es la aparición de un p-valor significativo en el test de rachas respecto a la mediana y correlación positiva de orden uno.

- **Daroca.** Existen indicios, no significativos, de estacionalidad en la duración y el exceso de los periodos no secos con localización en el punto inicial.

- **Huesca.** En las series de duración y déficit, figura 4.10, se observa un compor-

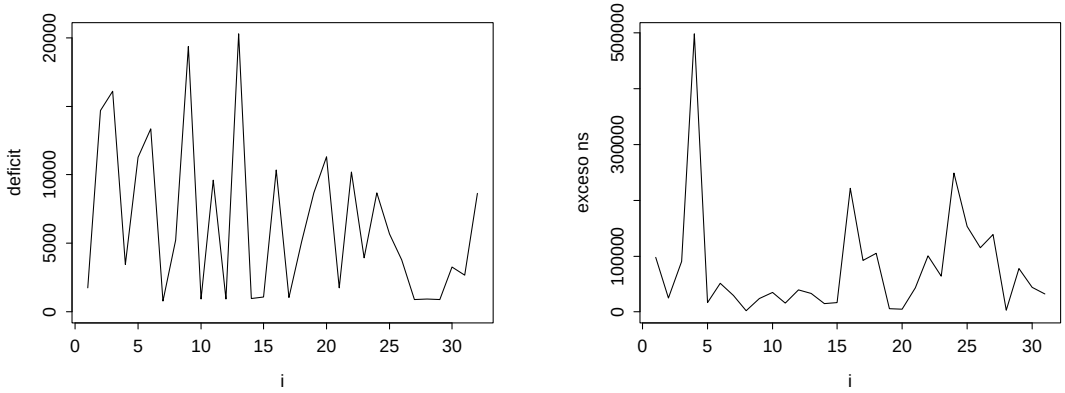

Figura 4.8: Serie del déficit y del exceso de los periodos no secos, seq10; Burgos.

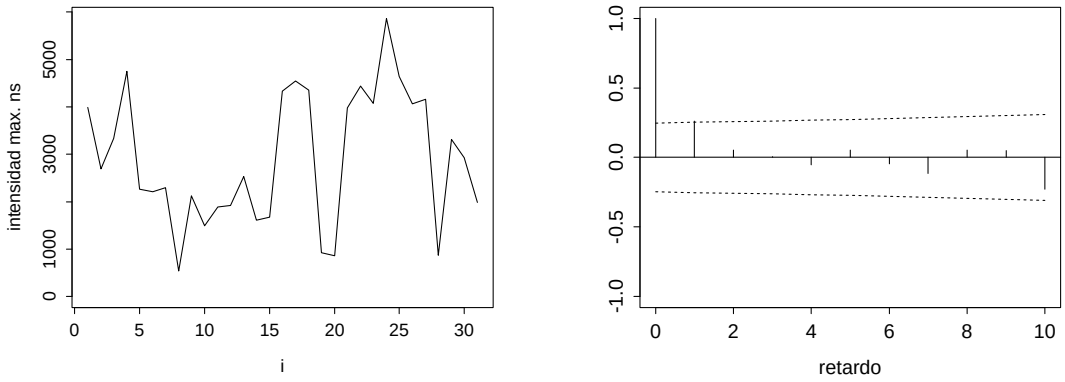

Figura 4.9: Serie de la intensidad máxima de los periodos no secos y correlograma de Kendall de esa serie, seq10; Burgos.

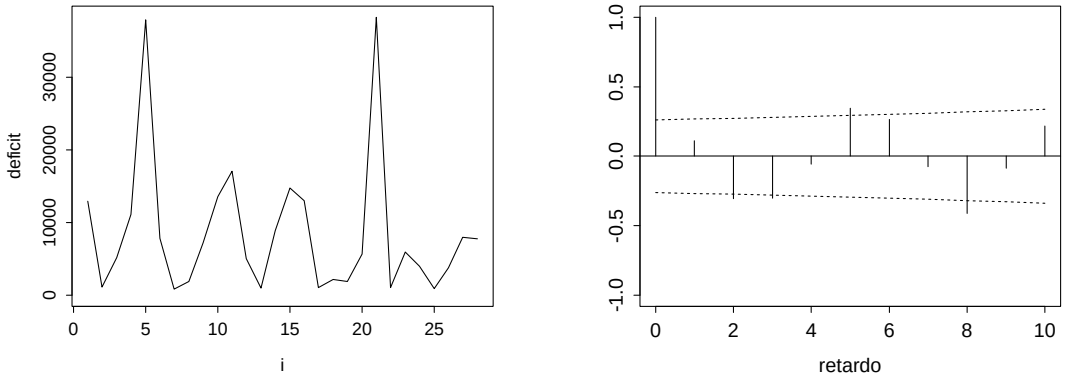

Figura 4.10: Serie del déficit y correlograma de Kendall de esa serie, seq10; Huesca.

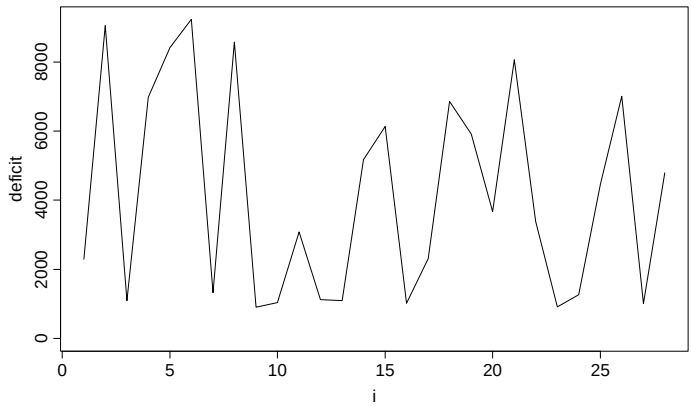

Figura 4.11: Serie del déficit, seq10; Madrid.

tamiento cíclico consecuencia de la alternancia de sequías graves -observaciones número 5, 11, 15 y 21- con otras de escasa importancia -observaciones número 2, 7, 13, 17 y 22.

- **Madrid.** En las series de duración y déficit, se observa un cambio de régimen, menos nítido que en el proceso de periodos secos, en torno a la observación número 11 que corresponde, aproximadamente, al año 1901, figura 4.11.

  El test de Kruskal-Wallis, así como el gráfico de burbujas y el de signos indican la existencia de un efecto estacional en todas las magnitudes, especialmente en el déficit y la duración; este efecto es apreciable en las tres localizaciones aunque no es significativo en la serie correspondiente al punto medio. Los gráficos de la figura 4.13 indican que las sequías que comienzan en verano son las de menor entidad.

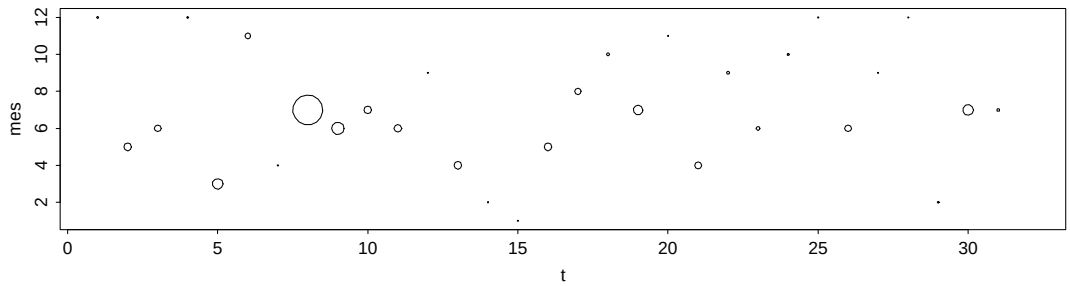

Figura 4.12: Gráfico de burbujas de la duración con localización punto medio, seq10; San Fernando.

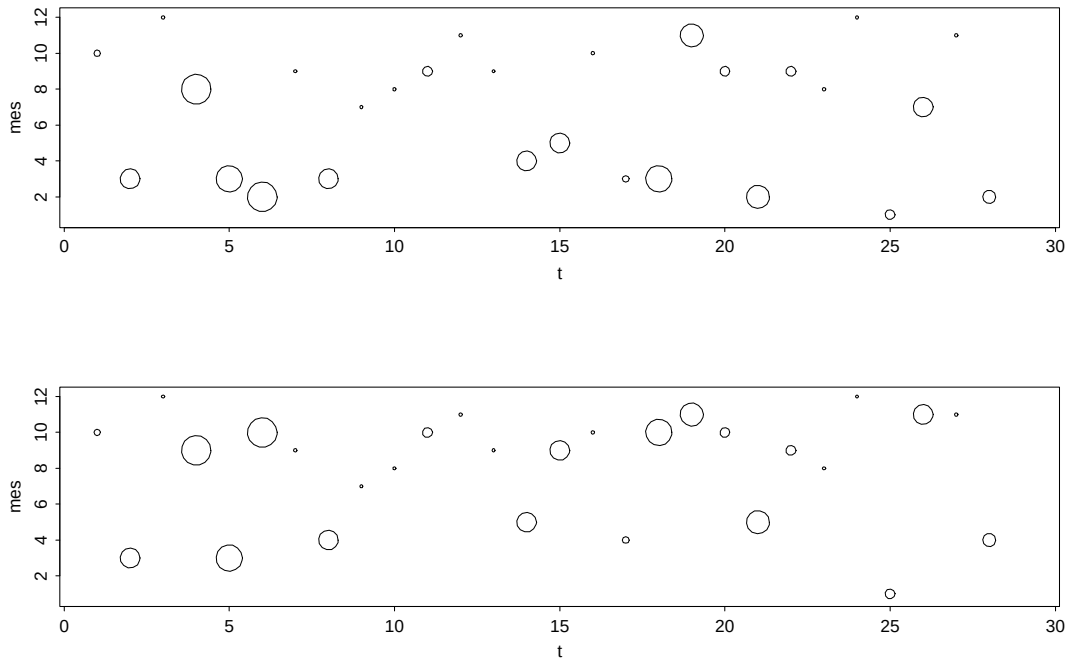

Figura 4.13: Gráficos de burbujas de la duración con localización en el punto inicial (sup.) y en el de intensidad máxima (inf.), seq6; Madrid.

- **Murcia.** No existen indicios significativos de anomalías en ninguno de los controles.

- **San Fernando.** Todas las series asociadas a los episodios secos presentan carácter estacional; como se aprecia en la figura 4.12, los episodios cuyo punto medio se localiza en los meses de verano son los de mayor duración.

A continuación se resumen las conclusiones del análisis.

- La introducción del concepto de sequía como agrupación de periodos secos próximos y su modelización con un PPClC resuelve el problema de dependencia observado en el vector de magnitudes asociado a los periodos secos. En este modelo los controles de incorrelación y aleatoriedad son aceptables en todos los observatorios.

- Respecto a las hipótesis restantes, las conclusiones son las mismas que en el proceso de los periodos secos: no se puede asegurar la existencia de tendencia en ninguna serie, y sólo las de Madrid y San Fernando presentan carácter estacional.

- Existe una fuerte dependencia entre las tres magnitudes asociadas a las sequías, aunque la estructura de correlación presenta algunas diferencias con la observada en los periodos secos; en este caso, la correlación más fuerte no se produce entre el déficit y la intensidad máxima, sino entre la duración y el déficit.

En conclusión, las series de Huesca, Burgos, y Daroca correspondientes al percentil 10 y la de Murcia definida con el percentil 7 satisfacen todas las hipótesis de un PPClC. A partir de 1901, la serie de Madrid correspondiente al percentil 6 también verifica las hipótesis de un PPClC, pero con vector de magnitudes no homogéneo estacionalmente. En el caso de San Fernando, un modelo adecuado para la serie definida con el percentil 10 podría ser un PPClC no homogéneo, dado el carácter estacional tanto del proceso de ocurrencia como del vector de magnitudes.

Comparación de los tiempos de recurrencia de los distintos observatorios Los resultados de este análisis son análogos a los obtenidos en la serie ps10, ya que los mayores tiempos de recurrencia de ambos procesos son, básicamente, los mismos; las pequeñas diferencias que se aprecian están provocadas por los desplazamientos en la localización de los puntos de ocurrencia de sequías y periodos secos.

4.5 Implementación en S-plus

- **Función: sseq.fun** (sseq.txt). La segunda parte de esta función -la primera se describió en el capítulo anterior- es opcional e identifica los periodos secos que forman cada cluster aplicando el criterio I. Establecida la composición de los clusters, calcula todas las magnitudes asociadas a este proceso, incluido el número de elementos que forman cada uno, y realiza un control opcional para detectar y eliminar los episodios despreciables de la muestra. Finalmente, ofrece la posibilidad de comprobar el carácter Poisson del proceso aplicando el procedimiento habitual con la función contip.fun, o realizar un análisis exploratorio utilizando sólo el gráfico de Castro.

  Argumentos:

  - datafsmo: base de datos que contiene la precipitación anual móvil

- nomb: etiqueta del nombre del observatorio
- otros argumentos de las subfunciones.

Subfunciones de la segunda parte:

- union.fun: calcula las magnitudes asociadas a un episodio de sequía formado por la unión de periodos secos.

- **Función: indnec.fun** (ajusnec.txt). Esta función analiza la serie del número de elementos por cluster; realiza un control de independencia y homogeneidad, con las series de cada una de las localizaciones.

  Argumentos:

  - dataf: base de datos con las magnitudes asociadas a los clusters
  - datafsmo: base de datos con la serie de precipitación móvil anual
  - nomb: etiqueta del nombre del observatorio
  - tipod: etiqueta con el tipo de dato: 'sequia' y el percentil que define el periodo seco.

- **Función: contmag.fun** (contmag.txt). Esta función realiza los controles de las hipótesis que debe verificar el vector de magnitudes; se debe especificar el argumento etiqueta tipod='sequia'. Los controles de estacionalidad se hacen por trimestres, diciembre-enero-febrero, marzo-abril-mayo, junio-julio-agosto y septiembre-octubre-noviembre.

Capítulo 5

Distribución del vector de magnitudes. Análisis de máximos

Una vez comprobado que el proceso de sequías verifica las hipótesis del PPClC y ajustado el modelo del proceso de ocurrencia, interesa analizar el vector de magnitudes que describe las características de los episodios observados, con el objeto de determinar valores útiles en la toma de decisiones y planificación de recursos hídricos, por ejemplo, la duración esperada de una sequía, la probabilidad de observar déficits superiores a uno dado, el periodo de retorno de episodios de cierta intensidad, o la duración y el déficit de la sequía más grave que cabe esperar en un determinado periodo de tiempo. Para dar respuesta a estas cuestiones se ajustan modelos paramétricos a las magnitudes asociadas a un episodio -duración, déficit, intensidad máxima- y se caracteriza la correspondiente distribución del máximo de una muestra. Se ajusta también una distribución al número de elementos por cluster.

5.1 Modelos paramétricos para las magnitudes

En este apartado se presentan distribuciones que pueden ser adecuadas para ajustar las magnitudes asociadas a las sequías, y criterios de comparación y validación de los ajustes que permitan seleccionar la mejor distribución en cada caso. La estimación de los parámetros se realiza, en todos los modelos, utilizando el procedimiento de

máxima verosimilitud.

5.1.1 Distribuciones más frecuentes en el análisis de fenómenos extremos

En primer lugar, revisaremos brevemente las distribuciones más utilizadas en estudios de carácter climático e hidrológico.

Duración: La distribución más utilizada para modelizar esta magnitud es la Exponencial; Zelenhasic & Salvai (1987) y Madsen & Rosbjerg (1998) ajustan esta distribución a la duración de episodios de sequías fluviales. Si el rango de valores de la variable es reducido se pueden considerar distribuciones discretas: Griffiths (1990) y Madsen et al. (1994) ajustan una distribución Geométrica a las duraciones de sequías basadas en la precipitación mensual y los flujos fluviales respectivamente, y Castro & Pérez-Abreu (1994) proponen una distribución Binomial negativa truncada para modelizar la duración de huracanes.

Déficit: Zelenhasic & Salvai (1987) y Madsen & Rosbjerg (1998) proponen la distribución Exponencial para ajustar esta magnitud. Madsen et al. (1994) utilizan la distribución PG para modelizar déficits a escala diaria, mensual y anual, obteniendo un buen ajuste salvo en casos con déficits muy extremos. Griffiths ajusta una distribución Gamma al déficit -exactamente al déficit dividido por la mediana anual- condicionado a la duración.

Intensidad Máxima: El ajuste de esta magnitud es habitual en las aproximaciones POT, en las que sólo se considera la máxima observación de cada cluster. En los primeros trabajos se utilizó la distribución Weibull y la Exponencial, (Artina & Todini 1985, Bardsley & Manly 1987). Posteriormente, debido a su justificación asintótica y a los buenos resultados que proporciona, se ha generalizado la utilización de la distribución PG: Smith (1989) ajusta esta distribución a niveles máximos de ozono en la atmósfera, Smith (1984) la utiliza para modelizar la altura máxima de olas, y Davison & Smith (1990) para los excesos del flujo fluvial.

Distribución	f. densidad $f(x)$	f. distribución $F(x)$
Exponencial(α)	$\alpha \exp(-\alpha t)$	$1-\exp(-\alpha t)$
Weibull(ν, α)	$\nu\alpha(\alpha t)^{\nu-1}\exp[-(\alpha t)^{\nu}]$	$1-\exp[-(\alpha t)^{\nu}]$
Gamma(μ, ν)	$\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} t^{\nu-1}\exp\left(-\frac{\nu}{\mu}t\right)$	–
Lognormal(μ, σ)	$\frac{1}{t\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right]$	–
Pareto Generalizada(γ, σ)	$\frac{1}{\sigma}\left(1+\gamma\frac{t}{\sigma}\right)^{-(1+1/\gamma)}$	$1-\left(1+\gamma\frac{t}{\sigma}\right)^{-1/\gamma}$
Binomial negativa(n, p)	$\binom{n+t-1}{n-1}(1-p)^t p^n$	–
Geométrica 0(p)	$p(1-p)^t$	$1-(1-p)^{t+1}$
Poisson(λ)	$\frac{\lambda^t \exp(-\lambda)}{t!}$	–

Tabla 5.1: Distribuciones propuestas para modelizar las magnitudes.

En la tabla 5.1 se resumen las distribuciones propuestas, que se pueden generalizar considerando un parámetro de localización μ. Dado que todas las magnitudes son variables positivas, se han considerado distribuciones con rango $(0, \infty)$, o $Z_+ \cup \{0\}$ en el caso de las discretas.

5.1.2 Análisis preliminar

Antes de formular un modelo paramétrico conviene estudiar las distribuciones posibles, para seleccionar entre ellas la más adecuada. El gran número de distribuciones existentes dificulta esta tarea, por lo que interesa disponer de criterios rápidos y simples para realizar una primera selección.

Un método exploratorio, utilizado en Fiabilidad y Análisis de Supervivencia, para determinar la adecuación de un modelo paramétrico son los gráficos lineales de la función de supervivencia $S(t) = \bar{F}(t)$. Dada la expresión de $S(t)$ de una distribución determinada, se buscan funciones continuas g_1 y g_2 que linealicen la relación entre t y $S(t)$. Se representa $g_1[\hat{S}(t)]$ frente a $g_2(t)$, siendo $\hat{S}(t)$ una estimación no paramétrica de la función de supervivencia, por ejemplo el estimador de Kaplan-Meier, de forma que si el modelo es correcto, la nube de puntos debe ser aproximadamente lineal. El ajuste de una recta de regresión a los puntos $(g_2(t_i), g_1[\hat{S}(t_i)])$ proporciona una medida de la adecuación de la distribución propuesta, el coeficiente R^2, y estimadores

Distribución	$g_2(t)$	$g_1[S(t)]$	Estimadores de los parámetros
Exponencial(α)	t	$\ln[S(t)]$	$\hat{\alpha} = -\hat{\beta}_1$
Weibull(ν, α)	$\ln(t)$	$\ln(-\ln[S(t)])$	$\hat{\alpha} = \exp\left(\hat{\beta}_0/\hat{\beta}_1\right)$, $\nu = \hat{\beta}_1$
Gamma(μ, ν) (aprox.)	$\sqrt{t}$	$\Phi^{-1}(1 - S(t))$	-
Lognormal(μ, σ)	$\ln(t)$	$\Phi^{-1}(1 - S(t))$	$\mu = -\hat{\beta}_0/\hat{\beta}_1$, $\sigma = 1/\hat{\beta}_1$
Geométrica0(p)(μ, σ)	t	$\ln[S(t)]$	$p = 1 - \exp(\hat{\beta}_1)$

Tabla 5.2: Funciones de los gráficos lineales de supervivencia de las distribuciones propuestas.

iniciales de los parámetros de la misma, a partir de los coeficientes de la recta ajustada. En la tabla 5.2 se muestran las funciones g_1 y g_2, y los estimadores iniciales de lo parámetros para algunas distribuciones. En la distribución Gamma se realiza una comprobación basada en que la distribución de $\sqrt{T}$, siendo T una variable con distribución Gamma, es aproximadamente Normal.

Para la distribución PG no existen funciones de linealización; en este caso se utiliza el gráfico del exceso medio, descrito en el análisis de selección del umbral, sección 3.2.3, que proporciona los siguientes estimadores de los parámetros.

$$\begin{aligned} \hat{\gamma} &= \frac{\hat{\beta}_1}{1 + \hat{\beta}_1} \\ \hat{\sigma} &= \frac{\hat{\beta}_0}{1 + \hat{\beta}_1}. \end{aligned}$$

Para analizar la adecuación de las distribuciones discretas se utiliza un gráfico basado en la función generatriz de probabilidades, $\psi(z) = E[z^X]$, propuesto por Nakamura & Pérez-Abreu (1993). Dada una muestra $(x_1, \ldots, x_n)$ y definiendo $Y(z) = \ln[\varphi(z)]$, se demuestra que la función,

$$\hat{Y}_n(z) = \ln[\hat{\varphi}_n(z)] = \ln\left(\frac{1}{n}\sum_{j=1}^{n} z^{x_i}\right)$$

converge a $Y(z)$. Si la muestra tiene una distribución Poisson(λ),

$$Y(z) = \lambda(z - 1)$$

y, en consecuencia, la nube de puntos $(z_i, Y_n(z_i))$ debe ser aproximadamente lineal con $\hat{\lambda} = \hat{\beta}_1$. Las funciones $Y(z)$ correspondientes a una distribución Binomial,

$B(n,p)$, y Binomial negativa, $BN(n,p)$, son funciones cóncava y convexa respectivamente,

$$\begin{aligned} Y(z) &= n\ln(1-p+pz) \\ Y(z) &= n\left(\ln(p)-\ln[1-(1-p)z]\right), \end{aligned}$$

de forma que si el gráfico no es lineal, su forma indicará qué distribución es más adecuada.

Una alternativa a los gráficos lineales de $S(t)$ consiste en representar transformaciones que linealicen la función de riesgo $h(t) = f(t)/S(t)$; por ejemplo, la propia h(t) si la distribución es Weibull o Exponencial, o $1/h(t)$, si es PG. El inconveniente de estos gráficos es que requieren una estimación no paramétrica de $h(t)$, que se puede obtener a partir de una estimación tipo núcleo de la función de densidad, pero que resulta más complicada que la de $S(t)$.

5.1.3 Criterios de comparación y de bondad de ajuste

Las distribuciones seleccionadas en el análisis preliminar se comparan aplicando los siguientes criterios.

- Comparación gráfica de la estimación paramétrica de la función de densidad con una estimación no paramétrica utilizando el núcleo de Epanechnikov.

- Comparación gráfica de la estimación paramétrica de la función de distribución con la función de distribución empírica.

- Evaluación de medidas de la distancia entre la función de distribución empírica y la estimación paramétrica; se definen tres distancias:

  D1: $\max_i |F_e(x_i) - \hat{F}(x_i)|$

  D2: $\sum_{i=1}^{n} |F_e(x_i) - \hat{F}(x_i)|$

  D3: $\sum_{i=1}^{n} [F_e(x_i) - \hat{F}(x_i)]^2$.

  Para las distancias 2 y 3 se define una versión, D2tp y D3tp, en la que el sumatorio recorre todos los enteros hasta el máximo observado en la muestra. Se consideran también las correspondientes medidas relativas, D2r, D3r,

D2tpr y D3tpr, en las que se evalúan las diferencias divididas por la estimación paramétrica.

- QQplot de la distribución ajustada.

- Test de Kolmogorov-Smirnov de bondad de ajuste. El resultado de este test es conservador ya que supone que los parámetros de la distribución son conocidos. Sólo se dispone de una versión del test con parámetros estimados para las distribuciones Normal y Exponencial.

- Comparación del criterio AIC de Akaike,

$$AIC = -2 * \ell\ell + 2p$$

siendo p el número de parámetros del modelo.

- Test de razón de verosimilitud. Este test permite comparar dos distribuciones cuando una contiene a la otra como caso particular; por ejemplo, la distribución Weibull, Gamma, y PG con la Exponencial, y la distribución Binomial negativa con la Geométrica0.

- Análisis de residuos. Cox & Snell (1968) proponen una definición general de residuo, válida en una gran variedad de modelos. La idea básica consiste en expresar el término de error del modelo como una función,

$$\varepsilon_i = g_i(\theta, y_i),$$

siendo θ un vector de parámetros desconocido e y_i el valor de la variable respuesta, cuya distribución sea conocida bajo la hipótesis nula; en estas condiciones, se define el residuo denominado de máxima verosimilitud,

$$e_i = g_i(\hat{\theta}, y_i),$$

siendo $\hat{\theta}$ el MLE de θ. Si el modelo es correcto, la distribución de los residuos será, aproximadamente, la del error. Definiendo,

$$e_i = -\ln[S(y_i; \hat{\theta})],$$

los residuos deberán tener una distribución aproximadamente Exponencial. Para analizar el carácter exponencial de los residuos se realiza un qqplot y el gráfico lineal de $S(t)$ correspondiente.

Distribución	Media	Desviación típica	V. retorno (k u.t.)
Exponencial	$1/\alpha$	$1/\alpha$	$\ln(k\lambda)/\alpha$
Weibull	$\Gamma(\nu^{-1}+1)/\alpha$	$\sqrt{\Gamma(1+2\nu^{-1})-\Gamma^2(1+\nu^{-1})}/\alpha$	$\ln^{1/\nu}(k\lambda)/\alpha$
Gamma	μ	$\mu/\sqrt{\nu}$	$F^{-1}\left(1-\frac{1}{k\lambda}\right)$
Lognormal	$\exp\left(\mu+\frac{\sigma^2}{2}\right)$	$e^{\mu}\sqrt{e^{\sigma^2}(e^{\sigma^2}-1)}$	$F^{-1}\left(1-\frac{1}{k\lambda}\right)$
Pareto G.	$\frac{\sigma}{(1-\gamma)}$ si $\gamma<1$	$\frac{\sigma}{(1-\gamma)\sqrt{1-2\gamma}}$ si $\gamma<0.5$	$[(k\lambda)^{\gamma}-1]\frac{\sigma}{\gamma}$
Binomial N.	$n(1-p)/p$	$\sqrt{n(1-p)}/p$	$F^{-1}\left(1-\frac{1}{k\lambda}\right)$
Geométrica0.	$(1-p)/p$	$\sqrt{1-p}/p$	$-1-\frac{\ln(k\lambda)}{\ln(1-p)}$
Poisson	λ	$\sqrt{\lambda}$	$F^{-1}\left(1-\frac{1}{k\lambda}\right)$

Tabla 5.3: Parámetros de las distribuciones propuestas.

5.1.4 Aplicaciones del modelo ajustado

Una de las ventajas de ajustar un modelo paramétrico es que permite hacer inferencia y predecir en un rango de valores mayor que el de la muestra, uno de los objetivos del análisis de datos extremos. A partir del modelo ajustado se pueden calcular los valores medios de las magnitudes, la desviación típica como medida de la variabilidad de la distribución, percentiles de distinto orden y el valor de retorno en k unidades de tiempo, tabla 5.3. En un proceso de ocurrencia $PP(\lambda)$, el valor de retorno, Vr, de una magnitud con función de distribución F, en k unidades de tiempo, se define:

$$Vr = F^{-1}\left(1-\frac{1}{\lambda k}\right).$$

Esta medida es muy utilizada en Climatología e Hidrología para expresar resultados, y corresponde al valor de la variable tal que, en un periodo de tiempo de longitud k, cabe esperar una observación mayor o igual que dicho valor.

5.1.5 Análisis de resultados

Se presentan los resultados de la serie de Huesca que, como se comprobó en el capítulo anterior, satisface las hipótesis de un PPClC y es una de las de mayor longitud. Se realiza un proceso de selección y ajuste para cada variable del vector de magnitudes del proceso seq10. En primer lugar se realiza un análisis gráfico exploratorio entre las distribuciones consideradas: Exponencial, Weibull, Gamma,

Distribución	Parámetros		R^2	AIC	pv χ^2	pv K-S	pv Λ
Exponencial	$\alpha = 0.147$		0.987	165.22	0.806	0.550	-
Weibull	$\alpha = 0.141$	$\nu = 1.12$	0.966	166.64	0.358	0.859	0.447
Gamma	$\mu = 6.786$	$\nu = 1.263$	0.984	166.34	0.265	0.773	0.345
Lognormal	$\mu = 1.470$	$\sigma = 0.997$	0.941	165.56	0.266	0.557	-
Pareto G.	$\gamma = -0.094$	$\sigma = 7.431$		167.04	-	-	0.663
Geométrica0	$p = 0.147$			160.88	0.872	0.564	-
Binomial N.	$n = 1$	$p = 0.122$		162.88	0.292	0.363	0.999
Poisson	$\lambda = 5.786$			260.27	0.000	0.002	-

Tabla 5.4: Distribuciones ajustadas a la duración, seq10; Huesca.

Valor medio	7.79 meses	
D. típica	6.79	
Percentil	p=0.25	3.0
	p=0.50	5.7
	p=0.75	10.4
V. retorno	50 años	17.5
	100	22.2
	200	26.9

Valor medio	7.79 meses	
D. típica	6.27	
Percentil	p=0.25	2
	p=0.50	5
	p=0.75	9
V. retorno	50 años	16
	100	20
	200	24

Tabla 5.5: Parámetros estimados de la distribución Exponencial (izda.) y Geométrica0 (dcha.) de la duración, seq10; Huesca.

Lognormal y PG, como distribuciones continuas, y Geométrica0, Binomial negativa y Poisson, como discretas. Posteriormente se ajustan todas las distribuciones que en el análisis anterior no hayan mostrado signos claros de falta de ajuste y se comparan utilizando los criterios expuestos. Finalmente se calculan algunos parámetros de interés: media, desviación típica, percentiles 25, 50 y 75 y valor de retorno en 50, 100 y 200 años.

Duración Dadas las características de esta variable se consideran distribuciones continuas y discretas. El rango de esta magnitud es $[1, \infty)$, por lo que las distribuciones se ajustan desplazadas una unidad o, equivalentemente, con parámetro de localización $\mu = 1$. El análisis preliminar indica que una distribución sin memoria, Exponencial o Geométrica, es suficiente para describir esta magnitud; como se observa en la figura 5.1, el gráfico de Nakamura indica un carácter BN, en particular Geométrico, de la muestra. Un análisis más detallado, tabla 5.4, confirma estos

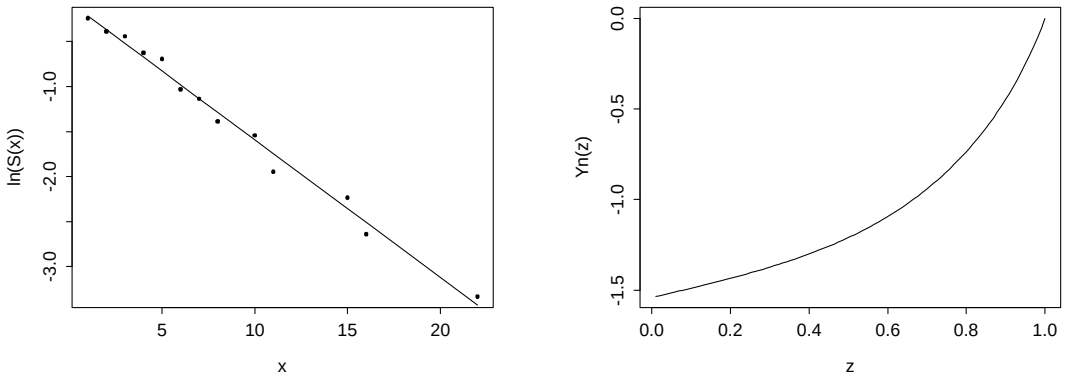

Figura 5.1: Gráfico lineal de $S(t)$ Exponencial (izda.) y gráfico de Nakamura de la duración (dcha.), seq10; Huesca.

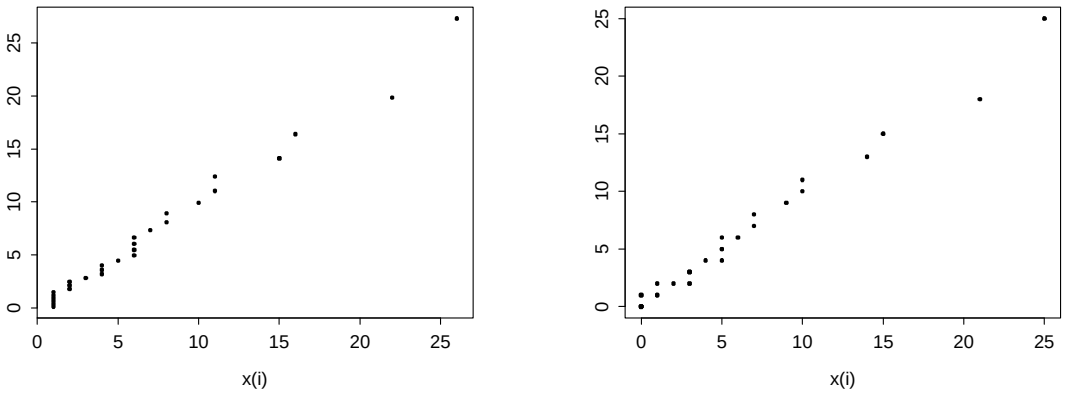

Figura 5.2: QQplot Exponencial (izda.) y Geométrico (dcha.) de la duración, seq10; Huesca.

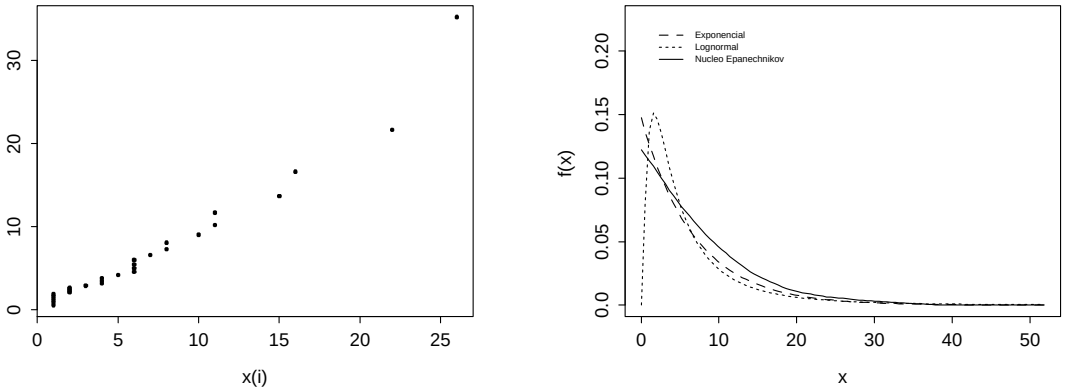

Figura 5.3: QQplot Lognormal (izda.) y gráfico de las funciones de densidad Lognormal y Exponencial ajustadas a la duración (dcha.), seq10; Huesca.

Distribución	Parámetros		R^2	AIC	pv χ^2	pv K-S	pv Λ
Exponencial	$\alpha = 0.00012$		0.902	565.26	0.785	>0.2	-
Weibull	$\alpha = 0.00011$	$\nu = 0.994$	0.928	567.26	0.630	0.926	0.967
Gamma	$\mu = 8588.143$	$\nu = 1.057$	0.953	567.20	0.818	0.862	0.815
Lognormal	$\mu = 8.516$	$\sigma = 1.093$	0.943	565.22	0.470	0.743	-
PG	$\gamma = 0.099$	$\sigma = 7735.921$		566.99	-	-	0.609

Tabla 5.6: Distribuciones ajustadas al déficit, seq10; Huesca.

Valor medio	858.8 l.	
D. típica	858.8	
Percentil	p=0.25	247.1
	p=0.50	595.3
	p=0.75	1190.6
V. retorno	50 años	2090.0
	100	2685.3
	200	3280.6

Valor medio	907.6 l.	
D. típica	1377.5	
Percentil	p=0.25	238.9
	p=0.50	499.3
	p=0.75	1043.7
V. retorno	50 años	2196.1
	100	3229.0
	200	4521.0

Tabla 5.7: Parámetros estimados de las distribuciones Exponencial (izda.) y Lognormal (dcha.) ajustadas al déficit, seq10; Huesca.

resultados. Aunque el valor del AIC correspondiente a la distribución Lognormal es competitivo con el de la Exponencial, los gráficos de las figuras 5.2 y 5.3 y los demás criterios de comparación indican que la distribución Exponencial proporciona un mejor ajuste. En consecuencia, se selecciona una distribución sin memoria, Exponencial o Geométrica0; la elección entre ambas dependerá, en cada caso, del tipo de análisis que se vaya a realizar. Las estimaciones de algunos parámetros obtenidas en estos ajustes se muestran en la tabla 5.5.

Déficit El análisis preliminar no rechaza ninguna distribución continua, y sugiere que la distribución Lognormal es la más adecuada, figura 5.4. Los resultados del test de razón de verosimilitudes, tabla 5.6, señalan que la distribución Exponencial es suficiente frente a sus posibles generalizaciones. Las distribuciones Exponencial y Lognormal proporcionan ajustes similares tanto numérica como gráficamente, figuras 5.5 y 5.6. Los residuos de ambos modelos presentan un comportamiento exponencial aceptable, figura 5.6. En la tabla 5.7 se resumen los parámetros estimados en los dos modelos.

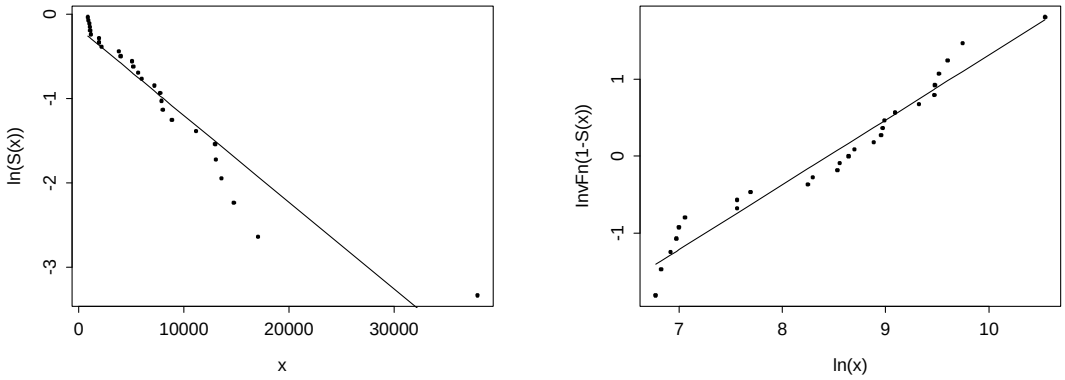

Figura 5.4: Gráfico lineal de $S(t)$ Exponencial (izda.) y Lognormal (dcha.) del déficit, seq10; Huesca.

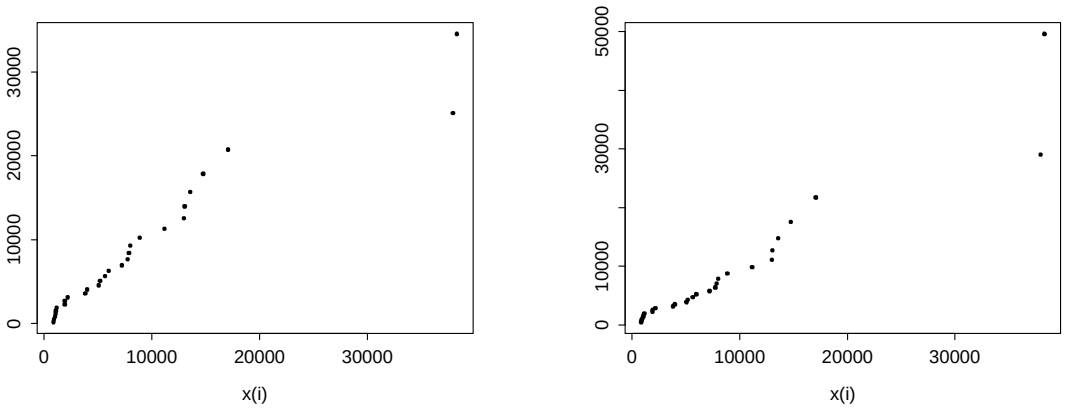

Figura 5.5: QQplot Exponencial y Lognormal del déficit, seq10; Huesca.

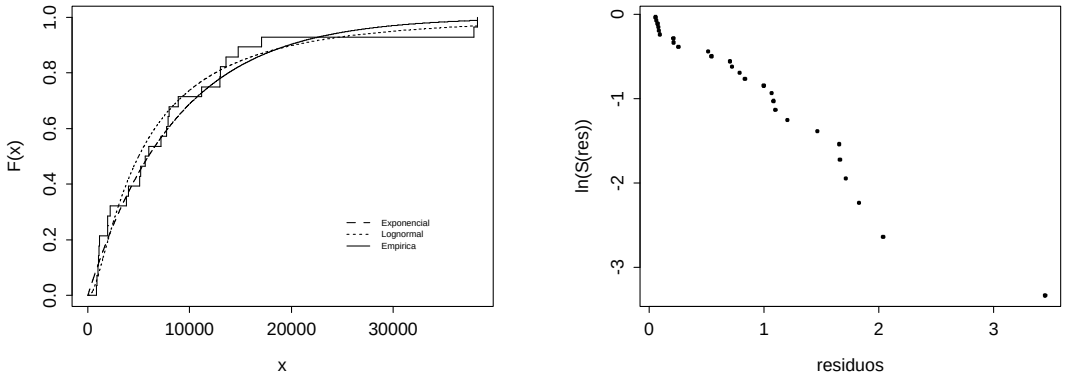

Figura 5.6: Gráfico de la f.d. Exponencial y Lognormal (izda.) y gráfico lineal de $S(t)$ Exponencial de los residuos del ajuste Lognormal (dcha.) del déficit, seq10; Huesca.

Distribución	Parámetros	R^2	AIC	pv χ^2	pv K-S	pv Λ
Exponencial	$\alpha = 0.002$	0.950	421.64	0.116	0.057	-
Weibull	$\alpha = 0.001$ $\nu = 1.28$	0.954	421.28	0.340	0.739	0.124
Gamma	$\mu = 660.893$ $\nu = 1.397$	0.979	421.89	0.144	0.581	0.186
Lognormal	$\mu = 6.095$ $\sigma = 1.081$	0.871	429.14	0.037	0.240	-
PG	$\gamma = -0.504$ $\nu = 1020.784$		419.76	-	-	0.049

Tabla 5.8: Distribuciones ajustadas a la intensidad máxima, seq10; Huesca.

Valor medio	67.9 l.	
D. típica	47.9	
Percentil	p=0.25	27.3
	p=0.50	59.7
	p=0.75	101.8
V. retorno	50 años	143.1
	100	160.6
	200	173.0

Valor medio	1.571 ps	
D. típica	0.756	
Percentil	p=0.25	1
	p=0.50	1
	p=0.75	2
V. retorno	50 años	3
	100	3
	200	3

Tabla 5.9: Parámetros estimados del ajuste PG de la intensidad máxima (izda.) y del ajuste Poisson del número de elementos por cluster, seq10; Huesca.

Intensidad máxima Según el análisis preliminar, las distribuciones Exponencial, Weibull, Gamma y PG son distribuciones posibles. Los resultados de los ajustes, tabla 5.8, indican que la distribución Exponencial es suficiente frente a la distribución Weibull y Gamma, pero no frente a la PG, conclusión que se confirma gráficamente, figuras 5.7 y 5.8. En consecuencia, se selecciona la distribución PG que presenta la máxima verosimilitud y cuya adecuación se justifica teóricamente. El análisis de bondad de ajuste de esta distribución es satisfactorio, figura 5.9. En la tabla 5.9 se resumen los parámetros estimados de la distribución.

Número de periodos secos por cluster Dado el rango de valores que toma esta magnitud, sólo se consideran distribuciones discretas desplazadas una unidad.

Distribución	Parámetros	AIC	pv χ^2	pv K-S	pv Λ
Geométrica0	$p = 0.636$	59.68	0.748	0.646	-
Binomial N.	$n = 3$ $p = 0.815$	61.20	-	0.282	0.486
Poisson	$\lambda = 0.571$	59.65	0.378	0.459	-

Tabla 5.10: Distribuciones ajustadas al número de elementos por cluster, seq10; Huesca.

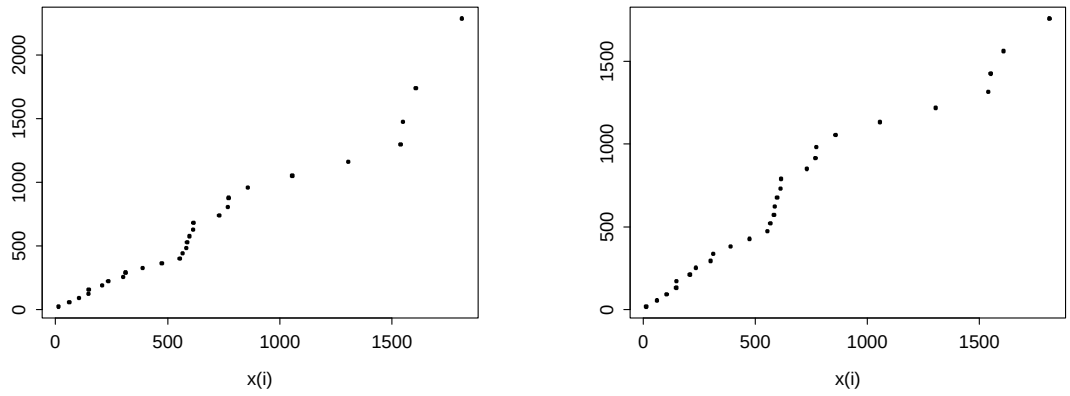

Figura 5.7: QQplot Weibull (izda.) y PG (dcha.) de la intensidad máxima, seq10; Huesca.

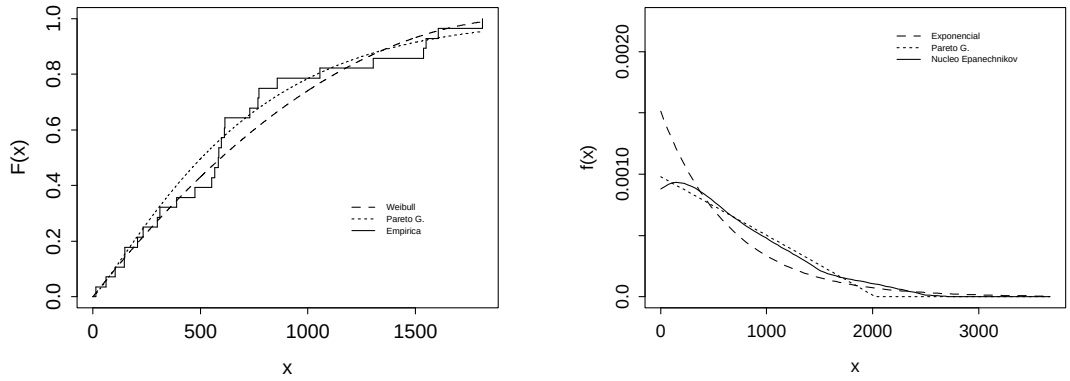

Figura 5.8: Gráfico de la f.d. Weibull y PG, (izda.) y de la f. de densidad Exponencial y PG (dcha.) de la intensidad máxima, seq10; Huesca.

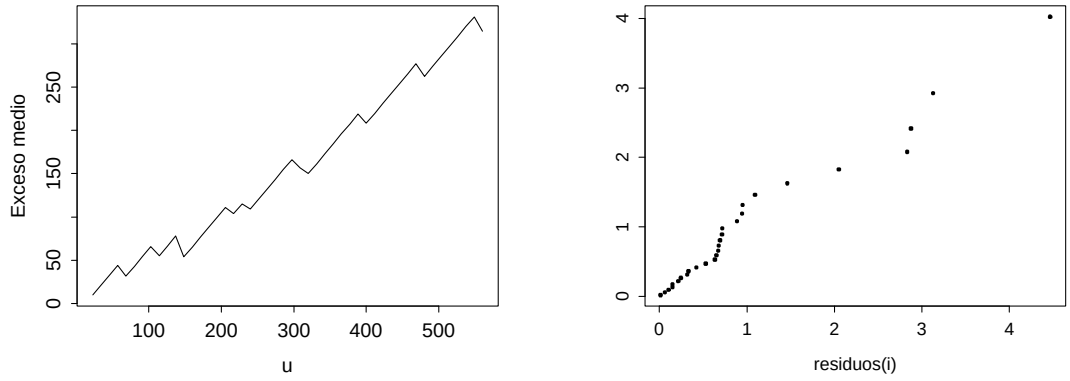

Figura 5.9: Gráfico del exceso medio (izda.) y qqplot Exponencial de los residuos del ajuste PG (dcha.) de la intensidad máxima, seq10; Huesca.

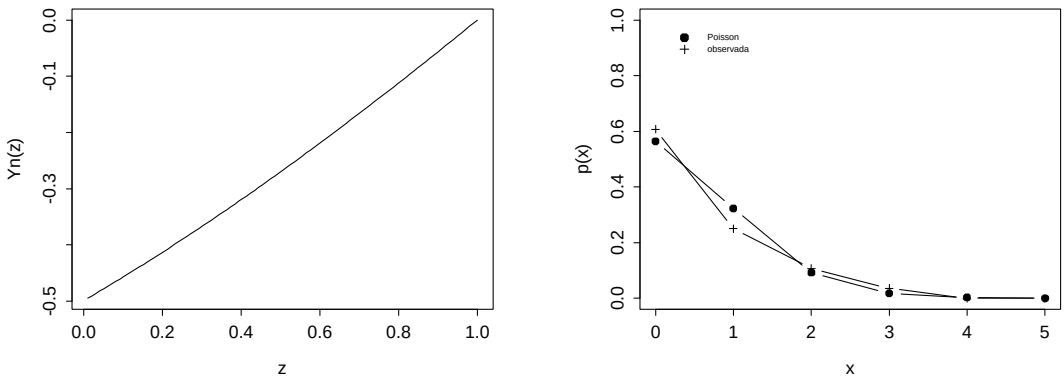

Figura 5.10: Gráfico de Nakamura y f. de probabilidad Poisson de n^o ps/cluster, seq10; Huesca.

Aunque el ajuste máximo verosímil de las distribuciones Poisson y Geométrica0 proporciona resultados similares, tabla 5.10, el gráfico de Nakamura, figura 5.10, indica que un ajuste Poisson es más adecuado, conclusión confirmada en los análisis de otros observatorios. Los estimaciones correspondientes a esta distribución se resumen en la tabla 5.9.

5.2 Distribución del máximo de una muestra

Una cuestión importante en el análisis de sequías, de utilidad en la planificación hidrológica, es la evaluación de las características de la peor sequía que cabe esperar en un cierto periodo de tiempo. Con este objetivo, se obtienen resultados que permiten caracterizar la distribución del máximo de una muestra, en la situación correspondiente al modelo de ocurrencia propuesto.

5.2.1 Máximo de muestras de tamaño no aleatorio

La distribución del máximo de una muestra de tamaño n con f.d. F es,

$$F_{M_n}(x) = F(x)^n,$$

por lo que su cálculo es sencillo si F es conocida y tiene una expresión simple; en otro caso, se puede aproximar utilizando argumentos asintóticos. Los resultados sobre max-dominios de atracción, expuestos en el capítulo segundo, determinan la

Distribución	a_n	b_n
Exponencial	$1/\alpha$	$\ln(n)/\alpha$
Weibull	$\frac{1}{\alpha\nu}\left[\ln(n)\right]^{1/\nu-1}$	$\frac{1}{\alpha}\left[\frac{\ln(n)}{}\right]^{1/\nu}$
Gamma	μ/ν	$\frac{\mu}{\nu}[\ln(n)+(\nu-1)\ln\ln(n)-\ln\Gamma(\nu)]$
Lognormal	$\sigma[2\ln(n)]^{-1/2}b_n$	$\exp\left[\mu+\sigma\left(\sqrt{2\ln(n)}-\frac{\ln(4\pi)+\ln\ln(n)}{2[2\ln(n)]^{1/2}}\right)\right]$
Pareto	$\sigma n^{1/\alpha}$	0
Beta*	$1/(n\alpha)$	1

Tabla 5.11: Constantes normalizadoras de las distribuciones.

distribución asintótica de M_n para un gran número de distribuciones, en particular para las ajustadas a las magnitudes de los episodios de sequía en la sección 5.1.

Teorema 5.1. *Las distribuciones Exponencial, Weibull, Gamma y Lognormal pertenecen al MDA de la distribución Gumbel.*

La distribución Pareto pertenece al MDA de la distribución Fréchet.

La distribución Beta pertenece al MDA de la distribución Weibull*.*

Las constantes de normalización correspondientes a estas distribuciones se muestran en la tabla 5.11.

5.2.2 Máximo de muestras de tamaño aleatorio

Nuestro objetivo es hacer inferencia sobre la máxima sequía en un periodo de tiempo de longitud n; en estas condiciones el tamaño de muestra será una variable aleatoria, $N(n)$, el número de episodios secos en ese periodo. Dado el modelo de ocurrencia propuesto, interesa establecer la distribución del máximo de una muestra de v.a. i.i.d., de tamaño aleatorio $N(n)$ con distribución $Poisson(\lambda(n))$. Por claridad, omitiremos el índice n en la notación de N y λ.

Resultados exactos

En primer lugar caracterizaremos la f.d. de M_N, el máximo de una muestra de tamaño N, en términos de la función de distribución de la muestra F.

Resultado 5.1. *La función de distribución de M_N, con $N \sim P(\lambda)$ es,*

$$F_{M_N}(x) = e^{-\lambda[1-F(x)]}.$$

Distribución	Gumbel
Esperanza	$\frac{1}{\alpha}C_{Eu} + \frac{1}{\alpha}\ln(\lambda)$ con $C_{Eu} = 0.577$
Desviación típica	$\frac{1}{\alpha}\frac{\pi}{\sqrt{6}}$
Percentil p	$\frac{1}{\alpha}\ln\left[\frac{-\lambda}{\ln(p)}\right]$

Tabla 5.12: Parámetros del máximo de una muestra Exponencial de tamaño $N \sim P(\lambda)$.

En efecto,

$$\begin{aligned} F_{M_N}(x) &= \sum_{n=0}^{\infty} P(N=n)F^n(x) \\ &= e^{-\lambda}e^{\lambda F(x)} = e^{-\lambda[1-F(x)]}. \end{aligned}$$

A continuación, y aplicando el resultado 5.1, se obtiene la función de distribución del máximo y las expresiones de algunos parámetros de interés -media, desviación típica y percentiles- de las distribuciones continuas propuestas.

Distribución Exponencial Si la distribución de las variables X_i es $Exp(\alpha)$ se obtiene la expresión,

$$\begin{aligned} F_{M_N}(x) &= \exp[-\lambda e^{-\alpha x}] \\ &= \exp[-e^{\ln(\lambda)}e^{-\alpha x}] \\ &= \exp[-\exp(\ln(\lambda) - \alpha x)] \\ &= \exp\left[-\exp\left(-\alpha\left[x - \frac{\ln(\lambda)}{\alpha}\right]\right)\right], \end{aligned}$$

de donde se deduce el siguiente resultado y los parámetros que se muestran en la tabla 5.12.

Resultado 5.2. *La distribución del máximo de una muestra aleatoria con distribución $Exp(\alpha)$ de tamaño $N \sim P(\lambda)$ es Gumbel$(1/\alpha, \ln(\lambda)/\alpha)$.*

Si la distribución de la muestra es $Exp(\alpha, \mu)$, este resultado sigue siendo válido y la distribución del máximo es Gumbel$(1/\alpha, \ln(\lambda)/\alpha + \mu)$.

Distribución	VE	
Esperanza	$\frac{\sigma}{\gamma}\left[\lambda^{\nu}\Gamma(1-\gamma)-1\right]$	si $\gamma<1$
Desviación típica	$\frac{\sigma\lambda^{\gamma}}{\gamma}\sqrt{\Gamma(1-2\gamma)-\Gamma^2(1-\gamma)}$	si $\gamma<1/2$
Percentil p	$\frac{\sigma}{\gamma}\left[\left(-\frac{\lambda}{\ln(p)}\right)^{\gamma}-1\right]$	

Tabla 5.13: Parámetros del máximo de una muestra PG de tamaño $P(\lambda)$.

Distribución Pareto Generalizada Si la distribución de las variables X_i de la muestra es $GP(\gamma,\sigma)$,

$$\begin{aligned} F_{M_N}(x) &= \exp\left[-\lambda\left(1+\frac{\gamma x}{\sigma}\right)^{-1/\gamma}\right] \\ &= \exp\left[-\left(\lambda^{-\gamma}+\lambda^{-\gamma}\frac{\gamma x}{\sigma}\right)^{-1/\gamma}\right] \\ &= \exp\left[-\left(1+\frac{x\gamma-(\sigma\lambda^{\gamma}-\sigma)}{\sigma\lambda^{\gamma}}\right)^{-1/\gamma}\right] \\ &= \exp\left[-\left(1+\gamma\frac{x-\frac{\sigma\lambda^{\gamma}-\sigma}{\gamma}}{\sigma\lambda^{\gamma}}\right)^{-1/\gamma}\right], \end{aligned}$$

de donde se deduce el siguiente resultado y los parámetros de la tabla 5.13.

Resultado 5.3. *La distribución del máximo de una muestra aleatoria con distribución $PG(\gamma,\sigma)$ de tamaño $N\sim P(\lambda)$, es $VE\left(\gamma,\sigma\lambda^{\gamma},\frac{\sigma\lambda^{\gamma}-\sigma}{\gamma}\right)$.*

En particular, si F es Pareto(α,σ), M_N tiene una distribución Fréchet$(\alpha,\sigma\lambda^{1/\alpha})$. Si la distribución de la muestra es $PG(\gamma,\sigma,\mu)$, el resultado sigue siendo válido, y en ese caso la distribución del máximo es $VE\left(\gamma,\sigma\lambda^{\gamma},\frac{\sigma\lambda^{\gamma}-\sigma}{\gamma}+\mu\right)$.

Distribución Weibull Si las variables X_i tienen distribución $W(\nu,\alpha)$,

$$\begin{aligned} F_{M_N}(x) &= \exp\left(-\lambda\exp\left[-(\alpha x)^{\nu}\right]\right) \\ &= \exp\left[-\exp\left(\ln(\lambda)-(\alpha x)^{\nu}\right)\right] \\ &= \exp\left[-\exp\left(-\frac{x^{\nu}-\ln(\lambda)/\alpha^{\nu}}{1/\alpha^{\nu}}\right)\right]. \end{aligned} \tag{5.1}$$

Esta función no corresponde a ninguna distribución conocida, aunque proponemos la siguiente caracterización que permite calcular de forma aproximada sus momentos.

Propiedad 5.1. *Si la f.d. de una variable X es de la forma 5.1, X^ν tiene una distribución Gumbel$(1/\alpha^\nu, \ln(\lambda)/\alpha^\nu)$.*

El cálculo exacto de los momentos de esta distribución es complicado; por ejemplo la esperanza,

$$E(M_N) = \int_0^\infty 1 - \exp\left[-\exp\left(-\frac{x^\nu - \ln(\lambda)/\alpha^\nu}{1/\alpha^\nu}\right)\right] dx,$$

requiere el cálculo de una integral que es convergente -basta aplicar el criterio integral y el criterio de la raíz de convergencia de series- pero cuyo valor se debe obtener por métodos numéricos.

Una alternativa consiste en calcular los momentos utilizando la propiedad 5.1, es decir, que la distribución de $Y_N = M_N^\nu$ es Gumbel$(1/\alpha^\nu, \ln(\lambda)/\alpha^\nu)$ y aplicar el método delta. Este método permite obtener, de forma aproximada, los momentos de la v.a. $g(X)$ mediante una aproximación de la función basada en los primeros términos de su desarrollo en serie de Taylor,

$$g(X) = \sum_{i=0}^{\infty} \frac{1}{i!} \frac{dg}{dx}^{(i)} (x_0)(X - x_0)^i.$$

Considerando $M_N = g(Y_N) = Y_N^{1/\nu}$ y utilizando una aproximación de segundo orden,

$$M_N = Y_N^{1/\nu} \approx y_0^{1/\nu} + \frac{1}{\nu} y_0^{1/\nu - 1}(Y_N - y_0) + \frac{1}{2\nu}\left(\frac{1}{\nu} - 1\right) y_0^{1/\nu - 2}(Y_N - y_0)^2;$$

tomando $y_0 = E(Y_N)$ y calculando la esperanza de la expresión resultante,

$$E(M_N) \approx y_0^{1/\nu} + \frac{1}{2\nu}\left(\frac{1}{\nu} - 1\right) y_0^{1/\nu - 2} Var(Y_N).$$

Como Y_N es Gumbel$(1/\alpha^\nu, \ln(\lambda)/\alpha^\nu)$, se sustituye $y_0 = [C_{Eu} + \ln(\lambda)]/\alpha^\nu$ y $Var(Y_N) = \pi^2/6\alpha^{2\nu}$, de forma que,

$$E(M_N) \approx \frac{[C_{Eu} + \ln(\lambda)]^{1/\nu}}{\alpha}\left[1 + \frac{1}{2\nu}\left(\frac{1}{\nu} - 1\right)\frac{\pi^2}{6[C_{Eu} + \ln(\lambda)]^2}\right].$$

Distribución	
Esperanza (aprox.)	$\frac{[C_{Eu}+\ln(\lambda)]^{1/\gamma}}{\alpha} + \left[1 + \frac{1}{2\gamma}\left(\frac{1}{\gamma} - 1\right)\frac{\pi^2}{6[C_{Eu}+\ln(\lambda)]^2}\right]$
Desviación típica (aprox.)	$\frac{1}{\sqrt{6}}\frac{\pi}{\gamma\alpha}[C_{Eu} + \ln(\lambda)]^{1/\gamma-1}$
Percentil p	$\frac{1}{\alpha}\left[\ln\left(\frac{-\lambda}{\ln(p)}\right)\right]^{1/\gamma}$

Tabla 5.14: Parámetros del máximo de una muestra Weibull de tamaño $P(\lambda)$.

Esta aproximación se puede mejorar hasta el orden deseado utilizando más términos del desarrollo en serie de Taylor, y calculando los momentos necesarios con la función generatriz de momentos de la distribución Gumbel.

La varianza también requiere el cálculo de una integral, cuya convergencia se prueba aplicando el criterio integral y el criterio del cociente de convergencia de series, que no se puede obtener de forma exacta. Utilizando una aproximación de orden uno en el desarrollo de Taylor se obtiene la expresión,

$$Var(M_N) \approx \left(\frac{1}{\nu}\right)^2 y_0^{2/\nu-2} Var(Y_N),$$

y sustituyendo y_0 y $Var(Y_N)$:

$$Var(M_N) \approx \frac{1}{6}\left(\frac{\pi}{\nu\alpha}\right)^2 [C_{Eu} + \ln(\lambda)]^{2/\nu-2}.$$

Puesto que la función de distribución de esta variable tiene una expresión explícita, los percentiles se pueden calcular de forma exacta. En la tabla 5.14 se resumen las expresiones obtenidas.

Distribución Gamma La función de distribución Gamma se expresa en términos de una integral que no es calculable de forma exacta, por lo que el procedimiento utilizado en los casos anteriores no resulta útil. Zelenhasic & Salvai (1987) proponen un resultado exacto para el máximo en $[0, t]$ de una muestra de tamaño Poisson con distribución Gamma(μ, ν), basado en el cálculo de la función generatriz de momentos del proceso,

$$\varphi_t(u) = e^{-\lambda t} + \left(\frac{\nu}{\mu}\right)^\nu \frac{\lambda t}{(\nu-1)!}\int_0^\infty x^{\nu-1}\exp\left(ux - \frac{\nu}{\mu}x - \lambda t e^{-x\nu/\mu}\sum_{i=0}^{\nu-1}\left(\frac{\nu}{\mu}x\right)^i\frac{1}{i!}\right)dx.$$

A partir de esta expresión deducen los momentos de orden j,

$$m_j(t) = \frac{\lambda t}{(\nu-1)!}\left(\frac{\mu}{\nu}\right)^j \int_0^\infty y^{\nu-1+j} \exp\left(-y - \lambda t e^{-y} \sum_{i=0}^{\nu-1} \frac{y^i}{i!}\right) dy$$

con $y = \nu x/\mu$. Este resultado, aunque exacto, proporciona una función complicada de evaluar, cuyo valor se puede aproximar utilizando métodos numéricos. Si ν es grande, los momentos se pueden aproximar por su valor asintótico:

$$m_j(t) \approx \lambda t \left(\frac{\mu}{\nu}\right)^j e^{-\lambda t}(\nu + j)^j.$$

Este procedimiento no permite calcular los percentiles.

Distribución Lognormal El problema de esta distribución es análogo al de la distribución Gamma; su función de distribución sólo se puede expresar en función de una integral y la expresión de F_{M_N} es demasiado complicada para realizar cálculos exactos.

Resultados asintóticos

La aplicación de resultados asintóticos sólo tiene interés en distribuciones para las que no existen resultados exactos o éstos implican cálculos demasiado complejos; por ejemplo: Gamma, Lognormal y Weibull.

El siguiente teorema, Galambos (1978), es básico para el desarrollo de las propiedades asintóticas del máximo de muestras de tamaño aleatorio.

Teorema 5.2. *Sea una muestra de tamaño aleatorio* $(X_1, \ldots, X_N)$ *tal que,*

i.- N es una v.a. con

$$\frac{N(n)}{n} \xrightarrow{p} \xi$$

siendo ξ *una v.a. positiva con función de distribución* F*.*

ii.- La distribución de la muestra es tal que M_n *verifica,*

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

donde H *es una función de distribución no degenerada.*

En estas condiciones, la distribución asintótica del máximo de la muestra es,

$$\lim_{n\to\infty} P\left(\frac{M_N - b_n}{a_n} < x\right) = \int_{-\infty}^{\infty} H^y(x)dF(y).$$

Aplicando este teorema se obtiene la distribución asintótica del máximo de una muestra de tamaño Poisson.

Resultado 5.4. *Dada una muestra de tamaño $N \sim P(\lambda)$, cuya distribución verifique la condición ii del teorema 5.2, la distribución asintótica de M_N es:*

$$\lim_{n\to\infty} P\left(\frac{M_N - b_n}{a_n} < x\right) = H^\lambda(x).$$

En efecto, si N es $P(\lambda)$ se verifica la condición i del teorema 5.2,

$$\frac{N(n)}{n} \xrightarrow{p} \lambda$$

con λ una v.a. positiva degenerada. La expresión integral se reduce a,

$$\int_{-\infty}^{\infty} H^y(x)dF(y) = H^\lambda(x)$$

obteniéndose la distribución límite buscada.

Utilizando las caracterizaciones del max-dominio de atracción, teorema 5.1, se pueden determinar las distribuciones que satisfacen la condición ii, y la forma de su función de distribución límite H; en particular las distribuciones Weibull, Gamma, Lognormal, pertenecen al MDA(Gumbel) y, por consiguiente, la distribución asintótica del máximo de todas ellas es del mismo tipo. Para caracterizar esta distribución es necesario el siguiente resultado previo.

Propiedad 5.2. *Si F es una función de distribución Gumbel$(1,0)$, F^a es una función de distribución Gumbel$(1,\ln(a))$.*

En efecto,

$$\begin{aligned} F^a(x) &= \left[\exp\left(-e^{-x}\right)\right]^a = \exp\left(-ae^{-x}\right) \\ &= \exp\left[-e^{-x+\ln(a)}\right] = \exp\left(-e^{-[x-\ln(a)]}\right). \end{aligned}$$

Distribución asintótica	-
Esperanza	$C_{Eu} + \ln(\lambda)$
Desviación típica	$\pi/\sqrt{6}$
Percentil p	$\ln\left[-\frac{\lambda}{\ln(p)}\right]$

Tabla 5.15: Parámetros de la distribución asintótica del máximo de una muestra de tamaño $P(\lambda)$ con $F \in$ MDA(Gumbel).

Resultado 5.5. *Si F pertenece al MDA(Gumbel), la distribución asintótica del máximo normalizado de una muestra con distribución F y tamaño $N \sim P(\lambda)$, es Gumbel$(1, \ln(\lambda))$; es decir,*

$$\lim_{n\to\infty} P\left(\frac{M_N - b_n}{a_n} \leq x\right) = G_0(x; 1, \ln(\lambda)).$$

En efecto, si $F \in$ MDA(Gumbel), existen constantes normalizadoras tales que la distribución del límite estandarizado de M_n es Gumbel$(1, 0)$,

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G_0(x).$$

Aplicando el resultado 5.4 se obtiene,

$$\lim_{n\to\infty} P\left(\frac{M_N - b_n}{a_n} \leq x\right) = G_0^{\lambda}(x; 1, 0).$$

Finalmente, como consecuencia de la propiedad 5.2 la f.d. $G_0^{\lambda}(x; 1, 0)$ corresponde a una Gumbel$(1, \ln(\lambda))$, con lo que se demuestra el resultado.

Conocida la distribución límite de M_N es inmediato obtener las expresiones asintóticas de los momentos y parámetros de interés, que se resumen en la tabla 5.15.

A continuación se presenta la distribución asintótica del máximo de las distribuciones que nos interesan. Para obtener estas caracterizaciones basta aplicar el resultado 5.5 y sustituir las correspondientes constantes normalizadoras expuestas en la tabla 5.11.

Distribución Gamma. *La distribución del máximo de una muestra i.i.d. con distribución Gamma(μ, ν) de tamaño $N \sim P(\lambda)$ se puede aproximar, si n es grande, por una distribución,*

Distribución asintótica	Gumbel
Esperanza	$\frac{\mu}{\nu}\left[C_{Eu}+\ln\left(\frac{n[\ln(n)]^{\nu-1}\lambda}{\Gamma(\nu)}\right)\right]$
Desviación típica	$\frac{\mu}{\nu}\frac{\pi}{\sqrt{6}}$
Percentil p	$\frac{\mu}{\nu}\ln\left(-\frac{n[\ln(n)]^{\nu-1}\lambda}{\ln(p)\Gamma(\nu)}\right)$

Tabla 5.16: Parámetros de los distribución asintótica del máximo de una muestra Gamma de tamaño $P(\lambda)$.

$$Gumbel\left[\frac{\mu}{\nu},\frac{\mu}{\nu}\ln\left(\frac{n[\ln(n)]^{\nu-1}\lambda}{\Gamma(\nu)}\right)\right].$$

Distribución Lognormal. *La distribución del máximo de una muestra de variables i.i.d. con distribución $Lnor(\mu,\sigma)$ de tamaño $N\sim P(\lambda)$ se puede aproximar, si n es grande, por una distribución,*

$$Gumbel\left[\frac{\sigma b_n}{\sqrt{2\ln(n)}},b_n\left(1+\frac{\sigma\ln(\lambda)}{\sqrt{2\ln(n)}}\right)\right],$$

con $b_n=\exp\left[\mu+\sigma\left(\sqrt{2\ln(n)}-\frac{\ln(4\pi)+\ln\ln(n)}{2\sqrt{2\ln(n)}}\right)\right]$.

Distribución Weibull. *La distribución del máximo de una muestra i.i.d. con distribución $W(\nu,\alpha)$ de tamaño $N\sim P(\lambda)$ se puede aproximar, si n es grande, por una distribución,*

$$Gumbel\left[\frac{[\ln(n)]^{1/\nu-1}}{\alpha\nu},\frac{1}{\alpha}[\ln(n)]^{1/\nu-1}\ln(n\lambda^{1/\nu})\right].$$

Distribución asintótica	Gumbel
Esperanza	$b_n\left(\frac{\sigma}{\sqrt{2\ln(n)}}[C_{Eu}+\ln(\lambda)]+1\right)$
Desviación típica	$\frac{\sigma b_n}{\sqrt{2\ln(n)}}\frac{\pi}{\sqrt{6}}$
Percentil p	$b_n\left[1+\frac{\sigma}{\sqrt{2\ln(n)}}\ln\left(-\frac{\lambda}{\ln(p)}\right)\right]$

Tabla 5.17: Parámetros de la distribución asintótica del máximo de una muestra Lognormal de tamaño $P(\lambda)$.

Distribución asintótica	Gumbel
Esperanza	$\frac{1}{\alpha\nu}[\ln(n)]^{1/\nu-1}\left[C_{Eu}+\ln\left(n^{\nu}\lambda\right)\right]$
Desviación típica	$\frac{1}{\alpha\nu}[\ln(n)]^{1/\nu-1}\frac{\pi}{\sqrt{6}}$
Percentil p	$\frac{1}{\alpha\nu}[\ln(n)]^{1/\nu-1}\ln\left[-\frac{n^{\nu}\lambda}{\ln(p)}\right]$

Tabla 5.18: Parámetros de la distribución asintótica del máximo de una muestra Weibull de tamaño $P(\lambda)$.

5.2.3 Penúltima aproximación del máximo

Una cuestión importante en los resultados de carácter asintótico es la velocidad de convergencia; interesa conocer si la aproximación que proporciona la distribución asintótica para un cierto tamaño de muestra es aceptable, y determinar distribuciones que proporcionen mejores aproximaciones si n no es suficientemente grande. Esta cuestión ha sido analizada por diversos autores para muestras de tamaño no aleatorio; algunos de los resultados obtenidos se resumen en el apartado siguiente.

Penúltima aproximación en muestras de tamaño no aleatorio

Fisher & Tippett (1928) señalan que la convergencia del máximo de n variables Normales independientes a su distribución límite, Gumbel, es extremadamente lenta y que la distribución del máximo está más próxima a una distribución Weibull*. Gomes (1984) confirma que, en el caso de variables Normales, la aproximación con una distribución Weibull* o Fréchet es mejor que la correspondiente a la distribución límite. Castillo (1988) propone también la utilización de lo que se denomina penúltima aproximación.

> **Resultado 5.6.** *La distribución del máximo M_n de variables i.i.d. con distribución $F \in MDA(VE)$ se aproxima mejor por una distribución VE, $H(x;\gamma_n,\sigma_n,\mu_n)$, la penúltima aproximación, que por su distribución límite $H(x;\gamma,\sigma_n,\mu_n)$, con γ constante.*

La inclusión de un grado de libertad más, asociado al parámetro de forma que varía con el tamaño de muestra, mejora la aproximación. Castillo (1988) propone estimar los parámetros γ_n, σ_n y μ_n por el método de los percentiles; es decir, igualando

tres percentiles calculados con la distribución exacta, $F^n(x)$, a los correspondientes calculados con la penúltima aproximación.

Aplicando este resultado a las distribuciones Weibull, Gamma y Lognormal, se obtiene:

Resultado 5.7. *La distribución del máximo de una muestra con distribución Weibull, Gamma o Lognormal se aproxima mejor por una distribución $VE(\gamma_n, \sigma_n, \mu_n)$ que por una $Gumbel(\sigma_n, \mu_n)$.*

Penúltima aproximación en muestras de tamaño aleatorio

Siendo los resultados sobre la penúltima convergencia de gran aplicación, interesa disponer de resultados análogos para la convergencia en muestras de tamaño aleatorio. En el siguiente resultado comprobamos que la penúltima aproximación también es aplicable cuando el tamaño de muestra, N, es Poisson.

Resultado 5.8. *Sea $F \in MDA(VE)$ tal que,*

$$\lim_{n\to\infty} P\left(\frac{M_n - a_n}{b_n} < x\right) = H(x;\gamma); \tag{5.2}$$

Entonces, si la función de distribución VE de la penúltima aproximación de M_n es $H(x;\gamma_n,\sigma_n,\mu_n)$, la función de distribución de la penúltima aproximación de M_N, con $N \sim Poisson(\lambda)$, será $H^\lambda(x;\gamma_n,\sigma_n,\mu_n)$.

Se debe probar que $\forall\varepsilon > 0\ \exists\ n_0\ t.q.\ \forall\ n > n_0$,

$$\left|P\left(M_N < x\right) - H^\lambda(x;\gamma_n,\sigma_n,\mu_n)\right| \leq \varepsilon$$

Para probar esta desigualdad se acota la diferencia por la suma de dos términos, que se comprobará que son menores que $\varepsilon/2$ a partir de un cierto valor de n.

$$\begin{aligned}&\left|P\left(M_N < x\right) - H^\lambda(x;\gamma_n,\sigma_n,\mu_n)\right| \leq \\ &\left|P\left(M_N < x\right) - H^\lambda(x;\gamma,a_n,b_n)\right| + \left|H^\lambda(x;\gamma,a_n,b_n) - H^\lambda(x;\gamma_n,\sigma_n,\mu_n)\right|.\end{aligned}$$

- En primer lugar se prueba que $\forall \varepsilon > 0 \; \exists \; n_1 \; t.q. \; \forall \; n > n_1$

$$\left| P\left(M_N < x\right) - H^{\lambda}(x; \gamma, a_n, b_n) \right| < \frac{\varepsilon}{2}. \tag{5.3}$$

  En efecto, por hipótesis $F \in \text{MDA(VE)}$ y aplicando el resultado 5.4,

$$\lim_{n \to \infty} P\left(\frac{M_N - b_n}{a_n} < x\right) = H^{\lambda}(x; \gamma)$$

  o, equivalentemente, $\forall \varepsilon > 0 \; \exists \; n_2 \; t.q. \; \forall \; n > n_2$

$$\left| P\left(\frac{M_N - b_n}{a_n} < x\right) - H^{\lambda}(x; \gamma) \right| < \frac{\varepsilon}{2}$$
$$\left| P\left(M_N < y\right) - H^{\lambda}\left(\frac{y - b_n}{a_n}; \gamma\right) \right| < \frac{\varepsilon}{2}$$

  de donde se deduce la desigualdad 5.3.

- El segundo sumando se acota comprobando que $\forall \varepsilon > 0 \; \exists \; n_3 \; t.q. \; \forall \; n > n_3$,

$$\left| H^{\lambda}(x; \gamma, a_n, b_n)) - H^{\lambda}(x; \gamma_n, \sigma_n, \mu_n) \right| < \frac{\varepsilon}{2}. \tag{5.4}$$

  En efecto: si $\lim_{n \to \infty} H_n(x) = H(x)$, entonces $\lim_{n \to \infty} H_n^{\lambda}(x) = H^{\lambda}(x)$; es decir, si $\forall \varepsilon > 0 \; \exists \; n_4 \; t.q. \; \forall \; n > n_4$,

$$|H(x; \gamma, a_n, b_n) - H(x; \gamma_n, \sigma_n, \mu_n)| < \frac{\varepsilon}{2}, \tag{5.5}$$

  entonces $\forall \varepsilon > 0 \; \exists \; n_5 \; t.q. \; \forall \; n > n_5$,

$$\left| H^{\lambda}(x; \gamma, a_n, b_n) - H^{\lambda}(x; \gamma_n, \sigma_n, \mu_n) \right| < \frac{\varepsilon}{2}.$$

  En consecuencia, para comprobar la desigualdad 5.4 basta demostrar 5.5. Esta diferencia se acota de nuevo por dos sumandos:

$$|H(x; \gamma, a_n, b_n) - H(x; \gamma_n, \sigma_n, \mu_n)| \leq$$
$$|H(x; \gamma, a_n, b_n) - P\left(M_n < x\right)| + |P\left(M_n < x\right) - H(x; \gamma_n, \sigma_n, \mu_n)| <$$
$$\frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

  La acotación del segundo término es inmediata porque $H(x; \gamma_n, \sigma_n, \mu_n)$ es la

penúltima convergencia de M_n. El primer sumando se obtiene de la igualdad 5.2; en efecto, dado $\varepsilon > 0$ $\exists$ n_6 *t.q.* $\forall$ $n > n_6$,

$$\begin{aligned}
\left| P\left(\frac{M_n - a_n}{b_n} < x\right) - H(x;\gamma) \right| &\leq \frac{\varepsilon}{4} \\
\left| P\left(M_n < y\right) - H\left(\frac{y - b_n}{a_n};\gamma\right) \right| &\leq \frac{\varepsilon}{4} \\
\left| P\left(M_n < y\right) - H(y;\gamma,a_n,b_n) \right| &\leq \frac{\varepsilon}{4}.
\end{aligned}$$

En conclusión, se obtiene que $\forall \varepsilon > 0$ $\exists$ $n_0 = \max(n_i)$ *t.q.* $\forall$ $n > n_0$,

$$\left| P\left(M_N < x\right) - H^{\lambda}(x;\gamma_n,\sigma_n,\mu_n) \right| \leq \varepsilon.$$

Antes de presentar algunas aplicaciones de este resultado generalizamos la propiedad 5.2 a distribuciones VE.

Propiedad 5.3. *Si F es una función de distribución $VE(\gamma,\sigma,\mu)$, la función de distribución F^a es,*

$$VE\left(\gamma, \sigma a^{\gamma}, \mu + \frac{(a^{\gamma}-1)\sigma}{\gamma}\right).$$

Además, el parámetro de forma de las dos distribuciones es el mismo y, en consecuencia, ambas son del mismo tipo.

Para obtener el resultado basta plantear la expresión de la función de distribución de la nueva variable,

$$\begin{aligned}
F^a(x) &= \left(\exp\left[-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right]\right)^a \\
&= \exp\left[-a\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right] \\
&= \exp\left[-\left(\left[1+\gamma\frac{x-\mu}{\sigma}\right]\frac{1}{a^{\gamma}}\right)^{-1/\gamma}\right] \\
&= \exp\left[-\left(1+\gamma\frac{x-\mu-\sigma a^{\gamma}/\gamma+\sigma/\gamma}{\sigma a^{\gamma}}\right)^{-1/\gamma}\right] \\
&= \exp\left[-\left(1+\gamma\frac{x-\left[\mu+\frac{(a^{\gamma}-1)\sigma}{\gamma}\right]}{\sigma a^{\gamma}}\right)^{-1/\gamma}\right].
\end{aligned}$$

Distribución asintótica	VE	
Esperanza (aprox.)	$\sigma_n \frac{\Gamma(1-\gamma_n)-1}{\gamma_n} + \mu_n$	si $\gamma_n < 1$
Desv. típica (aprox.)	$\frac{\sigma_n}{\gamma_n}\sqrt{\Gamma(1-2\gamma_n) - \Gamma^2(1-\gamma_n)}$	si $\gamma_n < 0.5$
Percentil p (aprox.)	$\mu_n + \frac{\sigma_n}{\gamma_n}\left([-\ln(p)]^{-\gamma_n} - 1\right)$	

Tabla 5.19: Parámetros de la penúltima aproximación del máximo de una muestra de tamaño $P(\lambda)$, con $F \in$ MDA(Gumbel).

Finalmente presentamos la penúltima aproximación de las distribuciones de mayor interés en el análisis de las sequías.

Resultado 5.9. *Si $F \in MDA(Gumbel)$, en particular si F es una distribución Weibull, Gamma o Lognormal, la penúltima aproximación de M_N, con $N \sim P(\lambda)$, es una distribución $VE(\gamma_n, \sigma_n, \mu_n)$.*

En efecto, por ser $F \in MDA(Gumbel)$, la penúltima aproximación de M_n es una distribución VE. Aplicando el resultado 5.8, se obtiene que la penúltima aproximación de M_N tiene una distribución F^λ con F una distribución $VE(\gamma'_n, \sigma'_n, \mu'_n)$. De acuerdo con la propiedad 5.3, ésta corresponde a una distribución:

$$VE\left(\gamma'_n, \sigma'_n\lambda^{\gamma'_n}, \mu'_n + \frac{(\lambda^{\gamma'_n} - 1)\sigma'_n}{\gamma'_n}\right),$$

y se obtiene el resultado con $\sigma_n = \sigma'_n\lambda^{\gamma'_n}$, $\mu_n = \mu'_n + (\lambda^{\gamma'_n} - 1)\sigma'_n/\gamma'_n$ y $\gamma_n = \gamma'_n$.

Los valores σ_n, μ_n y γ_n se pueden obtener por el método de los percentiles. En efecto, como $N \sim P(\lambda)$, la distribución exacta de la muestra es $F_{M_N}(x) = e^{-\lambda[1-F(x)]}$ y los percentiles,

$$x_p = F^{-1}\left[1 + \frac{1}{\lambda}\ln(p)\right].$$

Por otra parte, los percentiles de una distribución $VE(\gamma_n, \sigma_n, \mu_n)$ son:

$$z'_p = \mu_n + \frac{\sigma_n}{\gamma_n}\left([-\ln(p)]^{-\gamma_n} - 1\right)$$

Igualando estas expresiones,

T	50 años	100 años	200 años	500 años
Distribución	Gumbel	Gumbel	Gumbel	Gumbel
λ_P	11.4	22.8	45.6	114.0
μ	17.55	22.27	26.99	33.22
σ	6.80	6.80	6.80	6.80
$E(M_N)$ (mes)	21.48	26.20	30.91	37.15
$Dt(M_N)$	8.73	8.73	8.73	8.73
Mediana	15.1	19.8	24.5	30.7

Tabla 5.20: Distribución exacta de la duración máxima en un periodo T, $F \sim Exp$, seq10; Huesca.

T	50 años	100 años	200 años	500 años
Distribución	VE	VE	VE	VE
λ_P	11.4	22.8	45.6	114.0
γ	0.330	0.308	0.290	0.270
μ	21978.0	32304.1	45219.7	66988.6
σ	13191.2	16643.9	20671.5	26988.9
$E(M_N)$ (l.)	3589.7	4910.9	6535.0	9227.3
$Dt(M_N)$	3589.9	4168.9	4861.1	5957.5
Mediana	2711.7	3876.2	5321.3	7738.6

Tabla 5.21: Distribución asintótica del déficit máximo en un periodo T, $F \sim Lnor$, seq10; Huesca.

$$F^{-1}\left[1+\frac{1}{\lambda}\ln(p)\right] = \mu_n + \frac{\sigma_n}{\gamma_n}\left([-\ln(p)]^{-\gamma_n} - 1\right)$$

para tres valores de p, se obtiene un sistema de ecuaciones en σ_n, μ_n, y γ_n que se puede resolver numéricamente. En la tabla 5.19 se muestran los estadísticos de esta distribución.

5.2.4 Análisis de resultados

Se calcula la distribución del máximo en periodos de 50, 100, 200 y 500 años, de las magnitudes, duración, déficit e intensidad máxima, a partir de la distribución seleccionada y ajustada en la sección 5.1 y del proceso de ocurrencia Poisson ajustado en el cuarto capítulo. Como aplicación, se han calculado algunas medidas -media, desviación típica y mediana- correspondientes a cada una de esas distribuciones. Con las distribuciones Exponencial y PG se utilizan resultados exactos, y con las restantes resultados asintóticos basados en la penúltima aproximación.

T	50 años	100 años	200 años	500 años
Distribución	Gumbel	Gumbel	Gumbel	Gumbel
λ_P	11.4	22.8	45.6	114.0
μ	20280	26056	31833	39468
σ	8333	8333	8333	8333
$E(M_N)$ (l.)	2509.0	3086.6	3664.2	4427.8
$Dt(M_N)$	1068.8	1068.8	1068.8	1068.8
Mediana	1722.6	2300.2	2877.8	3641.4

Tabla 5.22: Distribución exacta del déficit máximo en un periodo T, $F \sim Exp$, seq10; Huesca.

T	50 años	100 años	200 años	500 años
Distribución	VE	VE	VE	VE
λ_P	11.4	22.8	45.6	114.0
γ	-0.50	-0.50	-0.50	-0.50
μ	1436.93	1614.03	1739.27	1850.39
σ	302.34	213.78	151.16	95.61
$E(M_N)$ (l.)	150.6	166.3	177.4	187.2
$Dt(M_N)$	28.0	19.8	14.0	10.9
Mediana	153.8	168.6	179.0	188.2

Tabla 5.23: Distribución exacta de la intensidad máxima en un periodo T, $F \sim PG$, seq10; Huesca.

Duración Suponiendo que la duración es Exponencial, la distribución exacta del máximo en un intervalo de tiempo dado es Gumbel; en la tabla 5.20 se resumen los estadísticos correspondientes. La sequía de mayor duración observada en la muestra, que corresponde a un periodo de tiempo de 136 años, es 26 meses.

Déficit No existen resultados exactos sobre la distribución del máximo de una muestra Lognormal, por lo que la distribución del máximo déficit, VE, se ha calculado utilizando la penúltima aproximación; los parámetros y medidas correspondientes se muestran en la tabla 5.21. La distribución exacta del máximo, suponiendo la distribución del déficit Exponencial, es Gumbel con los parámetros indicados en la tabla 5.22. El mayor déficit observado en la muestra es 3829.0 litros.

Intensidad máxima Para modelizar el máximo de la intensidad máxima se utiliza una distribución VE, que es la distribución exacta del máximo de una muestra PG; los parámetros y medidas de las distribuciones correspondientes a los distintos

periodos de tiempo se muestran en la tabla 5.23. La mayor intensidad observada en la muestra es 181.2 litros.

5.3 Implementación en S-plus

- **Función: seldist.fun** (seldist.txt). Esta función representa los gráficos exploratorios del grado de bondad de ajuste de las distribuciones consideradas: Exponencial, Weibull, Gamma, Lognormal, PG; en el caso de la duración, realiza también el gráfico de Nakamura.

  Argumentos:

  - x: serie de la magnitud que se quiere analizar
  - elige1: etiqueta para indicar la distribución de la que se quiere representar el gráfico. Toma los valores 'exp', 'wei', 'gam', 'lnor', 'pg' o 0 si se quieren representar todos los gráficos
  - nomb: etiqueta del nombre del observatorio analizado
  - tipod: etiqueta del tipo de episodios analizados ('p.seco' o 'sequia') y percentil con el que se define el umbral
  - magn: etiqueta de la magnitud, 'Duracion', 'Deficit' o 'Intensidad max.' que se ajusta.

- **Función: ajuste.fun** (ajuste.txt). Esta función realiza el ajuste MLE de distintas distribuciones continuas. Para maximizar la verosimilitud ℓ, en realidad minimizar $-\ell$, se utiliza la función nlminb de S-plus que permite minimizar una función no lineal imponiendo condiciones sobre el rango de sus parámetros. Se valida el ajuste y se calculan los estadísticos, media, desviación típica, percentiles 10, 25, 50, 75 y 90, y los valores de retorno en 50, 100, 200 y 500 años.

  Argumentos:

  - data: serie de la magnitud que se quiere ajustar
  - lambda: parámetro del PP de ocurrencia de los episodios a los que corresponde la magnitud ajustada
  - nomb, tipod, magn: etiquetas indicadas en la subfunción anterior.

Subfunciones: además de la función nperiop.fun ya descrita, se utilizan:

- ajusexp.fun (exponencial.txt). Realiza un ajuste exponencial; el estimador inicial del parámetro se calcula por el método de los momentos, que coincide con el MLE.
- ajuswei.fun (weibull.txt). Realiza un ajuste Weibull; el estimador inicial de ν es el propuesto por Menon (1963),

$$\nu = \frac{\pi}{\sqrt{6} s_{\ln(x)}},$$

y el de α se obtiene por el método de los momentos aplicado a X^ν, que tiene una distribución $Exp(\alpha^\nu)$. Llama a la subfunción weibull.fun que minimiza la función $-\ell\ell$ de este modelo.

La parametrización de esta distribución en S-plus utiliza un parámetro de escala $1/\alpha$.

- ajusgam.fun (gamma.txt). Realiza un ajuste Gamma; los estimadores iniciales de los parámetros se calculan por el método de los momentos: $\mu = \bar{x}$ y $\nu = \mu^2/s_x^2$. Llama a la subfunción gamma.fun que minimiza la función $-\ell\ell$ de este modelo.

La parametrización de esta distribución en S-plus utiliza un parámetro de escala ν/μ.

- ajuslnor.fun (lognormal.txt). Realiza un ajuste Lognormal; los estimadores iniciales de los parámetros se calculan por el método de los momentos aplicado al logaritmo de la muestra: $\mu = \overline{\ln(x)}$ y $\nu = s^2_{\ln(x)}$. La subfunción lnormal.fun minimiza la función $-\ell\ell$ de este modelo.
- ajuspar.fun (pareto.txt). Realiza un ajuste PG. Los estimadores iniciales son los proporcionados por el gráfico del exceso medio calculado por la función meplot; se comprueba que $\max(x_i) < \sigma/\nu$, condición requerida por el rango de la distribución, y si no la verifican se pueden utilizar los estimadores de momentos: $\nu = 0.5(\bar{x}^2/s_x^2 - 1)$ y $\sigma = 0.5\bar{x}(\bar{x}^2/s_x^2 + 1)$ o introducir otros. La subfunción gpd.fun (fpareto.txt) minimiza la función $-\ell\ell$ de este modelo.

Las funciones de densidad y distribución PG no están implementadas en S-plus.

- resplot.fun: (fajuste.txt) realiza los gráficos de residuos para validar el modelo.

- **Función: ajusdisc.fun** (ajusdisc.txt). Esta función, análoga a ajuste.fun, realiza el ajuste de las distribuciones discretas.

  Argumentos: Los mismos que la función ajuste.fun.

  Subfunciones:

  - ajusbn.fun (binomialn.txt). Ajusta una distribución BN utilizando la función glm.nb de la librería MASS. En S-plus el parámetro n de esta distribución sólo puede tomar valores enteros por lo que, una vez obtenido el MLE, se redondea al entero más próximo.
  - geomet0.fun (geometrica0.txt): ajusta una distribución Geométrica0. Estima el parámetro p por el método de los momentos, $p = 1/(\bar{x} + 1)$, que coincide con el MLE.
  - poiss.fun (poisson.txt): ajusta una distribución Poisson. Calcula el estimador MLE de λ, que coincide con el del método de los momentos $\lambda = \bar{x}$.

En las funciones descritas, al realizar el ajuste de la duración y el número de periodos secos por cluster, se ajustan distribuciones desplazadas: se resta uno a la muestra y posteriormente se recalculan los estadísticos correspondientes a la variable original.

- **Función: maxexp.fun** (maximos.txt). Esta función calcula los parámetros y algunos estadísticos de la distribución exacta del máximo de una muestra exponencial.

  Argumentos: además de las etiquetas habituales, nomb, tipod y magn, requiere,

  - landae: parámetro de la distribución exponencial de la muestra
  - landap: parámetro del PP de ocurrencia.

- **Función: maxpg.fun** (maximos.txt). Esta función calcula los parámetros y estadísticos de la distribución exacta del máximo de una muestra PG.

  Argumentos: las etiquetas habituales, nomb, tipod, magn y:

- gamma, sigma: parámetros de la distribución PG de la muestra
- landap: parámetro del PP de ocurrencia.

- **Función: maxmdaG.fun** (maximos.txt). Esta función calcula los parámetros y momentos de la distribución correspondiente a la penúltima aproximación del máximo de una muestra de tamaño Poisson con distribución perteneciente al MDA(Gumbel).

  Argumentos: las etiquetas habituales, nomb, tipod, magn y,

  - par1, par2: parámetros de la distribución de la muestra
  - landap: parámetro del PP de ocurrencia
  - n: longitud del periodo de tiempo en el que se ha observado el proceso de ocurrencia
  - muVE, sigmaVE, gamaVE: parámetros de la distribución VE de la penúltima aproximación
  - dist: etiqueta del nombre de la distribución de la muestra que puede tomar los valores Gamma o Lognormal.

  Los parámetros de la penúltima aproximación se calculan por el método de los percentiles, con valores de $p = 0.2,\ 0.5$ y 0.8. El correspondiente sistema de ecuaciones no lineales se resuelve utilizando el programa Mathematica.

Capítulo 6

Predicción en tiempo real

Uno de los aspectos de mayor utilidad e interés en relación con la sequía es la predicción de sus características en tiempo real, es decir cuando nos encontramos en un instante del episodio seco y la información disponible es la observada hasta ese instante. La obtención de modelos que permitan realizar esta predicción con un aceptable grado de éxito, sería de gran ayuda en la toma de decisiones y en la gestión de recursos hídricos durante un periodo de sequía. En este capítulo se proponen modelos de predicción para la duración restante de las sequías y para el riesgo de finalización de las mismas.

Antes de plantear posibles modelos es necesario establecer el tipo de información en que se va a basar la predicción. Los modelos propuestos utilizan, exclusivamente, la información aportada por el mismo proceso. Dado que nuestro objetivo es predecir en tiempo real, la información disponible sobre el estado del episodio se actualizará en cada instante, por lo que se deben considerar modelos que permitan introducir esta información, a través de covariables dependientes del tiempo u otros procedimientos.

6.1 Predicción de la duración restante de un episodio de sequía

El objetivo de esta sección es el desarrollo de modelos que permitan calcular la media y la mediana de la variable respuesta, ya que son estos valores los que se utilizarán

como predicción de la duración restante de una sequía en desarrollo.

Cada sequía está formada por un cluster de células o periodos secos, por lo que, en principio, planteamos dos posibles aproximaciones.

- Desarrollar directamente modelos para la duración restante de la sequía.
- Descomponer el proceso en dos pasos,
  a. Predicción sobre la duración restante del periodo seco actual.
  b. Predicción sobre el número de periodos secos restante y sus características.

Este último procedimiento es más complicado y debido a la gran variabilidad asociada a la segunda fase del proceso no mejora, en los ensayos efectuados, las predicciones finales. Sin embargo, dada la complejidad del fenómeno y el elevado número de variables que, a priori, pueden influir en la evolución de un episodio, hemos considerado que tiene interés realizar un estudio predictivo preliminar de los periodos secos. Los resultados de este análisis, menos complejo, aportarán información sobre el proceso de modelización y el tipo de covariables que se debe considerar en el análisis de sequías.

Antes de formular un posible modelo, planteamos un problema asociado a la predicción en tiempo real: la falta de información sobre el estado de la observación actual. Una sequía está formada por dos tipos de observaciones, las correspondientes a periodos secos -que no presentan problemas de identificación- y las pertenecientes a periodos de remisión dentro de la sequía. En tiempo real se pueden producir situaciones de indeterminación; en efecto, en las seis primeras observaciones posteriores a un periodo seco, si su valor se encuentra entre los dos umbrales U1 y U2 de definición, no hay información suficiente para asegurar si el estado de sequía ha terminado. Desde una de estas observaciones, la predicción se debe basar en la probabilidad de reentrada en un nuevo episodio seco, y la consecuente evolución. Este esquema de predicción, que combina elementos de los dos métodos de predicción propuestos en el capítulo, no está operativo en este momento, y presenta el problema adicional del pequeño tamaño de muestra disponible.

Como se ha comprobado en el capítulo cuarto, la duración de las sequías es una magnitud, aparentemente sin memoria, a la que se ajusta una distribución Geométri-

ca o Exponencial con rango $[1, \infty)$. Bajo esta hipótesis, la distribución de la duración restante en cada instante, Lr, tendrá la misma distribución con rango $[0, \infty)$. Dadas las características de esta variable, que toma valores enteros, generalmente, en un rango pequeño, consideramos que la distribución Geométrica0 es la más adecuada. Por otra parte, es verosímil suponer que el proceso estocástico que describe la sequía posee una estructura de dependencia markoviana: dada la situación actual, el comportamiento futuro es independiente del pasado; en consecuencia, las observaciones de la duración restante en cada instante se pueden considerar independientes condicionalmente a la trayectoria observada del episodio, que se resume en un conjunto adecuado de covariables.

Esencialmente, el problema que se plantea es predecir una magnitud con distribución Geométrica0, cuyo valor medio es una función de variables dependientes del tiempo. Este problema se puede plantear en el marco de los modelos lineales generalizados.

6.1.1 Modelos lineales generalizados

Definición 6.1. *Un modelo lineal generalizado, GLM, es un modelo que verifica las siguientes hipótesis:*

- *Existe una variable respuesta, y, observada de forma independiente en los valores de un conjunto de covariables $x_1, \ldots, x_p$.*

- *El efecto de las covariables sobre la respuesta se manifiesta únicamente a través del predictor lineal $\eta = \beta_1 x_1 + \ldots + \beta_p x_p$.*

- *La distribución de y pertenece a la familia Exponencial, es decir, tiene una función de densidad o probabilidad de la forma,*

$$f(y_i; \theta_i, \sigma) = \exp\left[A_i \frac{y_i \theta_i - \gamma(\theta_i)}{\sigma} + \tau\left(y_i, \frac{\sigma}{A_i}\right)\right],$$

  donde i es el índice de la observación, σ un parámetro de escala, θ_i un parámetro función del predictor lineal y A_i un peso conocido a priori.

- *La media μ es una función invertible del predictor lineal, $\mu = m(\eta)$. La inversa de esta función m es la función enlace, l.*

McCullagh & Nelder (1989) desarrollan en detalle las propiedades de este tipo de modelos. La herramienta principal en la modelización de los GLM es la desviación escalada, $\mathcal{D}^*$, medida de discrepancia definida como el estadístico del test de razón de log-verosimilitudes que compara el modelo saturado -en el que se alcanza la máxima verosimilitud ajustando una media distinta para cada observación- y el modelo planteado,

$$\begin{aligned}\mathcal{D}^* &= 2[\ell\ell_{max}(\tilde{\theta}) - \ell\ell(\hat{\theta})] \\ &= \sum_{i=1}^{n} 2A_i \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - \gamma(\tilde{\theta}_i) + \gamma(\hat{\theta}_i)}{\sigma} \\ &= \frac{\mathcal{D}}{\sigma},\end{aligned}$$

siendo $\hat{\theta}_i$ y $\tilde{\theta}_i$ los MLE del modelo propuesto y del saturado, respectivamente, y $\mathcal{D}$ la desviación, no escalada, del modelo.

Definido un GLM, comprobaremos que el problema de predicción planteado se puede incluir en este marco. En efecto,

- Se tienen observaciones condicionalmente independientes de la variable respuesta, Lr, y de un conjunto de covariables que expresan el estado del proceso en cada instante seco de cada episodio de la muestra.

- La variable Lr tiene una distribución Geométrica0 con función de probabilidad,

$$P(Lr_i;\theta_i) = p_i(1-p_i)^{Lr_i} = \exp[Lr_i \ln(1-p_i) + \ln(p_i)]$$

  que pertenece a la familia Exponencial, con $A_i = 1$ para todo i, parámetro de escala $\sigma = 1$, $\theta_i = \ln(1-p_i)$, $\gamma(\theta_i) = -\ln[1-\exp(\theta_i)]$, y función τ nula.

- El predictor lineal puede tomar valores en todo el rango real; sin embargo, la media de la distribución Geométrica0, $(1-p)/p$, es un valor positivo, por lo que será necesario utilizar una función m que garantice esta propiedad; por ejemplo $\mu = \exp(\eta)$, es decir una función de enlace logarítmica $\eta = \ln(\mu)$.

6.1.2 Definición de las covariables

La única restricción que se debe tener en cuenta al definir las covariables es que su valor en cada instante se pueda determinar en tiempo real, y no sea necesario haber observado todo el episodio para conocerlo, como sucede, por ejemplo, con la intensidad máxima total.

Proceso de los periodos secos Se han considerado las siguientes covariables básicas.

intt: intensidad o diferencia entre el valor del umbral y la precipitación observada en el instante t.

deft: déficit al umbral acumulado desde el comienzo del episodio hasta t, es decir, la suma de las intensidades de cada una de las observaciones del periodo hasta ese instante.

durt: duración del episodio hasta t o número de observaciones anteriores consecutivas con valor inferior al umbral.

difint: diferencia entre la intensidad observada en t y la intensidad en $t-1$.

imaxt: intensidad máxima observada hasta el instante t en ese episodio.

estimax: estado respecto a imaxt; variable binaria que toma el valor 0 en las observaciones cuya intensidad coincide con la máxima observada hasta ese instante, estado creciente, y 1 si es menor, estado decreciente.

dimaxt: distancia a la posición de la intensidad máxima observada en el episodio hasta ese instante; en las observaciones en estado creciente la distancia es 0.

rec: valor del último récord observado en el episodio hasta ese instante, es decir, valor de la intensidad en el último máximo local observado.

estrec: estado respecto a rec; variable binaria que toma el valor 0 en las observaciones que son récord, y 1 en otro caso.

direc: distancia a la posición del último récord del episodio hasta ese instante; en las observaciones que son récord, la distancia es 0.

Entre las covariables definidas existen dos factores: estimax y estrec; dado que el efecto de las restantes covariables puede ser distinto en cada uno de los niveles definidos por estos factores, se definen las correspondientes interacciones.

- Interacciones con el factor estimax: inttei, deftei, durtei, difintei, imaxtei, recei. No es necesario considerar la interacción con dimaxt y con direc ya que coincide con las variables.

- Interacciones con el factor estrec: intter, defter, durter, difinter, imaxter, dimaxter, recer; la interacción con direc coincide con ella misma.

No es necesario considerar la interacción entre los dos factores, ya que sólo se pueden producir tres combinaciones de las variables (estimax, estrec): $(0,0)$, $(1,0)$ y $(1,1)$; por lo que, en presencia de término independiente, el término de interacción entre ambos es linealmente dependiente.

La relación entre la respuesta y las covariables puede ser no lineal; por ello se han considerado potencias hasta orden tres de todas las covariables e interacciones. En total se definen 65 covariables.

Proceso de las sequías La información que describe el estado de las sequías es más compleja: además de definir las mismas covariables que en los periodos secos -asociadas en este caso a las magnitudes de la sequía observadas hasta el instante t- se ha considerado de interés añadir información relativa, exclusivamente, al estado hasta ese instante del periodo seco en el que se realiza la predicción, que denotaremos periodo seco actual. Se definen las siguientes variables:

defpst: déficit acumulado en el periodo seco actual hasta el instante t.

durpst: duración del periodo seco actual hasta el instante t.

imaxpst: intensidad máxima observada en el periodo seco actual hasta t.

estimaxpst: estado respecto a imaxpst.

dimaxpst: distancia a la posición de imaxpst.

Las covariables intt, difint, y las asociadas al récord son variables locales, por lo que no es necesario definir sus versiones relativas al periodo seco actual.

Se debe considerar también la posible interacción entre las covariables y el estado de los factores definidos:

- Interacciones con el factor estimax: inttei, deftei, durtei, difintei, imaxtei, recei, defpstei, durpstei, imaxpstei; la correspondiente interacción con dimaxt, direc y con dimaxpst coincide con ellas mismas.
- Interacciones con el factor estimaxpst: intteips, defteips, durteips, difinteips, imaxteips, dimaxteips, receips, defpsteips, durpsteips, imaxpsteips; las interacciones con direc y con dimaxpst no son necesarias.
- Interacciones con el factor estrec: intter, defter, durter, difinter, imaxter, dimaxter, recer, defpster, durpster, imaxpster, dimaxpster; la interacción con direc coincide con la variable.

Los únicos valores posibles de (estimax, estimaxpst, estrec) son, $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 1)$, y $(1, 1, 1)$; por consiguiente, los parámetros del término independiente y de los tres factores son suficientes para caracterizar todas las situaciones posibles, no siendo necesario ningún término de interacción.

De nuevo, la relación entre la respuesta y las covariables puede ser no lineal; dado el elevado número de covariables y los resultados del análisis de periodos secos, se han considerado únicamente potencias hasta orden dos de todas las covariables e interacciones. En total se definen 87 covariables.

Efecto del número de periodos secos Otra variable que puede ser útil en la predicción es **nper**, el número de periodos secos observados en la sequía hasta el instante de predicción. Para modelizar el efecto de esta variable se define un número de términos polinómicos igual a $n_c - 1$, siendo n_c el máximo muestral del número de episodios por cluster. Los términos polinómicos se parametrizan de forma ortogonal y se denotan f1, f2, ..., de acuerdo al orden de la potencia. Dado el carácter discreto de esta variable, la inclusión de estos términos es equivalente a introducir un factor con n_c categorías. La parametrización en forma de factor plantearía problemas al

predecir desde un instante en el que el número de periodos secos fuera mayor que el valor n_c observado en la muestra de estimación, mientras que con la parametrización polinómica, la predicción se interpreta como la proyección a un valor mayor de una relación establecida en el modelo.

La influencia de esta variable se puede manifestar individualmente y a través de otras covariables. Dado el elevado número de éstas, nper se ha interaccionado con las variables individuales, pero no con la interacciones con los factores binarios. A pesar de esta limitación, se definen 171 variables.

6.1.3 Selección de covariables

Dado el elevado número de covariables definidas es imprescindible disponer de un procedimiento que permita seleccionar las de mayor poder predictivo. Con este objetivo se ha utilizado un algoritmo tipo 'paso a paso' basado en el test de razón de verosimilitudes. En este algoritmo, a partir de un modelo inicial, se aplican alternativamente un paso de entrada y otro de salida hasta alcanzar un modelo en equilibrio, en el que todos los términos incluidos en el predictor lineal, y sólo ellos, son significativos. En un paso de entrada se decide la introducción de una nueva covariable comparando el modelo ajustado en ese instante con cada uno de los obtenidos al añadir una de las variables no incluidas; en un paso de salida se decide la exclusión de un término comparando el modelo actual con cada uno de los obtenidos al eliminar una de las variables ya incluidas.

El test utilizado como criterio de entrada-salida contrasta la hipótesis de que un modelo Mr con q parámetros, llamado modelo reducido, es un modelo suficiente respecto a un modelo más general, Mg, con p parámetros, $p > q$, en el que Mr está anidado. El estadístico del test es,

$$\Lambda = \mathcal{D}^*_{Mr} - \mathcal{D}^*_{Mg} = 2[\ell\ell_{Mg}(\hat{\theta}_g) - \ell\ell_{Mr}(\hat{\theta}_r)],$$

y, bajo la hipótesis nula, tiene una distribución aproximada χ^2 con $p - q$ grados de libertad.

En el proceso de selección de covariables, además de aplicar el test de razón de verosimilitudes, se comparan dos medidas de cada modelo M,

- El criterio AIC propuesto por Chambers & Hastie (1992),

$$AIC_{CH} = \mathcal{D}_M + 2p\sigma,$$

que en este caso, dado que $\sigma = 1$, es equivalente al criterio AIC de Akaike,

$$AIC = -2 * \ell\ell_M + 2p.$$

- El criterio BIC,

$$BIC = \mathcal{D}_M + \ln(n)p\sigma,$$

donde n es el tamaño de la muestra, que penaliza el número de parámetros más que el AIC y proporciona modelos más sencillos.

El algoritmo paso a paso requiere un modelo inicial como punto de partida. En general se utiliza el modelo nulo, que sólo tiene término independiente, pero se puede comenzar con un modelo lleno, que contenga todas las variables que puedan ser influyentes, o con un subconjunto de ellas que se crea relevante.

Esencialmente el objetivo del proceso de selección será buscar modelos que presenten valores próximos al mínimo en los criterios AIC y BIC. En la selección final entre los mejores modelos ajustados se considera un criterio adicional, la sencillez e interpretabilidad del modelo.

Control del algoritmo de búsqueda

Para comprobar si los modelos resultantes del algoritmo paso a paso presentan valores próximos al mínimo AIC y BIC global se propone un procedimiento basado en el análisis de componentes principales. Las componentes principales -combinaciones lineales de las variables originales, ortogonales entre sí- proporcionan un método de inspección de las direcciones de mayor variabilidad del espacio generado por las covariables, que permite reducir la dimensión del problema de estimación. Esta reducción de la dimensión se obtiene suponiendo que dependencias lineales 'casi exactas' entre las covariables, son exactas. Al eliminar las componentes cuya varianza es pequeña, se pierde poca información y se conserva un alto porcentaje de la variabilidad del espacio muestral original. La imposición de las restricciones lineales

son específicas de la muestra pero tienen algunas propiedades de máxima varianza (Greenberg 1975).

Generalmente, el objetivo de la aplicación de esta técnica en regresión es la estimación de los parámetros eliminando los sesgos provocados por la existencia de colinealidad entre las covariables; Judge et al. (1985) señalan, sin embargo, que los estimadores obtenidos con este procedimiento no tienen, en general, mejores propiedades que los de mínimos cuadrados obtenidos en el espacio original.

En este caso, los modelos basados en componentes principales se emplean como modelos de referencia para verificar la bondad de ajuste de los modelos propuestos, es decir, para medir la proximidad del espacio generado por las covariables de los modelos propuestos al correspondiente a todas las covariables definidas o, más exactamente, a la aproximación del mismo generada por un número de componentes principales suficiente. Se considera un número suficiente al que explica al menos el 90% de la variabilidad del espacio original. Dado que el rango de valores de las covariables originales es muy diferente y con el fin de que todas ellas tengan el mismo peso en el análisis, se calculan las componentes principales del espacio generado por las variables divididas por su desviación típica.

6.1.4 Predicción de la duración restante de un episodio

Ajustado el modelo y conocido el valor de las covariables en un instante l_0, el valor medio estimado de la duración restante se obtiene directamente,

$$E(Lr_{l_0}) = \exp[\nu(l_0)],$$

donde $\nu(l_0)$ es el predictor lineal del modelo evaluado en el instante l_0.

La mediana, o cualquier otro percentil, se obtienen a partir del parámetro p de la distribución,

$$p = \frac{1}{E(Lr_{l_0}) + 1},$$

y de la expresión del percentil de orden per,

$$P_{per}(Lr_{lo}) = \left[\frac{\ln(1 - per)}{\ln(1 - p)}\right]_{entera}.$$

6.1.5 Comparación y bondad de ajuste de los modelos

Puesto que el principal objetivo de los modelos desarrollados es la predicción, los criterios de comparación se basarán en la calidad de la misma. La forma más sencilla de evaluar la capacidad predictiva de un modelo es predecir observaciones cuyo valor real se conoce, y comparar los valores observados con los predichos; para evitar sesgos, conviene realizar las predicciones de comprobación utilizando métodos de validación cruzada, es decir, basados en la predicción de observaciones que no se hayan utilizado en el proceso de estimación. Dos posibles alternativas son:

- Separar la muestra disponible en dos partes: utilizar una para estimar los parámetros y otra para validar el modelo evaluando sus predicciones. Este procedimiento garantiza una valoración del modelo en condiciones reales de aplicación, pero sólo es aplicable si se dispone de series suficientemente largas.
- Predecir cada elemento de la muestra, en este caso cada episodio, a partir de un modelo cuyos parámetros se estimen eliminando de la muestra ese elemento; con este procedimiento se evitan los sesgos sin prácticamente reducir el tamaño de muestra. Los residuos correspondientes a este tipo de predicción se denotarán p-residuos.

Comparación de los modelos A continuación se proponen un conjunto de medidas que cuantifican el error cometido en la predicción. Su mayor inconveniente es la ausencia de valores de referencia absolutos y de resultados que permitan validar un modelo de forma individual; por el contrario, proporcionan criterios eficientes para comparar modelos alternativos de forma rápida y sencilla.

Como medida de error básica se considera un p-residuo de carácter general, el residuo de la respuesta,

$$e_i = y_i - \hat{y}_i.$$

A partir de ellos, se definen medidas de error individuales, asociadas a cada observación:

$$\begin{aligned} e1_i &= |y_i - \hat{y}_i| \\ e2_i &= (y_i - \hat{y}_i)^2, \end{aligned}$$

y una medida de carácter relativo, en la que se considera como valor de referencia $y_i + 1$,

$$er_i = \left| \frac{e_i}{y_i + 1} \right|.$$

Generalmente, el número de observaciones es elevado y no resulta fácil comparar directamente los errores individuales correspondientes a dos modelos. Para evitar esta dificultad se definen medidas de error relativas a cada episodio y medidas globales de toda la muestra. Se consideran medidas absolutas:

$$\begin{aligned} m1 &= \sum_{j=1}^{l_i} e1_j \\ m2 &= \max_{1 \le j \le l_i} e1_j \\ m3 &= \sum_{j=1}^{l_i} e2_j \end{aligned}$$

y relativas:

$$\begin{aligned} mr1 &= \sum_{j=1}^{l_i} er_j \\ mr2 &= \max_{1 \le j \le l_i} er_j, \end{aligned}$$

siendo l_i la duración del episodio i-ésimo. Como medidas globales se utilizan la suma, en todos los elementos de la muestra, de cada una de las cinco medidas propuestas.

Como herramienta complementaria se proponen métodos gráficos que facilitan la comparación de las medidas individuales.

- Gráfico de la sucesión de residuos. Se representa la serie temporal de los p-residuos en cada instante de cada episodio, señalando las divisiones entre ellos. Este gráfico, figura 6.1, además de proporcionar una idea global del error, permite determinar las observaciones que presentan mejor y peor ajuste y analizar la bondad del modelo según la longitud del episodio y el instante de predicción dentro del mismo.

- Gráfico de $e1$. Gráfico en el que se representa la serie temporal de la medida $e1$. El área bajo esta curva coincide con la medida global asociada a $m1$. Si se utilizan residuos relativos, el área bajo la curva coincide con la correspondiente a $mr1$.

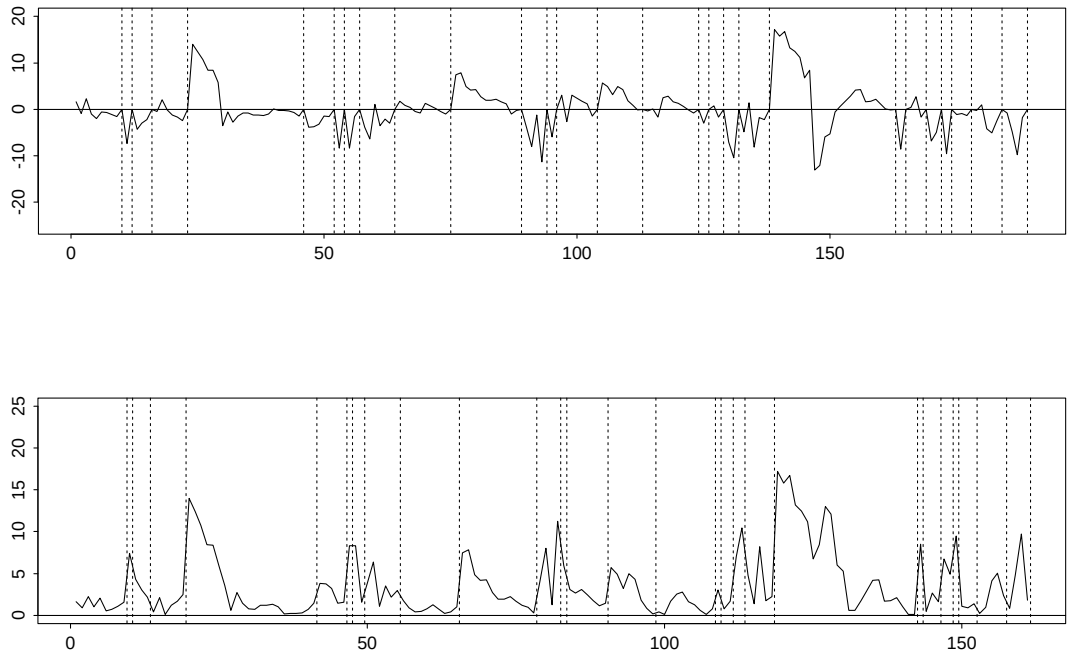

Figura 6.1: Gráfico de los residuos (sup.) y de los residuos absolutos (inf.) correspondientes al valor esperado del ajuste de la duración restante, Mseq*3, seq10; Huesca.

Bondad de ajuste Las medidas de bondad de ajuste utilizadas son:

- Desviación. La desviación es una medida de bondad de ajuste equivalente, en muchos sentidos, a la suma de residuos al cuadrado en el modelo lineal; en particular, si el término de error es Normal y el enlace la identidad, coincide con ese valor. En esas condiciones, la distribución de la desviación escalada es χ^2_{n-p}, siendo p el número de parámetros ajustado. Con otras distribuciones de error el resultado anterior es asintótico y no se conoce bien su comportamiento cuando el tamaño de muestra es pequeño. En el caso de distribuciones discretas la aproximación presenta problemas suplementarios, ya que la distribución exacta de la desviación es discontinua.

- Estadístico de Pearson generalizado. El estadístico de bondad de ajuste de Pearson es,

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var}(y_i)}.$$

  Si la distribución del error es Normal, su distribución es χ^2_{n-p}; con el resto de distribuciones, este resultado sólo es válido asintóticamente.

Pierce & Schafer (1986) establecen que la desviación es una medida de bondad de ajuste más adecuada, a pesar de que el estadístico X^2 presenta una mejor aproximación a la distribución χ^2. La desviación, además, permite comparar modelos anidados.

- Control del parámetro de escala. Un control sencillo para comprobar la existencia de sobredispersión en el modelo consiste en comparar el valor estimado del parámetro de escala,
$$\hat{\sigma} = \frac{\mathcal{D}_M}{(n-p)}$$
con, 1, su valor en la distribución Geométrica0.

El proceso de validación termina con un análisis de residuos. Los GLM requieren un concepto de residuo general, que sea aplicable con todas sus posibles distribuciones de error; los más utilizados son:

- Residuos de la desviación. Estos residuos se pueden interpretar como la contribución de cada observación a la desviación:
$$rd_i = sgn(y_i - \hat{\mu}_i)\sqrt{2A_i\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - \gamma(\tilde{\theta}_i) + \gamma(\hat{\theta}_i)\right]}.$$

- Residuos de Pearson. Son una versión estandarizada de los residuos respuesta,
$$rP_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}_i}};$$
y, salvo el signo, coinciden con la raíz cuadrada del sumando correspondiente en el estadístico X^2 de Pearson.

Pierce & Schafer (1986) analizan y comparan estos dos tipos de residuos en modelos con errores no Normales, tanto discretos como continuos, y concluyen que los residuos de la desviación son mejores, ya que presentan un comportamiento más próximo a la Normal que los de Pearson.

En el caso de la distribución Geométrica la expresión de los residuos de la desviación es,
$$rd_i = sgn\left[y_i - \left(\frac{1}{\hat{p}_i} - 1\right)\right]\sqrt{2\left(y_i \ln\left[\frac{y_i}{(1+y_i)(1-p_i)}\right] + \ln\left[\frac{1}{p_i(1+y_i)}\right]\right)}.$$

Varios autores, como Davison & Snell (1991) o Fahrmeir & Tutz (1994), aconsejan utilizar los residuos estandarizados de la desviación, que se obtienen dividiendo por una aproximación de su error estándar,

$$rds_i = \frac{rd_i}{\sqrt{1 - h_{ii}}}$$

donde h_{ii} son los elementos de la diagonal de la matriz de potencial H,

$$H = W^{1/2} X (X'WX)^{-1} X' W^{1/2},$$

con W, la matriz diagonal de pesos correspondiente a la última iteración del algoritmo de estimación. Los elementos h_{ii} constituyen una medida de la influencia que el valor de las covariables de cada observación ejerce en el ajuste global, aunque, a diferencia del modelo lineal clásico, la matriz H no depende únicamente de la matriz de diseño sino también del ajuste.

El análisis de residuos se basa principalmente en controles de tipo gráfico: representación de la serie de residuos y de los valores h_{ii}, de los residuos frente al valor ajustado y frente al instante de predicción en el episodio. Para comprobar la validez de las hipótesis relacionadas con la distribución de la respuesta se analiza el carácter Normal de los residuos utilizando un qqplot. Con el mismo objetivo se estudia el carácter exponencial de los residuos de máxima verosimilitud, definidos en el apartado 5.1.3, utilizando un qqplot y un gráfico lineal de supervivencia.

Se analiza también la existencia de correlación entre las observaciones mediante los tests habituales, un gráfico de los residuos frente a los residuos retardados y un análisis del correlograma. El modelo planteado supone que, dada la trayectoria pasada, las observaciones de la respuesta son independientes. Si las covariables del modelo no son suficientes para resumir esa trayectoria, existirá correlación entre los residuos de las observaciones pertenecientes al mismo episodio.

6.1.6 Análisis de resultados

En este apartado se presentan los modelos de predicción para el proceso de periodos secos y de sequías, definidos con el percentil décimo en la serie de Huesca. En primer lugar se realiza un proceso de selección de covariables con el procedimiento paso a

paso descrito, que proporciona varios modelos competitivos; se compara la bondad de predicción de los modelos obtenidos utilizando los gráficos y medidas indicadas, y se elige el más adecuado. Finalmente, se efectúa un análisis de validación del mejor modelo para episodios de sequía, basado en los residuos estandarizados de la desviación y los de máxima verosimilitud.

El modelo resultante del proceso de selección depende de la estrategia de modelización: modelo inicial utilizado, orden de entrada de las variables, criterios de selección de entrada-salida de las variables, etc. Se han ensayado distintos procedimientos con el objeto de compararlos y establecer criterios de modelización.

Los resultados obtenidos tras algunos ensayos confirman una propiedad conocida: la minimización del AIC produce modelos complejos, mientras que la del BIC proporciona modelos con menos covariables que en este caso han resultado, en general, insuficientes. Intentando evitar estos inconvenientes se ha establecido un procedimiento, que denominaremos AIC-BIC combinado, consistente en:

- Aplicar el procedimiento paso a paso buscando minimizar el AIC; es decir, se introduce o elimina una variable, si el AIC del modelo resultante es menor.
- Sobre el modelo obtenido -que contendrá todas las variables relevantes entre las definidas- se aplica de nuevo el algoritmo de entrada-salida, buscando ahora la minimización del BIC.

Periodos secos Algunos de los mejores modelos encontrados en el proceso de selección son:

Mps1: intt, estimaxt, dimaxt, deftei.

Mps1b: intt, estimaxt, dimaxt, difint2, deftei.

Mps2: deft, estimaxt, dimaxt.

Mps2b: imaxt, estimaxt, dimaxt, difint2, imaxt2ei, deft3ei, difint3ei.

Mps3: imaxt, dimaxt, difinter, deft2.

Mps3b: deft, difint, imaxt, estimaxt, dimaxt, durtei, difinter, deft2, durt2ei.

	$\mathcal{D}_M$	AIC	BIC	$\sum m1$	$\sum m2$	$\sum m3$	$\sum m1^*$	$\sum m2^*$	$\sum m3^*$
Mps1	159.3	169.3	184.7	121.0	186.9	486.1	86.2	146.0	317.6
				87.0	107.1		55.4	70.8	
Mps1b	155.0	167.0	185.5	117.1	180.7	467.0	83.0	144.0	301.6
				85.6	108.3		53.5	69.9	
Mps2	160.9	168.9	181.3	120.6	186.8	487.2	85.2	136.0	303.2
				88.5	108.4		55.5	69.3	
Mps2b	150.9	166.9	191.6	116.2	186.7	467.0	84.0	144.0	299.6
				85.0	109.2		54.0	70.4	
Mps3	156.7	168.7	182.2	114.9	181.3	449.2	87.1	146.0	311.7
				82.8	103.4		57.4	72.6	
Mps3b	148.8	166.8	199.7	114.9	185.3	506.6	82.5	147.0	314.7
				82.8	109.1		52.4	70.3	

Tabla 6.1: Medidas de bondad de ajuste de los modelos para la duración restante, ps10; Huesca.

Los modelos Mps1b, Mps2b y Mps3b son los de mínimo AIC resultantes del primer paso del algoritmo propuesto, partiendo de distintos modelos iniciales, y Mps1, Mps2 y Mps3 son los obtenidos, en cada caso, al final de la segunda fase. En la tabla 6.1 se muestran los valores de las medidas de bondad de ajuste de estos modelos: la desviación, el AIC, el BIC y las medidas de error global -absoluto en la primera línea, y relativo en la segunda- correspondientes a las predicciones basadas en la media, m, y la mediana, m^*.

Aunque los modelos seleccionados a partir de diferentes modelos iniciales no coinciden, se observa que algunas covariables aparecen de forma constante en todos ellos, bien en su forma original o a través de variables de la misma familia, las correspondientes a sus potencias o sus interacciones. En particular, en los modelos tipo b: deft, estimaxt, dimaxt y una variable de carácter local, difint; en los modelos más simples se mantienen deft, dimaxt, y es frecuente estimaxt.

Como se observa en la figura 6.2 y en la tabla 6.1, las predicciones de los modelos b son similares a las de sus correspondientes modelos finales. El modelo Mps2, el de mínimo BIC, presenta residuos más grandes que Mps1 y Mps3 en algún punto concreto, figura 6.3. Los resultados de predicción de Mps1 y Mps3 no presentan diferencias notables y es difícil establecer la superioridad de uno sobre el otro.

En el espacio generado por las 65 variables definidas, las 16 primeras componentes principales explican el 90% de la variabilidad total. La desviación del modelo que

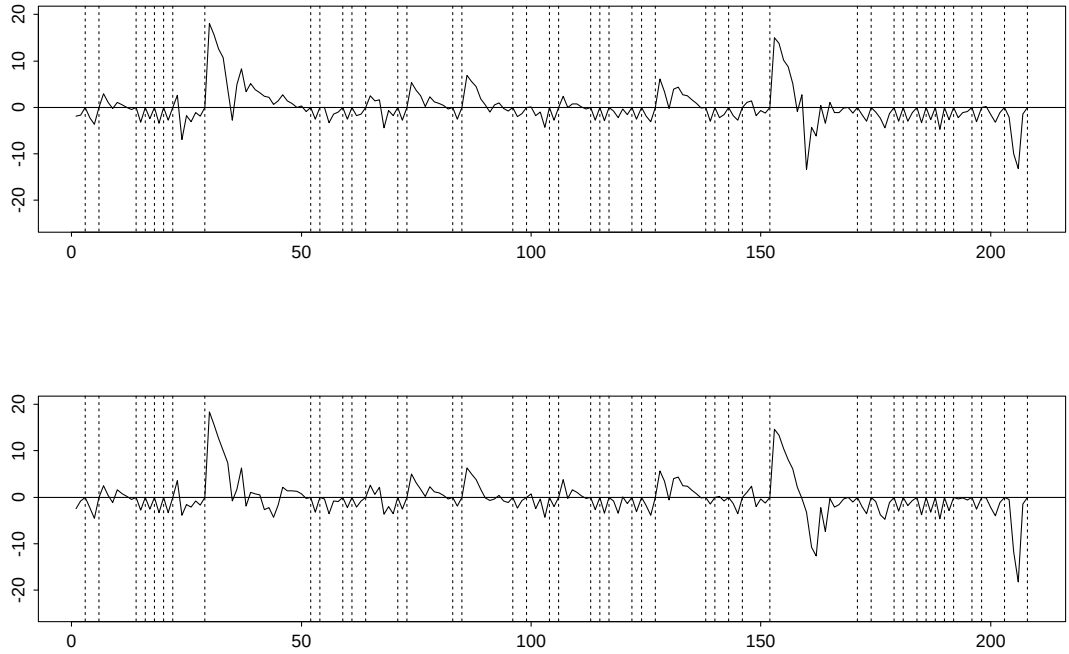

Figura 6.2: Gráfico de los residuos de los modelos Mps3 (sup.) y Mps3b (inf.), predicción con el valor esperado, ps10; Huesca.

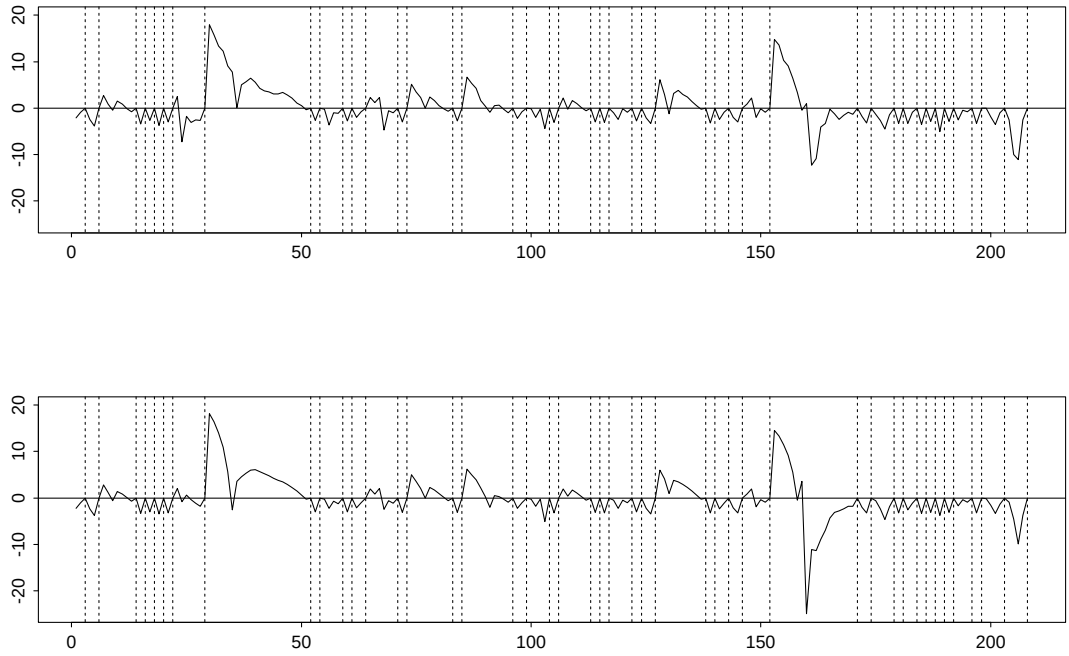

Figura 6.3: Gráfico de los residuos de los modelos Mps1 (sup.) y Mps2 (inf.), predicción con el valor esperado, ps10; Huesca.

	$\mathcal{D}_M$	AIC	BIC	$\sum m1$	$\sum m2$	$\sum m3$	$\sum m1^*$	$\sum m2^*$	$\sum m3^*$
Mseq1	131.6	139.6	152.0	101.4 63.2	188.0 99.6	581.5	81.5 43.2	162.0 65.9	417.9
Mseq1b	120.7	134.7	156.2	101.6 63.4	176.7 99.0	597.4	79.4 42.4	155.0 65.6	410.9
Mseq2	123.3	135.3	153.8	105.5 67.3	175.8 102.3	619.4	80.3 44.8	150.0 67.6	420.7
Mseq2b	117.8	133.8	158.5	108.4 67.0	182.5 98.1	665.6	80.4 43.3	152.0 63.8	428.5
Mseq*3	114.3	130.3	149.9	115.5 74.2	179.7 105.8	783.4	89.0 50.1	152.0 71.6	491.9
Mseq*3b	110.5	128.5	161.3	109.3 69.2	182.3 101.2	710.5	85.1 47.6	154.0 70.1	465.8
Mseq*4	117.3	129.3	147.8	107.7 68.2	173.3 102.7	663.0	83.7 47.5	151.0 71.3	450.6
Mseq*4b	108.3	124.3	148.9	103.7 64.4	183.0 98.2	662.2	83.1 44.5	156.0 67.0	441.9

Tabla 6.2: Medidas de bondad de ajuste de los modelos para la duración restante, seq10; Huesca.

tiene por covariables esas componentes principales es 153.6.

Episodios de sequía Dado el elevado número de posibles predictores, el algoritmo de selección se aplicó, en primer lugar, al conjunto de covariables sin considerar la familia nper. Posteriormente, se realizó un nuevo proceso de selección a partir de las 171 variables definidas, utilizando la información obtenida en esa primera selección y en el proceso de modelización de periodos secos para elegir los modelos iniciales. A continuación, se muestran algunos de los mejores modelos obtenidos en el primer proceso de selección, modelos Mseq, y en el segundo, que incluye a la familia nper, modelos Mseq*; las medidas de bondad de ajuste de cada uno de ellos se muestran en la tabla 6.2.

Mseq1: estimaxt, defpst, dimaxteips.

Mseq1b: estimaxt, defpst, difint2, recei, defpstei, dimaxteips.

Mseq2: dimaxt, estimaxpst, difint2, deftei, durtei.

Mseq2b: intt, dimaxt, estimaxpst, difint2, deftei, durtei, receips.

Mseq*3: dimaxt, estimaxpst, difint2, deftei, f1durt, f1durt2.

Mseq*3b: deft, dimaxt, estimaxpst, difint2, deftei, rec2ei, f1int, f1durt, f1durt2.

Mseq*4: dimaxt, estimaxpst, difint2, deftei, f1imaxt.

Mseq*4b: durt, dimaxt, estimaxpst, difint2, deftei, f1imaxt, f1imaxt2.

En todos los modelos de tipo b se encuentran presentes las covariables estimaxpst o estimax, dimaxt o dimaxteips, deftei o defpstei, y la variable local difint2; en los mejores modelos Mseq* ensayados aparece, además, alguna interacción del término lineal de la variable nper, denotadas f1. Estas covariables se mantienen en los modelos finales. Cabe señalar que en todos los modelos aparece alguna covariable referente a toda la sequía y alguna relativa al periodo seco actual.

Las predicciones de los modelos tipo b no son notablemente mejores que las de los modelos finales. Globalmente, el modelo de mínimo BIC de los ensayados es Mseq*4, y Mseq*4b el de mínima desviación y mínimo AIC, aunque las predicciones de todos ellos, en particular las de Mseq*3 y Mseq*4, son similares. En la figura 6.4, se pueden comparar las predicciones de Mseq1, el modelo más sencillo, con las de Mseq*4.

En el espacio generado por las 87 variables asociadas a los episodios de sequía, 19 componentes principales explican el 90% de la variabilidad total; considerando las variables de la familia nper, son necesarias 27 componentes para explicar el mismo porcentaje en el espacio generado por las 171 covariables. La desviación de los modelos de componentes principales es 114.7 y 108.0, respectivamente, valores próximos a los alcanzados en el proceso de modelización.

Los resultados más importantes del ajuste y selección de modelos de predicción para el proceso de periodo secos y sequías se resumen en las siguientes conclusiones.

- El grado de significación de las covariables durante el proceso de selección es muy sensible a las covariables presentes en el modelo; la causa de este comportamiento es la relación, o incluso coincidencia en algún caso, entre el valor de las covariables en algunas observaciones. Como consecuencia, el modelo con el que se alcanza el equilibrio depende en gran medida del modelo inicial.

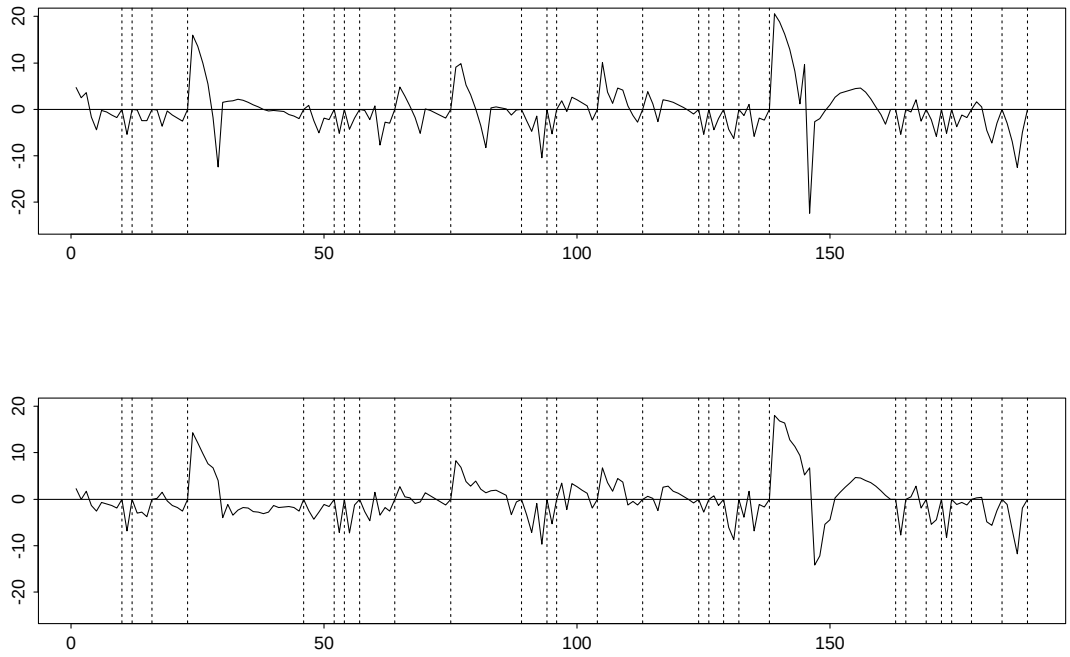

Figura 6.4: Gráfico de los residuos de los modelos Mseq1 (sup.) y Mseq*4 (inf.), predicción con el valor esperado, seq10; Huesca.

- Se pueden estimar modelos equivalentes, en cuanto a ajuste y predicción, con covariables distintas. Este efecto es consecuencia de la conclusión anterior, en concreto de la existencia de covariables, o subconjuntos de ellas, que aportan información muy parecida.

- A pesar del comportamiento señalado en los dos puntos anteriores, existe un conjunto de covariables cuya información parece ser necesaria, tanto en los modelos de periodos secos, como en los de sequías: dimaxt, estimaxt, difint y alguna de la familia deft.

  En el caso de las sequías esta información se complementa con un término que incluye el efecto lineal de nper. No son necesarios términos de esta variable de mayor orden.

  Es suficiente definir términos polinómicos de orden dos; los mejores modelos no incluyen términos de orden superior.

  La información que aportan las magnitudes asociadas a toda la sequía y las relativas al periodo seco actual es complementaria; los modelos seleccionados incluyen covariables de ambos tipos que, en general, no corresponden a la

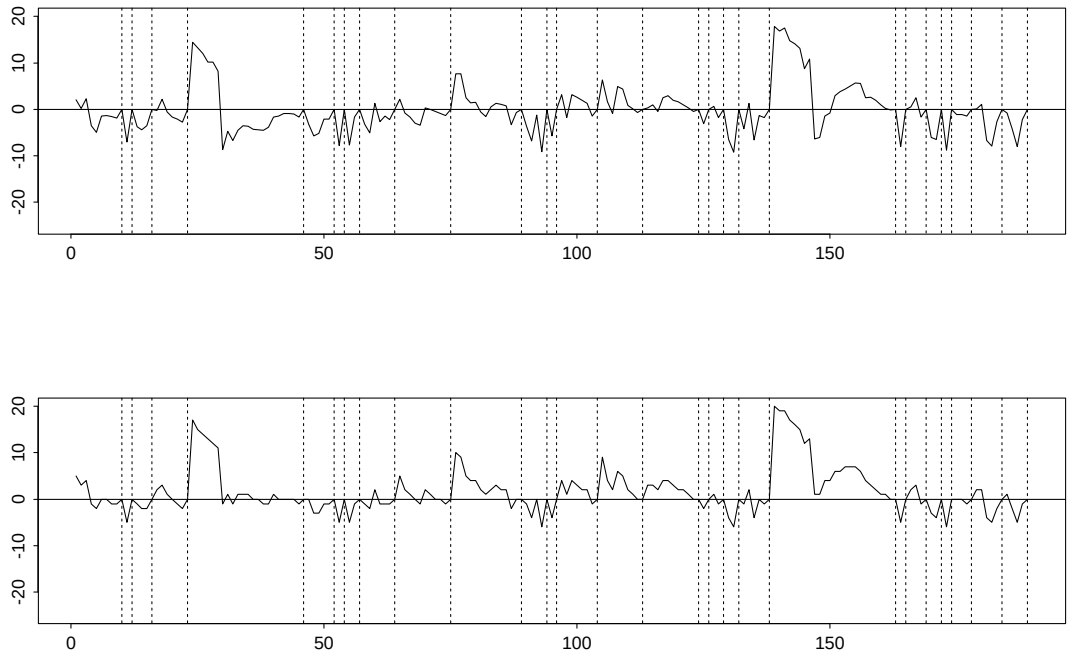

Figura 6.5: Gráfico de los residuos del modelo Mseq2, predicción con el valor esperado (sup.) y con la mediana (inf.), seq10; Huesca.

misma magnitud.

- La capacidad predictiva de los modelos depende del instante de predicción. Las predicciones en los primeros instantes de cada episodio son valores próximos a la media, debido a que las características iniciales de todos ellos, largos o cortos, son similares y, en consecuencia, el modelo no dispone de información suficiente que le permita discriminar. Con cada instante observado se acumula información, de forma que en los periodos secos más largos las predicciones mejoran progresivamente, como se aprecia en el gráfico de residuos.

  Este efecto es consecuencia de la información utilizada, relativa únicamente al propio episodio, y se manifiesta en todos los modelos seleccionados, que proporcionan las mejores y peores predicciones en los mismos instantes.

- Las series de los p-residuos correspondientes a las predicciones con el valor medio y con la mediana presentan un perfil similar en todos los modelos, figura 6.5, aunque las correspondientes a la mediana tienen una forma más suave y menores medidas de error.

- En general, los modelos obtenidos con el criterio AIC-BIC combinado, más sencillos que los obtenidos al minimizar el AIC, proporcionan predicciones muy próximas a las de éstos; la calidad de los ajustes y predicciones de los modelos de mínimo BIC es menor.

- Los modelos seleccionados presentan valores de la desviación próximos a los mínimos alcanzables en el espacio de covariables considerado.

- Los valores estimados del parámetro de escala en los modelos propuestos oscilan entre 0.8 y 1, indicando que no existe sobredispersión en los modelos.

Respecto a la metodología empleada se pueden señalar las siguientes observaciones.

- El número de covariables seleccionado está directamente relacionado con el criterio de entrada-salida de variables. El criterio de minimización del AIC proporciona resultados similares a la utilización del test de razón de verosimilitudes con un nivel de significación entre el 90 y el 95%. El procedimiento de selección de covariables AIC-BIC combinado proporciona modelos suficientes y no sobreparametrizados.

- La predicción con la mediana resulta menos sensible a valores particulares de la muestra. Sus medidas de error son algo mejores que las de la predicción con la media. La justificación de este hecho se basa en que el valor de la mediana en la distribución Geométrica, por ser asimétrica positiva, es siempre menor que el de la media; en consecuencia, proporciona mejores predicciones en las observaciones de menor valor, que son las más frecuentes.

- El gráfico de $e1$ proporciona una idea global de la calidad de predicción que permite comparar distintos modelos. El gráfico de la serie de residuos separados por episodios permite analizar la bondad de predicción en función de la posición de la observación dentro del episodio y comparar, de forma rápida, las predicciones de los modelos.

- El análisis gráfico de los residuos relativos es de menor interés ya que, en los instantes iniciales de los periodos largos, enmascaran las malas predicciones.

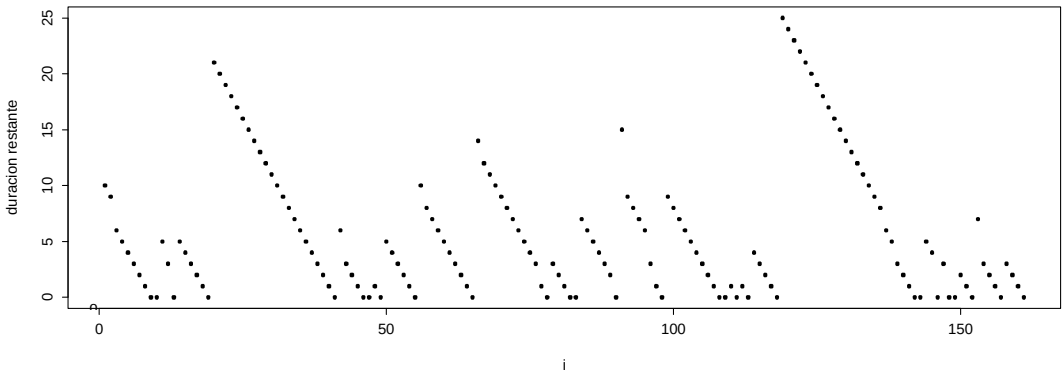

Figura 6.6: Gráfico de la serie de valores observados, seq10; Huesca.

Validación del modelo Mseq*4 En la figura 6.6 se representa la serie de observaciones de la variable respuesta, con el fin de mostrar la agrupación en periodos secos y episodios de la misma. El gráfico de la figura 6.7 corresponde a la serie de residuos estandarizados de la desviación. La hipótesis sobre la distribución de la respuesta, Geométrica, se ve confirmada por el carácter exponencial de los residuos de máxima verosimilitud, figura 6.7, y el carácter normal de los residuos de la desviación. En el gráfico de los residuos frente al instante de predicción, figura 6.8, se observa que en los primeros instantes de cada episodio la predicción es menos fiable, y cómo mejora progresivamente en el transcurso de los mismos.

No se detecta ninguna anomalía importante, excepto en el gráfico de correlación de los residuos, donde se observa en la zona de los residuos positivos

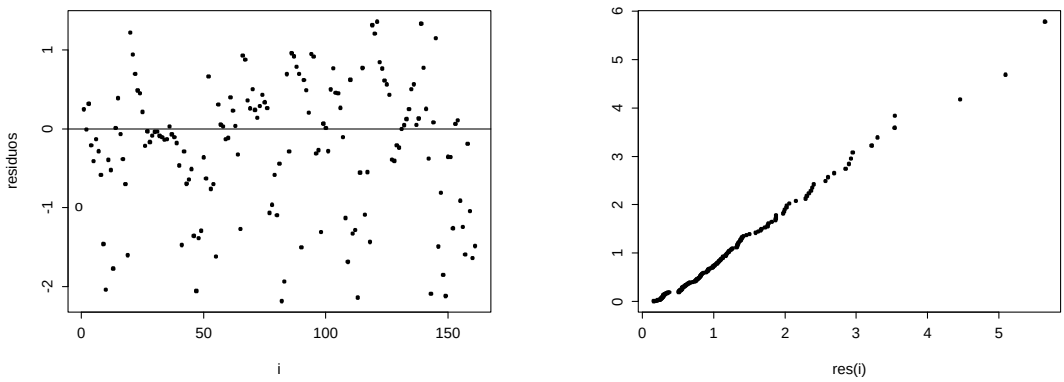

Figura 6.7: Gráfico de la serie de residuos estandarizados de la desviación (izda.) y qqplot Exponencial de los residuos de máxima verosimilitud (dcha.), Mseq*4 seq10; Huesca.

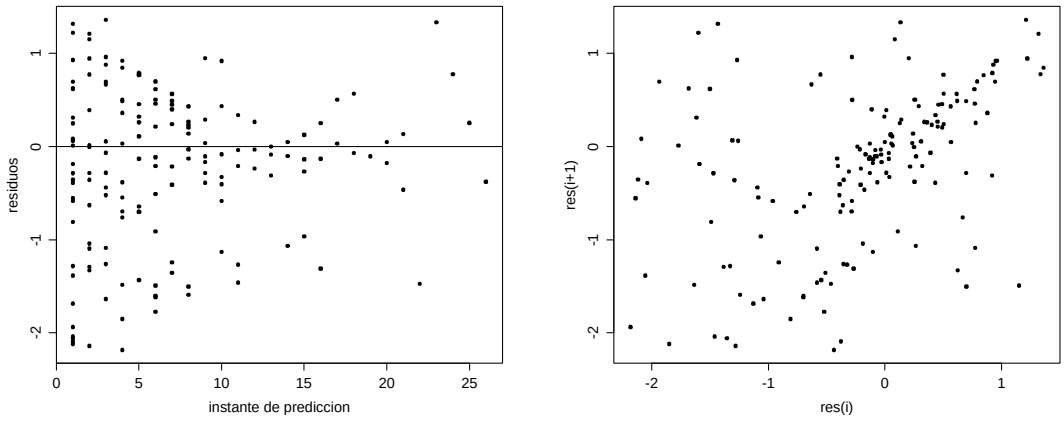

Figura 6.8: Gráfico de residuos frente al instante de predicción (izda.) y gráfico de correlación de los residuos (dcha.), Mseq*4 seq10; Huesca.

un conjunto de observaciones con clara correlación positiva, figura 6.8. En la figura 6.9 se representan únicamente los pares de observaciones pertenecientes al mismo episodio y, como cabía esperar, son algunas de éstas las que presentan correlación. Con el fin de determinar el origen de esta correlación, se realiza un análisis detallado de estos residuos en función de distintos factores como, el instante de ocurrencia, la magnitud de la respuesta y la duración total de los episodios. En el segundo gráfico de la figura anterior se muestran los pares de residuos correspondientes a las sequías de duración menor o igual que 9, grupo A; las observaciones de los episodios de mayor longitud se separan, a su vez, en dos grupos, en función de que la duración restante en ese instante sea menor, o mayor o igual que 8, grupos B y C respectivamente. Como se puede observar en estos gráficos, son las observaciones del grupo C, que corresponden al periodo inicial de los episodios más largos las que presentan correlación significativa, con un coeficiente de correlación de Kendall de 0.45, al que corresponde un p-valor de 0.00; por el contrario, las observaciones en la etapa final de los mismos episodios, grupo B, presentan la correlación menor, 0.13, con un p-valor de 0.21. En el grupo A, la correlación es 0.15 y tampoco resulta significativa, con un p-valor de 0.14.

De estos resultados se concluye que los dos factores que originan la correlación existente en el grupo C son:

- El peor ajuste del modelo en los primeros instantes de los episodios, de-

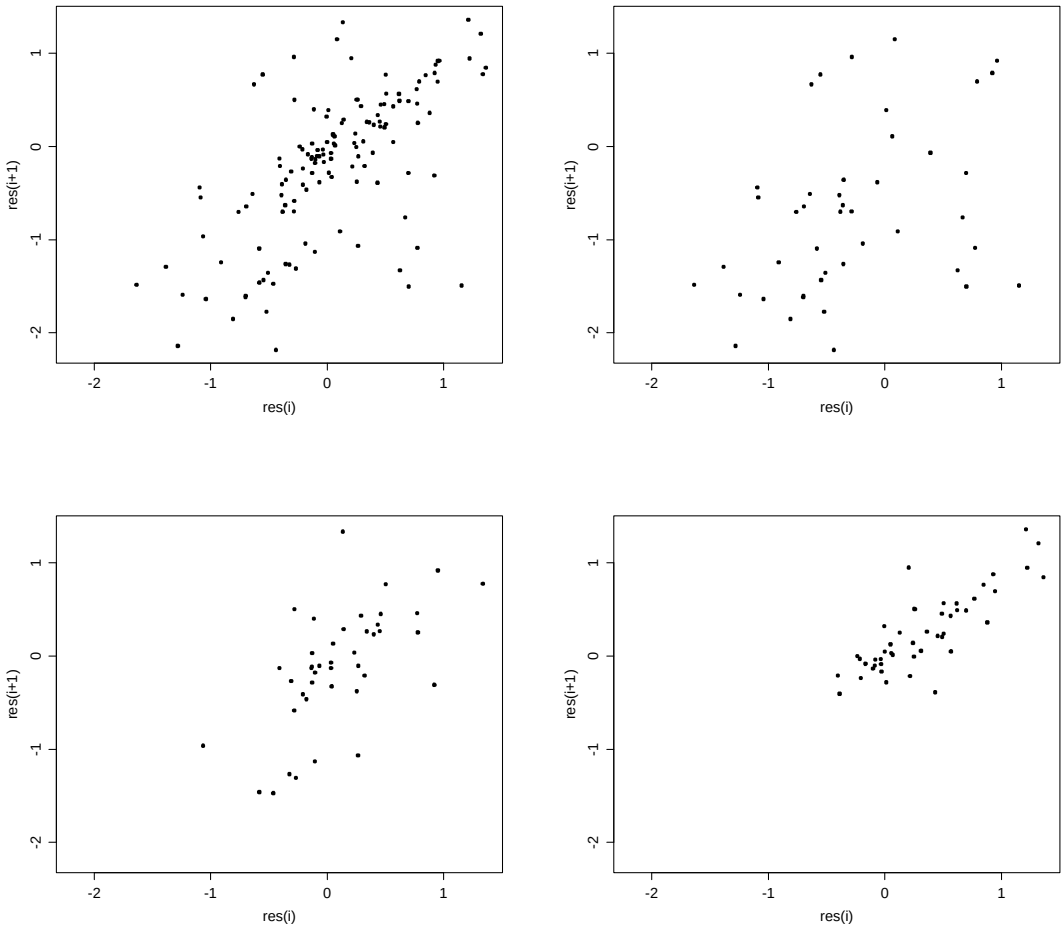

Figura 6.9: Gráfico de correlación de los residuos de las observaciones pertenecientes al mismo episodio (sup.izda.), del grupo A (sup.dcha.), B(inf.izda.) y C (inf.dcha.), Mseq*4 seq10; Huesca.

bido al tipo de covariables que se han considerado, relativas al propio episodio. En estos modelos, la información disponible en los primeros instantes no parece suficiente para garantizar la independencia condicional.

- La existencia de algún episodio de gran longitud. En la muestra existen dos sequías cuya duración es notablemente mayor que la del resto. En los primeros instantes de estos episodios se producen rachas de residuos positivos consecutivos, que dan lugar a la existencia de correlación positiva.

La solución al primer problema requerirá la introducción en el modelo de

información suficiente, desde el primer instante de cada episodio. Con este objetivo, se ha abierto una nueva línea de trabajo basada en la introducción de variables que caractericen un periodo, de longitud fija, anterior a cada observación, independientemente de que las observaciones de ese intervalo se encuentren por encima o por debajo del umbral. Con este planteamiento cabe esperar que también mejorará la independencia del grupo A. El segundo aspecto señalado es inherente a los datos y, en consecuencia, de difícil solución.

6.2 Predicción del riesgo de fallo y la finalización de un episodio

En esta sección se desarrollan modelos para predecir el riesgo de fallo y la finalización o no en un instante de un episodio seco activo, en función de las características del mismo observadas hasta ese momento.

La evolución de un episodio de sequía puede expresarse en términos de una variable binaria que describa su continuidad o finalización, tomando el valor 1 si la observación es la última del episodio, y 0 en otro caso. La predicción de esta variable en un instante se realiza a partir de la predicción del riesgo de finalización del episodio en ese momento, es decir del riesgo de la variable duración. La función riesgo de fallo es muy utilizada en Análisis de Supervivencia y Fiabilidad para estudiar el comportamiento de una variable positiva, T, que representa el tiempo hasta el fallo de un individuo y que puede estar sometida a la influencia de distintos factores.

6.2.1 Función riesgo de fallo

La función de riesgo $h(t)$ se define como la tasa de fallo instantánea de un individuo que no ha fallado hasta ese instante; en el caso de una distribución continua su expresión es:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t \mid T \geq t\right)}{\Delta t} = \frac{f(t)}{S(t)},$$

y en el de una variable discreta,

$$h(t_j) = P\left(T = t_j \mid T \geq t_j\right) = \frac{p(t_j)}{S(t_j^-)},$$

siendo $S(t) = P(T > t)$ la función de supervivencia.

En Análisis de Supervivencia es frecuente modelizar directamente la función de riesgo; este procedimiento permite determinar la combinación de variables explicativas que afecta al tiempo de fallo, ya que $h(t)$ caracteriza la distribución de T, y proporciona un estimador del riesgo en cada instante; éste tiene interés en sí mismo y permite estimar la función de supervivencia. Aunque los modelos resultantes son diferentes a los de regresión lineal, muchos de sus principios y procedimientos son aplicables.

Se proponen dos alternativas para modelizar la función de riesgo, una aproximación paramétrica y otra semiparamétrica; ambas permiten la inclusión de covariables dependientes del tiempo.

Modelos semiparamétricos de riesgo proporcional Estos modelos no requieren que la función de distribución de T tenga una forma determinada, la única condición necesaria es la denominada hipótesis de riesgo proporcional, que supone que el cociente del riesgo de dos individuos distintos es constante en el tiempo, es decir,

$$h(t) = h_0(t)g(x1, \ldots, xk),$$

donde $h_0(t)$ es la función de riesgo base o riesgo asociado a un individuo con covariables nulas, y $g(x1, \ldots, xk)$ se puede interpretar como el riesgo relativo respecto al riesgo base, de un individuo con covariables $x1, \ldots, xk$.

El modelo más utilizado de esta familia es el modelo de riesgo proporcional de Cox, en el que se define la función g como,

$$g(x1, \ldots, xk) = \exp\left(\sum_{j=1}^{k} \beta_j xj\right).$$

Recientemente se han desarrollado generalizaciones del modelo de Cox, como los modelos HARE (*hazard regression*), (Kooperberg, Stone & Truong 1995, Kooperberg & Clarkson 1997), que estiman la función de riesgo condicional utilizando *splines*

lineales y son aplicables en situaciones que no verifican la hipótesis de riesgo proporcional.

Modelos paramétricos En los modelos paramétricos se supone conocida la distribución de la variable T y, en consecuencia, la forma de la función de riesgo, excepto un vector de parámetros θ. Generalmente se plantean dos hipótesis simplificadoras:

i. Las covariables influyen en la variable T a través de los parámetros de su distribución, que son función del predictor lineal.

ii. El riesgo en el instante t no depende de toda la trayectoria anterior del vector de covariables dependientes del tiempo, que denotaremos $X(t)$, sino sólo del valor de dicho vector en ese instante, $x(t)$.

En estas condiciones, el riesgo se puede expresar como,

$$h\left[t_i \mid X(t_i)\right] = h\left[t_i \mid x(t_i)\right],$$

que es una función de t_i y $\theta[x(t_i)]$.

El procedimiento de estimación e inferencia que se presenta en los siguientes apartados, es válido para los dos tipos de modelos.

Función de riesgo de fallo de los episodios de sequía

La predicción del riesgo de finalización de una sequía se puede incluir en este marco, considerando cada episodio como un individuo cuyo fallo consiste en observar un valor de la señal de precipitación por encima del umbral. Con esta formulación, la duración es una variable tiempo de fallo cuyo riesgo en cada instante depende del estado del episodio en ese momento. En consecuencia, el problema se reduce a predecir la función de riesgo de una variable, a partir de un conjunto de covariables dependientes del tiempo. Los procedimientos del Análisis de Supervivencia se pueden simplificar en algunos aspectos, ya que en este caso no existen observaciones censuradas.

Este procedimiento es aplicable a la predicción de otras magnitudes que presentan el mismo tipo de estructura, por ejemplo, el riesgo de comienzo de una nueva

sequía cuando el proceso se encuentra en un periodo no seco, o de reentrada en un periodo seco en las observaciones en estado indeterminado.

6.2.2 Estimación e inferencia de los modelos para h(t)

Para estimar los modelos de la función de riesgo se emplea el método de máxima verosimilitud. Definiendo la función de verosimilitud en términos de $h(t)$ y denotando t_i a la observación i-ésima, la expresión de la función es,

$$\begin{aligned}
\ell &= \prod_{i=1}^{n} f\,(\,t_i \mid x(t_i)\,) \\
&= \prod_{i=1}^{n} h\,(\,t_i \mid x(t_i)\,)\, S\,(\,t_i \mid x(t_i)\,) \\
&= \prod_{i=1}^{n} h\,(\,t_i \mid x(t_i)\,) \exp\left(-\int_0^{t_i} h\,(\,u \mid x(u)\,)\, du\right).
\end{aligned}$$

Suponiendo que $h(t)$ queda completamente determinada por un vector de parámetros θ de dimensión m, las funciones de verosimilitud y logverosimilitud, $\ell\ell$, son,

$$\begin{aligned}
\ell(t,x;\theta) &= \prod_{i=1}^{n} h\,(\,t_i \mid x(t_i);\theta\,) \exp\left(-\int_0^{t_i} h\,(\,u \mid x(u);\theta\,)\, du\right) \\
\ell\ell(t,x;\theta) &= \sum_{i=1}^{n} \ln[h\,(\,t_i \mid x(t_i);\theta\,)] - \int_0^{t_i} h\,(\,u \mid x(u);\theta\,)\, du.
\end{aligned}$$

En el caso de distribuciones discretas,

$$\begin{aligned}
\ell(t,x;\theta) &= \prod_{i=1}^{n} \left(h\,(\,t_i \mid x(t_i);\theta\,) \prod_{t_k<t_i} [1 - h\,(\,t_k \mid x(t_k);\theta\,)] \right) \\
\ell\ell(t,x;\theta) &= \sum_{i=1}^{n} \left(\ln[h\,(\,t_i \mid x(t_i);\theta\,)] + \sum_{t_k<t_i} \ln[1 - h\,(\,t_k \mid x(t_k);\theta\,)] \right).
\end{aligned}$$

Esta expresión, incluso en el caso de distribuciones con función de riesgo sencilla como la Exponencial y la Geométrica, es complicada y para obtener el MLE del vector de parámetros θ se deben emplear métodos iterativos de maximización, por ejemplo, el de Newton-Raphson.

Los procedimientos de inferencia basados en la verosimilitud, (Kalbfleisch & Prentice 1980, Cox & Oakes 1984, Collett 1994), pueden dividirse en tres bloques.

- Basados en el vector de scores, $U(\theta) = [U^1(\theta), \ldots, U^m(\theta)]$, con

$$U^j(\theta) = \frac{\partial \ell(\theta)}{\partial \theta_j},$$

que tiene una distribución asintótica Normal con vector de medias nulo y matriz de covarianzas la matriz de información de Fisher o matriz de información esperada, $I(\theta)$, de elementos:

$$I_{ij}(\theta) = E\left(\frac{-\partial^2 \ell\ell(\theta)}{\partial \theta_i \partial \theta_j}\right).$$

- Basados en el MLE $\hat{\theta}$. Bajo ciertas condiciones de regularidad este estimador tiene una distribución asintótica Normal con vector de medias θ y matriz de covarianzas $I^{-1}(\theta)$; en consecuencia, la distribución asintótica del estadístico,

$$(\hat{\theta} - \theta)' I(\theta)(\hat{\theta} - \theta),$$

es χ^2_m. Si se sustituye la matriz $I(\theta)$ por un estimador consistente, como la matriz de información observada evaluada en $\hat{\theta}$, la correspondiente expresión presenta la misma distribución asintótica.

- Basados en el test de razón de verosimilitudes. Bajo la hipótesis, $\theta = \theta_0$, la distribución asintótica del estadístico,

$$\Lambda(\theta_0) = -2 \ln\left(\frac{\ell(\theta_0)}{\ell(\hat{\theta})}\right) = 2[\ell\ell(\hat{\theta}) - \ell\ell(\theta_0)]$$

es χ^2_m.

Los métodos que se proponen en los siguientes apartados se basan principalmente en el test de razón de verosimilitudes, el más recomendado en situaciones generales, que es invariante bajo reparametrizaciones, (Barndorff-Nielsen 1983, Barndorff-Nielsen & Cox 1984).

6.2.3 Covariables dependientes del tiempo

Se pueden distinguir dos tipos de covariables dependientes del tiempo, según su forma de dependencia,

i. Covariables definidas: variables cuya dependencia respecto al tiempo tiene una forma funcional conocida.

ii. Covariables estocásticas: variables que son la realización de un proceso estocástico. Dentro de esta clase de variables, Kalbfleisch & Prentice (1980) distinguen dos tipos,

ii.a. Covariables externas: aquéllas en cuya trayectoria no influye el valor de la variable tiempo de fallo, aunque ellas influyan en su valor.

ii.b.- Covariables internas: aquéllas que son el resultado de un proceso generado por el individuo, de forma que el valor de la covariable puede depender del valor de su tiempo de fallo.

El cálculo e interpretación de una verosimilitud en la que intervienen covariables internas no es sencillo. En efecto, una variable interna sólo puede ser observada mientras el individuo no ha fallado y, en consecuencia,

$$P\,(\,T \geq t \mid x(t)\,) = 1.$$

En esta situación, la función de riesgo no se puede plantear en términos de la función de supervivencia, y es necesario formular una expresión de la función de verosimilitud alternativa a la desarrollada en la sección 6.2.2.

6.2.4 Verosimilitud parcial

La introducción del concepto de verosimilitud parcial permite desarrollar técnicas de inferencia en presencia de parámetros 'inútiles'. Es frecuente que en la especificación de una distribución, además del vector de parámetros de interés θ, aparezcan parámetros irrelevantes. La idea básica de los métodos de verosimilitud parcial consiste en separar la función de verosimilitud en dos términos -uno, en el que la información sobre los parámetros de interés aparezca ligada a los parámetros inútiles, y otro que sólo sea función de θ- y simplificar la estimación considerando solamente el segundo término. No existe ningún resultado, con validez general, que permita asegurar la suficiencia de los estimadores obtenidos por este procedimiento y, obviamente, la función obtenida no es una verosimilitud en sentido estricto, ya que

no tiene una interpretación directa en términos probabilísticos; no obstante, existen argumentos heurísticos que sugieren que, en muchas situaciones, la aplicación de estos métodos no provoca una pérdida de información significativa en la estimación.

A continuación, desarrollamos la obtención de la función de verosimilitud asociada al modelo. Una función de verosimilitud se puede construir como el producto de las contribuciones de intervalos de la forma $[t, t+dt)$. Para calcular estas contribuciones se define $H(t)$, el proceso que contiene toda la información disponible sobre el fallo y las covariables hasta el instante t; $H(t)$ es un proceso de Markov, por lo que, utilizando la notación de Kalbfleisch & Prentice (1980), la verosimilitud se puede expresar como el producto integral de las probabilidades condicionales que aporta cada intervalo,

$$\ell = \mathcal{P}_0^{\infty} \left\{ P\left[H(t+dt) \mid H(t) \right] \right\}.$$

La información contenida en el proceso $H(t)$ se puede separar en dos partes: una relativa al proceso $D_t(dt)$, que contiene información sobre el fallo en el intervalo $[t, t+dt)$, y otra a la trayectoria de las covariables hasta t, $X(t)$. En consecuencia, cada factor se puede descomponer en dos términos,

$$P\left[H(t+dt) \mid H(t) \right] = P\left[X(t+dt) \mid H(t), D_t(dt) \right] P\left[D_t(dt) \mid H(t) \right],$$

y la verosimilitud se puede expresar como,

$$\ell = \mathcal{P}_0^{\infty} \left\{ P\left[D_t(dt) \mid H(t) \right] \right\} \mathcal{P}_0^{\infty} \left\{ P\left[X(t+dt) \mid H(t), D_t(dt) \right] \right\}.$$

donde el segundo factor corresponde a la contribución del proceso de covariables.

Habitualmente, la función de verosimilitud se define con el primer producto; si el segundo factor depende del vector de parámetros θ, como sucede en presencia de covariables internas, ese primer término es una verosimilitud parcial.

Bajo la hipótesis de que, dado $H(t)$, el mecanismo de fallo actúa de forma independiente en el intervalo $[t, t+dt)$,

$$P\left[D_t(dt) \mid H(t) \right] = \prod_{l \in D_t(dt)} h[t \mid x_l(t); \theta] dt \prod_{l \in R(t) - D_t(dt)} \left(1 - h\left[t \mid x_l(t); \theta \right] dt\right),$$

donde $x_l(t)$ es el vector de covariables del individuo l-ésimo evaluado en el instante t, y $R(t)$ el conjunto de individuos en riesgo en ese instante. Consecuentemente, la

función de verosimilitud parcial, ℓ_p, es,

$$\begin{aligned}\ell_p(\theta) &= \mathcal{P}_0^{\infty}\left\{\prod_{l\in D_t(dt)} h\,[\,t \mid x_l(t);\theta\,]\,dt \prod_{l\in R(t)-D_t(dt)} (1-h\,[\,t \mid x_l(t);\theta\,]\,dt)\right\} \\ &= \prod_{i=1}^{n} h\,(\,t_i \mid x_i(t_i);\theta\,)\;\mathcal{P}_0^{\infty}\left\{\prod_{l\in R(t)-D_t(dt)} (1-h\,[\,t \mid x_l(t);\theta\,]\,dt)\right\},\end{aligned}$$

donde $t_1 \leq t_2 \leq \ldots \leq t_n$ representan los tiempos de fallo de la muestra y $h\,(\,t_i \mid x_i(t_i);\theta\,)$ el riesgo del individuo i-ésimo evaluado en su tiempo de fallo.

La expresión del producto integral depende del tipo de distribución considerada; en el caso discreto,

$$\begin{aligned}&\mathcal{P}_0^{\infty}\left\{\prod_{l\in R(t)-D_t(dt)} (1-h\,[\,t \mid x_l(t);\theta\,]\,dt)\right\} \\ &= \prod_{i=1}^{t_n}\prod_{l\in R(i)-D(i)} (1-h\,[\,i \mid x_l(i);\theta\,]) \\ &= \prod_{i=1}^{t_n}\prod_{t_l>i} (1-h\,[\,i \mid x_l(i);\theta\,]) \\ &= \prod_{i=1}^{n}\prod_{l<t_i} (1-h\,[\,l \mid x_i(l);\theta\,]),\end{aligned}$$

donde $h[l \mid x_i(l);\theta]$ corresponde al riesgo del individuo i-ésimo de la muestra evaluado en el instante l, con l moviéndose en los enteros positivos menores que su tiempo de fallo t_i.

En el caso de distribuciones continuas la obtención de la expresión correspondiente es más compleja. Considerando una partición $0 = \tau_0 < \ldots < \tau_m < \infty$ tal que $\lim_{m\to\infty} \Delta\tau_i = 0$,

$$\begin{aligned}&\mathcal{P}_0^{\infty}\left\{\prod_{l\in R(t)-D_t(dt)} (1-h\,[\,t \mid x_l(t);\theta\,]\,dt)\right\} \\ &= \exp\left[\lim_{m\to\infty}\sum_{i=1}^{m} \ln\left(\prod_{l\in R(\tau_i)-D_{\tau_i}(\Delta\tau_i)} [1-h\,[\,\tau_i \mid x_l(\tau_i);\theta\,]\,\Delta\tau_i]\right)\right] \\ &= \exp\left[\lim_{m\to\infty}\sum_{i=1}^{m} \left(\sum_{l\in R(\tau_i)-D_{\tau_i}(\Delta\tau_i)} \ln\,[\,1-h\,[\,\tau_i \mid x_l(\tau_i);\theta\,]\,\Delta\tau_i\,]\right)\right]\end{aligned}$$

$$= \exp\left[-\lim_{m\to\infty}\sum_{i=1}^{m}\left(\sum_{l\in R(\tau_i)-D_{\tau_i}(\Delta\tau_i)} h\left[\tau_i \mid x_l(\tau_i);\theta\right]\Delta\tau_i\right)\right],$$

e intercambiando los sumatorios y aplicando la definición de integral,

$$= \exp\left[-\sum_{i=1}^{n}\int_0^{t_i} h(u \mid x_i(u);\theta)du\right].$$

En consecuencia la expresión de la verosimilitud parcial en el caso discreto es,

$$\ell_p(\theta) = \prod_{i=1}^{n}\left[h\left(t_i \mid x_i(t_i),\theta\right)\prod_{l<t_i}[1-h\left(l \mid x_i(l);\theta\right)]\right]$$

y en el caso continuo,

$$\ell_p(\theta) = \prod_{i=1}^{n} h\left(t_i \mid x_i(t_i),\theta\right)\ \exp\left[-\int_0^{t_i} h\left(u \mid x_i(u);\theta\right)du\right].$$

Como se ha comentado, en presencia de covariables internas la función de supervivencia no se puede interpretar de la forma habitual y, por lo tanto, la expresión obtenida no corresponde a

$$\prod_{i=1}^{n} h\left(t_i \mid x_i(t_i),\theta\right) S\left(t_i \mid x_i(t_i),\theta\right).$$

Inferencia basada en la verosimilitud parcial

Los métodos de inferencia máximo verosímiles siguen siendo aplicables en estas condiciones. Kalbfleisch & Prentice (1980) desarrollan resultados que justifican la aplicación de las propiedades asintóticas de la verosimilitud a la verosimilitud parcial, cuando la función está compuesta por un número elevado de términos, m, cada uno de los cuales aporta una pequeña cantidad de información sobre θ. Este es el caso de la función de verosimilitud parcial asociada a un problema con covariables internas. En particular, demuestran la validez de los siguientes resultados.

- La distribución asintótica del vector de scores, U, es Normal con vector de medias nulo y matriz de covarianzas estimada $I(\hat{\theta})$, bajo las condiciones necesarias para la aplicación del teorema central del límite, que son:

  - Independencia de las contribuciones U_j al score total U.
  - Varianzas similares de las componentes del vector de scores, $Var(U_j)$.
  - Convergencia a ∞ del sumatorio $\sum I_j$, con una velocidad adecuada.

- La distribución asintótica del MLE $\hat{\theta}$ es Normal con vector de medias θ y matriz de covarianzas $I^{-1}(\theta)$. Este resultado se obtiene aplicando desarrollos en serie de Taylor válidos bajo condiciones suaves sobre las derivadas parciales de tercer orden de la logverosimilitud parcial.

- Como consecuencia del resultado anterior, el test de razón de verosimilitudes también es aplicable.

Es importante señalar que la función de verosimilitud parcial planteada proporciona información, únicamente, sobre la tasa de fallo instantánea dado $X(t)$. Cualquier tipo de inferencia sobre la distribución marginal de T requerirá integración sobre $X(t)$.

6.2.5 Verosimilitud parcial de algunas distribuciones de interés

A partir de los resultados obtenidos en al apartado anterior se desarrolla la expresión de la verosimilitud parcial de las distribuciones Exponencial, Geométrica y Binomial negativa. De forma análoga, se puede obtener la expresión correspondiente a otras distribuciones.

Distribución Exponencial La distribución Exponencial se caracteriza por tener función de riesgo constante, igual a su parámetro α; para garantizar que se obtienen valores positivos del parámetro, se plantea un modelo de la forma,

$$h(t) = \alpha(t) = \exp[\nu(t;\beta)],$$

donde $\nu(t;\beta) = \sum_{i=1}^{k} \beta_i x_i(t)$. Como el valor de las covariables evoluciona de forma discreta, la función de riesgo se supone constante entre dos instantes consecutivos y, en consecuencia,

$$\int_0^{t_i} h\,(\,u \mid x_i(u);\beta\,)\,du = \sum_{l \le t_i} h\,(\,l \mid x_i(l);\beta\,)\,,$$

donde l es un índice en los enteros positivos. Sustituyendo este valor, se obtiene la expresión de la función de verosimilitud parcial,

$$\begin{aligned}
\ell_p(\beta) &= \prod_{i=1}^{n} h\,[\,t_i \mid x_i(t_i);\beta\,] \exp\left[-\sum_{l\le t_i} h\,(\,l \mid x_i(l);\beta\,)\right] \\
&= \prod_{i=1}^{n} \exp[\nu(t_i;\beta)] \exp\left(-\sum_{l\le t_i} \exp[\nu(l;\beta)]\right) \\
&= \prod_{i=1}^{n} \exp\left[\nu(t_i;\beta) - \sum_{l\le t_i} \exp[\nu(l;\beta)]\right]
\end{aligned}$$

y de logverosimilitud,

$$\ell\ell_p(\beta) = \sum_{i=1}^{n} \left(\nu(t_i;\beta) - \sum_{l\le t_i} \exp[\nu(t_j;\beta)]\right).$$

Distribución Geométrica0 La función de riesgo de la distribución Geométrica0 es constante e igual a su parámetro p. Para garantizar que el modelo proporciona un valor del parámetro en el rango adecuado, se utiliza la función de enlace logística,

$$h(t) = p(t) = \frac{\exp[\nu(t;\beta)]}{1+\exp[\nu(t;\beta)]}.$$

La función de verosimilitud parcial asociada es,

$$\begin{aligned}
\ell_p(\beta) &= \prod_{i=1}^{n} \left[h\,(\,t_i \mid x_i(t_i);\beta\,) \prod_{l<t_i} [1 - h\,(\,l \mid x_i(l);\beta\,)]\right] \\
&= \prod_{i=1}^{n} \left[\frac{\exp[\nu(t_i;\beta)]}{1+\exp[\nu(t_i;\beta)]} \prod_{l<t_i} \frac{1}{1+\exp[\nu(l;\beta)]}\right] \\
&= \prod_{i=1}^{n} \frac{\exp[\nu(t_i;\beta)]}{\prod_{l\le t_i}(1+\exp[\nu(l;\beta)])},
\end{aligned}$$

y la correspondiente función de logverosimilitud,

$$\ell\ell_p(\beta) = \sum_{i=1}^{n} \left[\nu(t_i;\beta) - \sum_{l\le t_i} \ln(1+\exp[\nu(l;\beta)])\right].$$

Distribución Binomial Negativa Para simplificar el modelo, supondremos que sólo el parámetro $p(t;\beta)$ de la distribución es función de las covariables. Para garantizar que este parámetro toma valores en $[0,1]$, se utiliza la misma función de enlace que en la distribución Geométrica. En estas condiciones la función de riesgo de la distribución es,

$$h(t) = \frac{Pb(t;n,\beta)}{1-\sum_{j=0}^{t-1} Pb(j;n,\beta)}$$

con $Pb(t;n,\beta) = \begin{pmatrix} n+t-1 \\ t \end{pmatrix} p(t;\beta)^n[1-p(t;\beta)]^t$ y la expresión de la verosimilitud parcial,

$$\ell_p(n,\beta) = \prod_{i=1}^{n} \left(\frac{Pb(t_i;n,\beta)}{1-\sum_{k=1}^{t_i-1} Pb(k;n,\beta)} \prod_{l<t_i} \left[1 - \frac{Pb(l;n,\beta)}{1-\sum_{k=1}^{l-1} Pb(k;n,\beta)} \right] \right).$$

La logverosimilitud correspondiente es,

$$\begin{aligned} \ell\ell_p(n,\beta) &= \sum_{i=1}^{n} \left(\ln[Pb(t_i;n,\beta)] - \ln\left[1 - \sum_{k=1}^{t_i-1} Pb(k;n,\beta) \right] \right. \\ &\quad \left. + \sum_{l<t_i} \ln\left[1 - \frac{Pb(l;n,\beta)}{1-\sum_{k=1}^{l-1} Pb(k;n,\beta)} \right] \right). \end{aligned}$$

Este modelo puede generalizarse permitiendo que el parámetro n sea también una función de las covariables aunque, generalmente, los modelos no mejoran bajo esta hipótesis.

6.2.6 Predicción de la probabilidad de fallo en el marco GLM

La evolución del estado de sequía se ha planteado en términos de una variable binaria que toma el valor 1, si la observación es la última del episodio, y 0 en otro caso. Bajo la hipótesis de dependencia markoviana, las observaciones de esta variable, dado el estado actual del proceso, serán independientes con distribución Bernoulli cuyo parámetro dependerá del estado del episodio a través de las covariables. Esta situación es análoga a la planteada en la predicción de la duración restante, sección

6.1, sólo que, en este caso, la respuesta es binaria. En consecuencia, la predicción de la probabilidad de finalización del episodio se puede plantear en el marco de los GLM utilizando un error Binomial. En efecto,

- Se tienen observaciones, en cada instante de cada episodio, de una variable respuesta I, indicador binario de la finalización del episodio, condicionalmente independientes dado el estado del proceso, así como el valor de un conjunto de covariables en esos instantes.

- La variable I tiene una distribución Bernoulli con función de probabilidad,

$$P(I_i;\theta_i) = p_i^{I_i} = \exp[I_i \ln(p_i)],$$

  que tiene la forma de la familia exponencial con $A_i = 1$ para todo i, parámetro de escala $\sigma = 1$, parámetro $\theta_i = \ln(p_i)$, y funciones γ y τ nulas.

- Las covariables influyen en la respuesta sólo a través de su valor medio, el parámetro p, que es una función del predictor lineal; para garantizar que el valor ajustado del parámetro toma valores en el rango $[0, 1]$ se utiliza una función de enlace logística, $\eta = \ln[p/(1-p)]$.

Bajo la hipótesis de que el tiempo de fallo tiene una distribución Geométrica, el riesgo coincide con la probabilidad de fallo en cada instante, y la expresión de la verosimilitud del modelo GLM planteado en este apartado,

$$\ell(\beta) \quad = \quad \prod_{i=1}^{n} \left[p\,(\,t_i \mid x_i(t_i);\beta\,) \prod_{l<t_i} [1 - p\,(\,l \mid x_i(l);\beta\,)] \right]$$

es la misma que la verosimilitud definida en la sección 6.2.5 para la distribución Geométrica. En consecuencia, en el caso de variables con esa distribución, como la duración de los episodios, la modelización y predicción del riesgo en un instante es equivalente a la de la probabilidad de fallo.

Se han programado los algoritmos de predicción del riesgo de otras distribuciones, como la BN, pero en el caso de la Geométrica, dada la equivalencia señalada y la mayor rapidez computacional del ajuste de modelos GLM, se ha utilizado esta aproximación en la modelización del riesgo.

6.2.7 Proceso de modelización y predicción

Las covariables consideradas son las mismas que las definidas en la modelización de la duración restante. Todas ellas son covariables internas, ya que se construyen a partir del proceso y, en consecuencia, la verosimilitud asociada es una verosimilitud parcial. La selección de covariables se realiza con un algoritmo paso a paso, basado en un test de razón de verosimilitudes y en los criterios AIC y BIC, de un modo análogo al descrito en la sección 6.1.3.

El modelo propuesto predice la probabilidad de finalización del episodio; sin embargo, interesa formular la predicción como una respuesta binaria que indique si el episodio finaliza o no, en ese instante. En consecuencia, se debe establecer un criterio que asigne un valor '1' -finalización- o '0' -continuación- de acuerdo a las probabilidades ajustadas.

El criterio más sencillo es determinar un umbral de referencia para la probabilidad de finalización, de forma que si el valor ajustado es mayor que el valor de referencia, la predicción será '1', y '0' en otro caso. En muestras con frecuencias de los valores de la respuesta equilibradas, el valor de referencia lógico será 0.5; en este caso, el número de observaciones con valor '1' es muy inferior al de las que tienen valor '0' y, en consecuencia, la estimación del valor del umbral adaptada a esta situación, sería la proporción muestral de las observaciones con valor '1'. Un criterio alternativo consiste en buscar el umbral que minimice una medida del error de predicción. Para evitar sesgos, las predicciones en las que se basará esta medida se calculan utilizando el método de validación cruzada descrito en la sección 6.1.5.

Para definir una medida del error de predicción es preciso establecer, en primer lugar, los distintos tipos de error y la importancia asignada a cada uno de ellos.

- Error 10: error que se produce cuando el valor observado es 1 y el predicho 0.

- Error 01: error que se produce cuando el valor observado es 0 y el predicho 1. Una predicción errónea de este tipo no tiene siempre las mismas consecuencias; es obvio que no debe considerarse tan grave predecir la finalización de un episodio un instante antes de que ésta se produzca, que predecirla mucho antes del final del episodio.

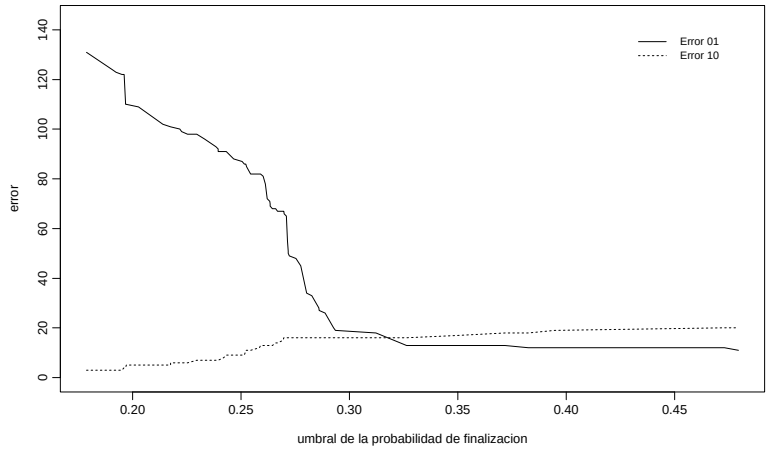

Figura 6.10: Gráfico de las curvas de error, Mrseq2, seq10; Huesca.

Como las covariables utilizadas se definen a partir del estado del propio episodio, en los instantes iniciales, al no disponer de información suficiente, la predicción será poco fiable independientemente de la bondad del modelo.

En consecuencia, se define una medida que considera todos los errores que se producen en la predicción, con diferentes pesos según su importancia: a los errores de tipo 01 se les asigna un peso proporcional a la distancia a la que se encuentran del final del episodio, excepto los que se producen en la etapa inicial del modelo, las tres primeras observaciones, que tienen peso 1. Los errores de tipo 10 corresponden siempre a la misma situación y se les asigna a todos peso 1.

Debido a la diferente proporción de los valores de la respuesta, la suma de las medidas de error tipo 01 y 10 no proporciona un criterio equitativo; en efecto, la mayor magnitud de la primera enmascara a la segunda y, en consecuencia, minimizar su suma es prácticamente equivalente a minimizar el error 01. El objetivo es minimizar las dos medidas de forma simultánea y, por consiguiente, se debe buscar el mínimo en el espacio bidimensional definido por ambas. Como la medida del error 10 es una función creciente del umbral, y la del error 01 decreciente, el punto donde se alcanza el mínimo corresponde al punto de corte de las dos funciones. El valor correspondiente a ese punto se puede calcular, de forma aproximada, representándolas gráficamente, figura 6.10.

En general, el número de observaciones de la muestra es grande y resulta difícil

establecer la capacidad predictiva de un modelo, o compararlo con modelos alternativos, analizando las predicciones en cada instante. Por ello proponemos la comparación de las medidas de error y el análisis de las p-predicciones utilizando el gráfico de la figura 6.11. En este gráfico se representa las respuestas y las probabilidades de finalización ajustadas en cada instante; se muestra también la magnitud de su diferencia, señalando las que corresponden a predicciones erróneas. El gráfico permite analizar también la bondad de las predicciones en función de su localización en el episodio. La bondad del modelo se puede evaluar globalmente, o analizando por separado los errores tipo 10 y 01, que corresponden, respectivamente, a los segmentos superiores e inferiores.

Con el objeto de facilitar la interpretación de la predicción de la probabilidad de finalización, se calcula un intervalo de confianza aproximado para el parámetro p, basado en la aproximación Normal de la distribución del estimador del predictor lineal, (Collett 1991), que se representa en un gráfico análogo al anterior, figura 6.12.

Bondad de ajuste El carácter binario de la respuesta invalida algunas de las técnicas de validación expuestas en la sección 6.1.5. En particular, en los modelos con respuesta binaria la desviación no se puede utilizar como medida de bondad de ajuste; en efecto, en estos modelos la desviación,

$$\mathcal{D} = -2\sum_{i=1}^{n}\left[\hat{p}_i \ln\frac{\hat{p}_i}{1-\hat{p}_i} + \ln(1-\hat{p}_i)\right],$$

depende de la respuesta sólo a través de las probabilidades ajustadas, por lo que no se puede interpretar como una medida de la concordancia entre los valores observados y los ajustados. En los modelos binarios la distribución de la desviación no es, ni siquiera de forma aproximada, χ^2; no obstante, la aproximación de la diferencia entre las desviaciones de dos modelos anidados sigue siendo válida, por lo que se mantiene su aplicación en la comparación de modelos, (Collett 1991).

En el análisis de residuos aparecen también algunas dificultades adicionales. Si el error es Bernoulli, los residuos de la desviación son,

$$rd_i = sgn(y_i - \hat{p}_i)\sqrt{-2\left[y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)\right]},$$

expresión que se reduce a $-[-2\ln(1-\hat{p}_i)]^{1/2}$, si la respuesta es 0, y $[-2\ln(\hat{p}_i)]^{1/2}$,

si es 1. En consecuencia, los residuos son positivos o negativos dependiendo del valor de la respuesta; además, la distribución de los residuos no se puede aproximar por una Normal aunque el modelo sea correcto. No obstante, el gráfico seminormal de los residuos junto con una envolvente calculada mediante simulación, (Atkinson 1981), aporta información sobre la existencia de datos atípicos y la adecuación del predictor lineal. Si el modelo es correcto, los puntos de la muestra deben encontrarse, con una probabilidad prefijada, dentro de la envolvente; en particular, con diecinueve simulaciones la probabilidad de que el mayor residuo absoluto de los datos se encuentre fuera de la envolvente, siendo el modelo correcto, es del 5%. Los residuos se estandarizan de la forma indicada en el caso general.

Otros gráficos también presentan algunas peculiaridades que son consecuencia del carácter binario de los datos. Por ejemplo, los gráficos de los residuos frente al predictor lineal no son informativos, ya que siempre muestran un patrón independientemente de que el modelo sea correcto o no. La interpretación del gráfico de la serie de residuos, de los residuos frente al instante de predicción y del de correlación sigue siendo la habitual.

En modelos con respuesta binaria los datos atípicos corresponden a observaciones $y_i = 1$ cuya probabilidad ajustada es próxima a 0, o a observaciones $y_i = 0$ con valor ajustado próximo a 1. Estas probabilidades ajustadas se obtienen con valores grandes, negativos y positivos respectivamente, del predictor lineal, por lo que todos los datos atípicos corresponderán a observaciones con valores extremos en las covariables, es decir, con un alto potencial; en consecuencia, los posibles datos anómalos se pueden identificar representando la serie h_{ii}.

6.2.8 Análisis de resultados

En este apartado se muestran modelos de predicción del riesgo de finalización de los periodos secos y de las sequías, de la serie de Huesca correspondiente al umbral definido con el percentil décimo. El proceso de modelización es el mismo que el descrito en los modelos de predicción de la duración restante.

Periodos secos Algunos de los mejores modelos obtenidos son,

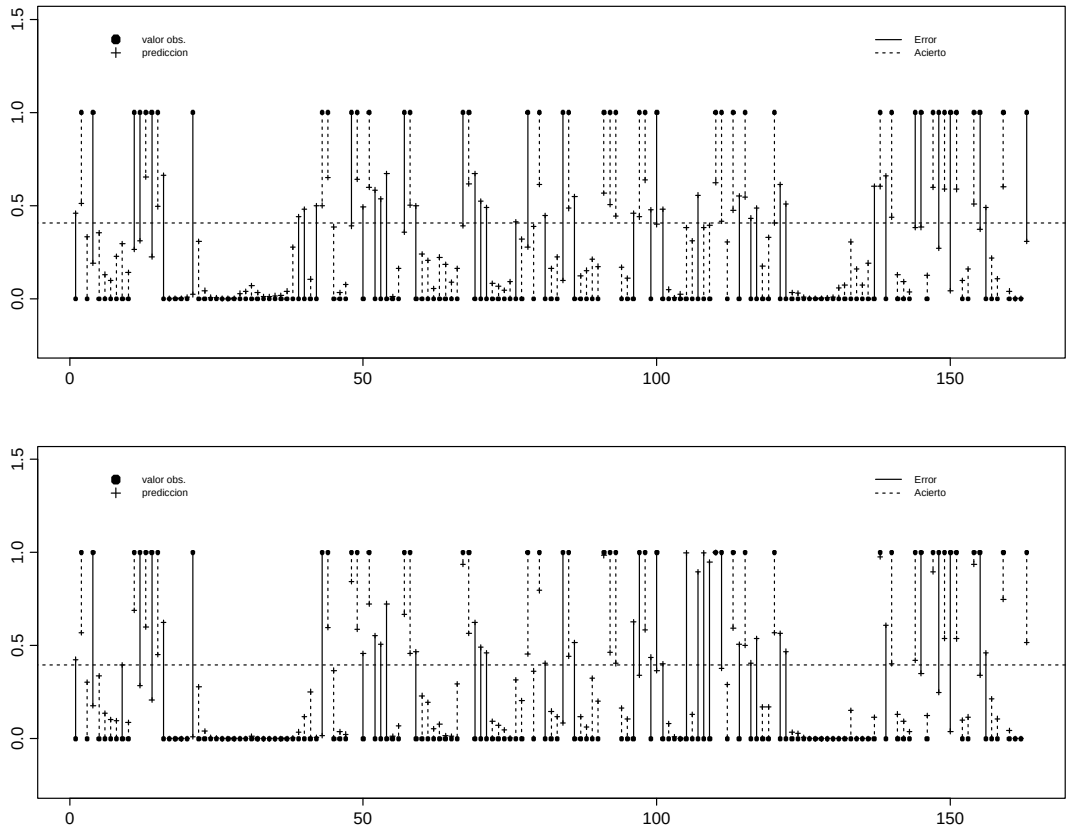

Figura 6.11: Gráfico de las predicciones de finalización, Mrps1 (sup.) y Mrps3 (inf.), ps10; Huesca.

Mrps1: intt.

Mrps2: intt, rec, defter, durter, dimaxter.

Mrps2b: intt, rec, direct, defter, durter, imaxter, dimaxter, intt2er, intt3ei.

Mrps3: imaxt, dimaxt, inttei, durter, recer, deft2.

Mrps3b: imaxt, dimaxt, inttei, durter, dimaxter, recer, deft2, durt3ei.

Mrps4: intt, rec2, deft2er, dimaxt2er.

Mrps4b: intt, rec, durer, imaxter, dimaxter, intt2er, deft2er, dimaxt3er.

En la tabla 6.3 se muestran los resultados de los ajustes de estos modelos. Los modelos de mínimo AIC son más complicados y presentan valores mayores de las

	$\mathcal{D}_M$	AIC	BIC	Error01	Error10	Errort
Mrps1	148.4	152.4	158.6	36	17	53
Mrps2	128.3	140.3	158.9	39	15	54
Mrps2b	118.8	138.8	169.7	41	14	55
Mrps3	125.7	139.7	161.4	35	13	48
Mrps3b	117.2	135.2	163.0	39	13	52
Mrps4	135.2	145.2	160.7	33	13	46
Mrps4b	118.8	136.8	164.7	35	14	49

Tabla 6.3: Modelos ajustados al riesgo de finalización de los periodos secos, ps10; Huesca.

medidas de error que los correspondientes al criterio AIC-BIC combinado, por lo que no se consideran competitivos.

Un rasgo a destacar es la importancia en estos modelos de la variable intt, que es, con diferencia, la covariable más significativa; el modelo Mrps1, que contiene únicamente esta covariable, es el de mínimo BIC. Su inclusión es difícil de sustituir por otra u otras covariables: ella o una de sus interacciones aparece siempre en los mejores modelos. Existen otras familias de covariables, menos significativas, cuya presencia también es constante: deft, dimaxt y rec; estas variables aparecen en su forma original o en interacción con estrec.

Como se observa en la figura 6.11 y la tabla 6.3, el modelo Mrps1 es competitivo con otros más complicados. Comparado con Mrps3, el que proporciona mejores resultados, predice peor en el instante final de los episodios ya que aunque sus probabilidades suelen ser mayores que las de las observaciones anteriores no llegan a superar el umbral de referencia; a cambio produce menos falsos avisos de finalización.

La desviación del modelo cuyo predictor lineal contiene las 16 componentes principales del espacio determinado por las 65 covariables es 128.8.

Episodios de sequía A continuación se presentan algunos de los mejores modelos seleccionados con el criterio AIC-BIC combinado, y los modelos correspondientes de mínimo AIC; las medidas de error de éstos últimos son mayores por lo que no se han considerado competitivos.

	$\mathcal{D}_M$	AIC	BIC	Error01	Error10	Errort
Mrseq1	130.6	134.6	140.8	17	26	43
Mrseq2	105.5	115.5	130.9	14	16	30
Mrseq2b	99.9	113.9	135.5	15	17	32
Mrseq*3	101.5	113.6	132.0	13	17	30
Mrseq*3b	99.6	113.6	135.2	17	18	35
Mrseq*4	77.9	101.9	138.9	14	16	30

Tabla 6.4: Modelos ajustados al riesgo de finalización de las sequías, seq10; Huesca.

Mrseq1: intt.

Mrseq2: intt2, rec2ei, dimaxter, defpsteips.

Mrseq2b: imaxt2, imaxtei, rec2ei, dimaxter, intteips, defpsteips.

Mrseq*3: dimaxt, intt2, deftei, recer, f1estimaxpst.

Mrseq*3b: dimaxt, intt2, deftei, recer, durt2er, f1estimaxpst.

Mrseq*4: intt, rec, inttei, defpsti, difint2ei, dimaxter, difinteips, f1difint2, f2intt, f2difint2, f3imaxpst2.

La variable intt es, individualmente, la más significativa de las covariables definidas, aunque, a diferencia de lo que ocurría en el proceso de los periodos secos, el modelo Mrseq1 no es el de mínimo BIC; es decir, existen otras covariables cuya inclusión en el modelo complementa, de forma significativa, la información aportada por esta variable. En concreto, analizando sus predicciones, tabla 6.4 y figura 6.12, se observa que la variable intt no es suficiente para discriminar la situación previa a la finalización del episodio; en consecuencia, el modelo Mrseq1 presenta una medida Error 10 alta. Existen familias de covariables cuya presencia es constante en los mejores modelos, en particular, deft, dimaxt, rec e intt.

La desviación del modelo cuyo predictor lineal contiene las 19 primeras componentes principales del espacio determinado por las 87 covariables es 92.2, y la del modelo ajustado con las 27 componentes que aproximan el espacio generado por el conjunto de las 171 covariables, 76.3. El modelo Mrseq*4, con 11 covariables, presenta un valor de la desviación comparable. Sin embargo, la disminución de la

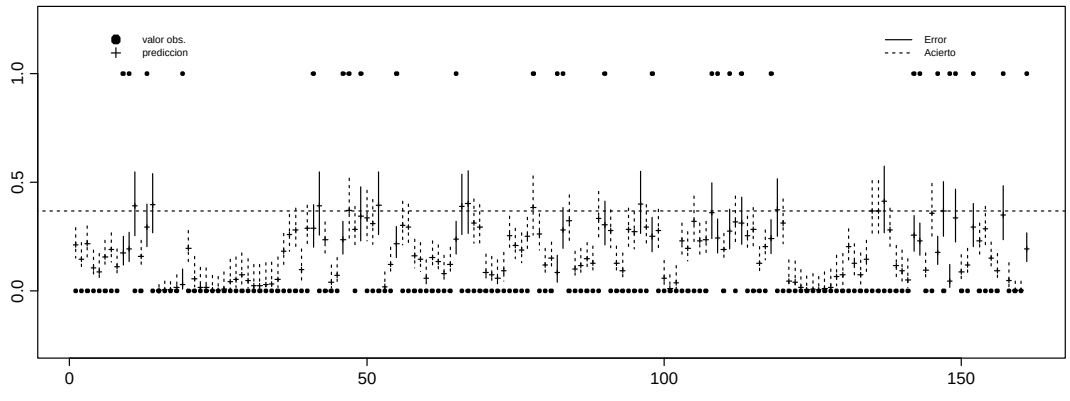

Figura 6.12: Gráfico de las predicciones de finalización y su error estándar, Mrseq1 seq10; Huesca.

desviación no implica una mejora en las medidas de error de la predicción; en este aspecto, Mrseq*4 es análogo a Mrseq2 y Mrseq*3, que tienen un menor número de parámetros.

Los medidas de error de los modelos Mrseq2 y Mrseq*3 son similares aunque sus predicciones presentan algunas diferencias, figura 6.13; Mrseq2 tiende a anticipar los finales de los episodios.

Los resultados de la modelización del riesgo de finalización de los procesos de periodos secos y sequías se resumen en las siguientes conclusiones.

- Las covariables de tipo local, como rec, intt o estrec, son más importantes en la predicción de finalización de un episodio que en la de la duración restante.

- Modelos con medidas de error similares pueden proporcionar predicciones mejores en distinto tipo de situaciones; en consecuencia, la elección del mejor modelo se puede condicionar al tipo de error que se considere más importante, o de peores consecuencias, dependiendo de la aplicación de la predicción.

- Obviamente, la disminución del umbral de referencia mejora el porcentaje de errores tipo 10 pero aumenta el de errores 01. En consecuencia, la selección del umbral también se puede condicionar al tipo de error que se quiere minimizar.

- Las probabilidades que toman valores en torno al umbral de referencia no permiten discriminar la respuesta binaria a la que corresponden; las predicciones

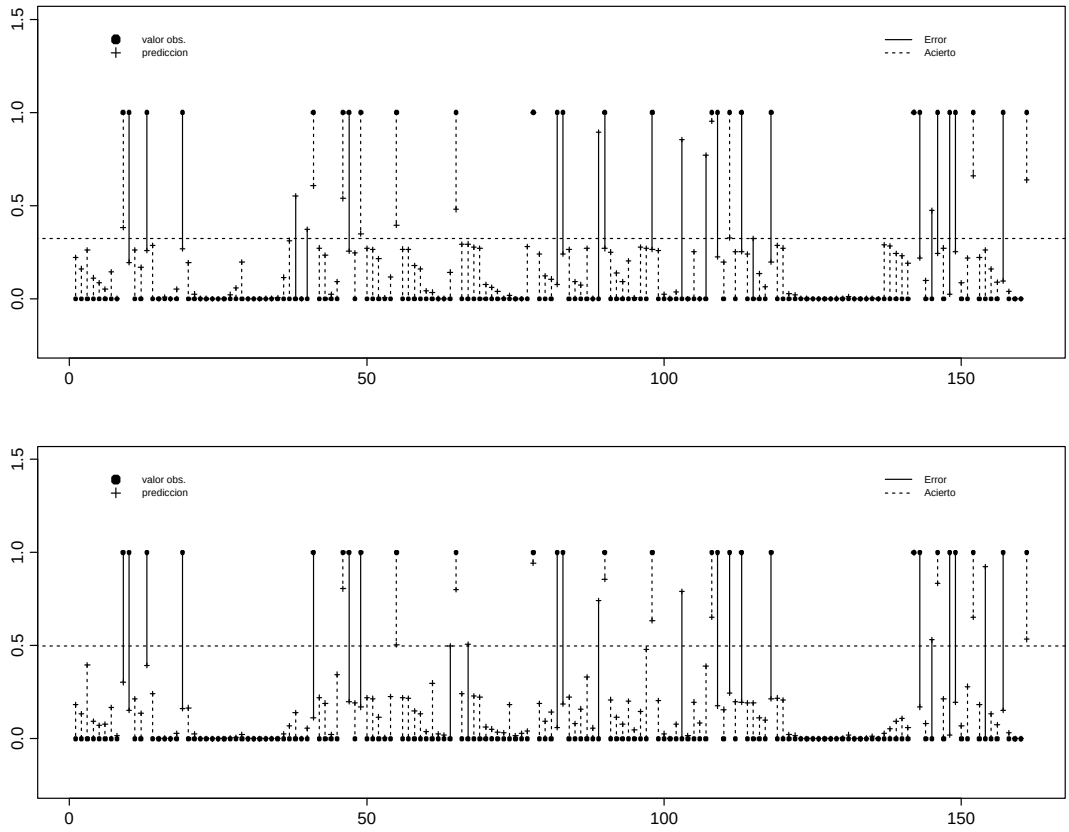

Figura 6.13: Gráfico de las predicciones de finalización, Mrseq2 (sup.) y Mrseq*3 (inf.) seq10; Huesca.

en los instantes iniciales de cada episodio suelen tomar valores en este rango, debido a la poca información disponible para realizar la predicción.

- Un inconveniente de la predicción binaria es que no diferencia la magnitud del error, éste es el mismo independientemente del valor de la probabilidad ajustada; por esta razón, en algunas situaciones, puede ser más interesante interpretar las probabilidades ajustadas como probabilidad de finalización de la sequía, que realizar una predicción 0-1.

Validación del modelo Mrseq*3 En la figura 6.14 se muestra la serie temporal de los residuos estandarizados de la desviación de este modelo. Debido al carácter binario de la respuesta, los residuos de las observaciones '0', más numerosas, son siempre negativos, mientras que los correspondientes a la observación final de ca-

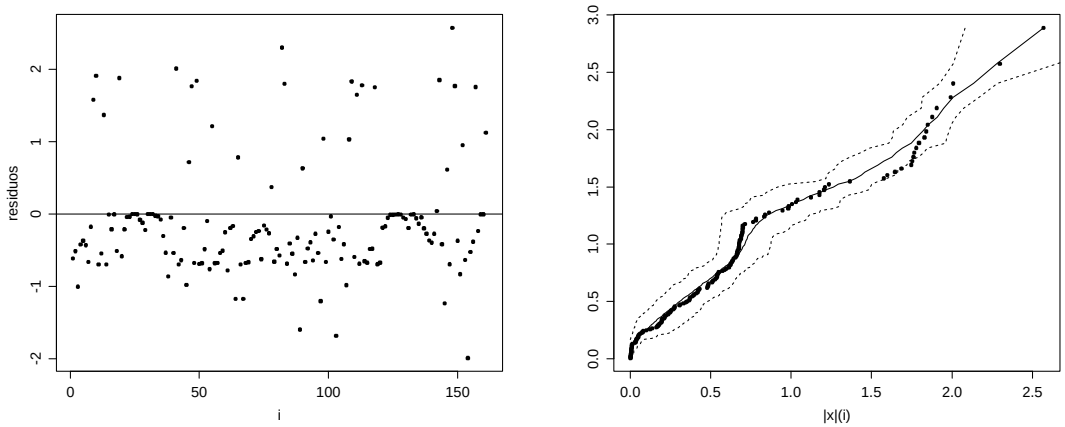

Figura 6.14: Gráfico de la serie temporal y gráfico seminormal de los residuos estandarizados de la desviación, Mrseq*3, seq10; Huesca.

da episodio son positivos. El gráfico seminormal, en la misma figura, no indica insuficiencias en el predictor lineal del modelo.

El gráfico de correlación de los residuos, figura 6.15, resulta de difícil interpretación debido a la formación de grupos dependiendo de la respuesta de las observaciones i e $(i+1)$-ésima. Los pares de observaciones pertenecientes al mismo episodio se localizan en la parte negativa del eje horizontal. Globalmente, no existe correlación, ya que el p-valor del test de Kendall es 0.97, aunque existe un pequeño grupo de observaciones que si presentan este problema. En los tres gráficos restantes de la figura se representan las observaciones de los grupos A, B y C definidos en los modelos para la duración restante. Los p-valores de los tests de correlación son 0.58, 0.72 y 0.0002, respectivamente; de nuevo, son las observaciones de la etapa inicial de los episodios más largos las que presentan dependencia significativa. El origen de esta estructura es el mismo que el indicado en los modelos Mseq, aunque debido a la mayor influencia en esta respuesta de los factores de carácter local, la estructura es menos marcada.

6.3 Implementación en S-plus

Modelos para la duración restante

- **Función: sdeptiem.fun y sdeptiems.fun** (sdeptiem.txt). Estas funciones

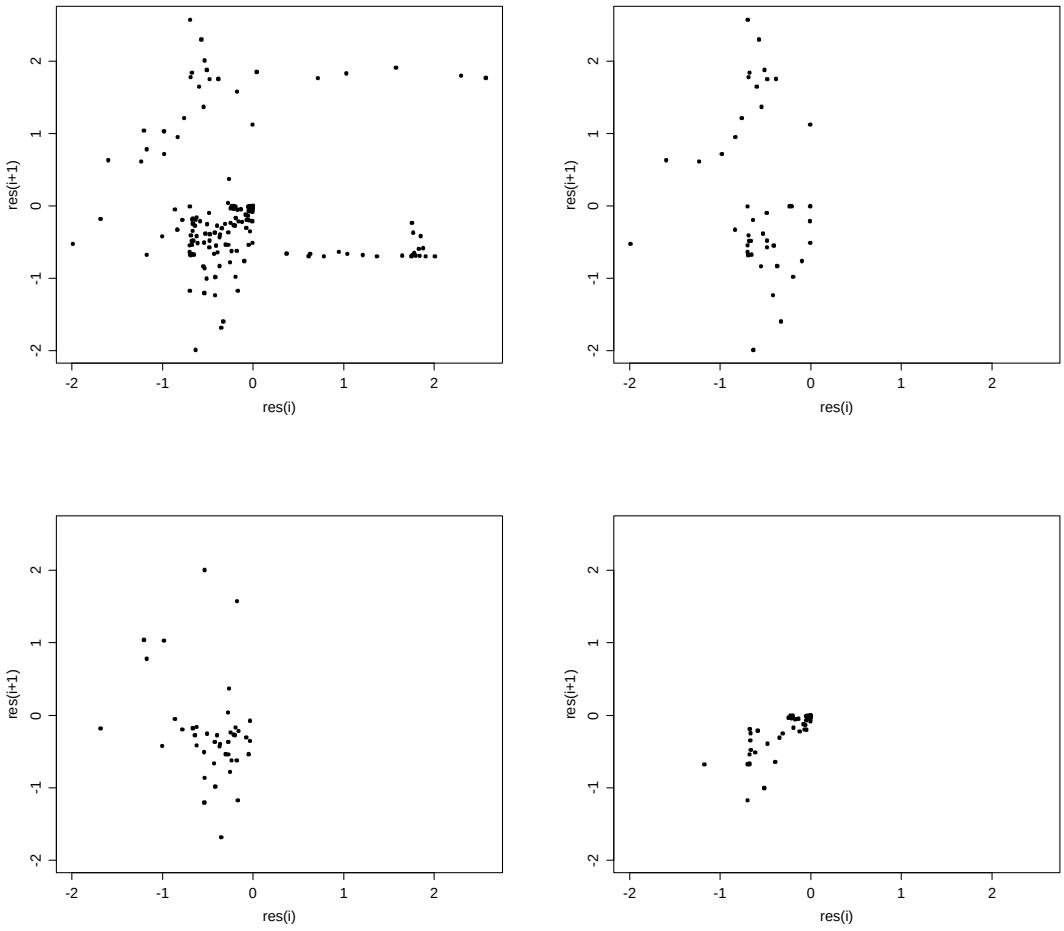

Figura 6.15: Gráfico de correlación de los residuos de todas las observaciones (sup.izda.), del grupo A (sup.dcha.), B (inf.izda.) y C (inf.dcha.), Mrseq*3 seq10; Huesca.

calculan las observaciones de la variable tiempo restante y de las correspondientes covariables, en cada instante de cada periodo seco y sequía respectivamente. También crean una variable que etiqueta las observaciones de episodios distintos. En el caso de las sequías no se incluyen las observaciones en estado indeterminado. El resultado de las funciones es una matriz de datos que se puede guardar, opcionalmente, en un fichero.

Argumentos:

- smo: serie de sumas móviles de la precipitación
- pos, def, dur: series de posición inicial, duración y déficit de cada episodio
- perc: percentil con el que se define el umbral que determina la entrada

en un periodo seco

- tipo: (sólo en sdeptiem.fun) etiqueta que indica el tipo de periodo 'ps' seco, o 'pns' no seco, al que corresponden los datos. Por defecto 'ps'
- percc: (sólo en sdeptiems.fun) percentil con el que se define U2. Por defecto 0.3.

- **Función: prepmodelos.fun, prepmodelos2.fun y prepmodelos3.fun** (faglmgeoc.txt). Estas funciones preparan las covariables e interacciones necesarias asociadas a los periodos secos, y a las sequías, sin considerar, y considerando la familia nper, respectivamente. El resultado es una lista con componentes, obs: respuesta observada en cada instante, data: matriz de covariables, ncolu: número de columnas de la matriz anterior, fac: etiqueta del número de episodio, y auxnomt: lista de nombres de las covariables.

  Argumentos:

  - dataf: base de datos con las covariables dependientes del tiempo
  - nomb: etiqueta del nombre del observatorio
  - tipod: etiqueta del tipo de dato que se está analizando.

- **Función: aglmgeoc.fun** (aglmgeoc.txt.). Esta función realiza un proceso de selección de covariables utilizando el algoritmo paso a paso descrito. Se puede aplicar a periodos secos y sequías. El resultado de la función es una lista con componentes, obs: vector de valores observados, datacov: matriz de covariables seleccionadas, dentro: vector de códigos de las covariables seleccionadas, fac: indicador del episodio correspondiente de cada observación y durt: respuesta o duración restante en cada instante.

  Argumentos:

  - listobs: lista con las componentes indicadas en el resultado de la función prepmodelos.fun
  - nomb: etiqueta del nombre del observatorio
  - tipod: etiqueta del tipo de dato que se está analizando.

  Subfunciones:

- geo.glm: definición de una familia glm, con error geométrico y enlace logarítmico.

- **Función: predglm.fun** (predglm.txt). Esta función calcula la predicción en cada instante de la muestra utilizando la media y la mediana. También calcula los p-residuos y las medidas de comparación propuestas.

  Argumentos:

  - listglm: lista con las componentes indicadas en el resultado de la función aglmgeoc.txt
  - nomb: etiqueta del nombre del observatorio
  - tipod: etiqueta del tipo de dato que se está ajustando
  - resmaxve, resminve, resrelmaxve, resrelminve: argumentos opcionales para indicar el rango, máximo y mínimo, para los gráficos de residuos y residuos relativos de la predicción basada en el valor esperado
  - resmaxm, resminm, resrelmaxm, resrelminm: argumentos opcionales para indicar el rango, máximo y mínimo, para los gráficos de residuos y residuos relativos de la predicción basada en la mediana.

  Subfunciones: (fpredglm.txt)

  - pressglm.fun: calcula la predicción en cada instante de cada episodio de la muestra utilizando el procedimiento de validación cruzada propuesto.
  - medba.fun: calcula las medidas de bondad de ajuste y los gráficos propuestos para comparar la predicción de un modelo.
  - resresp.fun: calcula y analiza gráficamente los residuos respuesta y los residuos respuesta relativos.
  - compglm.fun: compara dos modelos anidados con un test de razón de verosimilitudes.

- **Función: valglm.fun** (valglm.txt). Esta función realiza un análisis de validación del modelo; calcula los residuos estandarizados de la desviación y los de máxima verosimilitud, con los que realiza los gráficos descritos. Permite analizar la correlación de los residuos en función de su instante de predicción, la longitud de los episodios y la duración restante.

Argumentos: listglm, nomb y tipod, definidos en la función predglm.fun.

Modelos para la finalización de un episodio

Computacionalmente, la ejecución de los modelos GLM es más rápida que la modelización de la función de riesgo, por lo que se ha programado el ajuste de la probabilidad de finalización de un episodio, en este marco.

- **Función: aglmhc.fun** (aglmhc.txt.). Esta función realiza un proceso de selección de covariables utilizando el algoritmo paso a paso descrito. Se puede aplicar a periodos secos y sequías. El resultado de la función es una lista como la de la función aglmgeoc.fun

  Argumentos: los mismos que la función aglmgeoc.fun.

- **Función: predglmh.fun** (predglmh.txt). Esta función calcula la predicción en cada instante de la muestra de la probabilidad de finalización del episodio y la correspondiente variable binaria con el umbral de referencia que minimiza simultáneamente los errores 10 y 01. Realiza el análisis gráfico de validación del modelo.

  Argumentos: los mismos que la función predglm.fun.

  Subfunciones: (fpredglmh.txt)

  - pressglmh.fun: calcula la predicción de la probabilidad de finalización en cada instante de cada episodio de la muestra, utilizando el procedimiento de validación cruzada.

- **Función: valglmh.fun** (valglmh.txt). Esta función realiza un análisis de validación del modelo; calcula los residuos estandarizados de la desviación y representa los gráficos descritos. Permite realizar un análisis de la correlación de los residuos en función de su instante de predicción, la longitud de los episodios y la duración restante.

  Argumentos: los mismos que la función valglm.fun

  Subfunciones:

- seminormal.fun: representa el gráfico seminormal de los residuos del modelo indicado y la correspondiente envolvente.

En el caso general, la modelización de la función de riesgo se hace utilizando procedimientos de máxima verosimilitud; para ello se utiliza la función nlminb de S-Plus que permite minimizar funciones no lineales, imponiendo restricciones sobre los parámetros si es necesario, con un algoritmo tipo Newton-Raphson. Esta función calcula el gradiente y el hessiano de la función por un método de diferencias. Como nlminb es un algoritmo de minimización, para obtener los estimadores MLE de los parámetros se minimiza el opuesto de la función de verosimilitud correspondiente.

Para utilizar la función anterior sólo es necesario especificar la expresión de la función de verosimilitud, más exactamente de menos la función de logverosimilitud en términos de los parámetros que se quieren estimar.

- **Función: asepdeptc.fun**(apdeptc.txt). Esta función realiza un proceso de selección de las covariables utilizando un algoritmo paso a paso. El resultado de esta función es una lista con los siguientes elementos, marcf1: un marcador de la observación final de cada episodio, datacov: la matriz con las covariables seleccionadas en el modelo, dentro: variable con los códigos de las covariables, fac: indicador del número de episodio y vinic: el vector de valores iniciales de los parámetros del modelo.

  Argumentos:

  - dataf: base de datos con las covariables dependientes del tiempo que se obtiene como resultado de la función sdeptiem.fun
  - nomb: etiqueta del nombre del observatorio
  - tipod: etiqueta del tipo de dato que se está analizando.

  Subfunciones: (fapdeptc.txt) además de prepmodelos2.fun, ya citada,

  - llp.obj: calcula la función de verosimilitud que se quiere maximizar cambiada de signo.

- **Función: predriesgo.fun**(predriesgo.txt). Esta función calcula la predicción en cada instante de cada episodio, utilizando el criterio de validación cruzada propuesto

Argumentos:

- listriesgo: es una lista que se obtiene como salida de la función apdept.fun
- nomb: etiqueta del nombre del observatorio
- tipod: etiqueta del tipo de dato que se está analizando.

Subfunciones: (predriesgo.txt)

- elimunah.fun: estima el modelo indicado eliminando las observaciones correspondientes a un episodio, y calcula la predicción en todos los instantes de ese episodio.

.

.

Símbolos y abreviaturas

AIC	Criterio AIC de Akaike
$a_n,\ b_n$	Constantes de normalización
BIC	Criterio BIC
BN	Distribución Binomial negativa
$\mathcal{D}^*$, $\mathcal{D}^*_M$	Desviación escalada (del modelo M)
$\mathcal{D}$, $\mathcal{D}_M$	Desviación (del modelo M)
$Dt(X)$	Desviación típica de la variable X
EOT	Exceso sobre un umbral
$E(X)$	Esperanza de la variable X
Exp	Distribución Exponencial
f.d.	Función de distribución
$f(x)$	Función de densidad de una variable aleatoria
$F(x)$	Función de distribución de una variable aleatoria
$F_e(x)$	Función de distribución empírica
$\bar{F}(x)$, $S(x)$	Función de supervivencia
GLM	Modelo lineal generalizado
$G(x)$	Función de distribución VE
$h(x)$	Función de riesgo
$H(x)$	Función de distribución GP
i.i.d.	Independiente, idénticamente distribuido
$k_{j,F}$	Cumulante de orden j de la distribución F
l_0	Longitud actual de un episodio activo
ℓ	Función de verosimilitud
$\ell\ell$	Función de logverosimilitud
ℓ_p	Función de verosimilitud parcial
$\ell\ell_p$	Función de logverosimilitud parcial
(L, D, IM)	Vector de magnitudes ps: duración, déficit e intensidad máxima
$(Lns, Ens, IMns)$	Vector de magnitudes pns: duración, exceso e intensidad máxima

ln	Logaritmo neperiano
L_r	Longitud o duración restante
$m_{j,F}$	Momento de orden j de la distribución F
max	Máximo
MDA	Max-dominio de atracción
min	Mínimo
MLE	Estimador máximo verosímil
M_n	Máximo de una muestra de tamaño n
M_N	Máximo de una muestra de tamaño N aleatorio
Mps	Modelo duración restante de periodos secos
Mseq	Modelo duración restante de sequías
Mseq*	Modelo duración restante de sequías con familia nper
Mrps	Modelo riesgo de periodos secos
Mrseq	Modelo riesgo de sequías
Mrseq*	Modelo riesgo de sequías con familia nper
$N(n)$, N	Número de ocurrencias en un intervalo de longitud n
Np	Número de periodos secos por cluster
$\mathcal{P}_0^\infty\{\ \}$	Producto integral
$P(\lambda)$	Distribución Poisson de parámetro λ
PG	Distribución Pareto generalizada
POT	Picos sobre un umbral
pns	Periodo no seco
PP	Proceso de Poisson
PPC	Proceso de Poisson Compuesto
PPCL	Proceso de Poisson Cluster
PPClC	Proceso de Poisson Cluster Compuesto
PPH	Proceso de Poisson homogéneo
PPNH	Proceso de Poisson no homogéneo
PRA	Proceso de renovación alternante
PRAC	Proceso de renovación alternante compuesto
ps	Periodo seco
ps10	Proceso de periodos secos definido con el percentil 10
$\mathbb{R}$	Conjunto de los números reales
$R_{-\alpha}$	Conjunto de funciones de variación regular
r_n	Número de excesos en n observaciones

seq10	Proceso de sequías definidas con el percentil 10
$rang(\ )$	Rango de una observación en una muestra
$sgn(\ )$	Signo de una expresión
S_n	Suma de $X_1, X_2, \ldots, X_n$
T_i	Tiempo de ocurrencia i-ésimo
Tr	Tiempo de recurrencia
u	Umbral de un proceso
u_n	Sucesión de umbrales
U1	Umbral de definición de los periodos secos
U2	Umbral de separación de dos periodos secos
v.a.	Variable aleatoria
$Var(X)$	Varianza de la variable X
VE	Distribución Valor extremo
Vr	Valor de retorno
$\bar{X}$	Valor medio muestral
X_i	Variable i-ésima de la muestra
$X_{(i)}$	Variable i-ésima de la muestra ordenada
x_F	Punto extremo superior de la distribución F
X_u	Variable Exceso sobre el umbral u
$z*$	Longitud mínima de separación de dos clusters
$\Gamma(\ )$	Función Gamma
λ	Intensidad de un PPH
$\lambda(t)$	Función de intensidad de un PP
Λ	Estadístico del test de razón de verosimilitudes
μ	Parámetro de localización
η	Predictor lineal
$\Phi(x)$	Función de distribución $N(0,1)$
$\psi(z)$	Función generatriz de probabilidades
$\Psi(g)$	Funcional generador de probabilidades
ρ	Coeficiente de correlación
ρ_S	Coeficiente de correlación de Spearman
σ	Parámetro de escala
τ	Coeficiente de correlación de Kendall
$\varphi(z)$	Función generatriz de momentos
$\theta(\ )$	Vector de parámetros de un modelo

θ	Indice extremal
$[\ ]_{entera}$	Parte entera
$\xrightarrow{d}$	Convergencia en distribución
$\xrightarrow{p}$	Convergencia en probabilidad
$\overset{d}{=}$	Igualdad en distribución
$X \sim a$	Variable con distribución a
$\approx$	Aproximadamente

Bibliografía

Abaurrea, J. & Cebrián, A. C. (1998), Modelización de episodios de sequía, *in* 'Actas del XXIV Congreso Nacional de Estadística e Investigación Operativa', Vol. 1, Universidad de Almería, pp. 19–21.

Abaurrea, J. & Cebrián, A. C. (1999), Análisis de episodios de sequía utilizando un proceso de Poisson cluster compuesto, *in* 'Actas de la VII Conferencia Española de Biometría', Vol. 1, Universitat de les Illes Balears, pp. 109–113.

Almarza, C., López, J. A. & Flores, C. (1996), *Homogeneidad y variabilidad de los registros históricos de precipitación en España*, Instituto Nacional de Meteorología. Ministerio de Medio-ambiente.

Ansell, J. I. & Phillips, M. J. (1994), *Practical methods for reliability data analysis*, Oxford University Press.

Artina, S. & Todini, E. (1985), Alternative approaches to risk assessment in hydrological processes, *in* E. Plate & N. Buras, eds, 'Scientific procedures applied to the planning, design and management of water resources systems', Vol. 147, IAHS Publication, pp. 225–36.

Ascaso, A. & Casals, M. (1981), 'Periodos secos y sequías en la depresión central del Ebro', *Geographicalia* **11-12**, 55–71.

Atkinson, A. C. (1981), 'Two graphical displays for outlying and influential observations in regression', *Biometrika* **68**, 13–20.

Balkema, A. A. & Haan, L. (1974), 'Residual lifetime at great age', *Ann. Probab.* **2**, 792–804.

Bardsley, W. E. & Manly, B. F. J. (1987), 'Transformations for improved convergence of distributions of flood maxima to a Gumbel limit', *J. Hydrol.* **91**, 137–152.

Barndorff-Nielsen, O. (1983), 'On a formula for the distribution of the maximum likelihood estimator', *Biometrika* **70**, 343–65.

Barndorff-Nielsen, O. & Cox, D. R. (1984), 'Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator', *J. R. Statist. Soc.* **B46**, 483–95.

Bates, W. (1976), National economic effects of drought in Australia, *in* T. Chapman, ed., 'Drought, Australian UNESCO seminar', pp. 217–42.

Benito, A., Orellana, P. & Zurita, E. (1994), Análisis de la estabilidad temporal de los patrones de la precipitación en la Península Ibérica, *in* M. F. Pita & M. Aguilar, eds, 'Cambios y variaciones climáticas en España', Universidad de Sevilla, pp. 183–90.

Beser, V. & Tico, A. (1993), 'La excepcionalidad de la sequía de 1989 en el Pais Vasco', *Cuadernos de sección. Historia. Geografía* **20**, 235–47.

Buishand, T. (1985), 'The effect of seasonal variation and serial correlation on the extreme value distribution of rainfall data', *Journal of Climate and applied Meteorology* **24**, 154–60.

Buishand, T. (1989), 'Statistics of extremes in Climatology', *Statistica Neerlandica* **43**(1), 1–31.

Capel, J. J. (1989), 'La sequía del invierno 1988-1989 en España', *Papeles de Geografía* **15**, 9–17.

Castillo, E. (1988), *Extreme value theory in Engineering*, Academic Press, Inc.

Castro, J. & Pérez-Abreu, V. (1994), Some statistical analysis of the cluster point processes of hurricanes on the coasts of Mexico, *in* Barnett & Turkman, eds, 'Statistics for the environment 2: Water Related Issues', John Wiley and Sons.

Chambers, J. M. & Hastie, T. J. (1992), *Statistical models in S*, Chapman and Hall.

Coles, S. (1994), A temporal study of extreme rainfall, *in* V. Barnett & K. Turkman, eds, 'Statistics for the environment 2: Water related issues', John Wiley and Sons, pp. 61–78.

Collett, D. (1991), *Modelling binary data*, Chapman and Hall.

Collett, D. (1994), *Modelling survival data in medical research*, Chapman and Hall.

Cox, D. R. & Oakes, D. (1984), *Analysis of survival data*, Chapman and Hall.

Cox, D. R. & Snell, E. J. (1968), 'A general definition of residuals', *J. Roy. Statist. Soc. Ser B* **30**, 248–75.

Daniel, W. W. (1990), *Applied nonparametric statistics*, PWS-Kent.

Davison, A. C. & Smith, R. L. (1990), 'Models for exceedances over high thresholds', *J. R. Statist. Soc. B* **52**(3), 393–442.

Davison, A. C. & Snell, E. J. (1991), Residuals and diagnostics, *in* Hinkley & Snell, eds, 'Statistical theory and modelling', Chapman and Hall, pp. 83–106.

Embrechts, P., Kluppelberg, C. & Mikosch, T. (1997), *Modelling extremal events*, Springer.

Fahrmeir, L. & Tutz, G. (1994), *Multivariate statistical modelling based on generalized linear models*, Springer.

Fisher, R. A. & Tippett, L. H. C. (1928), 'Limiting forms of the frequency distributions of the largest or smallest member of a sample', *Proc. Cambridge Philos. Soc.* **24**, 180–90.

Galambos, J. (1978), *The asymptotic theory of extreme order statistics*, John Wiley and Sons.

Gertensgarbe, F. W. & Werner, P. C. (1989), 'A method for the statistical definition of extreme-value regions and their application to meteorological time series', *Z. Meteorol.* **39**, 224–6.

Gibbons, J. D. & Chakraborti, S. (1992), *Non-parametric statistical inference*, Dekker.

Gibbs, W. J. & Maher, J. V. (1967), 'Rainfall deciles as drought indicators', *Bureau of Meteorology Bulletin. Melbourne. Australia*.

Gomes, M. I. (1981), An i-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes, *in* C. Taillie, G. Patil & B. Baldessari, eds, 'Statistical distributions in scientific work', Vol. 6, Reidel, pp. 389–410.

Gomes, M. I. (1984), 'Penultimate limiting forms in extreme value theory', *Ann. Inst. Statist. Math.* **36**, 71–85.

Greenberg, E. (1975), 'Minimum variance properties of principal components regression', *J. Amer. Statist. Assoc.* **70**, 194–7.

Griffiths, G. A. (1990), 'Rainfall deficits: distribution of monthly runs', *J. Hydrol.* **115**, 219–29.

Gumbel, E. J. (1958), *Statistics of extremes*, Columbia Univ. Press.

Hsing, T., Husler, J. & Leadbetter, M. R. (1988), 'On the exceedance point process for a stationary sequence', *Probab. Theory Related Fields* **78**, 97–112.

Judge, G. G. et al. (1985), *The theory and practice of econometrics*, John Wiley and Sons.

Kalbfleisch, J. D. & Prentice, R. L. (1980), *The statistical analysis of failure time data*, John Wiley and Sons.

Kendall, D. R. & Dracup, J. A. (1992), 'On the generation of drought events using an alternating renewal-reward model', *Stochastic Hydrol. Hydraul.* **6**, 55–68.

Kendall, M. & Gibbons, J. D. (1990), *Rank correlation methods*, Oxford University Press.

Kooperberg, C. & Clarkson, D. B. (1997), 'Hazard regression with interval-censored data', *Biometrics* **53**, 1485–94.

Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995), 'Hazard regression', *J. Amer. Statist. Assoc.* **90**(429), 78–94.

Lall, U., Rajagopalan, B. & Tarboton, D. G. (1996), 'A nonparametric wet-dry spell model for resampling daily precipitation', *Water Resour. Res.* **32**(9), 2803–23.

Leadbetter, M. R., Lindgren, G. & Rootzén, H. (1983), *Extremes and related properties of random sequences and processes*, Springer-Verlag.

Madsen, H. & Rosbjerg, D. (1998), A regional Bayesian method for estimation of extreme streamflow droughts, *in* E. Parent, B. Bobée, P. Hubert & J. Miquel, eds, 'Studies and reports in Hydrology', Unesco, pp. 327–40.

Madsen, H., Rasmussen, P. F. & Rosbjerg, D. (1997), 'Comparison of annual maximum series and partial duration series methods for modelling extreme hydrologic events. 1. At site modelling', *Water Resourc. Res.* **33**(4), 747–57.

Madsen, H., Rosbjerg, D. & Harremoes, P. (1994), 'PDS-modelling and regional Bayesian estimation of extreme rainfalls', *Nordic Hydrol.* **25**(4), 279–300.

Martín, J., Conesa, C. & Moreno, M. C. (1992), Acerca de la bondad de las cadenas de Markov de primero, segundo y tercer órdenes en el análisis de las sequías del sureste de España, *in* 'Actas del V coloquio de Geografía Cuantitativa', Universidad de Zaragoza, pp. 485–500.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Lineal Models. 2^a Ed.*, Chapman and Hall.

Menon, M. V. (1963), 'Estimation of the shape and scale parameters of the Weibull distribution', *Technometrics* **5**, 175–82.

Moyé, L. A. & Kapadia, A. S. (1995), 'Predictions of drought length extreme order statistics using run theory', *J. Hydrol.* **169**, 95–110.

Moyé, L. A., Kapadia, A. S. & Cech, I. M. (1988), 'The theory of runs with application to drought prediction', *J. Hydrol.* **103**, 127–37.

Nakamura, M. & Pérez-Abreu, V. (1993), 'Exploratory data analysis for counts using the empirical probability generating function', *Commun. Statist.- Theory Meth.* **2**(3), 827–42.

Olcina, J. & Rico, A. (1994), 'Sequías en el sureste ibérico: hecho climático o hecho humano', *Serie Geográfica* **4**, 11–30.

Ortigosa, L. M. (1987), 'Las sequías climáticas en el extremo noroccidental de la depresión del Ebro (La Rioja)', *Estudios geográficos* **189**, 639–58.

Pickands, J. (1975), 'Statistical inference using extreme order statistics', *Ann. Statist.* **3**(1), 119–31.

Pierce, D. A. & Schafer, W. (1986), 'Residuals in generalized linear models', *J. Amer. Statist. Assoc.* **81**(396), 977–86.

Pita, M. F. (1995), *Las sequías: análisis y tratamiento*, Junta de Andalucía. Consejería de Medio Ambiente.

Pérez, A. J. (1988), 'Notas sobre el concepto, los métodos de estudio y la génesis de las sequías', *Cuadernos de Geografía* **44**, 139–44.

Pérez, A. J. & Escrivá, J. L. (1982), 'Aspectos climáticos de las sequías en el ámbito mediterráneo', *Cuadernos de Geografía* **30**, 112–122.

Pérez, C. et al. (1984), 'Estudio de rachas secas y lluviosas en Gijón y San Sebastián', *Revista de Geofísica* **40**, 73–80.

Rasmussen, P. F. & Rosbjerg, D. (1991), 'Evaluation of risk concepts in PDS.', *Stochastic Hydrol. Hydraul.* **5**, 1–16.

Rasmussen, P. F. et al. (1994), The POT method for flood estimation: a review, *in* K. Hipel, ed., 'Exteme values: floods and droughts', Vol. 1, Kluwer Academic Publishers, pp. 15–26.

Raso, J. M. et al. (1981), 'La sequía del año agrícola 1980-81 en España', *Notes de Geografía Física* **6**, 31–47.

Resnick, S. (1987), *Extreme values, point processes and regular variation*, Springer.

Rossi, G. (1983), Droughts of Sicily, *in* V. Yevjevich, L. Cunha & E. Vlachos, eds, 'Coping with droughts', Water Resources Publications, pp. 244–58.

Saldariaga, J. & Yevjevich, V. (1970), *Application of run-lengths to hydrologic series*, Hidrology Papers. Colorado State University.

Sales, V., Jambrino, T. & Juste, J. J. (1982), 'Análisis espacial y temporal de la sequía 1978-81 en España', *Cuadernos de Geografía* **30**, 1–24.

Sen, Z. (1976), 'Wet and dry periods of annual flow series', *J. Hydrol.* **35**, 311–24.

Sen, Z. (1989), 'The theory of runs with application to drought prediction. Comment', *J. Hydrol.* **110**, 383–91.

Sen, Z. (1990), 'Critical drought analysis by second order Markov chain', *J. Hydrol.* **120**, 183–202.

Sen, Z. (1991*a*), 'On the probability of the longest run length in an independent series', *J. Hydrol.* **125**, 37–46.

Sen, Z. (1991*b*), 'Probabilistic modelling of crossing in small samples and application of runs to hydrology', *J. Hydrol.* **124**, 345–62.

Smith, R. L. (1984), Threshold methods for sample extremes, *in* J. T. de Oliveira, ed., 'Statistical extremes and applications', NATO Advanced Study Institute, Reidel, pp. 621–38.

Smith, R. L. (1986), 'Extreme value theory based on the r largest annual events', *J. Hydrol.* **86**, 27–43.

Smith, R. L. (1989), 'Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone', *Statistical Science* **4**(4), 367–93.

Tawn, J. A. (1988), 'An extreme value theory model for dependent observations', *J. Hydrol.* **101**, 227–50.

Todorovic, P. & Zelenhasic, E. (1970), 'A stochastic model for flood analysis', *Water Resour. Res.* **6**(6), 1641–8.

Wald, A. & Wolfowitz, J. (1943), 'An exact test for randomness in the non-parametric case based on serial correlation', *Ann. Math. Statist.* **14**, 378–88.

Weissman, I. (1978), 'Estimation of parameters and large quantiles based on the k largest observations', *J. Amer. Statist. Assoc.* **73**, 812–15.

Yevjevich, V. (1967), *An objective approach to definitions and investigations of continental hydrologic droughts*, Hydrol. Papers, Colorado State University Publ.

Yevjevich, V. (1984), Extremes in hidrology, *in* J. T. de Oliveira, ed., 'Statistical extremes and applications', Reidel, pp. 197–220.

Zelenhasic, E. & Salvai, A. (1987), 'A method of streamflow drought analysis', *Water Resour. Res.* **23**(1), 156–68.