ANÁLISIS ESTADÍSTICO DE LAS PRECIPITACIONES ANUALES EXTREMAS EN URUGUAY

Florencia Santiñaque

Tesis presentada con el objetivo de obtener el título de Magíster en Ingeniería Matemática

VERSIÓN POST DEFENSA CORREGIDA

Instituto de Probabilidad y Estadística Prof. Rafael Laguarda
Facultad de Ingeniería
Universidad de la República

Director Académico: Juan Kalemkerian.
Director de Tesis: Juan Kalemkerian.
Co-Directora de Tesis: Madeleine Renom.

March 4, 2020

Tabla de contenidos

Agradecimientos

A mis hijos Paula y Agustín y a mi companẽro de vida Juan Manuel por el aguante, sacrificio de tiempo compartido y apoyo incondicional permanente.

A mi familia, quienes también me apoyaron incondicionalmente.

A Juan Kalemkerian por ser un excelente guía y docente, por su humildad , sabiduría y siempre buena onda.

A Madeleine Renom por ser la impulsora del tema de tesis, por su confianza, tiempo dedicado y conocimiento compartido.

A mis amigas/os y colegas por el aguante.

A la Comisión Académica de Posgrado (CAP) de UdelaR por el apoyo otorgado a través de la beca para finalización de Maestría.

Resumen

En meteorología, así como en otras ciencias, el estudio de fenómenos extremos representan grandes desafíos. Al trabajar con datos extremos, es importante tener en cuenta que las metodologías clásicas no son las más adecuadas, siendo la teoría de valores extremos el marco apropiado en estos casos. El objetivo principal del presente trabajo consiste en estudiar la existencia de patrones espaciales de las preciptaciones máximas anuales diarias dentro del territorio uruguayo. En una primer etapa se estudiarán las distribuciones límite marginales de los valores extremos en cada estación meteorológica. En una segunda etapa se aplicarán metodologías de spatial clustering. Encontrar patrones espaciales con métodos basados en desviaciones de la media utilizando la distancia Euclidiana (L2) puede no ser la estrategia más apropiada en el contexto estudio de valores extremos. Una estrategia interesante para enfrentar este reto es utilizar un algoritmo de agrupamiento denominado Partitioning Around Medoids (PAM) utililizando como distancia el F-madogram. Se cuenta con una base de datos de precipitaciones diarias en 20 localizaciones del Uruguay para el período de enero 1981 a diciembre 2013. Se trabajó en bloques de máximos anuales diarios. Se encontró que en 18 localizaciones de las estudiadias, la distribución GEV que mejor ajusta a los datos es del tipo Gumbel. Para llegar a dicha conclusión, luego de estimados los parámetros, se procedió a relizar un test de hipótesis de Cramér-von Mises recortado para testear $H_0 : X \sim \text{Gumbel}(\mu, \sigma)$ contra $H_1 : H_0$ no es cierto. Mercedes y Rocha sin embargo rechazaron la hipótesis nula, siendo ambas modeladas según la distribución de Fréchet. Respecto a las metodologías de clustering, si se agrupa en base a los parámetros estimados de las distribuciones GEV, se encuentran dos grupos, mientras que si se agrupa en función a los datos de lluvias extremas diarias anuales de cada año con PAM utilizando el F-madograma como distancia, las estructuras de grupos encontradas fueron débiles o el algoritmo no logró captarlas. Algunos resultados obtenidos en las distintas agrupaciones se correspondieron con los resultados de un test de independencia en base a ratios de recurrencia realizado a todos los pares de estaciones de estudio.

Palabras clave: Precipitaciones extremas anuales; teoría de valores extremos; spatial clustering.

Part I
Introducción

1 Introducción

Los impactos causados por eventos meteorológicos extremos pueden tener graves consecuencias tanto económicas como humanas. En 1988, el Centro de Investigación sobre Epidemiología de los Desastres (CRED [15]) lanzó la base de datos de eventos de emergencia (EM-DAT). EM-DAT fue creado con el apoyo inicial de la Organización Mundial de la Salud (OMS) y el Gobierno de Bélgica. Según su publicación [52] en el que sólo se analizaron desastres naturales, en 2018 se registraron 315 eventos de desastres naturales con 11.804 muertes, más de 68 millones de personas afectadas y US$ 131.7 mil millones en pérdidas económicas en todo el mundo. Los terremotos fueron el tipo de desastre más mortal que representó el 45% de muertes, seguido de inundaciones con 24%. Las inundaciones afectaron al mayor número de personas, representando el 50% del total afectado, seguido de tormentas, que representaron el 28%. En relación con la década anterior (2008-2017), en 2018 hubo menor cantidad de desastres en comparación con el promedio anual de 348 eventos, menos muertes en comparación con el promedio anual de 67.572, menos personas afectadas en comparación con el promedio anual de 198.8 millones personas afectadas y menores pérdidas económicas en comparación al promedio anual de US$ 166.7 mil millones. Algunos de los eventos que afectaron fuertemente durante el período 2008-2017 fueron el terremoto de 2010 en Haití (222.500 muertes); sequía en la India en 2015/2016 dejando 330 millones de personas afectadas, terremoto y tsunami en Japón en el 2011 con US$ 210 mil millones en daños y perjuicios. Sobre los eventos registrados en 2018 se puede destacar dos terremotos en Indonesia (4.904 muertes). Cabe destacar que las inundaciones han afectado a más personas que cualquier otro tipo de desastre en el siglo XXI, incluso en 2018 (127 eventos). Es importante destacar que CRED define un desastre como «una situación o evento que satura la capacidad local, lo que requiere un solicitud a nivel nacional o internacional para asistencia externa; un imprevisto y a menudo repentino evento que causa gran daño, destrucción y sufrimiento humano». Además para que el evento sea ingresado en dicha base de datos, se deben cumplir al menos alguno de los siguientes criterios: 10 o más personas reportadas fallecidas, 100 o más personas reportadas afectadas, declaración de estado de emergencia ó convocatoria de asistencia internacional. Los datos se compilan a partir de diversas fuentes, incluidas agencias de la ONU, organizaciones no gubernamentales, compañías de seguros, institutos de investigación y agencias de prensa.

El Foro Económico Mundial (World Economic Forum) [40], es una organización público-privado, sin ánimo de lucro que reúne a los principales mandatarios de organizaciones internacionales, dirigentes de varios países, líderes de empresas y personas de reputado prestigio a nivel mundial para analizar los principales riesgos, retos y oportunidades que ofrece el panorama internacional, así como las principales tendencias geopolíticas, económicas y sociales a nivel global. Anualmente realizan un informe denominado «Informe de riesgos mundiales» en donde se analiza la percepción de riesgos globales a partir de una encuesta en la que participan aproximadamente 1.000 expertos y tomadores de decisiones. En la misma

se solicita evaluar tanto la probabilidad como el impacto de los riesgos mundiales[1] en una escala del 1 al 5, donde en lo que a probabilidad refiere 1 representaba un riesgo con pocas probabilidades de ocurrir y 5 un riesgo con muchas probabilidades de ocurrir; mientras que para medir el impacto la escala va de 1 (impacto mínimo), 2 (impacto menor), 3 (impacto moderado), 4 (impacto severo) y 5 (impacto catastrófico). En la figura (1.1) se puede observar cómo los riesgos ambientales asociados a eventos extremos se encuentran relacionados a alta percepción de probabilidades de ocurrencia así como a niveles altos de impacto y consecuencias. Como conclusión de este informe, y en lo que tiene que ver con riesgos extremos asociados a la naturaleza, los eventos climáticos extremos (inundaciones, tormentas, etc.), fracaso de la mitigación y adaptación al cambio climático, así como grandes desastres naturales (terremoto, tsunami, erupción volcánica, etc.) se encuentran dentro del top 5 en lo que a percepción de probabilidad de ocurrencia refiere. Respecto a riesgos por impacto, los grandes desastres naturales también forman parte del top 5.

Desde la CEPAL (Comisión Económica para América Latina y el Caribe) también advierten de la ocurrencia e impacto de eventos extremos en lo que es Latinoamérica y el Caribe. Una de las líneas de investigación y trabajo de esta institución abarca la colaboración con cada país en el desarrollo y obtención de estadísticas e indicadores de cambio climático, eventos extremos y desastres naturales. En ese sentido en Uruguay se están realizando esfuerzos en conjunto con diversas entidades entre ellas, INE (Instituto Nacional de Estadística), SINAE (Sistema Nacional de Emergencias), INUMET (Instituto Uruguayo de Meteorología), CECOED (Centros Coordinadores de Emergencias Departamentales), Policía Nacional de Tránsito, MIDES (Ministerio de Desarrollo Social), DNIC (Dirección Nacional de Identificación Civil), DINAGUA (Dirección Nacional de Aguas) y el Ministerio del Interior. En febrero del 2019 se realizó el Taller CEPAL «Estadísticas e indicadores de cambio climático, eventos extremos y desastres» en el que se marcaron diversos objetivos, entre ellos la creación de un sistema de información integrado para el procesamiento y gestión de riesgos, denominado MIRA, con el que se pretende por un lado generar, integrar y procesar información de forma unificada y estandarizada así como la generación de indicadores y estadísticas de calidad para apoyar la toma de decisiones en la prevención, mitigación, respuesta y recuperación del riesgo de emergencia. También se ha propuesto un sistema de monitoreo (SENDAI) para el registro e integración internacional de indicadores referentes a daños y pérdidas atribuidos a desastres como ser mortalidad, personas afectadas, pérdida económica, sistemas de alerta temprana, entre otros.

[1] Se define «riesgo mundial» como un evento o una condición potencial que, si se produce, puede tener un impacto negativo significativo en varios países e industrias dentro de los próximos diez años según definición dada en [40].

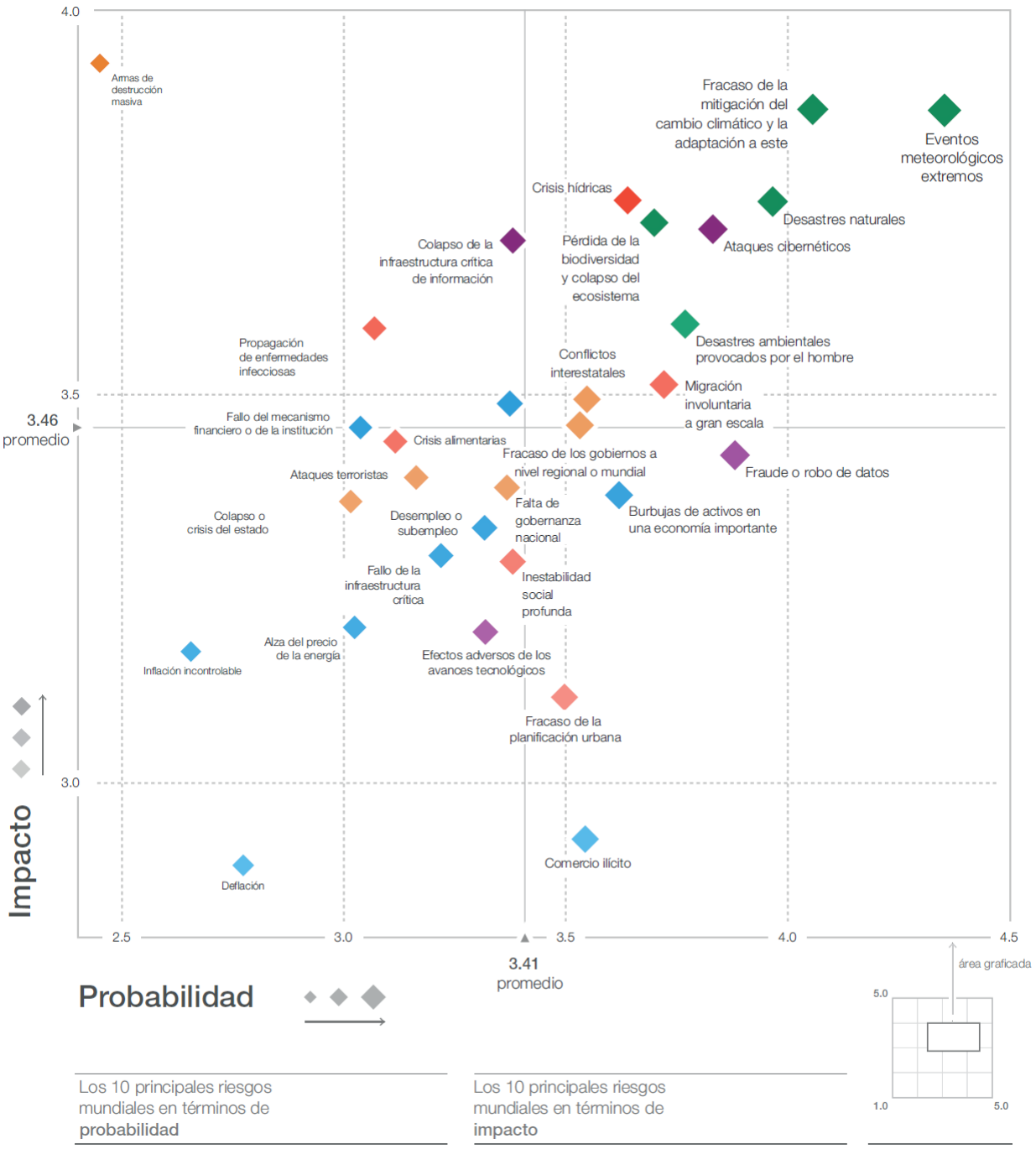


Figura 1.1: Percepción de riesgos globales. Fuente: Informe de riesgos mundiales 2019, World Economic Forum.

De la base de datos de eventos de emergencia (EM-DAT) surge que para Uruguay los mayores desastres naturales durante el período 1967 a 2014, ordenados en función a rangos relacionados con la población afectada y pérdidas humanas y económicas fueron: inundaciones (224.263 personas afectadas, 23 muertes y US$ 89.000.000 en pérdidas económicas), tormentas (2000 personas afectadas, 11 muertes y US$ 25.000.000 en pérdidas económicas). Temperaturas extremas ha dejado en dicho período 11 muertes y sequías han provocado alrededor de US$ 250.000.000 en pérdidas económicas al país.

El aumento de la variabilidad climática, a nivel mundial y en nuestro país, impone la necesidad de contar con herramientas que permitan analizar dichos fenómenos extremos para poder prevenir y/o minimizar los impactos causados por los mismos. Es por ello de suma importancia poder contar con estudios y análisis que otorguen herramientas e información que puedan ser utilizadas para una mejor comprensión de dichos fenómenos y es una de las principales motivaciones de la presente tesis.

2 Antecedentes

La exploración y análisis de datos meteorológicos extremos ha venido en aumento dado el crecimiento en la variabilidad climática. Diversos estudios académicos se han centrado en la utilización de la teoría de valores extremos para poder obtener conclusiones al respecto. Se puede mencionar por ejemplo el trabajo [42] que estudió las características de las precipitaciones extremas en La Rioja, España, analizando tanto la intensidad (mm máximos anuales diarios) así como también la acumulación de precipitaciones como consecuencia de la persistencia de lluvias, durante cierto período de tiempo. Se obtuvieron cartografías que reflejaron la máxima intensidad, magnitud y duración esperada. En [25] se estudiaron las precipitaciones extremas en Venezuela, ajustando modelos GEV a partir de la estimación de los parámetros, usando métodos Bayesianos. Los resultados arrojaron que los modelos Gumbel y Fréchet son los más apropiados para representar los máximos anuales en las localidades estudiadas. Sin embargo, en localizaciones con mesoclimas áridos o muy húmedos, el modelo Weibull es más apropiado. También se han analizado otro tipo de variables climáticas haciendo uso de dicha metodología, entre ellas se puede mencionar el trabajo de [4] en el que se estudia no sólo las precipitaciones máximas sino que se realiza un análisis de la tendencia de la temperatura tanto máxima como mínima en el Estado de Durango, México. Otra variable climática analizada utilizando este tipo de metodología es el viento, en este sentido se puede mencionar [16] el cual analiza la velocidad extrema de dicho fenómeno en el territorio cubano, ya que obtener este tipo de estimaciones es de suma importancia por ejemplo para el diseño estructural.

Otros trabajos han combinado la teoría de valores extremos con otro tipo de metodologías como ser cópulas, clustering, y otros. Se puede mencionar por ejemplo [39], cuyo objetivo fue obtener un modelo espacio-temporal para las precipitaciones diarias máximas en el Estado de Guanajuato, México. En dicho trabajo se implementó la teoría de valores extremos multivariados a través de procesos max-estables y también la modelización multidimensional de la dependecia a partir de cópulas extremas. En [3] se combina la teoría de valores extremos con metodologías de clustering, para analizar las precipitaciones extremas en Francia. Se comparan las metodologías de clustering *k-means* con PAM, utilizando como distancia la función F-madograma. En [2] se utiliza la teoría de valores extremos en dos etapas, para estimar las precipitaciones extremas en Francia a nivel espacial, a partir de información satelital de baja y alta resolución. En la primer etapa se ajusta una distribución estadística para vincular la resolución alta y baja de variables en algunos lugares dados. En la segunda etapa, dada esta función link, se simula un proceso max-estable condicional con el cual obtener una estimación de las precipitaciones máximas en cualquier punto de la región

de estudio. Otro trabajo que está en esta línea es [1], donde se utiliza los procesos max-estables condicionales a los valores observados en ciertos puntos de la región en estudio, para obtener una estimación de las precipitaciones extremas en aquellos lugares donde no se posee información. Dicho trabajo se centró en la región sur de Francia y el proceso max-estable condicional utilizado mostraba buena performance siempre que los supuestos de estacionariedad se cumplieran. En [53] se estudiaron las precipitaciones máximas en Bélgica también utilizando teoría de valores extremos e incorporando información respecto de la dependencia espacial, a través del madograma, en el análisis. En dicho trabajo se concluyó que el grado de dependencia de las precipitaciones extremas en ese país varía mucho según tres factores: la distancia entre dos estaciones, la estación (verano o invierno) y la duración de la acumulación de precipitación (por hora, día, mes, etc.).

Un trabajo aplicado a la región de América del Sur relacionados al estudio de eventos extremos es [46]. Estudia el desempeño de ocho modelos climáticos globales acoplados (IPCC AR4), en la simulación de índices anuales de eventos climáticos de temperatura extrema y precipitación en América del Sur. Se compararon dos índices de temperatura extrema y tres índices de precipitación extrema, a partir de información de estaciones meteorológicas durante 1961-2000. Se puede destacar dentro de las conclusiones que para las precipitaciones, el índice mejor representado por los modelos es el R95t, que relaciona las precipitaciones extremas, con el clima local. En [51] se estudió la variabilidad interdecadal observada en la distribución de los eventos de temperatura que superan un determinado umbral, en cinco estaciones meteorológicas de Argentina, durante el período 1941-2000, mediante la aplicación de la teoría de valores extremos. Los resultados arrojaron una disminución en la intensidad de eventos extremos cálidos durante todo el período de estudio, junto con un incremento en su frecuencia de ocurrencia durante los últimos 20 años del siglo XX. Los extremos fríos también muestran una disminución en intensidad. Sin embargo, los cambios en su frecuencia de ocurrencia no son tan consistentes entre las diferentes estaciones estudiadas.

En Uruguay sin embargo, son escasos los trabajos que han estudiado los fenómenos extremos meteorológicos o climáticos. Se destaca el trabajo de [14] en donde se estudió los vientos fuertes y pone en consideración que se revise y actualice la norma UNIT 50-84. Algunos resultados obtenidos resaltan por ejemplo que el comportamiento geográfico de los vientos fuertes difiere del indicado en el mapa nacional de vientos extremos dado por la norma de viento UNIT 50-84. Adicionalmente, se obtuvieron resultados que evidenciaron que la distribución de vientos extremos promediados en 10min para Montevideo puede ser modelada adecuadamente por una distribución Gumbel, mientras que la norma UNIT 50-84 propone una distribución Fréchet para las ráfagas de viento.

3 Objetivos

El trabajo se llevó a cabo con la finalidad de modelar las precipitaciones extremas anuales acumuladas en 24 horas (diarias) en Uruguay así como investigar la existencia de patrones espaciales de dicho fenómeno. Se contó con una base de datos diarios de precipitaciones en el período 1981 a 2013 de 19 estaciones meteorológicas y 1 estación pluviométrica. En ese marco se plantearon dos objetivos:

1. Estudiar la distribución de valores extremos de las precipitaciones en cada una de las estaciones meteorológicas, así como estimar los niveles de precipitaciones extremas para ciertos períodos de retorno como ser 10, 20, 50 o 100 años. La teoría de valores extremos proporciona un modelo teórico para representar el comportamiento de los máximos registrados en diferentes ubicaciones. El desafío que presenta estudiar estadísticamente los fenómenos extremos, radica en que las características respecto de los procesos espaciales implícitos no se comportan según una distribución Gaussiana, sino que por el contrario hay que acudir a la teoría de valores extremos para su caracterización. Diversas bibliografías hacen un desarrollo profundo respecto de teoría de valores extremos ([44], [13], [11], [10]), también se pueden encontrar diversos estudios relacionados a la estadística espacial ([9], [19]) entre otros.

2. Identificar patrones espaciales de las precipitaciones extremas en Uruguay. Para ello se recurre a metodologías de spatial clustering. El análisis de clustering clásico tiene por objetivo conformar grupos de acuerdo a ciertas caracterterísticas que pueden ser de interés, de manera que dentro de cada grupo los elementos sean lo más homogéneos posible. La homogeneidad (y heterogeneidad) estará medida a través de una distancia pre-definida, diversas bibliografías profundizan estas técnicas como por ejemplo [29] y [18]. Encontrar patrones espaciales con métodos basados en desviaciones de la media (i.e. varianza) y utilizando la distancia Euclidiana (L2), puede no ser la estrategia más apropiada en el contexto de estudio de valores extremos. La mayoría de los métodos clásicos de agrupaciones calculan los nuevos centroides en cada paso del algoritmo promediando las observaciones dentro de cada grupo. Los promedios de las observaciones normalmente distribuidas siguen siendo gaussianos, pero promedios de valores máximos no se comportan de igual manera. Una estrategia interesante para enfrentar este reto es utilizar un algoritmo de agrupamiento denominado Partitioning Around Medoids (PAM) propuesto por Kaufman y Rousseeuw (ver [34]). A su vez se utilizará una distancia basada en una medida de la dependencia espacial no paramétrica denominada F-madogram (propuesta por [7] y [41]).

4 Climatología de precipitaciones en Uruguay

Uruguay se encuentra situado en la parte oriental del cono sur americano. Limita al noreste con Brasil, al oeste y suroeste con Argentina y tiene costas sobre el Río de la Plata por el sur y costas sobre el océano Atlántico por el sureste. Posee una superficie terrestre de 176.215 km^2 siendo el segundo país más pequeño de Sudamérica, después de Surinam. El relieve de Uruguay se caracteriza por poseer regiones de penillanuras y llanuras surcadas por cuchillas. El clima es templado, con una temperatura media de 17,5 ºC, siendo enero el mes más cálido, con una media de 22,6 ºC, y julio el mes más frío, con una media de 10,6 ºC. Las lluvias son abundantes y varían de los casi 1000 mm acumulados anuales en el sur a los 1500 mm acumulados anuales en el norte, en la frontera con Brasil.

A continuación se presenta un mapa de las precipitaciones acumuladas a diciembre del año 2017:

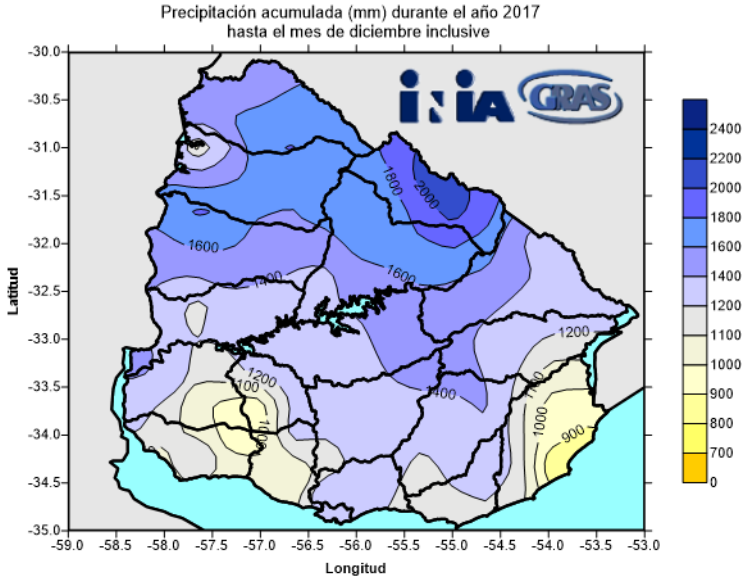


Figura 4.1: Precipitaciones acumuladas en el año 2017. Fuente: INIA

Como puede verse en Figura 4.1, a pesar de que Uruguay no posee cambios en relieve significativos, las precipitaciones no se comportan de la misma forma en dicho territorio. Por ejemplo en el año 2017, la zona norte presentó una mayor acumulación de precipitaciones, mientras que la zona sureste y suroeste son las zonas con menor presencia de lluvias durante ese año.

En esta tesis, se desea estudiar y analizar si el comportamiento de las lluvias extremas también cambia conforme a la localización geográfica o época del año, utilizando metodologías que han resultado exitosas en otros países. Uno de las mayores consecuencias de las precipitaciones extremas viene dada por las inundaciones que las mismas provocan.

Se puede destacar por ejemplo la inundación de 1959, hecho histórico ocurrido en abril de dicho año. Puntualmente, las lluvias comenzaron el 24 de marzo y fueron intensas y persistentes hasta fines de abril. La represa de Rincón del Bonete sufrió severas consecuencias. Según registros encontrados se dice que en el norte del país el promedio de lluvias mensual del mes de abril ascendió a 600 mm (cuando en general el promedio de lluvias acumuladas mensuales es de 112 mm aproximadamente). En la zona de Tacuarembó Chico se registró el máximo absoluto de 1.200 mm siendo superior al promedio anual en esta misma zona (1.100 mm).

Otras inundaciones que han afectado distintas localidades de Uruguay dentro del período de estudio (1981 a 2013), debido a las intensas lluvias, han sido las inundaciones de Agosto 1986, Junio 1992, Abril 1998, Mayo 2000, Junio 2001, Abril 2002, Mayo 2007, Noviembre 2009.

A continuación se muestran los registros de las precipitaciones acumuladas mensuales (Acum. Mensual) para algunos de los períodos de inundaciones mencionados anteriormente, para 20 estaciones meteorológicas del país. Dichos valores se comparan con los registros de precipitaciones máximas diarias del mes en cuestión (Max. Mensual) y con los valores de precipitaciones acumulados promedios del mes, obtenido en el período 1981 a 2013 (Prom. Mensual).

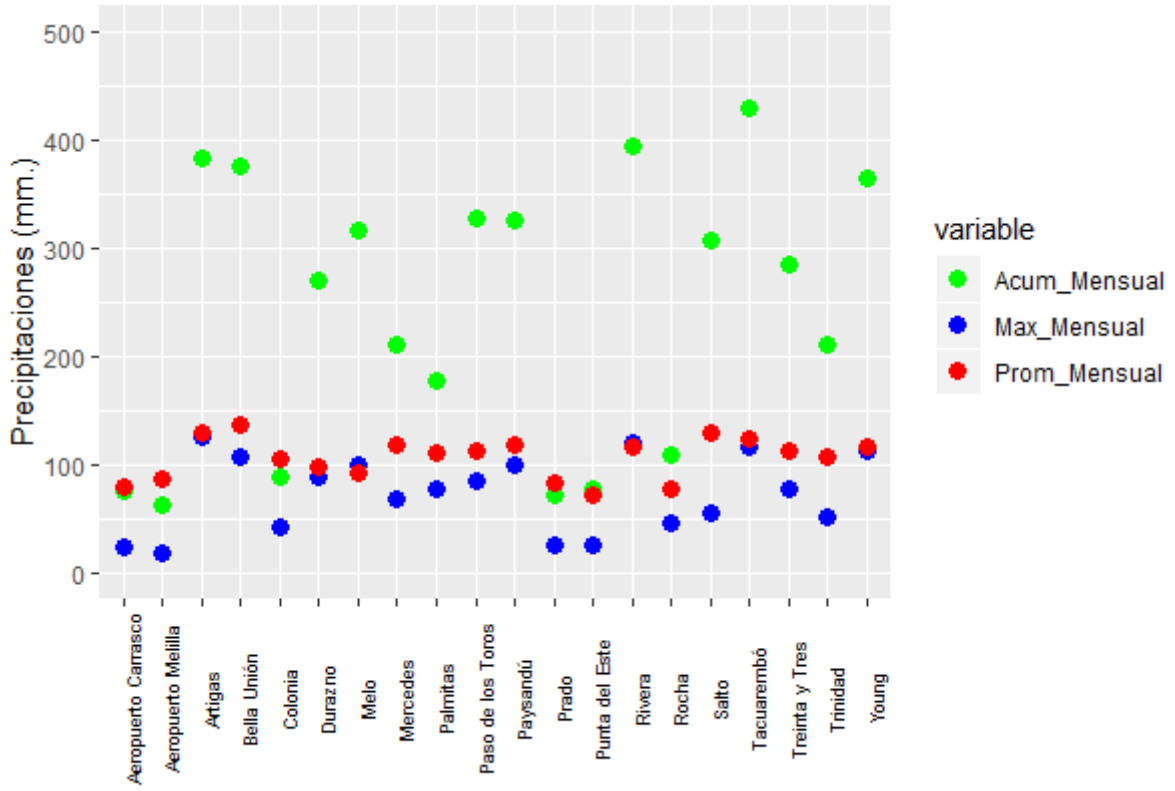


Figura 4.2: Inundaciones de Abril 2002. Fuente: elaboración propia en R.

Del gráfico 4.2 se visualiza cómo los valores de precipitaciones acumulados del mes superaron en la gran mayoría de las estaciones meteorológicas el valor promedio correspondiente a cada una por el doble o hasta el triple de dicho valor como es el caso de Tacuarembó. Sólo los registros de las estaciones de Aeropuerto de Carrasco y Melilla, Colonia y Prado no superaron el valor promedio respectivamente. En dichas regiones es donde menos afectación hubo de las precipitaciones de dicho año y corresponden a estaciones de la región sur del país. Mientras que los valores de lluvias promedios mensuales no superaron los 150 mm, en el año 2002 para ese mes, los valores acumulados alcanzaron valores cercanos a los 400 o más mm, siendo la región de Tacuarembó una de las más afectadas alcanzando el registro máximo de lluvia acumulada de 430.6 mm. También se destaca que en los casos de Artigas, Durazno, Melo, Rivera, Tacuarembó y Young el valor de precipitaciones máxima diaria alcanzó el valor de lluvia acumulada promedio del mes.

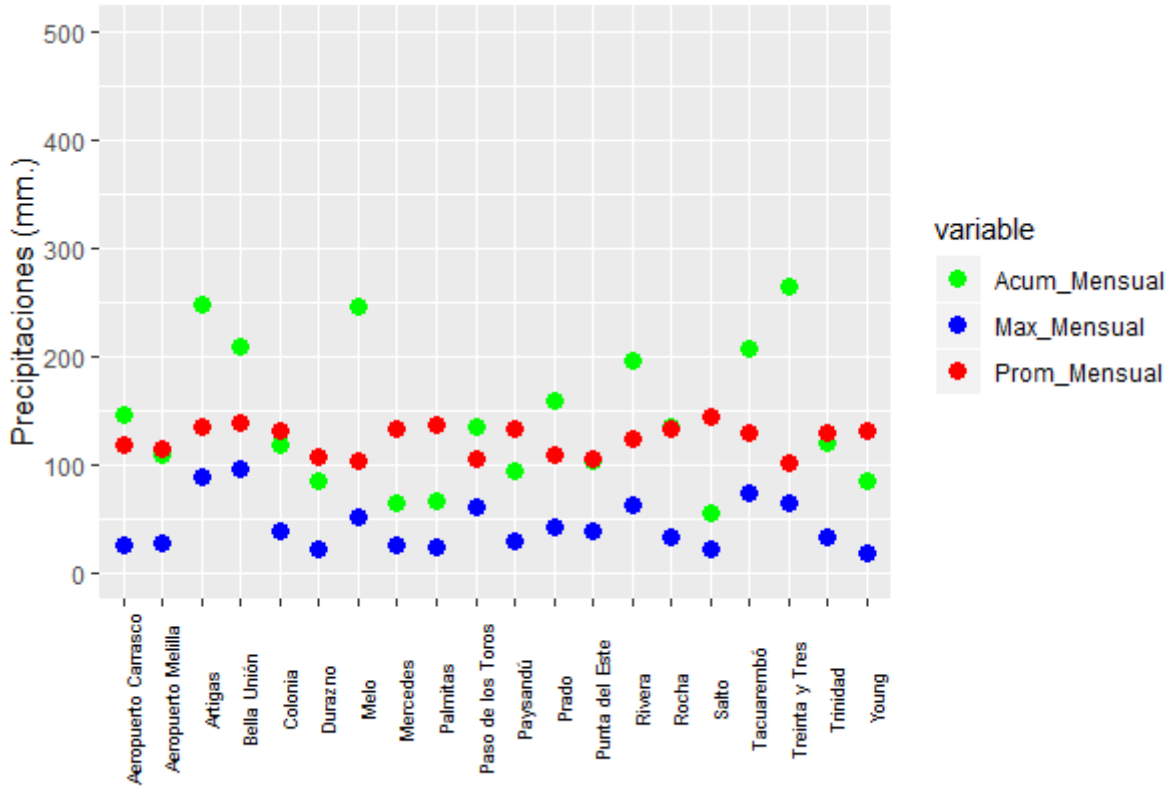


Figura 4.3: Inundaciones de Junio de 1992. Fuente: elaboración propia en R.

Del gráfico 4.3 se visualiza que el valor máximo acumulado mensual se registró en la estación Tacuarembó. También las precipitaciones acumuladas del mes en cuestión, para la mayoría de las estaciones de la región norte (Artigas, Bella Unión, Melo, Paso de los Toros, Rivera, Tacuarembó y Treinta y Tres), estuvieron por encima del promedio mensual correspondiente. De la región sur, sólo las estaciones meteorológicas de Aeropuerto de Carrasco y Prado, presentaron valores de lluvias acumuladas en el mes por encima del valor promedio. También es interesante notar que para las estaciones del norte del país, Artigas, Bella Unión, Tacuarembó y Treinta y Tres los valores de lluvias máximas diarias del mes se acercaron significativamente a los valores de los promedios mensuales correspondientes.

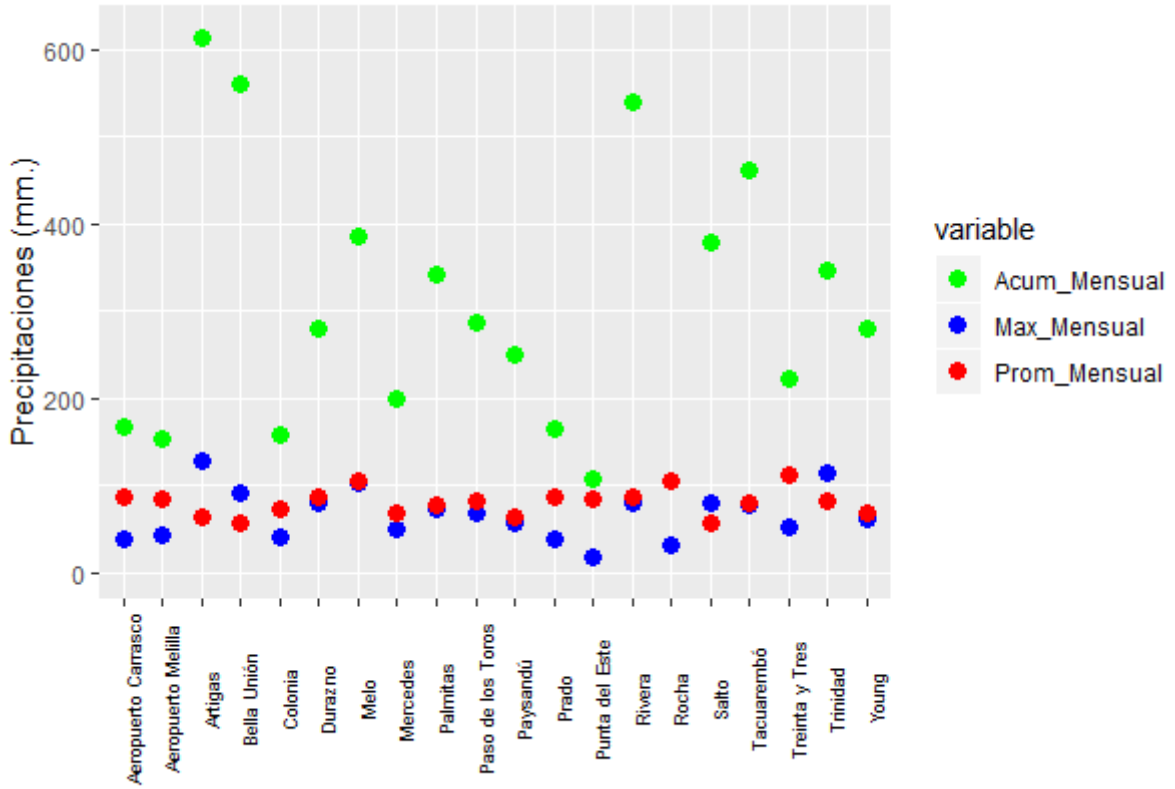


Figura 4.4: Inundaciones Noviembre de 2009. Fuente: elaboración propia en R.

De la Figura 4.4 se visualiza que en las 20 estaciones meteorológicas el valor acumulado en dicho mes superó ampliamente el valor acumulado promedio. Las localidades que presentaron valores acumulados por encima de los 250 mm fueron las de Artigas, Bella Unión, Durazno, Melo, Palmitas, Paso de los Toros, Paysandú, Salto, Tacuarembó y Trinidad, registrándose en Artigas el valor máximo acumulado de 613.3 mm. Notar que para este mes, en las localidades de Artigas, Bella Unión, Salto y Trinidad, los registros de lluvias máximas diarias de dicho mes, superaron el valor del promedio acumulado correspondiente.

Como se vió en gráficos anteriores, los comportamientos de las lluvias extremas son distintas en función de la localidad geográfica del país así como del mes del año en que se estudie. Esta tesis intentará aportar en ese sentido iniciando la investigación con el estudio de las lluvias diarias extremas anuales.

En la segunda parte, se verá el desarrollo metodológico, por una lado la teoría de valores extremos y por otro la teoría de análisis de clustering. En la tercer parte se presentarán los datos con los que se trabajará y los resultados obtenidos. Por último, en la cuarta parte, se presentarán las conclusiones obtenidas.

Part II
Metodologías

5 Teoría de valores extremos

5.1 Introducción

La teoría de valores extremos proporciona las herramientas necesarias para la correcta extrapolación de información relacionada a valores extremos de una variable. Es decir, dada cierta distribución desconocida, denotada como F, se busca realizar inferencia sobre los valores asociados a la cola de dicha distribución. Ello presenta varios inconvenientes dado que es probable que se cuente con pocas observaciones asociados a la cola de F y las técnicas clásicas de estimación resultan eficientes si se posee una cantidad suficiente de datos. También es posible que se quiera estimar valores por encima de los máximos de la propia muestra.

Uno de los intereses de esta tesis, se centra en estudiar la distribución del máximo de precipitación anual registrado en un día, para una estación dada. Es así que, fijada una estación, se define $M_n = max\{X_1, X_2, ..., X_n\}$ el máximo anual de precipitación registrada en $n = 365$ días, con X_i la precipitación acumulada en el día i en dicha estación. Sea $\{X_1, X_2, ..., X_n\}$ una cantidad finita de variables aleatorias que se suponen independientes e idénticamente distribuídas. Si se conociera la distribución exacta de X_i $(F(x))$ se puede calcular exactamente la distribución de M_n ya que

$$P(M_n \leq z) = P(X_1 \leq z, X_2 \leq z, ..., X_n \leq z) = P(X_1 \leq z)P(X_2 \leq z)...P(X_n \leq z) = (F(z))^n. \quad (5.1)$$

En la práctica no se conoce la distribución de X_i, y si así fuera la expresión analítica de F^n podría ser muy compleja. La teoría de valores extremos muestra que bajo ciertas condiciones se puede aproximar el comportamiento de M_n cuando $n \to \infty$.

Primero, probaremos que en el caso i.i.d la distribución del máximo converge casi seguramente a una distribución degenerada.

Proof. Definimos los puntos extremos del soporte de la distribución F como

$$\alpha(F) = inf\{x : F(x) > 0\} \geq -\infty.$$
$$\omega(F) = sup\{x : F(x) < 1\} \leq \infty.$$

$$\text{Caso A} : Si\, x < \omega(F) \Rightarrow F(x) < 1 \Rightarrow (F(x))^n \longrightarrow 0 \quad \text{cuando} \quad n \to \infty.$$
$$\text{Caso B} : Si\, x \geq \omega(F) \Rightarrow F(x) = 1 \Rightarrow (F(x))^n = 1.$$

Demostramos que $lim_{n\to\infty} P(|M_n - \omega(F)| > \epsilon) = 0 \quad \forall \epsilon > 0.$

$$|M_n - \omega(F)| > \epsilon \Leftrightarrow M_n - \omega(F) \leq -\epsilon \quad \text{ó} \quad M_n - \omega(F) \geq \epsilon.$$

Observar que $P(M_n - \omega(F) \leq -\epsilon) = P(M_n \leq \omega(F) - \epsilon)$ y por lo visto en el caso A, dicha probabilidad tiende a 0 cuando n tiende a infinito.

Similarmente $P(M_n - \omega(F) \geq \epsilon) = P(M_n \geq \omega(F) + \epsilon) = 1 - P(M_n < \omega(F) + \epsilon)$ y por lo visto en caso B, dicha probabilidad tiende a 0 cuando n tiende a infinito.

Entonces se puede afirmar que $lim_{n\to\infty} P(|M_n - \omega(F)| > \epsilon) = 0 \ \forall \epsilon > 0$, es decir que M_n converge en probabilidad al supremo del soporte de la distribución F.

Como M_n es una sucesión creciente entonces la convergencia en probabilidad implica convergencia casi segura. □

Para evitar la convergencia a una distribución degenerada, es que se normaliza convenientemente, la variable M_n de la siguiente forma

$$M_n^* = \frac{M_n - b_n}{a_n}, \tag{5.2}$$

siendo $\{a_n\}$ y $\{b_n\}$ sucesiones reales, tales que $a_n > 0$ para todo n.

Al igual que en el Teorema Central del Límite se busca sucesiones $\{a_n\}$ y $\{b_n\}$ tales que

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = (F(a_n x + b_n))^n \to H(x), \tag{5.3}$$

para $n \to \infty$ siendo $H(x)$ no degenerada.

Teorema 5.1. Si existen sucesiones de constantes $\{a_n\}$ y $\{b_n\}$, con $a_n > 0$ para todo n tales que

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to H(x), \tag{5.4}$$

cuando $n \to \infty$ siendo H una funcion distribución no degenerada, entonces H pertenece a alguna de las siguientes familias:

Tipo I: Distribución Gumbel

$$H_1\left(x; \mu, \sigma\right) = e^{-e^{\frac{\mu - x}{\sigma}}} \text{ para } \sigma > 0. \tag{5.5}$$

Tipo II: Distribución Fréchet

$$H_2\left(x; \mu, \sigma, \xi\right) = \begin{cases} e^{-\left(\frac{x-\mu}{\sigma}\right)^{-1/\xi}} & \text{si} \quad x > \mu \\ 0 & \text{si} \quad \text{no} \end{cases} \text{ para } \xi, \sigma > 0. \tag{5.6}$$

Tipo III: Distribución Weibull

$$H_3\left(x; \mu, \sigma, \xi\right) = \begin{cases} e^{-\left(\frac{\mu - x}{\sigma}\right)^{-1/\xi}} & \text{si} \quad x < \mu \\ 1 & \text{si} \quad \text{no} \end{cases} \text{ para } \sigma > 0, \ \xi < 0, \tag{5.7}$$

Este teorema fue originalmente planteado por Fisher y Tippett en 1928 [17] y formalizado luego por Gnedenko en 1943 [22].

Dichas distribuciones se denominan Distribuciones de Valores Extremos (DVE). Las mismas dependen de un parámetro de locación (ubicación) μ, un parámetro de escala $\sigma > 0$ y las distribuciones II y III dependen además de un parámetro de forma ξ.

Cabe destacar que dicho teorema no garantiza la existencia de sucesiones tales que se cumpla (5.3) para un límite no degenerado para M_n ni tampoco nos dice cuál es el límite cuando existe. Lo que sí nos indica el teorema anterior es que de existir dichas sucesiones tales que M_n^* es convergente, la distribución límite puede ser únicamente perteneciente a alguna de las tres familias mencionadas anteriormente y ésto es independiente de la distribución F de los datos.

Las tres familias mencionados anteriormente pueden resumirse en una única expresión

$$H\left(x;\mu,\sigma,\xi\right) = e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}} \text{ siendo } \sigma > 0,\ \xi \neq 0, \tag{5.8}$$

siendo $x_+ = \max\{0, x\}$ (parte positiva de x).

En el caso que $\xi > 0$ la distribución será Fréchet, mientras que si $\xi < 0$ la distribución será Weibull.

Notar que cuando $\xi \to 0$

$$\lim_{\xi\to 0}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi} = e^{-\left(\frac{x-\mu}{\sigma}\right)} = e^{\frac{\mu-x}{\sigma}},$$

de donde se deduce que si ξ tiende a cero en $H\left(x;\mu,\sigma,\xi\right)$ queda la fórmula de la distribución Gumbel $\left(\mu,\sigma\right)$.

De lo anterior se tiene que la expresión única que abarca las distribuciones de tipo I, II y III, queda de la siguiente forma:

$$\begin{aligned} &H\left(x;\mu,\sigma,\xi\right) = e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}} \text{ siendo } \sigma > 0,\ \xi \neq 0, \\ &H\left(x;\mu,\sigma,0\right) = e^{-e^{\frac{\mu-x}{\sigma}}}. \end{aligned} \tag{5.9}$$

Dicha fórmula simplificada recibe el nombre de Distribución de Valores Extremos Generalizada (GEV por sus siglas en inglés), y fue propuesta en [55] y [30]. Esta expresión simplifica en gran medida la implementación estadística ya que no es necesario realizar hipótesis sobre un tipo de distribución DEV particular sino que, a partir de la inferencia sobre ξ queda determinado el comportamiento de la cola de la distribución en cuestión.

Notar que cuando $\xi > 0$ o $\xi < 0$ la distribución límite es igual a la distribución Fréchet o Weibull respectivamente, pero no con los mismos μ y σ. El parámetro ξ es el mismo en la función $H\left(x,\mu,\sigma,\xi\right)$ que en la Fréchet o Weibull según sea el caso.

En el caso $\xi > 0$, vemos que el exponente de la ecuación (5.8) queda:

$$1+\xi\frac{x-\mu}{\sigma}=\frac{x-\mu+\sigma/\xi}{\sigma/\xi},$$

y según la ecuación (5.6) cuando $\xi>0$, la distribución $H\left(x;\mu,\sigma,\xi\right)$ queda igual a la de una Fréchet$\left(\mu-\sigma/\xi,\sigma/\xi,\xi\right)$.

De igual forma, para el caso $\xi<0$ queda lo siguiente:

$$1+\xi\frac{x-\mu}{\sigma}=\frac{\mu-\sigma/\xi-x}{-\sigma/\xi},$$

de donde según la ecuación (5.7) se deduce que $H\left(x;\mu,\sigma,\xi\right)$ =Weibull$\left(\mu-\sigma/\xi,-\sigma/\xi,\xi\right)$.

En suma se tiene que:

$$\begin{aligned}
H\left(x;\mu,\sigma,\xi\right) &= \text{Fréchet}\left(\mu-\sigma/\xi,\sigma/\xi,\xi\right) \text{ para } \xi>0,\\
H\left(x;\mu,\sigma,\xi\right) &= \text{Gumbel}\left(\mu,\sigma\right) \text{ para } \xi=0,\\
H\left(x;\mu,\sigma,\xi\right) &= \text{Weibull}\left(\mu-\sigma/\xi,-\sigma/\xi,\xi\right) \text{ para } \xi<0.
\end{aligned}$$

Observar que

$$\begin{aligned}
\text{Fréchet}\left(\mu,\sigma,\xi\right) &= \mu+\sigma\,\text{Fréchet}\left(0,1,\xi\right),\\
\text{Gumbel}\left(\mu,\sigma\right) &= \mu+\sigma\,\text{Gumbel}\left(0,1\right),\\
\text{Weibull}\left(\mu,\sigma,\xi\right) &= \mu+\sigma\,\text{Weibull}\left(0,1,\xi\right).
\end{aligned}$$

Se puede verificar directamente que mediante ciertas transformaciones, que se resumen a continuación, se puede obtener una distribución DVE a partir de otra DVE.

1. (Fréchet a Gumbel). Si $X\sim$Fréchet$\left(\mu,\sigma,\xi\right)$ entonces $\log\left(\frac{X-\mu}{\sigma}\right)^{1/\xi}\sim$Gumbel$(0,1)$.
2. (Fréchet a Weibull). Si $X\sim$Fréchet$\left(\mu,\sigma,\xi\right)$ entonces $\frac{X-\mu}{\sigma}\sim$Weibull$(0,1,-\xi)$.
3. (Weibull a Fréchet). Si $X\sim$Weibull$\left(\mu,\sigma,\xi\right)$ entonces $\frac{\sigma}{\mu-X}\sim$Fréchet$(0,1,-\xi)$.
4. (Weibull a Gumbel). Si $X\sim$Weibull$\left(\mu,\sigma,\xi\right)$ entonces $\log\left(\frac{\sigma}{\mu-X}\right)^{-1/\xi}\sim$Gumbel$(0,1)$.
5. (Gumbel a Weibull). Si $X\sim$Gumbel$\left(\mu,\sigma\right)$ entonces $-e^{\xi\left(\frac{X-\mu}{\sigma}\right)}\sim$Weibull$(0,1,\xi)$.
6. (Gumbel a Fréchet). Si $X\sim$Gumbel$\left(\mu,\sigma\right)$ entonces $e^{\xi\left(\frac{X-\mu}{\sigma}\right)}\sim$Fréchet$(0,1,\xi)$.

Si bien el Teorema 5.1 no brinda herramientas para determinar las sucesiones de normalización, resulta útil la siguiente equivalencia (válida para valores grandes de n):

$$P\left(\frac{M_n-b_n}{a_n}\leq z\right)\approx H(z), \tag{5.10}$$

que equivalente a

$$P(M_n \leq z) \approx H\left(\frac{z-b_n}{a_n}\right) = H^*(z), \tag{5.11}$$

siendo $H^*(z)$ otro miembro de la familia GEV cuando $n \to \infty$.

Esto permite aproximar la distribución de M_n por una familia de distribuciones con parámetros de ubicación y escala.

Respecto de la unicidad de las sucesiones de normalización, el Teorema de Convergencia a familias, de Gnedenko y Khinchin [23] nos brinda la respuesta:

Teorema 5.2. Sean $G(x)$ y $H(x)$ dos funciones de distribución ninguna de las cuales está concentrada en un punto. Supongamos que para $n \geq 0$, X_n son v.a. con funciones de distribución F_n. Sean, además, sucesiones $\{a_n\}$, $\{b_n\}$, $\{\alpha_n\}$ y $\{\beta_n\}$ pertenecientes a $\Re$, siendo $a_n > 0$, $\alpha_n > 0$ para todo n. Se plantean las siguientes tres afirmaciones

1. $F_n(a_n x + b_n) \to G(x)$ y $F_n(\alpha_n x + \beta_n) \to H(x)$ cuando $n \to \infty$ para todo x.
2. Existen $A > 0$ y $B \in \Re$ tales que $H(x) = G(Ax + B)$ para todo x.
3. $\frac{\alpha_n}{a_n} \to A$ y $\frac{\beta_n - b_n}{a_n} \to B$ cuando $n \to \infty$.

Si vale (1) entonces valen (2) y (3). Además si (3) vale, entonces cualquiera de las relaciones en (1) implica la otra y (2) vale.

Es decir, el teorema anterior demuestra que las sucesiones de normalización están determinadas por equivalencias asintóticas, y la distribución límite está determinada salvo por parámetros de ubicación y escala.

Por otro lado el siguiente corolario indica que dichas sucesiones no son únicas.

Corolario 5.2.1. Sea F_n una sucesión de funciones de distribución y $\{a_n\}$ siendo $a_n > 0$ para todo n y $\{b_n\}$ sucesiones de reales tales que

$$F_n(a_n x + b_n) \to G(x), \tag{5.12}$$

en todo punto de continuidad de G, que es una función distribución propia y no está concentrada en un punto. Sean $\{c_n\}$ siendo $c_n > 0$ para todo n y $\{d_n\}$ sucesiones de reales tales que:

$$\frac{a_n}{c_n} \to 1,$$

y

$$\frac{d_n - b_n}{a_n} \to 0.$$

Entonces (5.12) vale con $\{c_n\}$ y $\{d_n\}$ en lugar de $\{a_n\}$ y $\{b_n\}$.

La siguiente definición nos permite confirmar que la distribución del máximo es una distribución GEV:

Definición 5.1. Una función distribución F es max-estable si para cada n existen sucesiones $\{b_n\}$ y $\{a_n\}$ con $a_n > 0$ para todo n, tales que $(F(a_n x + b_n))^n = F(x)$ para todo $x \in \Re$.

Por lo tanto, una función distribución F es max-estable cuando para cada n existe una función lineal tal que la distribución del máximo de variables i.i.d con distribución F evaluada en dicha función lineal es F. La conexión con las DVE está dada por el siguiente teorema:

Teorema 5.3. Las únicas distribuciones max-estables son las Distribuciones de Valores Extremos.

La demostración de este teorema se puede encontrar en el Apéndice A de [39].

También es interesante conocer bajo qué condiciones sobre F se tiene convergencia a un límite no-degenerado. Es decir, investigar qué condiciones debe poseer F tal que $P(M_n \leq u_n)$ sea convergente cuando $n \to \infty$ para una sucesión apropiada.

Definición 5.2. Decimos que la función de distribución F está en el dominio de atracción de la distribución de valor extremo H ($F \in D(H)$) si existen sucesiones reales $\{a_n\}$ y $\{b_n\}$ tales que $a_n > 0$ para todo n y se cumpla que:

$$(F(a_n x + b_n))^n = P(M_n \leq a_n x + b_n) \to H(x), \quad \forall x \in R. \tag{5.13}$$

Resulta de interés obtener condiciones necesarias y suficientes para determinar si una F pertenece al dominio de atracción de alguna distribución DEV.

Proposición 5.3.1. F pertenece al dominio de atracción de la DVE H con sucesiones de normalización reales $\{a_n\}$ y $\{b_n\}$, con $a_n > 0$ para todo n, si y solo si:

$$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = -log(H(x)) \quad x \in R.$$

Cuando $H(x) = 0$ el límite se interpreta como ∞.

Definición 5.3. Dos funciones F y G son asintóticamente equivalentes si tienen el mismo extremo derecho, es decir $\omega(F) = \omega(G)$ y

$$\lim_{x\to\omega(F)} \frac{(1 - F(x))}{(1 - G(x))} = c \quad c \in (0, \infty).$$

Es interesante notar que $F, G \in D(H)$ si y sólo si son asintóticamente equivalentes y además es posible utilizar las mismas constantes de normalización.

5.2 Inferencia para las distribuciones GEV

Máximo por bloques

Dada una colección de datos medidos en distintos momentos de tiempo, los que agrupamos en conjuntos disjuntos de datos consecutivos y de igual longitud. Cada conjunto contiene la información correspondiente a un período de tiempo s, por ejemplo un año, un mes, etc. Por lo tanto los datos originales los observamos en bloques:

$$\begin{aligned}\mathbf{X}^{(1)} &= \left(X_1^{(1)}, ..., X_s^{(1)}\right)\\ \mathbf{X}^{(2)} &= \left(X_1^{(2)}, ..., X_s^{(2)}\right)\\ &\vdots\\ \mathbf{X}^{(n)} &= \left(X_1^{(n)}, ..., X_s^{(n)}\right).\end{aligned}$$

La elección del tamaño (s) del bloque resulta de un balance entre sesgo y varianza. Por un lado bloques de tamaño pequeño provocarán grandes sesgos mientras que bloques de gran tamaño provocarán aumento de varianza. Es casi una convención que se adopte un tamaño de bloque de 1 año, sobre todo considerando datos diarios de fenómenos meteorológicos, los cuales suelen ser estacionales intra anualmente. Para no violar el supuesto de equidistribución es que se toma como tamaño adecuado en estos casos un año.

El intervalo de tiempo se escoge de manera que los vectores $\mathbf{X}^{(i)}$ sean i.i.d (aunque los componentes de cada vector sean dependientes).

La muestra i.i.d. con la que se hará inferencia es:

$$M_i = max\{X_1^{(i)}, ..., X_s^{(i)}\} \quad i = 1, ..., n. \tag{5.14}$$

Las densidades correspondientes a las distribuciones de valores extremos son las siguientes:

- Tipo I: $h_1(x) = \frac{1}{\sigma} e^{\left(\frac{\mu - x}{\sigma}\right)} H_1(x)$.
- Tipo II: $h_2(x) = \frac{1}{\sigma}\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}-1} H_2(x)$ si $x \geq \mu$.
- Tipo III: $h_3(x) = \frac{1}{\sigma}\left(1 + \xi\frac{\mu-x}{\sigma}\right)^{-\frac{1}{\xi}-1} H_3(x)$ si $x < \mu$.

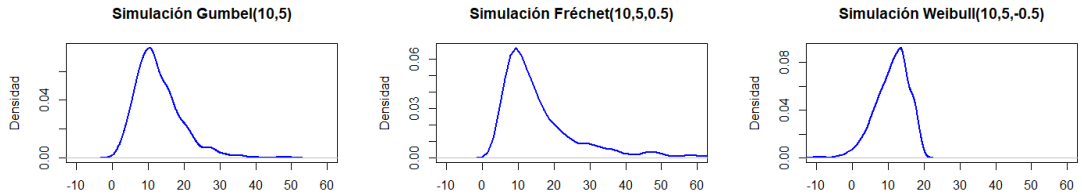
Siendo H_1, H_2 y H_3 según se vió en 5.1.

Si se parte de la ecuación (5.8) la expresión analítica de las densidades vista anteriormente sería:

$$h(x) = \frac{1}{\sigma}\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-(1+1/\xi)} H(x) \quad \text{si} \quad 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0 \quad \xi \neq 0. \tag{5.15}$$

Cuando $\xi \to 0$ entonces $h(x) \to h_1(x)$ con la ecuación según expresión anterior.

A continuación se muestran gráficas de densidades DEV simuladas:



Diversos métodos se pueden utilizar para hacer inferencia sobre la distribución de los máximos M_i. En este trabajo se detallará la estimación de parámetros por máxima verosimilitud, verosimilitud perfil y la estimación de parámetros por el método de los momentos pesados.

5.2.1 Estimación por máxima verosimilitud

Sea $Z_1, Z_2, ..., Z_n$ v.a i.i.d con distribución GEV con parámetros $\theta = (\mu, \sigma, \xi)$ y función densidad $h(x, \theta)$ definida en (5.15). La función verosimilitud se expresa de la siguiente forma:

$$L(\mu, \sigma, \xi | Z) = \prod_{i=1}^{n} h(z_i, \theta) I_{\{1+\xi(z_i-\mu)/\sigma>0\}}. \tag{5.16}$$

Entonces la log-verosimilitud denotada como $log(L(\mu, \sigma, \xi | Z)) = \ell(\mu, \sigma, \xi | Z)$ será:

$$\ell(\mu, \sigma, \xi | Z) = -nlog(\sigma) - (1 + 1/\xi)\Sigma_{i=1}^{n} log(1 + \xi(z_i - \mu/\sigma)) - \Sigma_{i=1}^{n}(1 + \xi(z_i - \mu/\sigma))^{-1/\sigma}.$$

En el caso que $\xi = 0$, la log-verosimilitud queda de la siguiente forma:

$$\ell(\mu, \sigma | Z) = -nlog(\sigma) - \Sigma_{i=1}^{n}(1 + (z_i - \mu/\sigma)) - \Sigma_{i=1}^{n} e^{-(z_i-\mu/\sigma)}.$$

El estimador de máxima verosimilitud para $\theta = (\mu, \sigma, \xi)$ es:

$$\hat{\theta}_n = argmáx_{\theta \in \Theta} \ell(\theta | Z). \tag{5.17}$$

Notar que no existe solución analítica de los estimadores máximos verosímiles de (μ, σ, ξ), aunque con algoritmos de optimización se logra obtener dichas estimaciones. En [48] se prueba que las condiciones de regularidad de los estimadores obtenidos mediante este método se mantienen si $\xi > -0.5$ y en ese caso los estimadores presentan las propiedades asintóticas deseadas (consistencia y normalidad asintótica). Si $-1 < \xi < -0.5$ se pueden obtener los estimadores máximo verosímiles pero no tendrán las propiedades asintóticas deseadas y si $\xi < -1$ tal vez no se puedan obtener los estimadores por este método.

Bajo las condiciones deseadas la distribución de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ para n suficientemente grande tiene aproximadamente una distribución normal multivariada de media (μ, σ, ξ) y matriz de covarianza igual al inverso de la matriz de información de Fisher observada, evaluada en los estimadores de máxima verosimilitud (ver Apéndice V).

A partir de las estimaciones obtenidas anteriormente, se pueden construir regiones de confianza, es decir aproximar la probabilidad de que dichas regiones contengan el verdadero valor del parámetro. Un método se basa en la función devianza definida como:

$$D(\theta) = 2\{\ell(\hat{\theta}_n) - \ell(\theta)\}. \tag{5.18}$$

La región de confianza se especifica como:

$$C = \{\theta : D(\theta) \leq c\}. \tag{5.19}$$

La elección de c tiene que ver con la idea de que con probabilidad $(1-\alpha)$ la región C debe contener el verdadero valor de θ. En este caso, se recurre a la distribución asintótica de la función devianza que nos da el siguiente teorema:

Teorema 5.4. Sea $Z_1, ..., Z_n$ una muestra i.i.d con dist. GEV con parámetros θ y sea $\hat{\theta} \in \Re^d$ el estimador EMV de θ para $n \to \infty$. Entonces se tiene que $D(\theta)$ converge en distribución a una v.a. χ^2_d.

Es así que c es elegido como el percentil $(1-\alpha)$ de la distribución χ^2_d, ver [6], obteniéndose de esta forma una región de confianza asintótica.

5.2.2 Estimación por verosimilitud perfil

Este método de estimación está basado en que se puede particionar $\theta = (\theta_i, \theta_{-i})$ siendo θ_{-i} todos los componentes (parámetros "estorbo") de θ excluyendo θ_i. De esta forma se estima θ_i de la siguiente forma:

$$\ell_p(\theta_i) = max_{\theta_{-i}} \ell(\theta_i, \theta_{-i}). \tag{5.20}$$

La verosimilitud perfil se obtiene maximizando la función de verosimilitud evaluada en los elementos de Θ con θ_i fijo. Si por ejemplo se desea estimar ξ, se fija $\xi = \xi_0$ y se maximiza la log-verosimilitud con respecto de los parámetros restantes. Se repite este procedimiento para un rango de valores de ξ_0 de forma de obtener los valores correspondientes a la log-verosimilitud perfil para ξ.

De forma similar a lo visto anteriormente, se puede proceder a la estimación de regiones de confianza asintóticas a partir de la función devianza:

$$D_p(\theta_i) = 2\{\ell(\hat{\theta}_n) - \ell_p(\theta_i)\} \quad \text{, la cual es aproximadamente} \quad \chi^2_k, \tag{5.21}$$

siempre que se cumplan que la muestra de v.a. sea i.i.d, $\hat{\theta}_n \in \Re^d$ el estimador EMV de $\theta = (\theta_i, \theta_{-i})$ con $\theta_i \in \Re^k$ y para $n \to \infty$.

Se puede obtener una región de confianza asintótica para θ_i de nivel $1-\alpha$, dada por $C_p = \{\theta_i : D_p(\theta_i) \leq c_p\}$ siendo c_p el percentil $(1-\alpha)$ de la distribución χ^2_k.

También esta idea de particionar el espacio Θ es útil para comparar familias DVE a través del Test Ratio de Verosimilitud, según se plantea en el siguiente teorema:

Teorema 5.5. Sea M_{-i} con parámetros θ_{-i} un submodelo de M con parámetros $\theta = (\theta_i, \theta_{-i})$ con $\theta_i = \mathbf{0}$. Sea $\ell_{-i}(M_{-i})$ y $\ell(M)$ las funciones máximo log-verosímiles de los modelos M_{-i} y M respectivamente. El siguiente test para validar el modelo M_{-i} frente al modelo M consiste

en: a un nivel de significación α rechazo M_{-i} en favor de M si $D = 2\{\ell_{-i}(M_{-i}) - \ell(M)\} > c_\alpha$ siendo c_α el percentil $(1-\alpha)$ de la distribución χ^2_k.

Las ventajas que tiene la verosimilitud perfil como tal, es que es la base de la estadística de prueba de la razón de las verosimilitudes. Además para el caso de valores extremos se ha probado que la función de verosimilitud perfil es robusta frente a cambios pequeños en los estimadores máximo verosímil restringidos de los parámetros estorbo. Otra ventaja de la verosimilitud perfil es que los intervalos de confianza son asimétricos lo cual podría ser más adecuado en un contexto de estudio de valores extremos. Dichos resultados pueden verse en [35].

5.2.3 Estimación por método de los momentos ponderados

Sea $X \sim F(X)$. Los momentos de probabilidad ponderados de X se definen de la siguiente manera:

$$M_{p,r,s} = E\left[X^p \{F(X)\}^r \{1-F(X)\}^s\right], \tag{5.22}$$

siendo p, r, s reales no negativos.

El método por momentos ponderados es introducido en [24]. Se utiliza cuando la inversa de la distribución F, denotado como $x(F)$, tiene una forma cerrada, en ese caso se podría escribir $M_{p,r,s}$ de la siguiente forma:

$$M_{p,r,s} = \int_0^1 \{x(F)\}^p F^r (1-F)^s \, dF. \tag{5.23}$$

Para estimar los parámetros de la distribución F se parte de:

$$M_{1,r,s} = E\left[X\{F(X)\}^r \{1-F(X)\}^s\right]. \tag{5.24}$$

En [26]se demuestra que si se denota a:

$$\beta_r = M_{1,r,0} = E\left[X\{F(X)\}^r\right] \quad r = 0, 1, 2..., \tag{5.25}$$

entonces dada una muestra de tamaño n con distribución F, el estimador de β_r, basado en los estadísticos de orden es:

$$b_r = n^{-1} \sum_{j=1}^{n} \frac{(j-1)(j-2)...(j-r)}{(n-1)(n-2)...(n-r)} X_j. \tag{5.26}$$

Se puede demostrar que el estimador dado por la ecuación anterior es insesgado.

Notar que dicho método es una generalización del método de los momentos clásico ya que si en ecuación (5.22) se toma $r = 0$ y $s = 0$ entonces $M_{p,0,0} = E[X^p]$ corresponden a los momentos de X para $p = 0, 1, 2, ...$

Como ejemplo, se desarrollará a continuación la estimación de los parámetros $\theta = (\mu, \sigma, \xi)$ para una distribución GEV cuando $\xi \neq 0$.

Como se vio anteriormente, si $X \sim GEV(\theta)$ cuando $\xi \neq 0$ la distribución se puede escribir según la ecuación general (5.8) como sigue:

$$H(x) = e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}} \text{ siendo } \sigma > 0,\ \xi \neq 0. \tag{5.27}$$

Por lo tanto, la inversa de dicha distribución se puede escribir como:

$$x(p) = \mu - \sigma/\xi\left[1-(-log(1-p))^{-\xi}\right]. \tag{5.28}$$

Para este caso, se puede escribir el momento de probabilidad ponderada como sigue:

$$\beta_r = (r+1)^{-1}\left[\mu + \sigma\{1-(r+1)^{-\xi}\Gamma(1+\xi)\}\xi^{-1}\right] \quad \xi > -1. \tag{5.29}$$

Siendo Γ la Gamma de Euler.

En [26] en su apéndice A se puede encontrar la demostración de (5.29).

Si por el contrario $\xi \leq -1$ entonces los momentos β_r para $r = 0,1,2,...$ no existen.

Utilizando la fórmula (5.29) para los casos $r = 0$ y $r = 1$ se obtiene:

$$\beta_0 = \mu + \sigma\{1-\Gamma(1+\xi)\}\xi^{-1}, \tag{5.30}$$

$$\beta_1 = 2^{-1}\left[\mu + \sigma\{1-2^{-\xi}\Gamma(1+\xi)\}\xi^{-1}\right]. \tag{5.31}$$

Si se combinan las ecuaciones anteriores se tiene que:

$$2\beta_1 - \beta_0 = \sigma\Gamma(1+\xi)(1-2^{-\xi})\xi^{-1}. \tag{5.32}$$

La estimación de θ corresponde a las soluciones de las ecuaciones anteriores, se obtienen reemplazando β_0, β_1 y β_2 por b_0,b_1 y b_2 respectivamente.

Primero se obtiene la estimación de ξ resolviendo numéricamente la siguiente ecuación:

$$\frac{3b_2 - b_0}{2b_1 - b_0} = \frac{1-3^{-\xi}}{1-2^{-\xi}}. \tag{5.33}$$

Dado $\hat{\xi}$ estimador obtenido en el paso anterior, se puede obtener $\hat{\sigma}$ sustituyendo ξ por $\hat{\xi}$ en la ecuación (5.32). Luego obtenido $\hat{\sigma}$, se obtiene $\hat{\mu}$ a partir de la ecuación (5.30), quedando los estimadores $\hat{\sigma}$ y $\hat{\mu}$ como siguen:

$$\hat{\sigma} = \frac{(2b_1 - b_0)\hat{\xi}}{\Gamma(1+\hat{\xi})(1-2^{-\hat{\xi}})}, \tag{5.34}$$

$$\hat{\mu} = b_0 - \hat{\sigma}\{\Gamma(1+\hat{\xi}) - 1\}\hat{\xi}^{-1}. \tag{5.35}$$

En [26] se demuestra que cuando $\xi < 0.5$ las distribuciones de los estimadores obtenidos a partir del método anterior, convergen a una distribución normal. Una ventaja frente a los métodos de máxima verosimilitud, es que su performance es buena aún en muestras pequeñas.

5.2.4 Test de bondad de ajuste de la distribución GEV

Dada una muestra $X_1, X_2, ..., X_n$ de una variable aleatoria con distribución F_X desconocida, se aplicará un test del tipo Cramér-von Mises recortado similar al planteado en [31], para testear $H_0 : X \sim$Gumbel(μ, σ) contra $H_1 : H_0$ no es cierto. Si bien el test planteado en [31] es para distribuciones Normales, en el presente trabajo se adaptó dicha idea para testear distribuciones del tipo Gumbel.

En las siguientes líneas se describe la idea del test.

Si $X \sim$ Gumbel (μ, σ) entonces $E(X) = \mu + \gamma$ siendo $\gamma \approx 0.577216$ la constante de Euler, y $V(X) = \pi^2\sigma^2/6$.

Por lo tanto, dada la muestra, se definen las nuevas variables $Y_i = \frac{\pi}{\sqrt{6}} \frac{X_i - \overline{X}_n}{S_n} + \gamma$, si H_0 es cierto y n es suficientemente grande, es de esperar que la distribución de las Y_i sea aproximadamente Gumbel estándar (Gumbel$(0,1)$) cuya función de distribución es $F(x) = e^{-e^{-x}}$.

Sea el proceso empírico tipificado, definido como:

$$\widehat{b}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{1}_{\{Y_i \leq x\}} - e^{-e^{-x}} \right). \tag{5.36}$$

Si H_0 es cierto entonces $\widehat{b_n}(x) \to 0$ para todo $x \in \Re$ cuando $n \to +\infty$ y es conocido que el proceso $\left\{\sqrt{n}\widehat{b}_n(x)\right\}_{x\in\Re}$ converge a un proceso gaussiano, que llamamos $\{b(x)\}_{x\in\Re}$.

Por lo tanto, se puede utilizar como estadístico para la región crítica de la prueba de hipótesis a:

$$T_n := n \int_{-\log n}^{\log n} \widehat{b}_n^2(x)\, dx, \tag{5.37}$$

que convergerá débilmente al proceso $\int_{-\infty}^{+\infty} \widehat{b}^2(x)\, dx$ y tomar como región crítica para la prueba a $\{T_n > cte\}$.

Los extremos de integración de $-\log n$ y $\log n$ pueden ser tomados directamente como $-\infty$ y $+\infty$, pero como se ha visto en [31] resulta más robusto en muchos casos si se considera sólo los casos en los cuales $Y_i \in (-\log n, \log n)$, tal como se demuestra a continuación:

Si $X \sim$Gumbel(μ, σ), entonces $F_X^{-1}(1 - 1/n) = -\log\left(-\log(1 - 1/n)\right) \sim \log n$ cuando $n \to \infty$.

Tal como se vio en sección 5.1, se puede realizar transformaciones que permiten el pasaje de una familia a cualquiera de las otras. Utilizando ello como insumo, se describirá el procedimiento para realizar un test de hipótesis del tipo Cramér-von Mises recortado para saber si la distribución en estudio pertenece a alguna de las tres posibles distribuciones límite.

El mismo puede ser llevado a cabo en los siguientes pasos:

1) Se estiman, por alguno de los métodos vistos anteriormente, los parámetros μ, σ y ξ.

2) Se realiza la prueba $H_0 : X \sim$Gumbel(μ, σ) contra $H_1 : H_0$ no es cierto.

3) Si se rechaza H_0 en el paso anterior, hay suficiente evidencia de que los datos no son bien modelados por la familia Gumbel. En dicho caso, se realiza la prueba de hipótesis de ajuste a la distribución Fréchet o Weibull según paso 4.

4) Partiendo de $\widehat{\xi}$, la estimación de ξ realizada en el paso 1 y según su signo se realiza los siguientes test:

- Si $\widehat{\xi} > 0$ entonces se testea $H_0 : X \sim$Fréchet(μ, σ, ξ) contra $H_1 : H_0$ no es cierto. Para realizar dicha prueba se aplica a los datos la transformación número 1 visto anteriormente (de Fréchet a Gumbel segun lo visto en 5.1) bajo la hipótesis de que H_0 es cierto y se aplica el test según paso 2. Si $\widehat{\xi} > 0$, es de esperar que no se rechace la hipótesis nula.
- Si $\widehat{\xi} < 0$ se aplica el test $H_0 : X \sim$ Weibull(μ, σ, ξ) contra $H_1 : H_0$ no es cierto, aplicando a los datos la transformación correspondiente (de Weibull a Gumbel segun lo visto en 5.1) y se lleva a cabo el test del paso 2. Si $\widehat{\xi} < 0$, es de esperar que no se rechace la hipótesis nula.

5.3 Diagnóstico de la estimación de modelos GEV

Sean $x_1, ..., x_n$ realizaciones independientes idénticamente distribuidas según una distribución F siendo F una GVE. Sea $\hat{F} = H(x, \hat{\mu}, \hat{\sigma}, \hat{\xi})$ la estimación de F con $\hat{\mu}, \hat{\sigma}, \hat{\xi}$ estimados mediante alguno de los métodos anteriormente mencionados. El objetivo de esta etapa consiste en evaluar si los x_i forman una muestra aleatoria de F.

Para llevar a cabo este análisis se utilizó el software R ([50]). La librería utilizada para el análisis de valores extremos fue *extRemes*. Se utilizó la función $fevd$ para realizar la estimación del modelo GEV para cada estación. El diagnóstico resulta en una salida de dicha función, con cuatro gráficos: P-P plot, Q-Q plot, densidad empírica vs densidad del modelo teórico y el gráfico de los valores de retorno correspondiente a períodos de retorno determinados. A continuación se detallarán brevemente los gráficos P-P plot y Q-Q plot que son de utilidad para evaluar la estimación obtenida y luego se detallará cómo se estiman los niveles de retorno. Primero se definirá la función de distribución empírica con la cual se construyen los gráficos a explicar.

Definición 5.4. Dada una muestra ordenada de observaciones provenientes de una distribución F:

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)},$$

la distribución empírica $(\overline{F})$ se define como sigue:

$$\overline{F(x)} = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \dfrac{i}{n+1} & \text{si } x_{(i)} \leq x < x_{(i+1)} \quad \text{si} \quad i = 1,2,...,n-1 \\ 1 & \text{si } x \geq x_{(n)}. \end{cases} \tag{5.38}$$

La estimación de F ($\hat{F}$) resulta razonable si $\hat{F}$ y $\overline{F}$ se parecen.

La comparación la llevamos a cabo a partir de dos gráficos: P-P plot y Q-Q plot.

5.3.1 Gráfico de probabilidad: P-P plot

Dicho gráfico resulta de graficar en el eje de las abscisas $\hat{F}(x_{(i)})$ y en el eje de las ordenadas la función de distribución empírica $\overline{F}(x_{(i)}) = \dfrac{i}{n+1}$ para $i = 1,2,...,n$.

$\hat{F}$ es una buena estimación de la distribución poblacional si los puntos de dicho gráfico se encuentran alineados sobre la diagonal $y = x$.

5.3.2 Gráfico de cuantiles: Q-Q plot

Resulta de graficar en el eje de las abscisas $\hat{F}^{-1}\left(\dfrac{i}{n+1}\right)$ y en el eje de las ordenadas $x_{(i)}$.

$\hat{F}$ es una buena estimación de la distribución poblacional si los puntos de dicho gráfico se encuentran alineados sobre la diagonal $y = x$.

Notar que tanto el gráfico P-P plot como el gráfico Q-Q plot contienen la misma información pero en distinta escala.

5.3.3 Estimación de niveles de retorno

En varias áreas de la ciencia (como ser meteorología, ingeniería, economía, y otras), se define el nivel de retorno, como el valor que se espera sea excedido una vez cada determinado período de tiempo. Dicho período de tiempo se denomina el período de retorno.

Estadísticamente se parte de $\{X_i\}$ sucesión de variables aleatorias i.i.d con función distribución F continua. Sea u un umbral dado. Sea $Y_i = I_{\{X_i > u\}}$ entonces $\{Y_i\}$ es una sucesión de variables aleatorias i.i.d Bernoulli tales que:

$$P(Y_i = 1) = 1 - F(u) = p.$$

El primer instante del primer éxito se puede denotar como:

$$L(u) = min\{i \geq 1 : X_i > u\},$$

entonces el instante de la primer excedencia del umbral u es una variable aleatoria Geométrica, es decir es la distribución de probabilidad del número de ensayos Bernoulli necesarios hasta obtener el primer éxito. La distribución de probabilidad puede describirse como sigue:

$$P(L(u) = k) = (1-p)^{k-1}p \qquad k = 1,2,... \tag{5.39}$$

El período de retorno de excedencia del umbral u de los eventos $\{Y_i\}$ se define como:

$$E[L(u)] = 1/p. \tag{5.40}$$

Es decir que, determinar el nivel correspondiente a un período de retorno de t años, corresponde a hallar u tal que $E[L(u)] = t$, es decir que $p = 1/t$ y $F(u) = 1-p$ o sea que $F(u) = 1-\frac{1}{t}$. Dicho en otras palabras, el nivel de retorno se espera sea excedido una vez cada $1/p$ períodos (años por ejemplo).

Dicho de otra forma, el nivel de retorno correspondiente a un período de t años es el cuantil $1-1/t$ de la distribución F.

Por ejemplo, puede llegar a ser de interés, calcular aquel valor de precipitaciones extremas tal que se espera sea excedido una vez cada 50 años. Dicho período de retorno de 50 años, corresponde a una probabilidad de excedencia de 1/50 = 0.02 o 2% para un año cualquiera (es decir, la probabilidad de excedencia para cada año será del 2%).

Para obtener dichos valores estimados, se procede invirtiendo la ecuación (5.8). De esta forma, se puede obtener los niveles de retornos z_p(cuantiles) de la distribución de máximos asociado a un período de retorno $t = 1/p$ tal que $F(z_p) = 1-p$, como sigue:

$$z_p = \begin{cases} \mu - \sigma/\xi\left[1-(-log(1-p))^{-\xi}\right] & \text{si } \xi \neq 0 \\ \mu - \sigma log(-log(1-p)) & \text{si } \xi = 0. \end{cases} \tag{5.41}$$

Si se define $y_p = -log(1-p)$ y se grafica z_p respecto de $log(y_p)$ entonces se puede obtener un gráfico como el que se muestra a continuación:

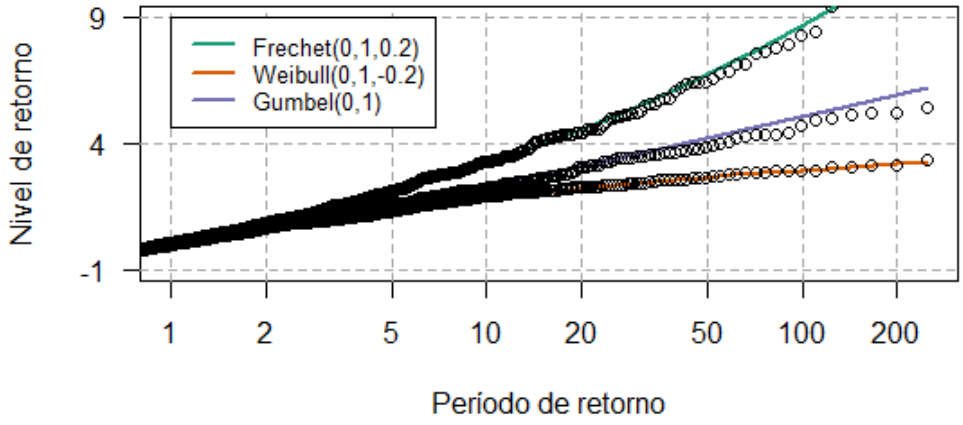


Figura 5.1: Gráfico de niveles de retorno. Fuente: elaboración propia en R.

Del gráfico anterior se puede observar que si $\xi = 0$, entonces la distribución GEV es Gumbel y en ese caso, la curva de los niveles de retorno respecto de los períodos de retorno tiende a ser lineal (observar el caso ejemplo graficado de una $Gumbel(0,1)$). Si $\xi < 0$, entonces el modelo será de la familia Weibull, y en ese caso el gráfico será convexo, mientras que si $\xi > 0$, el modelo será Fréchet y el gráfico de los niveles de retorno estimados se asemejará a una curva cóncava.

Este gráfico como se mencionó anteriormente también forma parte de la salida del diagnóstico de modelos.

5.4 Dependencia de valores extremos

Hasta ahora se ha analizado valores extremos en el contexto de series independientes y estacionarias. Sin embargo, este supuesto puede que no se cumpla y exista dependencia en los datos o la serie sea no estacionaria. A continuación se verá que a pesar que exista dependencia en la serie, si la serie es estacionaria y cumpliendo determinadas condiciones, los máximos de estas series seguirán las mismas leyes vistas anteriormente pero con parámetros diferentes.

En lo que sigue, se hará foco en el tipo de dependencia local es decir que cuando tomamos observaciones de las variables en momentos del tiempo suficientemente alejados entre sí, la serie se comporta como si fuera independiente.

Definición 5.5. Una sucesión estacionaria $\{X_n\}_{n\geq 1}$ satisface la condición $D(u_n)$ si para cualquiera $i_1 < i_2 < ... < i_p < j_1 < j_2 < ... < j_q$ con $j_1 - i_p > l_n$ se cumple que:

$$\left|P\left(X_{i_1} \leq u_n, ..., X_{i_p} \leq u_n, X_{j_1} \leq u_n, ..., X_{j_q} \leq u_n\right) - P\left(X_{i_1} \leq u_n, ..., X_{i_p} \leq u_n\right) P\left(X_{j_1} \leq u_n, ..., X_{j_q} \leq u_n\right)\right| \leq \alpha(n, l_n),$$

donde $\alpha(n, l_n) \to 0$ para alguna sucesión l_n tal que $l_n / n \to 0$ cuando $n \to \infty$.

Teorema 5.6. Sea $\{X_n\}_{n\geq 1}$ una sucesión estacionaria y sea $M_n = máx\{X_1, ..., X_n\}$. Entonces si $\{a_n\}$ con $a_n > 0$ para todo n y $\{b_n\}$ son sucesiones tales que:

$$P\left((M_n - b_n)/a_n \leq z\right) \to G(z), \tag{5.42}$$

donde $G(z)$ es no-degenerada y si se satisface la condición $D(u_n)$ con $u_n = a_n z + b_n$ para todo z real, entonces G está en la familia de distribuciones GVE.

5.4.1 Indice extremal

Como se comentó anteriormente el teorema anterior muestra que para sucesiones que sean asintóticamente independientes en sentido de la condición $D(u_n)$, los máximos de sucesiones estacionarias seguirán las mismas leyes asintóticas GEV como se vio para el caso independiente, pero con una corrección en los parámetros de posición y escala. El siguiente teorema muestra el vínculo existente entre las distribuciones límites para ambos escenarios e introduce el concepto de índice extremal:

Teorema 5.7. Sean $\{X_n\}_{n\geq 1}$ una sucesión estacionaria y $\{X_n^*\}_{n\geq 1}$ una sucesión de v.a. independientes tal que ambas sucesiones tienen la misma distribución marginal. Sean $M_n = máx\{X_1, ..., X_n\}$ y $M_n^* = máx\{X_1^*, ..., X_n^*\}$. Bajo ciertas condiciones de regularidad se tiene que existen $\{a_n\}$ con $a_n > 0$ para todo n y $\{b_n\}$ tales que cuando $n \to \infty$:

$$P\left(\frac{M_n^* - b_n}{a_n} \leq z\right) \to G(z), \tag{5.43}$$

si y solo si $\exists \theta \in [0, 1]$ tal que:

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G^\theta(z), \tag{5.44}$$

donde G es no degenerada. Al valor de θ se lo denomina índice extremal.

Notese que si G es una GVE con parámetros μ, σ y $\xi \neq 0$ entonces:

$$G^\theta(z) = \left(e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}}\right)^\theta = e^{-\theta\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}} = e^{-\left(1+\xi\frac{x-\mu^*}{\sigma^*}\right)^{-1/\xi}}, \tag{5.45}$$

siendo $\mu^* = \mu - \frac{\sigma}{\xi}\left(1 - \theta^{-\xi}\right)$ y $\sigma^* = \sigma\theta^\xi$ con iguales parámetros de forma.

Para el caso Gumbel, la relación entre los parámetros es $\mu^* = \mu + \sigma log\,(\theta)$ y $\sigma^* = \sigma$.

En base a un método de estimación propuesto por [27] se puede interpretar al índice extremal como el recíproco del tamaño promedio de los clusters.

Se observa que para series independientes, el índice extremal es $\theta = 1$. El recíproco, en cambio, no es cierto.

5.5 Valores extremos multivariados

Muchas veces se presenta el problema de analizar y modelar valores extremos de dos o más procesos en forma conjunta. Es decir, el problema ya deja de ser univariado y pasa a ser multivariado. Un ejemplo es el contexto analizado en esta tesis, es decir estudiar el comportamiento extremo de una variable (precipitaciones) observadas en distintas localizaciones geográficas. Otro tipo de estudio de extremos multivariado puede darse si por ejemplo se quieren modelar los extremos de dos variables para una misma ubicación geográfica, etc.

El problema entonces se transforma en un estudio multivariado y se debe analizar la posible dependencia que se genere entre los extremos de cada localización geográfica.

Se realizará el desarrollo del caso bivariado.

Máximo por bloques en el caso bivariado:

Sea una sucesión de n vectores aleatorios $(X_1, Y_1), ..., (X_n, Y_n)$ i.i.d con función distribución conjunta $F(x, y)$.

Los máximos por bloques de cada componente quedan definidos como:

$$\begin{aligned} M_{X,n} &= \max_{1\le i\le n}\{X_i\} \\ M_{Y,n} &= \max_{1\le i\le n}\{Y_i\}. \end{aligned} \tag{5.46}$$

Por ende el vector de máximos bivariado queda definido como $\mathbf{M}_n = (M_{X,n}, M_{Y,n})$.

Al igual que en el caso univariado, se puede obtener una distribución de extremos multivariada normalizando de forma adecuada el vector de máximos.

Definición 5.6. Dadas dos sucesiones $\{\mathbf{a}_n\}_{n\ge1} \subset \Re^2$ donde $\mathbf{a}_n = (a_n^{(1)}, a_n^{(2)})$ y $\mathbf{a}_n > 0$ para todo n y $\{\mathbf{b}_n\}_{n\ge1} \subset \Re^2$ donde $\mathbf{b}_n = (b_n^{(1)}, b_n^{(2)})$ tales que:

$$P\left(\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \le \mathbf{x}\right) = (F(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n))^n \to G(\mathbf{x}) \quad \text{cuando} \quad n \to \infty \quad \text{para todo} \quad \mathbf{x} = (x_1, x_2), \tag{5.47}$$

entonces se dice que G es una distribución de valores extremos multivariada (DVEM).

Notar que G es no degenerada en el sentido que todas sus marginales univariadas son no degeneradas y también son DVE. Análogamente al caso univariado, si dos funciones de distribución multivariadas F y G satisfacen (5.47), para ciertos vectores $\mathbf{a}_n$ y $\mathbf{b}_n$, se dice que F está en el dominio de atracción de G.

Notar que $\dfrac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n}$ denota $\left(\dfrac{M_{X,n} - b_n^{(1)}}{a_n^{(1)}}, \dfrac{M_{Y,n} - b_n^{(2)}}{a_n^{(2)}} \right)$.

A diferencia del caso univariado, no existe una caracterización paramétrica de la distribución límite multivariada, aunque si se transforman las variables de modo que las distribuciones marginales univariadas sean Fréchet unitarias[2] se podrá, sin pérdida de generalidad, separar el comportamiento marginal de la estructura de dependencia.

En el contexto multivariado existe el concepto de función de dependencia de n variables aleatorias. Dicha función es de especial interés ya que puede utilizarse para derivar la distribución límite de cierta distribución de valores extremos multivariada. La misma se define de forma genérica como sigue:

Definición 5.7. Sea $F(\mathbf{x})$ la función distribución de un vector aleatorio $\mathbf{X} = (X_1, X_2, ..., X_d)$ d-dimensional con distribuciones marginales univariadas $F_i(x_i)$, $i = 1, ..., d$. Se define la función de dependencia asociada a $F(\mathbf{x})$ con $\mathbf{x} = (x_1, x_2, ..., x_d)$, denotada por $D_F(y_1, ..., y_d)$ como:

$$D_F(y_1, ..., y_d) = F(F_1^{-1}(y_1), ..., F_d^{-1}(y_d)) = F(\mathbf{x}).$$

El siguiente teorema muestra la relación entre la función de dependencia y la distribución límite:

Teorema 5.8. Sean $\mathbf{X}_1, ..., \mathbf{X}_n$ vectores aleatorios d-dimensionales con distribución común F, entonces, existen vectores de constantes $\mathbf{a}_n, \mathbf{b}_n$ tales que:

$$\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \to G,$$

con G no degenerada, sí y solo sí, cada marginal pertenece al dominio de atracción de alguna $G_k(x)$[3] y si

$$\lim_{n\to\infty} D_{F^n}(y_1^{1/n}, ..., y_d^{1/n}) = D_G(y_1, ..., y_d)$$

Teorema 5.9. Una función de distribución d-dimensional G es una distribución límite de valores extremos sí y solo sí sus marginales univariadas pertenecen al dominio de atracción de $G_k(x)$ y su función de dependencia $D_G(y_1, ..., y_d)$ satisface la ecuación funcional $[D_G(y_1^{1/m}, ..., y_d^{1/m})]^m = D_G(y_1, ..., y_d)$ para todo $m \geq 1$.

La demostración de ambos teoremas anteriores se puede ver en [20].

El siguiente teorema es de importancia para caracterizar las familias de distribuciones límites de extremos multivariados vía procesos puntuales de Poisson como se mostrará en el siguiente teorema (ver demostración en [5], página 301):

[2] Si $X \sim$ Fréchet$(0,1,1)$, entonces se dice que X se distribuye como una Fréchet unitaria con función distribución $H(x,0,1,1) = e^{\left(\frac{1}{x}\right)} I_{x\geq 0}$.

[3] $G_k(x)$ con $k = X, Y$ es la distribución límite de $M_{X,n}$ y $M_{Y,n}$ respectivamente

Teorema 5.10. Sea $\mathbf{X}_1, \mathbf{X}_2, ...$ una sucesión de vectores aleatorios i.i.d d-dimensionales, con entradas no negativas y distribución común F, donde F pertenece al domino de atracción de una distribución multivariada de extremos G. Supongamos que las distribuciones marginales de F son Fréchet unitarias. Consideremos el proceso puntual

$$P_n = \left\{ \frac{\mathbf{X}_i}{n} : i = 1, 2, ..., n \right\},$$

entonces P_n converge en distribución a un proceso de Poisson no homogéneo P en $\Re_+^d - 0$ cuando $n \to \infty$, con medida de intensidad

$$\lambda(dr \times d\mathbf{w}) = m \frac{dr}{r^2} dS(\mathbf{w}),$$

donde r_i y w_{ij} son las coordenadas pseudo-polar y angular y S es una medida de probabilidad en el simplex unitario

$$S_d = \left\{ (w_1, ..., w_d) : \sum_{j=1}^{d} w_j = 1, w_j \geq 0, j = 1, 2, ..., d. \right\},$$

que satisface la condición

$$\int_{S^d} w_j dS(\mathbf{w}) = 1/d, j = 1, 2, ..., d.$$

El siguiente Corolario es una aplicación del teorema anterior y su importancia radica en que caracteriza todas las posibles distribuciones límite multivariadas. Se mostrará la aplicación siguiendo el contexto bivariado:

Corolario 5.10.1. Dados $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ i.i.d tales que X e Y son Fréchet unitarias. Sean sucesiones de normalización $\mathbf{a}_n = \mathbf{n}$ y $\mathbf{b}_n = \mathbf{0}$. Si se consideran $M_{X,n}^* = \frac{M_{X,n}}{n}$ y $M_{Y,n}^* = \frac{M_{Y,n}}{n}$, y si $P(M_{X,n}^* \leq x, M_{Y,n}^* \leq y) \to G(x, y)$ para una G no degenerada, entonces G tiene la siguiente forma:

$$G(x, y) = e^{-V(x,y)} \quad x, y > 0, \tag{5.48}$$

siendo $V(x, y)$ la llamada función de dependencia definida mediante:

$$V(x, y) = 2 \int_0^1 \max\left(\frac{w}{x}, \frac{1-w}{y} \right) dH(w),$$

donde H es una medida en $[0, 1]$ tal que $\int_0^1 w dH(w) = 1/2$.

La familia de distribuciones que surgen como límites en el teorema anterior se conocen como las distribuciones de valores extremos bivariadas.

Notar que $V(x, y)$ es homogénea de grado -1 ya que $V(a^{-1}x, a^{-1}y) = aV(x, y)$ con $a > 0$.

Aplicando dicha propiedad se tiene que:

$$(G(x,y))^n = e^{-nV(x,y)} = G(n^{-1}x, n^{-1}y). \tag{5.49}$$

Entonces G es una función max-estable ya que si la distribución de $\left(M^*_{X,n}, M^*_{Y,n}\right)$ tiende a G, entonces la distribución de $\mathbf{M}_n$ se aproxima a G, salvo por un cambio de escala de orden n^{-1}.

Si por el contrario, las distribuciones marginales no son Fréchet perp pertenecen a cualquiera de las otras dos familias, mediante transformaciones adecuadas vistas anteriormente, se pueden obtener distribuciones del tipo Fréchet y en ese caso aplica lo antes visto.

Es decir, supongamos que $X \sim GEV(\mu_X, \sigma_X, \xi_X)$ y $Y \sim GEV(\mu_Y, \sigma_Y, \xi_Y)$ entonces si aplicamos las siguientes transformaciones:

$$\begin{aligned} \widetilde{X} &= 1 + \xi_X \left(\frac{X - \mu_X}{\sigma_X} \right), \\ \widetilde{Y} &= 1 + \xi_Y \left(\frac{Y - \mu_Y}{\sigma_Y} \right), \end{aligned} \tag{5.50}$$

se tiene que $G(\widetilde{x}, \widetilde{y}) = e^{-V(\widetilde{x},\widetilde{y})} \quad \widetilde{x}, \widetilde{y} > 0$ siendo G una DVE bivariada cuyas marginales son Fréchet unitarias.

En la práctica, se escoge una subfamilia paramétrica de distribuciones DVEM, se estiman los parámetros por máxima verosimilitud por ejemplo, ajustando primero las distribuciones marginales y luego determinando los parámetros de la función de dependencia, o simultáneamente.

Para el caso bivariado, las familias paramétricas de distribuciones más conocidas que modelan la estructura de dependencia, son: distribuciones logísticas, distribuciones bilogísticas, entre otras.

- Modelo Logístico:
$$V(x,y) = \left(\frac{1}{x^{1/\alpha}} + \frac{1}{y^{1/\alpha}} \right)^{\alpha},$$
siendo $x, y > 0$ y $\alpha \in (0,1)$. Dicho modelo es simétrico. Si α es cercano a 1 indica que las variables son aproximadamente independientes, mientras que si α es cercano a 0 las variables presentan fuerte dependencia.

- Modelo Bilogístico:
$$V(x,y) = x\gamma^{1-\alpha} + y(1-\gamma)^{1-\beta},$$
siendo $\alpha, \beta \in (0,1)$ y γ solución de $(1-\alpha)x(1-\gamma)^{\beta} = (1-\beta)y\gamma^{\alpha}$. Los casos de independencia se dan cuando $\alpha = \beta \to 1$ o uno de los dos parámetros está fijo y el otro tiende a 1. Notar que si $\alpha = \beta$ queda el modelo Logístico. $\alpha - \beta$ determina la asimetría en la estructura de dependencia.

5.5.1 Procesos max-estables

En la sección (5.1) se vió el concepto de max-estabilidad asociada a variables aleatorias. En esta sección se extenderá dicho concepto a los proceso estocásticos y se mostrarán dos modelos utilizados propuestos por Smith y Schlather en [49] y [47] respectivamente.

En [12] se definen los procesos max-estables como sigue:

Definición 5.8. Sea K un subconjunto compacto de $\Re^d$ y sea Z un proceso estocástico con marginales no degeneradas con soporte compacto. Diremos que $Z = \{Z(t), t \in T\}$ es un proceso máx-estable si dadas n copias independientes de este proceso $Z_1, ..., Z_n$ existen constantes $a_n > 0$ y b_n tales que:

$$Z =^d \max_{i=1,...,n} \frac{Z_i - b_n}{a_n} \quad \forall n \in N.$$

Si Z es un proceso no degenerado, según la teoría de valores extremos unidimensional vista anteriormente, las distribuciones unidimensionales de Z tienen que ser una DVE cuya ecuación general es la vista en (5.8).

Un caso particular de lo anterior, es cuando el proceso máx-estable es denominado proceso max-estable simple. El mismo tiene la característica de que $P(Z(t) \leq z) = e^{-1/z} \quad z > 0$, es decir las marginales del proceso máx-estable simple son Fréchet unitarias. En este caso puntual si $a_n = n$ y $b_n = 0$ entonces el proceso máx-estable queda definido como sigue:

$$Z =^d \max_{i=1,...,n} \frac{Z_i}{n} \forall n \in N. \tag{5.51}$$

5.5.2 Familias de Procesos Max-Estables

En [49] y [47] se presentan formas de construir procesos max-estables con distribuciones bivariadas.

Modelo Smith

La primer forma de construcción de un proceso max-estable es el descripto en [49], conocido como el modelo Smith, y se presenta a través del siguiente teorema:

Teorema 5.11. Sea $\{(u_i, s_i) : i \in N\}$ un proceso puntual de Poisson definido en $(0, +\infty) \times S$ (se considera generalmente a S como $\Re^d$) con medida de intensidad $\lambda(du, ds) = \frac{du}{u^2} \times \nu(ds)$, con ν una medida de intensidad en $\Re^d$. Sea f una función definida en $\Re^d \times \Re^d$ continua y no negativa tal que:

$$\int_{\Re^d} f(s,t)\nu(ds) = 1 \quad \forall t \in \Re^d, \tag{5.52}$$

entonces el proceso $Z = \max_{i \geq 1} u_i f(s_i, t)$ es un proceso máx-estable definido en $\Re^d$, y todo proceso máx-estable admite una representación de este tipo.

En este modelo, se toma una familia conveniente de funciones $f(s,t)$ como función de s con t fijo, es decir, $\{f(s-t) : s, t \in \Re^d\}$, siendo una densidad normal d-dimensional con vector de medias t y matriz de varianzas y covarianzas Σ:

$$f(s,t) = \frac{1}{(2\pi)^{d/2}} e^{\{-1/2(s-t)'\Sigma(s-t)\}}.$$

De esta forma, dados dos elementos del proceso max-estable según el modelo de Smith, se tiene que la distribución conjunta queda determinada como:

$$P[Z(s) \leq z_1, Z(t) \leq z_2] = e^{\{-1/z_1\phi(a/2+1/alog(z_2/z_1))-1/z_2\phi(a/2+1/alog(z_1/z_2))\}}, \tag{5.53}$$

siendo $a^2 = (s-t)'\Sigma^{-1}(s-t)$ y ϕ la función de distribución normal estándar.

La ecuación anterior, representa una familia de distribuciones bivariada de valores extremos con función de dependencia como sigue:

$$D_F(w) = (1-w)\phi\big(a/2+1/alog((1-w)/w)\big) + w\phi\big(a/2+1/alog(w/(1-w))\big),$$

siendo a el parámetro de dependencia que representa una distancia generalizada entre los puntos s y t.

El proceso visto recién también se conoce como proceso gaussiano de valores extremos dado que se utiliza la distribución gaussiana para su construcción.

Este modelo es un proceso homogéneo en el espacio, es decir, no importa las ubicaciones de los puntos en el espacio sino la distancia entre ellos. Suele utilizarse para modelar lluvias convectivas, término asociado a lluvias que suelen producirse en zonas llanas o con pequeñas irregularidades topográficas, donde puede presentarse un ascenso de aire húmedo y cálido dando origen a nubes del tipo de cumulonimbos con lluvias intensas.

En [49] se propone una interpretación intuitiva respecto de la representación espectral: supongamos que S es una región en la cual ocurre una tormenta, ν representa la distribución de las tormentas sobre S. Cada u_i representa la magnitud de una tormenta y $u_i f(s_i, t)$ representa la cantidad de lluvia en la posición t de una tormenta de tamaño u_i centrada en s_i. La función f representa la forma de la tormenta (según distribución normal bivariada). Es decir, los proceso max-estables pueden pensarse como el máximo en cada ubicación de una cantidad infinita de tormentas.

En la práctica y contexto de esta tesis, para realizar un ajuste del modelo de Smith, basta con estimar la matriz de varianzas y covarianzas Σ ya que las posiciones de las tormentas (índices del proceso) vienen dados por las localizaciones de las estaciones meteorológicas.

Modelo Schlather

Otro método de construcción de un proceso max-estable está dada por [47] y se formaliza según el siguiente teorema demostrado en el mismo artículo. Se conoce como la caracterización de Schlather:

Teorema 5.12. Sea Y un proceso estacionario en $\Re^d$, sea $\mu = E[max\{0, Y(o)\}]$ donde o denota el origen y sea Π un proceso de Poisson en $(0,\infty)$ con medida de intensidad $d\Lambda(s) = \mu^{-1}s^{-1}ds$, entonces si $\{Y_s\}$ con $s \in (0,\infty)$ son copias i.i.d. de Y, entonces

$$Z(x) = \max_{s\in\Pi}\{sY_s(x)\} = \max_{s\in\Pi} s\max\{0, Y_s(x)\},$$

es un proceso max-estable estacionario con distribuciones marginales Fréchet unitarias.

Si se considera a Y un campo aleatorio estándar normal con función de correlación $\rho(t)$ y Π un proceso de Poisson en $(0,\infty)$ con medida de intensidad $d\Lambda(s) = \sqrt{2\pi}s^{-2}ds$, entonces aplicando el teorema anterior, se tiene:

$$-log\left(P[Z(o) \le s, Z(x) \le t]\right) = 1/2(1/t + 1/s)\left(1 + \sqrt{1 - 2(\rho(x)+1)st/(s+t)^2}\right). \tag{5.54}$$

Existen diversas alternativas de funciones de correlación según [9] como ser Exponencial, Power Exponencial, Esférica, Gaussiana, Mátern, y otras. A diferencia del modelo anterior, éste posee una forma de tormenta aleatoria, cuya estructura está dada por la función de correlación.

En la práctica, un ajuste del modelo de Schlather como modelo de un proceso max-estable, se reduce a la estimación de los parámetros asociados a $\rho(t)$.

Este proceso también se conoce como proceso extremal gaussiano y su interpretación según [47] es como sigue: sea sY_s la lluvia diaria, todos con la misma estructura de dependencia con la única diferencia en la magnitud s. Este tipo de modelos es aplicado a lluvias ciclónicas, es decir lluvias que ocurren en puntos aislados dispersos en la región de estudio.

De forma genérica, la distribución acumulada finito dimensional de un proceso máx-estable se puede determinar como sigue:

$$\begin{aligned}
F(z_1,...,z_k) &= P\left(\max_i\{\Psi Y_i(x_j)\} \le z_j, j = 1,...,k\right)\\
&= P\left(\Psi \le z_j / Y_i(x_j), i \ge 1, j = 1,...,k\right)\\
&= exp\left\{-\int_{\Re^d}\int_{\Re} I_{(\Psi > \min_j z_j/x_j)}\Psi^{-2}d\Psi dP^Y(y_1,...,y_k)\right\}\\
&= exp\left\{-\int_{\Re^d}\max_j y_j/z_j dP^Y(y_1,...,y_k)\right\}\\
&= exp\left\{-E\left(\max_j Y(x_j)/z_j\right)\right\}\\
&= exp\{-V(z_1,...,z_k)\} = exp\{-V(\mathbf{z})\},
\end{aligned} \tag{5.55}$$

siendo dP^Y la medida de probabilidad asociada al proceso Y y $E\left(\max_j Y(x_j)/z_j\right) = V(z_1,...,z_k)$.

5.5.3 Dependencia en el contexto multivariado

Coeficiente extremal:

A partir de la representación espectral de Schlather y su distribución finito dimensional (5.55) se vio que dados $\mathbf{x} = (x_1, ..., x_k)$ y $\mathbf{z} = (z_1, ..., z_k)$ se tiene que:

$$P(Z(\mathbf{x}) \leq \mathbf{z}) = e^{-V_{\mathbf{X}}(\mathbf{z})}. \tag{5.56}$$

Si todos los componentes de $\mathbf{z}$ son iguales entonces:

$$\begin{aligned} V_{\mathbf{X}}(\mathbf{z}) &= E\left(\max_{j=1,...,k} Y(x_j)/\mathbf{z}\right) \\ &= \frac{\theta(\mathbf{x})}{z}, \end{aligned} \tag{5.57}$$

siendo $\theta(\mathbf{x}) = E\left(\max_{j=1,...,k} Y(x_j)\right)$ el coeficiente extremal con $\theta(\mathbf{x}) \in \Re$ y representa una medida de dependencia entre los elementos del vector $Z(\mathbf{x})$.

En el caso bivariado, se conoce como la función de coeficiente extremal:

$$\theta(h) = E\left(\max\{Y(x), Y(x+h)\}\right), \tag{5.58}$$

con $\theta(h) \in [1,2]$. Si $\theta(h) = 1$ entonces los elementos $Y(x)$ y $Y(x+h)$ son completamente dependientes mientras que si $\theta(h) = 2$ se está frente al caso de independencia.
Demostramos a continuación esta afirmación:

$$P(Z(x+h) \leq z | Z(x) \leq z) = \frac{P(Z(x+h) \leq z, Z(x) \leq z)}{P(Z(x) \leq z)} = \frac{e^{-\frac{\theta(h)}{z}}}{P(Z(x) \leq z)} = P(Z(x+h) \leq z)^{\theta(h)-1}. \tag{5.59}$$

De aquí se deduce que si $\theta(h) = 1$ entonces los elementos $Y(x)$ y $Y(x+h)$ son completamente dependientes mientras que si $\theta(h) = 2$ se está frente al caso de independencia.

Al igual que para la función correlación definida previamente, para la función de coeficiente extremal existen diversas familias paramétricas:

- Modelo Smith: $\theta(h) = 2\phi\left(\frac{\sqrt{h^T \Sigma^{-1} h}}{2}\right)$.
- Modelo Schlather: $\theta(h) = 1 + \sqrt{\frac{1-\rho(h)}{2}}$.
- Modelo Brown-Resnick: $\theta(h) = 2\phi\left(\frac{\sqrt{V(Y(h))}}{2}\right)$.
- Modelo Geométrico-Gaussiano: $\theta(h) = 2\phi\left(\frac{\sqrt{\sigma^2(1-\rho(h))}}{2}\right)$.

- Modelo t-Extremal: $\theta(h) = 2T_{\nu+1}\left(\sqrt{\frac{\nu+1}{1-\rho^2(h)}}(1-\rho(h)) - \sqrt{\frac{1-\rho^2(h)}{\nu+1}}\rho(h)\right)$.

siendo ϕ la función de distribución normal estándar.

Variograma:

En fenómenos espaciales como ser variables meteorológicas, variables relacionadas a la ecología, económicas, entre otras, puede que exista cierto grado de dependencia espacial. El variograma es una herramienta que permite estudiar la dependencia espacial de una variable. Diversas bibliografías se pueden encontrar donde se realiza una explicación exhaustiva sobre geoestadística como ser [9], [21] o [19] entre otros.

Un proceso espacial se define como (ver [9]) $Z : Dx\Omega \to \Re$ medible, tal que:

$$\{Z(s) : s \in D\}$$

Donde:

- D es un subconjunto de $\Re^d$
- $s \in D$ es una posición genérica localizada en el espacio d-dimensional.
- $Z(s)$ una variable aleatoria localizada en s.

Definición 5.9. El variograma asociado al proceso $\{Z(s) : s \in D\}$ se define como:

$$2\gamma(h,s) = Var\left(Z(s+h) - Z(s)\right) \qquad \forall s \in D, \forall h, s+h \in D.$$

Madograma:

Si además, el proceso es intrínsecamente estacionario[4]:

$$2\gamma(h) = E\left(Z(s+h) - Z(s)\right)^2 \qquad s \in D, s+h \in D.$$

A la expresión anterior se la puede interpretar como el valor promedio de la diferencia al cuadrado de los valores que toma la variable de interés en dos puntos separados por una distancia h.

Para el contexto de valores extremos, puede que los momentos de segundo orden no existan y por ende no sería posible la obtención del variograma. En [38] y [8] se define una herramienta similar al variograma utilizando la distancia L_1 construyendo así el madograma:

$$\nu(h) = \frac{1}{2}E|Z(x+h) - Z(x)|. \tag{5.60}$$

[4]Se dice que $\{Z(s) : s \in D\}$ es intrínsecamente estacionario si satisface: 1) el valor esperado de las diferencias cumple que $E(Z(s+h)-Z(s)) = 0$ y 2) la varianza de las diferencias es finita y está dada por $Var(Z(s+h)-Z(s)) = 2\gamma(h)$, siendo la función $\gamma(.)$ el variograma.

F-Madograma:

Una adaptación a dicho concepto fue propuesta en [7] y [41] en el contexto de valores extremos, el *F-Madograma*:

Definición 5.10.

$$\nu^F(h) = \frac{1}{2}E|F(Z(x+h)) - F(Z(x))|, \tag{5.61}$$

siendo F la función de distribución de $Z \forall x$.

Una ventaja de la expresión anterior es que la esperanza calculada en el F-Madograma siempre existe por tratarse de variables acotadas.

En [7] se demuestra que el F-Madograma puede expresarse en términos del coeficiente extremal según:

$$\nu^F(h) = \frac{1}{2}\frac{\theta(h)-1}{\theta(h)+1}, \tag{5.62}$$

siendo $\theta(h)$ el coeficiente extremal que se definió anteriormente.

Cabe destacar que el F-Madograma es un concepto adimensional y solo describe la fuerza de la dependencia, ya que $F(Z(x))$ es una variable uniforme en $[0,1]$ y no depende de las leyes de las distribuciones marginales en cuestión. Es decir, no brinda información sobre el volumen de precipitaciones en un punto específico.

6 Análisis de clusters

6.1 Introducción

El análisis de clusters es un método exploratorio del tipo no supervisado. Dicha metodología es capaz de agrupar observaciones o variables en grupos lo más homogéneos que sea posible. Aquí se hará énfasis en clustering de observaciones. Es decir se dispone de una matriz de datos $X = (X_1, ..., X_p)$ con p variables de interés e I observaciones (individuos) de la que no se dispone de etiquetas de clase que identifiquen a las observaciones. El objetivo principal es agrupar las observaciones de manera que aquellas que pertenezcan a un mismo grupo sean más similares entre si que observaciones fuera de ese grupo. Es decir que el análisis de cluster tiene por objetivo conformar grupos de acuerdo a ciertas características que pueden ser de interés, de manera de que dentro de cada grupo los elementos sean lo más homogéneos posible. La homogeneidad (y heterogeneidad) estará medida a través de una distancia pre-definida.

La clasificación obtenida dependerá de las variables seleccionadas, pudiendo variar la estructura de grupos en el caso de que se consideren otras variables diferentes.

Los métodos de clasificación pueden ser jerárquicos o no jerárquicos. En los métodos jerárquicos, la agrupación se realiza mediante un proceso de agrupación (o desagrupación, según sean

métodos agregativos o divisivos) sucesiva, cuyo resultado final es una jerarquía de unión completa en la que cada grupo se une (o separa) en una determinada fase. En cambio, los métodos no jerárquicos o de partición, permiten la reasignación de elementos, es decir que las observaciones pueden variar de grupo en cada iteración. Los métodos no jerárquicos necesitan seleccionar la cantidad de grupos a agrupar a priori.

Ambos análisis no son competitivos, sino que se complementan, ya que el análisis de grupos jerárquico puede pensarse como una primer exploración de la estructura de los datos, mientras que el análisis de grupos no jerárquico se verá como una afinación de dichos resultados.

Dentro de los métodos jerárquicos se pueden mencionar el método del vecino más cercano (*single linkage*), método del vecino más lejano (*complete linkage*), método de mínima varianza (*Ward*), entre otros.

Dentro de los métodos no jerárquicos, se pueden encontrar diferentes técnicas como ser *k-means, PAM, Fuzzy*, entre otras.

Las medidas de disimilaridad entre los individuos $x_i = (x_{i1}, ..., x_{ip})$ y $x_j = (x_{j1}, ..., x_{jp}) \in \Re^p$ que se utilizan con mayor frecuencia son:

- Distancia Euclídea (L_2): $d(x_i, x_j) = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right)^{1/2}$.
- Distancia Manhattan (L_1): $d(x_i, x_j) = \sum_{k=1}^{p} \mid x_{ik} - x_{jk} \mid$.
- Distancia de Mahalanobis: $d(x_i, x_j) = (x_i - x_j)' \Sigma^{-1} (x_i - x_j)$ siendo Σ una matriz de varianzas y covarianzas no singular.

Las distancias anteriores suelen utilizarse en el caso de trabajar con variables cuantitativas. La distancia euclídea es la más utilizada, aunque la distancia Manhattan es más robusta frente a outliers. Cuando existe alta correlación entre las variables consideradas la distancia Mahalanobis es la más adecuada ya que considera las correlaciones de las variables en su fórmula. Dado que las medidas de disimilaridad no son invariantes frente a cambios de escala de las variables, es importante estandarizar los datos previamente.

El análisis de clusters no jerárquico, como se mencionó anteriormente, busca particionar el conjunto de datos en un número especificado de grupos K minimizando cierta función objetivo.

Cada algoritmo se diferencia según la función objetivo a utilizar. La función objetivo más común es la suma de las distancias de las observaciones de un grupo respecto de su centroide ($c(.)$), es decir, para cada partición $\zeta = C_1, ..., C_K$ encontrada se busca minimizar:

$$W(\zeta_K) = \sum_{h=1}^{K} \sum_{x_i \in C_h} d(x_i, c(h)), \tag{6.1}$$

siendo K el número de grupos.

Notar que si se utiliza la distancia Euclídea y el centroide es la media de las observaciones del grupo, la función objetivo corresponde a la suma de cuadrados intra cluster W que se detallará más adelante.

Para los métodos jerárquicos existe por un lado un gráfico que da cuenta de la estructura de particiones generada por dichos algoritmos denominado Dendrograma que es útil para definir la cantidad de grupos óptimos. Además existen ciertos indicadores que también resultan de utilidad para definir la cantidad de grupos. A continuación se describen brevemente cada uno de ellos:

- Dendrograma: gráfico en el que se puede visualizar la evolución de las agrupaciones, es decir se puede observar que grupos se van uniendo entre sí y los niveles al que lo hacen. Dicho gráfico se construye mediante la transformación de las distancias originales a distancias ultramétricas.[5]

- R^2: Establece la relación entre la variación explicada por la estructura de grupos determinada (VE) y la variación total (VT):

  $$R^2 = VE/VT = 1 - VNE/VT,$$

  siendo $VNE = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{J}\left(x_{ij(k)} - \overline{x}_{kj}\right)^2$ y $VT = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \overline{x}_j)^2$. Notar que $R^2 = 1$ cuando la cantidad de grupos es igual a la cantidad de observaciones y $R^2 = 0$ cuando la cantidad de grupos es igual a 1 (la nube original). Por lo que a medida que la cantidad de grupos crece, este indicador aumentará. Cuando dicho aumento deje de ser significativo sugiere la cantidad de grupos óptima.

- Regla de Calinski ($Pseudo_F$):

  $$Pseudo_F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}.$$

  Si dicho indicador crece monótonamente conforme crece la cantidad de grupos no se podría establecer una estructura clara de grupos. Análogamente, si dicho indicador disminuye conforme crece la cantidad de grupos (aunque en este caso podría existir una estructura jerárquica). Si dicho indicador crece, y luego disminuye entonces la cantidad de grupos óptima está dada en el valor de k que produce el cambio.

Para los métodos no jerárquicos una herramienta que se utiliza es el gráfico *silhouette* que fue propuesto en ([45]) y cuya idea mostramos a continuación:

Sean $\zeta = C_1, ..., C_K$ las particiones de los datos en K grupos, el valor *silhouette* de la $i-$ésima observación que pertenece al grupo C_k se define como sigue:

$$s_{ik} = \frac{b_i - a_i}{max\{a_i, b_i\}} \quad si \quad |C_k| > 1 \quad \forall k \in \{1, 2, ..., K\}, \tag{6.2}$$

5

Definición 6.1. Si d_{ij} es una distancia y además satisface la siguiente propiedad: $d_{ij} = \max\left(d_{ik}, d_{jk}\right) \quad \forall i, \forall j, \forall k$ entonces d_{ij} es una distancia ultramétrica.

y

$$s_{ik} = 0 \quad si \quad |C_k| = 1 \quad \forall k \in \{1, 2, ..., K\}, \tag{6.3}$$

siendo:

- $|C_k|$ la cantidad de observaciones en el grupo k para $k \in \{1, 2, ..., K\}$,
- a_i es la distancia promedio entre la observación $i-$ésima y el resto de las observaciones de su mismo grupo. Es decir, para cada observación $i \in C_k$ con $k \in \{1, 2, ..., K\}$:

  $$a_i = \frac{1}{|C_k| - 1} \sum_{j \in C_k, i \neq j} d(x_i, x_j).$$

  a_i puede interpretarse como una medida de cuan bien ha sido clasificada la observación $i-$ésima en el cluster al que pertenece, es decir, cuanto más pequeño el valor de a_i mejor clasificada está la observación $i-$ésima.

- b_i es la mínima distancia promedio entre la observación $i-$ésima y las demás observaciones pertenecientes a los demás grupos. Es decir, que para cada observación $i \in C_k$ con $k \in \{1, 2, ..., K\}$:

  $$b_i = mín_{i \neq j, k \neq h} \frac{1}{|C_j|} \sum_{j \in C_h} d(x_i, x_j).$$

Notar que $-1 \leq s_{ik} \leq 1$. Valores de s_{ik} cercanos a 1 se dan si $a_i \approx 0$ y en ese caso la observación fue bien clasificada al grupo al que pertenece. Mientras que valores negativos de s_{ik} indican que la observación fue mal clasificada ($b_i \approx 0$) al grupo al que pertenece. Si $s_{ik} \approx 0$ quiere decir que $a_i \approx b_i$ y en ese caso la observación podría tener un grado de pertenencia a más de un grupo ([28]).

El gráfico *silhouette plot* es un gráfico de barras horizontales, es decir que en el eje de las abscisas se visualiza el valor de s_{ik} para todas las observaciones de la base de datos, ordenadas según el grupo al que pertenecen.

También se puede visualizar en dicho gráfico la cantidad de observaciones de cada grupo, y dos medidas de resumen, por un lado el valor *silhouette* promedio de cada grupo y por otro el estadístico $\overline{s}_K$ (promedio de los s_{ik} de todas las observaciones de la base de datos). Este último, es un buen indicador de la *performance* del algoritmo de clustering. Incluso suele utilizarse dicho estadístico para determinar el número óptimo de clusters para iniciar el algoritmo, determinando el valor K como aquel que maximice $\overline{s}_K$. En ([33]) se define el coeficiente *silhouette* como $SC = max_K\{\overline{s}_K\}$ y su interpretación es como sigue:

- Si $SC \leq 0.25$ no hay una estructura de grupos definida en los datos o si la hay el algoritmo no logró captarla.
- Si $0.26 \leq SC \leq 0.50$ la estructura encontrada es débil e incluso artificial.
- Si $0.51 \leq SC \leq 0.70$ la estructura encontrada es razonable.

- Si $0.71 \leq SC \leq 1$ la estructura encontrada es fuerte.

Las etapas de implementación de este tipo de metodologías se pueden resumir como sigue: 1) Estandarización de la matriz de datos originales X de I individuos; 2) definición de la medida de disimilaridad entre individuos; 3) definición del algoritmo de agrupación; 4) consideración de reglas de detención; 5) selección e interpretación de los grupos en función de las p variables de interés.

En esta tesis se desarrollará en una primer instancia un análisis de cluster jerárquico Ward a los datos utilizando distintas distancias, y luego utilizando el número de grupos obtenido se iniciará el algoritmo PAM y se comparará con el número de grupos óptimo que se encuentre observando el estadístico SC.

6.2 Análisis de clusters jerárquico: Ward

El método Ward busca minimizar como se mencionó anteriormente la siguiente función objetivo:

$$W(\zeta_K) = \sum_{h=1}^{K} \sum_{x_i \in C_h} d(x_i, c(h)). \tag{6.4}$$

Como ya se mencionó también, si la distancia utilizada es la euclídea, en cada etapa, dicho algoritmo busca minimizar la dispersión o varianza intra grupos en la nueva partición. Es decir, si la distancia a utilizar es la distancia euclídea dicho algoritmo busca minimizar W que es la suma de los cuadrados residuales. Recordar que la suma de cuadrados totales SCT se puede descomponer como sigue:

$$\begin{aligned} \sum_{i=1}^{I} \sum_{j=1}^{p} \left(x_{ij} - \overline{x}_j\right)^2 &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left(x_{ij(k)} - \overline{x}_{kj}\right)^2 + \sum_{k=1}^{K} \sum_{j=1}^{p} n_k \left(\overline{x}_{kj} - \overline{x}_j\right)^2 = \\ &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i, G_k) + \sum_{k=1}^{K} n_k d^2(G_k, G), \end{aligned} \tag{6.5}$$

siendo:

- n_k el tamaño del cluster k,
- G_k el baricentro o centro de gravedad del cluster k,
- G el baricentro o centro de gravedad general de los datos.

La variación intra grupo a medida que se van produciendo las agregaciones va aumentando, de manera que dicho algoritmo busca minimizar el incremento de la variación intra grupos resultante de la unión de dos elementos A y B en cada etapa del proceso dada por

$$(A \cup B) = \frac{n_A n_B}{n_A + n_B} d^2(G_A, G_B).$$

En cada etapa del algoritmo se calcula dicho incremento y se construye la tabla de disimilaridades con la cual se van uniendo grupos con menor incremento de la variación intra grupo.

Este algoritmo tiende a juntar grupos con pequeño número de observaciones y tiende a formar grupos con igual cantidad de observaciones. Dicho método tiene como ventaja que es menos sensible a la presencia de valores atípicos.

A continuación se describen los pasos de este algoritmo aglomerativo:

---

1) Input: $X = \{x_i, i = 1, ..., I\}$ estandarizada,I=número de grupos iniciales (igual a la cantidad de observaciones).

2) Calcular $D^{(1)} = D_{I\times I} = d^2_{ij} = d^2(x_i, x_j) \quad i, j = 1, ..., I$ matriz de disimilaridades entre los I grupos iniciales.

3) Unir aquellos grupos I y J ($\{I, J\}$) tales que d^2_{IJ} es la distancia mínima de $D^{(1)}$.

4) Calcular $d^2_{IJ,K}$ entre el grupo formado en el paso anterior $\{I, J\}$ y todos los demás grupos $K \neq \{I, J\}$.

5) Calcular $D^{(2)} = D_{(I-1)\times(I-1)}$ reemplazando la fila y columna I y J por la fila y columna IJ con las disimilaridades calculadas en el paso anterior. Calcular $(IJ \cup K) = \frac{n_{IJ} n_K}{n_{IJ} + n_K} d^2(G_{IJ}, G_K)$ para todo K y unir los grupos $\{I, J\}$ y K con mínimo $(IJ \cup K)$.

6) Repetir los pasos 4 y 5 $I - 1$ veces.

7) Output: indicadores $Pseudo_F$ ó R^2 para determinar la cantidad de grupos óptimo, lista con la evolución de los grupos que se fueron uniendo en cada paso, dendrograma, descripción de los grupos.

Notar que para el $i-$ésimo paso, $D^{(i)}$ será una matriz simétrica de tamaño $(Ii + 1) \times (Ii + 1)$ con $i = 1, ..., I$. En el último paso $i = I$ se tiene que $D^{(I)} = 0$ ya que todas las observaciones se encuentran agrupadas en un único grupo.

6.3 Análisis de clusters no jerárquico: *K-means* y PAM

Dentro de los métodos no jerárquicos o de partición se encuentran los métodos *K-means* y PAM (*Partition Around Medoids*), entre otros. Dichos métodos se diferencian de los métodos jerárquicos en que se debe elegir la cantidad de grupos previamente. Estos métodos iterativos, clasifican en cada etapa a cada observación dentro de alguno de los K grupos determinados a priori. Es decir en cada paso van formando estructuras de grupos donde no hay una jerarquización de los mismos. Dado K la cantidad de grupos definida previamente, los algoritmos no jerárquicos buscan particionar los datos en K grupos de manera que los elementos dentro de cada grupo sean lo más similares posible y los elementos de diferentes grupos sean lo más disímiles posible.

6.3.1 *K-means*

Dentro de los métodos de cluster no jerárquicos, el más utilizado es el *K-means* [37]. El algoritmo de dicho método se puede resumir como sigue:

---

1) Input: $X = \{x_i, i = 1, ..., I\}$. K número de grupos.

2) Elegir alguna de estas inicilizaciones:

- Asignar aleatoriamente las observaciones a K clusters y calcular $\overline{x}_k$ el centroide de cada grupo con $k = 1, ..., K$.
- Pre-definir los K centroides $\overline{x}_k$ a partir de alguna agrupación realizada por otro método de clustering por ejemplo.

3) Calcular la distancia entre la observación i y el centroide del cluster al cual pertenece: $W(i) = \sum_{k=1}^{K} \sum_{c(i)=k} d^2(x_i, \overline{x}_k)$, siendo $c(i)$ el cluster que contiene a x_i.

4) Re-asignar cada observación al cluster más cercano de forma de miniminzar $W(i)$.

5) Re-calcular los centroides luego de haber realizado el paso 4.

6) Iterar pasos 3,4 y 5 hasta que no sea necesaria ninguna re-asignación.

7) Output: Estructura de K grupos óptima.

6.3.2 PAM

El algoritmo PAM podría verse como una variación o mejora del *K-means* [54]. Al igual que el algoritmo *K-means*, necesita de una partición inicial dada. Una diferencia respecto del algoritmo anterior, es que PAM se basa en la búsqueda de K "individuos representativos" (denominados medioides en lugar de centroides) entre el conjunto de observaciones, de manera que representen adecuadamente la estructura de los datos en cada partición. El medioide de un cluster se define como la observación de dicho grupo cuya disimilaridad a las demás observaciones de su grupo es mínima.

En general PAM es más robusto que *K-means* y requiere como argumento de entrada solamente la matriz de disimiliaridades entre observaciones y no los datos originales. Como desventaja, dicho método es menos eficiente computacionalmente, debido principalmente a la búsqueda de los medioides ([28]).

La idea principal del algoritmo PAM es encontrar los medioides óptimos que serán en general observaciones centralmente localizadas dentro de los clusters. El algoritmo comienza con una asignación a priori del conjunto de medioides denominado conjunto M. El complemento de este conjunto $\bar{M}$ contiene las demás observaciones que no han sido seleccionadas

como medioides, entonces la base de datos completa puede escribirse como $U = M \cup \bar{M}$.

El objetivo de PAM es minimizar la distancia promedio de las observaciones al medioide del grupo al que fue clasificado. Esto equivale a minimizar la suma de las distancias entre las observaciones que pertenecen al mismo grupo. El algoritmo PAM se implementa en dos fases:

1. La primer fase consiste en la selección de K observaciones que formaran el conjunto M.

2. La segunda fase consiste en intercambiar observaciones de M para mejorar la *performance* del algoritmo.

La estrategia de intercambiar el medioide fue introducida en [34].

La implementación del algoritmo se mostrará a continuación pero antes se presenta las dos medidas que serán clave en su desarrollo:

- D_i = distancia entre la observación i y el medioide más cercano $\forall i \in U$.
- E_i = distancia entre i y el segundo medioide más cercano $\forall i \in U$.

Dichos valores se actualizarán en cada iteración cuando los conjuntos M y $\bar{M}$ cambien.

Se describirán los pasos a seguir en cada fase. La primer fase se lleva a cabo de la siguiente manera:

---

1) Input: $D = (d_{ij})$ matriz de disimilaridad. K número de grupos. Inicializar M por ejemplo con aquellas observaciones tal que la suma de las distancias a las demás observaciones es mínima.

2) Considerar una observación $i \in \bar{M}$ como un candidato a incluir en el conjunto M.

3) Dada $j \in \bar{M} - \{i\}$ calcular D_j.

4) Si $D_j > d(i, j)$ entonces la observación j contribuirá a la decisión de seleccionar i como medioide. Sea $W_{ij} = máx\{D_j - d(i, j), 0\}$.

5) Calcular la ganancia total de incluir a i en el conjunto de medioides M como $g_i = \sum_{j \in U} W_{ij}$.

6) Elegir la observación i que maximice g_i como nuevo medioide y por tanto se incluirá en el conjunto M.

Se repiten los pasos 2 a 6, hasta seleccionar K observaciones.

Para la segunda fase se considerará los pares $(i, h) \in MxM$ y se calculará el efecto X_{ih} en la suma de las distancias entre las observaciones y el medioide más cercano causado por intercambiar i por h. La segunda fase se describe a continuación:

1) Dada $j \in \bar{M}$, se calcula Y_{jih} según el caso que corresponda como sigue:

- Si $d(i,j) > D_j$ se calcula $Y_{jih} = mín\{d(j,h) - D_j, 0\}$
- Si $d(i,j) = D_j$ se calcula $Y_{jih} = mín\{d(j,h), E_j\} - D_j$

2) Calcular $X_{ih} = \sum_{j \in \bar{M}} Y_{jih}$.

3) Seleccionar el par $(i,h) \in Mx\bar{M}$ que minimiza X_{ih}.

4) Si $X_{ih} < 0$ entonces se intercambia la observación i por la observación h. Se actualiza D_i y E_i para todas las observaciones y se vuelve a correr el paso 1.

Si el mínimo $X_{ih} > 0$ el valor de la función objetivo no puede decrecer más el algoritmo se detiene. Es decir que la regla de detención es que todos los valores X_{ih} sean positivos.

Como desventaja de PAM y *K-means* respecto de los métodos jerárquicos, se pueden nombrar: 1) la elección del número de clusters a priori y 2) son sensibles a la inicialización. La cantidad de clusters puede venir sugerida a partir de una partición jerárquica anterior visible a partir de un dendrograma. Respecto de la inicialización, dado que el algoritmo puede converger a mínimos locales, se propone como solución correr el algoritmo con diferentes particiones iniciales y elegir el resultado con la menor función objetivo.

6.4 Análisis de clusters PAM aplicado a valores extremos

Dados los conceptos de teoría de valores extremos desarrollada en la sección 5 y la teoría desarrollada sobre análisis de clusters, ambas metodologías serán conjugadas para obtener una aplicación de clustering para el caso de valores extremos.

El interés principal de trabajar con un algoritmo de clustering como PAM bajo el contexto de valores extremos radica en que otros algoritmos, como ser el *K-means*, construye los centros de los grupos en cada paso, promediando las observaciones dentro de cada grupo. En el contexto de valores extremos, el promedio de valores extremos deja de ser un valor extremo y ello violaría propiedades de max-estabilidad.

El algoritmo de clustering a aplicar a los datos de precipitaciones extremas en Uruguay, será siguiendo la línea del trabajo aplicado en [3].

Sean M_i y M_j los valores extremos observados en dos localizaciones i y j, donde los datos fueron transformados de modo que ambas marginales sean Fréchet unitarias, se puede escribir la distribución bivariada según se propone en [3] como sigue:

$$P(M_i \leq u, M_j \leq v) = e^{-V\left(\frac{-1}{log(F_i(u))}, \frac{-1}{log(F_j(v))}\right)}, \tag{6.6}$$

siendo $F_i(x) = P(M_i \leq x)$ la distribución marginal de M_i, y recordando que $V(.,.)$ es la función estructura de dependencia definida de manera que cumpla con las condiciones vistas en el Teorema 5.10.1.

Notar que si $u = v$ entonces $V(u,u) = \dfrac{V(1,1)}{u^2}$.

Dada la función de estructura de dependencia definida anteriormente se tiene que:

$$P(M_i \leq u, M_j \leq u) = \left[P(M_i \leq u)P(M_j \leq u)\right]^{\frac{-V(1,1)}{2}}, \tag{6.7}$$

siendo $V(1,1)$ el coeficiente extremal.

En [7] se demuestra que el F-madograma definido en la ecuación (5.61) se puede expresar como se vio anteriormente:

$$\nu^F = \frac{1}{2}\frac{V(1,1)-1}{V(1,1)+1}. \tag{6.8}$$

Además el F-madograma puede estimarse de la siguiente manera: dada una muestra de máximos $(M_i^{(t)}, M_j^{(t)})$ obtenidas en dos estaciones i y j en el momento t para $t = 1,2,...,T$, entonces se puede obtener la siguiente estimación en cuestión:

$$\hat{\nu}^F = \frac{1}{2T}\sum_{t=1}^{T} |\hat{F}_i(M_i^{(t)}) - \hat{F}_j(M_j^{(t)})|, \tag{6.9}$$

siendo $\hat{F}_i$ y $\hat{F}_j$ las funciones de distribución empíricas de las estaciones i y j respectivamente, definidas como:

$$\hat{F}_i(u) = \frac{1}{T}\sum_{t=1}^{T} I_{\{M_i^{(t)} \leq u\}}, \tag{6.10}$$

$$\hat{F}_j(u) = \frac{1}{T}\sum_{t=1}^{T} I_{\{M_j^{(t)} \leq u\}}. \tag{6.11}$$

Una manera sencilla de obtener una estimación del coeficiente extremal se obtiene reemplazando en la ecuación (6.8) la estimación $\hat{\nu}^F$ de (6.9) y despejando $V(1,1)$.

También es interesante notar que para la construcción del F-madograma no se requiere de ajustar ninguna distribución GEV específica a los datos, ya que el único insumo necesario es $\hat{F}_i$ y $\hat{F}_j$ y por ende hay un ahorro computacional significativo.

Dadas las condiciones anteriores, se combinará el concepto de la función F-madograma con el análisis de clusters PAM, en el sentido que la función F-madograma anteriormente descripta será la función distancia a utilizar en el algoritmo de clustering PAM.

6.5 Test de independencia basado en ratios de recurrencia

Como complemento al análisis de clustering descripto en la sección anterior, resulta interesante conocer si las precipitaciones extremas anuales diarias para distintas estaciones son independientes o no. Para explorar este aspecto, se aplicó un test de independencia que fue propuesto en [32].

En las siguientes líneas se describe dicho test:

Dada una muestra bivariada $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ de un par (X, Y) donde $X \in S_X$, $Y \in S_Y$ siendo S_X y S_Y espacios métricos cualesquiera.

Se plantea el test $H_0 : X$ e Y son independientes contra la alternativa $H_1 : X$ e Y no son independientes.

El test está basado en un estadístico que es un funcional que mide la diferencia entre el porcentaje conjunto de recurrencias entre X e Y y el producto de los porcentajes marginales de recurrencias de X y de Y. Los mismos se definen mediante:

$$\begin{aligned} RR_n^{X,Y}(r,s) &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}_{\{d(X_i,X_j)<r,\, d(Y_i,Y_j)<s\}}, \\ RR_n^{X}(r) &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}_{\{d(X_i,X_j)<r\}}, \\ RR_n^{Y}(s) &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}_{\{d(Y_i,Y_j)<s\}}, \end{aligned}$$

donde se utiliza la misma letra d para medir la distancia tanto en el espacio métrico X como en el de Y (sin riesgo a confusión).

Se observa que $RR_n^X(r)$ representa el porcentaje pares de elementos de la muestra $\{X_1, ..., X_n\}$ cuyas distancias entre sí son menores que r. Análogamente se define $RR_n^Y(s)$ mientras que $RR_n^{X,Y}(r,s)$ representa el porcentaje conjunto.

Si X e Y son independientes y n es suficientemente grande, es de esperar que $RR_n^{X,Y}(r,s) \cong RR_n^X(r) RR_n^Y(s)$ para todos $r, s > 0$.

La región crítica se define mediante $\{T_n \geq c\}$ donde el estadístico de prueba se define como:

$$T_n = n \int_0^{+\infty} \int_0^{+\infty} \left(RR_n^{X,Y}(r,s) - RR_n^X(r) RR_n^Y(s) \right)^2 g_X(r) g_Y(s)\, dr ds,$$

donde las funciones de pesos se definen mediante $g_X(r) = \varphi\left(\frac{r-\mu_X}{\sigma_X}\right)$ y $g_Y(s) = \varphi\left(\frac{s-\mu_Y}{\sigma_Y}\right)$ siendo φ la función de densidad de la normal típica, $\mu_X = \mathbb{E}(d(X_1, X_2))$, $\sigma_X^2 = \mathbb{V}(d(X_1, X_2))$ y análogamente se definen μ_Y y σ_Y.

En la práctica, como no se conocen μ_X ni σ_X se los estima por la media muestral y la desviación muestral de las distancias entre los elementos, o sea que se utilizan $m_X = \frac{1}{n(n-1)} \sum_{i \neq j} d\left(X_i, X_j\right)$ y $s_X^2 = \frac{1}{n(n-1)} \sum_{i \neq j} \left(d\left(X_i, X_j\right) - m_X\right)^2$ para estimar los valores de μ_X y σ_X^2 respectivamente. Análogamente se trabaja con m_Y y s_Y^2.

La forma en la cual se puede implementar el test estadístico, las propiedades teóricas y los resultados del test en diversas simulaciones se puede ver en ([32]).

Part III
Resultados

7 Datos

Se parte de información otorgada por Facultad de Ciencias referente a precipitaciones diarias de 47 estaciones meteorológicas y pluviométricas, localizadas en todo el territorio uruguayo, para el período de enero 1980 a diciembre 2013. Las estaciones meteorológicas son Aeropuerto de Carrasco, Aeropuerto de Melilla, Artigas, Bella Unión, Colonia, Durazno, Florida, Melo, Mercedes, Paso de los Toros, Paysandú, Prado, Punta del Este, Rivera, Rocha, Salto, Tacuarembó, Treinta y Tres, Trinidad y Young. Estaciones pluviométricas eran las siguientes: Acegua, Arbolito, Cañas, Casupa, Cerro Colorado, Chuy, Conchillas, Cuchilla de Dionisio, Cufre, Ecilda Paullier, El Cerro, Florencio Sanchez, Isidoro Noblia, Marincho, Mendoza, Minas de Corrales, Ombues de Lavalle, P. de Valentin (Rincón), Palmitas, Paso de la Cruz, Pueblo Biassini, San Gregorio, Santa Catalina, Tararíras, Tomas Gomensoro, Tres Boliches y Valle Eden.

En Uruguay existen 3 tipos de estaciones de captura de datos de fenómenos climáticos: las estaciones meteorológicas convencionales, las estaciones pluviométricas convencionales y las estaciones meteorológicas automáticas. La diferencia entre una estación meteorológica y una pluviométrica es que en las estaciones meteorológicas se miden diversos fenómenos como ser: precipitaciones, temperatura, humedad, granizo, etc. y además son gestionadas por personal específicamente preparado y calificado. Por otro lado, las estaciones pluviométricas solo miden las precipitaciones y en general la calidad de los datos de éstas es inferior a las de las estaciones meteorológicas. Las mismas forman la Red Pluviométrica Nacional y la Red Meteorológica Nacional. A la fecha existen 19 estaciones meteorológicas convencionales activas pertenecientes a la Red Meteorológica Nacional y 8 estaciones meteorológicas de INIA.

Uno de los inconvenientes encontrados, fueron los datos faltantes. En esta tesis no se profundizó sobre la imputación de datos faltantes ni su impacto en las estimaciones obtenidas. Para depurar la base de datos, se priorizó mantener la mayor cantidad años, sacrificando la eliminación de estaciones meteorológicas. Para obtener dicho objetivo se procedió de la siguiente manera:

- Se eliminaron las estaciones cuyo cociente entre cantidad de años con más del 20% de datos faltantes respecto de los años totales supere el 80%. Es decir, si $C = P/N$ siendo, P = cantidad de años con datos faltantes mayor al 20% y $N = 34$ (1980 al 2013), entonces si C es mayor a 0,8 , esa estación se elimina.
- Una vez eliminadas las estaciones con mayor número de datos faltantes, se decidió eliminar el año 1980 ya que es el único año en que alguna de las estaciones que quedaban poseían datos faltantes ese año, priorizando en este caso no eliminar ninguna otra estación meteorológica.

De esa forma se obtuvo la base de datos final de precipitaciones diarias desde enero 1981 a diciembre de 2013 para 19 estaciones meteorológicas y 1 pluviométrica (Palmitas) que se ven en gráfico a continuación:

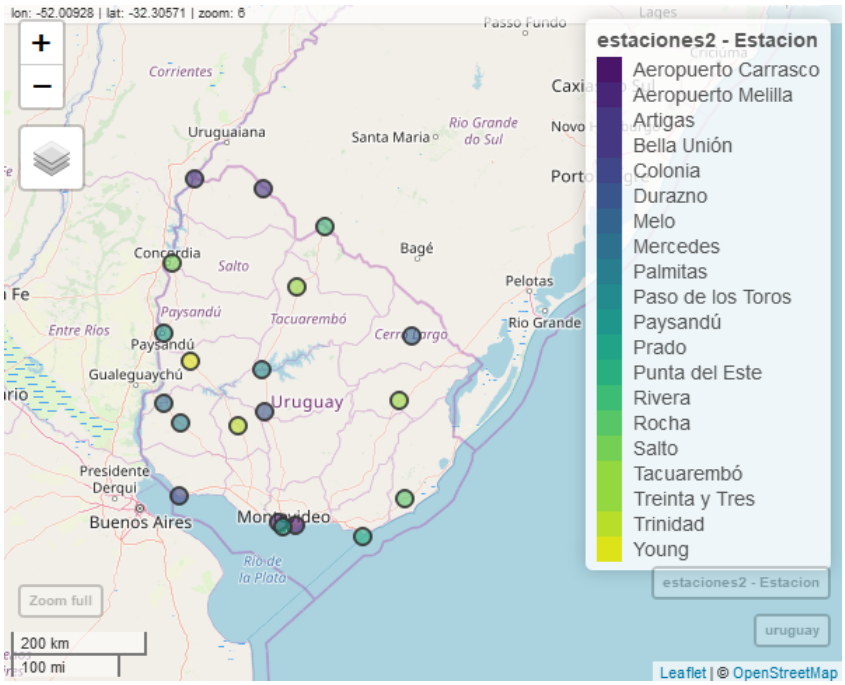


Figura 7.1: Estaciones meteorológicas y pluviométricas utilizadas. Fuente: elaboración propia en R.

A pesar de la distinción vista anteriormente respecto de una estación meteorológica y una estación pluviométrica, en la presente tesis se utilizará indistintamente el término estación meteorológica o estación para referirse a cualquiera de las 20 localidades con las que se trabajará de ahora en más.

Todos los resultados del presente trabajo han sido obtenidos con el software R [43]. Se han utilizado los siguientes paquetes: *readxl,cluster,plot3D,maptools, mapview,rgdal, rgeos, raster, ggplot2, rgl, spdep, caret, tmap, geospt, leaflet, plyr, DMwR, SpatialExtremes, extRemes, goft,xtable,evd,ismev,reshape2.*

8 Análisis de la base de datos

Para la aplicación de las distintas metodologías que se explicarán a partir de la sección siguiente, se trabajará con las precipitaciones máximas anuales diarias. De aquí en más al referirnos a precipitaciones o lluvias extremas, se dará por entendido que nos referimos a los valores máximos anuales diarios. En esta sección, se realizará un estudio descriptivo de las mismas. Comenzando por el gráfico 8.1 a continuación, en el que se visualizan las precipitaciones extremas anuales, desde 1981 a 2013 para cada una de las localidades en estudio:

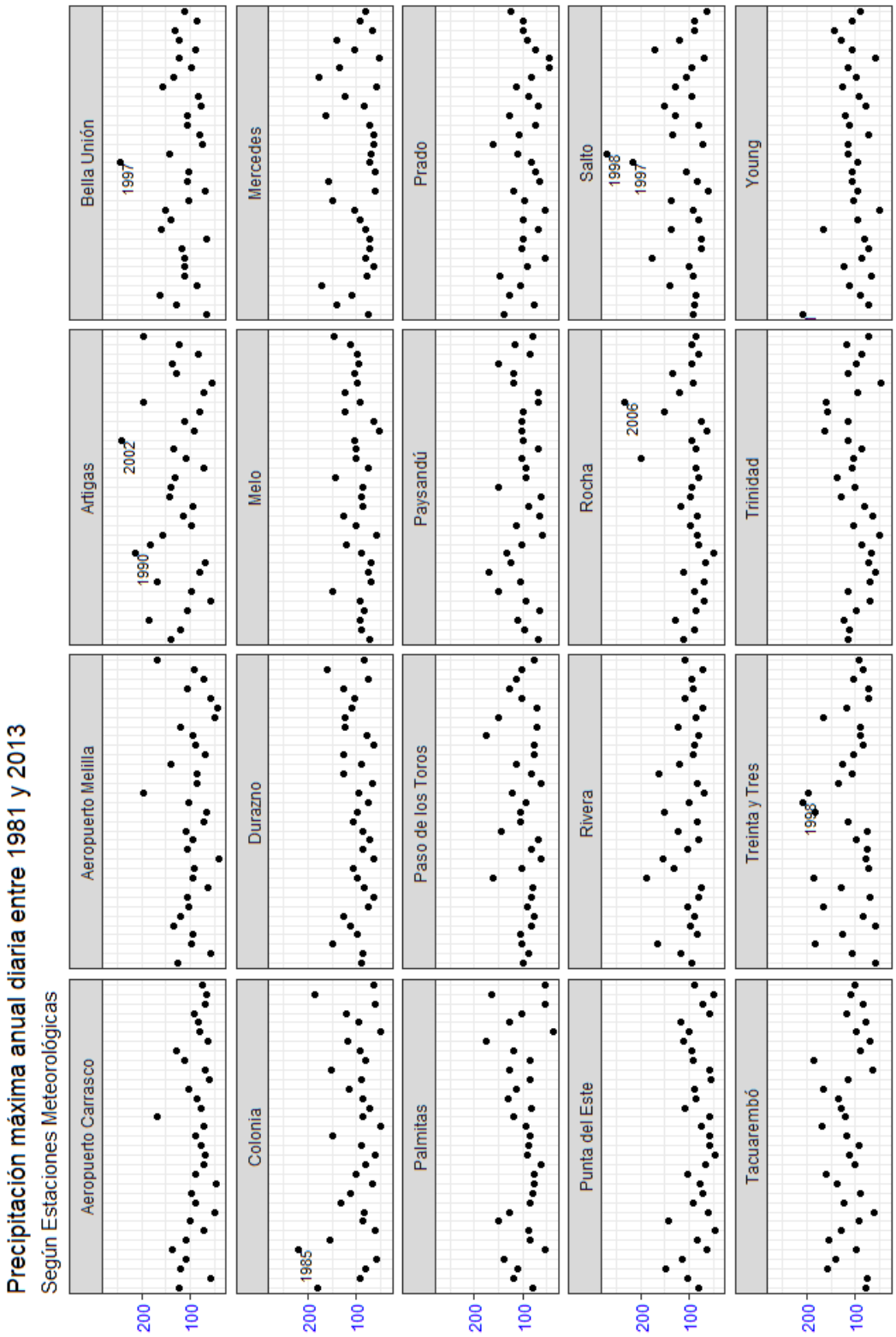


Figura 8.1: Precipitaciones extremas anuales según estación meteorológica. Fuente: elaboración propia en R.

En gráfico 8.1 se han destacado los valores extremos anuales que superaron los 200 mm. Se puede observar que hay estaciones meteorológicas en que su valor de precipitación máxima anual nunca superó dicho valor en el período de estudio 1981 a 2013. Las mismas son Aeropuerto de Carrasco, Aeropuerto de Melilla, Durazno, Melo, Mercedes, Palmitas, Paso de los Toros, Paysandú, Prado, Punta del Este, Rivera, Tacuarembó y Trinidad. Pero también existen estaciones cuyos valores de precipitaciones extremas anuales han superado los 200 mm, incluso más de una vez en el período estudiado. Las estaciones que alcanzaron o superaron dicho valor fueron: Artigas en el año 1990 y en 2002, Bella Unión en el año 1997, Colonia en el 1985, Rocha en el 2006, Salto en el 1997 y 1998, Treinta y Tres en 1998, Salto y Young en 1981. Como se observa, en los años 1997 y 1998 se registraron la mayor cantidad de estaciones afectadas por precipitaciones extremas.

Para tener una dimensión respecto de las lluvias extremas anuales se puede comparar con los valores acumulados por estación del año (otoño, invierno, primavera y verano[6]). En el gráfico 8.2 se puede observar para cada estación meteorológica un boxplot de las precipitaciones acumuladas en los trimestres otoño, invierno, primavera y verano.

Del gráfico 8.2 puede observarse por ejemplo que para la estación Artigas, el valor de la mediana de las precipitaciones acumuladas para invierno, primavera y verano es de 375 mm a 400 mm. Sin embargo, para en los años 1990 y 2002, según gráfico 8.1, existieron valores de precipitaciones extremas anuales que representaron en un solo día más del 50% del valor acumulado promedio en un trimestre. En Salto, los valores medios de precipitaciones acumuladas promedio en las distintas estaciones del año no superaron los 375 mm y sin embargo en los años 1997 y 1998 existieron valores de precipitaciones diarias que alcanzaron los 250 mm.

[6]En el presente trabajo se tomó como Otoño el período entre 21/03 al 20/06, Invierno entre 21/06 a 20/09, Primavera entre 21/09 al 20/12 y Verano entre 21/12 al 20/03

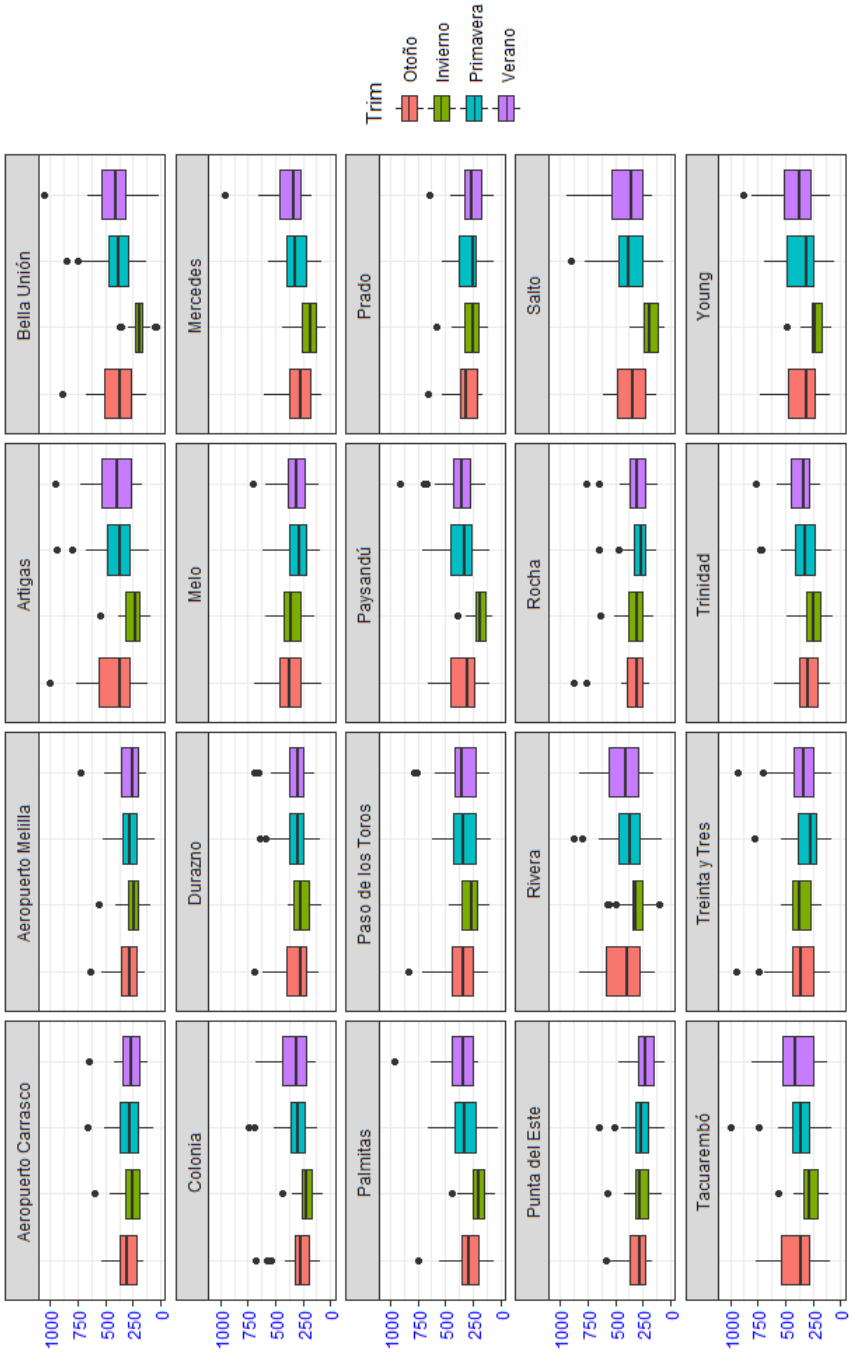


Figura 8.2: Boxplot de las precipitaciones acumuladas en otoño, invierno, primavera y verano por estación. Fuente: elaboración propia en R.

En el gráfico 8.3 se puede apreciar aún más las similaridades o disimilaridades del comportamiento de las lluvias extremas entre las estaciones involucradas, a partir de un boxplot generado con los 33 valores de las precipitaciones extremas anuales para cada localidad:

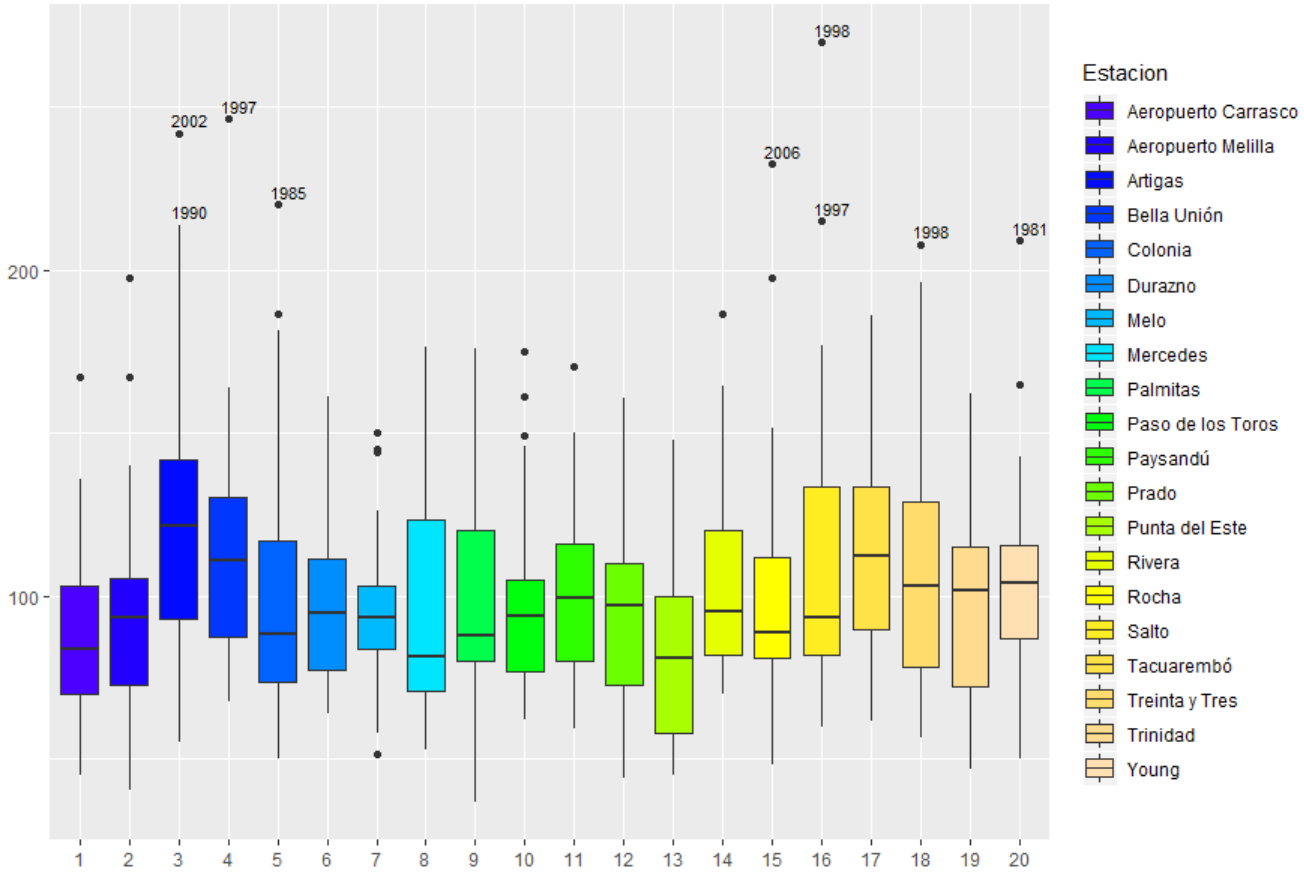


Figura 8.3: Boxplot de las lluvias extremas anuales por estación meteorológica. Fuente: elaboración propia en R.

Adicionalmente a lo que se visualiza en el gráfico (8.1), en el gráfico (8.3) se puede ver el comportamiento inter anual de las lluvias extremas para el período de estudio 1981 a 2013 para cada una de las localidades en estudio. Se puede por un lado observar que las estaciones de Artigas, Bella Unión, Colonia, Rocha, Salto, Treinta y Tres y Young son las que registraron precipitaciones extremas anuales por encima de los 200 mm en distintos años, dándose en el 1997 registros importantes en Bella Unión y Salto y en el 1998 en Salto y Treinta y Tres. Puede destacarse también Salto como la localidad que tuvo la mayor variabilidad inter anual en el comportamiento de dicho fenómeno. Por otro lado, si se comparan las estaciones en función de valores medios, como por ejemplo, en función a la mediana, puede observarse que Artigas, Bella Unión y Tacuarembó presentan valores de mediana por encima de las demás estaciones, es decir, promedialmente las lluvias extremas anuales en estas localidades estuvieron por encima que en demás regiones del país. Se destacan a Melo y Punta del Este como las estaciones con menor variabilidad inter anual, en cuyos casos, en ningún año del período en estudio, se superaron los 150 mm de precipitación extrema anual.

Tamién podría interesar estudiar el comportamiento intra anual de las lluvias extremas anuales. Para ello se presenta el gráfico 8.4 a continuación, en el cual se visualizan las lluvias extremas anuales para cada año en las distintas localidades de estudio. Se realiza un boxplot con los 20 valores extremos anuales de las distintas ubicaciones para un mismo año:

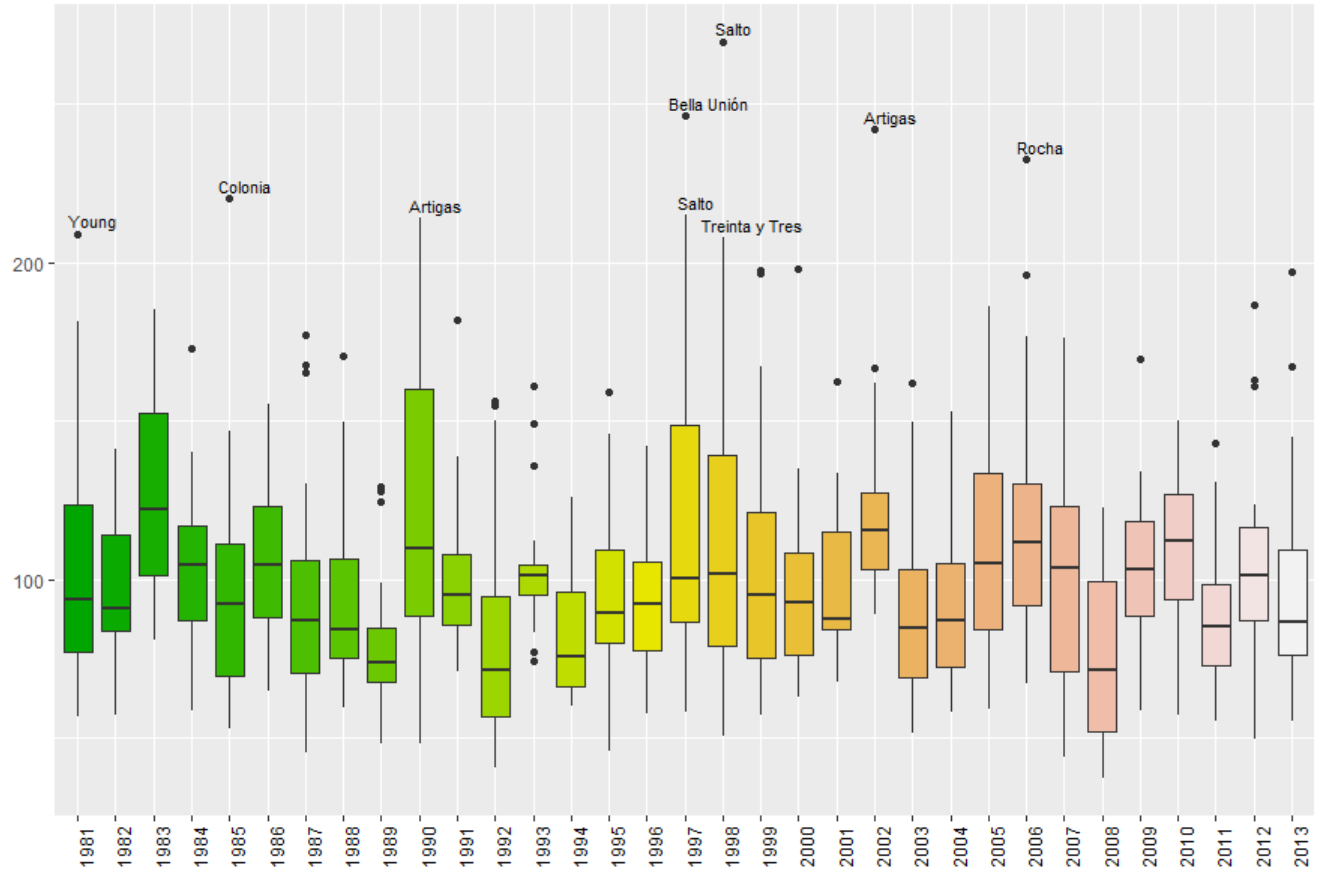


Figura 8.4: Boxplot de las precipitaciones extremas anuales por año. Fuente: elaboración propia en R.

Del gráfico anterior, 8.4, se destacan los años 1997 y 1998 con valores de precipitaciones máximas anuales con mayor variabilidad espacial que los restantes años. Por ejemplo, para el año 1997 mientras que el 75% de las estaciones registraron valores de precipitaciones máximas anuales menores o iguales a los 150 mm, en ese mismo año las localidades de Rivera, Treinta y Tres, Salto y Bella Unión tuvieron registros máximos de 152, 184, 215 y 246 mm respectivamente. En el año 1998 las estaciones más afectadas fueron Tacuarembó (170 mm), Treinta y Tres (208 mm) y Salto (270 mm), siendo éste último el registro más alto de todo el período en estudio. Otro año que se puede destacar, es el 2002, en donde se destacan dos estaciones con altos registros extremos, Tacuarembó (167 mm) y Artigas (242 mm) y a su vez fue el año cuyos registros de lluvias extremas en general fueron superiores a los demás años. Por el contrario el año con menores registros de lluvias extremas anuales en todo el territorio del país fue el 1989, en el que salvo para tres estaciones (Paysandú con 124 mm, Palmitas con

128 mm y Treinta y Tres con 129 mm), las demás no tuvieron registros de lluvias extremas anuales superiores a los 100 mm. En general, Artigas y Salto puede verse como las estaciones que más se repiten como las que se destacan con altos registros de lluvias máximas anuales. También se destaca el año 1990 en el que el percentil 75% se ubica por encima de los 150 mm, el más alto respecto a los demás años. Es decir, en dicho año la mayoría de las localidades en estudio fueron afectadas por precipitaciones intensas.

De los años mencionados anteriormente, se sabe que algunos de ellos, fueron afectados por el fenómeno de El Niño, como por ejemplo, los años 1997, 1998 y 2002. El fenómeno de El Niño es un evento climático relacionado al calentamiento del Pacífico oriental ecuatorial, el cual se manifiesta erráticamente cíclico (se habla de ciclos de entre tres a ocho años). Dicho fenómeno también conocido como ENSO (siglas en inglés correspondientes a El Niño Oscilación Sur). Este fenómeno se manifiesta de forma de lluvias intensas, afectando principalmente a la región costera del Pacífico de América del Sur ([56]). El Servicio Nacional del Clima (NOAA por sus siglas en inglés) es quien provee estadísticas históricas de dicho fenómeno. En [36] se puede observar el Índice Oceánico de El Niño (ONI[7] por sus siglas en inglés) que dicha asociación utiliza para identificar eventos cálidos (El Niño) y fríos (La Niña) en el océano Pacífico tropical.

Según dicho índice los años afectados por el fenómeno de El Niño, en el período de estudio, fueron: 1982 a 1983, 1986 a 1988, 1991 a 1992, 1994 a 1995, 1997 a 1998, 2002 a 2003, 2004 a 2005, 2006 a 2007, 2009 a 2010. Sin embargo, no todos esos años se reflejan valores extremos anuales de precipitaciones importantes. De hecho, existen otros años en los que las lluvias extremas anuales superaron los 200 mm y no fue en año Niño, como ser los años 1981, 1985 y 1990. Los años en que sí puede decirse que el fenómeno de El Niño pudo haber incidido en el comportamiento de las lluvias extremas es en los años 1997, 1998, 2002 y 2006.

Así como las precipitaciones en general tienen un comportamiento estacional, podría ser de interés analizar el comportamiento estacional de las precipitaciones extremas. Para ello una primer aproximación se realiza al graficar las lluvias extremas ocurridas en cada trimestre (estación del año) otoño, invierno, primavera o vearno. Como se mencionó anteriormente, se considera otoño desde el 21/Marzo al 20/Junio , invierno desde el 21/Junio al 20/Setiembre , primavera desde el 21/Setiembre al 20/Diciembre y verano desde el 21/Diciembre a 20/Marzo. A continuación, el gráfico 8.5 muestra las lluvias extremas ocurridas en cada trimestre, durante el período 1981 a 2013 (tomando todas las localidades del país juntas):

[7] El índice ONI se calcula como la media móvil de tres meses de las anomalías de la temperatura superficial del mar para la región El Niño 3.4 (es decir, la franja comprendida entre 5°N-5°S y 120°-170°W).

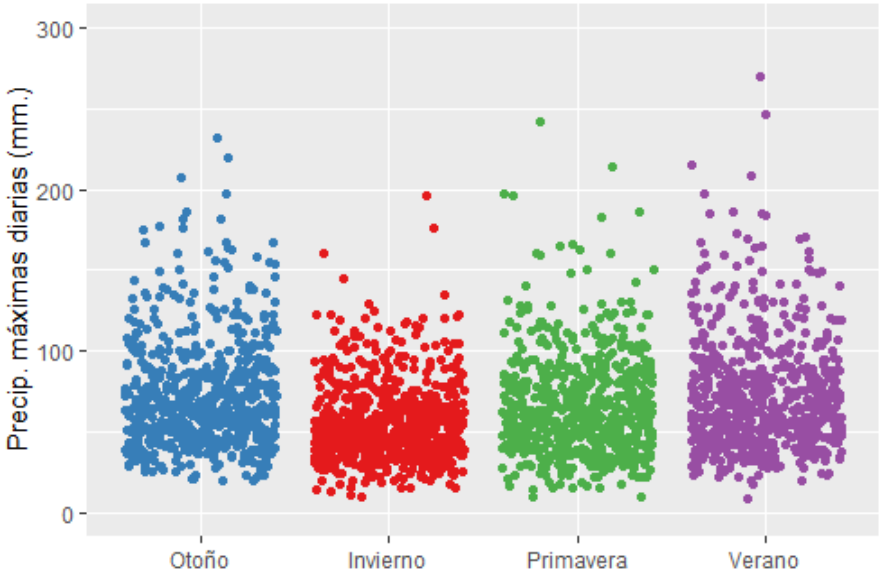


Figura 8.5: Lluvias extremas en cada estación del año. Fuente: elaboración propia en R.

Del gráfico 8.5 se puede observar que la estación del año con menor variabilidad respecto de las precipitaciones extremas es invierno, incluso dichos valores no superan los 200 mm diarios. En el sentido opuesto, se visualiza al trimestre verano como la estación con mayor dispersión respecto de los valores extremos de precipitaciones, incluso dándose en ese trimestre el mayor valor de precipitación extrema que supera los 250 mm (puntualmente ese hecho se dio en la estación Salto en el año 1998 llegando a un máximo trimestral diario de 270 mm). Las estaciones de otoño y primavera parecieran tener un comportamiento bastante similar entre sí en lo que a precipitaciones extremas refiere.

También como primer exploración de búsqueda de patrones espaciales, se divide a las estaciones en regiones, sur y norte, según se encuentren al sur o norte del Río Negro respectivamente. En el gráfico 8.6 se presentan las distribuciones de las estaciones de la base de datos según la región correspondiente:

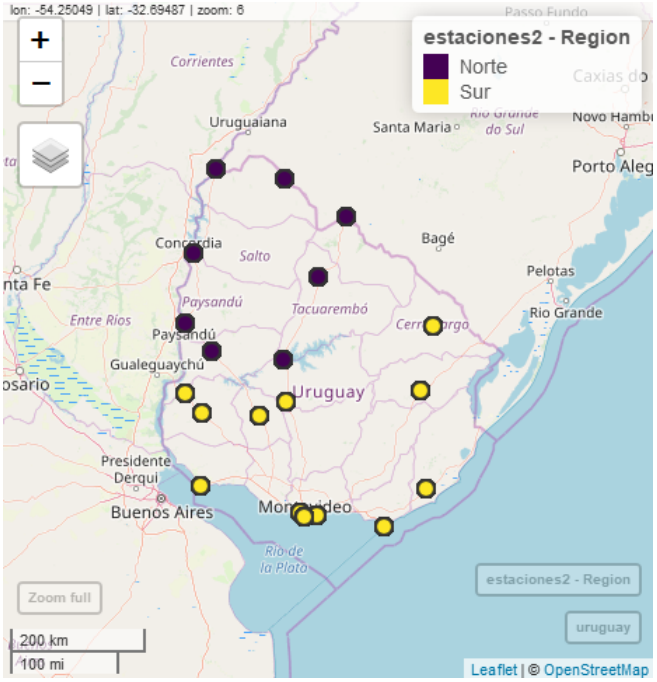
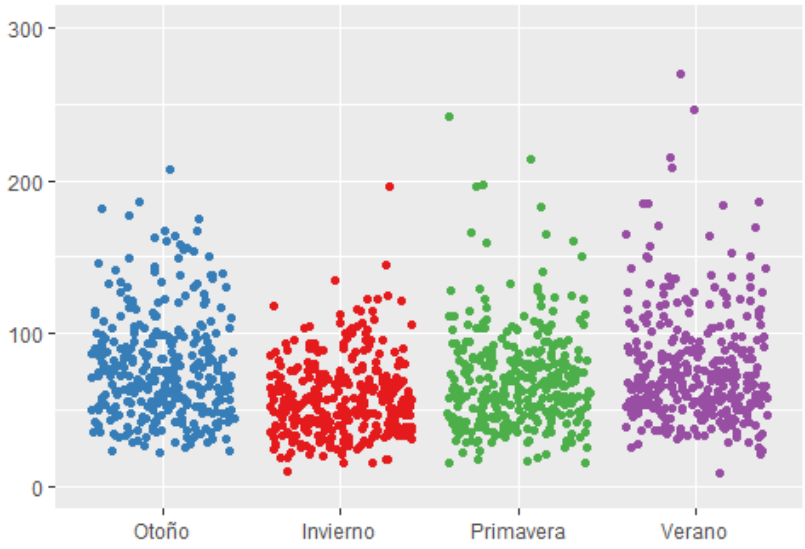
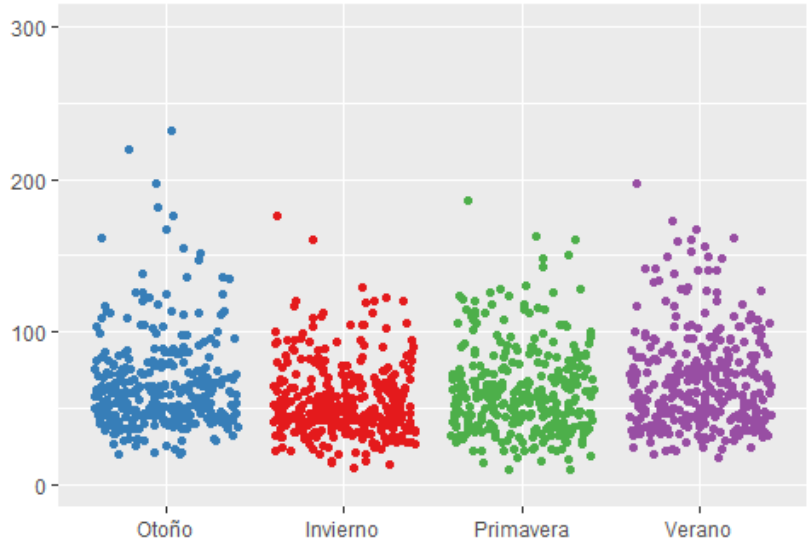


Figura 8.6: Estaciones meteorológicas según región.

Dada esta clasificación de las estaciones según gráfico 8.6, se procedió a comparar los valores de las precipitaciones extremas diarias en los trimestres otoño, invierno, primavera y verano, según cada región. En los siguientes gráficos se puede visualizar lo antes dicho:



(a) Lluvias extremas región Norte. Fuente: elaboración propia en R.



(b) Lluvias extremas región Sur. Fuente: elaboración propia en R.

De los gráficos anteriores se destaca que en la región norte, tal como se visualiza en la Figura 8.7(a), las precipitaciones extremas diarias de los trimestres de primavera y verano son mayores en comparación con los valores extremos en mismos trimestres de la región sur. Lo que puede explicarse tanto por mayor variabilidad inter anual como mayor aumento de variabilidad espacial respecto de las estaciones de la misma región. El trimestre invierno, en ambas regiones es la estación del año en donde los registros máximos alcanzados son menores que demás trimestres y además resulta ser el trimestre con menor dispersión. La estación otoño resultó más afectada en la región sur que la norte, registrándose en dicha región máximos cercanos a los 250 mm mientras que en la región norte apenas sobrepasan los 200 mm.

Lo resultados descriptivos arrojan suficiente evidencia como para pensar que las distribu-

ciones de valores extremos anuales de cada estación serán diferentes, con distintos parámetros de ubicación, escala y forma así como posiblemente pertenezcan a distintas familias de distribuciones GEV. En la próxima sección se analizarán las distribuciones GEV que mejor ajustan a cada estación.

9 Estimación de la distribución GEV

En esta sección se presentarán los resultados obtenidos de las estimaciones de las distribuciones marginales GEV para cada estación meteorológica.

Como se mencionó anteriormente, se trabajará con los valores de las precipitaciones máximas anuales diarias. Para esta sección justamente resulta de utilidad trabajar con bloques anuales, ya que de esa forma, se podrá evitar la estacionalidad intra anual, supuesto que violaría la correcta aplicación de dicha teoría tal como se detalló en 5.

El primer paso fue estimar los parámetros mediante los métodos de los momentos ponderados pesados, máxima verosimilitud y máxima verosimilitud perfil según se detalló en sección 5.2. Luego se calcularon intervalos de confianza para los mismos, y se procedió a realizar el test de hipótesis para el parámetro de forma ξ para determinar la familia óptima GEV a ajustar. Por último se realiza el diagnóstico de cada una de las densidades estimadas. Se comparan los resultados comparando los niveles de retorno estimados para cada caso.

9.1 Estimación de parámetros

En la tabla 9.1 a continuación, se presentan los parámetros estimados para cada una de las densidades marginales GEV estimadas de las estaciones a partir del método de máxima verosimilitud:

	Estación	Par_posicion	Par_escala	Par_forma
1	Aeropuerto Carrasco	76.20	21.16	-0.02
2	Aeropuerto Melilla	81.09	27.99	-0.07
3	Artigas	103.74	37.71	-0.02
4	Bella Unión	97.60	25.58	0.07
5	Colonia	80.38	26.37	0.20
6	Durazno	86.76	19.59	-0.01
7	Melo	87.07	21.46	-0.15
8	Mercedes	76.09	20.39	0.39
9	Palmitas	85.76	29.17	-0.16
10	Paso de los Toros	84.61	18.77	0.15
11	Paysandú	88.26	23.83	-0.08
12	Prado	82.82	26.77	-0.20
13	Punta del Este	70.21	20.74	0.00
14	Rivera	88.69	17.54	0.31
15	Rocha	83.63	22.13	0.12
16	Salto	89.08	24.23	0.29
17	Tacuarembó	100.28	28.78	-0.14
18	Treinta y Tres	89.79	28.35	0.21
19	Trinidad	88.17	27.80	-0.20
20	Young	90.03	24.54	-0.04

Tabla 9.1: Parámetros estimados a partir del método máxima verosimilitud. Fuente: elaboración propia en R.

Si bien se aplicaron los distintos métodos de estimación de los parámetros de las densidades GEV marginales mencionados anteriormente, no se encontraron diferencias significativas, por ende se decidió continuar en base a la estimación máximo verosímil. En Anexo VI se presentan los parámetros estimados de cada una de las distribuciones marginales extremas para cada estación según los métodos de los momentos ponderados y máxima verosimilitud perfil.

Se presentan a continuación gráficos de algunas de las densidades empíricas de las precipitaciones extremas anuales, comparando aquellas con mayor diferencia entre los valores del parámetro de posición, parámetro de escala y parámetro de forma.

Por ejemplo, si comparamos según las mayores diferencias respecto de los valores estimados del parámetro de posición, vemos por un lado que la estación de Artigas posee el mayor valor de dicho parámetro de posición ($\hat{\mu}$ = 103.74) mientras que la estación de Punta del Este posee el valor mínimo del parámetro de posición ($\hat{\mu}$ = 70.2068):

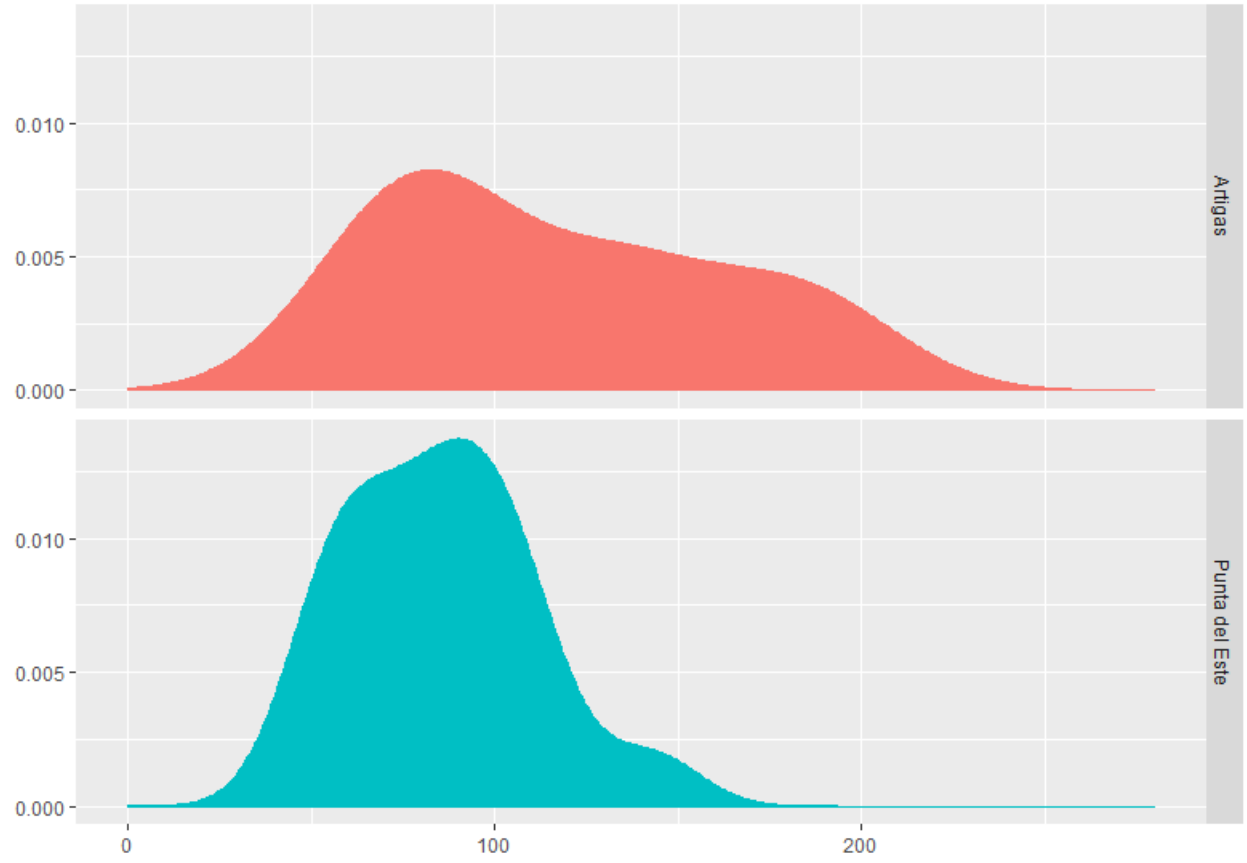


Figura 9.1: Comparación de densidades empíricas según distintos parámetros de posición. Fuente: elaboración propia en R.

El gráfico anterior sugiere que las estaciones con mayor valor estimado del parámetro de posición son densidades de colas más pesadas. Además los valores medios de precipitaciones extremas de las estaciones con mayores valores del parámetro de locación, son mayores que el resto.

Si comparamos estaciones respecto de los valores estimados de los parámetros de escala, se podrían comparar por ejemplo Artigas (con valor de parámetro de escala estimado de $\hat{\sigma} = 37.71$) contra la estación de Rivera (con valor de parámetro de escala estimado de $\hat{\sigma} = 17.54$):

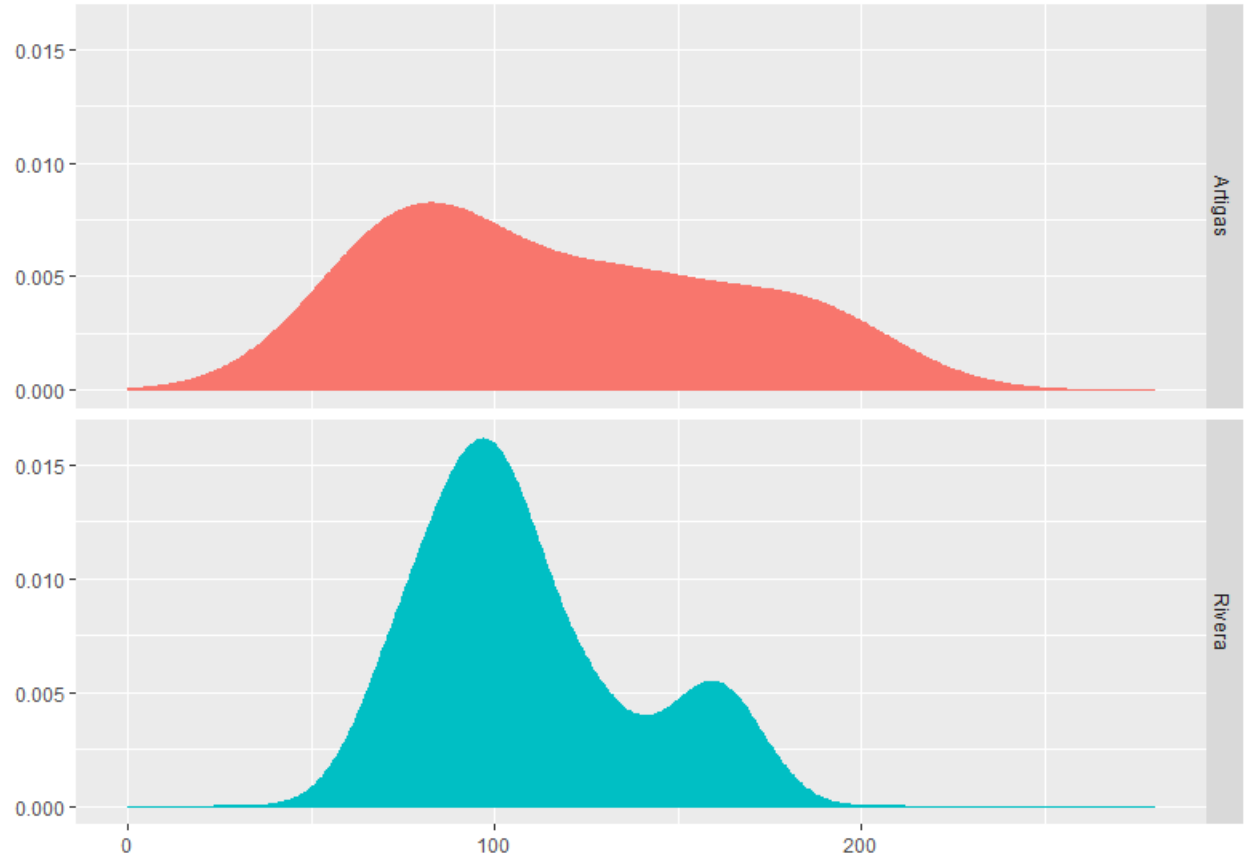


Figura 9.2: Comparación de densidades empíricas según el parámetro de escala. Fuente: elaboración propia en R.

La Figura 9.2 sugiere que las densidades con mayores valores del parámetro de escala tienen mayor variabilidad respecto de las otras con menores valores de dicho parámetro.

Por último, si comparamos las densidades según el parámetro de forma, sabemos que van a diferir en el tipo de familia GEV según sea Fréchet (con parámetro de forma positivo), Weibull (con parámetro de forma negativo) y Gumbel (con parámetro de forma que tiende a 0):

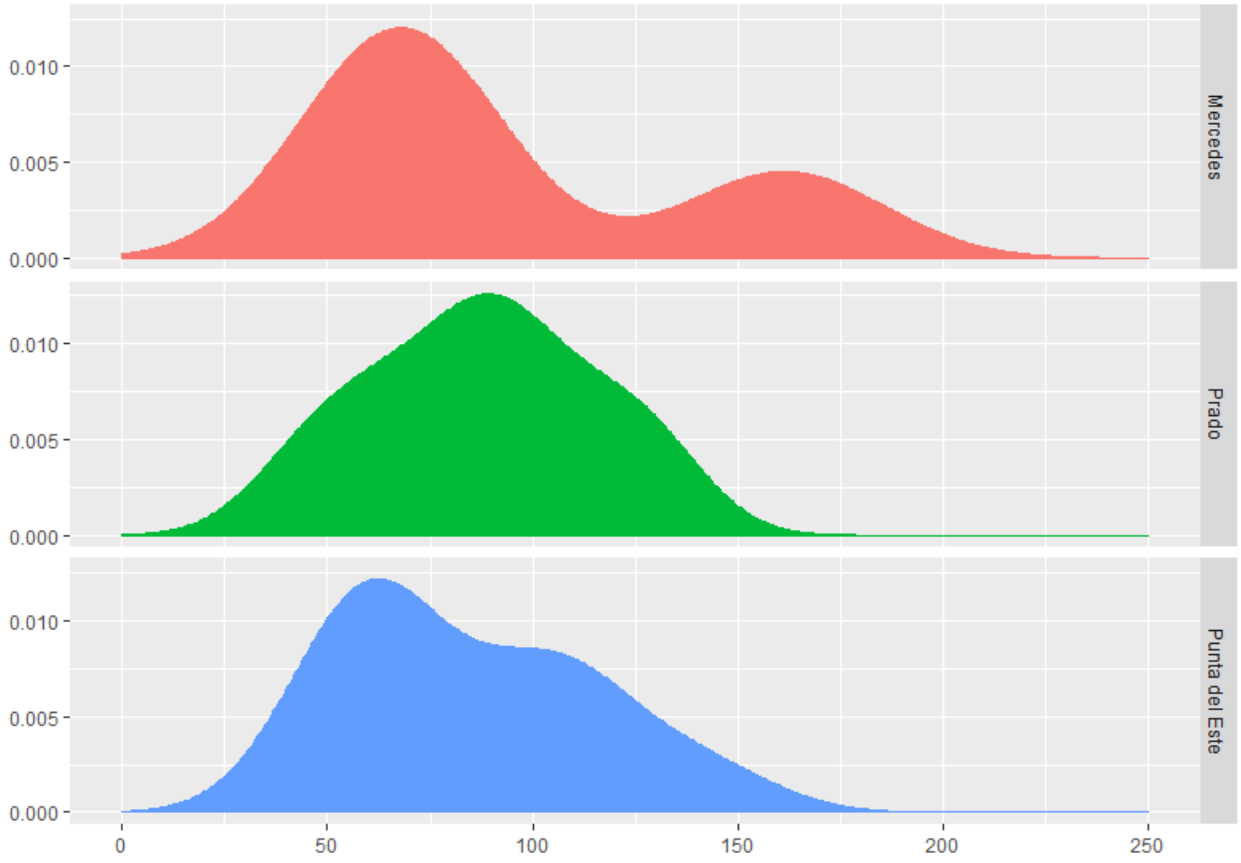


Figura 9.3: Comparación de densidades empíricas según el parámetro de forma. Fuente: elaboración propia en R.

En el gráfico 9.3 se compara la estación con mayor valor del parámetro de forma positivo (correspondiente a la estación de Mercedes con $\hat{\xi} = 0.3895$), la estación con el valor del parámetro más cercano a cero (correspondiente a la estación de Punta del Este con $\hat{\xi} = 0.0022$) y la estación con menor valor de dicho parámetro (correspondiente a la estación de Prado con $\hat{\xi} = -0.2023$).

9.2 Cálculo de intervalos de confianza

Se calcularon los intervalos al 95% de confianza para cada parámetro de cada estación meteorológica. Los intervalos de confianza se calcularon utilizando boostrap. Para ello, para cada estación, se siguen los siguientes pasos:

1. Se simula una muestra de tamaño n (tamaño de muestra original) con los parámetros estimados.
2. Se ajusta una distribución GEV a la muestra simulada.
3. Se repiten los pasos 1 y 2, m veces.
4. Se calculan los intervalos de confianza a partir de los pasos anteriores.

Si bien se continúan los análisis con las estimaciones máximo verosímiles, aquí se presenta una comparación gráfica de los intervalos de confianza calculados a partir de máxima verosimilitud o máxima verosimilitud perfil, para la estación Artigas. En los gráficos a continuación se comparan los intervalos de confianza para el parámetro a partir de los dos métodos (en color verde el intervalo de confianza calculado a partir del método máxima verosimilitud y en color azul el intervalo de confianza calculado a partir del método máxima verosimilitud perfil):

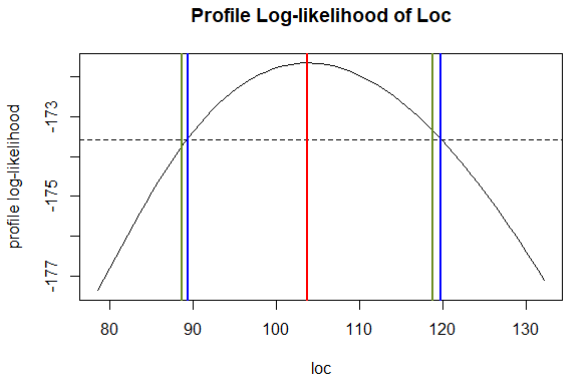


Figura 9.4: Comparación intervalos de confianza para el parámetro de locación. Fuente: elaboración propia en R.

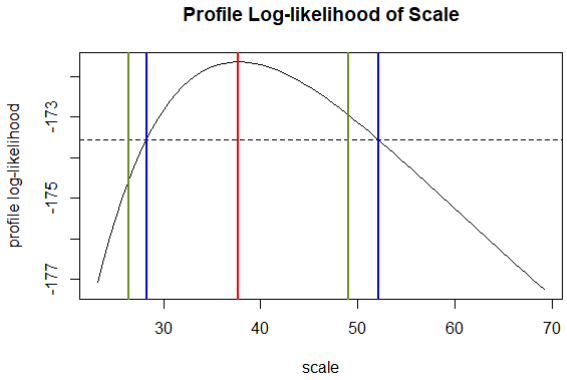


Figura 9.5: Comparación intervalos de confianza para el parámetro de escala. Fuente: elaboración propia en R.

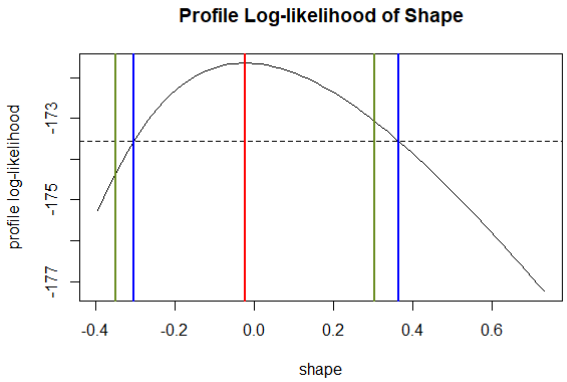


Figura 9.6: Comparación intervalos de confianza para el parámetro de forma. Fuente: elaboración propia en R.

De los gráficos anteriores se visualiza la asimetría de los intervalos de confianza calculados por el método de máxima verosimilitud perfil. Los intervalos de confianza del parámetro de locación, según se ve en Figura 9.4, resultaron muy similares. Mientras que los intervalos de confianza del parámetro de escala, según gráfico 9.5 calculado por el método de máxima verosimilitud perfil resulta asimétrico y sesgado a la derecha.

Si bien se calcularon intervalos de confianza para todos los parámetros, en particular hacemos foco en los resultados obtenidos para los parámetros de forma, ya que como se vio en sección (5.1) determinan el tipo de familia de distribución GEV. A continuación se muestra los intervalos de confianza para el parámetro ξ para cada estación meteorológica (mientras que los IC para los parámetros de ubicación y escala se pueden ver en Anexo VI):

Estación	Parámetros	2.5%	Estimación	97.5%
Aeropuerto Carrasco	Forma	-0.28	-0.02	0.23
Aeropuerto Melilla	Forma	-0.29	-0.07	0.15
Artigas	Forma	-0.35	-0.02	0.30
Bella Unión	Forma	-0.21	0.07	0.34
Colonia	Forma	-0.14	0.20	0.53
Durazno	Forma	-0.35	-0.01	0.33
Melo	Forma	-0.40	-0.15	0.10
Mercedes	Forma	0.02	0.39	0.76
Palmitas	Forma	-0.41	-0.16	0.10
Paso de los Toros	Forma	-0.19	0.15	0.48
Paysandú	Forma	-0.39	-0.08	0.23
Prado	Forma	-0.46	-0.20	0.05
Punta del Este	Forma	-0.35	0.00	0.35
Rivera	Forma	-0.08	0.31	0.69
Rocha	Forma	-0.09	0.12	0.33
Salto	Forma	-0.03	0.29	0.61
Tacuarembó	Forma	-0.47	-0.14	0.19
Treinta y Tres	Forma	-0.17	0.21	0.58
Trinidad	Forma	-0.47	-0.20	0.07
Young	Forma	-0.23	-0.04	0.16

Tabla 9.2: Intervalos de confianza al 95% para los parámetros de forma. Fuente: elaboración propia en R.

Como se puede apreciar en la Tabla 9.2, existen ciertas estaciones cuyo parámetro de forma estimado es cercano a cero e incluso el intervalo al 95% de confianza contiene al cero. Por lo tanto se propone realizar un test de hipótesis de Cramér-von Mises recortado segun se vió en sección (5.2.4) para testear si el parámetro de forma es cero, es decir que la distribución estimada es de la familia Gumbel.

9.3 Test de bondad de ajuste de la distribución GEV de cada estación

En la tabla a continuación se presentan los p-valores de 20 pruebas realizadas (una por cada localidad), obtenidos a partir 1000 réplicas del estadístico T_n y contabilizando el porcentaje de veces de entre las 1000 en los cuales (siendo Gumbel la variable observada) el valor de dicho estadístico supera al observado en la muestra en concreto, según se vio en 5.2.4.

A partir de la Tabla 9.3 se observa que, para un nivel de significación del 5%, salvo los casos de los departamentos de Mercedes y de Rocha, no se rechaza la hipótesis de que los datos se ajustan a una distribución del tipo de familia Gumbel.

Para las estaciones de Mercedes y Rocha, se propone realizar el mismo test para verificar si el modelo que mejor ajusta a los datos de cada una es del tipo de familia Fréchet, es decir se testea si el parámetro de forma es positivo. Para la aplicación del test en este caso, es necesario primero realizar las transformaciones correspondientes según visto en 5.1 (Fréchet a

Aer. Carrasco	Aer. Melilla	Artigas	Bella Unión	Colonia
1	0.32	0.89	0.89	0.46
Durazno	Melo	Mercedes	Palmitas	Paso de los Toros
0.6	0.45	0.03	0.28	0.54
Paysandú	Prado	Punta del Este	Rivera	Rocha
0.23	0.43	0.56	0.36	0.03
Salto	Tacuarembó	Treinta y Tres	Trinidad	Young
0.16	0.64	0.26	0.19	0.56

Tabla 9.3: p-valores de la prueba de bondad de ajuste de las distribuciones GEV marginales. Fuente: elaboración propia en R.

Gumbel). Una vez realizada dicha transformación se siguen los mismos pasos vistos anteriormente.

Los p-valores resultaron iguales a 0.099 y 0.335 respectivamente. Entonces al 5% de significación no se rechaza la hipótesis nula de que la distribución marginal GEV que mejor ajusta a los datos de ambas estaciones son del tipo de familia Fréchet.

También se realizó el test de ratio de verosimilitud según fue detallado en sección (5.2.2) donde se pudo observar que coinciden en que la estación de Mercedes rechaza la hipótesis nula de que los datos se ajusten a una distribución del tipo Gumbel. Sin embargo, para este último test, también resultó que la estación de Salto (y no Rocha) tampoco se ajustaba según Gumbel. En apéndice (VI) se pueden observar los p-valores que arrojó dicho test. Se seguirá adelante teniendo en cuenta los resultados arrojados por el test de Cramér-von Mises recortado.

Ya realizado el test mencionado anteriormente, para todas las estaciones, se vuelve a estimar los parámetros de todas las estaciones menos Mercedes y Rocha, estimando en este caso, sólo los parámetros de locación y escala de distribuciones GEV del tipo Gumbel.

A continuación se presenta los parámetros estimados a partir del método de máxima verosimilitud:

	Estacion	Par. posición	Par. escala
13	Punta del Este	70.22	20.75
1	Aeropuerto Carrasco	75.95	21.02
12	Prado	79.99	25.50
2	Aeropuerto Melilla	80.05	27.52
5	Colonia	83.22	28.95
9	Palmitas	83.40	28.15
19	Trinidad	85.30	26.41
7	Melo	85.34	20.79
10	Paso de los Toros	86.12	20.08
6	Durazno	86.66	19.52
11	Paysandú	87.30	23.18
20	Young	89.54	24.36
14	Rivera	91.81	20.68
18	Treinta y Tres	93.08	31.39
16	Salto	93.20	28.47
17	Tacuarembó	98.19	27.44
4	Bella Unión	98.51	26.26
3	Artigas	103.32	37.39

Tabla 9.4: Parámetros re-estimados por método MLE para las estaciones que no rechazaron la hipótesis nula de distribución Gumbel. Fuente: elaboración propia en R.

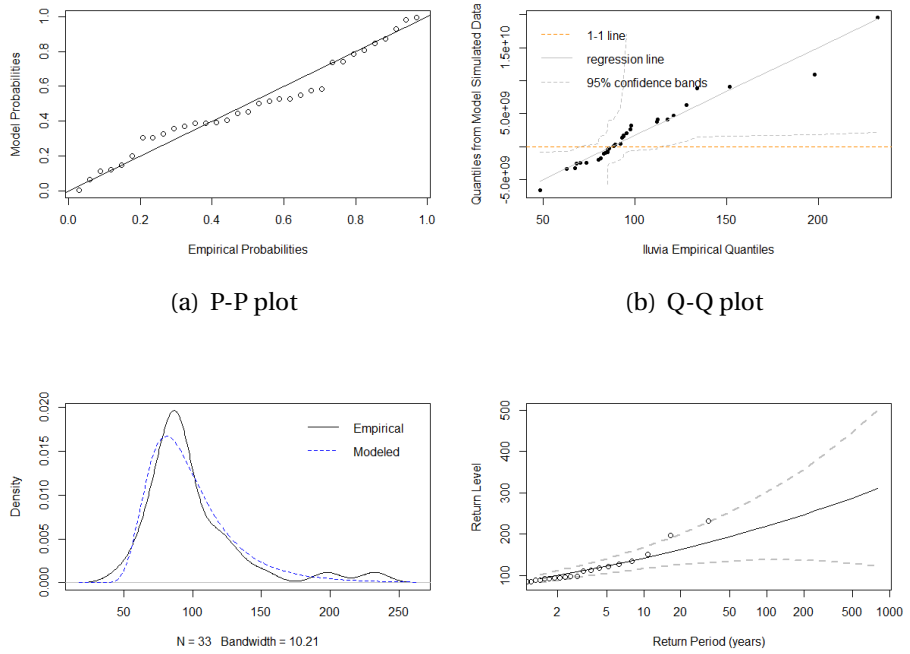
Los intervalos de confianza estimados se pueden ver en apéndice VI.

En conclusión los modelos ajustados a las distribuciones de las lluvias extremas anuales de las distintas localidades en estudio de Uruguay, fueron del tipo Gumbel y Fréchet. Solo las lluvias extremas de las estaciones Mercedes y Rocha se ajustaron mejor según modelo Fréchet, mientras que para las demás localidades el modelo que mejor ajustaba a dicho fenómeno fue la distribución Gumbel. En función a los parámetros estimados presentados en la tabla anterior, se puede observar que, de las localidades ajustadas según distribución Gumbel, las lluvias extremas anuales de las localidades del norte del país, como ser Artigas, Bella Unión, Tacuarembó y Salto presentan en general mayores valores de precipitaciones extremas anuales, resultando Artigas la de mayor variabilidad interanual. Por otro lado, las lluvias extremas anuales del sur y sur este del país, como ser Punta del Este, Aeropuerto de Carrasco y Prado, también modeladas según distribución GEV del tipo Gumbel, presentaron menores valores de parámetro de posición estimados, es decir son distribuciones de cola menos pesadas que las distribuciones marginales del norte del país. Además las distribuciones de las lluvias extremas anuales de las localidades de Mercedes y Rocha resultaron ser del tipo Fréchet con parámetros estimados según los valores de la tabla 9.1. Dichas distribuciones tienden a tener cola más pesada que las distribuciones Gumbel.

Lo anterior se reflejará también en los niveles de retorno estimados que se presentan en 9.5.

9.4 Gráficos de diagnóstico

Todos los modelos obtenidos para cada estación fueron sujetos a los análisis de diagnóstico según se detalló en 5.3. De forma satisfactoria, todas las distribuciones modeladas han mostrado buena performance. A continuación se presenta, como ejemplo, la etapa de diagnóstico del modelo GEV ajustado para la estación Rocha. El modelo ajustado para las lluvias extremas anuales de esta localidad resultó ser un modelo del tipo $Fréchet(\hat{\mu} = 83.63, \hat{\sigma} = 3.10, \hat{\xi} = 0.12)$ y los resultados de su etapa diagnóstica se presenta en los gráficos a continuación:



(a) P-P plot (b) Q-Q plot

(c) Densidad empírica (d) Niveles de retorno

Figura 9.7: Gráficos de diagnóstico de la modelización GEV para Rocha

El gráfico a) de 9.7 es el gráfico de P-P plot, en el que se observa por un lado los cuantiles empíricos en las abscisas y en el eje de ordenadas los cuantiles del modelo ajustado. Como se vio en sección (5.3) un buen ajuste se da cuando los puntos se encuentran sobre la línea $y = x$. Como se puede ver, el ajuste del modelo Fréchet con los parámetros estimados, es un buen ajuste para este caso.

El gráfico b) de 9.7 es similar al descripto anteriormente, con la diferencia que primero los datos se simulan a partir del modelo ajustado, y luego se realiza un gráfico Q-Q plot, pero en el eje Y se grafican los cuantiles obtenidos a partir de dichas simulaciones. También se representan las bandas de confianza al 95%. Un buen ajuste se refleja cuando los puntos se encuentran posicionados lo más aproximado a la línea de puntos $y = x$.

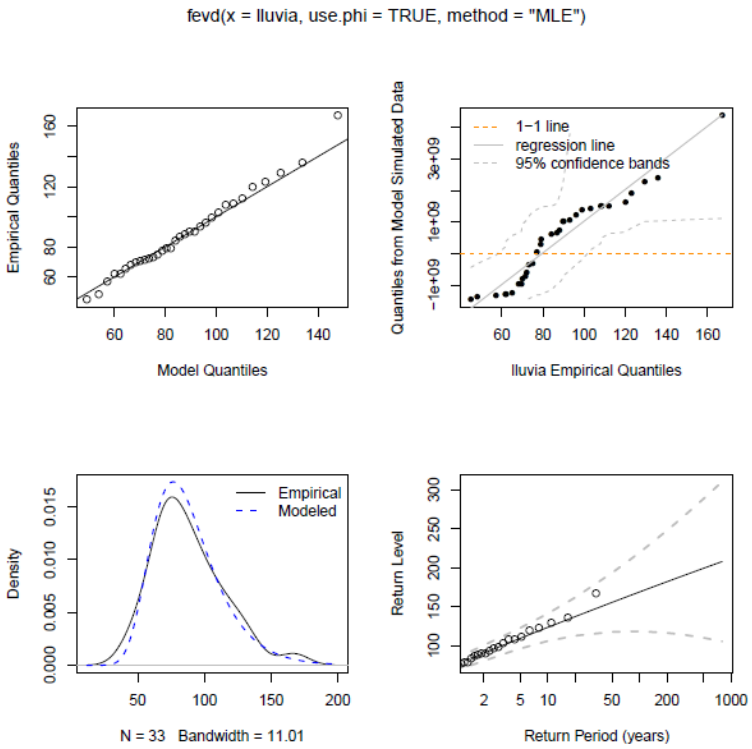
En el gráfico c) de 9.7 se visualiza la densidad empírica conjuntamente con la densidad del modelo. Puede notarse que ambas curvas son similares.

Por último, en el gráfico d) de 9.7 se visualizan los niveles de retorno según fue visto en sección (5.3.3). Como se puede observar, los puntos de dicho gráfico se acercan a la curva cóncava, es decir dan cuenta de que el el modelo que mejor ajusta a los datos es efectivamente del tipo de la familia Fréchet con los parámetros ($\hat{\mu} = 83.63, \hat{\sigma} = 3.10, \hat{\xi} = 0.12$).

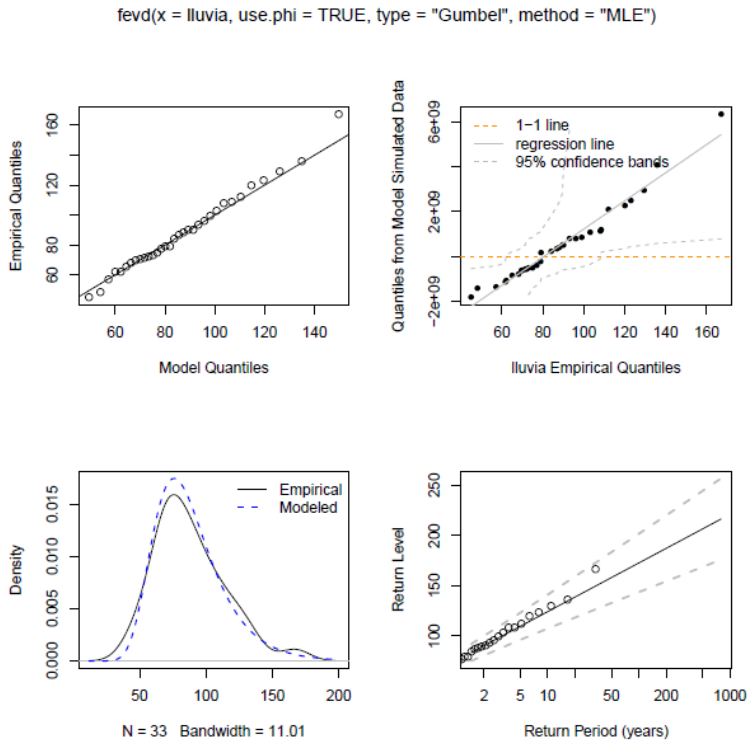
Como ejemplo de otros gráficos de diagnóstico se presenta el caso de la estación Aeropuerto de Carrasco. En dicha estación se vio que las primeras estimaciones de parámetros, según método MLE, arrojaron que los datos podrían ajustarse según un modelo GEV de la familia Weibull (dado que $\hat{\xi} = -0.02$), sin embargo, luego de realizado el testo de hipótesis Cramér-Mises recortado no rechaza la hipótesis de que las lluvias extremas anuales de dicha estación se ajustan mejor según un modelo GEV de la familia Gumbel.

En el gráfico 9.8 se compara las salidas de los gráficos de diagnóstico de dicha estación. El primero refleja la salida del ajuste realizado según un modelo Weibull y el segundo según un ajuste a los datos de un modelo Gumbel.

Como se ve en el gráfico 9.8, la mejora de la performance del modelo Gumbel a los datos, se ve reflejado por ejemplo, en el gráfico Q-Q plot 2 del gráfico b) respecto al a) donde los puntos están más sobre la línea uno a uno.



(a) Gráfico diagnóstico según ajuste Weibull



(b) Gráfico diagnóstico según ajuste Gumbel

Figura 9.8: Comparación de gráficos de diagnóstico de ajuste Weibull vs ajuste Gumbel

9.5 Niveles de retorno estimados

En esta sección se presentan los niveles de retorno estimados de las lluvias extremas anuales, para cada estación, a partir de los modelos ajustados anteriormente. Para cada caso, se obtuvieron los niveles de retorno para 20, 50, 100 y 200 años, es decir, se estiman para cada localidad, aquellos valores de precipitaciones extremas anuales diarias que se espera sea excedido una vez cada 20, 50, 100 o 200 años respectivamente. Dicho de otra manera, tal como se vió en 5.3.3, los niveles de retorno son el percentil 95%, 98%, 99% y 99.5% de la distribución GEV ajustada, respectivamente.

Como ejemplo, se presenta a continuación en la figura 9.9, los resultados específicos obtenidos para Rocha. En ella se visualizan las precipitaciones diarias de dicha localidad, resaltando los valores máximos anuales y los niveles de retorno estimados:

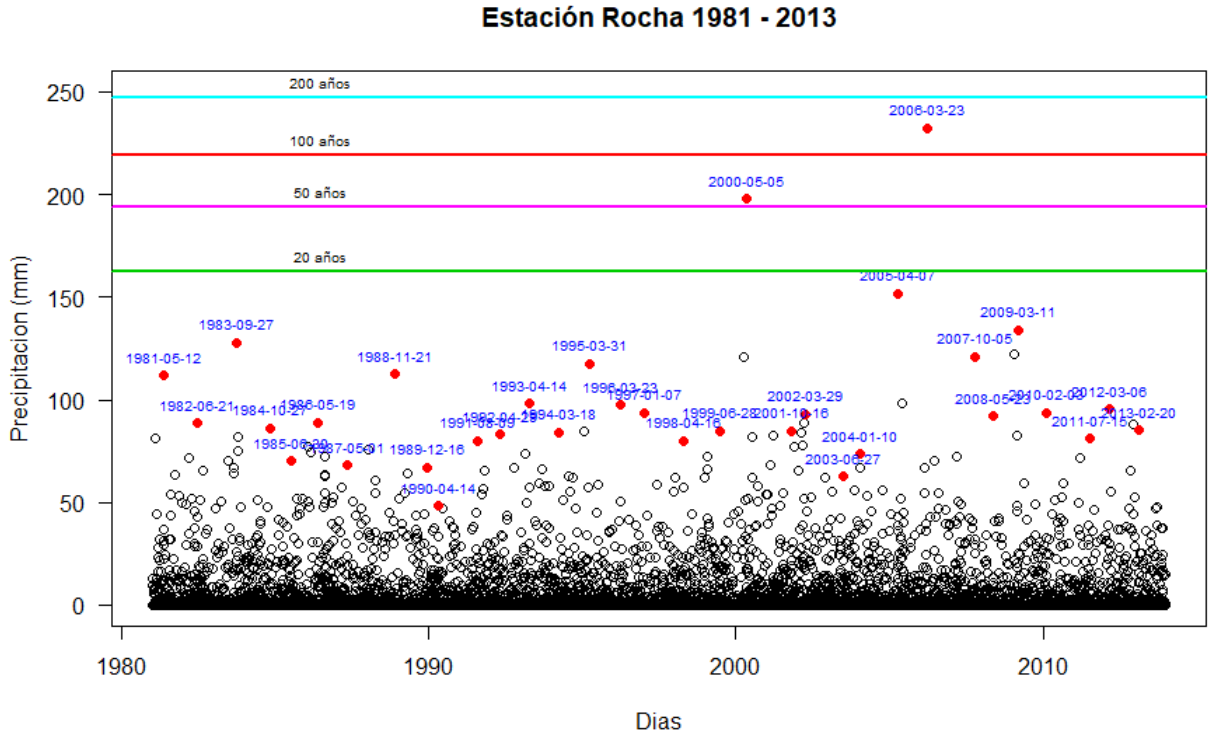


Figura 9.9: Niveles de retorno estimados para la localidad de Rocha. Fuente: elaboración propia en R.

Del gráfico anterior se desprende que los valores 162.52 mm, 193.61 mm, 219.28 mm y 247.07 mm resultaron ser los niveles de retorno estimados correspondientes a 20, 50, 100 o 200 años respectivamente, es decir, dichos valores resultan ser el percentil 95%, 98%, 99% y 99.5% de la distribución Fréchet ajustada. Resulta interesante notar que las lluvias máximas anuales ocurridas en los años 2000 y 2006 estuvieron por encima de los niveles de retorno asociados a períodos de retorno de 50 y 100 años respectivamente. Es decir, los valores de lluvias extremas diarias ocurridas esos años son valores que se espera ocurran 1 vez cada 50 o 100 años como mínimo.

En la tabla a continuación, se presentan las estimaciones de los niveles de retorno para cada estación, para los períodos de 20, 50, 100 y 200 años, ordenados de forma creciente según el nivel de retorno estimado a 200 años:

	Estación	20 años	50 años	100 años	200 años
13	Punta del Este	131.86	151.20	165.68	180.12
1	Aeropuerto Carrasco	138.38	157.96	172.64	187.26
6	Durazno	144.63	162.81	176.44	190.02
10	Paso de los Toros	145.77	164.48	178.50	192.47
7	Melo	147.11	166.48	181.00	195.46
14	Rivera	153.23	172.49	186.93	201.31
11	Paysandú	156.16	177.77	193.95	210.08
12	Prado	155.73	179.49	197.30	215.04
20	Young	161.89	184.58	201.59	218.53
19	Trinidad	163.76	188.37	206.82	225.19
2	Aeropuerto Melilla	161.80	187.44	206.65	225.80
9	Palmitas	167.00	193.23	212.88	232.47
5	Colonia	169.22	196.19	216.41	236.55
4	Bella Unión	176.49	200.96	219.29	237.56
17	Tacuarembó	179.68	205.25	224.41	243.49
16	Salto	177.77	204.30	224.18	243.98
15	Rocha	162.52	193.61	219.28	247.07
18	Treinta y Tres	186.31	215.55	237.47	259.30
3	Artigas	214.39	249.23	275.34	301.36
8	Mercedes	190.18	262.99	337.76	435.48

Tabla 9.5: Valores de retorno estimados.Fuente: elaboración propia en R.

Resulta curioso, notar que Mercedes, si bien posee un nivel de retorno menor a 20 años respecto de Artigas por ejemplo, resultó ser la localidad cuyos percentiles 98%, 99% y 99.5% de la distribución GEV ajustada fueron mayores que el resto. Luego le sigue Artigas y Treinta y Tres. Rocha queda en el cuarto puesto si se ordena en función al percentil 99.5%. Tal como se preveía, las distribuciones Fréchet son distribuciones de colas más pesadas que las distribuciones Gumbel. Por otra parte, las distribuciones con menores valores de retorno estimados resultaron ser Punta del Este y Aeropuerto de Carrasco.

10 Análisis de clusters

Como complemento a las modelizaciones obtenidas en sección anterior, se realizó una primer exploración a la existencia de patrones espaciales respecto de las lluvias extremas anuales diarias. Como se mencionó en sección (2), uno de las investigaciones que motivó el presente trabajo de tesis, fue el de [3], en el que se utilizaron métodos de clustering para explorar la conformación de zonas homogéneas de precipitaciones extremas en Francia. Dicho trabajo incorpora al análisis, la información sobre la dependencia espacial entre puntos cercanos a través del F-madograma.

En una primer etapa, se realizó análisis de clustering a las 20 estaciones meteorológicas en función de las estimaciones de los parámetros de las distribuciones GEV marginales correspondientes, obtenidos en la sección anterior 9.1.

En una segunda etapa, se exploró la conformación de grupos de estaciones meteorológicas en función de las precipitaciones máximas anuales diarias ocurridas en cada año (desde 1981 a 2013). Esta etapa se alinea con la idea propuesta en [3], es decir, se incorpora además la información de la dependencia espacial de las precipitaciones máximas a través de la utilización del F-madograma como distancia.

10.1 Clusters de parámetros estimados GEV

Para llevar a cabo esta etapa, se partió de los parámetros estimados de las distribuciones GEV marginales ajustadas, de ubicación, de escala y de forma, luego de aplicar el test de bondad de ajuste a las distribuciones GEV de cada estación. Es decir, para los casos en que el test test Cramér-von Mises recortado no rechazó la hipótesis nula de que la distribución sea Gumbel, se consideró el parámetro de forma igual a 0.

En los siguientes gráficos se presentan las estaciones meteorológicas coloreadas según los valores de las estimaciones de los parámetros de las distribuciones GEV marginales ajustadas, obtenidos según método MLE según Tabla 9.4:

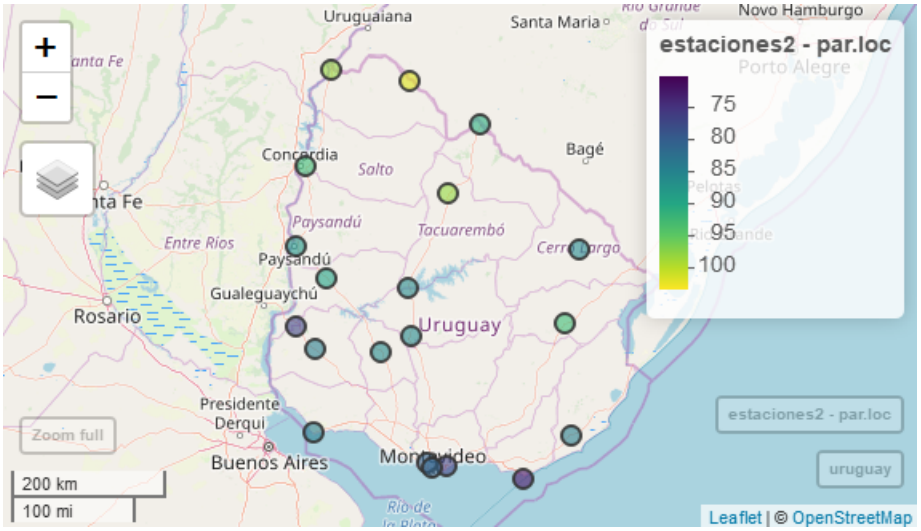


Figura 10.1: Según parámetro de ubicación. Fuente: elaboración propia en R.

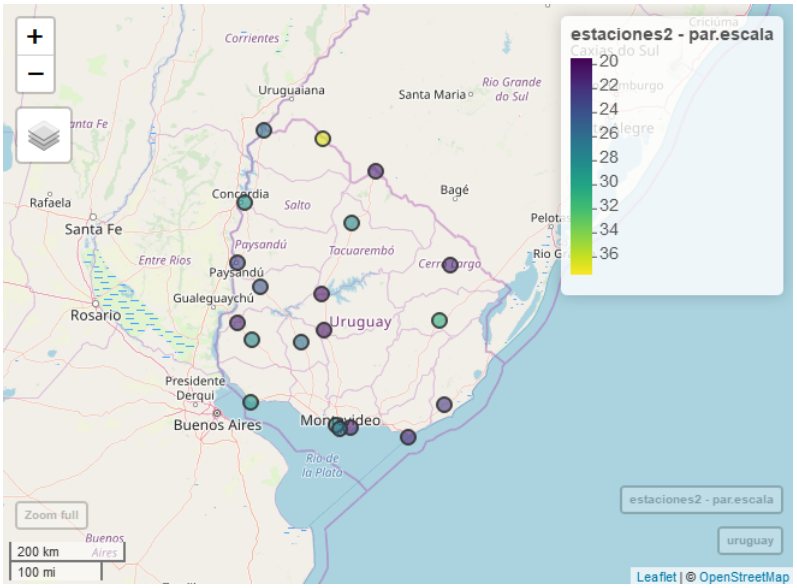


Figura 10.2: Según parámetro de escala. Fuente: elaboración propia en R.

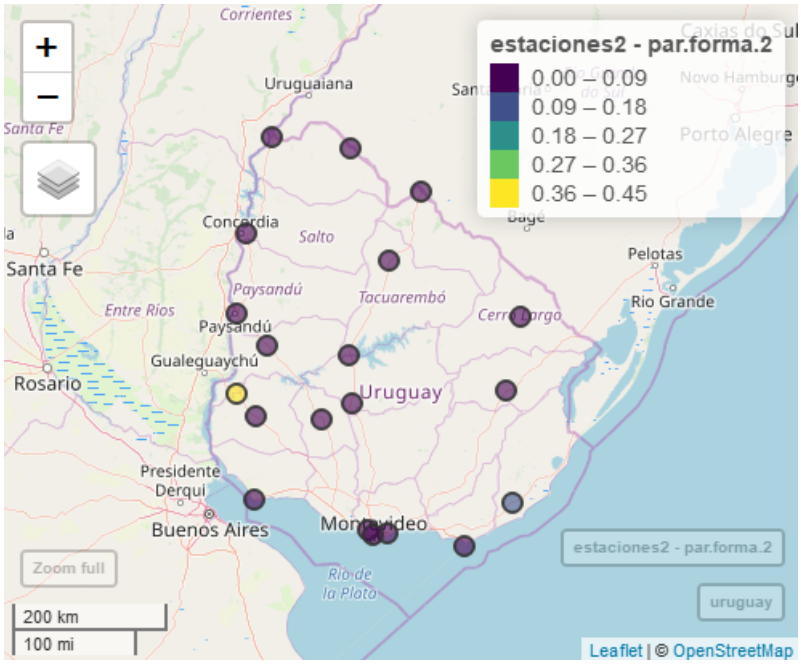


Figura 10.3: Según parámetro de forma.Fuente: elaboración propia en R.

A priori se podría agrupar a las estaciones en dos grupos, aquellas con distribución Gumbel (parámetro de forma cercano a 0) y aquellas estaciones con parámetro de forma positivo. Sin embargo, la cantidad de grupos no es trivial al incluir al análisis las variables parámetro de ubicación y escala.

Por un lado se aplicó la metodología Ward con distancia euclídea la que se comparó con los resultados de aplicar la metodología PAM utilizando el mismo tipo de distancia.

Tal como se vió en 6.1, para analizar la cantidad de grupos óptimos en un modelo jerárquico, se estudia el dendrograma así como los indicadores R^2 y $Pseudo_F$. Si se toma $K = 2$, el $R^2 = 99,96\%$ mientras que si $K = 3$ entonces $R^2 = 99,92\%$. Por otra parte, el indicador $Pseudo_F$ presenta un máximo en $K = 2$. A continuación se presentan los gráficos de dendrograma, agrupando las estaciones según dos (gráfico 10.4) o tres (gráfico 10.5) grupos:

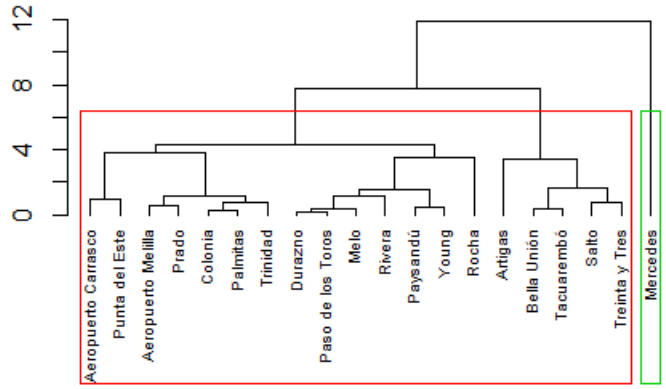


Figura 10.4: Dendrograma método Ward con distancia euclídea y $K = 2$.

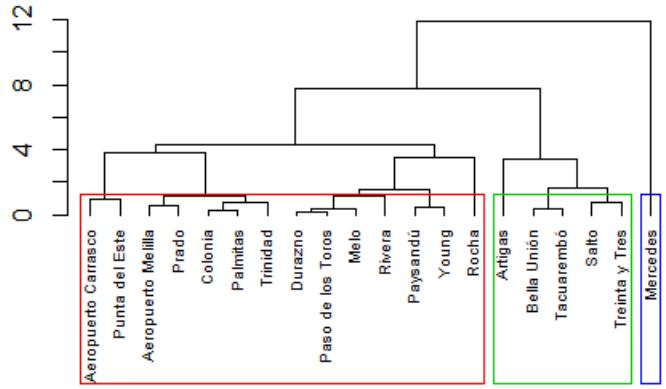


Figura 10.5: Dendrograma método Ward con distancia euclídea y $K = 3$.

De los resultados obtenidos mediante el método Ward con distancia euclídea tomando como variables los parámetros estimados, se puede observar que el R^2 y el $Pseudo_F$ sugieren en que la cantidad de grupos óptima es $K = 2$. Observando el dendrograma 10.4, se puede observar que dicho método agrupó todas las estaciones del país que rechazaron el test de bondad de ajuste de distribución Gumbel conjuntamente con Rocha. Mercedes por otra parte, quedó formando un único grupo. Sin embargo como se ve en el gráfico 10.5, si se

toma $K = 3$, el primer grupo se separa en dos, por un lado se agrupan las estaciones Artigas, Bella Unión, Tacuarembó, Salto y Treinta y Tres quedando las estaciones de más al sur y sur este agrupadas en otro cluster y el tercer grupo conformado por Mercedes.

Los resultados anteriores se comparan con los obtenidos mediante la aplicación del método de clustering PAM utilizando el mismo tipo de distancia. Como se vió en 6.1, éste método requiere que la cantidad de grupos sea definida previamente. Para ello, se calculó los valores silhouette variando la cantidad de grupos desde $K = 2$ a $K = 10$. Los resultados se presentan en forma de boxplot en el siguiente gráfico:

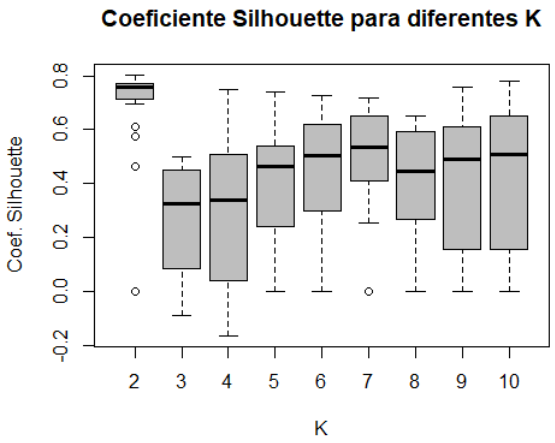
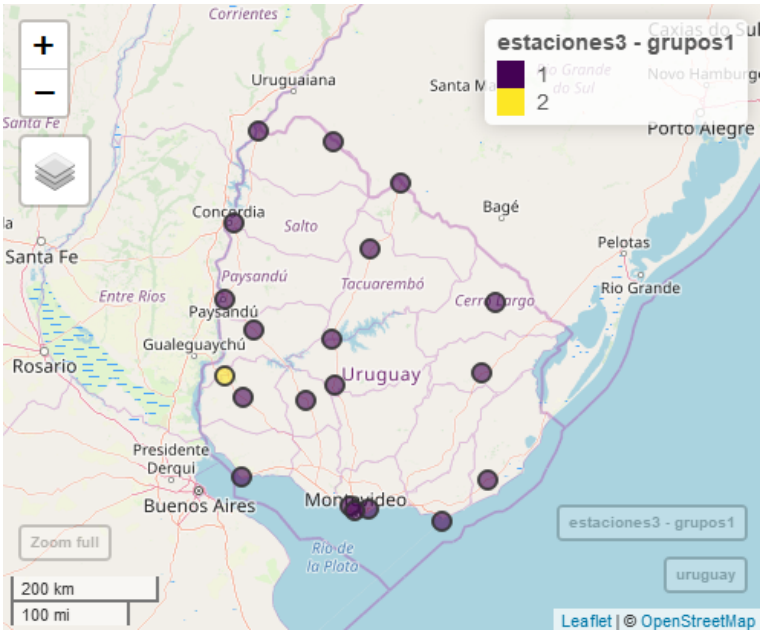


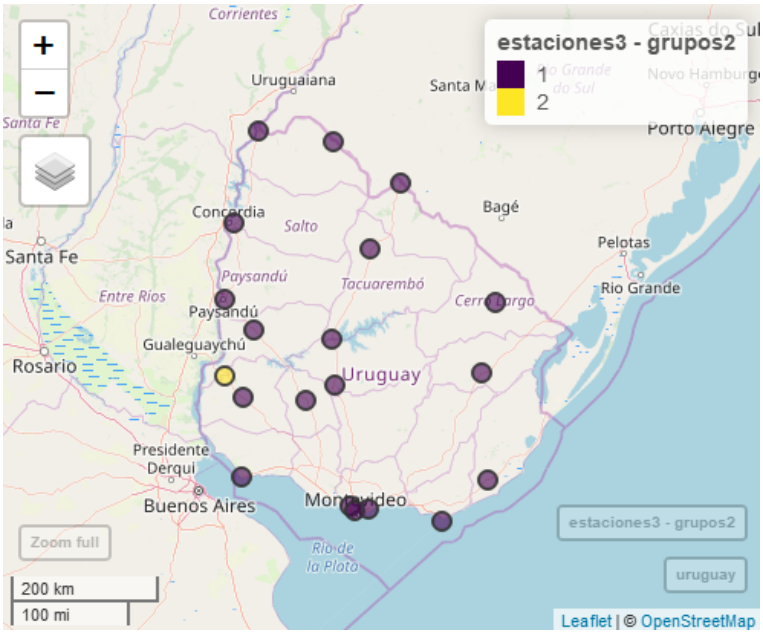
Figura 10.6: Valores silhouette variando K.

El gráfico anterior, sugiere que la cantidad óptima de grupos para inicializar el algoritmo en cuestión es $K = 2$. En este caso el coeficiente silhouette asciende a $SC = 0,6924$ indicando que la estructura de grupos encontrada para este caso es razonable. Para $K = 3$ el coeficiente silhouette desciende a 0,2813 indicando que en ese caso la estructura de grupos encontrada podría ser débil o incluso artificial. La conformación de grupos luego de aplicar PAM con $K = 2$ quedó igual a la del método Ward vista anteriormente.

En los gráficos a continuación se comparan los resultados obtenidos mediante Ward y PAM con $K = 2$ y $K = 3$.



(a) Método Ward con $K = 2$.



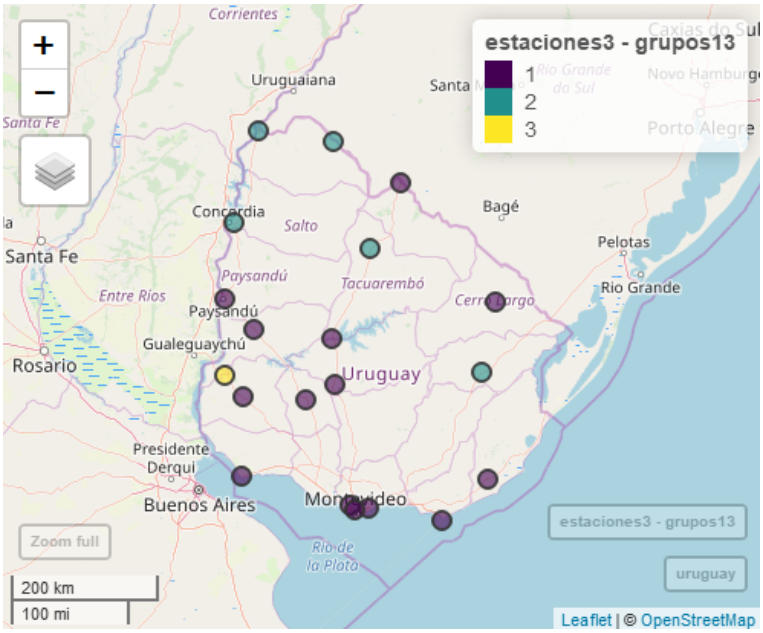
(b) Método PAM con $K = 2$.

Figura 10.7: Comparación de métodos clustering con $K = 2$.

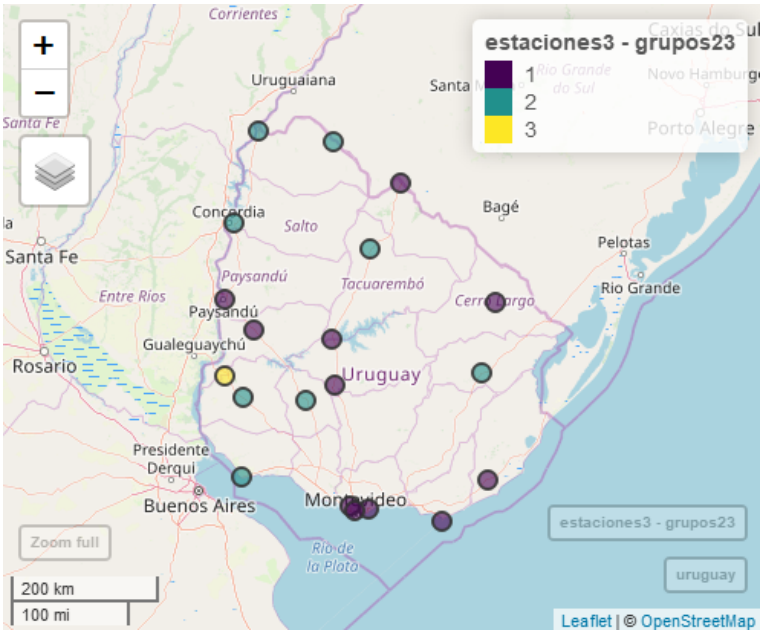
Los gráficos 10.7(a) y 10.7(b) arrojan resultados idénticos. La Figura 10.7(a) muestra espacialmente lo mismo que se visualizó a partir del dendrograma 10.4. Todas las estaciones que conforman el grupo 1, fueron clasificadas con distribución Gumbel según sección anterior salvo por la estación de Rocha que no rechazó el test de bondad de ajuste a la Fréchet. Los valores promedios de los parámetros estimados de las estaciones que se encuentran incluídas en este cluster son 87.10 mm para el parámetro de locación, 25.26 mm para el parámetro de escala. También se destaca que en general, las estaciones agrupadas bajo este cluster son las que tienen mayores valores de niveles de retorno de lluvias extremas a 200

años, destacándose Artigas y Treinta y Tres que están en el segundo y tercer puesto respectivamente. El segundo grupo, está conformado por Mercedes cuyos valores de parámetros de estimados de locación, escala y forma fueron 76.09, 20.39 y 0.39 respectivamente. Cabe destacar que Mercedes resultó ser la estación con mayor valor de nivel de retorno a 200 años, es decir resultó ser la distribución de cola más pesada.

A continuación se presentan los resultados de ambas metodologías tomando $K = 3$:



(a) Método Ward con $K = 3$.



(b) Método PAM con $K = 3$.

Figura 10.8: Comparación de métodos clustering con $K = 3$.

Los resultados de ambas metodologías para 3 grupos tienen algunas coincidencias. En el grupo 2 de ambos métodos, se encuentran la mayoría de las estaciones más al norte del Río Negro (Artigas, Bella Unión, Salto y Tacuarembó) incluyendo también a las estaciones Treinta y Tres, aunque el método PAM en dicho grupo también agrupa a las estaciones de Colonia, Trinidad y Palmitas que se encuentran diametralmente opuestas a las mencionadas anteriormente. Otra coincidencia de ambas metodologías es que Mercedes queda como única estación del grupo 3. En el método Ward el grupo 1 queda conformado en su mayoría por estaciones del sur del Río Negro, así como también por Young, Paysandú, Rivera y Cerro Largo. En ambos métodos Rivera y Cerro Largo quedaron en grupos distintos a la de las demás estaciones del norte del Río Negro, cuestión que llamó la atención.

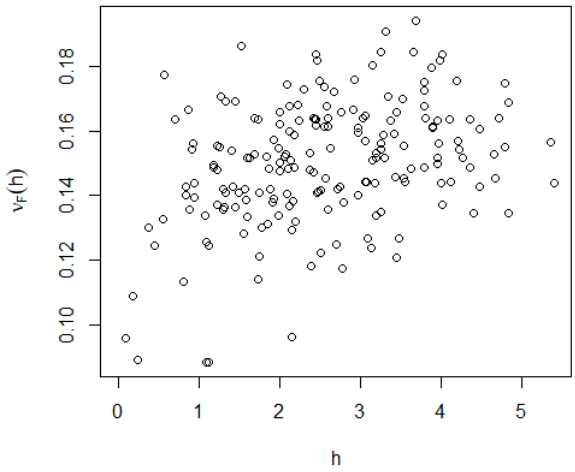
En suma, luego de comparar ambas metodologías para agrupar las estaciones o localidades en función a los parámetros estimados de las distribuciones GEV marginales, puede decirse que los resultados obtenidos a través del método Ward resultan ser más razonables espacialmente que los obtenidos por el método PAM. De hecho, bajo el método PAM las estructuras de grupos encontradas son débiles o incluso artificales. A diferencia de PAM, bajo el método Ward hay evidencia de que $K = 2$ resulta ser la cantidad de grupos óptima.

10.2 Clusters de máximos anuales

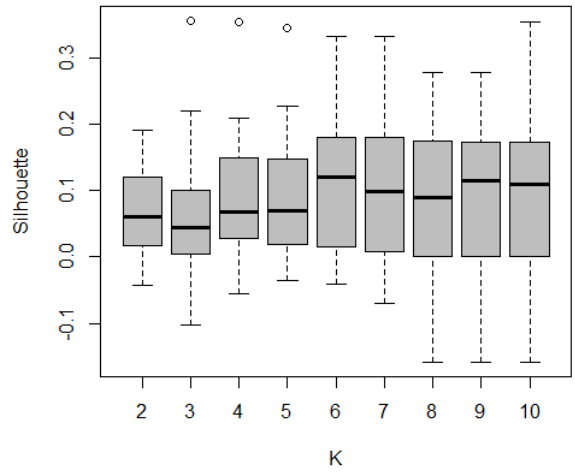
En esta sección es donde el F-madograma comienza a jugar un papel preponderante en el análisis de cluster. Tal como fue presentado en sección (5.61), dicha función será utilizado como insumo para realizar cluster PAM utilizando como base de datos, la matriz que contiene 20 filas correspondientes a las 20 localidades de estudio y tantas columnas como años del período de estudio (1981 a 2013). Es decir, se estudiará si existen patrones espaciales en función a las lluvias extremas anuales diarias, es decir se toma como variables los registros de las mismas en los diferentes años del período de estudio, es decir desde 1981 a 2013. Es decir, se realizará una agrupación a partir de 33 variables.

El F-madograma, fue utilizado como función distancia. Se trabajó con los datos de lluvias extremas anuales en logaritmo.

A continuación se muestra el F-madograma calculado a partir de los datos de precipitaciones máximas en logaritmo, convertidos previamente a Fréchet unitaria:

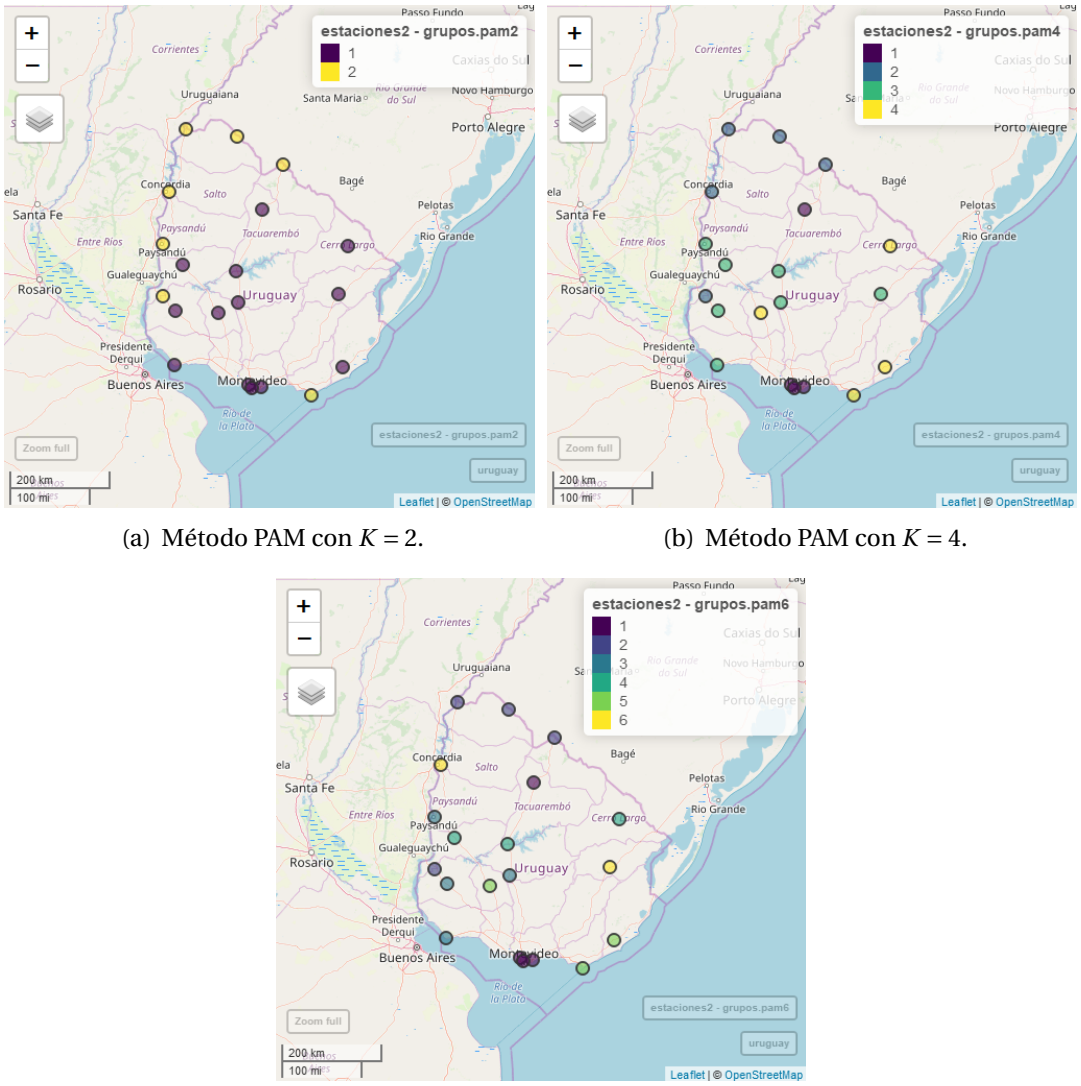


Al igual que en el caso del análisis de cluster en base a los parámetros estimados, al realizar un análisis en base a las precipitaciones extremas en logaritmo mediante metodología PAM, se estudió la cantidad óptima de clusters a determinar a partir de los valores silhouette arrojados variando K:



Observando el gráfico anterior, $K = 2$, $K = 4$ ó $K = 6$ podrían ser candidatos a cantidad de grupos para inicializar el algoritmo.

A partir de lo anterior, se muestran los resultados obtenidos para $K = 2$ y $K = 4$:



(a) Método PAM con $K = 2$.

(b) Método PAM con $K = 4$.

(c) Método PAM con $K = 6$.

Figura 10.9: Comparación de métodos clustering PAM con distancia F-Madograma.

La estructura de grupos más clara está dada en el gráfico 10.9(a), donde se visualiza que las estaciones del norte y litoral oeste del país quedaron agrupadas en un mismo cluster, y por otro lado las estaciones del centro y sur del país en un segundo cluster. Con la salvedad de Rocha que a pesar de la lejanía con el norte del país quedó en el mismo cluster que las primeras estaciones mencionadas. A diferencia de lo visto en las agrupaciones obtenidas a partir de los parámetros GEV estimados, aquí las estaciones Rocha y Mercedes quedaron en distintos cluster a pesar de pertenecer ambas a misma familia de distribución GEV.

Luego entre los gráficos 10.9(b) y 10.9(c) se puede ver que pese a sus diferencias, hay estaciones que en ambos casos quedan agrupadas en un mismo cluster, como ser Artigas, Bella Unión y Rivera así como las estaciones de Aeropuerto de Carrasco, Melilla y Prado.

En el caso de $K = 6$, se observa que estaciones cercanas, en el norte del Río Negro resultaron agrupadas en distintos cluster, destacando un grupo con las estaciones de Artigas,

Bella Unión y Rivera (pertenecientes al grupo 2 junto con Mercedes), otro grupo conformado por Young, Paso de los Toros (grupo 4 junto con Treinta y Tres). Por otro lado Tacuarembó quedó en el mismo grupo que Aeropuerto de Carrasco, Aeropuerto de Melilla y Prado (grupo 1). Paysandú y Melo pertenecen al grupo 6. Luego, respecto de las demás estaciones al sur del Río Negro, se puede observar que quedó un grupo conformado por Colonia, Palmitas y Durazno (grupo 3) y Trinidad, Punta del Este y Rocha quedaron agrupados en el grupo 5. Cabe destacar que para $K = 6$ se obtuvo el mayor valor del $SC = 0.11$.

De todas formas, el coeficiente silhouette SC no asciende demasiado para los distintos K que se probaron, dicho valor resulta en en el entorno de 0.1 para todos los K cuestión que daría la pauta que no hay una clara estructura de grupos en los datos o el algoritmo no está pudiendo captarla.

En suma, el SC para todos los K que se probaron no arroja un valor tal que refleje estructuras claras de grupos.

11 Test de independencia

Como complemento a al análisis de clustering visto anteriormente, y para evaluar algunos de los resultados respecto a las agrupaciones encontradas, se aplicó el test de independencia tal como se vió en 6.5. En este caso, la idea es utilizar el test de independencia basado en porcentajes de recurrencias en la cual consideraremos X = máximo anual en la estación X e Y = máximo anual en la estación Y. Por lo tanto en este caso los espacios métricos a considerar son $S_X = S_Y = \mathbb{R}$ y es razonable asumir que los valores máximos anuales son independientes de un año a otro, por lo que se tendría una muestra $(X_1, Y_1), (X_2, Y_2), ..., (X_{33}, Y_{33})$.

Dicho test se aplicó a todos los pares de estaciones de estudio. A continuación se muestran los p-valores obtenidos a partir de $m = 1000$ simulaciones realizado para cada test:

	Carr.	Mel.	Art.	B.U	Col.	Dur.	Melo	Mer.	Pal.	P.Toros	Pay.	Prado	P.Este	Riv.	Rocha	Salto	Tac.	T.Tres	Tri.
Mel.	0.00																		
Art.	0.24	0.01																	
B.U	0.11	0.45	0.55																
Col.	0.47	0.98	0.80	0.36															
Dur.	0.43	0.16	0.08	0.51	0.14														
Melo	0.11	0.84	0.20	0.86	0.51	0.07													
Mer.	0.61	0.96	0.31	0.18	0.58	0.05	0.23												
Pal.	0.91	0.24	0.13	0.28	0.04	0.17	0.68	0.61											
P.Toros	0.58	0.74	0.52	0.89	0.25	0.10	0.51	0.13	0.23										
Pay.	0.53	0.83	0.25	0.83	0.98	0.11	0.76	0.79	0.21	0.74									
Prado	0.00	0.00	0.02	0.77	0.06	0.12	0.59	0.67	0.05	0.46	0.28								
P.Este	0.72	0.10	0.25	0.41	0.53	0.09	0.71	0.52	0.16	0.24	0.24	0.76							
Riv.	0.57	0.64	0.00	0.00	0.24	0.18	0.11	0.15	0.18	0.46	0.29	0.24	0.56						
Rocha	0.42	0.11	0.80	0.41	0.15	0.21	0.67	0.69	0.06	0.03	0.06	0.54	0.02	0.84					
Salto	0.16	0.78	0.96	0.85	0.45	0.13	0.51	0.69	0.12	0.43	0.30	0.74	0.66	0.38	0.63				
Tac.	0.28	0.70	0.85	0.62	0.51	0.17	0.45	0.71	0.22	0.60	0.06	0.43	0.12	0.34	0.74	0.93			
T.Tres	0.41	0.49	1.00	0.08	0.09	0.15	0.39	0.05	0.20	0.29	0.19	0.09	0.71	0.91	0.64	0.33	0.70		
Tri.	0.71	0.96	0.22	0.93	0.58	0.04	0.17	0.19	0.04	0.01	0.23	0.43	0.56	0.72	0.01	0.06	0.68	0.16	
Young	0.09	0.04	0.19	0.39	0.05	0.20	0.36	0.55	0.07	0.00	0.14	0.18	0.09	0.88	0.81	0.43	0.72	0.25	0.01

Figura 11.1: Test de independencia de lluvias extremas anuales diarias.

De la tabla anterior 11.1, se observa que las estaciones meteorológicas ubicadas en zona metropolitana (Aeropuerto de Carrasco, Melilla y Prado) rechazan la hipótesis nula de que

las lluvias extremas anuales diarias son independientes. Esto reafirma los resultados de los métodos de clustering aplicados ya que agrupan a dichas estaciones en un mismo cluster (no así bajo el método PAM agrupando según estimaciones de parámetros GEV con distancia euclídea según se puede observar en 10.7(b) y 10.8(b)).

Respecto de las estaciones meteorológicas del norte del Río Negro, se observa en 11.1 que los pares de estaciones que resultaron rechazar la hipótesis nula de independencia fueron: Artigas - Rivera, Bella Unión - Rivera y Paso de los Toros - Young. Estos resultados no resultan en línea con los obtenidos en 10.7(a), donde la estación de Rivera quedó excluído del grupo al que pertenecen Artigas y Bella Unión, aunque sí refleja la dependencia entre Paso de los Toros y Young que pertenecen a un mismo grupo. Sin embargo, los resultados obtenidos en 10.9(c) se condicen con los resultados del test de independencia, ya que las estaciones Artigas, Bella Unión y Rivera pertenecen a un mismo grupo y Paso de los Toros y Young pertenecen a otro. De hecho, otros resultados obtenidos en 10.9(c) son favorables a los obtenidos bajo el test de independencia, esto son: Colonia - Palmitas (grupo 3), Punta del Este - Rocha y Rocha - Trinidad (grupo 5). A pesar que el *SC* para este caso reflejaba una estructura de grupos débil, las agrupaciones encontradas resultan coherentes con los obtenidos bajo el presente test de independencia.

Otros resultados obtenidos en 10.7(a), resultan coherentes con los resultados de dicho test entre las estaciones Durazno - Trinidad, Mercedes - Treinta y Tres, Palmitas - Trinidad, Paso de los Toros - Trinidad y Trinidad - Young (todos pertenecientes al grupo 1 de dicho método).

Los resultados de 6.5 respecto de Artigas - Prado y Artigas - Aeropuerto Melilla se reflejan en los resultados obtenidos en 10.7(b), sin embargo para ese caso, el coeficiente silohouette refleja que el algoritmo no logró capturar una estructura clara de grupos.

En suma, de la aplicación del test puede decirse que se refuerzan los resultados obtenidos tanto para el método Ward aplicado en base a los parámetros estimados GEV con distancia euclideana que se reflejan en el gráfico 10.7(a), así como para el resultado obtenido mediante la aplicación del método PAM es basado en las precipitaciones máximas anuales diarias de los años 1981 a 2013 reflejado en gráfico 10.9(c).

12 Procesos max-estables

En esta sección se realizará un ajuste espacial a los datos, a través de procesos máx estables tal como se vió en sección 5.1. Para ello, primero se transforman las distribuciones GEV marginales a Fréchet unitarias según las transformaciones vistas en 5.1. Es decir, las observaciones de cada estación serán transformadas utilizando los parámetros estimados de ubicación, escala y forma en 9.1.

Tal como se vió en gráfico del F-madograma en figura 5.61, se observa gran variabilidad espacial respecto de las lluvias extremas anuales aún en estaciones más cercanas.

Se ajustan diversos procesos máx estables. A partir de GEV marginales dadas, se estiman los parámetros de dependencia. El proceso max-estable óptimo se eligirá en función al criterio

AIC, es decir, el modelo que tenga menor AIC será el elegido. A continuación se presentan los coeficientes extremales estimados según diversos procesos max estables ajustados:

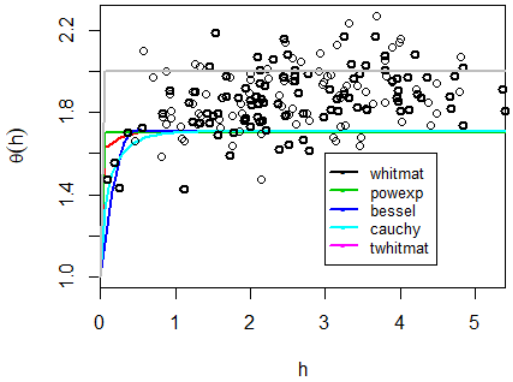


Figura 12.1: Estimaciones de la función coeficiente extremal según los distintos modelos max-estables ajustados.

A partir del criterio AIC, el modelo que mejor ajusta es el modelo de Schlather con función de covarianza del tipo Power Exponential. En la siguiente tabla se presentan los valores obtenidos para cada uno de los modelos ajustados:

	Parámetros de dependencia			
	Nugget	Range	Smooth	AIC
Schlather (Whittle-Matern)	0,78	0,08	3,50	148.501,50
Schlather (Power Exponential)	0,99	686,96	2,00	146.614,50
Schlather(Bessel)	0,01	0,06	4,12	148.667,20
Schlather(Cauchy)	1,37E-03	7,42E+10	1,78E+11	148.504,80
t-extremal (Whittle-Matern)	1,00	940,44	2,08	146.894,20

Figura 12.2: Estimación de los parámetros de los procesos máx-estables y AIC.

Part IV
Conclusiones

En Uruguay la aplicabilidad de la estadística espacial y la teoría de valores extremos a diversas áreas del conocimiento está en desarrollo. El proyecto de tesis planteado, abarcó una temática de actual importancia como lo es el estudio y caracterización espacial de los fenómenos meteorológicos extremos (en particular las precipitaciones extremas).

La teoría de valores extremos brindó las herramientas necesarias para llevar a cabo el primer objetivo planteado. Se estimaron las distribuciones asintóticas GEV a cada estación complementando. Los resultados obtenidos arrojaron que las distribuciones de las lluvias máximas anuales diarias del Uruguay, para las distintas localidades estudiadas, pueden modelarse según modelo Gumbel o Fréchet. Dichos resultados fueron validados al aplicar un test de bondad de ajuste no paramétrico del tipo Cramér-von Mises recortado. Estos resultados fueron validados también al comparar los gráficos diagnósticos, según Figura 9.8, en donde se observa que la distribución Gumbel ajusta mejor a los datos. Se obtuvieron niveles de retorno de lluvias extremas anuales para diferentes períodos de retorno para cada localidad, información que puede resultar de mucha utilidad tanto para la ingeniería como para otras actividades agrícolas, actividad aseguradora, y otros.

Respecto del segundo objetivo planteado, se estudiaron dos escenarios, por un lado la conformación de grupos tanto a partir de los parámetros de las familias GEV estimados según el objetivo anterior utilizando la distancia euclideana, y por otro la conformación de grupos tomando como variables las lluvias extremas anuales desde 1981 a 2013 en cada localidad utilizando la distancia definida a través del F-madograma.

Según el primer estudio, tomando como variables los parámetros de las distribuciones GEV estimados para cada estación meteorológica, se destaca que los resultados más contundentes fueron los obtenidos a partir de tomar como parámetro de forma igual a 0 para aquellas estaciones en que no se rechazó la hipótesis nula de que la distribución del fenómeno en estudio es del tipo Gumbel. Al realizar análisis de cluster tomando como parámetros estimados los obtenidos en una primer instancia según Tabla 9.1, los valores del coeficiente silhouette resultaron bajos, indicando que no existía una estructura de grupos clara.

Los resultados de la aplicación de las metodologías de clustering, confirman que a gran escala, el fenómeno de lluvias extremas anuales es bastante homogéneo en el país. Dichos resultados lo confirman los dos escenarios llevados a cabo. En el primero, se obtuvieron dos grupos, aunque uno de ellos estaba representado únicamente por Mercedes, que dada la opinión de expertos en el tema, Mercedes tiene un comportamiento atípico del resto de las localidades estudiadas. En este contexto, se realizó el ejercicio de eliminar a Mercedes del análisis, y se vuelve a repetir la conformación de dos grupos, quedando Rocha en un único grupo, por lo que el resto de las estaciones siguen teniendo un comporamiento homogéneo entre sí.

Bajo el segundo escenario, aplicando el método PAM con tantas variables como años en es-

tudio (1981 a 2013) para cada localidad, y utilizando como distancia el F-madograma, según el coeficiente silhouette no se encontró una estructura clara de grupos en los datos o el algoritmo no pudo captarla.

Sin embargo, se pueden destacar resultados localizados, como por ejemplo, que para $K = 2$ las estaciones más al norte del Río Negro y litoral del país, como ser Artigas, Bella Unión, Rivera, Salto y Paysandú quedan en un mismo grupo según ambos escenarios manejados. También algunas estaciones del centro y sur del país se muestran mayor similaridad bajo ambos escenarios (Young, Paso de los Toros, Durazno, Trinidad, Palmitas, Colonia, Prado, Aeropuerto de Melilla, Aeropuesto de Carrasco, Tacuarembó, Melo, Treinta y Tres y Rocha).

Como complemento a los análisis de clustering vistos anteriormente, se aplicó un test de independencia en base a ratios de recurrencia. De dicho test se puede destacar resultados tales como el de que la estación Mercedes rechazó la hipótesis de independencia con todas las localidades en estudio salvo con Treinta y Tres. También se visualiza que Rivera no rechaza la hipótesis nula de independencia ni con Artigas ni con Bella Unión. También las estaciones de la zona metropolitana (Aeropuerto de Melilla, Prado y Aeropuesto de Carrasco) no rechazan la hipótesis nula de independencia. Por otra parte, Rocha y Punta del Este no rechazan la hipótesis nula de independencia tampoco.

En suma, si se tomara como criterio de agrupación los resultados del test de independencia anteriormente descripto, se puede visualizar que varios de los resultados obtenidos mediante los anteriores métodos de clustering son reforzados mediante los resultados arrojados por dicho test.

Estudios similares podrían realizarse tomando bloques temporales equivalentes a trimestres o semestres, en lugar de bloques anuales. De esta forma podría estudiarse más en detalle la posible estacionalidad de dicho fenómeno. También sería de interés profundizar en la modelización según procesos max-estables mediante distintos modelos con distintas funciones de covarianzas o incluso modelando la dependencia a partir de cópulas extremas.

Part V
Apéndice A

Definición de matriz de información de Fisher

Sea una función densidad $f(x,\theta)$ que satisface las siguientes condiciones de regularidad:

- Existen las derivadas de primer y segundo orden de la función de verosimilitud.
- Además
$$E\left[\frac{\partial ln f(x;\theta)}{\partial \theta}\right] = 0 \quad \forall \theta.$$

El estimador máximo verosímil del parámetro θ es asintóticamente distribuído como

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta)).$$

Siendo $I(\theta)$ la matriz de información de Fisher evaluada en θ con

$$I(\theta) = -E\left[\frac{\partial^2 ln f(x;\theta)}{\partial \theta^2}\right].$$

12.1 Estimación de parámetros en el contexto multivariado

El primer paso a realizar es la estimación de los parámetros de las distribuciones marginales según se vio en (5.16). Una vez obtenidas dichas estimaciones, mediante las transformaciones mencionadas anteriormente se transforman las marginales GEV a Fréchet unitarias.

Estimar los parámetros del modelo de la familia multivariada seleccionada, tiene un costo computacional extremadamente alto. Un alternativa utilizada con frecuencia es la realización de inferencia mediante verosimilitudes compuestas, particularmente a través de la utilización de verosimilitudes a pares.

Dada una observación $\mathbf{z} \in \Re^k$, la log-verosimilitud compuesta a pares ponderada se define como sigue:

$$\ell(\psi; z) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij} log\left(f(z_i, z_j; \psi)\right), \tag{12.1}$$

siendo ψ el vector de parámetros del modelo a estimar, w_{ij} ponderadores que suelen utilizarse para hacer los cálculos más eficiente computacionalmente y $f(.,.;\psi)$ es la densidad bi-variada.

Se busca ψ que maximice la log-verosimilitud compuesta.

Las log-verosimilitudes compuestas vistas anteriormente son combinaciones lineales de log-verosimilitudes pero no son verdaderas log-verosimilitudes. Dicha estrategia, sin embargo,

permite obtener estimaciones insesgadas para ψ.

La estimación obtenida mediante el procedimiento anterior verifica la siguiente ley asintótica:

$$\sqrt{n}(\hat{\psi} - \psi_0) \to N(0, H^{-1}(\psi_0) J(\psi_0) H^{-1}(\psi_0)) \quad n \to \infty, \tag{12.2}$$

siendo $H(\psi_0) = -E(\nabla^2 \ell(\psi_0; \mathbf{Z}))$ y $J(\psi_0) = V(\nabla \ell(\psi_0; \mathbf{Z}))$.

Part VI
Apéndice B

Estimación de los parámetros GEV por método de los momentos ponderados

	Estación	Par. de posición	Par. de escala	Par. de forma
1	Aeropuerto Carrasco	75.56	21.36	0.01
2	Aeropuerto Melilla	81.13	28.52	-0.08
3	Artigas	103.32	39.96	-0.04
4	Bella Unión	97.67	26.96	0.03
5	Colonia	80.41	27.77	0.15
6	Durazno	86.78	21.07	-0.05
7	Melo	86.49	21.49	-0.12
8	Mercedes	78.13	24.54	0.17
9	Palmitas	84.78	29.16	-0.11
10	Paso de los Toros	84.66	19.94	0.10
11	Paysandú	88.04	24.87	-0.08
12	Prado	82.32	27.77	-0.19
13	Punta del Este	70.25	22.30	-0.04
14	Rivera	89.50	19.76	0.17
15	Rocha	81.99	18.81	0.27
16	Salto	89.05	25.16	0.25
17	Tacuarembó	99.53	30.22	-0.12
18	Treinta y Tres	90.92	31.77	0.09
19	Trinidad	87.91	29.05	-0.20
20	Young	89.97	24.60	-0.04

Estimación de los parámetros GEV por método de máxima verosimilitud perfil

	Estacion	Par. de posición	Par. de escala	Par. de forma
1	Aeropuerto Carrasco	76.18	21.14	-0.02
2	Aeropuerto Melilla	81.05	27.98	-0.07
3	Artigas	103.44	37.60	-0.02
4	Bella Unión	97.56	25.56	0.07
5	Colonia	80.32	26.34	0.20
6	Durazno	86.71	19.57	-0.01
7	Melo	87.05	21.46	-0.15
8	Mercedes	76.08	20.37	0.39
9	Palmitas	85.74	29.17	-0.16
10	Paso de los Toros	84.55	18.74	0.15
11	Paysandú	88.22	23.83	-0.08
12	Prado	82.82	26.77	-0.20
13	Punta del Este	70.19	20.72	0.00
14	Rivera	88.70	17.54	0.31
15	Rocha	83.64	22.14	0.12
16	Salto	89.12	24.29	0.29
17	Tacuarembó	100.22	28.77	-0.14
18	Treinta y Tres	89.80	28.36	0.21
19	Trinidad	88.17	27.80	-0.20
20	Young	90.03	24.55	-0.04

Intervalos de confianza según estimación por MLE

	Estación	Parametros	2.5%	Estimación	97.5%
1	Aeropuerto Carrasco	Locacion	68.05	76.20	84.34
2	Aeropuerto Carrasco	Escala	15.27	21.16	27.04
3	Aeropuerto Carrasco	Forma	-0.28	-0.02	0.23
4	Aeropuerto Melilla	Locacion	70.47	81.09	91.71
5	Aeropuerto Melilla	Escala	20.52	27.99	35.46
6	Aeropuerto Melilla	Forma	-0.29	-0.07	0.15
7	Artigas	Locacion	88.65	103.74	118.83
8	Artigas	Escala	26.44	37.71	48.99
9	Artigas	Forma	-0.35	-0.02	0.30
10	Bella Unión	Locacion	87.62	97.60	107.58
11	Bella Unión	Escala	18.08	25.58	33.08
12	Bella Unión	Forma	-0.21	0.07	0.34
13	Colonia	Locacion	69.81	80.38	90.95
14	Colonia	Escala	17.89	26.37	34.86
15	Colonia	Forma	-0.14	0.20	0.53
16	Durazno	Locacion	78.85	86.76	94.67
17	Durazno	Escala	13.63	19.59	25.56
18	Durazno	Forma	-0.35	-0.01	0.33
19	Melo	Locacion	78.86	87.07	95.28
20	Melo	Escala	15.68	21.46	27.23
21	Melo	Forma	-0.40	-0.15	0.10
22	Mercedes	Locacion	67.91	76.09	84.27
23	Mercedes	Escala	13.03	20.39	27.74
24	Mercedes	Forma	0.02	0.39	0.76
25	Palmitas	Locacion	74.56	85.76	96.97
26	Palmitas	Escala	21.21	29.17	37.14
27	Palmitas	Forma	-0.41	-0.16	0.10
28	Paso de los Toros	Locacion	77.09	84.61	92.14
29	Paso de los Toros	Escala	12.86	18.77	24.68
30	Paso de los Toros	Forma	-0.19	0.15	0.48
31	Paysandú	Locacion	78.81	88.26	97.72
32	Paysandú	Escala	16.86	23.83	30.80
33	Paysandú	Forma	-0.39	-0.08	0.23
34	Prado	Locacion	72.55	82.82	93.09
35	Prado	Escala	19.44	26.77	34.10
36	Prado	Forma	-0.46	-0.20	0.05
37	Punta del Este	Locacion	61.78	70.21	78.63
38	Punta del Este	Escala	14.33	20.74	27.15
39	Punta del Este	Forma	-0.35	0.00	0.35
40	Rivera	Locacion	81.50	88.69	95.88
41	Rivera	Escala	11.40	17.54	23.68
42	Rivera	Forma	-0.08	0.31	0.69
43	Rocha	Locacion	75.35	83.63	91.90
44	Rocha	Escala	16.00	22.13	28.27
45	Rocha	Forma	-0.09	0.12	0.33
46	Salto	Locacion	79.55	89.08	98.61
47	Salto	Escala	16.23	24.23	32.23
48	Salto	Forma	-0.03	0.29	0.61
49	Tacuarembó	Locacion	88.80	100.28	111.76
50	Tacuarembó	Escala	20.28	28.78	37.27
51	Tacuarembó	Forma	-0.47	-0.14	0.19
52	Treinta y Tres	Locacion	78.20	89.79	101.38
53	Treinta y Tres	Escala	18.92	28.35	37.79
54	Treinta y Tres	Forma	-0.17	0.21	0.58
55	Trinidad	Locacion	77.44	88.17	98.91
56	Trinidad	Escala	20.09	27.80	35.50
57	Trinidad	Forma	-0.47	-0.20	0.07
58	Young	Locacion	80.80	90.03	99.25
59	Young	Escala	18.10	24.55	30.99
60	Young	Forma	-0.23	-0.04	0.16

Intervalos de confianza según estimación por Máxima Verosimilitud Perfil

	Estación	Parametros	2.5%	Estimación	97.5%
1	Aeropuerto Carrasco	Locacion	68.24	76.1812	84.76
2	Aeropuerto Carrasco	Escala	16.29	21.1417	28.58
3	Aeropuerto Carrasco	Forma	-0.24	-0.0207	0.28
4	Aeropuerto Melilla	Locacion	70.53	81.0453	92.10
5	Aeropuerto Melilla	Escala	21.79	27.9791	37.38
6	Aeropuerto Melilla	Forma	-0.25	-0.0674	0.21
7	Artigas	Locacion	89.31	103.4441	119.77
8	Artigas	Escala	28.29	37.6008	52.10
9	Artigas	Forma	-0.30	-0.0196	0.37
10	Bella Unión	Locacion	87.99	97.5588	108.15
11	Bella Unión	Escala	19.20	25.5641	34.74
12	Bella Unión	Forma	-0.14	0.0659	0.44
13	Colonia	Locacion	70.52	80.3169	91.89
14	Colonia	Escala	19.22	26.3358	36.73
15	Colonia	Forma	-0.10	0.1961	0.60
16	Durazno	Locacion	79.23	86.7115	95.13
17	Durazno	Escala	14.53	19.5726	27.10
18	Durazno	Forma	-0.29	-0.0084	0.41
19	Melo	Locacion	78.84	87.0536	95.55
20	Melo	Escala	16.77	21.4598	29.09
21	Melo	Forma	-0.39	-0.1529	0.13
22	Mercedes	Locacion	68.74	76.0794	85.53
23	Mercedes	Escala	14.41	20.3746	29.65
24	Mercedes	Forma	0.03	0.3896	0.79
25	Palmitas	Locacion	74.57	85.7383	97.34
26	Palmitas	Escala	22.71	29.1674	39.66
27	Palmitas	Forma	-0.39	-0.1568	0.14
28	Paso de los Toros	Locacion	77.54	84.5483	92.73
29	Paso de los Toros	Escala	13.78	18.7429	26.01
30	Paso de los Toros	Forma	-0.14	0.1497	0.55
31	Paysandú	Locacion	79.10	88.2177	98.16
32	Paysandú	Escala	17.92	23.8298	32.73
33	Paysandú	Forma	-0.34	-0.0751	0.33
34	Prado	Locacion	72.48	82.817	93.33
35	Prado	Escala	20.82	26.7674	36.46
36	Prado	Forma	-0.43	-0.2022	0.11
37	Punta del Este	Locacion	62.22	70.1888	79.13
38	Punta del Este	Escala	15.30	20.7236	28.73
39	Punta del Este	Forma	-0.28	0.003	0.43
40	Rivera	Locacion	82.19	88.7034	96.79
41	Rivera	Escala	12.40	17.5436	25.07
42	Rivera	Forma	-0.04	0.3057	0.76
43	Rocha	Locacion	75.55	83.6422	92.44
44	Rocha	Escala	17.10	22.1437	29.84
45	Rocha	Forma	-0.06	0.1193	0.37
46	Salto	Locacion	80.32	89.1239	99.74
47	Salto	Escala	17.65	24.2902	34.27
48	Salto	Forma	0.01	0.288	0.66
49	Tacuarembó	Locacion	89.13	100.2167	112.27
50	Tacuarembó	Escala	21.81	28.766	40.06
51	Tacuarembó	Forma	-0.43	-0.1389	0.24
52	Treinta y Tres	Locacion	79.19	89.7995	102.75
53	Treinta y Tres	Escala	20.51	28.3613	40.06
54	Treinta y Tres	Forma	-0.15	0.2065	0.62
55	Trinidad	Locacion	77.44	88.1695	99.23
56	Trinidad	Escala	21.54	27.8005	38.10
57	Trinidad	Forma	-0.46	-0.2001	0.12
58	Young	Locacion	80.84	90.0274	99.63
59	Young	Escala	19.21	24.5463	32.61
60	Young	Forma	-0.20	-0.0361	0.22

Test de hipótesis según Likelihood Ratio Test:

A continuación se presentan los p-valores de aplicar el test de razón de verosimilitud para testear la hipótesis nula de que las distribuciones GEV marginales pertenecen a la familia Gumbel:

Estación	p-valor
Aeropuerto Carrasco	0.8738
Aeropuerto Melilla	0.574
Artigas	0.8992
Bella Unión	0.6184
Colonia	0.2128
Durazno	0.9575
Melo	0.2593
Mercedes	0.0358
Palmitas	0.2657
Paso de los Toros	0.3445
Paysandú	0.6505
Prado	0.1712
Punta del Este	0.9867
Rivera	0.0831
Rocha	0.2058
Salto	0.0411
Tacuarembó	0.4282
Treinta y Tres	0.2613
Trinidad	0.1897
Young	0.7328

Según dicho test, al 5% de significación, las estaciones que rechazan la hipótesis nula de que las distribuciones marginales asociadas a las estaciones son Gumbel, son Mercedes y Salto.

Cálculo de intervalos al 95% de confianza para las familias Gumbel:

	Estación	Parametros	2.5%	Estimación	97.5%
1	Aeropuerto Carrasco	Locacion	68.40	75.95	83.51
2	Aeropuerto Carrasco	Escala	15.42	21.02	26.61
3	Aeropuerto Melilla	Locacion	70.14	80.05	89.97
4	Aeropuerto Melilla	Escala	20.35	27.52	34.69
5	Artigas	Locacion	89.87	103.32	116.77
6	Artigas	Escala	27.27	37.39	47.52
7	Bella Unión	Locacion	89.09	98.51	107.92
8	Bella Unión	Escala	19.14	26.26	33.37
9	Colonia	Locacion	72.87	83.22	93.57
10	Colonia	Escala	20.82	28.95	37.09
11	Durazno	Locacion	79.64	86.66	93.67
12	Durazno	Escala	14.21	19.52	24.83
13	Melo	Locacion	77.84	85.34	92.85
14	Melo	Escala	15.45	20.79	26.14
17	Palmitas	Locacion	73.25	83.40	93.55
18	Palmitas	Escala	20.87	28.15	35.42
19	Paso de los Toros	Locacion	78.93	86.12	93.31
20	Paso de los Toros	Escala	14.49	20.08	25.67
21	Paysandú	Locacion	78.95	87.30	95.65
22	Paysandú	Escala	16.96	23.19	29.41
23	Prado	Locacion	70.78	79.99	89.20
24	Prado	Escala	18.89	25.50	32.11
25	Punta del Este	Locacion	62.76	70.22	77.68
26	Punta del Este	Escala	15.09	20.75	26.41
27	Rivera	Locacion	84.43	91.81	99.19
28	Rivera	Escala	14.76	20.68	26.59
31	Salto	Locacion	83.07	93.20	103.33
32	Salto	Escala	20.32	28.47	36.63
33	Tacuarembó	Locacion	88.30	98.19	108.07
34	Tacuarembó	Escala	20.12	27.44	34.75
35	Treinta y Tres	Locacion	81.84	93.08	104.31
36	Treinta y Tres	Escala	22.52	31.39	40.25
37	Trinidad	Locacion	75.77	85.30	94.84
38	Trinidad	Escala	19.54	26.41	33.29
39	Young	Locacion	80.78	89.54	98.31
40	Young	Escala	18.05	24.36	30.66

References

[1] BECHLER, A., BEL, L., AND VRAC, M. Conditional simulations of the extremal t process: application to fields of extreme precipitation. *Spatial statistics 12* (2015), 109–127.

[2] BECHLER, A., VRAC, M., AND BEL, L. A spatial hybrid approach for downscaling of extreme precipitation fields. *Journal of Geophysical Research: Atmospheres 120*, 10 (2015), 4534–4550.

[3] BERNARD, E., NAVEAU, P., VRAC, M., AND MESTRE, O. Clustering of maxima: Spatial dependencies among heavy rainfall in france. *Journal of Climate 26*, 20 (2013), 7929–7937.

[4] BLANCO, M., VAQUERA, H., VILLASEÑOR, J. A., VALDEZ-LAZALDE, J. R., ROSENGAUS, M., ET AL. Metodología para investigar tendencias espacio-temporales en eventos meteorológicos extremos: caso durango, México. *Tecnología y ciencias del agua 5*, 6 (2014), 25–39.

[5] CASTILLO, E., HADI, A. S., BALAKRISHNAN, N., AND SARABIA, J. M. *Extreme value and related models with applications in engineering and science*. John Wiley Sons, Inc., Hoboken, New Jersey, 2005.

[6] COLES, S., BAWA, J., TRENNER, L., AND DORAZIO, P. *An introduction to statistical modeling of extreme values*, vol. 208. Springer, 2001.

[7] COOLEY, D., NAVEAU, P., AND PONCET, P. Variograms for spatial max-stable random fields. In *Dependence in probability and statistics*. Springer, 2006, pp. 373–390.

[8] COOLEY, D., NAVEAU, P., AND PONCET, P. Variograms for spatial max-stable random fields. In *Dependence in probability and statistics*. Springer, 2006, pp. 373–390.

[9] CRESSIE, N. *Statistics for spatial data*, vol. 4. Wiley Online Library, 1992.

[10] DAVISON, A. C., PADOAN, S. A., AND RIBATET, M. Statistical modeling of spatial extremes. *Statistical Science 27*, 2 (2012), 161–186.

[11] DE HAAN, L. A characterization of multidimensional extreme-value distributions. *Sankhyā: The Indian Journal of Statistics, Series A* (1978), 85–88.

[12] DE HAAN, L. A spectral representation for max-stable processes. *The annals of probability 12*, 4 (1984), 1194–1204.

[13] DE HAAN, L., AND FERREIRA, A. *Extreme Value Theory. An Introduction*. Springer Science & Business Media, 2007.

[14] DURAÑONA, V. *Extreme wind climate of Uruguay*. PhD thesis, Facultad de Ingeniería - UdelaR - FI-IMFIA, 2015.

[15] EM-DAT, C. The ofda/cred international disaster database. universite catholique de louvain, brussels, belgium. 2016.

[16] FERNÁNDEZ, L. I., AND PARNÁS, V. B. E. Análisis de métodos de vientos extremos para calcular las velocidades básicas. *Revista Cubana de Ingeniería 7*, 2 (2016), 15–25.

[17] FISHER, R. A., AND TIPPETT, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society* (1928), Cambridge University Press, pp. 180–190.

[18] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. *The elements of statistical learning*. Springer series in statistics New York, 2001.

[19] GAETAN, C., AND GUYON, X. *Spatial statistics and modeling*, vol. 81. Springer, 2010.

[20] GALAMBOS, J. *The asymptotic theory of extreme order statistics*. Elsevier, 1977.

[21] GELFAND, A. E., DIGGLE, P., GUTTORP, P., AND FUENTES, M. *Handbook of spatial statistics*. CRC press, 2010.

[22] GNEDENKO, B. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics* (1943), 423–453.

[23] GNEDENKO, B. V., ALEKSANDR, I., AND KHINCHIN, A. *An elementary introduction to the theory of probability*, vol. 155. Courier Corporation, 1962.

[24] GREENWOOD, J. A., LANDWEHR, J. M., MATALAS, N. C., AND WALLIS, J. R. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water resources research 15*, 5 (1979), 1049–1054.

[25] HERNÁNDEZ, A., GUENNI, L., AND SANSÓ, B. Características de la precipitación extrema en algunas localidades de Venezuela. *Interciencia 36*, 3 (2011), 185–191.

[26] HOSKING, J. R. M., WALLIS, J. R., AND WOOD, E. F. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics 27*, 3 (1985), 251–261.

[27] HSING, T., HÜSLER, J., AND LEADBETTER, M. R. On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields 78*, 1 (1988), 97–112.

[28] IZENMAN, A. J. *Modern multivariate statistical techniques*. Springer, 2008.

[29] JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. *An introduction to statistical learning*, vol. 112. Springer, 2013.

[30] JENKINSON, A. F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society 81*, 348 (1955), 158–171.

[31] KALEMKERIAN, J. A truncated Cramér–von Mises test of normality. *Communications in Statistics-Theory and Methods* (2019), 48:16, 3956–3975, DOI: 10.1080/03610926.2018.1465093.

[32] KALEMKERIAN, J., AND FERNÁNDEZ, D. An Independence Test Based on Recurrence Rates. *arXiv preprint arXiv:1908.03305* (2019).

[33] KAUFMAN, L., AND ROUSSEEUW, P. *Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 344. John Wiley & Sons, 2009.

[34] KAUFMAN, L., AND ROUSSEEUW, P. J. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis* (1990), 68–125.

[35] LAOS, J. A. M. *La verosimilitud perfil en la Inferencia Estadística*. PhD thesis, Centro de Investigación en Matemáticas, 2008.

[36] LONG, C. S. Noaa, national weather service, national centers for environmental prediction. *Ultraviolet Index Verification Report-Indications of Surface Ultraviolet Radiation Observation Characteristics. Climate Prediction Center Report* (1996).

[37] MACQUEEN, J. B. Mathematical statistics and probability.

[38] MATHERON, G. Suffit-il, pour une covariance, d'être de type positif. *Sciences de la Terre, série informatique géologique 26* (1987), 51–66.

[39] MORENO, L. *Precipitaciones Máximas en el Estado de Guanajuato, México*. PhD thesis, Facultad de Ingeniería - UdelaR, 2013.

[40] MUNDIAL, F. E. World economic forum. Obtenido de Informe de Riesgos Mundiales: https://es.weforum.org/reports/the-global-risks-report-2019, 2019.

[41] NAVEAU, P., GUILLOU, A., AND COOLEY, D. Modelling pairwise dependence of maxima in space. *Biometrika 96*, 1 (2009), 1–17.

[42] PORTUGUÉS, S. B., SERRANO, S. M. V., AND MORENO, J. I. L. Distribución espacial y estacional de los eventos de precipitación en la rioja: intensidad, magnitud y duración. *Zubía* (2008), 169–186.

[43] R CORE TEAM. R: A language and environment for statistical computing. Vienna, Austria, 2019.

[44] RESNICK, S. I. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

[45] ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20* (1987), 53–65.

[46] RUSTICUCCI, M., MARENGO, J., PENALBA, O., AND RENOM, M. An intercomparison of model-simulated in extreme rainfall and temperature events during the last half of the twentieth century. part 1: mean values and variability. *Climatic Change 98*, 3-4 (2010), 493–508.

[47] SCHLATHER, M. Models for stationary max-stable random fields. *Extremes 5*, 1 (2002), 33–44.

[48] SMITH, R. L. Maximum likelihood estimation in a class of nonregular cases. *Biometrika 72*, 1 (1985), 67–90.

[49] SMITH, R. L. Max-stable processes and spatial extremes. *Unpublished manuscript 205* (1990).

[50] TEAM, R. C., ET AL. R: A language and environment for statistical computing.

[51] TENCER, B., AND RUSTICUCCI, M. Analysis of interdecadal variability of temperature extreme events in argentina applying evt. *Atmósfera 25*, 4 (2012), 327–337.

[52] UCLOUVAIN, CRED, U. Natural disasters 2018: An opportunity to prepare.

[53] VANNITSEM, S., AND NAVEAU, P. Spatial dependences among precipitation maxima over belgium. *Nonlinear Processes in geophysics 14*, 5 (2007), 621–630.

[54] VINOD, H. D. Integer programming and the theory of grouping. *Journal of the American Statistical association 64*, 326 (1969), 506–519.

[55] VON MISES, R. La distribution de la plus grande de n valuers. *Rev. math. Union interbalcanique 1* (1936), 141–160.

[56] WIKIPEDIA. El niño (fenómeno) — wikipedia, la enciclopedia libre. 2019. [Internet; descargado 25-agosto-2019].