



TESIS DOCTORAL

*Algunos problemas de
Estadística Computacional*

Mario Martínez Pizarro

Programa de doctorado
Modelización y Experimentación en Ciencia y Tecnología

Conformidad de la directora y los codirectores:

Dra. D^ª. María Isabel Parra Arévalo

Dra. D^ª. Eva Teresa López Sanjuán Dr. D. Jacinto Ramón Martín Jiménez

Esta tesis cuenta con la autorización de la directora y codirectores de la misma y de la Comisión Académica del programa. Dichas autorizaciones constan en el Servicio de la Escuela Internacional de Doctorado de la Universidad de Extremadura.

2023

*A mi familia.
A mis abuelos que hoy no están.*

Agradecimientos

Quiero comenzar estas líneas por quiénes me dieron la oportunidad para comenzar este viaje, mis directores. Muchas gracias a D^a. M^a. Isabel Parra Arévalo, D^a. Eva T. López Sanjuán y D. Jacinto R. Martín Jiménez, por animarme a emprender este camino. Gracias por compartir vuestra experiencia conmigo, por la dedicación y el apoyo constante, y por ayudarme principalmente a crecer en el ámbito personal y no solamente en el profesional. Mi más sincero agradecimiento, especialmente por esos largos ratos de conversación (y cafés) con vuestros mejores consejos.

A Maribel y Eva, por apoyarme diariamente y aconsejarme en todo momento, y a Jacinto por su insistencia para que comenzase este viaje, y acompañarme en todas las paradas.

A M. Carmen y Dani (Merino), mis compañeros de viaje. Gracias por esas charlas donde nos animábamos unos a otros para no rendirnos cuando parecía que no salía nada o que este viaje se hacía infinito. No quiero olvidarme tampoco de Dani (Morales), Carmen, José y Felipe por sus múltiples consejos. Muchas gracias por estar siempre ahí para escucharme y aconsejarme.

Quiero tener también unas palabras de agradecimiento para todos aquellos, de dentro o fuera del departamento, que me han ayudado y acompañado en este viaje durante estos años, especialmente a M^a. Ángeles, Javi, Agustín, Alfonso y Reme.

Quiero finalizar estas líneas con las personas que me acompañan en el día a día, mi familia y amigos.

Gracias al pilar fundamental de mi vida, mi familia. Sin su apoyo no habría empezado este largo viaje. A mis padres por su amor y constancia, por mostrarme que con trabajo y esfuerzo siempre se consigue llegar al destino final. A mi hermano, por su apoyo y ánimo, por estar siempre a mi lado por muy largo y duro que sea el viaje. A mis abuelos, por acompañarme en las múltiples paradas de este viaje y que les hubiera encantado verme llegar a esta parada final. Gracias a mi abuela, por su cariño y sus consejos en todo momento. A mis padrinos, tíos y primos por

acompañarme y apoyarme siempre.

No quiero dejar atrás a mi segunda familia, que me ha cuidado como un hijo y hermano más. Gracias a todos y cada uno de ellos por acompañarme y apoyarme en este viaje tan largo.

Otro de los pilares de mi vida, mis amigos: Pilar, María, Cristina, Alberto y Luisfe. Sin ellos, este viaje no hubiera sido el mismo, un viaje lleno de alegrías y risas, pero también de esfuerzo y constancia. Gracias a todos por mantenerme con los pies en la tierra y no dejarme abandonar en los momentos difíciles. Especialmente a Pilar, por ser el apoyo constante.

Muchísimas gracias a todos.

Badajoz, octubre de 2023

Mario Martínez Pizarro

Esta Tesis Doctoral ha sido financiada por:

- Ministerio de Ciencia e Innovación y Agencia Estatal de Investigación: proyectos MTM2017-86875-C3-2-R y PID2021-122209OB-C32.



- Junta de Extremadura: proyectos GR18108, GR21057 e IB16063.



Consejería de Economía, Ciencia y Agenda Digital

- European Union: proyectos GR18108, GR21057, IB16063, MTM2017- 86875-C3-2-R y PID2021-122209OB-C32.



EUROPEAN UNION
European Regional Development Fund

Resumen

La presente Tesis Doctoral se centra en plantear estrategias que ayuden a encontrar mejores soluciones para algunos problemas de estadística, empleando algoritmos computacionales. Se estructura en dos partes bien diferenciadas. En la primera parte se proponen modelos de Inferencia Bayesiana para estimar con mayor exactitud y precisión los parámetros de las distribuciones de datos extremos. Mientras que, en la segunda parte se presenta cómo usar regresión simbólica para optimizar la búsqueda de modelos matemáticos, aplicados a ciertas variables termodinámicas de interés.

La primera parte versa sobre la Teoría de Valores Extremos que habitualmente utiliza la información de una parte reducida del conjunto de datos para estimar los parámetros de la distribución límite, cuestión que implica que se produzcan grandes sesgos y errores en las estimaciones. El motivo principal es que emplea dos estrategias diferentes: el método de máximos de bloque y el método de excesos de un umbral, quedándose en ambos casos con pocos datos extremos. Para ambos enfoques, se dispone de sendos teoremas fundamentales que permiten caracterizar las distribuciones límites de valores extremos; es decir, la distribución de Valores Extremos Generalizada y la distribución de Pareto Generalizada, respectivamente, cuyos parámetros suele ser interesante estimar.

Bajo un enfoque frecuentista, los métodos de estimación habitualmente utilizados requieren de condiciones asintóticas en los parámetros. En cambio, desde un punto de vista Bayesiano se salva dicho problema, y además se aporta otra ventaja importante al permitir incluir información adicional a través de la distribución a priori elegida.

Existen multitud de trabajos sobre estimación Bayesiana para los parámetros de las distribuciones de valores extremos que consideran distribuciones a priori no informativas, y usan solo los datos extremos, desaprovechando la información que pudiera contener el resto de valores del conjunto de datos. Modificar estas condiciones puede permitirnos plantear nuevas estrategias que minimicen los errores de

estimación.

Con este objetivo en mente, la propuesta de este trabajo es usar distribuciones a priori altamente informativas. Éstas se han construido aprovechando la información disponible en todo el conjunto de observaciones, y no solamente en los valores extremos. En concreto, se han establecido relaciones teóricas y/o empíricas entre los parámetros de la distribución del conjunto total de observaciones y los parámetros de la distribución de valores extremos.

En el primer artículo que compone esta Tesis, se propone una nueva estrategia, con distribuciones a priori muy informativas, para estimar los parámetros de la distribución Gumbel. Ésta es un caso particular de la distribución de Valores Extremos Generalizada, y juega un papel importante en el análisis de valores extremos como modelo para máximos de bloque, ya que es idónea para describir eventos extremos de diversas distribuciones que aparecen de manera habitual en aplicaciones prácticas, como la distribución Exponencial o la Normal. Gracias al concepto de dominio de atracción, se pueden encontrar relaciones teóricas entre los parámetros de la distribución Gumbel para los máximos de bloque, y los parámetros de las distribuciones Exponencial y Normal para el conjunto completo de observaciones. En aquellos casos donde no sea posible establecer relaciones analíticas, se puede recurrir a hacerlo de manera empírica.

En el segundo artículo, se propone una estrategia equivalente a la anterior para estimar los parámetros de la distribución de Pareto Generalizada. Esta distribución permite modelar los valores que exceden un umbral; esto es, los valores que se encuentran en la cola de la distribución. En esta ocasión, se establecen relaciones teóricas y empíricas entre los parámetros de dicha distribución y los parámetros de las distribuciones estables para las observaciones originales, en particular las distribuciones Normal, Cauchy y Lévy.

Desde un punto de vista más aplicado, la Teoría de Valores Extremos se ha empleado con éxito en problemas de muy distinta naturaleza englobados en disciplinas tan variadas como la climatología (eventos extremos de temperatura, precipitaciones y climatología solar), o las finanzas y seguros (rendimientos sobre activos financieros y medidas de riesgo). Se comprobará cómo se trasladan las mejoras en la estimación propuestas sobre conjuntos de datos relativos a problemas enmarcados en estos ámbitos.

Concretamente, en climatología es interesante plantear modelos espaciales que caractericen y pronostiquen valores extremos multivariantes. Desde un enfoque Bayesiano, los más utilizados son los modelos Jerárquicos Bayesianos, que permiten estimar los parámetros del modelo dividiendo el proceso en varias etapas. Existen

múltiples trabajos que presentan estos modelos para eventos extremos de temperaturas o precipitaciones; sin embargo, en ellos se supone que las observaciones son independientes respecto al espacio, algo que no ocurre en situaciones reales, siendo necesario incluir una nueva herramienta que resuelva este problema. Sobre esto trata el tercer artículo de la presente Tesis Doctoral, donde se plantea un nuevo modelo que introduce el concepto de cópula para modelar la dependencia espacial existente entre las observaciones. Una cópula es una función de distribución multivariante que controla la dependencia entre las distintas variables unidimensionales. Además, permite construir una distribución multidimensional con las distribuciones marginales que se deseen. En particular, se emplea una cópula Gaussiana cuyas distribuciones marginales son distribuciones de Valores Extremos Generalizadas, puesto que se propone un modelo Jerárquico espacial para modelar temperaturas extremas en una región. Este nuevo modelo con cópula proporciona mejores resultados que el modelo clásico donde se considera que las observaciones son independientes respecto al espacio.

En los primeros tres artículos de esta Tesis, se han incluido exhaustivos estudios de simulación para valorar el comportamiento de los nuevos modelos propuestos frente a los modelos clásicos, partiendo de diversos escenarios, variando la distribución de las observaciones de partida según sea el caso, y número de observaciones, el tamaño del bloque o umbral, y número de réplicas. Todos los resultados obtenidos demuestran que las estimaciones proporcionadas por las estrategias de estimación propuestas son más precisas que las obtenidas utilizando las clásicas. Para completar cada estudio, también se han aplicado a distintos tipos de datos reales, destacando datos sobre contaminación atmosférica o temperaturas extremas.

La segunda parte incluye el cuarto artículo, que aborda un problema muy diferente como es la aplicación de algoritmos genéticos, en concreto, de métodos de regresión simbólica, para la búsqueda del modelo más adecuado para representar las relaciones existentes entre las variables de un conjunto de datos.

Los algoritmos genéticos son técnicas de optimización basadas en principios evolutivos estocásticos. Estos algoritmos permiten encontrar buenas soluciones para un problema concreto en tiempo razonable, mediante la evolución genética de una población de individuos que representen soluciones candidatas. En particular, la regresión simbólica permite obtener la estructura de una expresión que puede modelar un conjunto dado de datos, sin necesidad de asumir un formato de correlación específico para ellos. Tanto la forma analítica del modelo como sus coeficientes evolucionan automáticamente. La salida de este tipo de algoritmos proporciona las mejores expresiones matemáticas encontradas, atendiendo a algún criterio prefi-

jado. Esta situación permite analizar posteriormente dichos modelos propuestos, realizando, por ejemplo, un análisis de sensibilidad, e incluso pueden ser una base para generar mejores modelos para los datos. Para mostrar su viabilidad, se ha aplicado a datos de la tensión superficial, resultando fácilmente extrapolable a otras propiedades termodinámicas y a otros problemas de muy distinta naturaleza.

Objetivos de la Tesis Doctoral

Los objetivos propuestos se pueden resumir en dos principales:

- Mejorar las estrategias de estimación Bayesiana para los parámetros de las distribuciones límite de valores extremos que minimicen los errores de estimación.
- Emplear técnicas de regresión simbólica para descubrir el modelo subyacente satisfactorio para representar las relaciones entre las variables de un conjunto de datos experimentales.

Para conseguir éstos, se proponen los siguientes objetivos parciales:

1. Proponer nuevas estrategias de estimación Bayesiana con distribuciones a priori altamente informativas para los parámetros de las distribuciones límite de valores extremos, tanto máximos de bloque como excesos de un umbral.
2. Plantear un nuevo modelo espacial para valores extremos multivariantes que considere la dependencia espacial entre las observaciones.
3. Programar y analizar el uso de la regresión simbólica para la búsqueda de modelos matemáticos adecuados para describir las relaciones existentes entre variables termodinámicas.
4. Construir los algoritmos computacionales necesarios para poner en práctica y valorar los nuevos modelos propuestos, utilizando el lenguaje de programación R, de la manera más eficiente posible.
5. Valorar la viabilidad de las estrategias propuestas al ser aplicadas a problemas reales.

Índice general

Introducción General	1
I Inferencia Bayesiana para Valores Extremos	7
1. Introducción a los Valores Extremos	9
1.1. Máximos de Bloque	9
1.2. Excesos de un umbral	12
1.3. Modelos Jerárquicos Bayesianos	14
1.4. Introducción a la Teoría de Cópulas	15
1.5. Algoritmos para Inferencia Bayesiana	16
2. Artículo A: Métodos Base de inferencia Bayesiana en la distribución Gumbel	21
3. Artículo B: Métodos Base para la estimación de los parámetros de la distribución de Pareto Generalizada	41
4. Artículo C: Un modelo jerárquico Bayesiano espacial con cópula: Una aplicación a temperaturas extremas en Extremadura (España)	59
II Regresión Simbólica	77
5. Introducción a la Regresión Simbólica	79
6. Artículo D: Desarrollo de modelos de tensión superficial de alcoholes mediante Regresión Simbólica	85

Conclusiones Finales	95
Apéndice A. Trabajos en proceso de publicación	97
Apéndice B. Actividades desarrolladas durante la Tesis Doctoral	101
Bibliografía	111

Introducción General

El análisis clásico de conjuntos de datos se centra principalmente en estudiar la distribución de la parte central de los valores, pero suele ignorar los datos extremos, que por otra parte son generalmente escasos. Los valores extremos se corresponden con eventos que suceden excepcionalmente, y su estudio tiene gran interés en áreas de muy diversa naturaleza como la climatología, en el estudio de temperaturas extremas [1, 2], precipitaciones e inundaciones [3, 4] o en climatología solar [5, 6]. También es muy habitual el estudio de valores extremos en un contexto financiero [7, 8].

La Teoría de Valores Extremos (EVT) engloba el conjunto de herramientas estadísticas que permiten modelar y pronosticar las distribuciones que surgen cuando se estudian los valores extremos. Un estudio detallado puede consultarse en [9, 10, 11]. La EVT propone, principalmente, dos estrategias diferentes: el método de máximos de bloque (o block maxima, BM) y el método de excesos de un umbral (o peaks-over-threshold, POT), cuya diferencia se encuentra en el modo de clasificar las observaciones consideradas como valores extremos.

El método BM consiste en dividir las observaciones en bloques de igual tamaño y seleccionar el dato máximo para cada uno de los bloques como valores extremos. Resulta especialmente útil cuando las observaciones tienen un carácter estacional. El principal reto de este método es encontrar el tamaño de bloque óptimo, pues debe ser suficientemente grande para asegurar que todos los datos seleccionados son realmente extremos, pero no excesivamente, puesto que aumentar el tamaño del bloque implica una disminución en el número de datos.

En 1927, *Fréchet* [12] encontró la primera distribución asintótica para los máximos de bloque, y un año más tarde *Fisher* y *Tippet* [13] probaron la existencia de otras dos distribuciones asintóticas válidas. Posteriormente, en 1943, *Gnedenko* [14] demostró que estas tres distribuciones se podían agrupar en una única familia de distribuciones, conocida como distribución de Valores Extremos Generalizada (GEV), dando lugar a uno de los principales teoremas límite de la EVT.

En la década de los años 50, *Weibull* y *Gumbel* hicieron aportaciones a las distribuciones planteadas por *Fisher* y *Tippet*, por lo cual las distribuciones acabaron denominándose distribución Weibull y distribución Gumbel. En particular, *Gumbel* es uno de los grandes referentes en la Teoría de Valores Extremos debido al texto *Statistics of Extremes* publicado en 1958 [15]. La distribución que lleva su nombre juega un papel importante en la EVT como modelo para máximos de bloque, por ser idónea para describir valores extremos de distintas distribuciones que aparecen de manera habitual en datos reales.

Por otro lado, el método POT considera como valores extremos aquellos que exceden un cierto umbral, siendo la elección de éste el principal reto. Al igual que en el método BM, su elección debe ser un valor lo suficientemente alto para que los valores seleccionados puedan considerarse realmente como extremos, pero no tanto como para seleccionar una cantidad demasiado pequeña de observaciones. En 1974, *Balkema* y *de Haan* [16] y *Pickands* en 1975 [17] estudiaron las colas de las distribuciones, demostrando que los valores situados en ellas pueden modelarse asintóticamente a partir de la distribución de Pareto Generalizada (GPD), dando lugar a otro de los teoremas límite fundamentales para la EVT.

Resumiendo, la selección del tamaño de bloque y del valor del umbral, para el método BM y el método POT, respectivamente, debe ser suficientemente grande como para garantizar que los valores elegidos son extremos. Sin embargo, esta elección debe hacerse con sumo cuidado, ya que un valor demasiado alto implicaría un gran sesgo y falta de precisión en las estimaciones. Aún con una elección óptima, ambos métodos prescinden de la mayoría de los datos disponibles y, por tanto, de la información que pudieran contener.

Uno de los objetivos de la presente Tesis doctoral es proponer nuevas estrategias de estimación para los parámetros de las distribuciones límite desde un punto de vista Bayesiano. Este enfoque permite aliviar el problema que presentan los métodos clásicos, que requieren condiciones asintóticas en los parámetros, y aporta la ventaja de incluir información adicional a través de las distribuciones a priori de los parámetros. La idea de fondo será aprovechar la información contenida en los datos que son descartados por BM y POT para definir distribuciones a priori muy informativas.

En particular, se proponen nuevas estrategias de estimación para las distribuciones límite asociadas a ambos métodos de la EVT, basadas en las relaciones teóricas y/o empíricas existentes entre los parámetros de la distribución inicial de las observaciones, denominada *distribución base*, y los parámetros de la distribución de valores extremos correspondiente. Su implementación se realiza mediante

el algoritmo de Metropolis-Hastings (MH) que es una herramienta computacional que permite construir cadenas de Markov de Monte Carlo (MCMC) para los parámetros de los cuales se desea realizar inferencia. Un estudio más profundo del algoritmo puede encontrarse en [18, 19, 20, 21, 22]. Estas nuevas estrategias de estimación bayesiana se exponen para la distribución Gumbel en el artículo A [23] y para la distribución de Pareto Generalizada en el artículo B [24].

A partir de las relaciones encontradas entre los distintos parámetros, se han planteado dos nuevas estrategias de estimación. La primera de ellas consiste en estimar los parámetros de la distribución base, y posteriormente aplicar las relaciones mencionadas anteriormente para obtener estimaciones de los parámetros de la distribución de los valores extremos. Esta estrategia presenta un gran inconveniente, pues es necesario conocer previamente cuál es la distribución de las observaciones, cuestión que generalmente se desconoce en situaciones reales. La segunda consiste en construir distribuciones a priori altamente informativas, aprovechando la información del conjunto de datos completo y las relaciones entre los parámetros de las distribuciones base y de valores extremos correspondientes, para la inferencia bayesiana [25, 26].

Los métodos BM y POT de la Teoría de Valores Extremos tienen múltiples aplicaciones a situaciones reales [27, 28, 29, 30, 31, 32]. Cabe mencionar que el método BM suele emplearse principalmente en el campo de la climatología. Es por ello que en la presente Tesis Doctoral se muestra su aplicación en un contexto real.

En la actualidad, el estudio de los eventos que se producen asociados a temperaturas extremas tiene un gran interés por las consecuencias del cambio climático. El método BM permite modelar y pronosticar los eventos máximos a partir de la distribución de Valores Extremos Generalizada, dado su carácter estacional. Para ello, se han propuesto distintos modelos climáticos, destacando los modelos espaciales, dado que los eventos atmosféricos, como las temperaturas o las precipitaciones, tienen una clara dependencia de las coordenadas geográficas del lugar donde se observa dicho valor. Una de las propuestas más comunes para modelar esta dependencia es utilizar modelos Jerárquicos Bayesianos que permiten estimar los parámetros utilizando modelos espaciales [33, 34, 35, 36, 37, 38].

Un concepto fundamental, que puede aplicarse con los modelos jerárquicos, es el de cópula. Una cópula es una función de distribución multivariante cuyas distribuciones marginales son distribuciones uniformes, que permiten describir la dependencia entre las distintas variables aleatorias. En 1959, *Sklar* [39] estableció que cualquier función de distribución conjunta multivariante queda determinada de manera unívoca a partir de distribuciones marginales univariantes y una cópula

que describe la estructura de dependencia entre las variables. Como consecuencia de este célebre resultado es posible crear distribuciones multivariantes con las marginales deseadas, sin más que aplicar una cópula que permita modelar la dependencia. Se pueden estudiar en profundidad en [40, 41, 42].

En 2007, *Renard* [43] propuso introducir las cópulas en los modelos Jerárquicos Bayesianos, con el objetivo de controlar la dependencia existente entre las distintas observaciones. Esto ha permitido aplicar las cópulas en modelos espaciales, como se puede consultar en [44, 45, 46]. Con esta filosofía, en el artículo C [47] se presenta un modelo Jerárquico Bayesiano para predecir las temperaturas máximas en la región de Extremadura. Se utiliza una cópula Gaussiana para representar la dependencia espacial existente entre las temperaturas observadas en distintos lugares geográficos de la región, considerando una distribución de Valores Extremos Generalizada como distribución marginal, cuyos parámetros siguen modelos espaciales que dependen de las coordenadas geográficas [48].

La presente Tesis Doctoral lleva implícita una importante componente computacional, debido a la programación de los distintos algoritmos y la ejecución de simulaciones masivas, para valorar la calidad de las distintas estrategias de estimación propuestas. Todos los programas de la primera parte se han desarrollado en R por tratarse del entorno computacional dedicado a la estadística más utilizado actualmente, que facilita enormemente las tareas de programación. Se han creado sendas funciones para los algoritmos de estimación propuestos, a las que se puede acceder de forma libre y abierta [49] siguiendo la filosofía de R.

Otro problema interesante es tratar de encontrar modelos analíticos adecuados para variables de interés, sin necesidad de conocer previamente su estructura.

Los algoritmos genéticos se basan en principios evolutivos estocásticos con el objetivo de encontrar el extremo global de una función dada [50, 51]. En 1992, *Koza* [52] popularizó un caso particular de este tipo de algoritmos, conocido como programación genética.

La *regresión simbólica* es un tipo concreto de programación genética que permite obtener una expresión matemática óptima para modelar un conjunto de datos experimentales, sin necesidad de conocer su forma funcional. En esta técnica, tanto la configuración del modelo como sus coeficientes evolucionan automáticamente, empleando combinaciones genéticas entre las expresiones matemáticas consideradas como individuos de una población.

En el artículo D [53] de la presente Tesis Doctoral, se describe cómo emplear esta técnica estadística para encontrar un modelo de correlación óptimo para variables termodinámicas de cualquier familia de fluidos. En particular, se aplica a la

tensión superficial. Se trata de una propiedad termodinámica que juega un papel importante en el desarrollo, la producción y el rendimiento de productos químicos utilizados en farmacia, cosmética, desinfectantes, entre otros [54, 55, 56, 57].

A pesar de la gran cantidad de modelos que se han propuesto para la tensión superficial, no existe actualmente un modelo general para predecirla teóricamente. Por este motivo, se emplean modelos empíricos. Sin embargo, el principal reto se encuentra en determinar el modelo óptimo, puesto que según *O'Connell* [58], para la elección debe considerarse la complejidad, aplicabilidad, precisión y capacidad para predecir. En la actualidad, el número de nuevos modelos está creciendo muy rápidamente, como puede consultarse en [59, 60, 61, 62, 63, 64].

La regresión simbólica permite hacer un barrido masivo por distintos tipos de expresiones analíticas, permitiendo dar mayor peso a componentes deseables o forzando que aparezcan en ellas (para garantizar determinado comportamiento), definir qué operadores, variables y parámetros intervienen, así como el criterio de calidad del modelo. Se obtienen modelos generales muy sencillos con errores porcentuales medios inferiores al 7%.

Estructura de la Tesis Doctoral

La presente Tesis Doctoral consta de las siguientes 4 publicaciones científicas con aportaciones a la resolución de dos problemas de muy distinta naturaleza, pero con la componente común de abordarlos sirviéndose de estadística computacional.¹

- Parte I. Inferencia Bayesiana para Valores Extremos
 - *Baseline Methods for Bayesian Inference in Gumbel Distribution*: se propone una nueva estrategia de estimación bayesiana para los parámetros de la distribución Gumbel empleando distribuciones a priori altamente informativas. Éstas se construyen a partir de las relaciones existentes entre distribuciones muy conocidas, como la normal, con la distribución Gumbel.
 - *Baseline Methods for the Parameter Estimation of the Generalized Pareto Distribution*: se plantean estrategias bayesianas de estimación para los parámetros de la distribución de Pareto Generalizada utilizando las relaciones que existen entre ellos y los parámetros de las distribuciones

¹Las funciones programadas para poner en práctica las distintas estrategias propuestas están disponibles de forma abierta y libre en RPubS [49].

estables. Estas relaciones permiten construir distribuciones a priori muy informativas para los parámetros de escala y forma de la GPD.

- *A Bayesian Hierarchical Spatial Copula Model: An Application to Extreme Temperatures in Extremadura (Spain)*: se propone un modelo jerárquico Bayesiano para predecir temperaturas extremas, utilizando el concepto de cópula para construir distribuciones multivariantes. En particular, se emplea una cópula Gaussiana y la distribución de Valores Extremos Generalizada como distribuciones marginales.
- Parte II. Regresión Simbólica
 - *Development of models for surface tension of alcohols through symbolic regression*: se propone un modelo de correlación para la tensión superficial de alcoholes utilizando regresión simbólica. Este es un caso particular de algoritmos genéticos, que permite buscar un modelo satisfactorio empleando combinaciones genéticas de expresiones matemáticas que representan los individuos de una población.

Parte I

Inferencia Bayesiana para Valores
Extremos

Capítulo 1

Introducción a los Valores Extremos

Como comienzo de esta Tesis Doctoral, se presentan brevemente los conceptos teóricos necesarios para comprender las distintas estrategias que se emplean a lo largo de la primera parte para estimar los parámetros de las distribuciones de valores extremos. En primer lugar, se introduce el método de máximos de bloque de la Teoría de Valores Extremos, para continuar con el método de excesos de un umbral. Posteriormente, se introducen los conceptos de modelo Jerárquico Bayesiano y cópula. Se finaliza exponiendo algunos algoritmos computacionales muy empleados en estadística bayesiana.

1.1. Máximos de Bloque

El método de máximos de bloque consiste en dividir las observaciones en bloques no superpuestos del mismo tamaño y seleccionar el máximo para cada bloque. Esto permite dar la siguiente definición de valor extremo.

Definición 1.1. *Dada una sucesión de variables aleatorias independientes e idénticamente distribuidas (v.a.i.i.d.) Y_1, Y_2, \dots, Y_m , con función de distribución común F , denominada distribución base, y dado $k \in \mathbb{N}$ fijo, que denota el tamaño del bloque, se define el valor extremo máximo de bloque como*

$$X_i = \max_{(i-1)k < j \leq ik} Y_j, \quad i = 1, 2, \dots, n, \quad (1.1)$$

con $n = \left\lfloor \frac{m}{k} \right\rfloor$ ¹.

La función de distribución del máximo de bloque X_i es

$$P(X_i \leq x) = P(Y_{(i-1)k+1} \leq x, \dots, Y_{ik} \leq x) = \prod_{j=(i-1)k+1}^{ik} P(Y_j \leq x) = F(x)^k. \quad (1.2)$$

Como se observa en la ecuación (1.2), conocer la distribución de máximos de bloque de un conjunto de observaciones depende de F , que suele ser desconocida, especialmente en aplicaciones a problemas reales. Una posibilidad podría consistir en estimar F a partir de los datos observados y sustituirla posteriormente; el problema sería que el error cometido en F podría ser significativo en F^k . Otra opción sería asumir que F^k es desconocida y buscar qué familia de distribuciones puede aproximarla, utilizando únicamente los datos extremos. Por este motivo es de gran utilidad considerar la distribución asintótica. Para ello, se emplea uno de los resultados principales de la Teoría de Valores Extremos que puede consultarse en [13, 14], y se enuncia a continuación.

Teorema 1.2. (Fisher, Tipper y Gnedenko) *Cuando el tamaño de bloque es suficientemente alto, la distribución asintótica de los máximos de bloque puede aproximarse según la distribución de Valores Extremos Generalizada, con función de distribución*

$$GEV(x; \xi, \sigma, \mu) = \exp \left\{ - \left(1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right)^{-1/\xi} \right\}, \quad (1.3)$$

siendo $\xi, \mu \in \mathbb{R}, \sigma > 0$ los parámetros de forma, localización y escala, respectivamente, con dominio $1 + \xi \left(\frac{x - \mu}{\sigma} \right) > 0$.

Cuando $\xi = 0$, la parte derecha de la ecuación (1.3) se interpreta como

$$G(x; \sigma, \mu) = \exp \left\{ - \exp \left\{ - \left(\frac{x - \mu}{\sigma} \right) \right\} \right\}, \quad (1.4)$$

y se denomina distribución Gumbel con parámetros de localización $\mu \in \mathbb{R}$ y escala $\sigma > 0$, con $x \in \mathbb{R}$.

A continuación, se introduce el concepto de dominio de atracción que juega un papel importante en la Teoría de Valores Extremos.

¹[·] indica la parte entera.

Definición 1.3. Se dice que una función de distribución F pertenece al dominio de atracción de la distribución de valores extremos H , siendo H no degenerada, si existen sucesiones de números reales (a_k) y (b_k) , con $a_k > 0$, $b_k \in \mathbb{R}$, tales que

$$\lim_{k \rightarrow \infty} F(a_k \cdot x + b_k)^k = H(x), \quad (1.5)$$

para todo x punto de continuidad de H . A las sucesiones de números reales (a_k) y (b_k) se las denomina constantes de normalización, y representan los parámetros de escala y localización definidos en las ecuaciones (1.3) y (1.4), respectivamente, siendo asintóticamente k el tamaño del bloque.

Las constantes de normalización permiten relacionar la distribución base con la distribución de valores extremos máximo de bloque. Como caso particular, en Ferreira y de Haan [11] se pueden consultar los siguientes resultados que permiten determinar las constantes de normalización para cualquier distribución base F que pertenezca al dominio de atracción de una distribución de Valores Extremos Generalizada con parámetro de forma $\xi \in \mathbb{R}$.

Teorema 1.4. Sea F una función de distribución y x^* el mayor valor de su dominio. F pertenece al dominio de atracción de una distribución de Valores Extremos Generalizada H , con parámetro de forma $\xi \in \mathbb{R}$, si y solo si

a) Para $\xi > 0$: x^* es infinito y

$$\lim_{t \rightarrow \infty} \frac{1 - F(t \cdot x)}{1 - F(t)} = x^{-1/\xi}, \quad (1.6)$$

para todo $x > 0$.

b) Para $\xi < 0$: x^* es finito y

$$\lim_{t \rightarrow 0} \frac{1 - F(x^* - t \cdot x)}{1 - F(x^* - t)} = x^{-1/\xi}, \quad (1.7)$$

para todo $x > 0$.

c) Para $\xi = 0$: x^* puede ser finito o infinito y

$$\lim_{t \rightarrow x^*} \frac{1 - F(t + x \cdot h(t))}{1 - F(t)} = \exp\{-x\}, \quad (1.8)$$

para todo $x \in \mathbb{R}$, donde h es una función positiva adecuada.

Teorema 1.5. (Condición de Von Mises) Sean F una función de distribución y x^* el mayor valor de su dominio. Suponiendo que existe $F''(x)$, y que $F'(x)$ es positiva para todo x perteneciente a un entorno por la izquierda de x^* , si

$$\lim_{t \rightarrow x^*} \left(\frac{1 - F}{F'} \right)' (t) = \xi, \quad (1.9)$$

o equivalentemente

$$\lim_{t \rightarrow x^*} \frac{(1 - F(t)) \cdot F''(t)}{(F'(t))^2} = -\xi - 1, \quad (1.10)$$

entonces F pertenece al dominio de atracción de una distribución de Valores Extremos Generalizada con parámetro de forma $\xi \in \mathbb{R}$.

Corolario 1.6. Si F pertenece al dominio de atracción de una distribución Gumbel, entonces

$$\lim_{k \rightarrow \infty} F(a_k \cdot x + b_k)^k = \exp \{ -\exp \{ -x \} \}, \quad \forall x \in \mathbb{R}, \quad (1.11)$$

con

$$a_k := h(b_k), \quad b_k := F^{-1}(1 - k^{-1}), \quad (1.12)$$

y h definida como

$$h(t) := \frac{\int_t^{x^*} (1 - F(s)) ds}{1 - F(t)}, \quad \forall t \in \mathbb{R}, \quad t < x^*. \quad (1.13)$$

La condición de Von Mises (teorema 1.5) proporciona una forma de determinar la función positiva h como

$$h(t) := \frac{1 - F(t)}{F'(t)}. \quad (1.14)$$

1.2. Excesos de un umbral

El modelo de máximos de bloque presenta algunos problemas, por ejemplo cuando el comportamiento de los valores extremos no es estable, aunque quizás el más importante sea la cantidad de datos que menosprecia. Por ello, en la Teoría de Valores Extremos es más usado el modelo de excesos de un umbral que considera todos los valores que exceden un umbral fijado. Esta nueva metodología plantea un nuevo concepto de valor extremo a partir de la cola de la distribución de una variable aleatoria.

Definición 1.7. Sea X una variable aleatoria con función de distribución F , denominada distribución base. Fijado el valor $u > 0$ como el umbral, entonces se definen los valores extremos excesos de u como aquellos valores de X mayores que u . En particular, se define la variable aleatoria X_u como los valores de X que exceden el umbral u trasladados a 0; es decir, $(X - u | X > u)$, cuya función de distribución viene dada por:

$$F_{X_u}(x) = P(X_u \leq x) = P(X \leq x + u | X > u) = \frac{F(x + u) - F(u)}{1 - F(u)}, \quad (1.15)$$

para $0 \leq x \leq x^* - u$, siendo $x^* = \sup\{x : F(x) < 1\}$. Además, a F_{X_u} se la denomina función de distribución de excesos.

La distribución de Pareto Generalizada juega un papel importante en este contexto, ya que se comporta como la distribución asintótica de este tipo de valores extremos, cuando el umbral es suficientemente alto.

Definición 1.8. Dada una variable aleatoria X , se dice que sigue una distribución de Pareto Generalizada (GPD) cuando su función de distribución es

$$GPD(x; \xi, \sigma) = 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi}, \quad (1.16)$$

siendo $\xi \in \mathbb{R}, \sigma > 0$ los parámetros de forma y escala, respectivamente. Además, esta función de distribución es válida cuando $x \geq 0$ para $\xi \geq 0$, y cuando $0 \leq x \leq -\sigma/\xi$ si $\xi < 0$.

Cuando $\xi = 0$, el término de la derecha de la ecuación (1.16) se interpreta como

$$G(x; \sigma) = 1 - \exp\left\{-\frac{x}{\sigma}\right\}, \quad (1.17)$$

con parámetro de escala $\sigma > 0$.

Cabe destacar que la distribución de Pareto Generalizada con parámetro de forma nulo es la distribución Exponencial, con parámetro de escala $1/\sigma$.

El resultado asintótico para la distribución de excesos de un umbral se puede consultar en [16, 17], y se enuncia a continuación.

Teorema 1.9. (Balkema, de Haan, Pickands) Sea X_1, X_2, \dots, X_n una sucesión de variables aleatorias independientes con función de distribución común F , y sea F_{X_u} su función de distribución de excesos para el umbral $u > 0$. Para un valor del umbral u suficientemente alto, si F pertenece al dominio de atracción de una

distribución de Valores Extremos Generalizada con parámetro de forma $\xi \in \mathbb{R}$, entonces se verifica que

$$\lim_{u \rightarrow \infty} F_{X_u}(x) = GPD(x; \xi, \sigma), \quad (1.18)$$

para algún $\sigma > 0$.

Por otro lado, es importante mencionar que ambos métodos de valores extremos están relacionados a partir de las distribuciones asintóticas, tal y como se refleja en el siguiente resultado.

Proposición 1.10. Sean $H_\xi(x) = GEV(x; \xi, 1, 0)$ y $G_\xi(x) = GPD(x; \xi, 1)$, las funciones de distribución de valores extremos con el mismo parámetro de forma, $\xi \in \mathbb{R}$. Si $\ln H_\xi(x) > -1$, entonces se verifica que

$$G_\xi(x) = 1 + \ln H_\xi(x). \quad (1.19)$$

1.3. Modelos Jerárquicos Bayesianos

Los modelos Jerárquicos Bayesianos se utilizan frecuentemente para modelar observaciones que presentan características de dependencia espacial y/o temporal entre ellas. Estos modelos permiten involucrar otros modelos para las relaciones entre los parámetros de las distribuciones de las observaciones, introduciendo cierta jerarquía entre parámetros. Es decir, en los modelos jerárquicos se considera que las observaciones vienen dadas por distribuciones condicionadas a un conjunto de parámetros que pueden tener otras distribuciones con parámetros adicionales, denominados *hiperparámetros*.

Dar una definición formal de modelos jerárquicos suele ser complicado, a pesar de estar basados en una idea muy sencilla. Se trata de considerar a los parámetros del modelo $\theta_1, \dots, \theta_n$ como observaciones independientes de una distribución a priori dependiente de algún hiperparámetro φ , desconocido, y donde la distribución de las observaciones solo dependen de φ a través de los parámetros del modelo.

Una ventaja que presentan los modelos jerárquicos es la posibilidad de cuantificar la incertidumbre que existe en la estimación del modelo, ya que recoge la que presentan tanto los parámetros $\theta_1, \dots, \theta_n$ como el hiperparámetro φ .

En los modelos espacio-temporales, los modelos jerárquicos suelen dividirse en las siguientes tres etapas:

- Etapa 1. Modelo para las observaciones: $Y_j \sim F(Y_j; \theta_j | \varphi)$, $j = 1, \dots, n$.

- Etapa 2. Modelo para los parámetros: $\theta_j \sim \pi(\theta_j|\varphi)$, $j = 1, \dots, n$.
- Etapa 3. Modelo para los hiperparámetros: $\varphi \sim \pi(\varphi)$.

Es posible obtener la distribución a posteriori de los modelos gracias al teorema de Bayes, puesto que

$$\pi(\theta_1, \dots, \theta_n | Y_1, \dots, Y_n, \varphi) \propto \mathcal{L}(\theta_1, \dots, \theta_n | Y_1, \dots, Y_n, \varphi) \cdot \pi(\theta_1, \dots, \theta_n | \varphi) \cdot \pi(\varphi), \quad (1.20)$$

donde $\mathcal{L}(\theta_1, \dots, \theta_n | Y_1, \dots, Y_n, \varphi)$ es la función de verosimilitud de las variables Y_1, \dots, Y_n .

Este tipo de modelos permite una gran flexibilidad en el proceso de estimación de los parámetros, y para modelar y pronosticar observaciones de interés, como pueden ser los eventos extremos en climatología. Estos eventos presentan la característica especial de tener cierta dependencia espacial, es por ello necesario incluir modelos en los parámetros de las distribuciones de las observaciones climáticas en las que se recoja la dependencia de éstos con las coordenadas geográficas. Los modelos más sencillos consideran la distribución normal para los parámetros.

1.4. Introducción a la Teoría de Cópulas

El concepto de cópula permite modelar la dependencia espacial que existe entre los valores extremos.

Definición 1.11. *Para todo $n \geq 2$, una cópula n -dimensional es una función de distribución n -dimensional en $[0, 1]^n$ cuyas funciones de distribución marginales univariantes son distribuciones uniformes en $[0, 1]$.*

En la Teoría de Cópulas, existe un resultado de gran importancia que permite establecer una relación entre las distribuciones multivariantes y las cópulas, que se enuncia a continuación.

Teorema 1.12. (Sklar) *Sea F una función de distribución n -dimensional con funciones de distribución marginales univariantes F_1, F_2, \dots, F_n . Sea A_j el dominio de definición de F_j , para $j = 1, \dots, n$. Entonces, existe una cópula C tal que para todo $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$,*

$$F(x_1, x_2, \dots, x_n) = C(F_1(x_1), F_2(x_2), \dots, F_n(x_n)). \quad (1.21)$$

Si F_1, F_2, \dots, F_n son funciones de distribución continuas, entonces la cópula C es única. En otro caso, C está únicamente determinada en el conjunto $A_1 \times A_2 \times \dots \times A_n$.

Recíprocamente, si C es una cópula y F_1, F_2, \dots, F_n son funciones de distribución, entonces la función F definida por la ecuación (1.21) es una función de distribución n -dimensional con funciones de distribución marginales univariantes F_1, F_2, \dots, F_n .

Este resultado permite construir funciones de distribución multivariantes con las funciones de distribución marginales univariantes deseadas. Para ello, basta tomar una cópula que permita controlar la dependencia entre las variables aleatorias, y las funciones de distribución univariantes. Esto se muestra en el siguiente corolario. Los detalles pueden consultarse en [39].

Corolario 1.13. Sean F una función de distribución n -dimensional, F_1, F_2, \dots, F_n funciones de distribución univariantes, y C una cópula n -dimensional. Sean $F_1^{(-1)}, F_2^{(-1)}, \dots, F_n^{(-1)}$ las funciones quasi-inversas² de F_1, F_2, \dots, F_n , respectivamente. Entonces, para cualquier $(u_1, u_2, \dots, u_n) \in \text{Dom } C$,

$$C(u_1, u_2, \dots, u_n) = F\left(F_1^{(-1)}(u_1), F_2^{(-1)}(u_2), \dots, F_n^{(-1)}(u_n)\right) \quad (1.22)$$

1.5. Algoritmos para Inferencia Bayesiana

Se finaliza este primer capítulo exponiendo los principales métodos de generación de variables aleatorias con Cadenas de Markov que juegan un papel muy importante en la inferencia bayesiana.

Sea $f(x; \theta)$ la función de densidad de una variable aleatoria X de parámetro $\theta \in \Theta \subset \mathbb{R}^n$; $\mathcal{L}(\theta|x)$ la función de verosimilitud para θ y $\pi(\theta)$ la función de densidad de la distribución a priori de θ . El teorema de Bayes garantiza que la función de densidad de la distribución a posteriori de θ se puede escribir como

$$\pi(\theta|x) = \frac{\pi(\theta) \cdot \mathcal{L}(\theta|x)}{g(x)}, \quad (1.23)$$

donde

$$g(x) = \begin{cases} \int_{\Theta} \pi(\theta) \cdot \mathcal{L}(\theta|x) d\theta, & \text{en el caso continuo} \\ \sum_{\Theta} \pi(\theta) \cdot \mathcal{L}(\theta|x), & \text{en el caso discreto} \end{cases} \quad (1.24)$$

²Sea H una función de distribución. Se define su quasi-inversa $H^{(-1)}$ como la función $H^{(-1)}(u) = \inf\{t : H(t) \geq u\}$, $0 \leq u \leq 1$.

que no depende de θ , por lo que se cumple que

$$\pi(\theta|x) \propto \pi(\theta) \cdot \mathcal{L}(\theta|x). \quad (1.25)$$

Calcular $g(x)$ puede ser complicado o incluso imposible de tratar analíticamente, por lo que será necesario recurrir a simulaciones estocásticas. Para ello, resultan de gran utilidad las técnicas Monte Carlo basadas en Cadenas de Markov (MCMC) que permiten simular valores de distribuciones donde no podemos hacerlo directamente.

Cuando se desea generar una muestra de la distribución $\pi(x)$, con $x \in \mathcal{X} \subset \mathbb{R}^n$, pero no es posible hacerlo directamente, se puede considerar un proceso de Markov $p(\cdot, \cdot)$ en tiempo discreto, con espacio de estados \mathcal{X} del que sea sencillo muestrear, cuya distribución de equilibrio sea $\pi(x)$, y seguir la siguiente estrategia:

Escoger $X^{(0)}$ arbitrariamente, $i = 1$

Hasta que se juzgue convergencia,

generar $X^{(i)} \sim p(X^{(i-1)}, \cdot)$

hacer $i = i + 1$

Desde $j = 1$ hasta N ,

generar $X^{(i+j)} \sim p(X^{(i+j-1)}, \cdot)$

salir $X^{(i+j)}$

hacer $j = j + 1$

Su principal problema es cómo construir el proceso de Markov $p(\cdot, \cdot)$. A continuación, se describen dos de los algoritmos más populares: Metropolis-Hastings (MH) y Gibbs, tal como se han utilizado en la programación de las estrategias propuestas.

Algoritmo de Metropolis-Hastings

Para su implementación, solo es necesario poder evaluar puntualmente la distribución objetivo $\pi(\theta|x)$, salvo una constante, y generar observaciones candidatas de una distribución de prueba, $q(\cdot, \cdot)$, que satisfaga ciertas condiciones técnicas.

El algoritmo MH consiste en seguir los siguientes pasos:

1. $i=1$, inicializar $\theta^{(0)}$
2. Generar $\theta^{(can)} \sim q(\theta^{(can)}, \theta^{(i-1)})$

3. Evaluar la probabilidad de aceptación³

$$p(\theta^{(can)}, \theta^{(i-1)}) = \min \left\{ \frac{\pi(\theta^{(can)} | \theta^{(i-1)}) q(\theta^{(can)}, \theta^{(i-1)})}{\pi(\theta^{(i-1)} | \theta^{(can)}) q(\theta^{(i-1)}, \theta^{(can)})}, 1 \right\} \quad (1.26)$$

4. Generar $u \sim \mathcal{U}(0, 1)$

5. Si $u < p(\theta^{(can)}, \theta^{(i-1)})$ hacer $\theta^{(i)} = \theta^{(can)}$ y en caso contrario hacer $\theta^{(i)} = \theta^{(i-1)}$

6. $i=i+1$ y volver al paso 2

El problema está en escoger q , que en principio es arbitraria siempre que asegure la convergencia del método [21]. Si q tiene un comportamiento similar a $\pi(\theta|x)$, entonces la probabilidad de aceptación será mayor, y por tanto, también lo será la proporción de candidatos aceptados de la distribución de interés.

Debe ser una función de fácil generación (salvo constante), y suele escogerse simétrica para simplificar el cálculo de la probabilidad de aceptación que se reduce a

$$p(\theta^{(can)}, \theta^{(i-1)}) = \min \left\{ \frac{\pi(\theta^{(can)} | \theta^{(i-1)})}{\pi(\theta^{(i-1)} | \theta^{(can)})}, 1 \right\}. \quad (1.27)$$

Finalmente, si las distribuciones a priori son altamente informativas, entonces $\pi(\theta^{(can)}) \simeq \pi(\theta^{(i-1)})$, y la probabilidad de aceptación se reduce a

$$p(\theta^{(can)}, \theta^{(i-1)}) = \min \left\{ \frac{\mathcal{L}(\theta^{(can)} | \mathbf{x})}{\mathcal{L}(\theta^{(i-1)} | \mathbf{x})}, 1 \right\}. \quad (1.28)$$

Muestreador de Gibbs

Cuando se desea muestrear un conjunto de parámetros $\theta = (\theta_1, \dots, \theta_s)$, cuya distribución no es sencilla de simular, pero existen algoritmos eficientes para simular de la distribución a posteriori condicional, $\pi(\theta_j | \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_s)$, para todo $j = 1, \dots, s$, entonces se emplea el algoritmo del muestreador de Gibbs.

También puede verse como un caso particular del algoritmo MH en el que las distribuciones que se proponen coinciden con las distribuciones a posteriori condicionadas, y por tanto la probabilidad de aceptación siempre es uno.

El muestreador de Gibbs consiste en los siguientes pasos:

³La probabilidad se suele calcular en términos logarítmicos para evitar problemas computacionales.

1. $i=1$, inicializar $(\theta_1^{(0)}, \dots, \theta_s^{(0)})$
2. Generar $\theta_1^{(i+1)} \sim \pi(\theta_1 | \theta_2^{(i-1)}, \dots, \theta_s^{(i-1)})$
3. Generar $\theta_2^{(i+1)} \sim \pi(\theta_2 | \theta_1^{(i-1)}, \theta_3^{(i-1)}, \dots, \theta_s^{(i-1)})$
4. ...
5. Generar $\theta_s^{(i+1)} \sim \pi(\theta_s | \theta_1^{(i-1)}, \dots, \theta_{s-1}^{(i-1)})$
6. $i=i+1$ y volver al paso 2

Capítulo 2

Artículo A: Métodos Base de inferencia Bayesiana en la distribución Gumbel

Autores:

Jacinto Martín, M. Isabel Parra, Mario M. Pizarro, Eva L. Sanjuán

Departamento de Matemáticas, Universidad de Extremadura

Revista: Entropy, 22(11), 1267, 2020

DOI: [10.3390/e22111267](https://doi.org/10.3390/e22111267)

Resumen:





En este trabajo se desarrolla un modelo para hacer uso de todos los datos disponibles en un marco de valores extremos. La clave es aprovechar la relación existente entre los parámetros de la distribución base y los parámetros de la distribución de máximos de bloque. En particular, se proponen dos métodos de estimación Bayesiana, el método Baseline distribution (BDM) que permite estimar los parámetros de la distribución base con todos los datos y al realizarles una transformación se pueden calcular estimaciones para los parámetros de la distribución de máximos de bloque. El método Improved Baseline distribution (IBDM) asigna una mayor importancia a los datos máximos de bloque que a los valores base, y permite obtener los parámetros de la distribución de máximos de bloque con distribuciones a priori altamente informativas construidas a partir de las relaciones existentes con los parámetros de la distribución base. Finalmente, se comparan empíricamente, mediante un amplio estudio de simulación, ambos métodos con el método

Bayesiano estándar con a priori no informativas, considerando tres distribuciones base (Gumbel, Exponencial y Normal) que conducen a una distribución de valores extremos Gumbel. Se comprueba empíricamente que las estimaciones son más exactas y precisas, especialmente para tamaños muestrales pequeños, con los métodos propuestos.



Article

Baseline Methods for Bayesian Inference in Gumbel Distribution

Jacinto Martín ^{1,†} , María Isabel Parra ^{1,†} , Mario Martínez Pizarro ^{2,†}  and Eva L. Sanjuán ^{3,*,†} 

¹ Departamento de Matemáticas, Facultad de Ciencias, Universidad de Extremadura, 06006 Badajoz, Spain; jmartin@unex.es (J.M.); mipa@unex.es (M.I.P.)

² Departamento de Matemáticas, Facultad de Veterinaria, Universidad de Extremadura, 10003 Cáceres, Spain; mariomp@unex.es

³ Departamento de Matemáticas, Centro Universitario de Mérida, Universidad de Extremadura, 06800 Mérida, Spain

* Correspondence: etlopez@unex.es

† The authors contributed equally to this work.

Received: 28 September 2020; Accepted: 4 November 2020; Published: 7 November 2020



Abstract: Usual estimation methods for the parameters of extreme value distributions only employ a small part of the observation values. When block maxima values are considered, many data are discarded, and therefore a lot of information is wasted. We develop a model to seize the whole data available in an extreme value framework. The key is to take advantage of the existing relation between the baseline parameters and the parameters of the block maxima distribution. We propose two methods to perform Bayesian estimation. Baseline distribution method (BDM) consists in computing estimations for the baseline parameters with all the data, and then making a transformation to compute estimations for the block maxima parameters. Improved baseline method (IBDM) is a refinement of the initial idea, with the aim of assigning more importance to the block maxima data than to the baseline values, performed by applying BDM to develop an improved prior distribution. We compare empirically these new methods with the Standard Bayesian analysis with non-informative prior, considering three baseline distributions that lead to a Gumbel extreme distribution, namely Gumbel, Exponential and Normal, by a broad simulation study.

Keywords: Bayesian inference; highly informative prior; Gumbel distribution; small dataset

1. Introduction

Extreme value theory (EVT) is a widely used statistical tool for modeling and forecasting the distributions which arise when we study events that are more extreme than any previously observed. Examples of these situations are natural rare events in climatology or hydrology, such as floods, earthquakes, climate changes, etc. Therefore, EVT is employed in several scientific fields, to model and predict extreme events of temperature [1–3], precipitation [4–10] and solar climatology [11–13], as well as in the engineering industry to study important malfunctions (e.g., [14], finance (study of financial crisis), insurance (for very large claims due to catastrophic events) and environmental science (concentration of pollution in the air).

Overall fitting method tries to fit all the historical data to several theoretical distributions and then choose the best one, according to certain criteria. However, since the number of extreme observations is usually scarce, overall fitting works well in the central distribution area, but it can poorly represent the tail area.

In EVT, there are two main approaches, block maxima (BM) method and peak-over-threshold (POT), which are differentiated by the way each model classifies the observations considered as extreme events and then uses them in the data analysis process. In the POT method, the extreme data are the ones over a certain threshold, while, for BM method, data are divided into blocks of equal size, and the maximum datum for each block is selected. BM method is preferable to POT when the only available information is block data, when seasonal periodicity is given or the block periods appear naturally, such as in studies for temperature, precipitation and solar climatology. The biggest challenge in BM method is deciding the size of the blocks when they are not obvious.

The Gumbel distribution plays an important role in extreme value analysis as model for block maxima, because it is appropriate to describe extreme events from distributions such as Normal, Exponential and Gumbel distributions [15]. To estimate maximum data distribution, both frequentist and Bayesian approaches have been developed [16,17]. However, the knowledge of physical constraints, the historical evidence of data behavior or previous assessments might be an extremely important matter for the adjustment of the data, particularly when they are not completely representative and further information is required. This fact leads to the use of Bayesian inference to address the extreme value estimation [18].

Practical use of Bayesian estimation is often associated with difficulties to choose prior information and prior distribution for the parameters of the extreme values distribution [19]. To fix this problem, several alternatives have been proposed, either by focusing exclusively on the selection of the prior density [20,21] or by improving the algorithm for the estimation of the parameters [22]. However, the lack of information still seems to be the weakness when referring to extreme value inference.

Examples for its application are the modeling of annual rainfall maximum intensities [23], the estimation of the probability of exceedence of future flood discharge [24] and the forecasting of the extremes of the price distribution [25]. Some of these works are focused on the construction of informative priors of the parameters for which data can provide little information. Despite these previous efforts, it is well understood that some constraints to quantify qualitative knowledge always appear when referring to construct informative priors.

Therefore, this paper focuses on techniques to employ all the available data in order to elicit a highly informative prior distribution. We consider several distributions that lead to a Gumbel extreme distribution. The key is to take advantage of the existing relation between the baseline parameters and the parameters of the block maxima distribution. The use of the entire dataset, instead of the selected block maximum data, results to be adequate and it is advisable when dealing with very shortened available data.

We employ MCMC techniques, concretely a Metropolis–Hastings algorithm. Several statistical analyses are performed to test the validity of our method and check its enhancements in relation to the standard Bayesian analysis without this information.

2. Domains of Attraction of Gumbel Distribution

As is well known, the BM approach consists on dividing the observation period into non overlapping periods of equal size and select the maximum observation in each period. Given a sequence of i.i.d. random variables Y_1, Y_2, \dots, Y_m with common distribution function F , and given a fixed $k \in \mathbb{N}$ (block size), we define the block maxima

$$X_i = \max_{(i-1)k < j \leq ik} Y_j, \quad i = 1, 2, \dots, n. \quad (1)$$

Hence, the total observations, $m = k \times n$, are divided into n blocks of size k . The extreme values depend upon the full sample space from which they have been drawn through its shape and size. Therefore, extremes variate according to the initial distribution and sample size ([26]). Then, the cumulative distribution function of X_i is

$$P(X_i \leq x) = P(Y_{(i-1)k+1} \leq x, \dots, Y_{ik} \leq x) = \prod_{j=(i-1)k+1}^{ik} P(Y_j \leq x) = F(x)^k \tag{2}$$

This result depends on our knowledge of F , in which we could be lacking. Therefore, it is useful to consider the asymptotic distribution.

According to the Gnedenko [27] and Fisher and Tippett [28] theorems, the asymptotic distribution of block maxima of random i.i.d. variables can be approximated by a generalized extreme value distribution, with distribution function

$$GEV(x; \xi, \mu, \sigma) = \exp \left\{ - \left(1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right)^{-1/\xi} \right\} \tag{3}$$

with $\xi, \mu \in \mathbb{R}$, $\sigma > 0$ and $1 + \xi \left(\frac{x - \mu}{\sigma} \right) > 0$.

When $\xi = 0$, the right-hand side of Equation (3) is interpreted as

$$G(x; \mu, \sigma) = \exp \left\{ - \exp \left\{ - \left(\frac{x - \mu}{\sigma} \right) \right\} \right\} \tag{4}$$

and it is called Gumbel distribution with parameters μ (location) and $\sigma > 0$ (scale).

Definition 1. We say that the distribution function F is in the domain of attraction of a extreme value Gumbel distribution when there exist sequences $\{a_k\}$ and $\{b_k\}$, with $a_k > 0, b_k \in \mathbb{R}$ such that

$$\lim_{k \rightarrow \infty} F^k(a_k x + b_k) = G(x), \quad x \in \mathbb{R} \tag{5}$$

Sequences $\{a_k\}$ and $\{b_k\}$ are called normalizing constants. We usually call the distribution F baseline or underlying distribution. Normalizing constants correspond to the parameters of scale and location of the limit Gumbel distribution, therefore they allow us to establish a relation between this distribution and the baseline distribution.

Moreover, Ferreira and de Haan [29] showed theoretical results which allow determining the normalizing constants for many baseline distributions in the domain of attraction of a Gumbel distribution:

Theorem 1. When F belongs to the domain of attraction of a Gumbel distribution, there is a positive function h that verifies

$$a_k = h(b_k), \quad b_k = F^{-1}(1 - k^{-1}), \quad \forall k. \tag{6}$$

To determine function h , the following condition is very useful.

Theorem 2 (Von-Mises condition). When $F''(x)$ and $F'(x)$ exist, and F' is positive for all x belonging to a neighborhood at the left of x^* (right endpoint of F), and

$$\lim_{t \rightarrow x^*} \left(\frac{1 - F}{F'} \right)'(t) = 0, \tag{7}$$

or equivalently

$$\lim_{t \rightarrow x^*} \frac{(1 - F(t)) \cdot F''(t)}{(F'(t))^2} = -1, \tag{8}$$

then F belongs to the domain of attraction of the Gumbel distribution. In this case, function h is determined by

$$h(t) = \frac{1 - F(t)}{F'(t)} \tag{9}$$

Distributions whose tails decrease exponentially produce a Gumbel distribution when taking the block maxima. Besides the Exponential distribution, the class of distributions which belong to the domain of attraction of the Gumbel includes the Normal distribution, and many others, such as Log-normal, Gamma, Rayleigh, Gumbel, etc.

We also use the following result:

Proposition 1. *If X is a random variable belonging to the domain of attraction of a Gumbel distribution, then $Y = \mu + \sigma X$ also belongs to the same domain of attraction. The normalization constants are:*

$$\tilde{a}_k = \sigma a_k, \quad \tilde{b}_k = \mu + \sigma b_k. \quad (10)$$

where a_k and b_k are the normalization constants of X .

2.1. Gumbel Baseline Distribution

If $Y \sim \mathcal{G}(\mu, \sigma)$, then block maxima distribution of size k , denoted by X , is also a Gumbel distribution, because

$$F(x)^k = \exp \left\{ -\exp \left(-\frac{x - \mu}{\sigma} \right) + k \right\} = \exp \left\{ -\exp \left(-\frac{x - (\mu + \sigma \ln k)}{\sigma} \right) \right\}, \quad (11)$$

therefore $X \sim \mathcal{G}(\mu + \sigma \ln k, \sigma)$.

2.2. Exponential Baseline Distribution

Let $Y \sim \text{Exp}(\lambda)$ with distribution function

$$F(y) = 1 - e^{-\lambda y}, \quad y \geq 0, \quad (12)$$

Exponential distribution belongs to the domain of attraction of the Gumbel, with $h(t) = \lambda^{-1}$. As $F^{-1}(u) = \lambda^{-1} \ln(1 - u)$, the normalization constants are

$$a_k = \lambda^{-1}, \quad b_k = \lambda^{-1} \ln k, \quad (13)$$

and they settle a relation that allow us to make estimations for Gumbel limit distribution, when there is an exponential baseline distribution for k big enough.

2.3. Normal Baseline Distribution

When the baseline distribution is a Standard Normal distribution, normalizing constants can be computed, making use of asymptotic limit and results showed before.

Let $\mathcal{Z} \sim \mathcal{N}(0, 1)$, with distribution function F and density function f . It is easy to show that F verifies von Mises condition (8):

$$\begin{aligned} \lim_{t \rightarrow x^*} \frac{(1 - F(t)) \cdot F''(t)}{(F'(t))^2} &= \lim_{t \rightarrow x^*} \frac{-(1 - F(t)) \cdot f(t)t}{(f(t))^2} = \lim_{t \rightarrow x^*} \frac{-(1 - F(t)) \cdot t}{f(t)} \\ &= \lim_{t \rightarrow x^*} \frac{-(1 - F(t)) + f(t) \cdot t}{f'(t)} = \lim_{t \rightarrow x^*} \frac{f(t) \cdot t}{-t \cdot f(t)} = -1 \end{aligned}$$

using L'Hôpital and noticing that $f'(t) = -t \cdot f(t)$. Therefore, $1 - F(t) \approx f(t) \cdot t^{-1}$, and, consequently, the function h defined as (9) verifies

$$\lim_{t \rightarrow x^*} h(t) = t^{-1}$$

Besides, by (6), $F(b_k) = 1 - k^{-1}$. Therefore, $\ln(1 - F(b_k)) = -\ln k$, or

$$\ln f(b_k) - \ln b_k = -\ln k,$$

so

$$b_k^2 + \ln(2\pi) + 2\ln b_k = 2\ln k. \tag{14}$$

Defining the function $g(b_k) = b_k^2 + \ln(2\pi) + 2\ln b_k - 2\ln k$, and developing its Taylor series around $(2\ln k)^{1/2}$, we obtain

$$\begin{aligned} g(b_k) &= g\left((2\ln k)^{1/2}\right) + g'\left((2\ln k)^{1/2}\right) \cdot \left(b_k - (2\ln k)^{1/2}\right) + \mathcal{O}\left((2\ln k)^{1/2}\right) \\ &= [\ln \ln k + \ln 4\pi] + 2\left[(2\ln k)^{1/2} + (2\ln k)^{-1/2}\right] \cdot \left(b_k - (2\ln k)^{1/2}\right) \\ &\quad + \mathcal{O}\left((2\ln k)^{1/2}\right), \end{aligned}$$

so, as $g(b_k) = 0$, for k big enough

$$b_k = (2\ln k)^{1/2} - 2^{-1} (2\ln k)^{-1/2} [\ln(\ln k) + \ln(4\pi)]. \tag{15}$$

In addition, as $a_k = h(b_k) \approx b_k^{-1}$ and

$$b_k^{-1} = \frac{1}{(2\ln k)^{1/2} - 2^{-1} (2\ln k)^{-1/2} [\ln(\ln k) + \ln(4\pi)]} \approx (2\ln k)^{-1/2} \tag{16}$$

a_k can be taken as

$$a_k \approx (2\ln k)^{-1/2}. \tag{17}$$

Besides, as a consequence of Theorem 10, if $Y \sim \mathcal{N}(\mu, \sigma)$, for k big enough, the normalization constants are, approximately,

$$a_k = \sigma (2\ln k)^{-1/2}, \quad b_k = \mu + \sigma \left[(2\ln k)^{1/2} - 2^{-1} (2\ln k)^{-1/2} [\ln \ln k + \ln(4\pi)] \right]. \tag{18}$$

2.4. Other Baseline Distributions

This way of working can be extended to other baseline distributions, whose block maxima limit is also a Gumbel, by using existing relations between baseline and limit parameters. In Table 1, normalization constants computed for the most employed distribution functions in the domain of attraction of the Gumbel distribution are shown. Constants a_N and b_N are the normalization constants for Standard Normal distribution, given by (15) and (17), respectively.

Table 1. Normalization constants computed for the most employed distribution function.

Baseline Distribution F	a_k	b_k
Exponential (λ)	λ^{-1}	$\lambda^{-1} \ln k$
Gamma (α, β)	β	$\beta [\ln k + (\alpha - 1) \ln \ln k - \ln \Gamma(\alpha)]$
Gumbel (μ, σ)	σ	$\mu + \sigma \ln k$
Log-Normal (μ, σ)	$a_N \sigma e^{\mu + \sigma b_N}$	$e^{\mu + \sigma b_N}$
Normal (μ, σ)	$\sigma (2\ln k)^{-1/2}$	$\mu + \sigma \left[(2\ln k)^{1/2} - \frac{\ln \ln k + \ln 4\pi}{2(2\ln k)^{1/2}} \right]$
Rayleigh (σ)	$\sigma (2\ln k)^{-1/2}$	$\sigma (2\ln k)^{1/2}$

3. Bayesian Estimation Methods

3.1. Classical Bayesian Estimation for the Gumbel Distribution

To make statistical inferences based on the Bayesian framework, after assuming a prior density for the parameters, $\pi(\theta)$, and combining this distribution with the information brought by the data which are quantified by the likelihood function, $L(\theta|x)$, the posterior density function of the parameters can be determined as

$$\pi(\theta|x) \propto L(\theta|x)\pi(\theta) \quad (19)$$

The remaining of the inference process is fulfilled based on the obtained posterior distribution.

The likelihood function for $\theta = (\mu, \sigma)$, given the random sample $\mathbf{x} = (x_1, \dots, x_n)$ from a Gumbel(μ, σ) distribution, with density function given by

$$f(x|\mu, \sigma) = \frac{1}{\sigma} \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma}\right) \quad (20)$$

where $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*$, is

$$L(\mu, \sigma|\mathbf{x}) = \frac{1}{\sigma^n} \exp(\Delta), \quad (21)$$

with

$$\Delta = -\sum_{i=1}^n \exp\left(-\frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^n \frac{x_i - \mu}{\sigma}. \quad (22)$$

In the case of the Gumbel distribution, Rostami and Adam [21] selected eighteen pairs of priors based on the parameters' characteristics, assumed independence, and compared the posterior estimations by applying Metropolis–Hastings (MH) algorithm, concluding that the combination of Gumbel and Rayleigh is the most productive pair of priors for this model. For fixed initial hyper-parameters $\mu_0, \sigma_0, \lambda_0$

$$\begin{aligned} \pi(\mu) &\propto \exp\left(-\exp\left(-\frac{\mu - \mu_0}{\sigma_0}\right) - \frac{\mu - \mu_0}{\sigma_0}\right) \\ \pi(\sigma) &\propto \sigma \exp\left(-\frac{\sigma^2}{2\lambda_0^2}\right). \end{aligned} \quad (23)$$

The posterior distribution is

$$\pi(\mu, \sigma|\mathbf{x}) \propto \frac{1}{\sigma^{n-1}} \exp\left(\Delta - \exp\left(-\frac{\mu - \mu_0}{\sigma_0}\right) - \frac{\mu - \mu_0}{\sigma_0} - \frac{\sigma^2}{2\lambda_0^2}\right), \quad (24)$$

and conditional posterior distributions are

$$\begin{aligned} \pi(\mu|\sigma, \mathbf{x}) &\propto \exp\left(\Delta - \exp\left(-\frac{\mu - \mu_0}{\sigma_0}\right) - \frac{\mu - \mu_0}{\sigma_0}\right) \\ \pi(\sigma|\mu, \mathbf{x}) &\propto \frac{1}{\sigma^{n-1}} \exp\left(\Delta - \frac{\sigma^2}{2\lambda_0^2}\right) \end{aligned} \quad (25)$$

Then, an MCMC method is applied through the MH algorithm.

1. Draw a starting sample $(\mu^{(0)}, \sigma^{(0)})$ from starting distributions, $\pi(\mu), \pi(\sigma)$, respectively, given by Equation (23).
2. For $j = 0, 1, \dots$, given the chain is currently at $\mu^{(j)}, \sigma^{(j)}$,

- Sample candidates μ^*, σ^* for the next sample from a proposal distribution,

$$\mu^* \sim \mathcal{N}(\mu^{(j)}, v_\mu) \text{ and } \sigma^* \sim \mathcal{N}(\sigma^{(j)}, v_\sigma)$$

- Calculate the ratios

$$r_\mu = \frac{\pi(\mu^* | \sigma^{(j)}, \mathbf{x})}{\pi(\mu^{(j)} | \sigma^{(j)}, \mathbf{x})}, \quad r_\sigma = \frac{\pi(\sigma^* | \mu^{(j)}, \mathbf{x})}{\pi(\sigma^{(j)} | \mu^{(j)}, \mathbf{x})} \tag{26}$$

- Set

$$\mu^{(j+1)} = \begin{cases} \mu^*, & \text{with probability } \min\{1, r_\mu\} \\ \mu^{(j)}, & \text{otherwise} \end{cases} \tag{27}$$

$$\sigma^{(j+1)} = \begin{cases} \sigma^*, & \text{with probability } \min\{1, r_\sigma\} \\ \sigma^{(j)}, & \text{otherwise} \end{cases} \tag{28}$$

3. Iterate the former procedure. Notice that

$$\begin{aligned} r_\mu &= \exp \left\{ \frac{n}{\sigma^{(j)}} (\mu^* - \mu^{(j)}) + \frac{\mu^{(j)} - \mu^*}{\sigma_0} + \exp \left(-\frac{\mu^{(j)} - \mu_0}{\sigma_0} \right) - \exp \left(-\frac{\mu^* - \mu_0}{\sigma_0} \right) \right. \\ &\quad \left. + \sum_{i=1}^n \left[\exp \left(-\frac{x_i - \mu^{(j)}}{\sigma^{(j)}} \right) - \exp \left(-\frac{x_i - \mu^*}{\sigma^{(j)}} \right) \right] \right\} \\ r_\sigma &= \left(\frac{\sigma^{(j)}}{\sigma^*} \right)^{n-1} \exp \left\{ \frac{(\sigma^{(j)})^2 - (\sigma^*)^2}{2\lambda_0^2} + \left(\frac{1}{\sigma^{(j)}} - \frac{1}{\sigma^*} \right) \sum_{i=1}^n (x_i - \mu^{(j)}) \right. \\ &\quad \left. + \sum_{i=1}^n \left[\exp \left(-\frac{x_i - \mu^{(j)}}{\sigma^{(j)}} \right) - \exp \left(-\frac{x_i - \mu^{(j)}}{\sigma^*} \right) \right] \right\}. \end{aligned}$$

Therefore, we obtain a Markov chain that converges to the posterior distributions for the parameters μ and σ . We call this method Classical Metropolis–Hastings method (MHM).

3.2. Baseline Distribution Method

In Baseline distribution method (BDM), we take all the information available to determine posterior baseline distribution. We denote $\mathcal{B}(\theta)$ as the baseline distribution with parameter vector θ .

Then, we can apply Bayesian inference procedures to estimate the posterior distribution of the baseline distribution, denoted by $\pi_b(\theta | \mathbf{y})$ and, therefore, to obtain estimations for the parameters of the baseline distribution θ , with all the data provided by \mathbf{y} .

Afterwards, making the transformation given by the relations we obtained in previous section, we can obtain new estimations for the parameters of block maxima distribution, which is the Gumbel in this case. We explain the procedure for the three baseline distributions considered in this paper: Gumbel, Exponential and Normal distribution.

3.2.1. Gumbel Baseline Distribution

When the baseline distribution $Y \sim \mathcal{G}(\mu_b, \sigma_b)$, it is known that the limit distribution $X \sim \mathcal{G}(\mu_b + \sigma_b \ln k, \sigma_b)$. Therefore, MH algorithm can be applied to the whole dataset, \mathbf{y} , to find estimations for μ_b and σ_b . Afterwards, we make the adequate transformation to compute estimations for the parameters of X .

3.2.2. Exponential Baseline Distribution

When the baseline distribution $Y \sim \text{Exp}(\lambda_b)$, we consider a Gamma distribution with parameters α_0 and β_0 as prior distribution

$$\pi(\lambda_b) \propto \lambda_b^{\alpha_0 - 1} \exp(-\beta_0 \lambda_b). \tag{29}$$

Therefore, the posterior distribution is

$$\pi(\lambda_b | \mathbf{y}) \sim \Gamma\left(\alpha_0 + m, \beta_0 + \sum_{j=1}^m y_j\right), \quad (30)$$

thus Gibbs algorithm can be employed to generate samples of posterior distribution $\pi(\lambda_b | \mathbf{y})$. Once the estimation of λ_b is obtained, k th power of the distribution function will be the estimation for block maxima distribution function of size k .

3.2.3. Normal Baseline Distribution

Finally, when the baseline distribution $Y \sim \mathcal{N}(\mu_b, \sigma_b)$, we employ Normal and inverse Gamma prior distributions

$$\begin{aligned} \pi(\mu_b) &\propto \exp\left(-\frac{\sigma_0}{2\sigma_b^2}(\mu - \mu_0)^2\right), \\ \pi(\sigma_b^2) &\propto \left(\frac{1}{\sigma_b^2}\right)^{\alpha_0-1} \exp\left(-\frac{\beta_0}{\sigma_b^2}\right). \end{aligned} \quad (31)$$

Therefore, posterior distributions are

$$\begin{aligned} \pi(\mu_b | \mathbf{y}, \sigma_b^2) &\sim \mathcal{N}\left(\frac{\sigma_0 \mu_0 + \sum_{j=1}^m y_j}{\sigma_0 + m}, \frac{\sigma_b^2}{\sigma_0 + m}\right), \\ \pi(\sigma_b^2 | \mathbf{y}, \mu_b) &\sim \text{Inv}\Gamma\left(\frac{m}{2} + \alpha_0, \beta_0 + \frac{1}{2} \sum_{j=1}^m (y_j - \mu_b)^2\right), \end{aligned} \quad (32)$$

and we can employ Gibbs algorithm to generate samples of posterior distribution, and, afterwards, the k th power of the distribution function, as in the previous case.

3.3. Improved Baseline Distribution Method

Finally, we propose a new method, called Improved Baseline distribution method (IBDM), to import the highly informative baseline parameters into the Bayesian inference procedure. Here, we take into account the spirit of classical EVT, which grants more weight to block maxima data than to baseline data.

The method consists on applying BDM to obtain the posterior distribution for the parameters of the baseline distribution $\pi(\theta | \mathbf{y})$, and then uses it to build a prior distribution for the parameters of the Gumbel. Therefore, we have a highly informative prior distribution.

As the priors are highly informative, $\pi(\theta^*) = \pi(\theta^{(j)})$, the ratio in the j th step of MH algorithm is

$$r_\theta = \frac{L(\mu^*, \sigma^* | \mathbf{x})}{L(\mu^{(j)}, \sigma^{(j)} | \mathbf{x})} = \left(\frac{\sigma^{(j)}}{\sigma^*}\right)^n \exp\left\{\sum_{i=1}^n \left[\exp\left(-\frac{x_i - \mu^{(j)}}{\sigma^{(j)}}\right) - \exp\left(-\frac{x_i - \mu^*}{\sigma^*}\right) + \frac{x_i - \mu^{(j)}}{\sigma^{(j)}} - \frac{x_i - \mu^*}{\sigma^*}\right]\right\}.$$

For every iteration of the algorithm, we first make an iteration of Baseline Distribution method, resulting θ_b as estimation of the posterior distribution $\pi(\theta_b | \mathbf{y})$. Afterwards, a candidate θ^* is generated using a Normal distribution $\mathcal{N}(f(\theta_b), \nu_\theta)$ with the adequate transformation $f(\theta_b)$, given by Equations (11), (13) and (18) in the case of Gumbel, Exponential or Normal baseline distributions, respectively.

Obviously, this method is quite similar to BDM when block size is big and, consequently, there are few maxima data. It approaches the classical Bayesian method as the block size gets smaller (more maxima data).

4. Simulation Study

We made a simulation study for the three distributions analyzed above, which belong to the domain of attraction of the Gumbel distribution: Gumbel, Exponential and Normal.

For each distribution selected (once its parameters are fixed), we generated $m_{ij} = n_i \times k_j$ values, where

- n_i is the number of block maxima, $n_i = 2^i, i = 1, 2, \dots, 7$; and
- k_j is the block size, $k_j = 10^j, j = 1, 2, 3$.

Therefore, the sample sizes vary from 20 to 128,000. Besides, each sequence is replicated 100 times. Consequently, 2100 sequences of random values were generated for each combination of parameters of each baseline distribution.

To guarantee the convergence of the MCMC algorithm, we must be sure that the posterior distribution has been reached. Some proceedings are advisable to be performed.

- Burn-in period: Eliminate the first generated values.
- Take different initial values and select them for each sample.
- Make a thinning to assure lack of autocorrelation.

These proceedings were made using library coda [30] for R software, taking 3000 values for the burn-in period, 50 values for the thinning and selecting initial values for each sample. Finally, to get the posterior distribution for each parameter, a Markov chain of length 10,000 was obtained. Therefore, 53,000 iterations were made for each sequence.

There are some common features for the baseline distributions considered when comparing the three methods MHM, BDM and IBDM.

1. To choose an estimator for the parameters, we compared mean- and median-based estimations. They were reasonably similar, due to the high symmetry of posterior distributions. Therefore, we chose the mean of the posterior distribution to make estimations of the parameters.
2. MHM usually provides high skewed estimations for the posterior distributions. BDM is the method that shows less skewness.
3. BDM is the method that offers estimations for posterior distribution with less variability. IBDM provides higher variability, but we must keep in mind that this method stresses the importance of extreme values, therefore more variability is expectable than the one provided by BDM. The method with highest variability is MHM.
4. The election of the most suitable method also depends on the characteristics of the problem. When block maxima data are very similar to the baseline distribution, BDM provides the best estimations and the lowest measures of error. On the contrary, when extreme data differ from baseline data, IBDM offers the lowest errors. IBDM is the most stable method: regardless of the differences between extreme data and baseline data, it provides reasonably good measures of error.

4.1. Gumbel Baseline Distribution

We considered the baseline $\mathcal{G}(\mu_b, \sigma_b)$ distribution. As the localization parameter has no effect on the variability of data, its value was fixed as $\mu_b = 0$ for easiness. Scale parameter does affect variability, so we considered values $\sigma_b = 2^j, j = -2, -1, 0, 1, 2$.

One important point of the simulation study is to observe how the estimation of the parameters vary for a fixed block size as the number of block maxima n (the amount of information we have) is

changing. Regardless of the chosen method, as n increases, variability and skewness decreases. However, for small values of n , BDM and IBDM provide more concentrated and less skewed distributions than the ones offered by MHM. We can appreciate this in Figure 1, where the probability density functions (pdf) are shown for the 100 estimations of the mean for the parameters μ (left) and σ (right) for block maxima Gumbel distributions, with the three methods. The baseline distribution is $\mathcal{G}(0, 4)$ and fixed block size $k = 1000$. Therefore, block maxima distribution is $\mathcal{G}(27.63, 4)$ (from (11)). Scales are very different in this charts, due to a better visualization of distributions. For example, for MHM, highest value of the pdf for $\hat{\sigma}$ is around 1.5 (for $n = 128$), but it is over 40 for BDM.

This behavior is shown qualitatively for all the values of the parameters employed in the simulation.

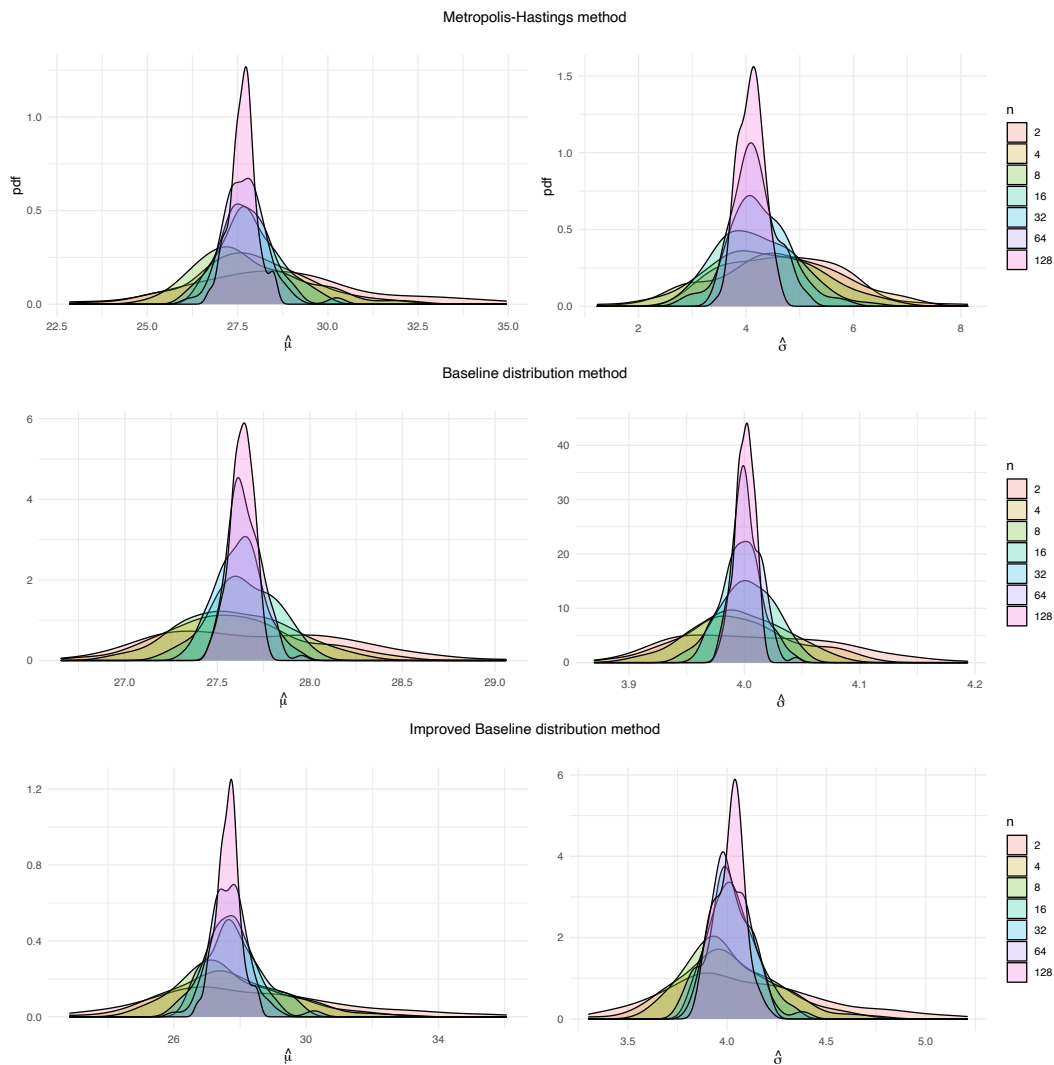


Figure 1. Probability density functions for 100 estimations of the block maxima parameters μ (left) and σ (right), obtained for the three methods, with $k = 1000$ and different values of n , from $\mathcal{G}(0, 4)$ baseline distribution.

To compute measures of error, in the case of Gumbel baseline distribution, we can employ absolute errors $AE_i = |\hat{\theta}_i - \theta|$, where $\hat{\theta}_i$ is the estimation obtained from i th sample and θ is the real value of the estimated parameter. We can then define

- Mean error:

$$ME = \frac{1}{M} \sum_{i=1}^M (\hat{\theta}_i - \theta).$$

- Root mean square error:

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M (\hat{\theta}_i - \theta)^2}.$$

- Mean absolute error:

$$MAE = \frac{1}{M} \sum_{i=1}^M |\hat{\theta}_i - \theta|,$$

where M is the number of samples.

The three methods provide estimations with low absolute errors AE when the number of maxima n is high, and especially when block size k is high. When both numbers are small, BDM and IBDM get closer estimations and differ from MHM.

In Tables 2 and 3, we show values for ME and RMSE for the estimations of parameters μ and σ , respectively, for some values of k , n and σ_b , for a $\mathcal{G}(0, \sigma_b)$ baseline distribution. We can see that BDM is the method that offers lower values for both measures of error, followed by IBDM. The method that provides highest errors is MHM.

Table 2. ME for μ , with RMSE in brackets, for a baseline distribution $\mathcal{G}(0, \sigma_b)$.

	k									σ_b
	10			100			1000			
n	MHM	BDM	IBDM	MHM	BDM	IBDM	MHM	BDM	IBDM	
2	-0.0879	0.0752	-0.0038	-0.1056	0.0044	-0.0450	-0.0495	0.0061	0.0159	1/4
	(0.2290)	(0.1594)	(0.1848)	(0.2136)	(0.0635)	(0.1658)	(0.1659)	(0.0343)	(0.1782)	
	0.3128	0.3376	0.3350	0.3326	0.1017	0.2866	0.1578	0.0354	0.1137	1
	(0.9906)	(0.6637)	(0.7905)	(0.8535)	(0.3046)	(0.6906)	(0.6573)	(0.1464)	(0.6850)	4
	1.6813	1.0156	1.3856	0.2687	-0.0001	0.3466	1.0252	0.0747	0.6297	
	(4.6198)	(2.2176)	(2.6874)	(3.1694)	(1.1367)	(2.0640)	(2.6346)	(0.4899)	(2.8024)	
16	0.0040	0.0036	-0.0026	-0.0018	-0.0050	-0.0083	0.0118	-0.0002	0.0014	1/4
	(0.0677)	(0.0477)	(0.0680)	(0.0645)	(0.0260)	(0.0636)	(0.0678)	(0.0119)	(0.0636)	
	0.0712	0.0505	0.0528	0.0454	-0.0033	0.0218	0.0499	0.0006	0.0219	1
	(0.2556)	(0.1966)	(0.2470)	(0.2821)	(0.1019)	(0.2657)	(0.2536)	(0.0441)	(0.2446)	4
	0.2096	0.1711	0.2450	0.0625	-0.0010	0.0861	0.2545	0.0178	0.2324	
	(1.3466)	(0.8447)	(1.1989)	(1.0148)	(0.3901)	(1.0165)	(0.8414)	(0.1779)	(0.8607)	
128	0.0018	0.0023	0.0012	0.0012	0.0001	0.0006	-0.0038	-0.0005	-0.0045	1/4
	(0.0225)	(0.0170)	(0.0224)	(0.0233)	(0.0088)	(0.0230)	(0.0208)	(0.0039)	(0.0206)	
	0.0258	0.0155	0.0233	0.0070	0.0006	0.0041	-0.0037	0.0004	-0.0067	1
	(0.0992)	(0.0650)	(0.0982)	(0.0958)	(0.0355)	(0.0953)	(0.0944)	(0.0170)	(0.0948)	4
	-0.0077	0.0098	0.0002	0.0070	0.0007	0.0052	0.0215	0.0001	0.0097	
	(0.3712)	(0.2236)	(0.3623)	(0.3897)	(0.1425)	(0.3879)	(0.3650)	(0.0636)	(0.3632)	

Table 3. ME for σ , with RMSE in brackets, for a baseline distribution $\mathcal{G}(0, \sigma_b)$.

n	k									σ_b
	10			100			1000			
	MHM	BDM	IBDM	MHM	BDM	IBDM	MHM	BDM	IBDM	
2	2.4739	0.0309	0.0258	2.6628	0.0013	0.0048	2.7177	0.0007	0.0024	1/4
	(2.6201)	(0.0584)	(0.0560)	(2.6967)	(0.0119)	(0.0216)	(2.7371)	(0.0045)	(0.0155)	
	2.3399	0.1121	0.1324	2.3384	0.0206	0.0445	2.4139	0.0043	0.0130	1
	(2.4169)	(0.2259)	(0.2591)	(2.3760)	(0.0596)	(0.1085)	(2.4441)	(0.0204)	(0.0774)	
16	0.8238	0.3104	0.4727	0.6607	0.0056	0.0790	0.7284	0.0117	0.0905	4
	(1.6431)	(0.7600)	(0.9749)	(1.1945)	(0.2174)	(0.4192)	(1.2739)	(0.0680)	(0.3978)	
	0.0292	0.0006	0.0031	0.0229	-0.0008	-0.0037	0.0462	0.0000	0.0070	1/4
	(0.0577)	(0.0165)	(0.0279)	(0.0560)	(0.0051)	(0.0264)	(0.0788)	(0.0016)	(0.0293)	
128	0.1520	0.0152	0.0367	0.1339	-0.0003	0.0032	0.1495	0.0001	0.0058	1
	(0.2835)	(0.0604)	(0.1234)	(0.2740)	(0.0198)	(0.0959)	(0.2712)	(0.0059)	(0.0605)	
	0.3237	0.0742	0.1261	0.3159	-0.0010	0.0176	0.2071	0.0023	0.0294	4
	(0.9959)	(0.2826)	(0.4765)	(0.8719)	(0.0755)	(0.2238)	(0.7576)	(0.0237)	(0.1211)	
2	0.0074	0.0009	0.0055	0.0036	0.0001	0.0018	0.0027	-0.0001	0.0009	1/4
	(0.0181)	(0.0058)	(0.0162)	(0.0203)	(0.0017)	(0.0184)	(0.0179)	(0.0005)	(0.0161)	
	0.0300	0.0041	0.0220	0.0313	0.0003	0.0203	0.0138	0.0000	0.0037	1
	(0.0778)	(0.0221)	(0.0693)	(0.0756)	(0.0069)	(0.0616)	(0.0695)	(0.0023)	(0.0518)	
16	0.0557	0.0062	0.0287	0.0415	-0.0020	0.0024	0.0884	-0.0001	0.0123	4
	(0.2423)	(0.0759)	(0.1843)	(0.2805)	(0.0274)	(0.1286)	(0.2522)	(0.0086)	(0.0753)	

4.2. Exponential Baseline Distribution

Assume now we have another baseline distribution F , which is not a Gumbel. Notice that, for methods MHM and IBDM, we are approaching block maxima distribution by a Gumbel. However, for BDM, we employ the k th power of the baseline distribution function. When the baseline distribution is a Gumbel, we know that the k th power is also a Gumbel. However, for another baseline distributions, this is not true.

For this reason, we have to define different measures of error to evaluate the quality of estimations. We compared estimated distribution functions (H) with real ones (F^k) through their mean distance (D), mean absolute distance (AD) and root square distance (RSD). As analytical computation is not possible, we made a Monte-Carlo computation employing sample size $s = 10^4$. Then,

$$D_j = \frac{1}{s} \sum_{i=1}^s (H(x_i; \hat{\theta}_j) - F(x_i; \theta)^k) \tag{33}$$

$$AD_j = \frac{1}{s} \sum_{i=1}^s |H(x_i; \hat{\theta}_j) - F(x_i; \theta)^k| \tag{34}$$

and

$$RSD_j = \sqrt{\frac{1}{s} \sum_{i=1}^s (H(x_i; \hat{\theta}_j) - F(x_i; \theta)^k)^2}, \tag{35}$$

with $j = 1, \dots, M$, where M is the number of samples. $H(x; \hat{\theta})$ denotes the estimated distribution function for block maxima, for the baseline parameter $\hat{\theta}$, which is $\hat{\theta} = \hat{\lambda}_b$ if we have an exponential baseline distribution and $\hat{\theta} = (\hat{\mu}_b, \hat{\sigma}_b)$ for the Normal baseline distribution.

The measures of error were:

- Mean error:

$$ME = \frac{1}{M} \sum_{j=1}^M D_j. \tag{36}$$

- Root mean square error:

$$RMSE = \frac{1}{M} \sum_{j=1}^M RSD_j. \tag{37}$$

- Mean absolute error:

$$MAE = \frac{1}{M} \sum_{j=1}^M AD_j. \tag{38}$$

We considered the baseline $Exp(\lambda_b)$ distribution. In this case, for k big enough, $X \approx \mathcal{G}(\lambda_b^{-1} \ln k, \lambda_b^{-1})$. We took $\lambda_b = 2^j$, with $j = -2, -1, 0, 1, 2$.

As in the Gumbel baseline distribution, MHM shows high skewness when the number of blocks n is very small, compared to IBDM (see Figure 2). In addition, if we compute measures of error, we can see that, for small block sizes, the three methods offer similar values (see Table 4). For bigger values of k , BDM and IBDM provide better results, and usually BDM is the best method.

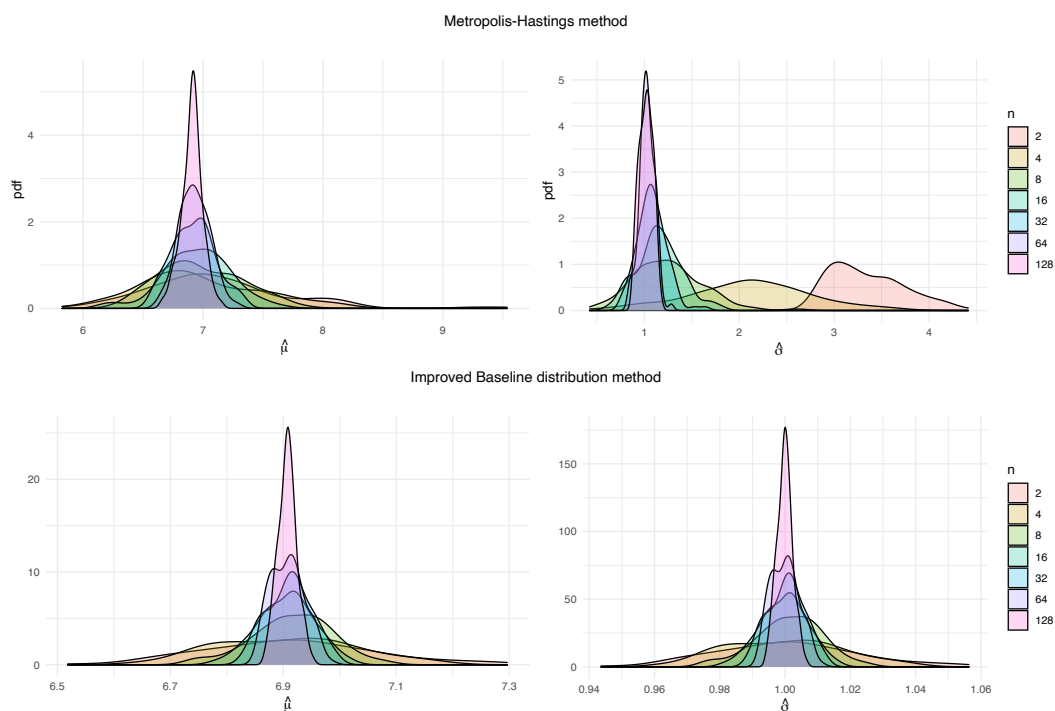


Figure 2. Probability density functions for M estimations of the block maxima parameters μ (left) and σ (right), obtained for the methods MHM and IBDM, with $k = 1000$ and different values of n , from $Exp(1)$ baseline distribution.

Table 4. ME for σ , with RMSE in brackets, for a baseline distribution $Exp(\lambda_b)$.

n	k									λ_b
	10			100			1000			
	MHM	BDM	IBDM	MHM	BDM	IBDM	MHM	BDM	IBDM	
2	-0.0645	0.0287	0.0207	-0.0563	0.0098	0.0061	-0.0652	-0.0004	-0.0011	1/2
	(0.1569)	(0.1297)	(0.1278)	(0.1485)	(0.0880)	(0.0878)	(0.1471)	(0.0339)	(0.0339)	
	-0.0818	0.0157	0.0084	-0.0915	-0.0078	-0.0115	-0.0945	0.0002	-0.0006	1
	(0.2022)	(0.1409)	(0.1382)	(0.2032)	(0.0780)	(0.0782)	(0.2012)	(0.0330)	(0.0330)	
16	-0.0156	-0.0006	0.0096	-0.0221	-0.0008	-0.0002	-0.0159	0.0008	0.0008	1/2
	(0.0680)	(0.0440)	(0.0452)	(0.0832)	(0.0302)	(0.0302)	(0.0750)	(0.0110)	(0.0110)	
	-0.0109	0.0019	0.0120	-0.0230	-0.0027	-0.0021	-0.0220	0.0001	0.0122	1
	(0.0741)	(0.0429)	(0.0455)	(0.0750)	(0.0294)	(0.0293)	(0.0743)	(0.0132)	(0.0132)	
128	-0.0039	-0.0057	0.0069	-0.0020	0.0002	0.0013	-0.0007	0.0002	0.0003	1/2
	(0.0244)	(0.0183)	(0.0218)	(0.0243)	(0.0100)	(0.0101)	(0.0227)	(0.0036)	(0.0036)	
	-0.0010	-0.0022	0.0103	-0.0002	0.0012	0.0024	-0.0004	0.0012	0.0013	1
	(0.0243)	(0.0153)	(0.0216)	(0.0267)	(0.0113)	(0.0115)	(0.0269)	(0.0044)	(0.0044)	

4.3. Normal Baseline Distribution

Finally, assume the baseline distribution is $\mathcal{N}(\mu_b, \sigma_b)$. As the mean has no effect on the variability of data, its value was fixed as $\mu_b = 0$ for easiness. Standard deviation was taken as $\sigma_b = 2^j$, $j = -2, -1, 0, 1, 2$.

The conclusions we obtain are quite similar to the previous case. We illustrate skewness and variability for MHM employing similar graphs (see Figure 3). Variability is especially pronounced when n is small, and we are estimating σ . In addition, errors are shown in Table 5.

Table 5. ME for σ , with RMSE in brackets, for a baseline distribution $\mathcal{N}(0, \sigma_b)$.

n	k									σ_b
	10			100			1000			
	MHM	BDM	IBDM	MHM	BDM	IBDM	MHM	BDM	IBDM	
2	-0.1097	-0.2932	-0.1942	-0.1120	-0.1183	-0.1204	-0.1101	-0.0132	-0.0319	1/4
	(0.2923)	(0.3274)	(0.2210)	(0.2975)	(0.1356)	(0.1481)	(0.2991)	(0.0443)	(0.0559)	
	-0.1023	-0.0049	-0.0222	-0.1009	0.0078	-0.0077	-0.1053	0.0047	-0.0150	1
	(0.2481)	(0.1139)	(0.1404)	(0.2609)	(0.0718)	(0.0876)	(0.2704)	(0.0394)	(0.0482)	
16	-0.0822	0.0252	-0.0079	-0.0803	0.0079	-0.0103	-0.0824	-0.0045	-0.0227	1/4
	(0.1643)	(0.1197)	(0.1374)	(0.1790)	(0.0892)	(0.1073)	(0.1876)	(0.0393)	(0.0497)	
	-0.0089	-0.0650	-0.0435	-0.0131	-0.0166	-0.0278	-0.0046	-0.0006	-0.0198	1/4
	(0.0736)	(0.0765)	(0.0644)	(0.0785)	(0.0326)	(0.0500)	(0.0638)	(0.0134)	(0.0283)	
128	-0.0085	0.0024	-0.0030	-0.0092	-0.0042	-0.0164	-0.0109	0.0012	-0.0184	1/4
	(0.0750)	(0.0546)	(0.0645)	(0.0658)	(0.0282)	(0.0423)	(0.0724)	(0.0144)	(0.0292)	
	-0.0060	0.0033	-0.0049	-0.0171	-0.0022	-0.0185	-0.0028	-0.0006	-0.0194	1
	(0.0679)	(0.0427)	(0.0574)	(0.0719)	(0.0338)	(0.0472)	(0.0675)	(0.0142)	(0.0281)	
128	0.0054	-0.0101	-0.0116	0.0041	-0.0011	-0.0154	0.0021	-0.0008	-0.0207	1/4
	(0.0323)	(0.0188)	(0.0299)	(0.0282)	(0.0116)	(0.0272)	(0.0235)	(0.0044)	(0.0260)	
	0.0090	-0.0002	-0.0038	0.0058	0.0014	-0.0126	0.0016	-0.0002	-0.0204	1
	(0.0336)	(0.0172)	(0.0296)	(0.0301)	(0.0116)	(0.0263)	(0.0249)	(0.0050)	(0.0258)	
128	0.0052	-0.0020	-0.0063	0.0052	-0.0006	-0.0147	0.0020	0.0001	-0.0200	1
	(0.0357)	(0.0188)	(0.0320)	(0.0288)	(0.0116)	(0.0272)	(0.0277)	(0.0048)	(0.0255)	4

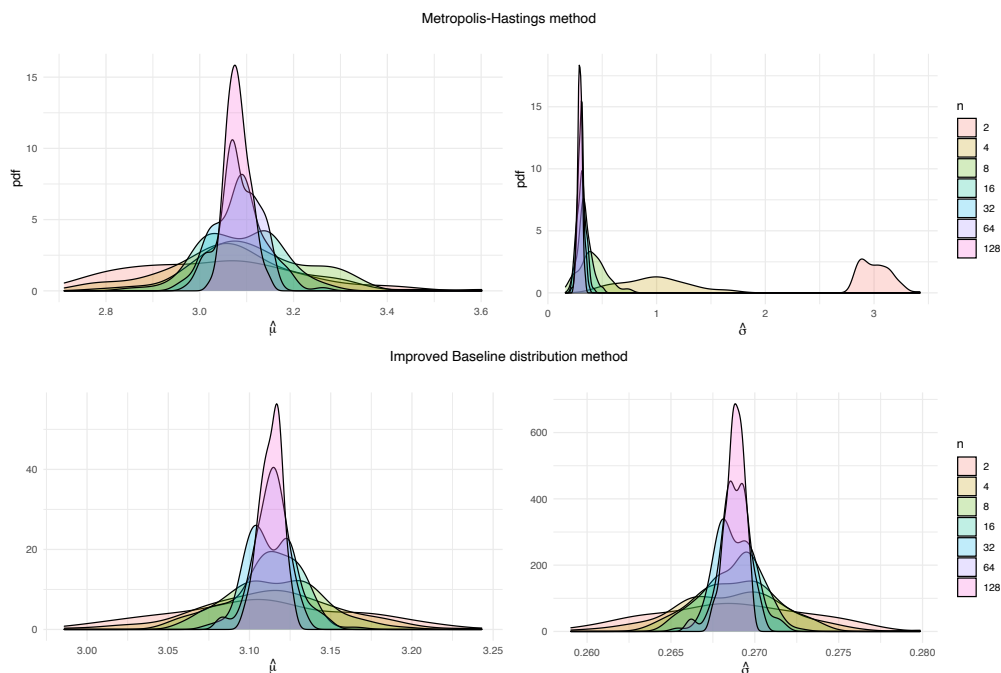


Figure 3. Probability density functions for M estimations of the block maxima parameters μ (left) and σ (right), obtained for the methods MHM and IBDM, with $k = 1000$ and different values of n , from $\mathcal{N}(0, 1)$.

In practical situations, data might not adjust to a concrete distribution and some perturbations (noise) could appear. To get a quick overview of how differences between baseline distribution data and block maxima data can affect the choice of the best method, we simulated a simple situation, when data come from a mixture of normal variables. Concretely,

$$Y = 0.9 \cdot Z + 0.1 \cdot Y_1, \quad Z \sim \mathcal{N}(0, 1), \quad Y_1 \sim \mathcal{N}(1, 1.5).$$

In Figure 4, we can see how MAE vary for the three methods when we vary the number of block maxima n , for a block size $k = 100$. In this case, IBDM offers the lowest errors, because it stresses the importance of extreme data. When the extreme data are scarce, both new methods, BDM and IBDM, improve MHM meaningfully.

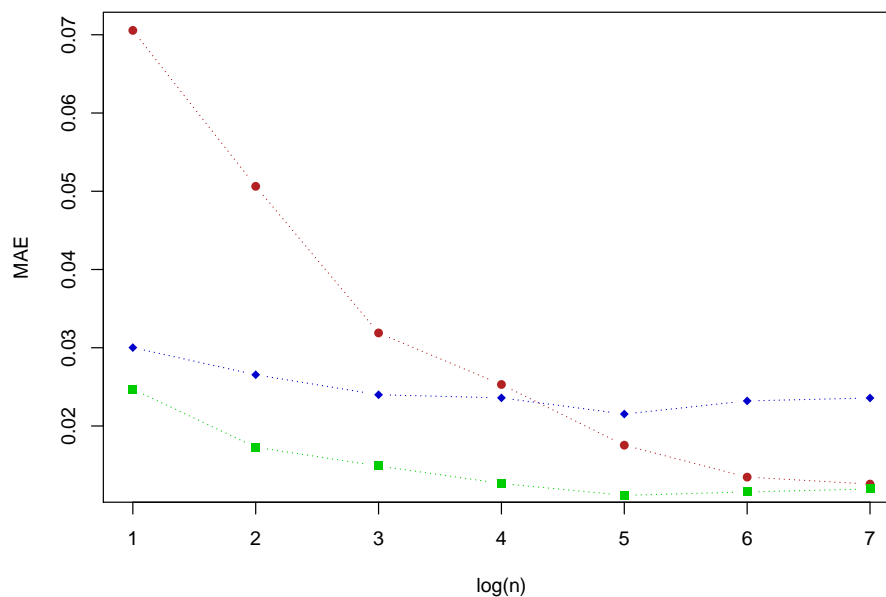


Figure 4. MAE for the three methods MHM (red), BDM (blue) and IBDM (green) with $k = 100$ and different values of n , for the baseline distribution Y .

5. Conclusions

1. One of the most common problems in EVT is estimating the parameters of the distribution, because the data are usually scarce. In this work, we considered the case when block maxima distribution is a Gumbel, and we developed two bayesian methods, BDM and IBDM, to estimate posterior distribution, making use of all the available data of the baseline distribution, not only the block maxima values.
2. The methods were proposed for three baseline distributions, namely Gumbel, Exponential and Normal, but the new strategy can easily be applied to some other baseline distributions, following the relations shown in Table 1.
3. We performed a broad simulation study to compare BDM and IBDM methods to classical Metropolis–Hastings method (MHM). The results are based on numerical studies, but theoretical support still needs to be provided.
4. We obtained that posterior distributions of BDM and IBDM are more concentrated and less skewed than MHM.
5. In general, the results obtained show that the methods which offer lower measures of error are BDM and IBDM, as they leverage all the data. The classical method, MHM, shows the worst results, especially when extreme data are scarce.
6. IBDM is the most stable method: regardless of the differences between extreme data and baseline data, it provides reasonably good measures of error. When the extreme data are scarce, both new methods, BDM and IBDM, improve MHM meaningfully.

Author Contributions: All authors contributed equally to this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by Junta de Extremadura, Consejería de Economía e Infraestructuras FEDER Funds IB16063, GR18108 project from Junta de Extremadura and Project MTM2017-86875-C3-2-R from Ministerio de Economía, Industria y Competitividad de España.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Nogaj, M.; Yiou, P.; Parey, S.; Malek, F.; Naveau, P. Amplitude and frequency of temperature extremes over the North Atlantic region. *Geophys. Res. Lett.* **2006**, *33*. [[CrossRef](#)]
2. Coelho, C.A.S.; Ferro, C.A.T.; Stephenson, D.B.; Steinskog, D.J. Methods for Exploring Spatial and Temporal Variability of Extreme Events in Climate Data. *J. Clim.* **2008**, *21*, 2072–2092. [[CrossRef](#)]
3. Acero, F.J.; Fernández-Fernández, M.I.; Carrasco, V.M.S.; Parey, S.; Hoang, T.T.H.; Dacunha-Castelle, D.; García, J.A. Changes in heat wave characteristics over Extremadura (SW Spain). *Theor. Appl. Climatol.* **2017**, *1–13*. [[CrossRef](#)]
4. García, J.; Gallego, M.C.; Serrano, A.; Vaquero, J. Trends in Block-Seasonal Extreme Rainfall over the Iberian Peninsula in the Second Half of the Twentieth Century. *J. Clim.* **2007**, *20*, 113–130. [[CrossRef](#)]
5. Re, M.; Barros, V.R. Extreme rainfalls in SE South America. *Clim. Chang.* **2009**, *96*, 119–136. [[CrossRef](#)]
6. Acero, F.J.; García, J.A.; Gallego, M.C. Peaks-over-Threshold Study of Trends in Extreme Rainfall over the Iberian Peninsula. *J. Clim.* **2011**, *24*, 1089–1105. [[CrossRef](#)]
7. Acero, F.J.; Gallego, M.C.; García, J.A. Multi-day rainfall trends over the Iberian Peninsula. *Theor. Appl. Climatol.* **2011**, *108*, 411–423. [[CrossRef](#)]
8. Acero, F.J.; Parey, S.; Hoang, T.T.H.; Dacunha-Castelle, D.; García, J.A.; Gallego, M.C. Non-stationary future return levels for extreme rainfall over Extremadura (southwestern Iberian Peninsula). *Hydrol. Sci. J.* **2017**, *62*, 1394–1411. [[CrossRef](#)]
9. Wi, S.; Valdés, J.B.; Steinschneider, S.; Kim, T.W. Non-stationary frequency analysis of extreme precipitation in South Korea using peaks-over-threshold and annual maxima. *Stoch. Environ. Res. Risk Assess.* **2015**, *30*, 583–606. [[CrossRef](#)]
10. García, A.; Martín, J.; Naranjo, L.; Acero, F.J. A Bayesian hierarchical spatio-temporal model for extreme rainfall in Extremadura (Spain). *Hydrol. Sci. J.* **2018**, *63*, 878–894. [[CrossRef](#)]
11. Ramos, A.A. Extreme value theory and the solar cycle. *Astron. Astrophys.* **2007**, *472*, 293–298. [[CrossRef](#)]
12. Acero, F.J.; Carrasco, V.M.S.; Gallego, M.C.; García, J.A.; Vaquero, J.M. Extreme Value Theory and the New Sunspot Number Series. *Astrophys. J.* **2017**, *839*, 98. [[CrossRef](#)]
13. Acero, F.J.; Gallego, M.C.; García, J.A.; Usoskin, I.G.; Vaquero, J.M. Extreme Value Theory Applied to the Millennial Sunspot Number Series. *Astrophys. J.* **2018**, *853*, 80. [[CrossRef](#)]
14. Castillo, E.; Hadi, A.S.; Balakrishnan, N.; Sarabia, J.M. *Extreme Value and Related Models with Applications in Engineering and Science*; Wiley: Hoboken, NJ, USA, 2004.
15. Castillo, E. Estadística de valores extremos. Distribuciones asintóticas. *Estad. Esp.* **1987**, *116*, 5–35.
16. Smith, R.L.; Naylor, J.C. A Comparison of Maximum Likelihood and Bayesian Estimators for the Three-Parameter Weibull Distribution. *Appl. Stat.* **1987**, *36*, 358. [[CrossRef](#)]
17. Coles, S.; Pericchi, L.R.; Sisson, S. A fully probabilistic approach to extreme rainfall modeling. *J. Hydrol.* **2003**, *273*, 35–50. [[CrossRef](#)]
18. Bernardo, J.M.; Smith, A.F.M. (Eds.) *Bayesian Theory*; Wiley: Hoboken, NJ, USA, 1994. [[CrossRef](#)]
19. Kotz, S.; Nadarajah, S. *Extreme Value Distributions: Theory and Applications*; ICP: London, UK, 2000.
20. Coles, S.G.; Tawn, J.A. A Bayesian Analysis of Extreme Rainfall Data. *Appl. Stat.* **1996**, *45*, 463. [[CrossRef](#)]
21. Rostami, M.; Adam, M.B. Analyses of prior selections for gumbel distribution. *Matematika* **2013**, *29*, 95–107. [[CrossRef](#)]
22. Chen, M.H.; Shao, Q.M.; Ibrahim, J.G. *Monte Carlo Methods in Bayesian Computation*; Springer: New York, NY, USA, 2000. [[CrossRef](#)]
23. Vidal, I. A Bayesian analysis of the Gumbel distribution: An application to extreme rainfall data. *Stoch. Environ. Res. Risk Assess.* **2013**, *28*, 571–582. [[CrossRef](#)]
24. Lye, L.M. Bayes estimate of the probability of exceedance of annual floods. *Stoch. Hydrol. Hydraul.* **1990**, *4*, 55–64. [[CrossRef](#)]
25. Rostami, M.; Adam, M.B.; Ibrahim, N.A.; Yahya, M.H. Slice sampling technique in Bayesian extreme of gold price modelling. *Am. Inst. Phys.* **2013**, *1557*, 473–477. [[CrossRef](#)]
26. Gumbel, E.J. *Statistics of Extremes (Dover Books on Mathematics)*; Dover Publications: New York, NY, USA, 2012.
27. Gnedenko, B. Sur la distribution limite du terme maximum d’une serie aleatoire. *Ann. Math.* **1943**, *44*, 423–453. [[CrossRef](#)]

28. Fisher, R.A.; 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*; Cambridge University Press: Cambridge, UK, 1928; Volume 24, pp. 180–190.
29. Ferreira, A.; de Haan, L. *Extreme Value Theory*; Springer: Berlin/Heidelberg, Germany, 2006.
30. Plummer, M.; Best, N.; Cowles, K.; Vines, K. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **2006**, *6*, 7–11.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Capítulo 3

Artículo B: Métodos Base para la estimación de los parámetros de la distribución de Pareto Generalizada

Autores:

Jacinto Martín, M. Isabel Parra, Mario M. Pizarro, Eva L. Sanjuán

Departamento de Matemáticas, Universidad de Extremadura

Revista: Entropy, 24(2), 178, 2022

DOI: [10.3390/e24020178](https://doi.org/10.3390/e24020178)

Resumen:

En este trabajo, se plantean dos nuevos métodos Bayesianos para estimar los parámetros de la distribución de Pareto Generalizada (GPD) aprovechando la información proporcionada por todas las observaciones. Para ello, se tienen en cuenta las relaciones existentes entre los parámetros de la distribución base y los parámetros de la distribución límite GPD definiendo distribuciones a priori altamente informativas. En particular, se comparan estos métodos con los resultados obtenidos al emplear el algoritmo de Metropolis-Hastings (MH) sobre los datos que exceden el umbral, cuando la distribución base es una distribución estable. Éstas permiten reducir el problema a estudiar las distribuciones estándar y proponer nuevos estimadores para los parámetros de la GPD. En concreto, se consideran tres distribuciones estables (Normal, Lévy y Cauchy) como principales ejemplos

de los diferentes comportamientos de las colas de una distribución. Estas nuevas estrategias se pueden aplicar a otras distribuciones mediante la búsqueda de relaciones entre sus parámetros y los de la GPD a través de estudios de simulación. Finalmente, se aplican los nuevos métodos a datos reales de contaminación atmosférica distribuidos según una Gamma, mostrando que las estimaciones son más precisas que las dadas cuando se utiliza el algoritmo MH.

Article

Baseline Methods for the Parameter Estimation of the Generalized Pareto Distribution

Jacinto Martín ^{1,†} , María Isabel Parra ^{1,†} , Mario Martínez Pizarro ^{2,†}  and Eva López Sanjuán ^{2,*,†} 

¹ Departamento de Matemáticas, Facultad de Ciencias, Universidad de Extremadura, 06006 Badajoz, Spain; jrmartin@unex.es (J.M.); mipa@unex.es (M.I.P.)

² Departamento de Matemáticas, Centro Universitario de Mérida, Universidad de Extremadura, 06800 Mérida, Spain; mariomp@unex.es

* Correspondence: etlopez@unex.es

† The authors contributed equally to this work.

Abstract: In the parameter estimation of limit extreme value distributions, most employed methods only use some of the available data. Using the peaks-over-threshold method for Generalized Pareto Distribution (GPD), only the observations above a certain threshold are considered; therefore, a big amount of information is wasted. The aim of this work is to make the most of the information provided by the observations in order to improve the accuracy of Bayesian parameter estimation. We present two new Bayesian methods to estimate the parameters of the GPD, taking into account the whole data set from the baseline distribution and the existing relations between the baseline and the limit GPD parameters in order to define highly informative priors. We make a comparison between the Bayesian Metropolis–Hastings algorithm with data over the threshold and the new methods when the baseline distribution is a stable distribution, whose properties assure we can reduce the problem to study standard distributions and also allow us to propose new estimators for the parameters of the tail distribution. Specifically, three cases of stable distributions were considered: Normal, Lévy and Cauchy distributions, as main examples of the different behaviors of the tails of a distribution. Nevertheless, the methods would be applicable to many other baseline distributions through finding relations between baseline and GPD parameters via studies of simulations. To illustrate this situation, we study the application of the methods with real data of air pollution in Badajoz (Spain), whose baseline distribution fits a Gamma, and show that the baseline methods improve estimates compared to the Bayesian Metropolis–Hastings algorithm.

Keywords: Bayesian inference; generalized Pareto distribution; Metropolis–Hastings algorithm; stable distributions; extreme value theory



Citation: Martín, J.; Parra, M.I.; Pizarro, M.M.; Sanjuán, E.L. Baseline Methods for the Parameter Estimation of the Generalized Pareto Distribution. *Entropy* **2022**, *24*, 178. <https://doi.org/10.3390/e24020178>

Academic Editor: Antonio M. Scarfone

Received: 17 November 2021

Accepted: 21 January 2022

Published: 25 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Extreme value theory (EVT) is a set of statistical tools employed for modeling and predicting the occurrence of rare events outside the range of available data. It has been widely used to study events that are more extreme than any previously observed, e.g., in disciplines such as climatology: extreme events of temperature [1–4], precipitations [5–10], and solar climatology [11–13]; finance and insurance: applications to risk theory [14–20]; and engineering: design for modern buildings [21].

There are two approaches for modeling an extreme value problem. The first one is the block maxima method: dividing the sample space into blocks of equal size, the maxima values of each block follow, under a certain domain of attraction conditions, approximately a Generalized Extreme Value (GEV) distribution [22]. The second way to deal with an extreme value data set attempts to make use of the available information about the upper tail of the distribution than just the block maxima. The so-called *peaks-over-threshold* (POT) method is due to hydrologists trying to model floods. Loosely speaking, References [23,24] showed that when we consider the distribution of data above a certain threshold u , it can

usually be approximated by a properly scaled Generalized Pareto distribution (GPD) as u tends to the endpoint of the distribution. This is the point of view considered in this article.

Due to its importance, several methods have been proposed for estimating the shape and scale parameters of the GPD. Classical methods include the method of moments, probability weighted moments, maximum likelihood and others. An exhaustive review of them can be consulted in [25,26]. Other researchers have proposed generalizations of the GPD: Reference [27] proposed a three-parameter Pareto distribution and employed POT to make estimations of Value at Risk and [28] introduced an extension of GPD and performed parametric estimation. However, classical methods might be inappropriate in certain situations, as explained in [26]. That is why Bayesian inference could be advisable. There are not many approaches to the GPD parameter estimation through Bayesian techniques. We can cite [29], who recommended the use of conjugate prior distributions; Reference [30] estimated the shape parameter when it is positive, and the computation of the posterior distribution was implemented via the Markov Chain Monte Carlo (MCMC) method with Gibbs sampling; Reference [31] employed Gamma and Pareto priors via MCMC; Reference [32] proposed a Bayesian mixture model, where the threshold was also a random parameter; Reference [33] employed Jeffrey's prior and Metropolis–Hastings (MH) method and [34] employed the GPD distribution itself as the prior density.

In this paper, we aim at seizing all the available information coming from data in order to estimate the parameters of the GPD in a way as accurate as possible. A similar idea was also implemented in [35], for the estimation of the parameters of the Gumbel distribution when the baseline distribution is Gumbel, Exponential or Normal.

We will take into account all the data of the baseline distribution and study the relation between the baseline parameters and the parameters of the limit GPD in order to incorporate such relation into the sketch of new methods to make estimations. Concretely, we propose two methods and compare them with the classical MH method for data over the threshold. In addition, we will analyze four particular examples of underlying distribution: Exponential, Lévy, Cauchy and Normal, and make a special study of stable distributions.

2. Generalized Pareto Distribution and Its Relation with Extreme Value Theory

Let X be a random variable with distribution function F . Define u as the threshold value and let $X_u = X - u | X > u$ be the random variable with distribution function

$$F_{X_u}(x) = P[X_u \leq x] = P[X \leq x + u | X > u] = \frac{F(x + u) - F(u)}{1 - F(u)},$$

for $0 \leq x \leq x_F - u$, being x_F the right endpoint of F , that is, $x_F := \sup\{x : F(x) < 1\}$.

Notice that X_u is the random variable that we obtain when we consider the distribution of data above the threshold, which we usually call the tail distribution.

Given a random variable X , we say that it follows a Generalized Pareto Distribution (GPD) when its distribution function is

$$G(x|\zeta, \sigma) = \begin{cases} 1 - \left(1 + \frac{\zeta x}{\sigma}\right)^{-1/\zeta}, & \zeta \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right), & \zeta = 0 \end{cases} \quad (1)$$

where $\sigma > 0$ and $\zeta \in \mathbb{R}$ are the scale and shape parameters, respectively. Equation (1) is valid when $x \geq 0$ for $\zeta \geq 0$, and for $0 \leq x \leq -\sigma/\zeta$ for $\zeta < 0$.

The fundamental result that connects EVT and GPD distribution belongs to [23,24], and it establishes that under certain mild conditions, for a random variable X , the distribution of X_u for a sufficiently high threshold u follows a properly scaled Generalized Pareto distribution (GPD).

We will call the distribution function of X , F , the baseline distribution of the GPD. The parameters ζ and σ will depend on the value of the threshold u , and on the baseline distribution. For example, ζ is determined by the shape of the upper tail of the baseline distribution F . Positive values of the shape parameter correspond to heavy tails, while

negative ones come from light tails. The special case $\xi = 0$ will appear when the upper tail of the distribution tends to an exponential distribution of parameter $1/\sigma$.

In this work, we will consider several types of baseline distributions for the GPD. Traditionally, estimation for the parameters of the limit GPD in Extreme Value Theory has been made taking into account only values above the threshold, but this information might be scarce. We propose a new strategy consisting in seizing all the data from the baseline distribution. We will show how this strategy can produce accurate estimations for the parameters of the GPD.

In the case when the baseline distribution is an exponential distribution with parameter λ , with distribution function $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$, for every $u \geq 0$,

$$\begin{aligned} F_{X_u}(x) &= \frac{1 - e^{-\lambda(x+u)} - (1 - e^{-\lambda u})}{e^{-\lambda u}} \\ &= 1 - e^{-\lambda x} = F(x), \forall x \geq 0. \end{aligned}$$

Consequently, the tail distribution X_u is the same as the underlying distribution in the exponential case. Therefore, we must employ all the available data to estimate the parameter $\lambda = 1/\sigma$ in the definition of the GPD (1).

The case when $\xi \neq 0$ is different. In this paper, we will consider some of the most employed distributions as underlying distributions: normal distribution, which has light tails ($\xi < 0$); and the Cauchy and Lévy distributions, which lead to heavy tails ($\xi > 0$). Those distributions are stable; therefore, they have additional properties that will be helpful to estimate the parameters of the GPD. We will study such properties below.

With respect to the threshold, it can be settled as a known value, which has a physical meaning, depending on the characteristics of the data, or it can be defined as an upper order statistic. It is generally defined as a p -quantile of the underlying distribution q_p , for appropriate values of p .

3. Stable Distributions

Stable distributions are a rich class of probability distributions characterized by [36] and they have been proposed as a model for many types of physical and economic systems because of their interesting properties. They are the attractors of sums of independent, identically distributed distributions whether or not the mean or variance is finite. Good references to study them are [16] or [37].

Let Z be a random variable with parameters defined by its characteristic function:

$$E[e^{itZ}] = \begin{cases} \exp\left\{-|t|^\alpha \left(1 - i\beta \tan \frac{\pi\alpha}{2} \text{sign}(t)\right)\right\}, & \text{if } \alpha \neq 1 \\ \exp\left\{-|t| \left(1 + i\beta \frac{2}{\pi} \text{sign}(t) \log |t|\right)\right\}, & \text{if } \alpha = 1 \end{cases} \quad (2)$$

where the parameter $\alpha \in (0, 2]$ is called the index of stability, and $\beta \in [-1, 1]$ is the skewness parameter. When $\beta = 0$, the distribution is symmetric.

A random variable X is said to follow a stable distribution with parameters $a > 0$ and $b \in \mathbb{R}$ if it satisfies that

$$X = aZ + b. \quad (3)$$

Generally, densities can be expressed only by complicated special functions, but there are three special cases of stable distributions that have probability density functions, which can be expressed analytically:

- When $\alpha = 1/2$ and $\beta = 1$, we obtain Lévy distributions, $Z \sim \mathcal{L}(0, 1)$. If $X \sim \mathcal{L}(\gamma, \delta)$, then $a = \delta$ and $b = \gamma$ in (3).
- When $\alpha = 1$ and $\beta = 0$, we obtain the family of Cauchy distributions, $Z \sim \mathcal{C}(0, 1)$. If $X \sim \mathcal{C}(\gamma, \delta)$, then $a = \delta$ and $b = \gamma$ in (3).
- When $\alpha = 2$ and $\beta = 0$, we obtain the normal distribution, $N(0, \sqrt{2})$. If $X \sim N(\mu, \sigma)$, then $a = \sigma$ and $b = \mu$ in (3). As usual, in this case we will denote $Z \sim N(0, 1)$.

In particular, as we can standardize the stable distributions of a family, the p -quantiles q_p for a stable distribution X , can be expressed in terms of the p -quantiles z_p of the standard distribution Z , as

$$q_p = az_p + b. \quad (4)$$

Let us assume that X follows a stable distribution with parameters a and b for Equation (3), and fix $u = q_p$ as the threshold for the problem of extreme value. Then, for u large enough, z_p from (4) is also large, and, consequently, denoting $u_Z = z_p$, Theorem 1 guarantees that $Z_{u_Z} \sim \text{GPD}(\xi_Z, \sigma_Z)$. Therefore, as

$$\frac{X_u}{a} = Z_{u_Z} \quad (5)$$

and

$$\begin{aligned} F_{X_u}(x) &= P(X_u \leq x) = P(X_u/a \leq x/a) \\ &\approx G\left(\frac{x}{a} \mid \xi_Z, \sigma_Z\right) = 1 - \left(1 + \frac{\xi_Z x}{\sigma_Z a}\right)^{-1/\xi_Z}, \end{aligned} \quad (6)$$

then

$$X_u \sim \text{GPD}(\xi_Z, a\sigma_Z). \quad (7)$$

The parameter ξ_Z will remain constant for all the random variables of the same stable family, whatever the parameters of the baseline variable are, while the scale parameter is obtained through the product of the standardization parameter a and the scale parameter σ_Z for the GPD limiting distribution of Z_{u_Z}

$$\xi = \xi_Z, \sigma = a\sigma_Z. \quad (8)$$

In the case when the underlying distribution is a Cauchy or a Lévy, or any stable distribution X with the index $0 < \alpha < 2$, the tail of the distribution is considered to be "heavy", therefore it leads to a GPD where $\xi > 0$.

Stable distributions with index of stability $\alpha \neq 2$, (all of them except normal distribution) also verify an interesting property. As we can see in [36], given the standard distribution Z , its survival function \bar{F} can be approximated by:

$$\bar{F}(x) \sim (1 + \beta)C_\alpha x^{-\alpha}, \quad x \rightarrow \infty \quad (9)$$

where $C_\alpha = \frac{1}{\pi} \Gamma(\alpha) \sin\left(\frac{\alpha\pi}{2}\right)$.

From this approximation, we can infer that the shape of the tail of the distribution will only depend on the index of stability α . Therefore, if we consider the GPD that models Z_{u_Z} , the shape parameter ξ_Z will be fixed. We will estimate it through simulation.

Proposition 1. *When the baseline distribution is a standard stable distribution Z with $\alpha < 2$, the relation between the parameters of Z and the parameters of the GPD that models the distribution above the p -quantile of Z , u_Z , is:*

$$\hat{\xi}_Z = \frac{1}{\alpha}, \quad \hat{\sigma}_Z = \frac{1}{\alpha} \left(\frac{C_\alpha(1 + \beta)}{1 - p} \right)^{1/\alpha} \quad (10)$$

Proof. From Theorem 1, for u_Z that is big enough,

$$\bar{F}_{u_Z}(x) \sim \left(1 + \frac{\xi_Z x}{\sigma_Z}\right)^{-1/\xi_Z}$$

and by Proposition 1 (9), also for big values of u_Z

$$\bar{F}_{u_Z}(x) \sim \frac{(1 + \beta)C_\alpha x^{-\alpha}}{\bar{F}_Z(u_Z)} = \frac{(1 + \beta)C_\alpha x^{-\alpha}}{1 - p}.$$

Then, making equal both expressions,

$$\left(1 + \frac{\xi_Z}{\sigma_Z} x\right)^{-1/\xi_Z} = \left(\left[\frac{(1 + \beta)C_\alpha}{1 - p}\right]^{-1/\alpha} x\right)^{-\alpha}$$

Therefore, we can take $\hat{\xi}_Z = 1/\alpha$ as an estimator for ξ_Z (notice that the shape of the tail of the distribution depends only on the value of α). \square

Substituting ξ_Z by $1/\alpha$, we have

$$1 + \frac{1}{\alpha\sigma_Z} x = \left[\frac{(1 + \beta)C_\alpha}{1 - p}\right]^{-1/\alpha} x,$$

so

$$1 = \left(\left[\frac{(1 + \beta)C_\alpha}{1 - p}\right]^{-1/\alpha} - \frac{1}{x}\right)\alpha\sigma_Z \sim \left[\frac{(1 + \beta)C_\alpha}{1 - p}\right]^{-1/\alpha} \alpha\sigma_Z$$

as $1/x$ can be negligible. Therefore, we define

$$\hat{\sigma}_Z = \frac{1}{\alpha} \left(\frac{C_\alpha(1 + \beta)}{1 - p}\right)^{1/\alpha}.$$

In Section 5, we will assure the accuracy of these estimators through an extensive simulation study.

4. Metropolis–Hastings (MH) Method

In this section, we will explain how to apply the Markov chain Monte Carlo (MCMC) method through the Metropolis–Hastings (MH) algorithm to make the estimations for stable distributions. We have to distinguish between light tails ($\xi < 0$) and heavy tails ($\xi > 0$). Let us assume $X_u \sim GPD(\xi, \sigma)$ and that we dispose of m values.

Let $\mathbf{x} = (x^1, \dots, x^m)$ be a sample of n values from X and $\mathbf{x}_u = (x_u^1, \dots, x_u^m)$ be a sample of m values from X_u .

4.1. Light Tails $\xi < 0$

Take $k = -\xi$, and $\delta = -\frac{\sigma}{\xi}$, so $X_u \sim GPD(-k, k\delta)$, with the likelihood function

$$L(k, \delta | \mathbf{x}_u) = k^{-m} \delta^{-m} \prod_{i=1}^m \left(1 - \frac{x_u^i}{\delta}\right)^{-1+1/k}$$

Considering $\Gamma(a_0, b_0)$ and $\Gamma(a_1, b_1)$ as prior distributions for both parameters. Then, the MH algorithm is applied:

1. Draw a starting sample $(k^{(0)}, \delta^{(0)})$
2. For $j = 0, 1, \dots$
 - Sample candidates k^*, δ^* from the proposal distributions

$$k^* \sim \mathcal{N}(k^{(j)}, \nu_k), \delta^* \sim \mathcal{N}(\delta^{(j)}, \nu_\delta)$$

- Calculate the ratios

$$r_k = \frac{\pi(k^* | \delta^{(j)}, \mathbf{x}_u)}{\pi(k^{(j)} | \delta^{(j)}, \mathbf{x}_u)}, r_\delta = \frac{\pi(\delta^* | k^{(j)}, \mathbf{x}_u)}{\pi(\delta^{(j)} | k^{(j)}, \mathbf{x}_u)}$$

- Set

$$k^{(j+1)} = \begin{cases} k^*, & \text{with probability } \min\{1, r_k\} \\ k^{(j)}, & \text{otherwise} \end{cases}$$

$$\delta^{(j+1)} = \begin{cases} \delta^*, & \text{with probability } \min\{1, r_\delta\} \\ \delta^{(j)}, & \text{otherwise} \end{cases}$$

3. Iterate the former procedure.

Notice that

$$r_k = \left(\frac{k^*}{k^{(j)}}\right)^{a_0-m-1} \exp\left\{b_0(k^* - k^{(j)}) + \left(\frac{1}{k^*} - \frac{1}{k^{(j)}}\right) \sum_{i=1}^m \ln\left(1 - \frac{x_u^i}{\delta^{(j)}}\right)\right\}$$

$$r_\delta = \left(\frac{\delta^*}{\delta^{(j)}}\right)^{a_1-m-1} \exp\left\{b_1(\delta^* - \delta^{(j)}) + \left(\frac{1}{k^{(j)}} - 1\right) \sum_{i=1}^m \ln\left(1 - \frac{x_u^i}{\delta^*}\right) - \left(\frac{1}{k^{(j)}} - 1\right) \sum_{i=1}^m \ln\left(1 - \frac{x_u^i}{\delta^{(j)}}\right)\right\}$$

Finally, we obtain estimations for ξ and σ from $\xi = -k$ and $\sigma = k\delta$.

4.2. Heavy Tails $\xi > 0$

In this case, the likelihood function is

$$L(\xi, \sigma | \mathbf{x}_u) = \sigma^{-m} \prod_{i=1}^m \left(1 + \xi \frac{x_u^i}{\sigma}\right)^{-(1+1/\xi)}$$

Taking a type I Pareto (a_0, b_0) as prior distribution for ξ and $\text{Inv}\Gamma(a_1, b_1)$ for σ ,

$$\pi(\xi) \propto \xi^{-(a_0+1)}, \text{ with } \xi > b_0$$

$$\pi(\sigma) \propto \sigma^{-(a_1+1)} \exp\left\{-\frac{b_1}{\sigma}\right\}$$

Posterior conditional distributions are

$$\pi(\xi | \sigma, \mathbf{x}_u) \propto \xi^{-(a_0+1)} \prod_{i=1}^m \left(1 + \xi \frac{x_u^i}{\sigma}\right)^{-(1+1/\xi)}$$

$$\pi(\sigma | \xi, \mathbf{x}_u) \propto \sigma^{-(m+a_1+1)} \exp\left\{-\frac{b_1}{\sigma}\right\} \prod_{i=1}^m \left(1 + \xi \frac{x_u^i}{\sigma}\right)^{-(1+1/\xi)}$$

Then, MH algorithm is applied, as in the previous case.

Notice that

$$r_\xi = \left(\frac{\xi^{(j)}}{\xi^*}\right)^{a_0+1} \exp\left\{\left(1 + \frac{1}{\xi^{(j)}}\right) \sum_{i=1}^m \ln\left(1 + \xi^{(j)} \frac{x_u^i}{\sigma^{(j)}}\right) - \left(1 + \frac{1}{\xi^*}\right) \sum_{i=1}^m \ln\left(1 + \xi^* \frac{x_u^i}{\sigma^{(j)}}\right)\right\}$$

$$r_\sigma = \left(\frac{\sigma^{(j)}}{\sigma^*}\right)^{m+a_1+1} \exp\left\{b_1\left(\frac{1}{\sigma^{(j)}} - \frac{1}{\sigma^*}\right) + \left(1 + \frac{1}{\xi^{(j)}}\right) \sum_{i=1}^m \ln\left(1 + \xi^{(j)} \frac{x_u^i}{\sigma^{(j)}}\right) - \left(1 + \frac{1}{\xi^{(j)}}\right) \sum_{i=1}^m \ln\left(1 + \xi^{(j)} \frac{x_u^i}{\sigma^*}\right)\right\}$$

5. Baseline MH Method (BMH)

In this section, we will introduce Baseline Metropolis–Hastings (BMH) method, designed according to the objectives of seizing all the available data from the baseline distribution and making use of the existing relations between the baseline parameters and the limit GPD parameters. The method consists of:

1. Applying the MH algorithm to estimate the parameters for the baseline distribution θ .
2. Making use of the relations between the baseline parameters θ and the GPD parameters ζ and σ to compute estimations for ζ and σ .

In the case of stable distributions, these relations have been explained in previous sections and are given by (8). We will detail below the application of the BMH method for the three selected stable distributions.

For the rest of the baseline distributions, the strategy to find such relations would be made thorough studies of simulation, in order to establish correspondences between the baseline parameters and the tail GPD parameters. At the moment, there are no studies in the literature about this subject; therefore, it would be interesting to perform them in future research.

Then, in the case of stable distributions, the application of BMH would be:

1. Apply the MH algorithm to estimate scale parameter a from the stable baseline distribution.
2. Make use of the relation (8) to compute estimations for ζ and σ , using estimators for ζ_Z and σ_Z that we detail below.

5.1. Estimations for ζ_Z and σ_Z

In order to provide good estimations for ζ_Z and σ_Z , we made a thorough simulation study for the three baseline distributions we have considered: Lévy, Cauchy, and Normal distribution. We took values for $p \in [0.990, 0.995]$ with increments of 0.001, and set the threshold $u_Z = q_p$. For each distribution, and for each value of the threshold, $m = 1000$ values from Z_{u_Z} were generated. This sequence was repeated 100 times. Therefore, we obtained 100 point estimations for each p .

To guarantee the convergence of the MCMC algorithm, we must be sure that the posterior distribution has been reached. These proceedings were made using library coda [38] for R software, taking 10,000 values for the burn-in period, 25 values for the thinning, and selecting initial values for each sample. Finally, to obtain the posterior distribution for each parameter, a Markov chain of length 10,000 was obtained and we considered the mean as the estimator. The results of the simulation study are shown in Figure 1.

5.1.1. Lévy and Cauchy Baseline Distribution

By Proposition 1 (10), we had estimations for ζ_Z and σ_Z . Concretely, these are

$$\hat{\zeta}_Z = 2, \quad \hat{\sigma}_Z = \frac{4}{\pi}(1 - p)^{-2}, \tag{11}$$

for the Lévy distribution and,

$$\hat{\zeta}_Z = 1, \quad \hat{\sigma}_Z = \frac{1}{\pi}(1 - p)^{-1} \tag{12}$$

for the Cauchy distribution.

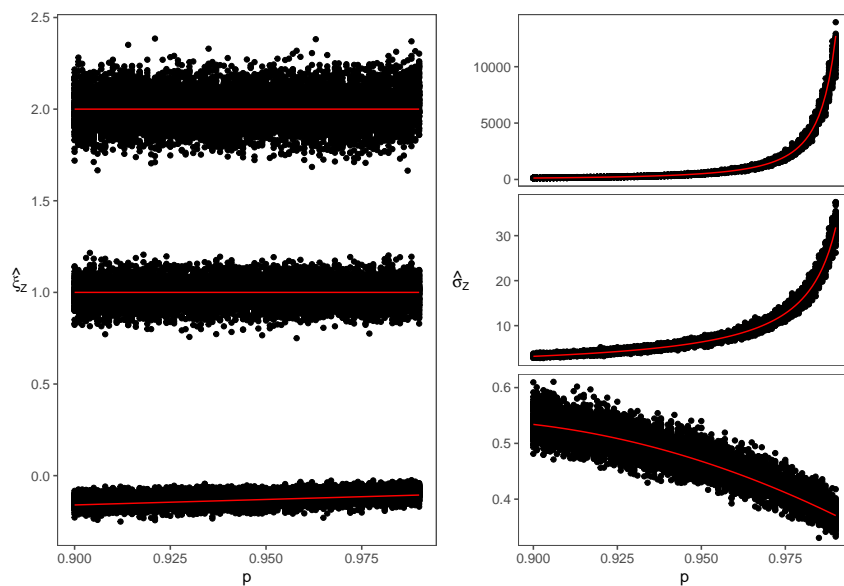


Figure 1. Estimations for ξ_Z (left) and σ_Z (right) for Lévy (upper charts), Cauchy (middle charts), Normal (lower charts) and estimators from Equations (11)–(13) plotted in red for $p \in [0.900, 0.995]$.

5.1.2. Normal Baseline Distribution

In the case of the normal distribution, property (10) is not verified. In this case, the adjustment from the simulation study is:

$$\hat{\xi}_Z = -0.7 + 0.61p, \quad \hat{\sigma}_Z = 0.34 + 3.18(1 - p) - 12.4(1 - p)^2. \quad (13)$$

Now, we will estimate the scale parameter a of the baseline distribution X . Notice that parameter b for the stable distribution X defined in (3) does not have any influence on the estimation of the parameters of the GPD, as we have shown before. Consequently, we will assume $b = 0$ from now on.

5.2. Lévy Distribution

Likelihood function for Lévy distribution is:

$$L(a|\mathbf{x}) \propto a^{n/2} \exp\left\{-\frac{a}{2} \sum_{i=1}^n \frac{1}{x^i}\right\}$$

Taking a prior distribution $\Gamma(a_0, b_0)$ for a , and making use of (8) and (11), we obtain estimations $\hat{\xi}$ and $\hat{\sigma}$.

5.3. Cauchy Distribution

Likelihood function for Cauchy is:

$$L(a|\mathbf{x}) \propto a^{-n} \prod_{i=1}^n \left(1 + \left(\frac{x^i}{a}\right)^2\right)^{-1}$$

Taking $\Gamma(a_0, b_0)$ as prior distribution and, making use of (8) and (12), we obtain estimations $\hat{\xi}$ and $\hat{\sigma}$.

5.4. Normal Distribution

In this case, we consider $\text{Inv}\Gamma(a_0, b_0)$ as prior distribution for a^2 , and making use of (8) and (13), we obtain estimations $\hat{\xi}$ and $\hat{\sigma}$.

6. Informative Priors Baseline MH (IPBMH)

Finally, we propose an MH method to estimate the parameters ζ and σ , where highly informative priors are employed, seizing the estimations computed before. Employing estimations of ζ_Z, σ_Z and a obtained in previous sections, we can settle priors for ζ and σ , which are very informative. Notice that this way of proceeding keeps the original idea of Extreme Value Theory, granting more weight to tail values because they are employed twice: to compute estimations for ζ_Z and σ_Z and also through the likelihood function. As we commented before, this method could also be implemented for other baseline distributions once we have found relations between baseline and GPD parameters.

For stable distributions, highly informative priors are

$$\zeta \sim N(\zeta_Z, b_1), \sigma \sim N(a \cdot \sigma_Z, b_2)$$

Then, for the three distributions studied, a is estimated through BMH and ζ_Z, σ_Z are estimations computed through (11)–(13). In addition,

- b_1 is constant, being 0.03, 0.065 and 0.1 for Normal, Cauchy and Lévy baseline distributions, respectively.
- $b_2 = \exp\{c_1 p^2 + c_2 p + c_3\}$, where values are given in Table 1.

Table 1. Values for c_1, c_2, c_3 for the three baseline distributions.

Distribution	c_1	c_2	c_3
Lévy	500.2	−900.9	408.2
Cauchy	323.57	−588.51	266.13
Normal	−46.24	83.55	−41.58

The Joint posterior distribution is

$$\pi(\zeta, \sigma | \mathbf{x}_u) \propto \sigma^{-m} \exp\left\{-\frac{1}{2b_1^2}(\zeta - \zeta_Z)^2 - \frac{1}{2b_2^2}(\sigma - a \cdot \sigma_Z)^2\right\} \prod_{i=1}^m \left(1 + \zeta \frac{x_u^i}{\sigma}\right)^{-(1+1/\zeta)}$$

and marginal distributions are

$$\begin{aligned} \pi(\zeta | \sigma, \mathbf{x}_u) &\propto \exp\left\{-\frac{1}{2b_1^2}(\zeta - \zeta_Z)^2\right\} \prod_{i=1}^m \left(1 + \zeta \frac{x_u^i}{\sigma}\right)^{-(1+1/\zeta)} \\ \pi(\sigma | \zeta, \mathbf{x}_u) &\propto \sigma^{-m} \exp\left\{-\frac{1}{2b_2^2}(\sigma - a \cdot \sigma_Z)^2\right\} \prod_{i=1}^m \left(1 + \zeta \frac{x_u^i}{\sigma}\right)^{-(1+1/\zeta)} \end{aligned}$$

Then, we apply the MH algorithm with

$$\begin{aligned} r_\zeta &= \exp\left\{\frac{1}{2b_1^2} \left((\zeta^{(j)} - \zeta_Z)^2 - (\zeta^* - \zeta_Z)^2 \right) \right. \\ &\quad \left. - \left(1 + \frac{1}{\zeta^*}\right) \sum_{i=1}^m \ln\left(1 + \zeta^* \frac{x_u^i}{\sigma^{(j)}}\right) + \left(1 + \frac{1}{\zeta^{(j)}}\right) \sum_{i=1}^m \ln\left(1 + \zeta^{(j)} \frac{x_u^i}{\sigma^{(j)}}\right)\right\} \\ r_\sigma &= \left(\frac{\sigma^{(j)}}{\sigma^*}\right)^m \exp\left\{\frac{1}{2b_2^2} \left((\sigma^{(j)} - a \cdot \sigma_Z)^2 - (\sigma^* - a \cdot \sigma_Z)^2 \right) \right. \\ &\quad \left. - \left(1 + \frac{1}{\zeta^{(j)}}\right) \sum_{i=1}^m \ln\left(1 + \zeta^{(j)} \frac{x_u^i}{\sigma^*}\right) + \left(1 + \frac{1}{\zeta^{(j)}}\right) \sum_{i=1}^m \ln\left(1 + \zeta^{(j)} \frac{x_u^i}{\sigma^{(j)}}\right)\right\} \end{aligned}$$

7. Simulation Study

Now, we will develop a thorough simulation study in order to compare the accuracy of the three MH methods of estimation: MH, BMH and IPBMH.

We fixed $p = 0.9$ and the threshold $u = q_p$. Furthermore, we take $b = 0$. We sample $n = 2^i$, with $i = 5, \dots, 10$ values from the three baseline distributions considered, with scale $a = 2^j$, $j = -2, -1, 0, 1, 2$. We obtained an MCMC with length 10,000, taking 10,000 values for the burn-in period, 25 values for the thinning and selecting initial values for each sample. Finally, this sequence was repeated 100 times and we considered the mean as the estimator.

In Figure 2, we can see the posterior distribution of the parameters ξ_Z and σ_Z for the different sample sizes n , when the baseline distribution is $\mathcal{L}(0, 1)$ (left), $\mathcal{C}(0, 1)$ (middle) and $\mathcal{N}(0, 1)$ (right), for the methods MH and BMH. For the first one, the distribution is right skewed, although skewness becomes smaller as n increases. BMH offers a point estimation, plotted as a vertical red line.

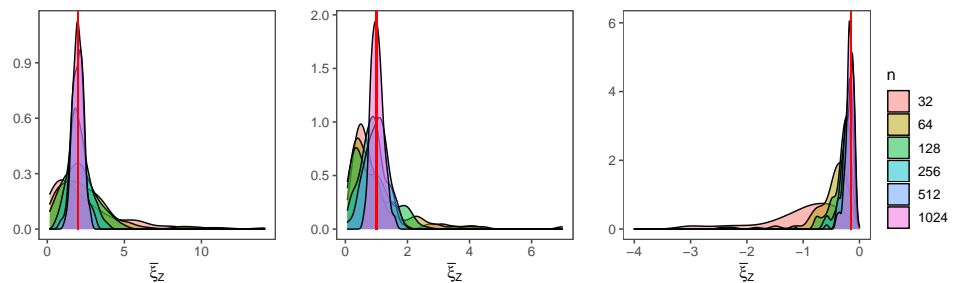


Figure 2. Posterior distribution of ξ_Z for the different values of n , when the baseline distribution is $\mathcal{L}(0, 1)$ (left), $\mathcal{C}(0, 1)$ (middle) and $\mathcal{N}(0, 1)$ (right), for the MH method. The estimation of BMH is plotted as a vertical red line.

In Figure 3, we can see the posterior distribution for σ_Z , for both methods. MH (upper charts) shows much more skewness, such as in the case of ξ_Z , while estimations from BMH (lower charts) are less skewed and dispersed.

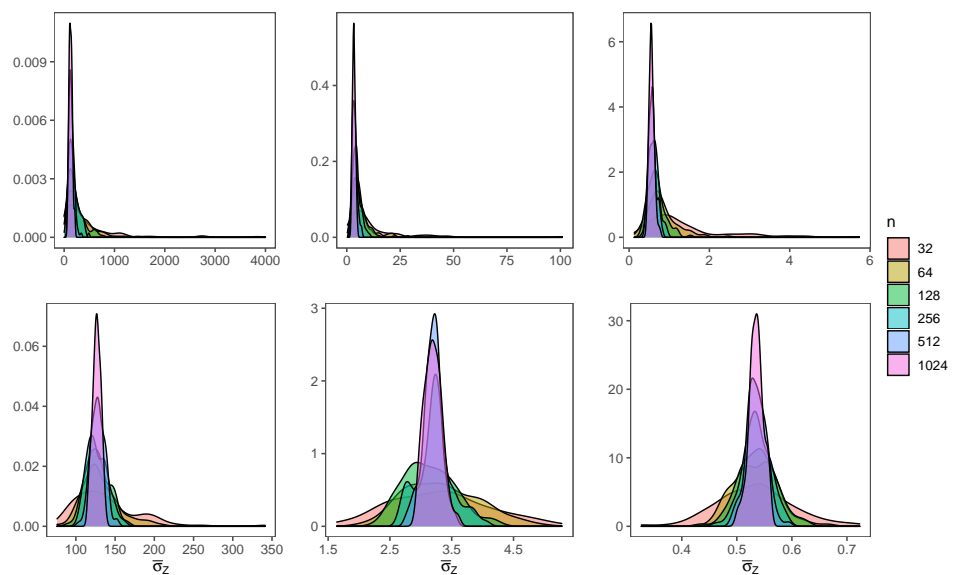


Figure 3. Posterior distribution of σ_Z for the different values of n , when the baseline distribution is $\mathcal{L}(0, 1)$ (left), $\mathcal{C}(0, 1)$ (middle) and $\mathcal{N}(0, 1)$ (right), for the method MH (upper charts) and BMH (lower charts).

Now, we will compare mean absolute errors (MAE) for MH and BMH when $a = 0.25, 1, 4$, and for sample sizes $n = 2^5, 2^7, 2^9$. In Figures 4–6, we can see how BMH provides smaller errors than MH.



Figure 4. MAE for MH and BMH when $a = 0.25, 1, 4$, and for sample sizes $n = 2^5, 2^7, 2^9$, for the baseline distribution $\mathcal{L}(0, 1)$.

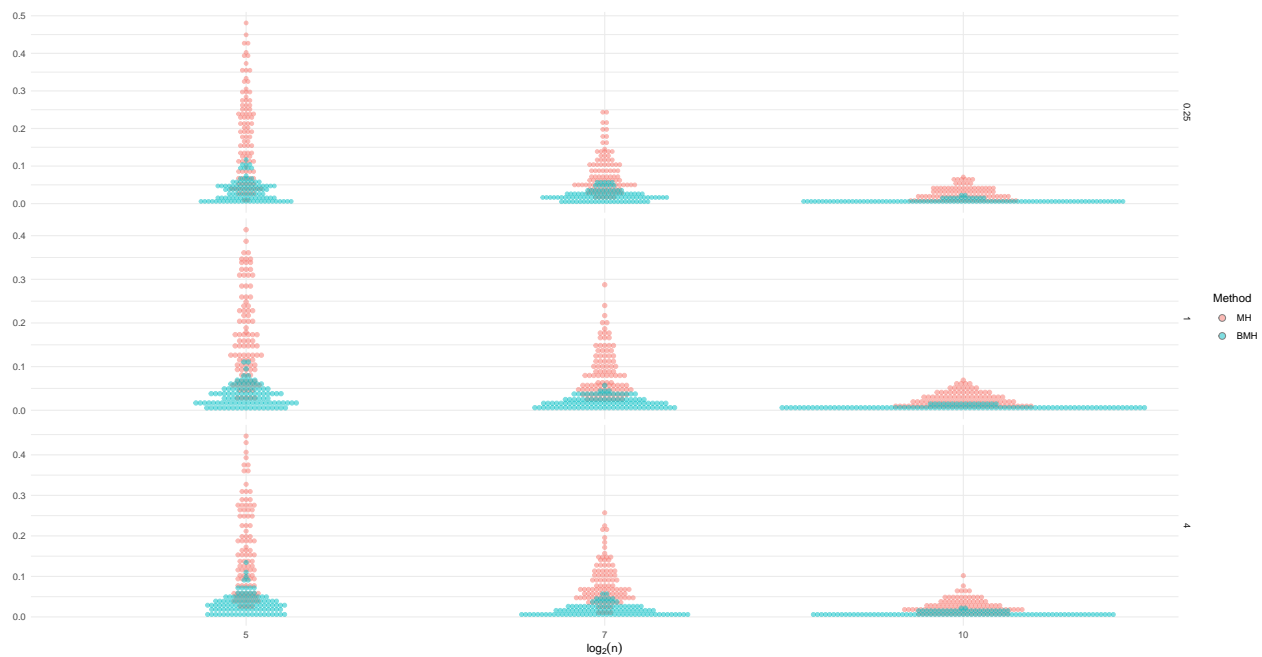


Figure 5. MAE for MH and BMH when $a = 0.25, 1, 4$, and for sample sizes $n = 2^5, 2^7, 2^9$, for the baseline distribution $\mathcal{C}(0, 1)$.

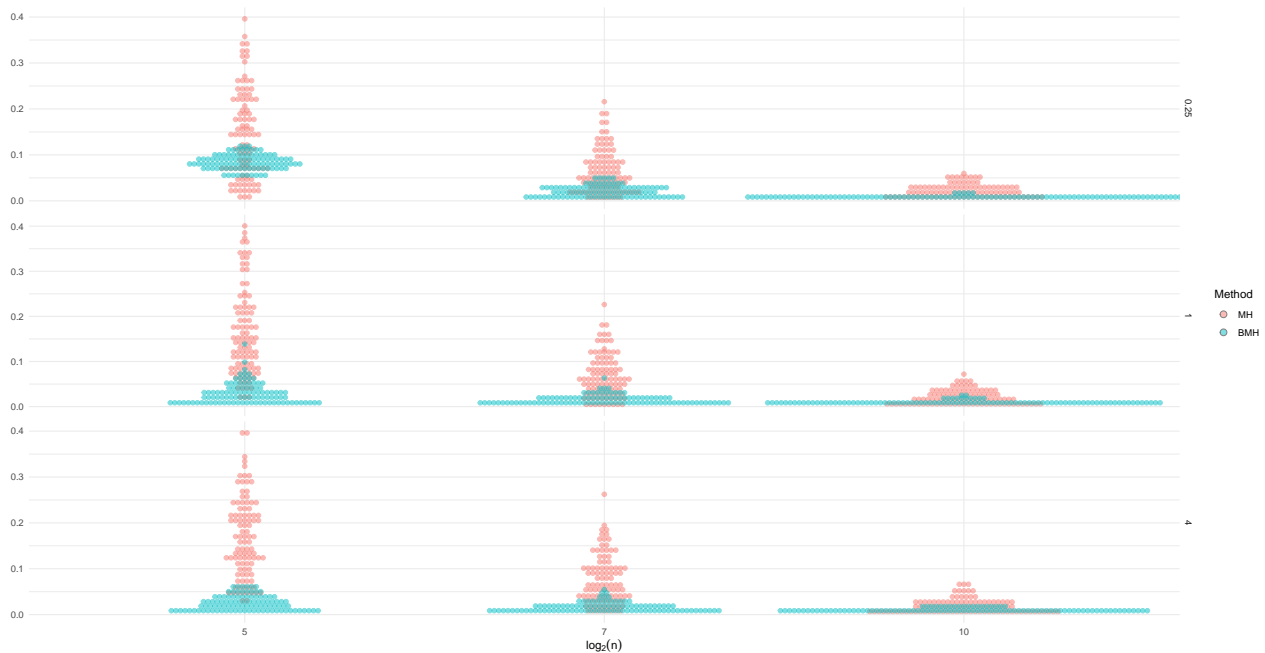


Figure 6. MAE for MH and BMH when $a = 0.25, 1, 4$, and for sample sizes $n = 2^5, 2^7, 2^9$, for the baseline distribution $\mathcal{N}(0, 1)$.

Clearly, BMH provides more accurate estimations for the parameters of the GPD when the baseline distribution is stable and known. However, in practical situations, we might not know which is the baseline distribution, or data could better fit a mixture of distributions rather than a simple one. In these situations, the use of highly informative priors, built with the information available from all the data, could be advisable. To develop this idea, we simulated from different mixtures and computed values of MAE for the three methods, finding that IPBMH is the method that shows the best behavior when data differ from the simple distributions.

In Figure 7, we can see MAE for the three methods, in the case of the mixtures employed ($\alpha = 0.5$): $\alpha F(0, 1/2) + (1 - \alpha)F(0, 2)$ (left charts), $\alpha F(0, 1) + (1 - \alpha)F(1, 1)$ (middle charts) and $\alpha F(0, 1) + (1 - \alpha)F(1, 2)$ (right charts), for the three baseline distributions considered (Lévy, Cauchy, Normal). In general, IPBMH is the most advisable method, especially for the case when data are scarce. Notice that when data approaches one of the pure stable distributions, for example, in the case of Cauchy mixtures (which are still quite similar to a simple Cauchy) and the second mixture for the Normal distribution, BMH and IPBMH show very similar results. However, when the mixtures differ significantly from the simple distribution, the method IPBMH and MH are more advisable than BMH.

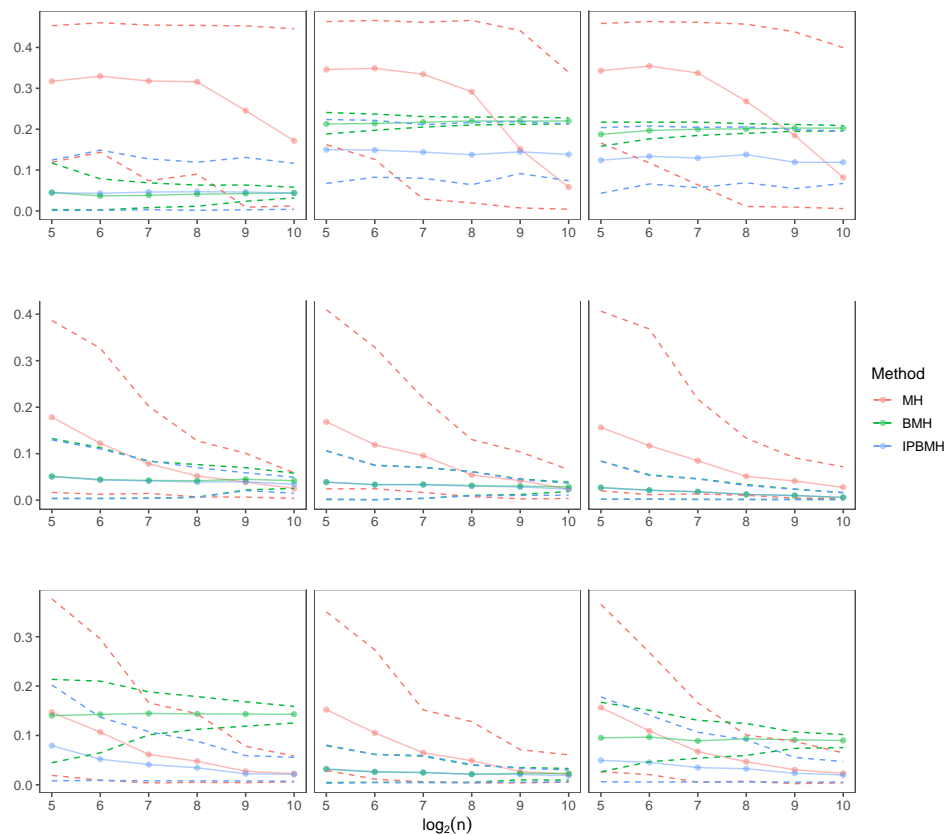


Figure 7. MAE and 2.5%, 97.5% quantiles for the three methods, for $\alpha F(0, 1/2) + (1 - \alpha)F(0, 2)$ (left charts), $\alpha F(0, 1) + (1 - \alpha)F(1, 1)$ (middle charts) and $\alpha F(0, 1) + (1 - \alpha)F(1, 2)$ (right charts), for Lévy (upper charts), Cauchy (medium charts), Normal (lower charts).

8. An Application: PM 2.5 in Badajoz (Spain) during the Period 2011–2020

As we mentioned before, both baseline methods can be generalized for other baseline distributions by studying the relations between the parameters of the baseline distributions and the parameters of the limit GPD. We show an example, employing real data whose baseline distribution can be fitted by a Gamma distribution.

Data from measurements of the levels for diverse air pollutants in many municipalities in Spain are publicly available on the website <https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/>, accessed on 15 January 2022. Particulate matter is a mixture of solid particles and liquid droplets that can be inhaled and cause serious health problems. In particular, we have selected particulate matter less than 2.5 micrometers in diameter, known as PM 2.5, because it is considered to be especially dangerous to human health. In this context, studying the tail distribution above a certain threshold is essential because the World Health Organization recommends not to exceed an average daily value of $25 \mu\text{g}/\text{m}^3$ and not to exceed an average annual value of $10 \mu\text{g}/\text{m}^3$. We studied levels of PM 2.5 measured in $\mu\text{g}/\text{m}^3$ for the last ten years available, 2011–2020, from Badajoz. There are $n = 1066$ observations, and, as can be seen in Figure 8, data can be fitted by a Gamma distribution. As in the previous simulations, $p = 0.90$ and the threshold $u = q_p$, resulting $u = 15 \mu\text{g}/\text{m}^3$.

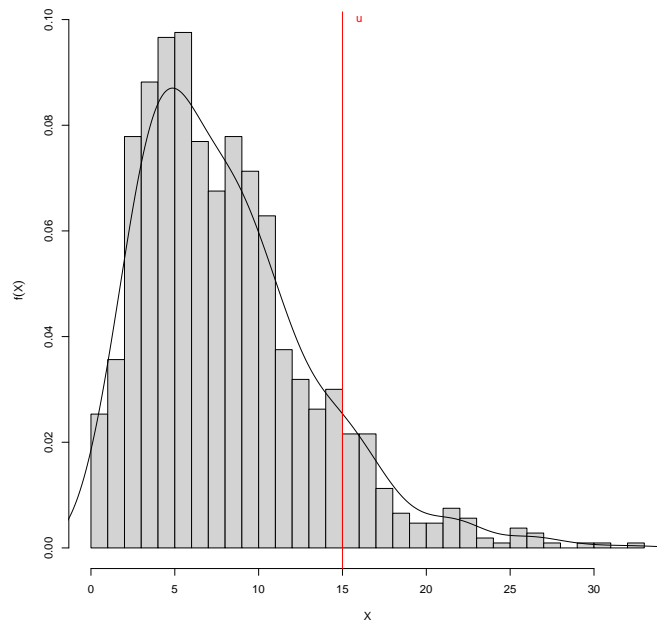


Figure 8. Histogram of PM 2.5 (in $\mu\text{g}/\text{m}^3$) in Badajoz for 2011–2020, threshold $u = q_p = 15$, $p = 0.90$ and density (black curve).

Making a simulation study, similar to the previous ones made for Normal, Lévy and Cauchy, we obtained the following relation between the parameters of the baseline distribution $\Gamma(\alpha, \beta)$ and the parameters of limit GPD(ξ, σ):

$$\hat{\xi} = 0, \hat{\sigma} = \frac{1}{\beta} (1 + 0.22 \log_2 \alpha) \quad (14)$$

Then, we randomly selected 50 data and applied the three methods MHM, BDM and IPBDM to fit the tail data. This proceeding was repeated many times, and we found three possible behaviors, as shown in Figure 9. In the first case (left chart), there is an example of the most usual situation: IPBDM offers intermediate estimations, between MHM and BDM. When there are scarce tail data (middle chart), MH differs significantly from the real density, while BDM and IPBDM provide better estimations. Finally, in the right chart, a common situation is shown in which IPBDM clearly offers the best estimations.

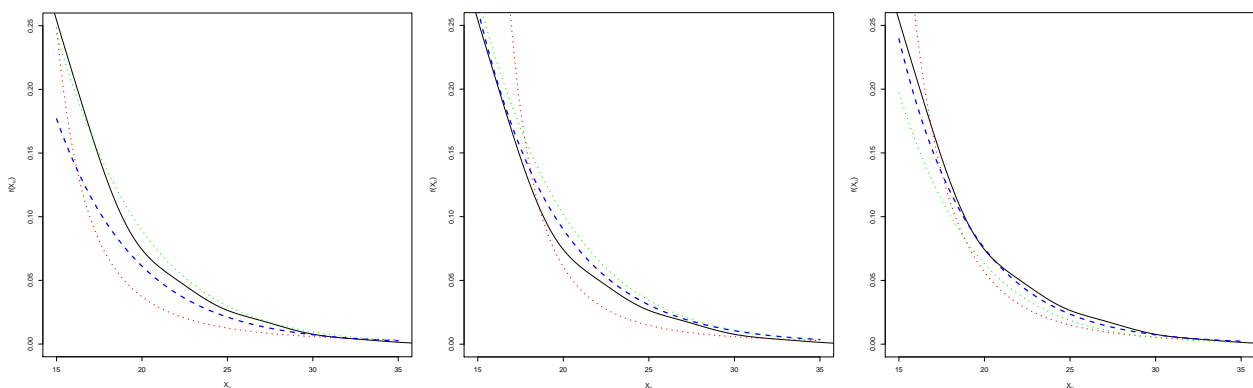


Figure 9. Tail distribution of data, density (black curve), estimations for MHM (red colored), BDM (green colored) and IPBDM (blue colored).

9. Conclusions

1. In the parameter estimation of GPD, usual EVT methods waste a lot of information. We have proposed two MH methods that make the most of all the information available from the data set. They are based on making use of the existing relations between baseline and GPD parameters, through informative priors.
2. When considering the GPD coming from stable baseline distributions, we employed singular properties of stable distributions to simplify the problem (reducing to standard cases) and to provide estimators for the parameters of the GPD.
3. We have achieved very accurate estimations for the parameters of the GPD when the baseline distribution is Cauchy, Normal or Lévy, making use of the properties of stable distributions and MH methods.
4. We have studied the goodness of the estimations for classical MH method and BMH when the baseline distribution is standard Cauchy, Normal or Lévy. Clearly, BMH provides more accurate estimations than MH.
5. In most real situations, data do not fit a simple distribution. We simulated some examples of mixtures of stable distributions and showed that IPBMH provides more accurate estimations than the other methods.
6. These proposals could be generalized for other baseline distributions by studying the relations between the parameters of the baseline distributions and the parameters of the limit GPD. We provide an application with real data to illustrate this situation.

Author Contributions: Investigation, J.M., M.I.P., M.M.P. and E.L.S.; Methodology, J.M., M.I.P., M.M.P. and E.L.S.; Writing—review & editing, J.M., M.I.P., M.M.P. and E.L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Junta de Extremadura, Consejería de Economía, Ciencia y Agenda Digital GR21057, partially supported by Fondo Europeo de Desarrollo Regional (FEDER).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Nogaj, M.; Yiou, P.; Parey, S.; Malek, F.; Naveau, P. Amplitude and frequency of temperature extremes over the North Atlantic region. *Geophys. Res. Lett.* **2006**, *33*. [[CrossRef](#)]
2. Coelho, C.A.S.; Ferro, C.A.T.; Stephenson, D.B.; Steinskog, D.J. Methods for Exploring Spatial and Temporal Variability of Extreme Events in Climate Data. *J. Clim.* **2008**, *21*, 2072–2092. [[CrossRef](#)]
3. Acero, F.J.; Fernández-Fernández, M.I.; Carrasco, V.M.S.; Parey, S.; Hoang, T.T.H.; Dacunha-Castelle, D.; García, J.A. Changes in heat wave characteristics over Extremadura (SW Spain). *Theor. Appl. Climatol.* **2017**, *133*, 605–617. [[CrossRef](#)]
4. García, J.A.; Pizarro, M.M.; Acero, F.J.; Parra, M.I. A Bayesian Hierarchical Spatial Copula Model: An Application to Extreme Temperatures in Extremadura (Spain). *Atmosphere* **2021**, *12*, 897. [[CrossRef](#)]
5. García, J.; Gallego, M.C.; Serrano, A.; Vaquero, J. Trends in Block-Seasonal Extreme Rainfall over the Iberian Peninsula in the Second Half of the Twentieth Century. *J. Clim.* **2007**, *20*, 113–130. [[CrossRef](#)]
6. Re, M.; Barros, V.R. Extreme rainfalls in SE South America. *Clim. Chang.* **2009**, *96*, 119–136. [[CrossRef](#)]
7. Acero, F.J.; Gallego, M.C.; García, J.A. Multi-day rainfall trends over the Iberian Peninsula. *Theor. Appl. Climatol.* **2011**, *108*, 411–423. [[CrossRef](#)]
8. Acero, F.J.; García, J.A.; Gallego, M.C. Peaks-over-Threshold Study of Trends in Extreme Rainfall over the Iberian Peninsula. *J. Clim.* **2011**, *24*, 1089–1105. [[CrossRef](#)]
9. Acero, F.J.; Parey, S.; Hoang, T.T.H.; Dacunha-Castelle, D.; García, J.A.; Gallego, M.C. Non-stationary future return levels for extreme rainfall over Extremadura (southwestern Iberian Peninsula). *Hydrol. Sci. J.* **2017**, *62*, 1394–1411. [[CrossRef](#)]
10. Wi, S.; Valdés, J.B.; Steinschneider, S.; Kim, T.W. Non-stationary frequency analysis of extreme precipitation in South Korea using peaks-over-threshold and annual maxima. *Stoch. Environ. Res. Risk Assess.* **2015**, *30*, 583–606. [[CrossRef](#)]
11. Ramos, A.A. Extreme value theory and the solar cycle. *Astron. Astrophys.* **2007**, *472*, 293–298. [[CrossRef](#)]

12. Acero, F.J.; Carrasco, V.M.S.; Gallego, M.C.; García, J.A.; Vaquero, J.M. Extreme Value Theory and the New Sunspot Number Series. *Astrophys. J.* **2017**, *839*, 98. [[CrossRef](#)]
13. Acero, F.J.; Gallego, M.C.; García, J.A.; Usoskin, I.G.; Vaquero, J.M. Extreme Value Theory Applied to the Millennial Sunspot Number Series. *Astrophys. J.* **2018**, *853*, 80. [[CrossRef](#)]
14. Longin, F.M. Value at Risk and Extreme Values. *IFAC Proc. Vol.* **1998**, *31*, 45–49. [[CrossRef](#)]
15. Singh, A.K.; Allen, D.E.; Powell, R.J. Value at risk estimation using extreme value theory. In Proceedings of the 19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011.
16. Embrechts, P.; Klüppelberg, C.; Mikosch, T. *Modelling Extremal Events for Insurance and Finance*; Springer: Berlin, Germany; New York, NY, USA, 1997.
17. Trzpiot, G.; Majewska, J. Estimation of Value at Risk: Extreme value and robust approaches. *Oper. Res. Decis.* **2010**, *20*, 131–143.
18. Magnou, G. An Application of Extreme Value Theory for Measuring Financial Risk in the Uruguayan Pension Fund. *Compend. Cuad. Econ. Adm.* **2017**, *4*, 1–19.
19. Gilli, M.; Këllezli, E. An application of extreme value theory for measuring financial risk. *Comput. Econ.* **2006**, *27*, 207–228. [[CrossRef](#)]
20. Bali, T.G. A generalized extreme value approach to financial risk measurement. *J. Money Credit Bank.* **2007**, *39*, 1613–1649. [[CrossRef](#)]
21. Castillo, E.; Hadi, A.S.; Balakrishnan, N.; Sarabia, J.M. *Extreme Value and Related Models with Applications in Engineering and Science*; Wiley-Interscience: Hoboken, NJ, USA, 2004.
22. Gumbel, E.J. *Statistics of Extremes (Dover Books on Mathematics)*; Dover Publications: Mineola, NY, USA, 2012.
23. Balkema, A.A.; De Haan, L. Residual life time at great age. *Ann. Probab.* **1974**, *2*, 792–804. [[CrossRef](#)]
24. Pickands, J. Statistical inference using extreme order statistics. *Ann. Stat.* **1975**, *3*, 119–131.
25. De Zea Bermudez, P.; Kotz, S. Parameter estimation of the generalized Pareto distribution—Part I. *J. Stat. Plan. Inference* **2010**, *140*, 1353–1373. [[CrossRef](#)]
26. De Zea Bermudez, P.; Kotz, S. Parameter estimation of the generalized Pareto distribution—Part II. *J. Stat. Plan. Inference* **2010**, *140*, 1374–1388. [[CrossRef](#)]
27. Korkmaz, M.Ç.; Altun, E.; Yousof, H.M.; Afify, A.Z.; Nadarajah, S. The Burr X Pareto Distribution: Properties, Applications and VaR Estimation. *J. Risk Financ. Manag.* **2018**, *11*, 1. [[CrossRef](#)]
28. Ihtisham, S.; Khalil, A.; Manzoor, S.; Khan, S.A.; Ali, A. Alpha-Power Pareto distribution: Its properties and applications. *PLoS ONE* **2019**, *14*, e0218027. [[CrossRef](#)]
29. Arnold, B.C.; Press, S.J. Bayesian estimation and prediction for Pareto data. *J. Am. Stat. Assoc.* **1989**, *84*, 1079–1084. [[CrossRef](#)]
30. Diebolt, J.; El-Aroui, M.A.; Garrido, M.; Girard, S. Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling. *Extremes* **2005**, *8*, 57–78. [[CrossRef](#)]
31. De Zea Bermudez, P.; Turkman, M.A. Bayesian approach to parameter estimation of the generalized Pareto distribution. *Test* **2003**, *12*, 259–277. [[CrossRef](#)]
32. Behrens, C.N.; Lopes, H.F.; Gamerman, D. Bayesian analysis of extreme events with threshold estimation. *Stat. Model.* **2004**, *4*, 227–244. [[CrossRef](#)]
33. Castellanos, M.E.; Cabras, S. A default Bayesian procedure for the generalized Pareto distribution. *J. Stat. Plan. Inference* **2007**, *137*, 473–483. [[CrossRef](#)]
34. Zhang, J.; Stephens, M.A. A new and efficient estimation method for the generalized Pareto distribution. *Technometrics* **2009**, *51*, 316–325. [[CrossRef](#)]
35. Martín, J.; Parra, M.I.; Pizarro, M.M.; Sanjuán, E.L. Baseline Methods for Bayesian Inference in Gumbel distribution. *Entropy* **2020**, *22*, 1267. [[CrossRef](#)]
36. Lévy, P. *Calcul des Probabilités*; Gauthier Villars: Paris, France, 1925.
37. Nolan, J.P. *Stable Distributions—Models for Heavy Tailed Data*; Birkhauser: Boston, MA, USA, 2018; Chapter 1; Available online: <http://fs2.american.edu/jpnolan/www/stable/stable.html> (accessed on 15 January 2022).
38. Plummer, M.; Best, N.; Cowles, K.; Vines, K. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **2006**, *6*, 7–11.

Capítulo 4

Artículo C: Un modelo jerárquico Bayesiano espacial con cópula: Una aplicación a temperaturas extremas en Extremadura (España)

Autores:

J. Agustín García¹, Mario M. Pizarro², F. Javier Acero¹, M. Isabel Parra²

¹ Departamento de Física, Universidad de Extremadura

² Departamento de Matemáticas, Universidad de Extremadura

Revista: Atmosphere, 12(7), 897, 2021

DOI: [10.3390/atmos12070897](https://doi.org/10.3390/atmos12070897)

Resumen:

En este trabajo se propone un modelo jerárquico Bayesiano con una cópula Gaussiana y distribución de Valores Extremos Generalizada (GEV) como distribuciones marginales, para describir la dependencia espacial existente en los datos. Este modelo de cópula espacial se aplica a temperaturas extremas durante el periodo de verano observadas en estaciones climatológicas de la región de Extremadura, durante el periodo 1980-2015, y se compara con el modelo espacial sin cópula. El modelo jerárquico Bayesiano se implementa con cadenas de Markov de Monte Carlo (MCMC), permitiendo así estimar la distribución de los parámetros del modelo.

Los resultados muestran que el parámetro de forma de la distribución GEV toma valores negativos constantes, el parámetro de localización depende de la altitud y los valores del parámetro de escala se concentran alrededor del mismo valor en toda la región. Además, el modelo con cópula elegido presenta valores del Criterio DIC más bajos cuando se asumen distribuciones espaciales para los parámetros de localización y escala de la distribución GEV, que cuando se toma este último parámetro como constante en la región.



Article

A Bayesian Hierarchical Spatial Copula Model: An Application to Extreme Temperatures in Extremadura (Spain)

J. Agustín García ^{1,†}, Mario M. Pizarro ^{2,*}, F. Javier Acero ^{1,†} and M. Isabel Parra ^{2,†}

¹ Departamento de Física, Universidad de Extremadura, Avenida de Elvas, 06006 Badajoz, Spain; agustin@unex.es (J.A.G.); fjacero@unex.es (F.J.A.)

² Departamento de Matemáticas, Universidad de Extremadura, Avenida de Elvas, 06006 Badajoz, Spain; mipa@unex.es

* Correspondence: mariomp@unex.es

† These authors contributed equally to this work.

Abstract: A Bayesian hierarchical framework with a Gaussian copula and a generalized extreme value (GEV) marginal distribution is proposed for the description of spatial dependencies in data. This spatial copula model was applied to extreme summer temperatures over the Extremadura Region, in the southwest of Spain, during the period 1980–2015, and compared with the spatial noncopula model. The Bayesian hierarchical model was implemented with a Monte Carlo Markov Chain (MCMC) method that allows the distribution of the model's parameters to be estimated. The results show the GEV distribution's shape parameter to take constant negative values, the location parameter to be altitude dependent, and the scale parameter values to be concentrated around the same value throughout the region. Further, the spatial copula model chosen presents lower deviance information criterion (DIC) values when spatial distributions are assumed for the GEV distribution's location and scale parameters than when the scale parameter is taken to be constant over the region.

Keywords: Bayesian hierarchical model; extreme temperature; Gaussian copula; generalized extreme value distribution



Citation: García, J.A.; Pizarro, M.M.; Acero, F.J.; Parra, M.I. A Bayesian Hierarchical Spatial Copula Model: An Application to Extreme Temperatures in Extremadura (Spain). *Atmosphere* **2021**, *12*, 897. <https://doi.org/10.3390/atmos12070897>

Academic Editors: Kreso Pandzic, Tanja Likso and Ognjen Bonacci

Received: 31 May 2021

Accepted: 7 July 2021

Published: 10 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Extreme events tend to occur naturally as topics of importance in several sciences—climatology, hydrology, engineering, etc.—but also in finance (financial crisis studies) and in the insurance industry. Extreme Value Theory (EVT) is a widely used statistical tool with which to address their study. It is used in several particular scientific fields to model and predict extreme events of precipitation [1–5], temperature [6–8], solar climatology [9,10], and financial crises [11,12].

In climatology, given the spatial nature of extreme events, it is useful to apply a spatial theory as this will improve the accuracy when estimating parameter distributions by sharing information from similar sites [13,14]. Several theories are used to address the problem of spatial extremes, some examples are max-stable processes [15], Bayesian hierarchical models [16–18], and copula theory [19–22].

Copula theory is increasingly being used in multivariate extreme value models in climatology [23], since Sklar's theorem [24] allows the construction of joint distributions using the desired univariate marginal distributions. Of the different copula families, there stand out the Archimedean copulas, extreme value copulas, and elliptical copulas (in particular, the Gaussian and the Student-t copulas). These copulas constitute a convenient tool for modeling dependence in a non-Gaussian and high-dimensional context [25]. Indeed, copula theory is based on describing the dependence structure regardless of any marginal distributions [26], which makes copulas an ideal candidate for modeling dependence.

Wikle et al. [27] proposed a general Bayesian hierarchical framework in which to describe the spatial variability of the distribution of some environmental variable. Their

model comprises various layers. For example, in the first layer, it is assumed that the data follow a distribution with unknown parameters, while in a second layer, the variability of these parameters is modeled spatially by regression techniques. This kind of model has been used in many extreme rainfall studies [28–31] and also in temperature studies [32,33]. In most of these works however, it is assumed that the data are spatially independent given the values of the parameters of their distribution.

As pointed out by Cooley et al. [14], the spatial dependence among the parameters is related to what the authors called spatial climate dependence since the climate of a location or region is described by the statistical distribution of the variable of interest—in our case, the temperature. For readers interested in a climate parameter such as the return period or the return level, the hierarchical model used in the previously cited papers is enough to estimate those parameters. However, if someone is interested in, for example, transferring information from a gauged site to an ungauged site, the spatial dependence among the data must be taken into account in what Renard [20] called spatial weather dependence, because both dependencies are involved in the transferring information process. The spatial climate dependence allows the transfer of information from one place to another. However, the correlation between observatories decreases the content of the information available in nearby observatories [20]. Therefore, there must be an increase in uncertainty in the information transferred to the ungauged site when the correlation between observatories is taken into account. This key point of considering the spatial dependence among the data can be accomplished by means of a copula model.

In this sense, the main goal of the present work is to address spatial dependence by incorporating a copula into this model—specifically, a Gaussian copula (GC). An application of the proposed model is presented with the maximum temperatures recorded in the Extremadura Region, in southwestern Spain, for the period 1980–2015.

This communication is structured as follows. Section 2 describes the hierarchical model with copula, and Section 3 describes the proposed posterior distribution model and its use when inferring the GEV distribution parameters. The selected data are briefly described in Section 4, and Section 5 presents the results of applying the model to extreme temperatures in the Extremadura Region (Spain). Finally, some conclusions are drawn in Section 6.

2. Statistical Model

A Bayesian approach was chosen to address the spatial extremes problem because of its flexibility, the possibility of adding further elements or layers, and its adaptability to situations in which complex variations in the parameters of the extremes values distribution appear. In addition, one of the disadvantages shown in [18] of assuming conditional independence among observed data is resolved by introducing a Gaussian copula to control the existence of dependence, with this fact being the main difference with that model.

The proposed theoretical Bayesian hierarchical framework can be factored into the three stages shown in Figure 1. The first stage (Section 2.1) allows one to model the observations' joint distribution with a Gaussian copula and at-site marginal GEV distributions. The second stage (Section 2.2) models the GEV parameter variances with a latent process by means of a conditional probability. In the third stage (Section 2.3), prior distributions are given for the model's parameters. Thus, the posterior distribution of the parameters is calculated using Bayes' theorem (see Section 3). In the following, the model's stages will be described in more detail.

Stage 1 (data level) $P(\text{data} \mid \text{process, parameters})$

Stage 2 (process level) $P(\text{process} \mid \text{parameters})$

Stage 3 (prior distribution) $P(\text{parameters})$

Figure 1. Stages of the Bayesian hierarchical model.

We shall denote by Y_{st} the block maximum of the variable of interest, with $s \in S = \{s_1, \dots, s_M\}$ and $t \in T = \{t_1, \dots, t_N\}$. For simplicity and consistency with the application, we shall refer to s as the site and t as the time. Further, for each s and t , Y_{st} follows a specific distribution whose parameters vary spatially.

2.1. Data Level

According to the theorems of Gnedenko [34] and Fisher and Tippett [35], asymptotically, Y_s has a GEV distribution

$$Y_s \sim GEV(\mu_s, \sigma_s, \xi_s) \tag{1}$$

with cumulative distribution function (cdf)

$$P(Y_s \leq y) = \exp\left\{-\left[1 + \xi_s\left(\frac{y - \mu_s}{\sigma_s}\right)\right]^{-1/\xi_s}\right\}, \tag{2}$$

where $1 + \xi_s((y - \mu_s)/\sigma_s) \geq 0$ and in which μ , σ , and ξ are the location, scale, and shape parameters, respectively.

The location parameter explains the mean values of the extreme value distribution, the scale one is referred to the variability, and the shape parameter determines the rate of decay of the upper tail of the distribution. Shape parameter values below zero indicate that the distribution has an upper bound showing that the maximum values are not getting large, and values above or equal to zero indicate that the distribution has no upper limit showing that maximum values are getting infinitely large [36].

For any set of M sites, an M -dimensional Gaussian copula is assumed for the multivariate distribution of the data, with pairwise correlation matrix \mathcal{C} and marginal distributions $\{GEV(\mu_s, \sigma_s, \xi_s)\}_{s \in S}$

$$(Y_{s_1 t}, \dots, Y_{s_M t}) \sim GC_M(\mathcal{C}, \{\mu_s, \sigma_s, \xi_s\}_{s \in S}), \tag{3}$$

for which the probability density function (pdf) is shown in Appendix A.

Geostatistical models use correlation matrices that capture the positive relationship between different stations through their distance. In this case, the Gaussian copula collects this information in the correlation matrix \mathcal{C} which is positively defined. Then, it is assumed that the correlation between data from two sites depends on the distance between them. In particular, the elements of the correlation matrix are defined as

$$C_{ij} = \begin{cases} c_0 \cdot \exp\left\{-\frac{\|\vec{x}(s_i) - \vec{x}(s_j)\|}{c_1}\right\}, & i \neq j \\ 1, & i = j \end{cases} \tag{4}$$

where c_0 and c_1 are unknown parameters, $\vec{x}(s)$ is the geographical position (longitude, latitude) of the site $s \in S$, and $\|\vec{x}(s_i) - \vec{x}(s_j)\|$ are the distances between sites s_i and s_j , for $i, j = 1, \dots, M$. Moreover, independence is assumed between observations $(Y_{s_1 t}, \dots, Y_{s_M t})$ and $(Y_{s_1 t'}, \dots, Y_{s_M t'})$ for $t, t' \in T$ such that $t \neq t'$.

The multivariate distribution of the data thus comprises the following components: at-site distribution (1), Gaussian copula to model the spatial dependence (3) and correlation matrix (4).

2.2. Process Level

The second stage of the hierarchical model consists of describing the variance of the at-site GEV parameters (location— μ , scale— σ , and shape— ξ) through a Gaussian spatial process whose mean depends on covariates that describe the site’s characteristics. A spatial regression model as described by Garcia et al. [18] is used:

$$\begin{aligned}\mu(\mathbf{s}) &= \mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_\mu + W_\mu(\mathbf{s}) + \epsilon_\mu(\mathbf{s}) \\ \sigma(\mathbf{s}) &= \mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_\sigma + W_\sigma(\mathbf{s}) + \epsilon_\sigma(\mathbf{s}) \\ \zeta(\mathbf{s}) &= \mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_\zeta + W_\zeta(\mathbf{s}) + \epsilon_\zeta(\mathbf{s})\end{aligned}\quad (5)$$

where $\mathbf{s} = (s_1, \dots, s_M)$ denotes a site vector and p_μ, p_σ , and p_ζ denote the number of regression parameters in each case. To simplify use of the notation, the parameters μ, σ , or ζ will be represented by k , so that $\mathbf{X}(\mathbf{s})$ represents p_k spatial covariates (geographic coordinates), $\boldsymbol{\alpha}_k$ is a set of p_k regression parameters (including the intercept), $W_k(\mathbf{s})$ represents a spatial model that captures the dependencies between different sites, and $\epsilon_k(\mathbf{s})$ is the noise not included in the spatial model.

In general, we shall denote the proposed models by BHGCM- $p_\mu p_\sigma p_\zeta$, and the noncopula models described by Garcia et al. [18] by BHM- $p_\mu p_\sigma p_\zeta$.

With regard to Equations (5), for $p_k = 1$, we shall assume that no covariate associated with the site characteristics is involved, i.e.,

$$\mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_k = \alpha_{k1}, \quad (6)$$

and for $p_k = 2$, the only covariate is h_s , the altitude of site s , i.e.,

$$\mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_k = \alpha_{k1} + \alpha_{k2} \cdot h(\mathbf{s}). \quad (7)$$

As a particular case, for $p_k = 0$, we shall assume that the parameter k is constant, and hence, the spatial model does not intervene, i.e., $W_k(\mathbf{s}) = \mathbf{0}$.

In addition, a position-independent Gaussian model, i.e., $\mathcal{N}(0, \tau_k^2)$, is adopted for the pure noise effect $\epsilon_k(\mathbf{s})$. The spatial term $W_k(\mathbf{s})$ was considered to be a random variable with an M -dimensional normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$, where the covariance matrix $\boldsymbol{\Sigma}_k$ follows the exponential model

$$\boldsymbol{\Sigma}_k(i, j) = \beta_{k0} \cdot \exp\left\{-\frac{\|\bar{\mathbf{x}}(s_i) - \bar{\mathbf{x}}(s_j)\|}{\beta_{k1}}\right\}, \quad i, j = 1, \dots, M, \quad (8)$$

where β_{k0} (the sill) and β_{k1} (the range) are unknown parameters.

The random variables on the left-hand side of Equation (5) are assumed to have an M -dimensional normal distribution:

$$\begin{aligned}P(\mu(\mathbf{s})) &= \mathcal{N}\left(\mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_\mu + W_\mu(\mathbf{s}), \tau_\mu^2 \cdot \text{Id}_M\right) \\ P(\sigma(\mathbf{s})) &= \mathcal{N}\left(\mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_\sigma + W_\sigma(\mathbf{s}), \tau_\sigma^2 \cdot \text{Id}_M\right) \\ P(\zeta(\mathbf{s})) &= \mathcal{N}\left(\mathbf{X}(\mathbf{s}) \cdot \boldsymbol{\alpha}_\zeta + W_\zeta(\mathbf{s}), \tau_\zeta^2 \cdot \text{Id}_M\right)\end{aligned}\quad (9)$$

2.3. Prior Distribution

The Bayesian framework requires the prior distributions of the parameters included in the model to be specified. The prior distribution for the spatial regression parameters $\boldsymbol{\alpha}_k$ was a p_k -dimensional normal distribution with hyperparameters chosen such that the distribution was either non- or only weakly informative. Inverse gamma distributions were taken for the sill β_{k0} and the variance τ_k^2 parameters, and a gamma distribution for the range β_{k1} . For the Gaussian copula parameters, c_0 and c_1 , uniform prior distributions were assumed— $\mathcal{U}(0,1)$ and $\mathcal{U}(0, 1000)$, respectively. A normal distribution with mean 0 was taken for the shape parameter. The parameters were assumed to be mutually independent.

3. Estimation

We shall apply two proposed models to the extreme temperature data (see Section 5). In the first, denoted BHGCM-200, the scale and shape parameters are constant and the location parameter is modeled as in Equation (5). In the second, denoted BHGCM-210, the shape parameter is constant and the location and scale parameters are modeled as in

Equation (5), with $p_\mu = 2$ and $p_\sigma = 1$, respectively. The model that best fits the data will be compared with the equivalent noncopula version.

As mentioned above, Bayes' theorem allows one to calculate the posterior distribution of a proposed model as being proportional to the product of the probabilities described in Figure 1. To simulate the posterior distribution of each of the proposed models, a Markov Chain Monte Carlo (MCMC) method was applied, in particular, using a Gibbs sampler with embedded Metropolis–Hastings steps [37] and—as appropriate given the characteristics of these methods—the Gelman–Rubin diagonal convergence test [38]. For this last test, the CODA package [39] of the R language was used.

Four parallel chains with sizes of 30,000 values were constructed starting at different points. For each chain, 10,000 values were used as burn-in, leaving 20,000 values less 10 taken for thinning. The last 2000 values of each chain were combined to form a single chain of 8000 values with which to construct the posterior distribution.

The code used to carry out the simulations was written in FORTRAN, closely following the procedure set out. Maps were prepared using the R package `ggplot2` [40], with the geographical coordinates provided by Spain's National Centre for Geographic Information (Centro Nacional de Información Geográfica, CNIG) [41].

3.1. Posterior Distribution

For each GEV parameter that changes in the model, the hierarchical framework described in Section 2 estimates the following unknowns: Gaussian copula parameters (c_0 and c_1), regression parameters (α_k), sill parameter (β_{k0}), range parameter (β_{k1}), and variance parameter (τ_k^2).

For the BHGCM-200 model, the posterior distribution is

$$P\left(c_0, c_1, \mu(\mathbf{s}), \sigma, \xi, W_\mu(\mathbf{s}), \alpha_\mu, \beta_{\mu 0}, \beta_{\mu 1}, \tau_\mu^2 \mid Y, h(\mathbf{s})\right), \tag{10}$$

and for the BHGCM-210 model, it is

$$P\left(c_0, c_1, \mu(\mathbf{s}), \sigma(\mathbf{s}), \xi, W_\mu(\mathbf{s}), W_\sigma(\mathbf{s}), \alpha_\mu, \alpha_{\sigma 1}, \beta_{\mu 0}, \beta_{\mu 1}, \beta_{\sigma 0}, \beta_{\sigma 1}, \tau_\mu^2, \tau_\sigma^2 \mid Y, h(\mathbf{s})\right). \tag{11}$$

Assuming independence between the observations $(Y_{s_1 t}, \dots, Y_{s_M t})$ and $(Y_{s_1 t'}, \dots, Y_{s_M t'})$ ($t, t' \in T$ with $t \neq t'$), the likelihood function of the observations is given by

$$L(c_0, c_1, \mu(\mathbf{s}), \sigma(\mathbf{s}), \xi \mid Y) := P(Y \mid c_0, c_1, \mu(\mathbf{s}), \sigma(\mathbf{s}), \xi) = \prod_{t \in T} f_{GC}(y_{s_1 t}, \dots, y_{s_M t}). \tag{12}$$

This shows the details of Equations (A3)–(A5) in Appendix B.

3.2. Assessment of the Models' Goodness-Of-Fit

The deviance information criterion (DIC) described by Spiegelhalter et al. [42] was used to choose the model that best fits the observed data. The best model has the lowest DIC value. The parameter values needed to calculate the DIC were those determined through the MCMC procedure. The criterion is defined as

$$DIC = \bar{D}_\theta + p_\theta, \tag{13}$$

where

- (a) θ is the parameter vector of interest in the model (GEV parameters in a BHM model, and GEV and Gaussian copula parameters in a BHGCM model).
- (b) $\bar{D}_\theta = E[D(\theta)]$ measures the model's goodness-of-fit, where the deviance $D(\theta) = -2 \cdot \ln L(\theta \mid Y)$, i.e., -2 times the logarithm of the likelihood of the random variable Y under study. In a BHGCM model, the likelihood is defined by Equation (A5), and in a BHM model, by the GEV pdf.

- (c) $p_\theta = \bar{D}_\theta - D(\bar{\theta})$ is a parameter that controls the complexity of the model (effective number of parameters), where $D(\bar{\theta})$ is the deviance of the posterior mean $\bar{\theta}$ of the parameter of interest.

In particular, the goodness-of-fit was used to compare the proposed copula models (BHGC M - $p_\mu p_\sigma p_\zeta$), and the resulting model with the lowest DIC value will be contrasted with the equivalent noncopula model described by García et al. [18].

3.3. Inference

The MCMC method gave the posterior distribution of the GEV distribution's parameters at the gauged sites, \mathbf{s} . A set of replicates of size $nsim$ was generated from the posterior distribution for the parameter vector $(c_0^{(l)}, c_1^{(l)}, \boldsymbol{\alpha}_\mu^{(l)}, \boldsymbol{\beta}_\mu^{(l)}, \tau_\mu^{(l)}, \boldsymbol{\alpha}_\sigma^{(l)}, \boldsymbol{\beta}_\sigma^{(l)}, \tau_\sigma^{(l)})_{l=1 \dots nsim}$. This sample was then used to infer the GEV distribution's parameters at an ungauged site \tilde{s} by applying the following algorithm (Algorithm 1):

Algorithm 1 Ungauged Site

Do for $l = 1 \dots nsim$:

- Using the well-known formula for conditional Gaussian distributions, generate $W_k(\tilde{s})$ with a normal distribution of mean μ_{cond} and standard deviation σ_{cond} given by

$$\begin{aligned}\mu_{cond} &= \boldsymbol{\Omega}^{(l)} \cdot (\boldsymbol{\Sigma}^{(l)})^{-1} \cdot W_k^{(l)}(\mathbf{s}) \\ \sigma_{cond} &= \beta_{k0} - \boldsymbol{\Omega}^{(l)} \cdot (\boldsymbol{\Sigma}^{(l)})^{-1} \cdot (\boldsymbol{\Omega}^{(l)})^t\end{aligned}\quad (14)$$

where $W_k^{(l)}(\mathbf{s})$ is the $1 \times M$ vector generated with the gauged sites \mathbf{s} , $\boldsymbol{\Sigma}_k$ is the $M \times M$ covariance matrix of the spatial model for the gauged sites \mathbf{s} , and $\boldsymbol{\Omega}^{(l)}$ is the $1 \times M$ vector of covariances between \tilde{s} and the gauged sites \mathbf{s} .

- Compute the GEV parameters for an ungauged site \tilde{s} , $(\mu_{\tilde{s}}^{(l)}, \sigma_{\tilde{s}}^{(l)}, \zeta_{\tilde{s}}^{(l)})$, from the regression model (5).
-

In addition, the proposed theoretical model provides observations at ungauged sites with a spatial dependence on other sites that is controlled by a Gaussian copula. Algorithm 2 is the scheme used to simulate these observations. This algorithm provides a sample of the posterior predictive distribution (PPD) of observations defined by Gelman [43] as

$$p(\tilde{y}|Y) = \int p(\tilde{y}|\theta) \cdot L(\theta|Y) d\theta, \quad (15)$$

where θ is the parameter vector.

Algorithm 2 Observations

Do for $l = 1 \dots nsim$:

- Calculate the GEV distribution's parameters for an ungauged site \tilde{s} with Algorithm 1, taking into account the GEV parameters of the gauged sites \mathbf{s} obtained with the MCMC generated sample.
 - Generate a value of the Gaussian copula $(u_{s_1}^{(l)}, \dots, u_{s_M}^{(l)}, u_{\tilde{s}}^{(l)})$ with the $(M+1) \times (M+1)$ correlation matrix, $\mathcal{C}^{(l)}$, between \tilde{s} and the gauged sites \mathbf{s} , given by Equation (4).
 - Invert the value of the copula in accordance with the relationship $y_s^{(l)} = F_{GEV}^{-1}(u_s^{(l)} | \mu_s^{(l)}, \sigma_s^{(l)}, \zeta_s^{(l)})$ for all $s = s_1, \dots, s_M, \tilde{s}$, thus yielding a vector of observations $\mathbf{y}^{(l)} = (y_{s_1}^{(l)}, \dots, y_{s_M}^{(l)}, y_{\tilde{s}}^{(l)})$.
-

Thus, the sample $(\mathbf{y}^{(l)})_{l=1 \dots nsim}$ is a realization of the posterior predictive distribution of observations.

4. Data

The data used in this study are annual maximum observed temperatures at a set of meteorological observatories distributed over the Extremadura Region (Spain), from 1980

to 2015. These time series were provided by Spain’s State Meteorological Agency (Agencia Estatal de Meteorología, AEMET). Figure 2 shows the location of this region within Spain and the spatial distribution of the observatories considered. In particular, there are $M = 28$ meteorological observatories, each providing a time series of $N = 36$ extreme temperatures.

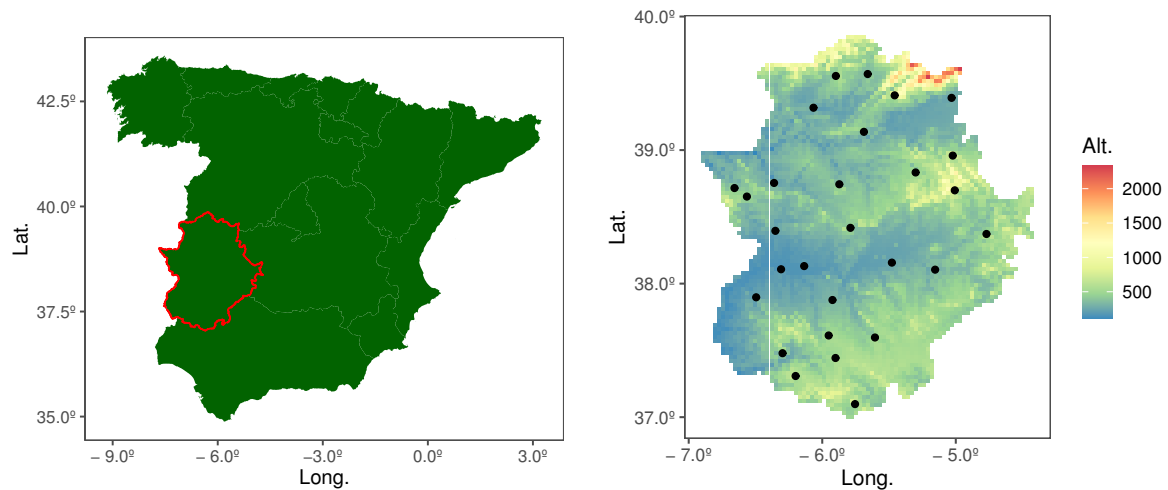


Figure 2. Location of the Extremadura Region within Spain (left). Topographic map of Extremadura together with the locations of the meteorological observatories used in this study (right).

For a site $s \in S$, temperature and altitude are correlated, as higher altitudes mean lower temperatures, i.e., temperature decreases with increasing altitude. This reason, together with the fact that Extremadura is not a large region, led us to take the altitude of the sites as being the only covariate in the regression model (5). The altitude was standardized as follows:

$$\tilde{h}_s = \frac{h_s^{obs} - \min_s h_s}{\max_s h_s - \min_s h_s} \tag{16}$$

where h_s^{obs} is the altitude above mean sea level of site $s \in S$, and $\max_s h_s$ and $\min_s h_s$ are the maximum and minimum altitudes of all the sites, respectively.

5. Results

The Bayesian spatial copula model was applied to the described data set (see Section 4). In particular, the BHGCM-200 and BHGCM-210 models were compared using the DIC as a measure of the goodness-of-fit (see Section 3.2). The parameters of the model that best fits the data were compared with the equivalent noncopula model (BHM).

5.1. Evaluation of the Models

As noted above, the DIC was employed to compare the two candidate spatial copula models. The results are presented in Table 1. One observes that the copula model in which a spatial model intervenes in the location and scale parameters’ regressions, i.e., BHGCM-210, has a lower DIC value than the model in which the scale parameter is constant (BHGCM-200).

Table 1. Results of using the DIC for the copula models. Boldface indicates the better model.

Model	\bar{D}_θ	$D(\bar{\theta})$	p_θ	DIC
BHGCM-200	3697.57	3668.49	29.08	3726.64
BHGCM-210	3589.41	3536.57	52.83	3642.24

Therefore, the spatial copula model chosen is BHGCM-210. Table 2 gives the results of applying the goodness-of-fit to the BHM-210 model. Recall that the parameter p_θ indicates the complexity of the model, and, as can be observed in Table 2, it is greater in the BHGCM model, which was to be expected given that the proposed models are intrinsically more complex. However, this increase in complexity over the models proposed by García et al. [18] is acceptable since the variance in the models proposed in the present work is less than in the BHM cases.

Table 2. Results of using the DIC for models BHM-210 and BHGCM-210.

Model	\bar{D}_θ	$D(\bar{\theta})$	p_θ	DIC
BHM-210	4046.96	3997.51	49.46	4096.42
BHGCM-210	3589.41	3536.57	52.83	3642.24

5.2. Parameter Estimates

In this subsection, the regression parameters of the BHM-210 model and the proposed BHGCM-210 model are compared.

Figure 3 shows the estimate of the posterior distribution density function from the selected BHGCM-210 model (red line) versus that estimated by the BHM-210 model (blue line) for the different regression coefficients α . The regression coefficient $\alpha_{\sigma 1}$ (lower left panel) of the scale parameter, σ , has a mean value of 0.78 °C (SD: 0.51 °C) for the BHGCM model. Moreover, this model provides an estimate of the density function that is less concentrated than that given by the BHM model.

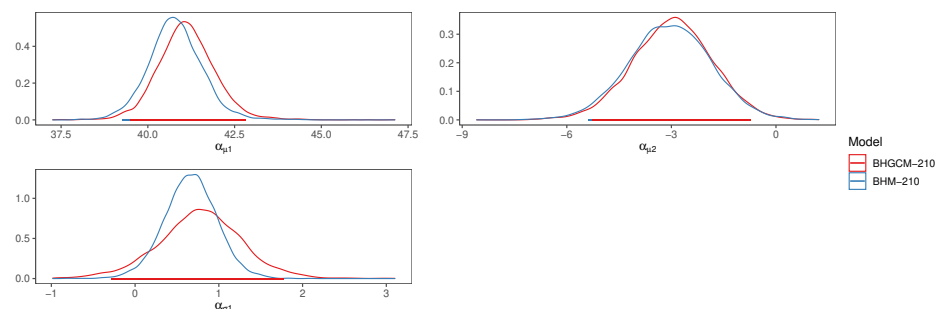


Figure 3. Estimated posterior distribution density functions of the regression coefficients α_{k1} (left) and α_{k2} (right) for the location (top row) and scale (bottom row) parameters for the models BHM-210 (blue line) and BHGCM-210 (red line). The red horizontal lines show the 0.025 to 0.975 quantiles.

With regard to the regression coefficients for the location parameter, μ (top row), the two models give qualitatively similar posterior density function estimates. In particular, the coefficient $\alpha_{\mu 1}$ (top left panel) has a mean of 41.12 °C (SD: 0.84 °C), while the coefficient $\alpha_{\mu 2}$ (top right panel) is clearly negative with a mean of -2.97 °C km $^{-1}$ (SD: 1.15 °C km $^{-1}$). These negative values are consistent with the fact that temperature decreases with altitude.

Another interesting result is the covariance function of the GEV parameters given by Equation (8). Table 3 lists the medians and (2.5%, 97.5%) quantiles of the sill (β_{k0}) and range (β_{k1}) coefficients for the location and scale parameters in models BHM-210 and BHGCM-210. The values of these coefficients are of similar orders of magnitude for the two models. Since the range is a measure of the strength of spatial dependence for the location and scale parameters, one observes that, in both models, this dependence is weaker for the location parameter than for the scale parameter.

Table 3. Median (2.5%, 97.5%) of the sill and range coefficients for the location and scale parameters of models BHM-210 and BHGCM-210.

Model	Location Sill	Location Range	Scale Sill	Scale Range
BHM-210	0.52 (0.18, 2.08)	395.07 (132.90, 885.62)	0.25 (0.11, 0.62)	554.84 (235.55, 1110.02)
BHGCM-210	0.53 (0.18, 2.15)	389.82 (125.59, 872.49)	0.27 (0.12, 0.71)	562.47 (233.41, 1133.02)

5.3. Validation of the Models

The posterior predictive distribution (PPD) provides temperature values for the measured observatories that can be compared to the observed temperatures. The error (E) and the absolute error (AE) were used to validate the BHGCM-210 model. These errors were calculated using the values obtained from the PPD. They define as

$$E_{st} = \hat{Y}_{st} - Y_{st}$$

$$AE_{st} = |\hat{Y}_{st} - Y_{st}| \tag{17}$$

where \hat{Y}_{st} is the predictive temperature, Y_{st} is the observed temperature, $s \in S = \{s_1, \dots, s_M\}$, and $t \in T = \{t_1, \dots, t_N\}$.

Figure 4 shows the density function of E (left panel) and AE (right panel) for the BHM-210 model (black color) and BHGCM-210 model (red color). The error’s density functions are symmetric around their mean values of 0.23 and 0.06 for the BHGCM-210 and BHM-210 models, respectively. Note that the mean values represent the bias of the models. The precision, measured as the absolute error, is slightly better in the copula model than in the noncopula model. This is related with the increase in uncertainty, previously mentioned in the Introduction.

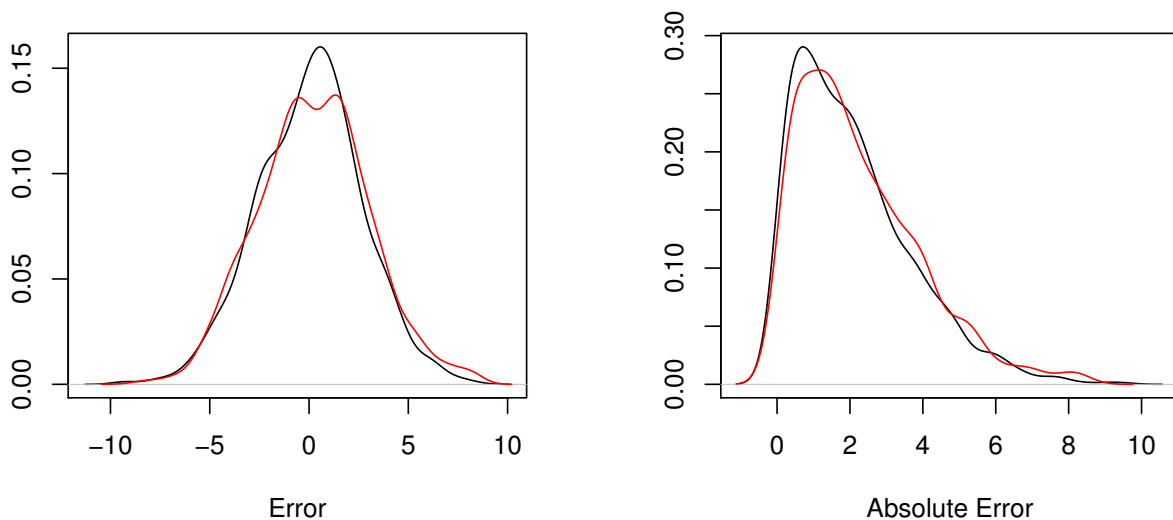


Figure 4. Validation of the BHM-210 (black color) and BHGCM-210 (red color) models for all the observatories: E (left) and AE (right) density functions.

Figure 5 shows the Q-Q plots for all observatories between the theoretical and empirical quantile of Y_{st} and \hat{Y}_{st} , respectively, for BHM-210 (left panel) and BHGCM-210 (right panel) models. It can be seen that the ratio between both is close to the red 1:1 line.

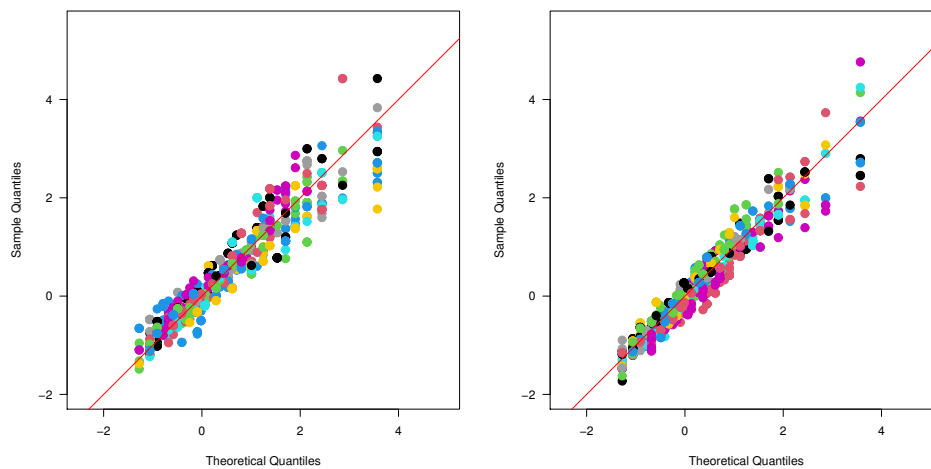


Figure 5. Validation of the BHM-210 (left) and BHGCM-210 (right) models for all the observatories: Q-Q plots.

5.4. Inference

The chosen model (BHGCM-210) estimates the location, scale, and shape parameters of the at-site GEV distribution. These parameters are predicted at the ungauged sites with Algorithm 1. Figure 6 shows the estimated posterior distribution density function for the shape parameter. This parameter takes clearly negative values, with a symmetric and homogeneous distribution around the mean value -0.38 and a standard deviation of 0.03 , indicating that there is an upper bound on how high extreme temperatures can be. This leads to a quick decrease of the rate of decay of the extreme temperature distribution, so it does not increase infinitely. Moreover, it is relevant to highlight the low value of the standard deviation obtained when compared with that obtained for the shape parameter estimated in each location individually. Using the maximum-likelihood fit, the standard deviation of the shape parameter varies from 0.09 to 0.12 —meaning more than three times that obtained with the spatial model. Therefore, pooling nearby data leads to a decrease in the statistical uncertainty of the parameters and implies an important advantage of using spatial models.

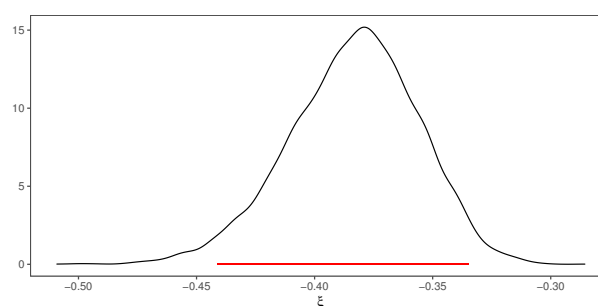


Figure 6. Estimated posterior distribution density function for the shape parameter. The red horizontal line shows the 0.025 to 0.975 quantile.

Figure 7 shows the spatial posterior distributions of the means and standard deviations of the location and scale parameters. The spatial posterior distribution of the location parameter shows its dependence on altitude, with mean values between 39.29 °C and 41.12 °C and standard deviations between 1.31 °C and 1.41 °C. The areas with low values of the location parameter correspond to those of higher altitude, since the temperature is highly determined by the orography (see Figure 2). The spatial posterior distribution of the scale parameter shows no spatial dependence, taking very similar values over the entire

region (between 2.96 °C and 3.24 °C), and the standard deviations are very concentrated throughout the region.

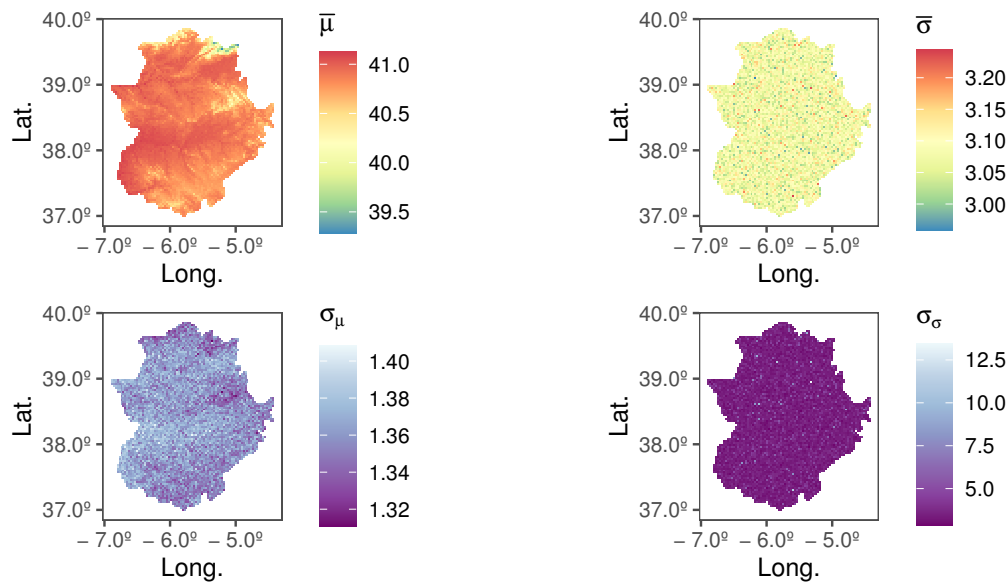


Figure 7. Spatial posterior distributions of the mean (left) and the standard deviation (right) of the location (top row) and scale (bottom row) parameters.

Figure 8 shows the estimation of the maximum temperatures in Extremadura through the posterior predictive distribution. These estimations are consistent with the estimated values for the location parameter given in the previous figure.

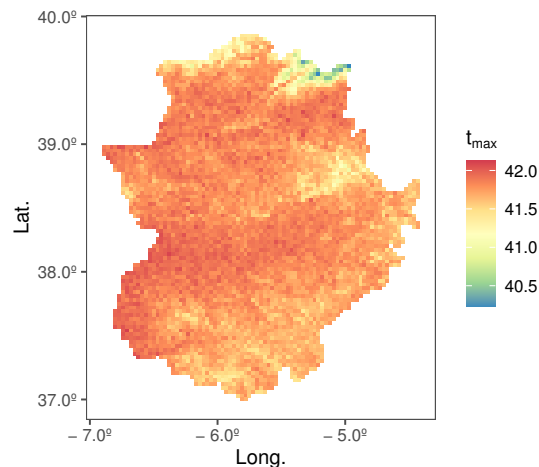


Figure 8. Spatial posterior predictive distribution of the maxima temperature.

6. Conclusions

1. Bayesian hierarchical models, BHM, proposed by García et al. [18] present the problem of assuming spatial independence between observations at different sites. The present work has addressed this problem by introducing a copula.
2. A Gaussian copula is assumed as a joint distribution with at-site GEV marginal distributions. In this way, the spatial dependence of observations from different sites

is represented by a correlation matrix. In addition, spatial regression models of the GEV parameters are proposed.

3. Two BHGCM models are proposed: BHGCM-200 takes a spatial regression model for μ while the parameters σ and ξ are constant; BHGCM-210 takes spatial models for μ and σ , while the parameter ξ is constant.
4. The BHGCM-210 model has a better DIC goodness-of-fit value than the BHGCM-200 model and the noncopula BHM-210 model.
5. For the GEV distribution's location parameter, the BHGCM-210 and BHM-210 models give qualitatively similar estimates of the regression parameter posterior distributions.
6. For the GEV distribution's scale parameter, the BHGCM-210 model gives a distribution with greater variance than that given by the BHM-210 model.
7. In the BHGCM-210 model, the GEV shape parameter takes negative values, and its posterior distribution is symmetrical and highly concentrated around -0.38 . Therefore, the extreme temperature distribution is not expected to increase too much.
8. The BHGCM-210 model gives a spatial posterior distribution for the location parameter that is strongly dependent on altitude, unlike the scale parameter. The location parameter's mean values in the region lie between 39.29°C and 41.12°C .
9. In the BHGCM-210 model, the scale parameter's spatial posterior distribution is very concentrated, taking very similar values throughout the region.

Author Contributions: All the authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript

Funding: This work was supported by FEDER / Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación / Proyecto MTM2017-86875-C3-2-R, and by the Junta de Extremadura, FEDER Funds, GR18108, GR18097 and IB16063 (Consejería de Economía e Infraestructuras).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study belong to Spain's State Meteorological Agency (Agencia Estatal de Meteorología: www.aemet.es) (accessed on 9 July 2021). The code is available upon request to the author.

Acknowledgments: Thanks are due to the Spanish State Meteorological Agency for providing the daily temperature time series used in this study.

Conflicts of Interest: The authors declare that there were no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in writing the manuscript, or in the decision to publish the results

Appendix A. Gaussian Copula

Definition A1. For every $M \geq 2$, an M -dimensional copula is an M -variate distribution function on $[0, 1]^M$ whose marginals are uniformly distributed on $[0, 1]$.

In copula theory, the most important theorem is Sklar's Theorem, because it permits establishing a relation between multivariate distributions and copulas.

Theorem A1 (Sklar's Theorem). Let F be a M -dimensional distribution function with univariate F_1, F_2, \dots, F_M . Let $A_j := F_j(\mathbb{R})$ denote the range of F_j ($j = 1, 2, \dots, M$). Then, there exists a copula C such that for all $(x_1, x_2, \dots, x_M) \in \mathbb{R}^M$,

$$F(x_1, x_2, \dots, x_M) = C(F_1(x_1), F_2(x_2), \dots, F_M(x_M)). \quad (\text{A1})$$

Such a C is uniquely determined on $A_1 \times A_2 \times \dots \times A_M$, hence, it is unique when F_1, F_2, \dots, F_M are all continuous.

An elliptical copula is a family of copula. Gaussian copula is one of them. This copula has probability density function (pdf)

$$f_{GC}(y_1, \dots, y_M) = \frac{\prod_{i=1}^M f_{GEV}(y_i | \mu_{s_i}, \sigma_{s_i}, \xi_{s_i})}{\prod_{i=1}^M f_N(u_i)} \cdot f_{N_M}(u_1, \dots, u_M | \mathcal{C}), \quad (A2)$$

where f_{GEV} represents the GEV distribution's pdf (1), f_N the pdf of the standard normal distribution, f_{N_M} the pdf of the M -dimensional normal distribution with correlation matrix \mathcal{C} , and u_i the quantile of $P(Y_{s_i t} \leq y_i)$ (2) of the standard normal distribution for $i = 1, \dots, M$.

Appendix B. Estimation

The posterior distributions of different models are

$$\begin{aligned} &P(c_0, c_1, \mu(\mathbf{s}), \sigma, \xi, W_\mu(\mathbf{s}), \alpha_\mu, \beta_{\mu 0}, \beta_{\mu 1}, \tau_\mu^2 | Y, h(\mathbf{s})) \propto P(Y | c_0, c_1, \mu(\mathbf{s}), \sigma, \xi) \\ &\cdot P(\mu(\mathbf{s}) | h(\mathbf{s}), W_\mu(\mathbf{s}), \alpha_\mu, \beta_{\mu 0}, \beta_{\mu 1}, \tau_\mu^2) \cdot P(\sigma) \cdot P(\xi) \cdot P(W_\mu(\mathbf{s}) | \beta_{\mu 0}, \beta_{\mu 1}) \\ &\cdot P(c_0) \cdot P(c_1) \cdot P(\alpha_\mu) \cdot P(\beta_{\mu 0}) \cdot P(\beta_{\mu 1}) \cdot P(\tau_\mu^2), \end{aligned} \quad (A3)$$

for the BHGCM-200 model; for the BHGCM-210 model, it is

$$\begin{aligned} &P(c_0, c_1, \mu(\mathbf{s}), \sigma(\mathbf{s}), \xi, W_\mu(\mathbf{s}), W_\sigma(\mathbf{s}), \alpha_\mu, \alpha_{\sigma 1}, \beta_{\mu 0}, \beta_{\mu 1}, \beta_{\sigma 0}, \beta_{\sigma 1}, \tau_\mu^2, \tau_\sigma^2 | Y, h(\mathbf{s})) \\ &\propto P(Y | c_0, c_1, \mu(\mathbf{s}), \sigma(\mathbf{s}), \xi) \cdot P(\mu(\mathbf{s}) | h(\mathbf{s}), W_\mu(\mathbf{s}), \alpha_\mu, \beta_{\mu 0}, \beta_{\mu 1}, \tau_\mu^2) \\ &\cdot P(\sigma(\mathbf{s}) | W_\sigma(\mathbf{s}), \alpha_{\sigma 1}, \beta_{\sigma 0}, \beta_{\sigma 1}, \tau_\sigma^2) \cdot P(\xi) \cdot P(W_\mu(\mathbf{s}) | \beta_{\mu 0}, \beta_{\mu 1}) \cdot P(W_\sigma(\mathbf{s}) | \beta_{\sigma 0}, \beta_{\sigma 1}) \\ &\cdot P(c_0) \cdot P(c_1) \cdot P(\alpha_\mu) \cdot P(\beta_{\mu 0}) \cdot P(\beta_{\mu 1}) \cdot P(\tau_\mu^2) \cdot P(\alpha_{\sigma 1}) \cdot P(\beta_{\sigma 0}) \cdot P(\beta_{\sigma 1}) \cdot P(\tau_\sigma^2). \end{aligned} \quad (A4)$$

Additionally, the likelihood function given in (12) can be written in logarithmic terms as

$$\begin{aligned} \ln L(c_0, c_1, \mu(\mathbf{s}), \sigma(\mathbf{s}), \xi | Y) = &-N \cdot \sum_{i=1}^M \ln \sigma_{s_i} - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^M \sum_{j=1}^N \ln \left(1 + \xi \cdot \frac{y_{s_i t_j} - \mu_{s_i}}{\sigma_{s_i}}\right) \\ &- \sum_{i=1}^M \sum_{j=1}^N \left(1 + \xi \cdot \frac{y_{s_i t_j} - \mu_{s_i}}{\sigma_{s_i}}\right)^{-1/\xi} + \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N u_{s_i t_j}^2 - \frac{1}{2} \cdot N \cdot \ln |\mathcal{C}| \\ &- \frac{1}{2} \sum_{j=1}^N (u_{s_1 t_j}, \dots, u_{s_M t_j}) \mathcal{C}^{-1} (u_{s_1 t_j}, \dots, u_{s_M t_j})^t. \end{aligned} \quad (A5)$$

For the BHGCM-200 model, the above equation is simpler since $\sigma_{s_i} = \sigma$ for $i = 1, \dots, M$.

References

- García, J.; Gallego, M.C.; Serrano, A.; Vaquero, J. Trends in Block-Seasonal Extreme Rainfall over the Iberian Peninsula in the Second Half of the Twentieth Century. *J. Clim.* **2007**, *20*, 113–130.
- Re, M.; Barros, V.R. Extreme rainfalls in SE South America. *Clim. Chang.* **2009**, *96*, 119–136.
- Acero, F.J.; García, J.A.; Gallego, M.C. Peaks-over-Threshold Study of Trends in Extreme Rainfall over the Iberian Peninsula. *J. Clim.* **2011**, *24*, 1089–1105.
- Acero, F.J.; Parey, S.; Hoang, T.T.H.; Dacunha-Castelle, D.; García, J.A.; Gallego, M.C. Non-stationary future return levels for extreme rainfall over Extremadura (SW Iberian Peninsula). *Hydrol. Sci. J.* **2017**, *62*, 1394–1411.
- Wi, S.; Valdés, J.B.; Steinschneider, S.; Kim, T.W. Non-stationary frequency analysis of extreme precipitation in South Korea using peaks-over-threshold and annual maxima. *Stoch. Environ. Res. Risk Assess.* **2016**, *30*, 583–606.
- Nogaj, M.; Yiou, P.; Parey, S.; Malek, F.; Naveau, P. Amplitude and frequency of temperature extremes over the North Atlantic region. *Geophys. Res. Lett.* **2006**, *33*, doi:10.1029/2005GL02425.

7. Coelho, C.A.S.; Ferro, C.A.T.; Stephenson, D.B.; Steinskog, D.J. Methods for Exploring Spatial and Temporal Variability of Extreme Events in Climate Data. *J. Clim.* **2008**, *21*, 2072–2092.
8. Acero, F.J.; Fernández-Fernández, M.I.; Carrasco, V.M.S.; Parey, S.; Hoang, T.T.H.; Dacunha-Castelle, D.; García, J.A. Changes in heat wave characteristics over Extremadura (SW Spain). *Theor. Appl. Climatol.* **2018**, *133*, 605–617.
9. Ramos, A.A. Extreme value theory and the solar cycle. *Astron. Astrophys.* **2007**, *472*, 293–298.
10. Acero, F.J.; Carrasco, V.M.S.; Gallego, M.C.; García, J.A.; Vaquero, J.M. Extreme Value Theory Applied to the Millennial Sunspot Number Series. *Astrophys. J.* **2018**, *853*, 80.
11. Longin, F.M. From value at risk to stress testing: The extreme value approach. *J. Bank. Financ.* **2000**, *24*, 1097–1130.
12. Castillo, E.; Hadi, A.S.; Balakrishnan, N.; Sarabia, J.M. *Extreme Value and Related Models with Applications in Engineering and Science*; Wiley: Hoboken, NJ, USA, 2004.
13. Casson, E.; Coles, S. Spatial regression models for extremes. *Extremes* **1999**, *1*, 449–468.
14. Cooley, D.; Nychka, D.; Naveau, P. Bayesian spatial modeling of extreme precipitation return levels. *J. Am. Stat. Assoc.* **2007**, *102*, 824–840.
15. Portero, J.; Acero, F.J.; García, J.A. Analysis of Extreme Temperature Events over the Iberian Peninsula during the 21st Century Using Dynamic Climate Projections Chosen Using Max-Stable Processes. *Atmosphere* **2020**, *11*, 506.
16. Davison, A.C.; Padoan, S.A.; Ribatet, M. Statistical modeling of spatial extremes of spatial extremes. *Stat. Sci.* **2012**, *27*, 161–186.
17. Acero, F.J.; García, J.A.; Gallego, M.C.; Parey, S.; Dacunha-Castelle, D. Trends in summer extreme temperatures over the Iberian Peninsula using nonurban station data. *J. Geophys. Res. Atmos.* **2014**, *119*, 39–53, doi:10.1002/2013JD020590
18. García, A.; Martín, J.; Naranjo, L.; Acero, F.J. A Bayesian hierarchical spatio-temporal model for extreme rainfall in Extremadura (Spain). *Hydrol. Sci. J.* **2018**, *63*, 878–894.
19. Renard, B.; Lang, M. Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Adv. Water Resour.* **2007**, *30*, 897–912.
20. Renard, B. A Bayesian hierarchical approach to regional frequency analysis. *Water Resour. Res.* **2011**, *47*, 11513.
21. Liu, Y.R.; Li, Y.P.; Ma, Y.; Jia, Q.M.; Su, Y.Y. Development of a Bayesian-copula-based frequency analysis method for hydrological risk assessment—The Naryn River in Central Asia. *J. Hydrol.* **2020**, *580*, 124349.
22. Beck, N.; Genest, C.; Jalbert, J.; Mailhot, M. Predicting extreme surges from sparse data using a copula-based hierarchical Bayesian spatial model. *Environmetrics* **2020**, *31*, e2616.
23. Salvadori, G.; De Michele, C.; Kottegoda, N.T.M.; Rosso, R. *Extremes in Nature: An Approach Using Copulas*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007; Volume 56.
24. Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **1959**, *8*, 229–231.
25. Genest, C.; Favre, A.C.; Bêliveau, J.; Jacques, C. Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resour. Res.* **2007**, *43*, W09401.
26. Favre, A.C.; El Adlouni, S.; Perreault, L.; Thiémonge, N.; Bobée, B. Multivariate hydrological frequency analysis using copulas. *Water Resour. Res.* **2004**, *40*, doi:10.1029/2003WR002456.
27. Wikle, C.K.; Berliner, M.L.; Cressie, N. Hierarchical Bayesian space-time models. *Environ. Ecol. Stat.* **1998**, *5*, 117–154.
28. Sun, X.; Thyer, M.; Renard, B.; Lang, M. A general regional frequency analysis framework for quantifying local-scale climate effects: A case study of ENSO effects on Southeast Queensland rainfall. *J. Hydrol.* **2014**, *512*, 53–68.
29. Dyrddal, A.V.; Lenkoski, A.; Thorarinsdottir, T.L.; Stordal, F. Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics* **2015**, *26*, 89–186.
30. Ragulina, G.; Reitan, T. Generalized extreme value shape parameter and its nature for extreme precipitation using long time series and Bayesian approach. *Hydrol. Sci. J.* **2017**, *62*, 863–879.
31. Barlow, A.M.; Rohrbeck, C.; Sharkey, P.; Shooter, R.; Simpson, E.S. A Bayesian spatio-temporal model for precipitation extremes-STOR team contribution to the EVA2017 challenge. *Extremes* **2018**, *21*, 431–439.
32. Craigmile, P.F.; Guttorp, P. Can a regional climate model reproduce observed extreme temperatures? *Statistica* **2013**, *73*, 103–122.
33. Daraio, J.A.; Amponsah, A.O.; Sears, K.W. Bayesian Hierarchical Regression to Assess Variation of Stream Temperature with Atmospheric Temperature in a Small Watershed. *Hydrology* **2017**, *4*, 44.
34. Gnedenko, B. Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.* **1943**, *44*, 423–453.
35. Fisher, R.A.; 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*; Cambridge University Press: Cambridge, UK, 1928; Volume 24, pp. 180–190.
36. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*; Springer: London, UK, 2001; Volume 208
37. Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. Introducing Markov Chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*; Chapman & Hall: New York, NY, USA, 1996.
38. Cowles, M.K.; Carlin, B.P. Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Am. Stat. Assoc.* **1996**, *91*, 883–904.
39. Martyn, P.; Nicky, B.; Kate, C.; Karen, V. CODA: Convergence diagnosis and output analysis for MCMC. *R News* **2006**, *6*, 7–11.
40. Wickham, H. *ggplot2. Elegant Graphics for Data Analysis*; Version, 2 (1); Springer-Verlag: New York, NY, USA, 2016;
41. Centro Nacional de Información Geográfica. Modelo Digital del Terreno 2015 CC-BY 4.0. Available online: scn.es (accessed on 9 July 2021).

42. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; van der Linde, A. The deviance information criterion: 12 years on. *J. R. Stat. Soc.* **2014**, *76*, 485–493.
43. Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. *Bayesian Data Analysis*, 2nd ed.; Texts in Statistical Science; Chapman and Hall: New York, NY, USA, 1995; 696p.

Parte II

Regresión Simbólica

Capítulo 5

Introducción a la Regresión Simbólica

Los algoritmos evolutivos constituyen una línea de la inteligencia artificial que tiene numerosas aplicaciones prácticas. Fundamentalmente, contribuyen a encontrar soluciones razonablemente buenas a problemas reales de forma rápida. Se trata de métodos de optimización y búsqueda de soluciones basándose en la evolución natural.

Resultan especialmente útiles en la búsqueda de soluciones a problemas no lineales, principalmente cuando intervienen diversas variables y la búsqueda es extensa. Un modo de implementar algoritmos evolutivos es usando algoritmos genéticos, y más concretamente la programación genética. Poli et al. [65] la definen del siguiente modo:

Definición 5.1. *La programación genética es una colección de técnicas de computación evolutiva que permiten a los ordenadores resolver problemas automáticamente sin requerir que el usuario especifique previamente la forma o estructura de la solución.*

En líneas generales, la programación genética parte de un conjunto (o población) de N soluciones (o individuos), $X^{(1)} = \{x_1^1, \dots, x_N^1\}$, que evoluciona mediante operadores evolutivos, manteniéndose su tamaño en las siguientes iteraciones, $X^{(i)} = \{x_1^i, \dots, x_N^i\}$ (población en la iteración i -ésima).

Los operadores habituales son:

- *Selección de buenos individuos:* se escogen individuos de la población $X^{(i)}$, con reemplazamiento, de modo que aquellos individuos con mejores valores de la función objetivo f tienen mayor probabilidad asignada.

- *Cruce de individuos*: se parten aleatoriamente dos individuos e intercambian sus partes dando lugar a nuevos individuos, a los que se suele llamar descendientes.
- *Mutación*: se cambian aleatoriamente algunos elementos de la codificación de una solución.
- *Selección de malos individuos*: se escogen individuos de $X^{(i)}$, sin reemplazamiento, otorgándose mayor probabilidad a aquellos con peores valores de f .

La estructura general de un algoritmo genético, tras elegir la población inicial $X^{(1)}$ del conjunto de soluciones factibles y tomar el mejor individuo como solución óptima inicial, x^* , mientras no se cumpla la condición de parada, es:

1. Seleccionar $2M$ buenos individuos de la población actual.
2. Formar pares y realizar su cruce.
3. Aplicar una posible mutación a los descendientes e insertarlos en el lugar de $2M$ malos individuos de la población.
4. Actualizar la solución óptima alcanzada, caso de ser necesario.

El esquema de un algoritmo evolutivo es:

Seleccionar $X^{(1)} = \{x_1^1, \dots, x_N^1\} \subset X$

Hacer $x^* = \arg \min \{f(x_1^1), \dots, f(x_N^1)\}$, $f^* = \min \{f(x_1^1), \dots, f(x_N^1)\}$, $i = 1$

Hasta satisfacer el criterio de parada,

Seleccionar $2M$ buenos individuos $\{y_1, \dots, y_{2M}\}$ de $X^{(i)}$

Desde $k = 1$ hasta M ,

cruzar los pares y_{2k-1}, y_{2k} obteniendo los pares z_{2k-1}, z_{2k}

Desde $k = 1$ hasta $2M$,

mutar z_k obteniendo $w_k : M_i = \{w_1, \dots, w_{2M}\}$

Seleccionar $2M$ malos individuos $B_i = \{v_1, \dots, v_{2M}\}$ de $X^{(i)}$

Hacer $i = i + 1$, $X^{(i)} = (X^{(i-1)} \setminus B_i) \cup M_i$

Actualizar f^*, x^*

Seguramente el mayor reto al que haya que enfrentarse a la hora de poner en práctica estos algoritmos sea la codificación del problema.

La *regresión simbólica* es un caso particular de la programación genética que emplea estructuras de árboles para representar a los individuos de la población.

En este caso, se pueden representar las expresiones matemáticas utilizando *árboles sintácticos* cuyos nodos serán funciones matemáticas, símbolos asociados a las variables y/o constantes del modelo. En concreto, los nodos terminales van asociados a las variables y/o parámetros ajustables del modelo, mientras que en los nodos restantes se sitúan funciones con uno o dos argumentos (por ejemplo, $+$, $-$, $*$, $/$, $^$, \sin , \cos , ...), operadores booleanos (AND, OR, NOT) o condicionales (IF-THEN-ELSE). En la Figura 5.1 se representa la expresión analítica $\sigma_0 \cdot \left(1 - \frac{T}{T_C}\right)^{n_0}$, a modo de ejemplo.

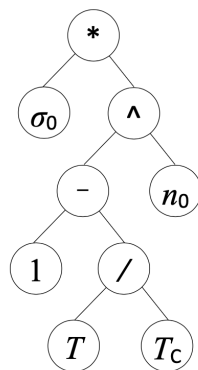


Figura 5.1: Árbol sintáctico para la expresión analítica $\sigma_0 \cdot \left(1 - \frac{T}{T_C}\right)^{n_0}$.

En el caso particular del problema presentado en esta Tesis Doctoral, los individuos son expresiones matemáticas tratando de generar soluciones lo más idóneas posible, ya que el objetivo es encontrar aquella relación que mejor reproduzca la relación entre las variables.

Una función objetivo de uso habitual en la regresión simbólica es la función de errores relativos, cuya definición se muestra a continuación.

Definición 5.2. *Dado un conjunto de d datos experimentales, y dado un individuo F , se define el error relativo correspondiente a F como*

$$f_F = \frac{100}{d} \sum_{i=1}^d \left| \frac{F(i)^{fit} - F(i)^{exp}}{F(i)^{exp}} \right|, \quad (5.1)$$

donde $F(i)^{exp}$ es el i -ésimo dato experimental y $F(i)^{fit}$ representa el valor que predice el individuo (o función) F para el i -ésimo dato.

Para elegir a los buenos y malos individuos, dado que la función objetivo es positiva, se definen las siguientes funciones de probabilidad.

Definición 5.3. Para cada individuo F_i de una población de tamaño N la probabilidad de ser seleccionado como progenitor es

$$p_{F_i} = \frac{1/f_{F_i}}{\sum_{j=1}^N 1/f_{F_j}}, \quad i = 1, \dots, N, \quad (5.2)$$

y la probabilidad de ser eliminado es

$$p'_{F_i} = \frac{f_{F_i}}{\sum_{j=1}^N f_{F_j}}, \quad i = 1, \dots, N. \quad (5.3)$$

Así, se eligen con mayor probabilidad los buenos individuos como progenitores y los de mala calidad para ser sustituidos.

Una vez seleccionados los progenitores, el cruce y la mutación se aplican según ciertas probabilidades p_c y p_m , respectivamente, introduciendo mayor diversificación cuanto mayores sean y mayor intensificación en caso contrario. En el cruce, dos progenitores intercambian partes de sus árboles proporcionando dos nuevos individuos. En la Figura 5.2 se muestra el proceso de cruce para dos progenitores (árboles superiores) que producen dos descendientes (árboles inferiores).

Para la mutación también existen varias estrategias, como por ejemplo escoger un nodo, y generar bajo él otro árbol, o bien cambiar por otro de la misma clase; o simplemente intercambiar dos nodos del mismo tipo. La mutación es útil para introducir diversidad en la población cuando se tiene una alta presión de selección, esto es, cuando la probabilidad de elegir buenos individuos es mucho mayor que la que se asigna a los peores miembros de la población. La Figura 5.3 muestra el individuo original (izquierda) y el individuo mutado (derecha) al intercambiar dos nodos del mismo tipo.

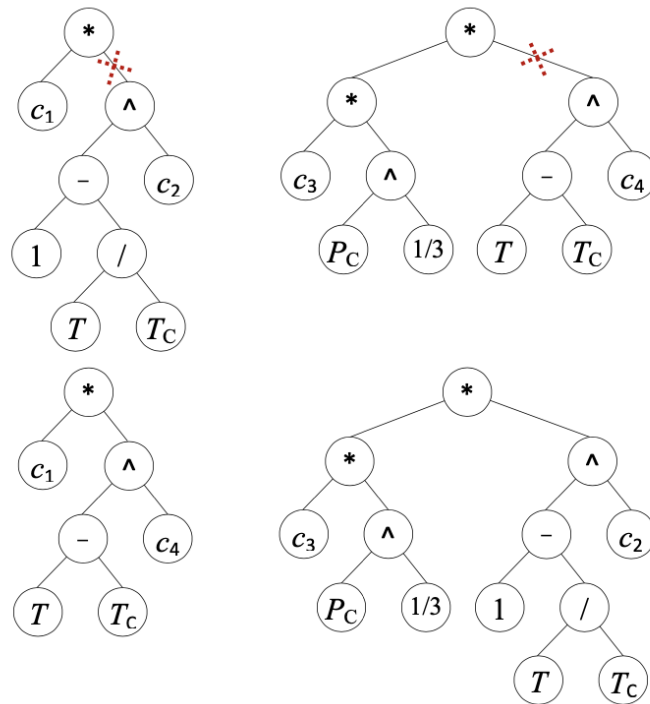


Figura 5.2: Árboles sintácticos correspondientes a los progenitores seleccionados para el cruce, y los descendientes obtenidos. Los progenitores representan las expresiones $c_1 \cdot (1 - T/T_C)^{c_2}$ (superior izquierda) y $c_3 \cdot P_C^{1/3} \cdot (T - T_C)^{c_4}$ (superior derecha). Los descendientes representan las expresiones $c_1 \cdot (T - T_C)^{c_4}$ (inferior izquierda) y $c_3 \cdot P_C^{1/3} \cdot (1 - T/T_C)^{c_2}$ (inferior derecha).

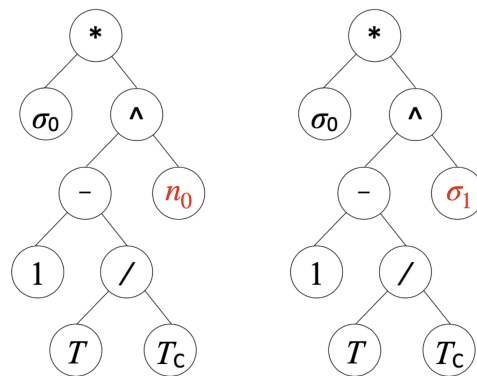


Figura 5.3: Mutación de árboles sintácticos. El individuo original (izquierda) representa la expresión analítica $\sigma_0 \cdot (1 - T/T_C)^{n_0}$ y el individuo mutado (derecha) la expresión $\sigma_0 \cdot (1 - T/T_C)^{\sigma_1}$.

Capítulo 6

Artículo D: Desarrollo de modelos de tensión superficial de alcoholes mediante Regresión Simbólica

Autores:

Eva L. Sanjuán, M. Isabel Parra, Mario M. Pizarro

Departamento de Matemáticas, Universidad de Extremadura

Revista: Journal of Molecular Liquids, 298, 111971, 2020

DOI: [10.1016/j.molliq.2019.111971](https://doi.org/10.1016/j.molliq.2019.111971)

Resumen:

En este trabajo, se aplica la regresión simbólica (SR) para encontrar un modelo que establezca las relaciones existentes entre una serie de variables regresoras y una variable respuesta. En concreto, se explica cómo encontrar modelos de correlación para la tensión superficial de alcoholes; aunque esta metodología es aplicable a cualquier otra propiedad termodinámica y/o familia de fluidos. Se utiliza una base de datos de 87 alcoholes, con un total de 3570 datos que se han seleccionado y filtrado en trabajos anteriores. Se consideran todos los parámetros que tienen un significado físico para el modelo y se realiza un estudio de correlación para seleccionar los más representativos: temperatura, temperatura crítica, presión crítica, volumen crítico, volumen molar y factor acéntrico. Posteriormente, se comparan los mejores modelos obtenidos con SR, atendiendo a la precisión y complejidad del modelo, con los modelos de regresión polinómica empleados habitualmente. Finalmente, se analizan y optimizan los modelos ofrecidos por SR que incluyen

el menor número de parámetros posible y proporcionan la mayor precisión. Los mejores modelos encontrados ofrecen valores inferiores a 7.8% para la desviación absoluta porcentual media (MAPD) y por debajo del 0.07 para $1 - R^2$, mejorando considerablemente los aportados por otros modelos generales publicados por otros autores.



Contents lists available at ScienceDirect

Journal of Molecular Liquids

journal homepage: www.elsevier.com/locate/molliq

Development of models for surface tension of alcohols through symbolic regression

E.L. Sanjuán*, M.I. Parra, M.M. Pizarro

Department of Mathematics, University of Extremadura, Spain



ARTICLE INFO

Article history:

Received 10 June 2019

Received in revised form 7 October 2019

Accepted 18 October 2019

Available online 5 November 2019

ABSTRACT

The study of models for the correlation of the surface tension of fluids is a crucial issue in various fields of chemical industries. Usually, seeking for appropriate models is an arduous process that relies too much on the researcher's inventiveness. The main advantage of symbolic regression (SR) is being a straightforward method which automatizes the searching for models, and provides precise and trustworthy models for the researcher to work with. Although in this work we apply SR to develop correlation models for the surface tension of alcohols, the methodology is clearly applicable to any other thermodynamic property or any other family of fluids. We employ a database set of 87 alcohols, with a total quantity of 3570 data that have been carefully selected and filtered in previous works. We consider all the parameters which have a physical meaning for the model, and make a correlation study in order to select the most representative ones: temperature, critical temperature, critical pressure, critical volume, molar volume and acentric factor. Then we make the comparison for the best models obtained with SR (attending to accuracy and complexity of the model) and the usually employed polynomial regression models, obtaining lower errors ($1-R^2$) with SR. Afterwards, we analyze and optimize the models offered by SR which include the least number of parameters possible and provide the highest accuracy. The best models offer values under 7.8% for MAPD (the lowest one is 6.8%), and under 0.07 for $1-R^2$ (the lowest one is 0.04).

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Surface tension is a physical property of great importance [1], whose understanding and knowledge provides the possibility of many industrial engineering applications, given that it influences several chemical processes, such as gas absorption, distillation, extraction, bubble and droplets formation, etc ... [2–5].

Concretely, the study of the behavior of the surface tension is crucial for the development, production and performance of different chemical products such as pharmaceutical, food, biomaterial, paints, cosmetics, aerosols, detergents, disinfectants, fuel additives and others [6–10].

There are several experimental, theoretical and empirical methods to determine surface tension for a fluid. Among them, experimental measurements are the most precise way of estimation. However, conducting a surface tension experiment is time wasting, expensive, and it depends on the availability of appropriate instruments and fluid sample quantity. Consequently, the application of empirical correlations for fast estimation of surface tension has been investigated by several researchers in the last decades.

In spite of the considerable amount of models that have been proposed, up to the present day there is not a general model to predict

surface tension theoretically. That is why we have to use empirical or semiempirical models. However, one of the great issues in empirical modeling is choosing the most proper model. The number of new proposed models is growing fast, making the choice of a model or state equation that fits a given property and fluid very difficult, in order to be the basis for industrial processes [11–22]. The decision has to take into account, at least, its complexity, applicability, accuracy and predictable capacity [23].

By previous studies for surface tension, as well as for another thermodynamical properties, the main problem is that most of the models need to employ macroscopical parameters, which might not be known or that should be estimated with some uncertainty. Besides, the goodness of the models is usually limited to certain temperature ranges and/or to a concrete family of fluids [24–33].

The problem of finding a definite and universal method for a certain property or for a certain family of fluids is far from being resolved [34,35]. Usually, the process of looking for suitable models is an arduous and complex process that often relies on the researcher's intuition and inventiveness. Therefore, it is essential to dispose of an algorithmic and straightforward method to determine the best correlation models that fit experimental data, without the need to assume its functional form.

A lot of different models have been proposed, although they usually have two weak points: they can not be applied in general, and desired accuracy is not always guaranteed. In this context, the main objective

* Corresponding author.

E-mail addresses: etlopez@unex.es (E.L. Sanjuán), mipa@unex.es (M.I. Parra), mariomp@unex.es (M.M. Pizarro).

of this work is showing how to seek for appropriate models which avoid those disadvantages, cover the temperature range, and make the parameters of the models to have a physical meaning. The tool that will allow us to perform it is symbolic regression.

In this work, we will apply symbolic regression techniques in order to find a better correlation model, because it allows us to obtain the structure of an expression that can model a given set of experimental data, with no need of assuming an specific correlation format, and both the configuration and the coefficients of the model evolve automatically. Although there are another heuristic techniques, such as artificial neural networks [36,37], which can reproduce the shape of the function with great accuracy, nevertheless the main disadvantage is that they do not provide a user-friendly mathematical equation that could be employed by all the researchers. In fact, only the owner of the neural network can make use of it. This problem does not appear by using symbolic regression, because the algorithm leads us to clear mathematical expressions. Moreover, neural networks do not give a clue as to why and how the model was obtained, and there is no opportunity to improve the model. Besides, symbolic regression models can be analyzed afterwards (we can perform sensitivity analysis, error propagation ...) and can even turn out into the basis for the generation of better models.

A great amount of software related to symbolic regression has been developed until this moment, either for general languages or for specific ones. For this work, we chose Mathematica, because of being widely known and also its easy use. Mathematica itself does not have a function for symbolic regression. However, a third party product, Data Modeler package, developed by Evolve Analytics, provides an infrastructure for managing symbolic regression problems.

2. Correlation models

Surface tension appears from a certain temperature value, called triple point T_t , which depends on the fluid. Values for the surface tension are only positives, and decrease as the temperature increases, tending to zero. When the temperature reaches the critical point, $T = T_c$, liquid and vapor are indistinguishable, and as a consequence, surface tension is zero. Therefore, the function corresponding to surface tension is defined in the interval $[T_t, T_c]$.

The analysis of the published models, together with the development of our own ones, have taught us that one of the best strategies is basing them on universal scale laws, introducing the properties for the triple and critical points as entry parameters. In this way, a correct behavior in all the temperatures range is guaranteed [38,39].

In previous works, we have made advances in the study and construction of theoretical molecular models (valid only for Lennard-Jones fluids [40,41]) and for real pure fluids [34], proposing new coefficients for the Somayajulu model [12,42] and analyzing some families of fluids in detail, like liquid oxides [43] or alcohols [44]. Particularly, we have thoroughly studied the suitability of the model employed by National Institute of Standard Technologies (NIST) in REFPROP program [45], for the surface tension σ of pure fluids

$$\sigma(T) = \sum_i \sigma_i = \sigma^k - 1 \sigma_i \left(1 - \frac{T}{T_c}\right)_i^n \quad (1)$$

where σ_i and n_i are coefficients fitted by regression and T is the temperature. We showed that it could be improved using more representative data samples [34,35,44]. The new version of the program [45] uses our coefficients [34,35].

3. Symbolic regression

Genetic algorithm is an optimization technique based on stochastic, evolutionary principles that is used to find global extreme of a given function [46,47]. Genetic programming is a particular case of genetic

algorithm, developed in the nineties in the United States. J. Koza [48] popularized these techniques and settled their basis. Poli et al. [49] defined it as a collection of evolutionary computation techniques that allow computers to solve problems automatically without requiring the user to know or specify the form or structure of the solution in advance. The objective of genetic programming is to find the best solution for a particular problem, by genetically recombining a population of individuals that portray candidate solutions. Developed originally to automatically generate computer programs, genetic programming has been used to solve a wide range of practical problems in a variety of fields, e.g. finance [50], electronic design [51], signal processing [52], system identification [53], modeling of chemical systems [54], thermal engineering [55], among others.

In genetic programming we work with a population of individuals (mathematical expressions) which evolve as the search proceeds, and the objective is to find the most qualified one.

To implement the algorithm, we use tree-structured representations of the individuals. This is a universal way of representation, based on a free-context grammar. Branch nodes may be operators with one or two arguments, such as +, -, *, /, ^, sin, cos, exp, log..., or may be boolean (such as AND, OR, NOT) or conditional (IF-THEN-ELSE) operators. Leaf or terminal nodes, on the other hand, are the variables of a particular problem, or constants to be determined.

For example, the expressions $\sigma_0(1-T/T_c)_0^n$ and $\sigma_0(1-T/T_c)_0^n + \sigma_1(1-T/T_c)_1^n$ are codified as we can see in Fig. 1.

The parsimony principle compels us to try to make the quantity of operators and functions used as small as possible, but they have to be able to code all the possible expressions for the solution of the problem.

The quality of an individual is measured by the fitness function. When we have a problem of symbolic regression, the objective is to fit the experimental data. Therefore, it is quite usual to use the relative errors to define it: that is, for an individual F (an expression or a proposed model), the value of the fitness function will be

$$f_F = \frac{100}{N} \sum_i = \frac{1^N |F(i)^{fit} - F(i)^{exp}|}{F(i)^{exp}} \quad (2)$$

where N is the number of experimental data, i is the i -th datum, and the superscripts exp and fit indicate the experimental data and the fitted values by the expression F , respectively.

We can randomly generate J individuals, in a careful way so that they are syntactically correct, that will form the initial population. Maximum size of the population and individuals will be important parameters to choose, as they will influence the evolution of the algorithm and the complexity of the solution. Another option is starting the algorithm

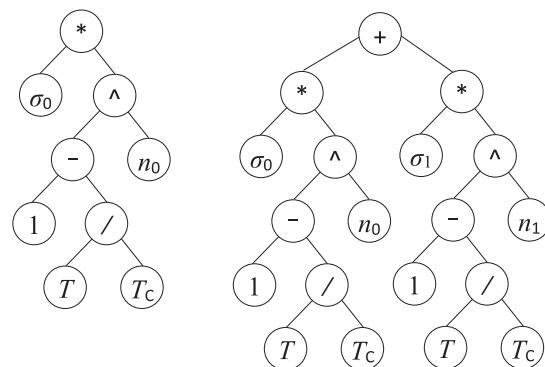


Fig. 1. Example of trees.

with known expressions for certain models, or reasonable modifications.

To select the individuals for the crossovers and/or mutations, there is a number of selection strategies in the literature, e.g. fitness proportionate or elitism, tournament method [47]. In this work, we will attend to their quality, measured by the fitness function. This is a way to make sure that the most suitable individuals have the highest probability of being selected as parents. It seems reasonable, then, to assign the values

$$\frac{p_{Fj}}{\sum_i 1/f_{Fi}} = \frac{1/f_{Fj}}{\sum_i 1/f_{Fi}} \quad (3)$$

$$\frac{p_{Fj}'}{\sum_i 1/f_{Fi}'} = \frac{f_{Fj}}{\sum_i 1/f_{Fi}'} \quad (4)$$

as the probability for an individual F_j to be chosen as a parent, being J the actual size of the population; and the probability for F_j to be selected as one of the eliminated individuals of the population, respectively, $j = 1, \dots, J$.

Once the parents are selected, crossover and mutation are applied according to preselected probabilities p_c and p_m , respectively. In crossover, two parents interchange parts of their trees to produce two offspring following a process of cutting and grafting. Mutation is applied on a node-by-node basis by random alteration of a branch or a terminal node. It consists in generating a subtree, or changing a subtree by a subtree of the same class. It is useful to introduce diversity in the population, when we have a high selective pressure (that is, when the probability of choosing good individuals is much more higher than the one assigned to worse members of the population).

In Figs. 2 and 3, we can see example of crossover, related to the problem that we will analyze later (here, T, T_C and P_C are part of the variables set, and c_1, c_2, c_3, c_4 are constants).

During the search, the expressions may not have optimal values for the constants, that could cause the search path to deviate from the optimum as the searching proceeds. Thus it is necessary to complement the algorithm with an optimization of the constants involved in the expression.

With respect to the stopping condition of the algorithm, it is recommended to fix a maximum number of iterations for the algorithm, but it must be based on the reaching of a fixed quality for the solution. Maybe the algorithm will not lead us to the optimal solution, but it will find a good one in a reasonable time.

The idea of the algorithm is quite simple but, unfortunately, implementing it turns out to be an arduous and complex task. It requires a high computing power, that strongly depends on the implementation and on the chosen parameters and operators. Initially, genetic programming was implemented in LISP, but presently we can find some programming in C++, MatLab, Python, Mathematica or R, among others. For this work, as we explained in the Introduction, we used

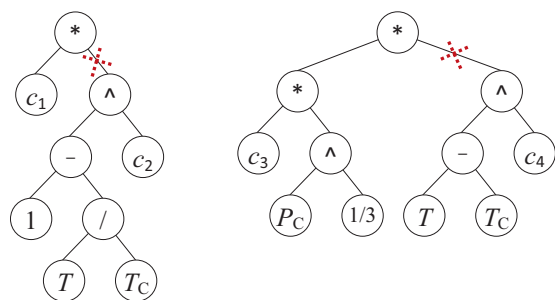


Fig. 2. Parents $c_1(1-T/T_C)_2^2$ (left) and $c_3P_C^{1/3}(T-T_C)_4^2$ (right).

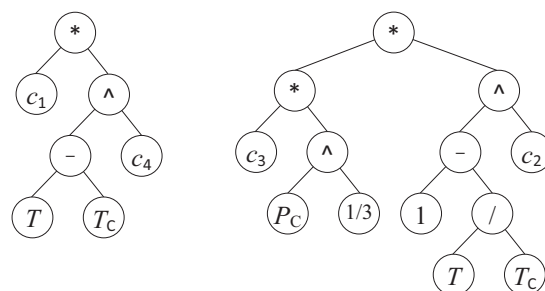


Fig. 3. Offspring from crossover of (2).

DataModeler, a package powered by Mathematica, designed by Evolved Analytics L.L.C., which provides an infrastructure for data exploration, model development, model exploration and model management. It also produces interactive editable and executable reports and representations, while it offers unique visual diagnosis of the data. Besides, it is a very easy tool that can be employed correctly even with very basic knowledge of symbolic regression or Mathematica.

4. Data collection

As we have stated, the theoretical study of the surface tension is quite complex, and includes the use of computer simulations, state equations or predictions based in the study of microscopical interactions. Anyway, they have to be validated through the comparison with experimental data.

The first step for a good model's development is to be provided with a high quality database. The main sources that we considered to obtain reliable experimental data are the databases DIPPR-801 [56] and DETHERM [57], completing them with the data compiled in Wohlfarth and Wohlfarth's book [58].

DIPPR (Design Institute for Physical Properties, Technical Society of the American Institute of Chemical Engineers) is a critically evaluated thermo-physical and environmental property data. The main advantage of the DIPPR database is that it compiles data from a wide range of sources and evaluates them critically. The database is subjected to an extensive GOLD STANDARD evaluation methodology which provides insights into the currency and quality of the database as well guidance for measuring values if data is not available or considered of poor quality or would help improve an estimation method. For the surface tension, this estimation has got a special relevance, because it is often estimated through Sugden's model from 1924 [11]. DIPPR data are used by leading chemical, petroleum, and pharmaceutical companies throughout the world and used extensively in third-party software.

The DETHERM database is produced from the DECHEMA e.V. in cooperation with the DDBST GmbH, Oldenburg. It provides thermophysical property data, and it contains literature values, together with bibliographical information, descriptors and abstracts. The complete database consists of sets of property orientated packages, which are maintained and produced by external experts. This guarantees high quality and checked data.

As noted above, in addition to these three databases we also used data reported in Ref. 58. This volume includes surface tension data for pure liquids and binary liquid mixtures at different temperatures and pressures. In a great number of cases, the data included are also present in either the DIPPR or DETHERM databases.

Some previous considerations have to be made before using the datasets provided by the databases. First, filtering and validating data were necessary, because outlier data and repeated sources appear often. Besides, for certain fluids, there are only available data for intermediate temperatures, or the temperature range is too reduced. Finally,

the update of the databases fails sometimes. And, for many fluids, the only available data come from the predictions made by Sugden [11]. In previous works [34,35], we found that these predictions could significantly disagree from the values provided by experimental measurements and more recent correlation models.

However, these disadvantages are justified, because the achievement of these kind of data is slow, expensive and it involves some uncertainty, not always insignificant. Especially, in limit conditions for temperature (in our case, in the neighborhood of T_C and T_t), for very high (or very low) pressures, when an inadequate design of the devices or experimental techniques has been performed, or because of human mistakes through the measurements, badly calibrated instruments ...

The process of filtering and validating data for surface tension of alcohols from these databases was already made in previous paper (Ref. 44), and that filtered database is the one we employed in this work, discarding the fluids for which at least 90% of the data come from DIPPR predictions using Sugden's model (those marked with S in Ref. 44). Finally, we applied the algorithm to a database composed of the surface tension for 87 alcohols. The final data set was composed of 3570 data.

5. Results

The modeling process explores the trade-off model complexity and model error. A "good enough" model captures the response dynamics without being inappropriately complex.

First of all, we selected variables corresponding to physical parameters which can be part of the model, because they are intrinsically related to surface tension, and they frequently appear in the existing models in the literature. These are the following ones: Temperature (T); Triple, boiling and critical temperatures (T_t, T_b, T_C); Triple and critical pressures (PT_t, P_C); Critical compressibility factor (Z_C); Critical volume (V_C), Molar volume (V_m), Molecular weight (M_w); Radius of gyration (RG) and Acentric factor (ω).

It is not advisable to perform the searching for models taking into account all the variables. It may be too complicated and expensive, and mainly, we want to find a model where the variables have a physical meaning. The parsimony principle compels us to try to make the quantity of operators and functions used as small as possible. Therefore, as a general rule, it is recommendable to explore the data before embarking on a modeling problem. Although visualization of multi-dimensional data is intrinsically difficult, there are several charts we can perform which are targeted at this.

One quite useful is the correlation matrix, which plots the standardized covariance matrix associated with the data. Fig. 4 shows the coupling between inputs as well as our targeted response variable, σ . We can also analyze the correlation chart showed in Fig. 5. For both graphs, positive correlations are color-coded in blue, and negative ones appear in red color. In spite of not being necessary to perform symbolic regression, a rule of thumb for linear modeling is that the input variables should be uncorrelated.

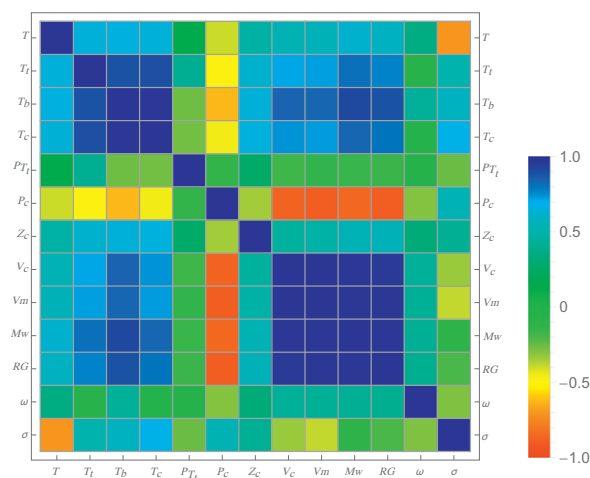


Fig. 4. Correlation matrix.

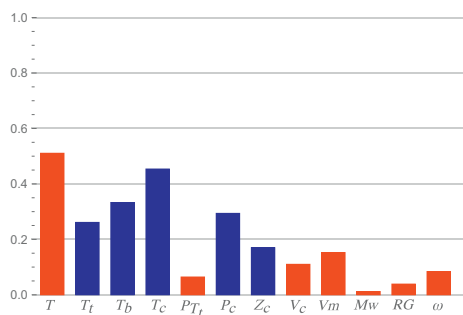


Fig. 5. Correlation chart for all the variables versus σ . Positive correlations are color-coded in blue, and negative ones appear in red color.

Therefore, we arrived at the following conclusions:

Due to the high correlation between the triple, boiling and critical temperatures, we must choose one of them to be part of the model. The chosen one was T_C , because we know that $\sigma(T_C) = 0$, and consequently this is a desirable characteristic for the models that will be built.

We decided to discard the variables PT_t and Z_C , because there is no significant correlation of them with the surface tension.

Besides, the group of variables V_C, V_m, M_w and RG are strongly correlated. We decided to include V_C and V_m in the seeking of the model, discarding M_w and RG . To justify this election, we selected the optimal models with minimum complexity, found through symbolic regression employing all the variables initially selected. There were a total number of 157 models, and we looked for the most frequently employed variables. Those variables were. The exact percentages are shown in Table 1.

In DataModeler, the function Symbolic Regression[] provides a single repository for many modeling controls. The primary decision to use this function is the amount of time which should be devoted (TimeConstraint), the number of independent model searches (IndependentEvolutions) and the mathematical operators (FunctionPatterns) to be used. We employed the default behavior, which performs a least-squares fitness of the model to the data, as well as merges the results from the IndependentEvolutions. We can also specify an InitialPopulation to incorporate baseline models into the search.

Afterwards, in order to make a comparison with traditional models, we also performed a polynomial (linear) models analysis, using the set of variables $T, T_C, P_C, V_C, V_m, \omega$. Strictly speaking, there is no need to study linear models, because symbolic regression procedure will explore such models, and, if they are appropriated, they will be considered. However, polynomial model building has been a staple of empirical models for many years. We used the set of every possible combination of the six chosen variables up to third order with no cross-terms. The main problem is the high complexity of the models, because of the large set of variables. The complexity of an expression is a measure of the computational cost of the expression.

To make the comparisons, we will employ the regression coefficient $1 - R^2$, which is defined as follows for a model $\hat{\sigma}$ and a data set for N individuals (in the case studied here, $N = 3570$):

$$1 - R^2 = \frac{\sum_{j=1}^N (\hat{\sigma}(j) - \sigma(j)^{exp})^2}{\sum_{j=1}^N (\sigma(j)^{exp} - \sigma^e xp)^2} \quad (5)$$

where $\sigma(j)^{exp}$, $j = 1, \dots, N$ are the experimental values for the surface tension, $\sigma^e xp$ is the mean of those values and $\hat{\sigma}(j)$ is the prediction of the model $\hat{\sigma}$ for each individual j .

$1 - R^2$ quantifies the proportion of variance in the surface tension which is not predictable from the model.

If we delete non significant terms, the model is highly simplified, but it is still quite complex. In fact, the results obtained for linear models of different orders can be seen in Table II. The selection criterion was $1 - R^2 < 0.1$.

Table 1
Most frequently variables found in the models using symbolic regression.

Variable	Number of models	Percentage of Models (%)
T	156	100
T_C	156	100
V_m	155	99.4
P_C	137	87.8
ω	93	59.6
V_C	17	10.9

Table 2
Regression coefficient for polynomial models.

Order	Complexity	1-R ²
1	51	0.0578
2	288	0.0325
3	1090	0.0223
4	3256	0.0069
5	8272	0.0049
6	18646	0.0045

Table 3
Measures of error for the six models studied.

Model	Complexity	1-R ²	MAD	MaxAD	MAPD (%)	MaxAPD (%)
10	11	0.739	6.182 × 10 ⁻³	38.54 × 10 ⁻³	47.035	12471.5
11	20	0.211	3.050 × 10 ⁻³	29.97 × 10 ⁻³	13.0381	279.096
12	15	0.214	3.184 × 10 ⁻³	29.73 × 10 ⁻³	14.483	354.905
13	25	0.067	1.567 × 10 ⁻³	32.18 × 10 ⁻³	7.360	165.325
14	21	0.068	1.580 × 10 ⁻³	32.21 × 10 ⁻³	7.771	354.137
15	42	0.043	1.245 × 10 ⁻³	30.73 × 10 ⁻³	6.802	531.964

We can also measure the goodness of the model by computing the absolute and absolute relative percentage deviations of experimental data from the fitted ones. Usual parameters employed are Mean Absolute Deviation (MAD) and the Mean Absolute Percentage Deviation (MAPD), that are defined as follows, for a model $\hat{\sigma}$ and a number of N data.

$$MAD = \frac{1}{N} \sum_j | \hat{\sigma}(j) - \sigma(j)^{exp} | \tag{6}$$

$$MAPD = \frac{1}{N} \sum_j \frac{| \hat{\sigma}(j) - \sigma(j)^{exp} |}{\sigma(j)^{exp} \times 100} \tag{7}$$

Also, Maximum Absolute Deviation (MaxAD) and Maximum Percentage Deviation (MaxAPD) can provide information about the suitability of a model.

$$MaxAD = \max | \hat{\sigma}(j) - \sigma(j)^{exp} | : j = 1, \dots, N \tag{8}$$

$$MaxAPD = \max \frac{| \hat{\sigma}(j) - \sigma(j)^{exp} |}{\sigma(j)^{exp}} \times 100 : j = 1, 2, \dots, N \tag{9}$$

However, values for MaxAPD are especially high for surface tension models (as we can see in Table III), due to the fact that, for temperatures close enough to critical temperature

of the fluid T_c , surface tension takes values close to zero. Consequently, small absolute deviations provide very high relative deviations. In this sense, MAD and MAPD are more representative indicators of the accuracy of a model than MaxAD and MaxAPD.

Then, we compared these polynomial models with the ones obtained through symbolic regression. In Fig. 6, we can see the 558 models found through symbolic regression, showing the coefficient $1-R^2$ versus the complexity of the model. The larger red points correspond to the 157 best models found for a fixed complexity. These models are optimal in the sense that, for a given level of accuracy ($1-R^2$), there is not a simpler model. Conversely, for a fixed complexity, there is not a more accurate model. Notice that we clearly have a knee in the curve and that we achieved high accuracy models.

If we make a comparison between the models obtained by both methods (polynomial regression and symbolic regression), it is obvious that the symbolic regression always provides less complex models for a given level of accuracy (Fig. 7).

Finally, after several executions of the algorithms, we focused on the models with a complexity equal or lower than 50 which include the least number of parameters possible, have a physical meaning and reasonable extrapolation.

The final selected models were

$$\sigma = 4.50 \times 10^{-2} - 5.81 \times 10^{-5} T \tag{10}$$

$$\sigma = 3.39 \times 10^{-3} - (9.52 \times 10^{-5}) T + (8.83 \times 10^{-5}) T_c \tag{11}$$

$$\sigma = 9.19 \times 10^{-5} (T_c - T) \tag{12}$$

$$\sigma = -4.448 \times 10^{-4} + 2.637 \times 10^{-5} (T_c - T) P_c^{1/3} \tag{13}$$

$$\sigma = 2.596 \times 10^{-5} (T_c - T) P_c^{1/3} \tag{14}$$

$$\sigma = (-145.364 - 1.114T + 1.426T_c + 1.741TVm - 3.408T_cVm + 0.323V_c\omega) \times 10^{-4} \tag{15}$$

In Model 11, only the variables T and T_c appear, and the independent term is very close to zero. As we know that the term $T_c - T$ is an usual term for surface tension models (it is a simple transformation of temperature), we built Model 12 as a modification of Model 11 to optimize the complexity.

Given that the constant in Model 13 seems to be negligible, 14 is a modification of Model 13.

Results in Table III confirm the suitability of the best models 13, 14 and 15. It is left to the user the choice of which model to use, but our recommendation would be 14, because of its simplicity and good accuracy results.

6. Conclusions

In this work, symbolic regression (SR) was applied to a database set of 87 alcohols, employing a total quantity of 3570 data that have been carefully selected and filtered in previous works, in order to develop a correlation model for surface tension.

We considered all the physical parameters which had a physical meaning in order to build the model, and made a correlation study to

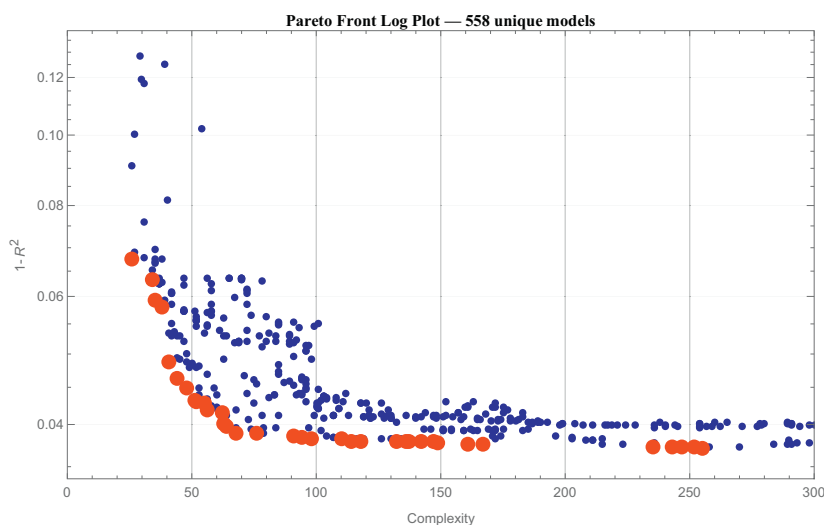


Fig. 6. Comparison between symbolic regression models. Larger red points correspond to the best 157 models.

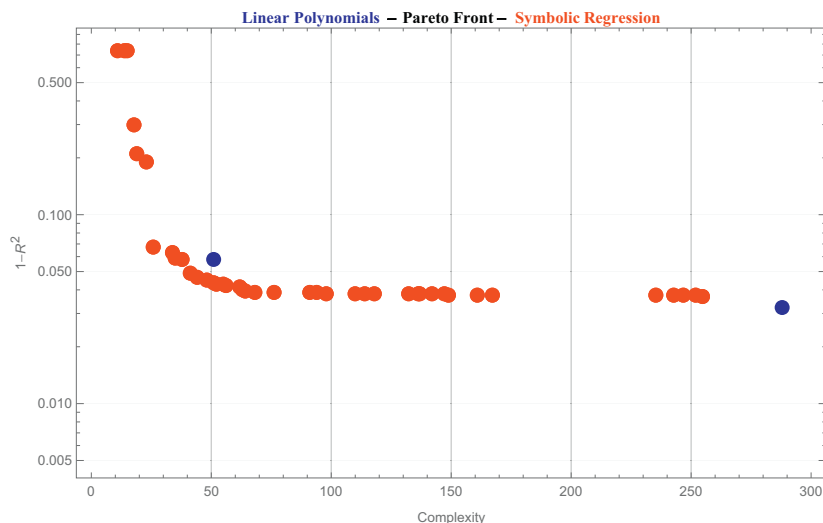


Fig. 7. Comparison between symbolic regression models and polynomial models.

select the most representative ones. The chosen parameters were temperature, critical temperature, critical pressure, critical volume, molar volume and acentric factor.

The comparison for the best models obtained with SR (attending to accuracy and complexity of the model) and the usually employed polynomial regression models, was made through the values of $1-R^2$, always obtaining better results for SR models.

Finally, the best SR models with complexity under 50 were studied and optimized, and different measures of error were computed to show the accuracy of the models.

Notice that this algorithm could be applied to every type of fluids in order to obtain a model providing accuracy and easiness of use. Besides, it can be applied to study any thermodynamic property.

Acknowledgments

This work was supported by GR18108 project from Junta de Extremadura.

References

- [1] C. Kling, *Surface Tension*, vol. 1, , Tell-Tale Press, 2012.
- [2] B.E. Poling, J.M. Prausnitz, J.P. O'connell, *The Properties of Gases and Liquids*, vol. 5, , McGraw-hill, New York, 2001.
- [3] C. Miqueu, D. Broseta, J. Satherley, B. Mendiboure, J. Lachaise, A. Graciaa, An extended scaled equation for the temperature dependence of the surface tension of pure compounds inferred from an analysis of experimental data, *Fluid Phase Equilib.* 172 (2) (2000) 169–182.
- [4] S. Hartland, *Surface and Interfacial Tension: Measurement, Theory, and Applications*, CRC Press, 2004.
- [5] A.W. Neumann, R. David, Y. Zuo, *Applied Surface Thermodynamics*, CRC press, 2010.
- [6] F. Biscay, A. Ghoufi, V. Lachet, P. Malfreyt, Prediction of the surface tension of the liquid-vapor interface of alcohols from Monte Carlo simulations, *J. Phys. Chem. C* 115 (17) (2011) 8670–8683.
- [7] J. Vijande, M. Pineiro, J. García, J.L. Valencia, J.L. Legido, Density and surface tension variation with temperature for heptane+ 1-alkanol, *J. Chem. Eng. Data* 51 (5) (2006) 1778–1782.
- [8] F.A.M.M. Gonçalves, A.R. Trindade, C.S.M.F. Costa, J.C.S. Bernardo, I. Johnson, I.M.A. Fonseca, A.G.M. Ferreira, P.vt, viscosity, and surface tension of ethanol: new measurements and literature data evaluation, *J. Chem. Thermodyn.* 42 (8) (2010) 1039–1049.
- [9] P.G. Aleiferis, Z.R. Van Romunde, An analysis of spray development with iso-octane, n-pentane, gasoline, ethanol and n-butanol from a multi-hole injector under hot fuel conditions, *Fuel* 105 (2013) 143–168.
- [10] C.W. Ye, J. Li, Density, viscosity, and surface tension of n-octanol-phosphoric acid solutions in a temperature range 293.15–333.15 K, *Russ. J. Phys. Chem. A* 86 (10) (2012) 1515–1521.
- [11] S. Sugden, The variation of surface tension with temperature and some related functions, *J. Chem. Soc. Trans.* 125 (1924) 32–41.
- [12] G.R. Somayajulu, A generalized equation for surface tension from the triple point to the critical point, *Int. J. Thermophys.* 9 (9) (1988) 559–566.
- [13] S.R.S. Sastri, K.K. Rao, A simple method to predict surface tension of organic liquids, *Chem. Eng. J. Biochem. Eng. J.* 59 (2) (1995) 181–186.
- [14] G. Di Nicola, C. Di Nicola, M. Moglie, A new surface tension equation for refrigerants, *Int. J. Thermophys.* 34 (12) (2013) 2243–2260.
- [15] G. Di Nicola, M. Moglie, A generalized equation for the surface tension of refrigerants, *Int. J. Refrig.* 34 (4) (2011) 1098–1108.
- [16] G. Di Nicola, M. Pierantozzi, Surface tension prediction for refrigerant binary systems, *Int. J. Refrig.* 36 (2) (2013) 562–566.
- [17] G. Di Nicola, M. Pierantozzi, A new scaled equation to calculate the surface tension of ketones, *J. Therm. Anal. Calorim.* 116 (1) (2014) 129–134.
- [18] F. Gharagheizi, A. Eslamimanesh, A.H. Mohammadi, D. Richon, Use of artificial neural network-group contribution method to determine surface tension of pure compounds, *J. Chem. Eng. Data* 56 (5) (2011) 2587–2601.
- [19] F. Gharagheizi, A. Eslamimanesh, A.H. Mohammadi, D. Richon, Determination of parachor of various compounds using an artificial neural network- group contribution method, *Ind. Eng. Chem. Res.* 50 (9) (2011) 5815–5823.
- [20] F. Gharagheizi, A. Eslamimanesh, B. Tirandazi, A.H. Mohammadi, D. Richon, Handling a very large data set for determination of surface tension of chemical compounds using quantitative structure–property relationship strategy, *Chem. Eng. Sci.* 66 (21) (2011) 4991–5023.
- [21] A. Roosta, P. Setoodeh, A. Jahanmiri, Artificial neural network modeling of surface tension for pure organic compounds, *Ind. Eng. Chem. Res.* 51 (1) (2011) 561–566.
- [22] F. Gharagheizi, A. Eslamimanesh, M. Sattari, A.H. Mohammadi, D. Richon, Development of corresponding states model for estimation of the surface tension of chemical compounds, *AIChE J.* 59 (2) (2013) 613–621.
- [23] J.P. O'Connell, R. Gani, P.M. Mathias, G. Maurer, J.D. Olson, P.A. Crafts, Thermodynamic property modeling for chemical process and product engineering: some perspectives, *Ind. Eng. Chem. Res.* 48 (10) (2009) 4619–4637.
- [24] I. Cachadiña, A. Mulero, M.I. Parra, Prediction of the enthalpy of vapourisation for anhydrides, formates, acetates, propionates, butyrates, esters, and ethers, *Phys. Chem. Liq.* 46 (5) (2008) 564–573.
- [25] A. Mulero, I. Cachadiña, F. Cuadros, Comparison of predictive correlations for the normal boiling density of nonpolar fluids, *Chem. Eng. Commun.* 193 (11) (2006) 1445–1456.
- [26] A. Mulero, I. Cachadiña, F. Cuadros, Calculation of the vaporization enthalpy of nonpolar fluids at the standard temperature, *Chem. Eng. Commun.* 193 (2) (2006) 192–205.
- [27] A. Mulero, I. Cachadiña, M.I. Parra, Liquid saturation density from predictive correlations based on the corresponding states principle. Part 1: results for 30 families of fluids, *Ind. Eng. Chem. Res.* 45 (5) (2006) 1840–1848.
- [28] A. Mulero, I. Cachadiña, M.I. Parra, Liquid saturation density from predictive correlations based on the corresponding states principle. 2. results for 49 families of fluids, *Ind. Eng. Chem. Res.* 45 (20) (2006) 6864–6873.
- [29] A. Mulero, I. Cachadiña, M.I. Parra, Comparison of corresponding-states-based correlations for the prediction of the vaporization enthalpy of fluids, *Ind. Eng. Chem. Res.* 47 (20) (2008) 7903–7916.
- [30] A. Mulero, C.A. Galán, M.I. Parra, F. Cuadros, A. Mulero, Theory and Simulation of Hard-Sphere Fluids and Related Systems, vol. 3, , Springer-Verlag, 2008.
- [31] A. Mulero, M.I. Parra, Improving the prediction of liquid saturation densities from models based on the corresponding states principle, *Phys. Chem. Liq.* 46 (3) (2008) 263–277.

- [32] A. Mulero, M.I. Parra, F. Cuadros, A new analytical model for the prediction of vapor-liquid equilibrium densities, *Int. J. Thermophys.* 27 (2006) 1435–1448.
- [33] M. Pierantozzi, G. Di Nicola, Surface tension correlation of carboxylic acids from liquid viscosity data, *Fluid Phase Equilib.* 482 (2019) 118–125.
- [34] A. Mulero, I. Cachadiña, M.I. Parra, Recommended correlations for the surface tension of common fluids, *J. Phys. Chem. Ref. Data* 41 (4) (2012), 043105.
- [35] A. Mulero, I. Cachadiña, Recommended correlations for the surface tension of several fluids included in the REFPROP program, *J. Phys. Chem. Ref. Data* 43 (2) (2014), 023104.
- [36] A. Mulero, M. Pierantozzi, I. Cachadiña, G. Di Nicola, An artificial neural network for the surface tension of alcohols, *Fluid Phase Equilib.* 449 (2017) 28–40.
- [37] G. Di Nicola, M. Pierantozzi, Surface tension of alcohols: a scaled equation and an artificial neural network, *Fluid Phase Equilib.* 389 (2015) 16–27.
- [38] A. Mulero, M.I. Parra, K.K. Park, F.L. Roman, Vaporization enthalpy of pure refrigerants: comparative study of eighteen correlations, *Ind. Eng. Chem. Res.* 49 (10) (2010) 5018–5026.
- [39] M.I. Parra, A. Mulero, A Mathematica program for the accurate correlation of different thermodynamic properties of saturated pure fluids, *Chem. Eng. Commun.* 200 (3) (2013) 317–326.
- [40] C.A. Galán, A. Mulero, F. Cuadros, Calculation of the surface tension and the surface energy of Lennard–Jones fluids from the radial distribution function in the liquid phase, *Mol. Phys.* 103 (4) (2005) 527–535.
- [41] C.A. Galán, A. Mulero, F. Cuadros, Calculation of the surface tension and the surface energy of Lennard–Jones fluids from the radial distribution function in the interface zone, *Mol. Phys.* 104 (15) (2006) 2457–2464.
- [42] A. Mulero, M.I. Parra, I. Cachadiña, The Somayajulu correlation for the surface tension revisited, *Fluid Phase Equilib.* 339 (2013) 81–88.
- [43] A. Mulero, M.I. Parra, E.L. Sanjuán, I. Cachadiña, Analysis of specific correlations and general models for the surface tension of six liquid oxides, *Fluid Phase Equilib.* 358 (2013) 60–67.
- [44] A. Mulero, I. Cachadiña, E.L. Sanjuán, Surface tension of alcohols. data selection and recommended correlations, *J. Phys. Chem. Ref. Data* 44 (3) (2015), 033104.
- [45] NIST, NIST reference fluid thermodynamic and transport properties database (REFPROP): version 9.1, <https://www.nist.gov/srd/refprop>.
- [46] J. Holland, *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan, USA, 1975.
- [47] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Massachusetts, 1989.
- [48] J.R. Koza, *Genetic Programming: on the Programming of Computers by Means of Natural Selection*, vol. 1, MIT press, 1992.
- [49] R. Poli, W.B. Langdon, N.F. McPhee, J.R. Koza, *A Field Guide to Genetic Programming*, Creative Commons Licenses (2008).
- [50] S.H. Chen, C.H. Yeh, Toward a computable approach to the efficient market hypothesis: an application of genetic programming, *J. Econ. Dyn. Control* 21 (6) (1997) 1043–1063.
- [51] J.F. Miller, D. Job, V.K. Vassilev, Principles in the evolutionary design of digital circuits. part i, *Genet. Program. Evolvable Mach.* 1 (1–2) (2000) 7–35.
- [52] K. Uesaka, M. Kawamata, Synthesis of low-sensitivity second-order digital filters using genetic programming with automatically defined functions, *IEEE Signal Process. Lett.* 7 (4) (2000) 83–85.
- [53] V. Arkov, C. Evans, P.J. Fleming, J.P. Hill, C. D. I. Pratt, D. Rees Norton, K. Rodriguez-Vazquez, System identification strategies applied to aircraft gas turbine engines, *Annu. Rev. Contr.* 24 (2000) 67–81.
- [54] B. McKay, M. Willis, G. Barton, Steady-state modelling of chemical process systems using genetic programming, *Comput. Chem. Eng.* 21 (9) (1997) 981–996.
- [55] W. Cai, A. Pacheco-Vega, M. Sen, K.T. Yang, Heat transfer correlations by symbolic regression, *Int. J. Heat Mass Transf.* 49 (23) (2006) 4352–4359.
- [56] DIPPR, The design institute for physical properties, <http://www.aiche.org/dippr>.
- [57] DETHERM, Gesellschaft für chemische technik und biotechnologie, <http://dechema.de/en/detherm.html>.
- [58] C. Wohlfarth, B. Wohlfarth, *Surface Tension of Pure Liquids and Binary Liquid Mixtures*, Springer, New York, 1997.

Conclusiones Finales

Como conclusiones finales de la presente Tesis Doctoral se pueden destacar las siguientes:

- Se proponen nuevas estrategias de inferencia bayesiana, para aportar mejoras acerca del problema de estimación de los parámetros de las distribuciones límite de valores extremos. Las ideas están basadas en el aprovechamiento de la información contenida en el conjunto completo de observaciones para aliviar el problema de la escasez de datos extremos.
- El modo de ejecutar estas ideas es construir distribuciones a priori altamente informativas, aprovechando las relaciones que se puedan establecer entre los parámetros de la distribución de los datos y los parámetros de la distribución de valores extremos correspondiente. Se trata de una estrategia general que puede extrapolarse más allá de las distribuciones y problemas abordados en esta Tesis.
- Siempre que sea posible, se establecen relaciones analíticas entre los parámetros de ambas distribuciones. Solo para aquellos casos en los que no es posible hacerlo, se recurre a encontrarlas empíricamente a partir de simulaciones masivas, dándole un carácter muy general a la estrategia.
- Se han considerado tres estrategias diferentes para la estimación de los parámetros de las distribuciones límite de valores extremos (máximos de bloque o excesos de un umbral) correspondientes: la primera consiste en estimarlos usando solo los datos extremos considerando una distribución a priori no informativa y aproximando la distribución a posteriori con un algoritmo MH; la segunda consiste en estimar los parámetros de la distribución base con todos los datos y aprovechar la relación existente con los parámetros de la distribución límite; la tercera también se sirve de la relación existente entre

los parámetros de ambas distribuciones, pero en este caso para construir una distribución a priori altamente informativa.

- Simulaciones masivas, considerando distintos tamaños para muestras, bloques y umbrales, así como valores para los parámetros de las distribuciones base, permiten valorar empíricamente hasta qué punto las nuevas estrategias mejoran la exactitud y precisión de las estimaciones. Las mejoras son más acusadas cuanto menor es el tamaño de la muestra considerada y mayor es su variabilidad.
- La incorporación de una cópula en modelos jerárquicos bayesianos, para reproducir la dependencia espacial entre datos extremos, presenta la gran ventaja de reducir los errores de estimación obtenidos cuando se considera la independencia de las observaciones respecto al espacio.
- El modelo con cópula óptimo para valores extremos multivariantes de máximos de bloque en el problema considerado proporciona un modelo de regresión espacial para los parámetros de localización y escala, pero constante para el parámetro de forma de la distribución de Valores Extremos Generalizada.
- El lenguaje de programación R facilita enormemente el trabajo de programación de algoritmos adecuados para poner en práctica las distintas estrategias propuestas. Quizás uno de sus mayores problemas sea su lentitud, pero se puede aliviar bastante con la paralelización y la programación vectorial. En este contexto se han desarrollado algoritmos eficientes, que, siguiendo el espíritu de ser un software libre y abierto, se han publicado en RPubs.
- La regresión simbólica es una técnica de optimización poco extendida, que permite encontrar expresiones analíticas para modelar relaciones entre variables sin necesidad de conocer su estructura funcional, explorando intensivamente en el conjunto de todas las posibles expresiones en tiempo razonable. Su aplicación a datos experimentales de la tensión superficial de alcoholes proporciona modelos generales muy sencillos, con errores porcentuales medios de 7.8 % y $R^2 \geq 0,93$.

Apéndice A. Trabajos en proceso de publicación

Estimación eficiente de medidas de riesgo

Una aplicación del método POT en la Teoría de Riesgo es la estimación de medidas de riesgo financieras. Estas medidas están asociadas a las observaciones que exceden un cierto umbral, y por lo tanto, pueden determinarse de forma aproximada empleando la distribución de Pareto Generalizada. Las medidas de riesgo más conocidas y empleadas son el Valor en Riesgo (VaR) y el Valor en Riesgo Condicional (CVaR), que permiten describir la cola de la función de pérdidas en un contexto financiero.

Definición 6.1. *Sea X una variable aleatoria continua que representa las pérdidas de una inversión en un determinado horizonte temporal. Dado un parámetro $0 < p < 1$, el Valor en Riesgo (VaR_p) de X es el p -cuantil de la distribución de X , esto es,*

$$VaR_p(X) = \inf\{c : P(X \leq c) \geq p\}, \quad (6.1)$$

y el Valor en Riesgo Condicional ($CVaR_p$) de X es la esperanza condicional

$$CVaR_p(X) = E[X | X \geq VaR_p(X)]. \quad (6.2)$$

La estimación de ambas medidas plantea todo un reto, puesto que requiere el conocimiento de la cola de la distribución. Empleando las relaciones existentes entre los parámetros de la distribución base y los parámetros de la GPD propuestas en el artículo B de la presente Tesis Doctoral, se pueden construir distribuciones a priori altamente informativas con el fin de obtener mejores estimaciones para el VaR y CVaR que cuando se consdieran únicamente los valores extremos.

En el trabajo [66] se aplica el algoritmo de Metropolis-Hastings (MH) sobre muestras de distintos tamaños con distribuciones exponencial, Normal, Cauchy y

Gamma, y diferentes valores para sus parámetros. Se realiza un estudio de simulación intensivo para comparar la precisión y exactitud proporcionadas por tres métodos diferentes comprobando cómo mejoran las estimaciones al tener en cuenta la información que contiene el conjunto de datos completo. Además, se analizan datos de un ejemplo real de índices financieros para mostrar la aplicación práctica de los métodos, cuando la distribución base es desconocida, que seguramente sea la mayor debilidad de la estrategia propuesta.

Con este trabajo se muestra que las mejoras observadas con la metodología desarrollada en el artículo B se mantienen cuando se utilizan para la estimación de medidas de riesgo financiero como el VaR y CVaR. En este caso se observa una mayor robustez considerando la estimación de dichas medidas a partir de la distribución a posteriori de los parámetros.

Uso de R para docencia e investigación

R es un lenguaje y un entorno para la computación estadística y la generación de gráficos que proporciona una gran variedad de herramientas ya programadas junto a técnicas gráficas sorprendentes y la gran ventaja de ser altamente extensible. Además, el hecho de ser código abierto facilita la programación de nuevas funciones adaptadas y más eficientes para situaciones concretas.

En la actualidad, es el software por excelencia en cursos de estadística para cualquier nivel y una herramienta casi imprescindible para cualquier trabajo de investigación que implique algún tipo de análisis o visualización de datos. Sin duda, se trata de uno de los lenguajes de programación más empleados no solo en estadística sino incluso para propósitos generales.

R crece muy rápidamente y la cantidad y variedad, tanto de librerías como de entornos gráficos disponibles actualmente, hace muy complicado elegir cuáles se adaptan mejor a nuestras necesidades. De hecho, es habitual encontrar funciones con igual nombre, distintos argumentos y salidas, tal vez sobre un mismo algoritmo (o similar). En este contexto se enmarca el trabajo [67] donde se describen y analizan críticamente distintos recursos de R ofreciendo recomendaciones sobre cómo incorporar esta herramienta para mejorar la investigación y/o docencia. En particular, se muestra cómo iniciarse en su uso, localizar recursos específicos, y programar de manera eficiente simulaciones que sirvan de soporte a la investigación.

Este lenguaje de programación facilita enormemente el trabajo de programación de algoritmos adecuados, como los presentados en los artículos A, B y C.

Quizás uno de sus mayores problemas sea su lentitud, pero se puede aliviar bastante con la paralelización y la programación vectorial. Con esta filosofía se han desarrollado todos los algoritmos necesarios para obtener los resultados presentados en los distintos trabajos que componen la presente Tesis Doctoral de manera lo más eficiente posible. Siguiendo el espíritu de ser un software libre y abierto, se han publicado en RPubS [\[49\]](#) aquellos algoritmos que pudieran ser útiles a otros investigadores, transformados en funciones.

Apéndice B. Actividades desarrolladas durante la Tesis Doctoral

Artículos en revistas JCR

1. Sanjuán, E.L., Parra, M.I. & Pizarro, M.M. (2020). Development of models for surface tension of alcohols through symbolic regression. *Journal of Molecular Liquids*, 298, 111971. doi: [10.1016/j.molliq.2019.111971](https://doi.org/10.1016/j.molliq.2019.111971). Factor de impacto JCR: 6.165 con posición 4/37 (90.54 %) Q1 en la categoría Physics, Atomic, Molecular & Chemical.
2. Martín, J., Parra, M.I., Pizarro, M.M. & Sanjuán, E.L. (2020). Baseline Methods for Bayesian Inference in Gumbel Distribution. *Entropy*, 22 (11), 1-18, 1267. doi: [10.3390/e22111267](https://doi.org/10.3390/e22111267). Factor de impacto JCR: 2.524 con posición 38/86 (56.40 %) Q2 en la categoría Physics, Multidisciplinary.
3. García, J.A., Pizarro, M.M., Acero, F.J. & Parra, M.I. (2021). A Bayesian Hierarchical Spatial Copula Model: An Application to Extreme Temperatures in Extremadura (Spain). *Atmosphere*, 12 (7), 897. doi: [10.3390/atmos12070897](https://doi.org/10.3390/atmos12070897). Factor de impacto JCR: 3.110 con posición 169/279 (39.61 %) Q3 en la categoría Environmental Sciences.
4. Martín, J., Parra, M.I., Pizarro, M.M. & Sanjuán, E.L. (2022). Baseline Methods for the Parameter Estimation of the Generalized Pareto Distribution. *Entropy*, 24 (2), 178. doi: [10.3390/e24020178](https://doi.org/10.3390/e24020178). Factor de impacto JCR: 2.700 con posición 40/85 (53.50 %) Q2 en la categoría Physics, Multidisciplinary.

Artículos bajo revisión en revistas JCR

- Martín, J., Parra, M.I., Sanjuán, E.L. & Pizarro, M.M. New Bayesian method for estimation of Value at Risk and Conditional Value at Risk. *Econometrics & Statistics*. Bajo revisión: revisión inicial. doi: [10.48550/arXiv.2306.12202](https://doi.org/10.48550/arXiv.2306.12202).
- Parra, M.I., Sanjuán, E.L., Robustillo, M.C. & Pizarro, M.M. Using R for teaching and research. *Axioms*. Bajo revisión: revisión inicial. doi: [10.48550/arXiv.2306.12200](https://doi.org/10.48550/arXiv.2306.12200)

Congresos y Seminarios

- Pizarro, M.M., Martín, J., Sanjuán, E.L. & Parra, M.I. (2021). Baseline methods for Bayesian inference in Gumbel copula. En *Proceedings of the 63th ISI World Statistics Congress*, Volumen 11 (p. 16).
- Pizarro, M.M. (2021). Teoría de Valores Extremos: Distribución de Máximos de Bloque. Seminario en Estadística y Actuaría organizado por la Universidad Nacional Autónoma de México.
- Pizarro, M.M., Parra, M.I., García, J.A. & Acero, F.J. (2021). A Bayesian Hierarchical spatial copula model. En *Proceedings of the 14th International Conference of the ERCIM Working Group on Computational and Methodological Statistics*, p. 114.
- Pizarro, M.M., Sanjuán, E.L. & Parra, M.I. (2022). Nuevas estrategias para estimar los parámetros de la GPD. En *Actas del XXXIX Congreso Nacional de Estadística e Investigación Operativa*, p.115.

Participación en Proyectos

- Proyecto *Desarrollo y aplicación de un modelo de predicción de temperaturas extremas en Extremadura*, IB16063, cofinanciado por la Junta de Extremadura y los Fondos FEDER de la Unión Europea.
- Proyecto *Modelización Estocástica Avanzada para el análisis de riesgos. Gestión de riesgo medioambientales*, MTM2017-86875-C3-2-R, cofinanciado por el Ministerio de Ciencia e Investigación, Agencia Estatal de Investigación y los Fondos FEDER de la Unión Europea.

-
- Proyecto *Grupo de Decisión e Inferencia Bayesiana*, GR18108, cofinanciado por la Junta de Extremadura y los Fondos FEDER de la Unión Europea.
 - Proyecto *Ayuda al Grupo de Investigación denominado "Grupo de Decisión e Inferencia Bayesiana"*, GR21057, cofinanciado por la Junta de Extremadura y los Fondos FEDER de la Unión Europea.
 - Proyecto *Modelización estocástica para el aprendizaje automático. Aplicaciones al diagnóstico asistido por ordenador*, PID2021-122209OB-C32, cofinanciado por el Ministerio de Ciencia e Investigación, Agencia Estatal de Investigación y los Fondos FEDER de la Unión Europea.

Bibliografía

- [1] F. J. Acero, M. I. Fernández-Fernández, V. M. S. Carrasco, S. Parey, T. T. H. Hoang, D. Dacunha-Castelle, J. A. García, Changes in heat wave characteristics over extremadura (sw spain), *Theoretical and Applied Climatology* (2017) 1–13. doi:[10.1007/s00704-017-2210-x](https://doi.org/10.1007/s00704-017-2210-x).
- [2] J. Portero Serrano, F. J. Acero Díaz, J. A. García García, Analysis of extreme temperature events over the iberian peninsula during the 21st century using dynamic climate projections chosen using max-stable processes, *Atmosphere* 11 (5) (2020) 506. doi:<https://doi.org/10.3390/atmos11050506>.
- [3] S. Wi, J. B. Valdés, S. Steinschneider, T.-W. Kim, Non-stationary frequency analysis of extreme precipitation in south korea using peaks-over-threshold and annual maxima, *Stochastic Environmental Research and Risk Assessment* 30 (2) (2015) 583–606. doi:[10.1007/s00477-015-1180-8](https://doi.org/10.1007/s00477-015-1180-8).
- [4] F. J. Acero, S. Parey, T. T. H. Hoang, D. Dacunha-Castelle, J. A. García, M. C. Gallego, Non-stationary future return levels for extreme rainfall over extremadura (southwestern iberian peninsula), *Hydrological Sciences Journal* 62 (9) (2017) 1394–1411. doi:[10.1080/02626667.2017.1328559](https://doi.org/10.1080/02626667.2017.1328559).
- [5] F. J. Acero, V. M. S. Carrasco, M. C. Gallego, J. A. García, J. M. Vaquero, Extreme value theory and the new sunspot number series, *The Astrophysical Journal* 839 (2) (2017) 98. doi:[10.3847/1538-4357/aa69bc](https://doi.org/10.3847/1538-4357/aa69bc).
- [6] F. Acero, M. Gallego, J. García, I. Usoskin, J. Vaquero, Extreme value theory applied to the millennial sunspot number series, *The Astrophysical Journal* 853 (1) (2018) 80. doi:[10.3847/1538-4357/aaa406](https://doi.org/10.3847/1538-4357/aaa406).
- [7] A. K. Singh, D. E. Allen, R. J. Powell, Value at risk estimation using extreme value theory, in: *19th International Congress on Modelling and Simulation*,

- Australian Mathematical Sciences Institute, Modelling and Simulation Society of Australia and New Zealand, 2011.
- [8] G. Magnou, An application of extreme value theory for measuring financial risk in the uruguayan pension fund, *Compendium: Cuadernos de Economía y Administración* 4 (7) (2017) 1–19.
- [9] S. Kotz, S. Nadarajah, *Extreme Value Distributions: Theory and Applications*, ICP, 2000.
- [10] S. Coles, J. Bawa, L. Trenner, P. Dorazio, An introduction to statistical modeling of extreme values, Vol. 208, Springer, 2001. doi:[10.1007/978-1-4471-3675-0](https://doi.org/10.1007/978-1-4471-3675-0).
- [11] A. Ferreira, L. de Haan, *Extreme value theory*, Springer Science+ Business Media, LLC, 2006. doi:[10.1007/0-387-34471-3](https://doi.org/10.1007/0-387-34471-3).
- [12] M. Fréchet, Sur la loi de probabilité de l'écart maximum, *Ann. Soc. Math. Polon.* 6 (1927) 93–116.
- [13] R. A. Fisher, L. H. C. Tippett, Limiting forms of the frequency distribution of the largest or smallest member of a sample, in: *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 24(2), Cambridge University Press, 1928, pp. 180–190. doi:[10.1017/S0305004100015681](https://doi.org/10.1017/S0305004100015681).
- [14] B. Gnedenko, Sur la distribution limite du terme maximum d'une serie aleatoire, *Annals of mathematics* (1943) 423–453. doi:[10.2307/1968974](https://doi.org/10.2307/1968974).
- [15] E. J. Gumbel, *Statistics of Extremes* (Dover Books on Mathematics), Dover Publications, 2012.
- [16] A. A. Balkema, L. De Haan, Residual life time at great age, *The Annals of probability* (1974) 792–804. doi:[10.1214/aop/1176996548](https://doi.org/10.1214/aop/1176996548).
- [17] J. Pickands III, Statistical inference using extreme order statistics, *Annals of Statistics* 3 (1975) 119–131.
- [18] J. M. Bernardo, A. F. M. Smith (Eds.), *Bayesian Theory*, John Wiley & Sons, Inc., 1994. doi:[10.1002/9780470316870](https://doi.org/10.1002/9780470316870).
- [19] W. R. Gilks, S. Richardson, D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Vol. 1, 1995. doi:[10.1201/b14835](https://doi.org/10.1201/b14835).

- [20] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian data analysis, Chapman and Hall/CRC, 1995. doi:[10.1201/b16018](https://doi.org/10.1201/b16018).
- [21] M. K. Cowles, B. P. Carlin, Markov chain monte carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association* 91 (434) (1996) 883–904. doi:[10.1080/01621459.1996.10476956](https://doi.org/10.1080/01621459.1996.10476956).
- [22] M.-H. Chen, Q.-M. Shao, J. G. Ibrahim, Monte Carlo Methods in Bayesian Computation, Springer New York, 2000. doi:[10.1007/978-1-4612-1276-8](https://doi.org/10.1007/978-1-4612-1276-8).
- [23] J. Martín, M. I. Parra, M. M. Pizarro, E. L. Sanjuán, Baseline methods for bayesian inference in gumbel distribution, *Entropy* 22 (11) (2020) 1267. doi:[10.3390/e22111267](https://doi.org/10.3390/e22111267).
- [24] J. Martín, M. I. Parra, M. M. Pizarro, E. L. Sanjuán, Baseline methods for the parameter estimation of the generalized pareto distribution, *Entropy* 24 (2) (2022) 178. doi:[10.3390/e24020178](https://doi.org/10.3390/e24020178).
- [25] M. M. Pizarro, J. Martin, E. L. Sanjuán, M. I. Parra, Baseline methods for bayesian inference in gumbel copula, in: *Proceedings 63rd ISI World Statistics Congress*, Vol. 11, 2021, p. 16.
- [26] M. M. Pizarro, E. L. Sanjuán, M. I. Parra, Nuevas estrategias para estimar los parámetros de la GPD, in: *Proceedings XXXIX Congreso Nacional de Estadística e Investigación Operativa*, 2022, p. 115.
- [27] M. Nogaj, P. Yiou, S. Parey, F. Malek, P. Naveau, Amplitude and frequency of temperature extremes over the north atlantic region, *Geophysical Research Letters* 33 (10) (2006). doi:[10.1029/2005gl024251](https://doi.org/10.1029/2005gl024251).
- [28] M. Re, V. R. Barros, Extreme rainfalls in SE south america, *Climatic Change* 96 (1-2) (2009) 119–136. doi:[10.1007/s10584-009-9619-x](https://doi.org/10.1007/s10584-009-9619-x).
- [29] I. Vidal, A bayesian analysis of the gumbel distribution: an application to extreme rainfall data, *Stochastic Environmental Research and Risk Assessment* 28 (3) (2013) 571–582. doi:[10.1007/s00477-013-0773-3](https://doi.org/10.1007/s00477-013-0773-3).
- [30] K. Chinghamu, C.-K. Huang, C.-S. Huang, D. Chikobvu, et al., Extreme risk, value-at-risk and expected shortfall in the gold market, *International Business & Economics Research Journal (IBER)* 14 (1) (2015) 107–122. doi:[10.19030/iber.v14i1.9035](https://doi.org/10.19030/iber.v14i1.9035).

- [31] S. Van der Merwe, D. Steven, M. Pretorius, Bayesian extreme value analysis of stock exchange data, arXiv preprint (2018). doi:[10.48550/arXiv.1804.01807](https://doi.org/10.48550/arXiv.1804.01807).
- [32] M. H. Park, J. H. Kim, Estimating extreme tail risk measures with generalized pareto distribution, Computational Statistics & Data Analysis 98 (2016) 91–104. doi:[10.1016/j.csda.2015.12.008](https://doi.org/10.1016/j.csda.2015.12.008).
- [33] C. K. Wikle, L. M. Berliner, N. Cressie, Hierarchical bayesian space-time models, Environmental and ecological statistics 5 (1998) 117–154. doi:[10.1023/A:1009662704779](https://doi.org/10.1023/A:1009662704779).
- [34] D. Cooley, D. Nychka, P. Naveau, Bayesian spatial modeling of extreme precipitation return levels, Journal of the American Statistical Association 102 (479) (2007) 824–840. doi:[10.1198/016214506000000780](https://doi.org/10.1198/016214506000000780).
- [35] A. C. Davison, S. A. Padoan, M. Ribatet, Statistical modeling of spatial extremes (2012). doi:[10.1214/11-STS376](https://doi.org/10.1214/11-STS376).
- [36] J. A. Daraio, A. O. Amponsah, K. W. Sears, Bayesian hierarchical regression to assess variation of stream temperature with atmospheric temperature in a small watershed, Hydrology 4 (3) (2017) 44. doi:[10.3390/hydrology4030044](https://doi.org/10.3390/hydrology4030044).
- [37] A. M. Barlow, C. Rohrbeck, P. Sharkey, R. Shooter, E. S. Simpson, A bayesian spatio-temporal model for precipitation extremes—stor team contribution to the eva2017 challenge, Extremes 21 (2018) 431–439. doi:[10.1007/s10687-018-0330-z](https://doi.org/10.1007/s10687-018-0330-z).
- [38] J. García, J. Martín, L. Naranjo, F. Acero, A bayesian hierarchical spatio-temporal model for extreme rainfall in extremadura (spain), Hydrological sciences journal 63 (6) (2018) 878–894. doi:[10.1080/02626667.2018.1457219](https://doi.org/10.1080/02626667.2018.1457219).
- [39] M. Sklar, Fonctions de repartition an dimensions et leurs marges, Publ. inst. statist. univ. Paris 8 (1959) 229–231.
- [40] R. B. Nelsen, An introduction to copulas, Springer science & business media, 2007. doi:[10.1007/0-387-28678-0](https://doi.org/10.1007/0-387-28678-0).

- [41] G. Salvadori, C. De Michele, N. T. Kottegoda, R. Rosso, *Extremes in nature: an approach using copulas*, Vol. 56, Springer Science & Business Media, 2007. doi:[10.1007/1-4020-4415-1](https://doi.org/10.1007/1-4020-4415-1).
- [42] P. Jaworski, F. Durante, W. K. Hardle, T. Rychlik, *Copula theory and its applications*, Vol. 198, Springer, 2010. doi:[10.1007/978-3-642-12465-5](https://doi.org/10.1007/978-3-642-12465-5).
- [43] B. Renard, M. Lang, Use of a gaussian copula for multivariate extreme value analysis: Some case studies in hydrology, *Advances in Water Resources* 30 (4) (2007) 897–912. doi:[10.1016/j.advwatres.2006.08.001](https://doi.org/10.1016/j.advwatres.2006.08.001).
- [44] B. Renard, A bayesian hierarchical approach to regional frequency analysis, *Water Resources Research* 47 (11) (2011). doi:[10.1029/2010WR010089](https://doi.org/10.1029/2010WR010089).
- [45] Y. Liu, Y. Li, Y. a. Ma, Q. Jia, Y. Su, Development of a bayesian-copula-based frequency analysis method for hydrological risk assessment—the naryn river in central asia, *Journal of Hydrology* 580 (2020) 124349. doi:[10.1016/j.jhydrol.2019.124349](https://doi.org/10.1016/j.jhydrol.2019.124349).
- [46] N. Beck, C. Genest, J. Jalbert, M. Mailhot, Predicting extreme surges from sparse data using a copula-based hierarchical bayesian spatial model, *Environmetrics* 31 (5) (2020) e2616. doi:[10.1002/env.2616](https://doi.org/10.1002/env.2616).
- [47] J. A. García, M. M. Pizarro, F. J. Acero, M. I. Parra, A bayesian hierarchical spatial copula model: An application to extreme temperatures in extremadura (spain), *Atmosphere* 12 (7) (2021) 897. doi:[10.3390/atmos12070897](https://doi.org/10.3390/atmos12070897).
- [48] M. M. Pizarro, M. I. Parrra, J. A. García, F. J. Acero, A bayesian hierarchical spatial copula model, in: *Proceedings 14th International Conference of the ERCIM Working Group on Computational and Methodological Statistics*, 2021, p. 114.
- [49] M. M. Pizarro, *Inferencia bayesiana para valores extremos* (2023). URL <https://rpubs.com/mariomp/1058194>
- [50] J. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, University of Michigan, USA, 1975.
- [51] D. Goldberg, *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, Massachusetts, 1989.

- [52] J. Koza, Genetic programming: on the programming of computers by means of natural selection, Vol. 1, MIT press, 1992.
- [53] E. Sanjuán, M. Parra, M. Pizarro, Development of models for surface tension of alcohols through symbolic regression, *Journal of Molecular Liquids* 298 (2020) 111971. doi:10.1016/j.molliq.2019.111971.
- [54] F. Gonçalves, A. Trindade, C. Costa, J. Bernardo, I. Johnson, I. Fonseca, A. Ferreira, Pvt, viscosity, and surface tension of ethanol: New measurements and literature data evaluation, *The Journal of Chemical Thermodynamics* 42 (8) (2010) 1039–1049. doi:10.1016/j.jct.2010.03.022.
- [55] F. Biscay, A. Ghoufi, V. Lachet, P. Malfreyt, Prediction of the surface tension of the liquid- vapor interface of alcohols from monte carlo simulations, *The Journal of Physical Chemistry C* 115 (17) (2011) 8670–8683. doi:10.1021/jp1117213.
- [56] C. Ye, J. Li, Density, viscosity, and surface tension of n-octanol-phosphoric acid solutions in a temperature range 293.15-333.15 K, *Russian Journal of Physical Chemistry A* 86 (10) (2012) 1515–1521. doi:10.1134/S0036024412100263.
- [57] P. G. Aleiferis, Z. R. Van Romunde, An analysis of spray development with iso-octane, n-pentane, gasoline, ethanol and n-butanol from a multi-hole injector under hot fuel conditions, *Fuel* 105 (2013) 143–168. doi:10.1016/j.fuel.2012.07.044.
- [58] J. O’Connell, R. Gani, P. M. Mathias, G. Maurer, J. D. Olson, P. A. Crafts, Thermodynamic property modeling for chemical process and product engineering: some perspectives, *Industrial & Engineering Chemistry Research* 48 (10) (2009) 4619–4637. doi:10.1021/ie801535a.
- [59] A. Roosta, P. Setoodeh, A. Jahanmiri, Artificial neural network modeling of surface tension for pure organic compounds, *Industrial & Engineering Chemistry Research* 51 (1) (2011) 561–566. doi:10.1021/ie2017459.
- [60] G. Di Nicola, M. Moglie, A generalized equation for the surface tension of refrigerants, *International journal of refrigeration* 34 (4) (2011) 1098–1108. doi:10.1016/j.ijrefrig.2011.02.008.

- [61] F. Gharagheizi, A. Eslamimanesh, A. H. Mohammadi, D. Richon, Determination of parachor of various compounds using an artificial neural network-group contribution method, *Industrial & Engineering Chemistry Research* 50 (9) (2011) 5815–5823. doi:[10.1021/ie102464t](https://doi.org/10.1021/ie102464t).
- [62] G. Di Nicola, M. Pierantozzi, Surface tension prediction for refrigerant binary systems, *International Journal of Refrigeration* 36 (2) (2013) 562–566. doi:[10.1016/j.ijrefrig.2012.10.004](https://doi.org/10.1016/j.ijrefrig.2012.10.004).
- [63] F. Gharagheizi, A. Eslamimanesh, M. Sattari, A. H. Mohammadi, D. Richon, Development of corresponding states model for estimation of the surface tension of chemical compounds, *AIChE Journal* 59 (2) (2013) 613–621. doi:[10.1002/aic.13824](https://doi.org/10.1002/aic.13824).
- [64] G. Di Nicola, M. Pierantozzi, A new scaled equation to calculate the surface tension of ketones, *Journal of Thermal Analysis and Calorimetry* 116 (1) (2014) 129–134. doi:[10.1007/s10973-013-3555-8](https://doi.org/10.1007/s10973-013-3555-8).
- [65] R. Poli, W. Langdon, N. McPhee, J. Koza, *A Field Guide to Genetic Programming*, 2008.
- [66] J. Martín, M. I. Parra, E. L. Sanjuán, M. M. Pizarro, New bayesian method for estimation of value at risk and conditional value at risk, arXiv preprint (2023). doi:[10.48550/arXiv.2306.12202](https://doi.org/10.48550/arXiv.2306.12202).
- [67] M. I. Parra, E. L. Sanjuán, M. C. Robustillo, M. M. Pizarro, Using R for teaching and research, arXiv preprint (2023). doi:[10.48550/arXiv.2306.12200](https://doi.org/10.48550/arXiv.2306.12200).

