UNIVERSITÀ DEGLI STUDI DI TRIESTE

DIPARTIMENTO DI SCIENZE ECONOMICHE, AZIENDALI, MATEMATICHE E STATISTICHE

Corso di Laurea magistrale in Scienze Statistiche e Attuariali

Modelli di Markov per la valutazione del rischio di ondate di calore: un case-study in condizioni di dipendenza estrema

Relatore:

Professor Francesco Pauli

Laureanda:

Valentina Caliri

ANNO ACCADEMICO 2021-2022

Sommario

I. Introduzione

Nel contesto del rapido cambiamento climatico che sta caratterizzando il nostro pianeta, uno dei fenomeni più diffusi ed impattanti sono le ondate di calore: con il continuo aumento delle temperature globali, sempre più aree geografiche sono vulnerabili a questo tipo di evento meteorologico estremo che può causare danni alla salute umana, alla produzione agricola e, in generale, all'ambiente. In questo contesto, Gorizia, come molte città italiane, è stata interessata da diverse ondate di calore negli ultimi anni, con conseguenti ripercussioni sulla qualità della vita e sull'organizzazione dell'attività umana. La tesi del presente documento è di esaminare le caratteristiche delle ondate di calore verificatesi nella provincia e prevederne un'evoluzione futura, utilizzando un approccio statistico di "eccesso sopra la soglia".

Sotto l'ipotesi che le temperature massime giornaliere durante i mesi da giugno ad agosto seguano un processo di Markov omogeneo del primo ordine con uno spazio degli stati continuo, la distribuzione congiunta delle osservazioni verrà modellata separando la modellazione delle distribuzioni marginali dalla struttura di dipendenza. La distribuzione degli eccessi oltre la soglia sarà modellata usando l'approccio Peak Over Threshold della teoria dei valori estremi mentre la struttura di dipendenza sarà modellata utilizzando tre approcci diversi: parametrico, semi-parametrico e non parametrico. Una volta ottenuta la legge del processo, si potrà procedere a generare scenari di evoluzione delle temperature massime giornaliere dalle quali sarà possibile studiare le caratteristiche delle ondate di calore.

La tesi è strutturata come segue. Il primo capitolo fornisce la definizione di ondata di calore, gli impatti negativi sulla salute umana e sull'ecosistema globale assieme ad una panoramica degli eventi più rilevanti verificatisi in Europa. Nel secondo capitolo vengono

introdotti i dati utilizzati per la modellazione, ovvero la serie storica delle temperature massime osservate nella città di Gorizia dal 1989 al 2022. Il terzo capitolo è dedicato ad una panoramica della metodologia alla base dei modelli che sono utilizzati per prevedere le ondate di calore. Il quarto capitolo è dedicato alla verifica delle assunzioni alla base dei modelli, alla stima dei parametri e all'identificazione della soglia estrema rispetto a cui è eseguita la modellazione. Nel quinto capitolo, infine, sono analizzati in dettaglio gli algoritmi utilizzati per simulare scenari futuri di ondate di calore e sono discussi i risultati delle stime ottenute con ciascuno dei modelli considerati.

Capitolo 1

1.1 Le ondate di calore

Le scienze meteorologiche descrivono le ondate di calore come periodi prolungati durante i quali le temperature risultano insolitamente elevate rispetto a quelle mediamente sperimentate per una data regione. Questi periodi sono spesso associati a elevati tassi di umidità, forte irraggiamento solare e scarsa ventilazione. Si tratta di eventi metereologici che caratterizzano da sempre la storia del pianeta, ma che a causa dei cambiamenti climatici stanno diventando più frequenti, lunghi e intensi in molte parti del mondo, costituendo una minaccia per l'ecosistema globale e la salute della popolazione. Per queste ragioni, molti autori (come, ad esempio, Domeisen et al. 2023, Winter et al. 2017, Winter et al. 2016, Reich BJ et al. 2014, Abaurrea et al. 2007) si sono posti l'obiettivo di studiare il fenomeno delle ondate di calore al fine di sviluppare modelli per predire la frequenza e l'intensità con cui questi eventi potrebbero verificarsi nel futuro.

1.1.1 Definizioni

Viste le diverse peculiarità e caratteristiche che le ondate di calore possono presentare non è possibile fornirne una definizione univocamente accettata e universalmente valida, in quanto le definizioni proposte in letteratura differiscono a seconda degli scopi e dei contesti. L'Intergovernmental Panel on Climate Change (IPCC) e l'Organizzazione mondiale della Sanità (OMS) ne forniscono due, applicabili all'interno di un perimetro di ricerca statistica:

- L'IPCC definisce un'ondata di calore come un periodo di caldo anomalo rispetto ad una soglia di temperatura considerata estrema. Tale soglia può essere definita come il 90-esimo percentile della distribuzione delle temperature o da un valore assoluto come una temperatura di 35°C. In pratica, ciò significa che si verifica un'ondata di calore quando le temperature superano la media stagionale di un dato luogo in modo significativo;
- L'OMS definisce un'ondata di caldo quando la temperatura massima giornaliera supera una temperatura media di riferimento[1] di 5°C per almeno cinque giorni consecutivi.

Entrambe queste definizioni risultano utili per valutare l'impatto delle ondate di calore sulla salute umana e sull'ambiente: la definizione dell'OMS è più orientata verso gli effetti sulla salute, in quanto tiene conto del fatto che il corpo umano ha bisogno di tempo per adattarsi alle temperature elevate mentre la definizione dell'IPCC è più orientata ad osservarne gli effetti sull'ambiente, in quanto si basa su una soglia di temperatura estrema che può causare problemi agli ecosistemi e alle colture.

1.1.2 I rischi correlati

Le condizioni metereologiche estreme sperimentate durante un'ondata di calore possono avere un impatto negativo sulla salute umana e sulla fauna selvatica oltre a causare perdite nella produzione agricola, aumentare il rischio di incendi boschivi e siccità, incrementare la domanda di energia per il raffreddamento degli ambienti e altri fenomeni correlati.

La salute umana risulta fortemente condizionata da questi fenomeni in quanto le ondate di calore impattano sulla termoregolazione corporea; infatti, per mantenere la temperatura in un range di 36°C-37°C, il corpo umano adotta un complesso sistema di meccanismi biologici volti a bilanciare la produzione e la perdita di calore. Il calore prodotto durante l'attività metabolica può essere ceduto all'esterno in diversi modi: per radiazione (attraverso il riscaldamento dell'aria o dell'acqua intorno al corpo), per conduzione (attraverso il contatto con corpi solidi più freddi come il pavimento), mediante respirazione (in quanto l'aria inalata è solitamente più fredda e secca dell'aria espirata), e per evaporazione (ovvero attraverso la sudorazione). Tuttavia, se la temperatura dell'aria

[1] Di solito si assume come benchmark la temperatura massima del periodo 1961-1990

supera o si avvicina a quella corporea, il complesso sistema che regola la termoregolazione può essere compromesso e il calore viene assorbito anziché dissipato. Questa condizione di surriscaldamento può portare all'insorgere di malattie da calore tra cui crampi, eruzioni cutanee, colpi di calore, sincope da calore, disidratazione e aumento del carico cardio vascolare che a sua volta può aggravare malattie cardiovascolari preesistenti. In generale, gli impatti sulla salute del caldo estremo e sostenuto sono stati ampiamente studiati (si veda, ad esempio, da D'Ippoliti et al. 2010, FitzGerald et al. 2016, l'Organizzazione Mondiale della sanità), e tutti gli studi mostrano che queste condizioni metereologiche aumentano sensibilmente il rischio di mortalità soprattutto nelle persone anziane (≥ 65), a causa dei cambiamenti intrinsechi del sistema regolatorio e / o alla presenza di farmaci che possono interferire con la normale omeostasi corporea. Tuttavia, l'entità delle stime varia in base alle diverse definizioni di ondata di calore, alla durata e all'intensità dell'evento esaminato.

Parallelamente alle complicazioni sulla salute umana, le ondate di calore hanno impatti anche sul settore energetico e agricolo a causa delle siccità che spesso le accompagnano e che stanno diventando sempre più rilevanti: la riduzione della capienza dei bacini idrici dovuta alla mancanza di pioggia riduce l'efficienza dei sistemi di raffreddamento delle centrali termiche mentre, la mancanza di acqua danneggia la quantità e la qualità dei raccolti agricoli. Negli ultimi anni, diversi studi hanno evidenziato che le produzioni di mais, frumento, olio e vino sono diminuite nei periodi di caldo prolungato a causa dell'eccessiva secchezza del terreno, delle bruciature, della proliferazione di insetti e di altri eventi correlati (per maggiori dettagli sugli impatti delle ondate di calore sul settore agricolo si veda Fraga et al. 2020, e Bras et al. 2021, IPCC Sixth Assessment Report).

A queste considerazioni si può aggiungere che le ondate di calore aumentano il rischio di incendi boschivi che, incrementando le emissioni di carbonio, contribuiscono a potenziare le emissioni di gas serra considerate da molti studiosi (IPCC Sixth Assessment Report) le principali responsabili dei cambiamenti climatici in atto. Si tratta di un fenomeno che in Europa, dal 2000 al 2017, ha distrutto 8,5 milioni di ettari (circa mezzo milione di ettari ogni anno), ha causato una perdita di coltivazione di più di 54 miliardi di euro (Faivre et al. 2018) e, solo nel corso dell'ultima estate, ha distrutto più di 758 mila ettari di boschi in totale nei paesi dell'Unione Europea (corrispondenti a una superficie di circa 7.580

chilometri quadrati, quasi pari a quella della regione Friuli-Venezia Giulia). Questo dato rappresenta più del doppio della media dei terreni bruciati negli ultimi quindici anni.

1.2 Anticicloni e zone ad alta pressione

Le ondate di calore sono causate principalmente dalla presenza di un'area di alta pressione atmosferica, nota anche come anticiclone.

Per comprendere meglio i concetti di alta e bassa pressione generati dai movimenti dell'aria e delle condizioni atmosferiche associate alle ondate di calore può essere utile fare riferimento all'esempio della brezza marina: quando l'aria entra in contatto con la superficie sabbiosa si riscalda a causa dell'irraggiamento solare e le molecole iniziano a muoversi (in accordo con la legge di Boltzman) e ad allontanarsi, riducendone la densità. A questo punto, l'aria divenuta più leggera, si solleva verso l'alto e, mentre sale, si raffredda in media di circa 1°C ogni 150 metri. Raffreddandosi le molecole si muovono di meno e tendono a raggomitolarsi aumentando nuovamente la densità della massa d'aria, che torna a scendere verso il basso. Tuttavia, anziché scontrarsi con l'aria calda che sale, la massa d'aria fredda viene deviata orizzontalmente dai venti di alta quota verso zone con temperature iso-temperate (verso il mare nell'esempio della brezza marina). Qui, non trovando masse che la contrastano, scende verso la superficie dove, via via, riprende a riscaldarsi e, per effetto della brezza marina, viene nuovamente deviata orizzontalmente verso una zona dalla temperatura simile (la spiaggia nell'esempio in esame) e riprende il ciclo. In base a quanto appena descritto, è possibile identificare le zone di alta e bassa pressione: si avrà bassa pressione dove l'aria è calda in quanto pesando di meno, la forza peso esercitata a parità di superficie è più piccola, e alta pressione dove l'aria è più fredda.

Sebbene il fenomeno della brezza marina possa fornire un esempio locale, esso può essere espanso a livello generale per comprendere i concetti di alta e bassa pressione. Nelle zone di bassa pressione, anche chiamate cicloni, l'aria calda viene spinta verso l'alto e il vapore acqueo contenuto in essa si condensa, generando precipitazioni. Al contrario, nelle zone di alta pressione, anche chiamate anticicloni, l'aria fredda più densa che si trova in alto viene spinta verso la superficie. Questa spinta (chiamata subsidenza) riscalda la massa d'aria per compressione e, assieme all'irraggiamento del terreno, contribuisce al riscaldamento dell'aria circostante.

Volendo riportare un esempio, durante la stagione estiva il bacino del Mediterraneo è colpito da diversi anticicloni, tra cui quello sub-tropicale africano che può favorire il verificarsi di ondate di calore. Esso, infatti, è caratterizzato da temperature e tassi di umidità elevati, ventilazione debole e precipitazioni scarse.

1.3 Ondate di calore in Europa

Negli ultimi anni l'Europa è stata colpita da numerose ondate di calore, delle quali la più disastrosa si verificò nel 2003 in Europa occidentale. Questo evento climatico estremo fu eccezionale sia per l'intensità che, soprattutto, per la sua durata. Si stima che causò circa 40.000 morti in più di 16 paesi. Si trattò di un evento climatico estremo caratterizzato da un calo delle precipitazioni fino a 300 mm (Trenberth et al., 2007). La produzione di mais nella pianura padana registrò un calo del 36%, mentre in Francia, il Paese che risentì di più delle alte temperature[2], diminuì del 30% (Ciais et al., 2005). Sempre in Francia, la produzione di frutta e foraggio subì un calo del 25% e quella di vino fu in generale la più bassa registrata in Europa negli ultimi 10 anni (Copa Cogeca, 2003b). Si stima che complessivamente le perdite economiche per il settore agricolo furono di 13.1 miliardi di euro, di cui le maggiori in Francia, pari a circa 4 miliardi di euro (Sénat, 2004). Oltre alle perdite nel settore agricolo, l'ondata di caldo mise a dura prova i sistemi energetici francesi, poiché l'aumento delle temperature dei fiumi assieme alla riduzione della capienza dei bacini idrici ridusse l'efficienza di raffreddamento delle centrali termiche (convenzionali e nucleari) e sei centrali elettriche vennero completamente chiuse (Létard et al., 2004). Se l'ondata di caldo fosse continuata, fino al 30% della produzione nazionale di energia elettrica sarebbe stata a rischio (Létard et al., 2004).

L'ondata di caldo del 2010 che colpì l'Europa orientale superò in termini di estensione spaziale quella del 2003 (Barriopedro et al., 2011) e causò circa 55.000 morti oltre a una perdita economica totale di circa 15 miliardi di dollari (Peters et al., 2010).

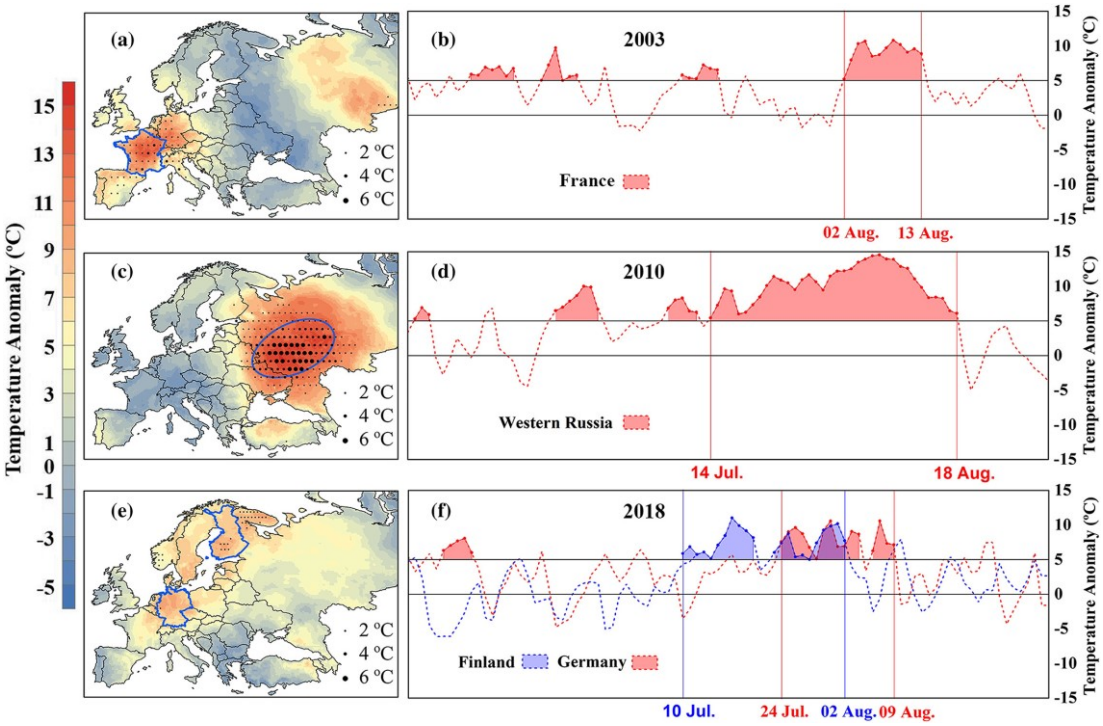
Negli anni successivi sono state registrate diverse altre ondate di calore in Europa, tra le quali quella del 2018 che è stata particolarmente intensa. Questa ondata colpì soprattutto l'Europa centro settentrionale e fu causata da una combinazione di prolungati deficit di

[2] Robine et al. 2008

precipitazioni e temperature insolitamente elevate già a partire dal mese di maggio (Somini, 2018).

A titolo comparativo, nello studio Liu et al. (2020) sono state confrontate le caratteristiche spaziotemporali e le durate delle ondate di calore del 2003, 2010 e 2018. La figura sotto riportata mostra l'evoluzione spaziotemporale delle temperature massime durante i tre periodi[3] di caldo intenso.

Figura 1: Confronto tra le ondate di calore del 2003, 2010 e 2018[4]: in (a), (c), (e) i punti neri indicano le giornate con temperature da record, la loro dimensione indica l'intensità dell'evento mentre, i poligoni blu rappresentano il centro dell'ondata. Invece, in (b), (d), (f) è riportata l'evoluzione nell'anomalia della temperatura (i.e., la differenza tra la temperatura massima in una data giornata ed una assunta come benchmark[5]).



In (a), (c), (e) i punti neri indicano le giornate con temperature molto elevate, la loro dimensione indica l'intensità con cui si sono verificati i superamenti della temperatura massima storica (assunta come benchmark) mentre i poligoni blu rappresentano le zone più colpite dall'ondata. Invece, in (b), (d), (f) è riportata l'evoluzione nell'anomalia della temperatura (i.e., la differenza tra la temperatura massima in una data giornata ed una

[3] Le ondate di calore sono state definite come un periodo in cui per almeno cinque giorni consecutivi la temperatura massima giornaliera supera la media del periodo assunto come riferimento di almeno 5°C.
[4] L'immagine è tratta dallo studio compativo realizzato da Liu et al. 2020.
[5] Temperatura relativa al periodo 1978-2002

assunta come benchmark[6]) in Francia, Russia occidentale, Finlandia e Germania. Le linee verticali indicano l'inizio e la fine delle diverse ondate di calore.

Come si evince dal grafico, la Francia, in assoluto il paese più colpito durante l'ondata del 2003, sperimentò quattro ondate di caldo con l'evento più lungo durato 12 giorni dal 2 al 13 agosto e temperature medie superiori di 9.6 °C rispetto alle medie storiche.

L'ondata di caldo del 2010, maggiore sia per estensione che per durata, colpì soprattutto l'Europa orientale. Durante questo periodo la Russia occidentale sperimentò quattro ondate, di cui la più lunga verificatasi dal 14 luglio al 18 agosto, (36 giorni), registrando una temperatura di 13.4 °C superiore alla media.

Infine, l'ondata di caldo del 2018 interessò soprattutto Finlandia e Germania. In Finlandia l'ondata di caldo durò dal 9 luglio al 2 agosto (20 giorni) e la temperatura massima raggiunse i 33.1°C, un dato estremamente insolito per una regione situata vicino al circolo polare artico con una temperatura media estiva di 15 °C. La Germania, invece, sperimentò l'ondata di caldo dal 23 luglio al 9 agosto (18 giorni). Questo eccezionale evento causò la chiusura di molte centrali nucleari, poiché l'acqua calda dei fiumi non era in grado di raffreddare adeguatamente i reattori (Vogel et al., 2019), in aggiunta, i ridotti livelli di ossigeno nell'acqua calda causarono la morte di molti pesci d'acqua dolce (Leistung, 2018). Secondo l'Associazione tedesca degli agricoltori, nel 2018 il raccolto di colza diminuì del 30%, mentre quello di cereali del 20% (Grieshaber, 2018). Sebbene l'anomalia durante l'ondata di caldo del 2018 sia stata inferiore a quella degli altri due eventi, la sua persistenza (con una temperatura media superiore a quella storica di 5°C), fu sufficiente a causare gravi danni, in particolare nel nord Europa.

Diverse ondate di caldo hanno segnato anche l'estate del 2022 con episodi che si sono distinti, oltre che per la loro persistenza e intensità, anche per la loro precocità. I dati del Copernicus Climate Change Service (C3S) mostrano che nell'ultimo anno la prima ondata si è abbattuta sull'Europa occidentale già nel mese di maggio, anche se l'ondata più calda è arrivata durante la metà di luglio. Durante questo periodo il Regno Unito ha sperimentato per la prima volta temperature superiori ai 40°C e in Portogallo nella città di Pinhão si è osservata una delle temperature più alte mai registrata in Europa pari a

[6] L'anomalia è stata calcolata perdendo come riferimento il periodo 1978-2002 (per ulteriori dettagli si veda Liu et al. 2020).

47.0°C. Un'ulteriore ondata si è abbattuta in Europa nel mese di agosto, l'ultima, di durata inferiore rispetto alle precedenti, ha colpito la Francia in settembre con temperature che hanno raggiunto i 40.1°C.

1.3.1 Ondate di calore nel mondo

In generale, l'aumento della frequenza e dell'intensità delle ondate di calore non è un fenomeno circoscritto al continente Europeo, ma diverse altre regioni stanno sperimentando cambiamenti simili[7]: l'Africa, per esempio, sta vivendo cambiamenti significativi, con fronti caldi che durano più a lungo rispetto a tre decenni fa[8]. Altre regioni come Mongolia, Cina, America, India[9] e Australia[10] hanno visto un'intensificazione nella frequenza e nell'intensità delle ondate di calore anche se i cambiamenti sono stati meno significativi rispetto all'Europa.

1.4 I cambiamenti climatici

Secondo numerosi studiosi, il principale responsabile dei cambiamenti climatici in atto è l'attività umana, in particolare l'aumento delle emissioni di gas serra. L'effetto serra, causato dalla presenza di gas come anidride carbonica (CO_2), metano (CH_4) e ossido di azoto (N_2O) nell'atmosfera, trattiene il calore del sole e causa un riscaldamento globale del pianeta.

Alcuni autori sostengono che molti dei cambiamenti non sarebbero stati possibili senza l'influenza umana sul clima. Ad esempio, con riferimento al fenomeno delle ondate di calore, Rahmstorf e Coumou (2011) hanno suggerito che il record di calore dell'ondata del 2018 non si sarebbe verificato senza il riscaldamento climatico causato dall'uomo.

Inoltre, dall'IPCC Sixth Assessment Report emerge che la maggior parte dei modelli testati nell'ensemble multi-modello CMIP6 concordano sul fatto che sia le temperature massime che quelle minime aumenteranno e che questi cambiamenti saranno amplificati a livelli di riscaldamento globale più elevati. Quindi con la variazione della distribuzione

[7] Perkins (2015), Bom et al. (2016), e Alexander e Arblaster (2017).
[8] Fontaine et al., 2013; Mouhamed et al., 2013; Ceccherini et al., 2016, 2017; Forzieri et al., 2016; Moron et al., 2016; Russo et al., 2016.
[9] Erdenebat e Sato, 2016; You et al., 2017; Xie et al., 2020, Ratnam et al., 2016; Rohini et al., 2016.
[10] Russo et al., 2015; Forzieri et al., 2016; Sánchez-Benítez et al., 2020.

delle temperature, e più specificatamente di quelle massime, sarà ragionevole aspettarsi anche un incremento dei picchi di temperature durante la stagione estiva e, più in generale, delle intensità con cui si verificano le ondate di calore.

Oltre al riscaldamento globale, esistono anche altri fattori che influenzano le ondate di calore e l'aumento delle temperature, come l'umidità del suolo. Ad esempio, Stefanon et al. (2014) stimano che la diminuzione dell'umidità del suolo abbia causato fino al 20% delle anomalie di temperatura nell'Europa occidentale durante le varie ondate di calore e secondo l'IPCC Sixth Assessment Report la deforestazione potrebbe aver contribuito a circa un terzo del riscaldamento degli estremi caldi in alcune regioni di media latitudine dall'epoca preindustriale.

L'azione dell'aerosol è un altro aspetto che molti studiosi considerano rilevante nell'aumento delle temperature e nell'aumento della frequenza e intensità delle ondate di caldo estremo. Infatti, negli anni '50 -'80 circa, la presenza di elevate quantità di aerosol nell'atmosfera ha portato ad un fenomeno chiamato "oscuramento globale", che ha provocato un raffreddamento delle temperature in alcune zone del pianeta con conseguente diminuzione della radiazione solare globale. Successivamente, a partire dagli anni '80, il fenomeno opposto chiamato "schiarimento globale", ha portato ad un aumento della radiazione solare e quindi ad un aumento delle temperature. È interessante notare che uno studio condotto da Wild et al. nel 2005 ha dimostrato che il raffreddamento indotto dall'aerosol ha ritardato la velocità con cui le temperature avrebbero raggiunto livelli estremi. Tuttavia, studi successivi evidenziano che la diminuzione dei carichi di aerosol a partire dagli anni '90 ha indotto un fenomeno opposto, portando ad un riscaldamento accelerato delle temperature estreme in alcune regioni del pianeta. In aggiunta, secondo Dong et al. (2017), una parte significativa del riscaldamento degli estremi più caldi nell'Europa occidentale dalla metà degli anni '90 è causato dalla diminuzione delle concentrazioni di aerosol nella regione.

Anche l'urbanizzazione contribuisce al riscaldamento delle temperature nelle aree abitate, a causa dell'isola di calore urbana, che determina temperature più elevate rispetto alle zone circostanti.

Al contrario, la destinazione d'uso del suolo può influire sulle temperature estreme: diversi

studi hanno dimostrato che, ad esempio, l'agricoltura senza lavorazione[11] e l'irrigazione possono contribuire a ridurre le temperature estreme (Davin et al., 2014) e quindi, se non la frequenza, almeno l'intensità delle ondate di calore.

In conclusione, la maggior parte degli studi concorda sul fatto che il riscaldamento globale porterà ad un aumento della frequenza e dell'intensità delle ondate di calore, anche se potrebbero esserci alcune differenze a livello regionale (legate alle caratteristiche del suolo, la concentrazione di aerosol, il livello di urbanizzazione, ecc). Di conseguenza, la necessità di studiare le ondate di calore sta diventando sempre più urgente per poter monitorare, prevenire e mitigare gli effetti negativi causati da queste condizioni climatiche estreme.

[11] L'agricoltura senza lavorazione è una tecnica agronomica che evita di utilizzare tutte quelle pratiche orientate al deterioramento del suolo.

Capitolo 2

2.1 Within-cluster o Over-cluster

In letteratura esistono diversi approcci per l'analisi e la previsione delle ondate di calore, tra cui l'approccio "within-cluster", che si focalizza sul singolo evento con lo scopo di studiare il comportamento dell'ondata di calore "tipo" che potrebbe interessare una area geografica, e l'approccio "over-cluster" che considera tutte le possibili ondate di calore che potrebbero colpire una certa zona in un determinato periodo di tempo (come, ad esempio, la stagione estiva). A seconda della prospettiva scelta, diverse grandezze possono risultare di interesse, come ad esempio l'intensità dell'evento e la sua durata. Entrambe possono essere analizzate sia nella prospettiva "within-cluster" che "over-cluster". Nel primo caso, si calcolano rispettivamente la probabilità che durante un'ondata di calore la temperatura raggiunga un valore massimo fissato, e l'arco temporale per cui persiste il singolo evento. Nel secondo, invece, si valutano rispettivamente la probabilità che in un periodo di riferimento si verifichino diverse ondate di calore le cui temperature raggiungono o superano delle soglie fissate, e la durata complessiva degli eventi che hanno colpito il territorio nel periodo considerato. Un'altra quantità di interesse è l'estensione spaziale che misura l'area geografica colpita dall'evento o, più in generale, l'intera area interessata dalle diverse ondate di calore lungo la stagione.

Per la prospettiva "over-cluster" possono essere individuate due ulteriori misure rilevanti: la frequenza, ovvero il numero di ondate che si verificano in una stagione e la temporalità,

determinata dalla data del primo evento registrato e conclusa alla data dell'ultimo giorno dell'ultimo evento registrato.

Queste grandezze sono utili per comprendere gli impatti delle ondate di calore sulla salute umana e sulle infrastrutture, nonché per valutare i fabbisogni energetici. Inoltre, aiutano a comprendere più a fondo le interazioni con altri fenomeni, come ad esempio la siccità, in quanto all'aumentare della durata, della frequenza e dell'intensità delle ondate di calore aumenta anche la probabilità del verificarsi di una o più siccità durante la stagione.

2.2 Obiettivi

Lo scopo dell'analisi condotta nei capitoli seguenti è quello di esaminare e prevedere le caratteristiche delle ondate di calore nella provincia di Gorizia e, utilizzando un approccio statistico di "eccesso sopra la soglia" (in linea con quanto suggerito dall'Intergovernmental Panel on Climate Change), si utilizzerà la definizione di ondata di calore come un periodo di lunghezza arbitraria di giorni non per forza consecutivi le cui temperature massime superano una soglia considerata estrema. Inizialmente saranno esaminate le proprietà delle ondate di calore secondo la prospettiva "within-cluster", concentrandosi soprattutto sulla durata e sull'intensità, per poi prendere in considerazione la prospettiva "over-cluster" esaminando gli eventi che si verificano in una stagione estiva al fine di calcolare il numero atteso di ondate, nonché la loro durata e intensità.

Più nel dettaglio siano:

- v la soglia usata per definire l'ondata di calore;
- η una soglia di temperatura considerata ancora più estrema di v (i.e., $\eta > v$);
- N il numero di giorni che eccedono la soglia;
- N_C il numero di giorni consecutivi che eccedono la soglia;
- M la temperatura massima osservata durante l'ondata di calore.

Partendo da queste definizioni e sotto la prospettiva "within-cluster" si stima dapprima la $Pr(N = n | N \geq 1)$, ovvero la probabilità che un'ondata di calore duri n giorni sapendo che l'ondata di calore si è verificata, in quanto almeno un'osservazione ha ecceduto la soglia considerata estrema. Successivamente si andrà a stimare la $Pr(N_C = n | N_C \geq 1)$ ovvero la probabilità che un'ondata di calore duri n giorni ma che queste eccedenze si

verifichino in modo consecutivo. Le due distribuzioni sopra citate forniscono informazioni sulla durata dell'ondata di calore senza fornire, però, alcun dettaglio sull'intensità con cui si presenta. Per questo motivo sarà necessario calcolare anche le due seguenti probabilità $Pr(N = n| M = \eta)$ e $Pr(N_C = n| M = \eta)$, rispettivamente le probabilità che l'ondata di calore abbia una durata di n giorni sapendo che la massima temperatura è stata pari a η, e ancora più interessante, la probabilità che l'ondata di calore duri n giorni consecutivi sapendo che la temperatura massima è stata pari a η. In aggiunta, utilizzando queste come punto di partenza, si calcoleranno anche le probabilità che l'ondata di calore duri n giornate, consecutive o non consecutive, sapendo che il massimo è stato almeno pari a η, ovvero $Pr(N = n| M \geq \eta)$ e $Pr(N_C = n| M \geq \eta)$.

Estendendo, poi, i risultati dalla prospettiva "within-cluster" a quella "over-cluster" si calcoleranno due ulteriori metriche: il numero atteso di ondate di calore durante la stagione estiva e la probabilità che durante la stagione si verifichi almeno un'ondata di calore con certe caratteristiche specificate, quali ad esempio: la probabilità che nella stagione si verifichi almeno un'ondata di calore con una certa durata, piuttosto che la probabilità che si verifichi almeno un'ondata di calore con k giornate che eccedono la soglia v e la cui temperatura massima osservata sia almeno pari a η.

2.2.1 Il modello in sintesi

Ai fini del calcolo delle quantità descritte nel precedente paragrafo supporremo che le temperature massime giornaliere durante i mesi da giugno ad agosto seguano un processo di Markov omogeneo del primo ordine con uno spazio degli stati continuo. Sotto questa ipotesi, per modellare la distribuzione congiunta delle osservazioni verrà utilizzata la teoria delle copule, separando la modellazione delle distribuzioni marginali dalla struttura di dipendenza. In particolare, la distribuzione degli eccessi oltre una soglia (che ci permetterà di ottenere la distribuzione marginale delle temperature massime giornaliere) sarà modellata usando l'approccio Peak Over Threshold della teoria dei valori estremi. La struttura di correlazione sarà invece modellata utilizzando tre approcci diversi: parametrico, semi-parametrico e non parametrico. Una volta ottenuta la legge del processo, si potrà procedere a generare le simulazioni delle traiettorie del processo delle temperature massime giornaliere con le quali, successivamente, identificare le ondate di calore e stimare le grandezze desiderate.

Nel prossimo capitolo, presenteremo il modello e tutti i riferimenti teorici utilizzati mentre in quanto segue sono presentati i dati che saranno alla base delle analisi.

2.3 I dati

Il modello descritto nel paragrafo precedente sarà applicato alle temperature massime osservate nella città di Gorizia, provincia del Friuli-Venezia Giulia, situata a 86m sul livello del mare. I dati utilizzati per le analisi sono stati forniti dal dott. Rodolfo Gratton, membro dell'Unione Meteorologica del Friuli-Venezia Giulia e responsabile del sito internet MeteoGo.it.

Il dataset a disposizione comprende osservazioni delle temperature massime dal 1952 al 2022, malgrado i dati raccolti tra il 1952 e il 1988 e quelli dal 1989 al 2022 presentano alcune differenze nella modalità di rilevamento: le temperature del periodo dal 1952 al 1988 sono approssimate al grado intero più vicino mentre le successive forniscono anche la prima cifra decimale; l'intervallo di rilevazione delle temperature del primo periodo, per ciascuna giornata, iniziava alle 9:00 a.m. del giorno precedente e terminava ventiquattro ore dopo mentre, nel secondo lasso temporale, l'inizio è stato fissato alle 0:00, sempre per ventiquattro ore. La discrepanza nell'orario di inizio della rilevazione, concettualmente di poco conto, ha come effetto una diversa tabulazione della temperatura massima e minima di giornata: solitamente, infatti, la temperatura minima si osserva di notte, entro le 24 ore dall'inizio della rilevazione e nel giorno di calendario al quale sono associate le rilevazioni, mentre la massima, tipicamente osservata all'ora del pranzo, è rilevata il giorno precedente poche ore dopo l'inizio della rilevazione. Al fine di evitare possibili distorsioni nel giorno di osservazione della temperatura massima, si è scelto di limitare la serie storica alle solo osservazioni dal 1989 al 2022 in modo da garantire coerenza e omogeneità del set di dati su cui basare le analisi e ridurre al minimo le possibili problematiche legate alle diverse modalità di rilevamento tra i dati del periodo precedente (1952-1988) e quello successivo (1989-2022).

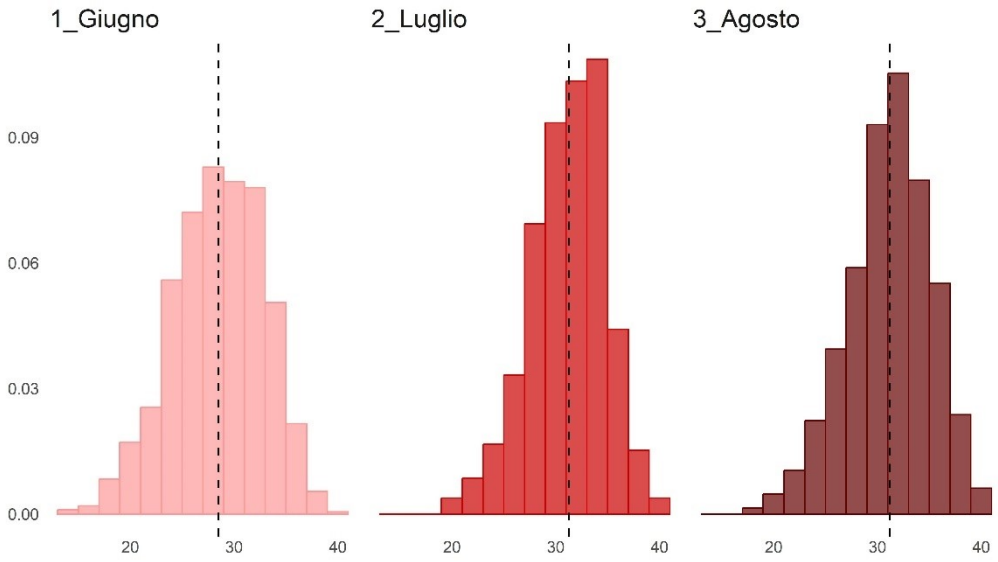
2.3.1 Analisi dei dati

In linea con quanto si trova comunemente nella letteratura, si è scelto di identificare la stagione estiva con i soli mesi di giugno, luglio e agosto. La scelta di escludere maggio e settembre dipende dall'effetto di stagionalità interna alla stagione (dovuto alle minori

temperature osservate) che verrebbe introdotto nel modello e che, se non convenientemente eliminato, potrebbe portare a effetti distorsivi sulla scelta delle soglie che saranno successivamente usate per la modellazione.

Di seguito sono riportati gli istogrammi dei tre mesi estivi considerati. La linea tratteggiata rappresenta la temperatura media massima relativa a ciascun periodo.

Figura 2: Istogrammi delle temperature per i mesi di giugno, luglio e agosto costruiti a partire dalla serie storica delle temperature massime osservate a Gorizia dal 1989 al 2022. La linea tratteggiata rappresenta la temperatura media massima di ciascun mese.



Come si osserva dai grafici, tutte e tre le distribuzioni presentano una leggera asimmetria a sinistra ma, nonostante ciò, in tutti i mesi considerati si osservano temperature superiori ai 34°C / 35°C con picchi fino a oltre 40°C.

Nella tabella che segue sono riportati alcuni valori sintetici della distribuzione delle temperature massime giornaliere assieme ai principali quantili e la media della distribuzione sul periodo 1989-2022.

Tabella 1: Quantili della distribuzione delle temperature massime giornaliere (1989-2022)

Min	1st Qu.	Mediana	3 st. Qu	Max
14.20	27.70	30.79	33.30	40.50

Tabella 2: Indici sintetici della distribuzione delle temperature massime giornaliere (1989-2022)

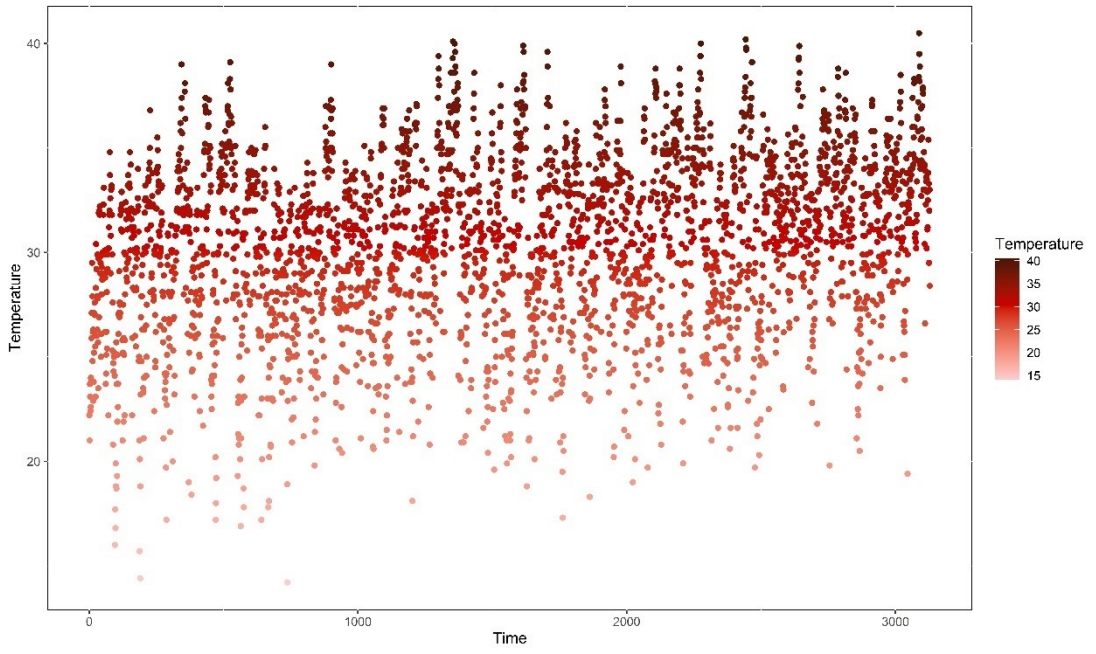
Sd	CV	Media	Skewness	Kurtosis
4.22	0.14	30.35	-0.42	0.02

La distribuzione delle temperature massime giornaliere presenta una leggera asimmetria a sinistra: la media è leggermente minore della mediana e l'indice di asimmetria assume valore negativo e in particolare pari a -0.42. Nonostante l'asimmetria a sinistra, una quota sostanziosa di osservazioni assume determinazioni elevate e in particolare superiori ai 34°C (i.e., circa il 20% delle osservazioni).

La temperatura media del periodo è di 30.4°C con un valore massimo di 40.5°C e uno minimo di 14.2°C osservati rispettivamente nel corso dell'estate del 2022 (22-luglio-2022) e nel corso dell'estate del 1997 (2-giugno-1997). In termini di temperature massime, le estati del 2003 e quella del 2022 sono state le più calde mai sperimentate con una temperatura media massima sui 3 mesi considerati di 34.2°C e 33.6°C, rispettivamente di 3.8°C e 3.2°C più calde rispetto al dato di lungo periodo pari a 30.4°C.

In ultimo, il grafico a dispersione sotto riportato mostra l'evoluzione delle temperature massime giornaliere rispetto al tempo ed è utile per cogliere la presenza di tendenze nelle osservazioni. Un eventuale tendenza potrebbe essere il risultato dei cambiamenti climatici in atto, e in particolare del riscaldamento globale, che stanno causando una traslazione della distribuzione della temperatura medie terrestre verso valori più elevati e di riflesso anche della distribuzione delle temperature massime e minime.

Figura 3: Grafico a dispersione delle temperature massime giornaliere nella città di Gorizia dal 1989 al 2022: sull'asse delle ascisse sono riportati i punti ordinati per data (che per semplicità di rappresentazione sono identificati con un indice da 0 a 3128) e sull'asse delle ordinate le temperature massime osservate.



Da un'attenta analisi grafica del fenomeno, emerge una velata tendenza lineare a conferma dei cambiamenti climatici in atto. Tuttavia, non trattandosi di un trend eccessivamente marcato e al fine di consentire una più diretta interpretazione dei risultati, trascureremo questa tendenza e assumeremo che il set di osservazioni sia assimilabile ad un processo stazionario, senza operare alcuna de-trendizzazione dei dati.

Capitolo 3

3.1 Modelli per quantificare i rischi connessi alle ondate di calore

Una tecnica particolarmente adatta allo studio delle ondate di calore[12] prevede la costruzione e la modellazione di cluster di valori estremi. In letteratura, un cluster di valori estremi viene definito come un insieme di osservazioni dal valore eccezionalmente elevato che sono localmente raggruppate ma non necessariamente consecutive.

In linea con quanto accade in molti fenomeni fisici, la dipendenza temporale tra gli eventi determina la loro inclinazione a manifestarsi in gruppi. Si pensi, ad esempio, ad una giornata di caldo intenso: è probabile che i giorni di calendario prossimi a quello eccezionale siano anch'essi caldi; come è intuibile, però, che questa informazione non influisce sulla probabilità del verificarsi di eventi estremi in un momento più distante nel tempo. Quindi, ai fini di una corretta rappresentazione del fenomeno è necessario introdurre modelli statistici che tengano adeguatamente conto della dipendenza temporale presente nei dati.

Sotto l'ipotesi che le osservazioni seguano un processo di Markov omogeneo del primo ordine con uno spazio degli stati continuo, si è scelto di modellare la distribuzione congiunta delle osservazioni sulla base della teoria delle copule, separando la modellazione

[12] Si considera l'accezione di ondata di calore come un periodo di lunghezza arbitraria di giornate anche non consecutive in cui le temperature massime superano una determinata soglia predefinita.

delle distribuzioni marginali dalla struttura di dipendenza: per la prima viene utilizzato l'approccio Peak Over Threshold, mentre per la seconda sono proposti tre approcci diversi: parametrico, semi-parametrico e non parametrico.

Nei paragrafi seguenti, a completamento della trattazione, verrà introdotta la teoria dei valori estremi, utilizzata per la modellazione marginale e, successivamente, si procederà a definire i cluster di valori estremi e le tecniche utilizzate per stimare la struttura di dipendenza.

3.2 La teoria dei valori estremi

La teoria dei valori estremi si occupa di costruire modelli per la valutazione di eventi estremi, ovvero di eventi che si presentano con bassa frequenza ma che assumono determinazioni molto distanti dalla maggior parte delle osservazioni disponibili nel campione. Differentemente dall'inferenza statistica parametrica classica, questa teoria offre diversi vantaggi: non richiede di assegnare a priori la distribuzione della serie storica, perché la teoria, stimati i parametri, individua già la sua forma funzionale e, in aggiunta, modella la coda della distribuzione basandosi solo su osservazioni considerate estreme anziché su tutto il campione disponibile.

Considerato un campione di n osservazioni $X_1,\dots,X_n$ indipendenti e identicamente distribuite, in letteratura sono proposti due approcci per modellare gli eventi estremi: l'approccio classico, o anche detto Block Maxima, e l'approccio Peak over Threshold.

3.2.1 Approccio Block Maxima

L'approccio Block Maxima studia il comportamento del massimo

$$M_n = max(X_1 ,\dots,X_n)$$

dove $X_1 ,\dots,X_n$ sono una sequenza di variabili aleatorie indipendenti e identicamente distribuite con comune funzione di ripartizione F. In via teorica si potrebbe determinare la distribuzione del massimo per ogni valore di x e di n secondo l'uguaglianza:

$$Pr(M_n \leq z) = \{F(z)\}^n.$$

Tuttavia, questo risultato non ha alcuna utilità pratica in quanto, indipendentemente dalla distribuzione dei dati, la distribuzione di M_n converge in distribuzione alla variabile

aleatoria degenere $\bar{X} = inf\{x \in R: F(x) = 1\}$ per $n \to \infty$. Questa limitazione può essere superata normalizzando la variabile aleatoria M_n come descritto dal seguente teorema:

Teorema 3.1 (Teorema dei Tre Tipi, Tippet-Fisher 1928, Gnedenko 1943): Siano $X_1, \ldots, X_n$ una sequenza di osservazioni indipendenti e identicamente e sia $M_n = max(X_1, \ldots, X_n)$, se esistono $a_n > 0,\ b_n \in R$ e una variabile aleatoria L con funzione di ripartizione non degenere H tale che

$$a_n M_n + b_n \ \to^d \ L$$

allora, a meno di un cambio di locazione e scala, H deve essere o una Gumbel o una Frèchet o una Weibull negativa. Le tre distribuzioni possono essere riassunte sotto un'unica forma funzionale detta distribuzione generalizzata dei valori estremi (GEV):

$$G_\xi(x) = exp\left(-\left[1 + \xi \frac{x - \mu}{\sigma}\right]_+^{-\frac{1}{\xi}}\right),$$

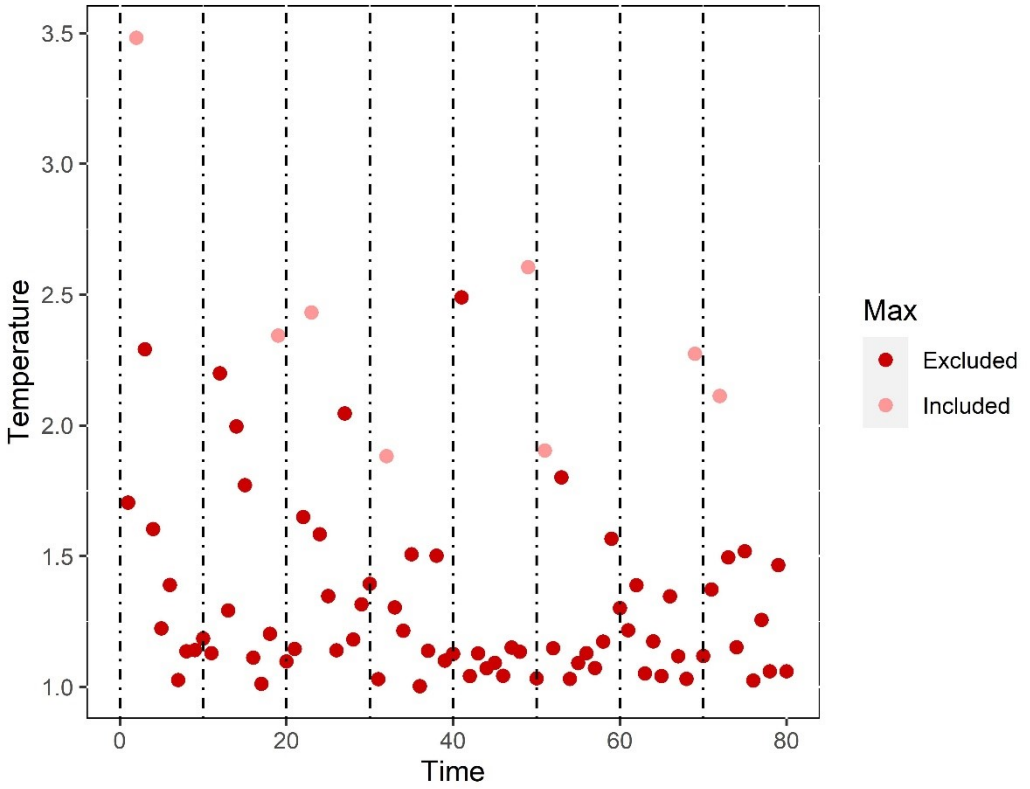
dove la parte positiva è definita dalla funzione $c_+ = max(c, 0)$. Se la comune funzione di ripartizione delle osservazioni è dotata di densità, ovvero le variabili aleatorie sono assolutamente continue, allora si può dimostrare che esiste una successione $a_n > 0,\ b_n \in R$ che rende soddisfatto il teorema.

Nota la forma funzionale, la distribuzione del massimo può essere stimata mediante la seguente procedura: si suddivide il dataset in blocchi e si costruisce un nuovo campione di punti, che sarà usato per la stima dei parametri, estraendo il massimo da ogni blocco.

La principale limitazione del metodo così descritto è rappresentata dalla scelta del numero di blocchi che comporta un compromesso tra distorsione e varianza degli stimatori ottenuti: aumentando la dimensione dei blocchi si aumenta la probabilità di selezionare osservazioni che sono davvero estreme ma, allo stesso tempo, la riduzione della dimensione campionaria produce un aumento della varianza degli stimatori; riducendo la dimensione dei blocchi, aumenta la numerosità campionaria ma si possono includere nel campione anche osservazioni che non sono davvero estreme, con conseguente rischio di ottenere stime distorte.

La figura che segue riporta il nuovo campione di punti ottenuto con il metodo Block Maxima.

Figura 4: Campione di punti selezionato mediante l'approccio Block Maxima: il dataset viene suddiviso in blocchi e si costruisce un nuovo campione di punti, che sarà usato per la stima dei parametri, estraendo il massimo da ogni blocco (punti in rosa) mentre gli altri sono esclusi (punti in rosso).



Le classi di funzioni F che rendono soddisfatto il teorema dei tre tipi formano i domini di attrazione delle distribuzioni dei valori estremi:

$$D = \{F: \ \exists\, a_n > 0, b_n \in R: \ a_n M_n + b_n \ \rightarrow^d L \text{ con } L \sim GEV\}$$

A seconda della pesantezza della coda destra, le varie distribuzioni vengono attratte da uno dei tre domini, in particolare:

- Il dominio di attrazione della Gumbel contiene distribuzioni a coda leggera, la cui funzione delle code converge a zero come un esponenziale e questa caratteristica le consente di avere momenti finiti di ogni ordine. Alcuni esempi di distribuzioni che appartengono a questa categoria sono la Normale, la Log-normale, l'Esponenziale e le sue generalizzazioni come la Gamma e la Weibull;
- Il dominio di attrazione della Frèchet contiene distribuzioni a coda pesante, la cui coda converge a zero come una potenza, solitamente hanno supporto superiormente illimitato, e proprio a causa della pesantezza delle code non tutti i

momenti sono finiti. Appartengono a questa classe la Pareto, la t di Student, la Cauchy e la f di Fisher;

- Il dominio di attrazione della Weibull negativa contiene distribuzioni senza coda con supporto superiormente limitato. Fanno parte di questa categoria, ad esempio, la Beta e l'Uniforme.

3.2.2 Approccio Peak over Threshold

L'approccio Peak over Threshold si concentra sullo studio del comportamento delle osservazioni che eccedono una soglia fissata u:

$$Y_n = X_n - u \mid X_n > u$$

dove $X_1, \ldots, X_n$ è una sequenza di variabili aleatorie indipendenti e identicamente distribuite con comune funzione di ripartizione F e posto $X \triangleq X_n$ e $Y \triangleq Y_n$, la distribuzione dell'eccesso oltre la soglia può essere ricavata come segue:

$$F_u(y) = Pr(Y \leq y) = Pr(X - u \leq y | X > u) = \frac{Pr(u < X \leq y + u)}{Pr(X > u)}$$

$$= \frac{F(y+u) - F(u)}{1 - F(u)}$$

dove la penultima uguaglianza segue dal teorema delle probabilità composte.

Se la funzione di ripartizione delle osservazioni fosse nota, attraverso la formula sopra riportata si potrebbe ricavare anche la distribuzione degli eccessi oltre la soglia. Tuttavia, nella maggior parte dei casi pratici, la funzione di ripartizione non è nota e va stimata a partire dal campione, il quale, nel suo complesso, può fornire una visione poco precisa della coda.

Per tenere adeguatamente conto delle informazioni in essa contenute, infatti, l'approccio Peak over Threshold propone di stimare la distribuzione del solo eccesso di soglia, non considerando la restante parte del campione, e per fare ciò ci si avvale dei risultati del seguente teorema.

Teorema 3.2 (Balkema & de Haan 1974, Pickands 1975): Siano $X_1, \ldots, X_n$ indipendenti e identicamente e $M_n = max(X_1, \ldots, X_n)$; se esistono $a_n > 0, b_n \in R$ e una

variabile aleatoria L non degenere con distribuzione generalizzata dei valori estremi (i.e., GEV):

$$G_\xi(x) = exp\left(-\left[1+\xi\,\frac{x-\mu}{\sigma}\right]_+^{-\frac{1}{\xi}}\right), \qquad c_+ = max(c,0)$$

tale che

$$a_n M_n + b_n \xrightarrow{d} L$$

allora la distribuzione degli eccessi per $u \to \bar{x}$ (estremo superiore di X definito da $\bar{x} = sup\{x: F(x) < 1\}$) converge a:

$$F_u(y) \xrightarrow{d} W_\xi\left(\frac{y}{\sigma_u}\right) \qquad \forall y: 0 \le y \le \bar{x} - u$$

con W_ξ distribuzione di Pareto Generalizzata (i.e., GPD) e $\sigma_u = \sigma + \xi(u-\mu)$ con la seguente forma funzionale:

$$W_\xi\left(\frac{y}{\sigma_u}\right) = 1 - \left[1+\xi\,\frac{y}{\sigma_u}\right]_+^{-\frac{1}{\xi}} \qquad y \ge 0,\ c_+ = max(c,0)$$

dove se:

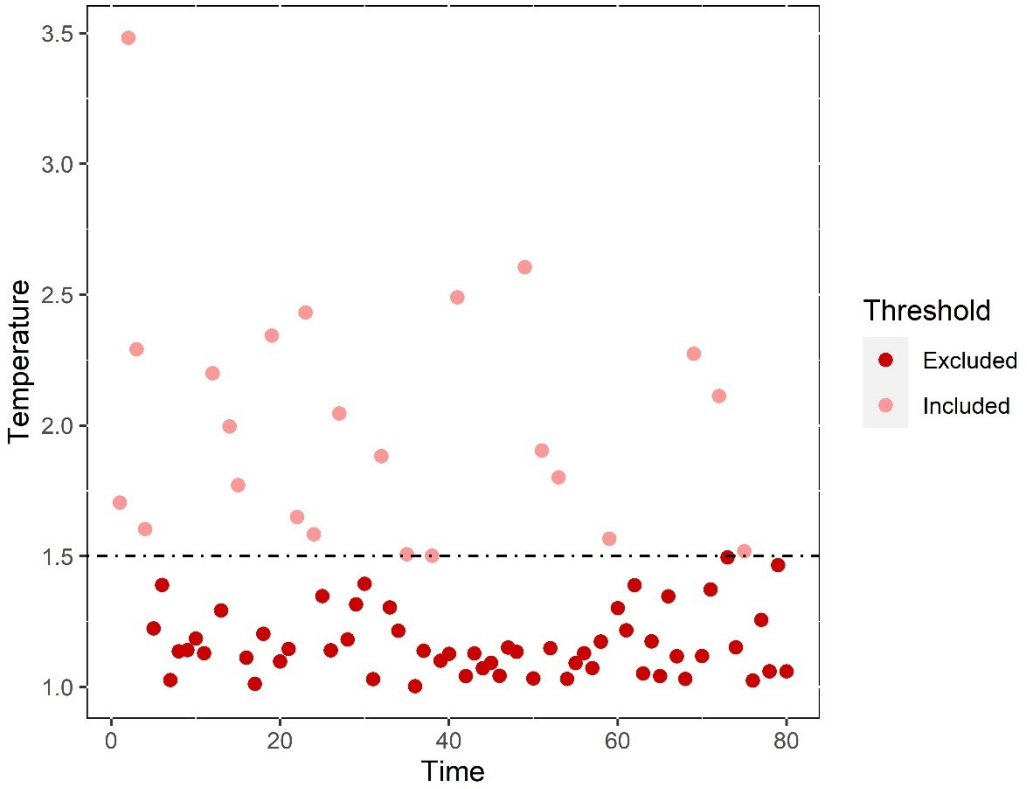
- $\xi \to 0$ la distribuzione è esponenziale;
- $\xi > 0$ è una Pareto;
- $\xi < 0$ è una Beta.

In conclusione, nota la forma funzionale, la distribuzione degli eccessi oltre la soglia può essere stimata fissando la soglia e costruendo un nuovo campione di punti da utilizzarsi per la stima dei parametri, considerando, dunque, solo le osservazioni che eccedono il limite considerato.

Come nel metodo Block Maxima, anche per il Peak over Threshold la principale limitazione è legata al valore della soglia selezionato che implica un compromesso tra distorsione e varianza; una soglia elevata riduce la dimensione campionaria e quindi causa un aumento della varianza degli stimatori, ma aumenta la probabilità di selezionare osservazioni che sono davvero estreme; un valore piccolo aumenta la numerosità campionaria ma allarga il campione, aumentando il rischio di includervi osservazioni non estreme e quindi di ottenere stime distorte.

Nella figura che segue è riportato il nuovo campione di punti secondo l'approccio Peak over Threshold.

Figura 5: Campione di punti selezionato mediante l'approccio Peak Over Threshold: il dataset viene suddiviso in due, i valori che eccedono la soglia (punti in rosa) che sono utilizzati per la stima dei parametri e i valori che non eccedono la soglia (punti in rosso) che sono scartati.



3.2.2.1 Proprietà del metodo POT

La distribuzione di Pareto Generalizzata gode delle seguenti proprietà statistiche che saranno utilizzate anche nel seguito. Sia $Y \sim$ GPD (u, σ_u, ξ) allora:

I. $E(Y) = \frac{\sigma_u}{1-\xi}$ $\xi < 1$, per $\xi \geq 1$ la media non è definita

II. $\forall\, v > 0,\ Y - v \mid Y > v \sim \text{GPD}(v,\ \sigma_u + v\,\xi,\ \xi)$.

In aggiunta, la proprietà *II* implica che la Pareto Generalizzata si conserva quando si considerano i suoi eccessi oltre una soglia e dunque, la distribuzione risultante è ancora una Pareto Generalizzata con diverso parametro di scala $\sigma_v = \sigma_u + \xi v$ e stesso parametro di forma ξ. L'ultimo risultato segue da:

$$Pr(Y - v > y | Y > v) = \frac{Pr(Y > v + y)}{Pr(Y > v)} = \frac{\left[1 + \xi \frac{v + y}{\sigma_u}\right]_+^{-\frac{1}{\xi}}}{\left[1 + \xi \frac{v}{\sigma_u}\right]_+^{-\frac{1}{\xi}}} = \left[\frac{\sigma_u + \xi v + \xi y}{\sigma_u + \xi v}\right]_+^{-\frac{1}{\xi}}$$

$$= \left[1 + \xi \frac{y}{\sigma_u + \xi v}\right]_+^{-\frac{1}{\xi}}$$

dove la prima uguaglianza segue dal teorema delle probabilità composte mentre le successive dall'ipotesi che $Y \sim$GPD.

Dalle proprietà *I* e *II*, inoltre, è possibile individuare due criteri utili per la determinazione della soglia:

- **Mean Residual Life Plot:** sia $Y \sim$ GPD(u, σ_u, ξ), allora riesce che la mean excess function:

  $$e(v) = E[Y - v | Y > v] = \frac{\sigma_u + \xi v}{1 - \xi} \quad \xi < 1$$

  è una funzione lineare della soglia $\forall\, v > 0$. Pertanto, la soglia u può essere fissata pari al più piccolo valore per cui il grafico del seguente stimatore è lineare da quel punto in avanti:

  $$\widehat{e(u)} = \frac{1}{k} \sum_{j=1}^{k} (x_j - u)$$

  dove k è il numero di osservazioni che eccedono la soglia.

- **Parameter Stability Plot:** sia $Y \sim$ GPD(u, σ_u, ξ), allora per la proprietà *II* riesce che la distribuzione degli eccessi oltre una soglia v è ancora una GPD di parametri (v, σ_v, ξ), con ξ invariato e $\sigma_v = \sigma_u + \xi v$. Ponendo $\sigma_{mod} = \sigma_v - \xi v = \sigma_u$ riesce che ξ, σ_{mod} sono costanti $\forall\, v > 0$. Pertanto, la soglia u può essere fissata in modo tale che i parametri modificati siano approssimativamente costanti da lì in avanti.

Individuata la soglia, la stima dei parametri si ottiene con il metodo della massima verosimiglianza. Supposto che $y_1, \dots, y_k$, siano i k eccessi oltre la soglia, la log verosimiglianza per $\xi \neq 0$ è pari a:

$$l(\sigma_u, \xi) = -k\, log(\sigma_u) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} log\left[1 + \xi \frac{y_i}{\sigma_u}\right]^{-\frac{1}{\xi}}$$

se $1 + \xi \frac{y_i}{\sigma_u} > 0$ per $i = 1, \dots, k$, altrimenti $l(\sigma_u, \xi) = -\infty$. Nel caso in cui $\xi = 0$ la log verosimiglianza si ottiene come:

$$l(\sigma_u) = -k\, log(\sigma_u) - \frac{1}{\sigma_u} \sum_{i=1}^{k} y_i.$$

Ottimizzando numericamente tale funzione, si ricavano le stime dei parametri e la matrice hessiana che, in base alle proprietà asintotiche degli stimatori di massima verosimiglianza, può essere utilizzata per approssimare le varianze degli stimatori.

Inoltre, dalla distribuzione degli eccessi oltre la soglia si può ricavare la distribuzione empirica delle osservazioni sotto l'assunto che la GPD(u, σ_u, ξ) costituisca un modello adeguato per le eccedenze sopra la soglia di una variabile X. Dal teorema di disintegrabilità della probabilità segue che per $x > u,\ 1 + \xi \frac{x-u}{\sigma_u} > 0$:

$$\begin{aligned} Pr(X > x) &= Pr(X > x | X > u) Pr(X > u) + Pr(X > x | X \leq u) Pr(X \leq u) \\ &= Pr(X > x | X > u) Pr(X > u) = Pr(X - u > x - u | X > u) Pr(X > u) \\ &= \left[1 + \xi \frac{x-u}{\sigma_u}\right]^{-\frac{1}{\xi}} P(X > u) = \left[1 + \xi \frac{x-u}{\sigma_u}\right]^{-\frac{1}{\xi}} \varsigma_u \end{aligned}$$

dove ς_uè solitamente stimato come $\frac{k}{n}$ (i.e., numero di osservazioni che eccedono la soglia sul numero totale di osservazioni). Dal momento che il numero di eccedenze oltre la soglia segue una distribuzione binomiale di parametri n e ς_u, $\frac{k}{n}$ risulta essere anche la stima di massima verosimiglianza.

Dalla distribuzione delle osservazioni si può anche definire il livello che si eccede in media una volta ogni m osservazioni come soluzione della seguente equazione:

$$\left[1 + \xi \frac{x_m - u}{\sigma_u}\right]^{-\frac{1}{\xi}} \varsigma_u = \frac{1}{m}$$

e, riarrangiando i termini, per m sufficiente grande da assicurare che $x_m > u$, si ottiene

$$x_m = \begin{cases} u + \frac{\sigma_u}{\xi} \left[(m\varsigma_u)^{\xi} - 1\right] & \xi \neq 0 \\ u + \sigma_u\, log(m\varsigma_u) & \xi = 0 \end{cases}$$

Il grafico che si ottiene ponendo sull'asse delle ascisse m e sull'asse delle ordinate i livelli di ritorno x_m dà informazioni sulla pesantezza delle code, in particolare la curva sarà una retta se $\xi = 0$, concava se $\xi > 0$ e convessa se $\xi < 0$.

3.3 Estensione a variabili dipendenti

Per modellare fenomeni naturali, spesso, è necessario estendere la teoria dei valori estremi, solitamente applicata ad osservazioni indipendenti e identicamente distribuite, ad una sequenza di variabili dipendenti e più precisamente ad un processo stazionario.

3.3.1 Processo stazionario

Si considerino $X_1, \dots, X_n, \dots$ variabili aleatorie. Queste costituiscono un processo stazionario se $\forall n, h \; e \; (t_1, \dots, t_n) \in N$ riesce che:

$$F(x_{t_1}, \dots, x_{t_n}) = F(x_{t_1+h}, \dots, x_{t_n+h}).$$

Quindi, la distribuzione rimane invariata se calcolata su due sequenze di punti che distano h, cioè purché la stessa traslazione h sia applicata a tutti gli indici. Si parla anche di invarianza per traslazioni rigide.

Considerando ad esempio (X_1, X_2) e $h = 1$, questa coppia di variabili aleatorie avrà la medesima distribuzione di (X_{101}, X_{102}) ma non di (X_{101}, X_{103}).

Inoltre, l'assunzione di stazionarietà implica che le singole variabili del processo $X_1, \dots, X_n, \dots$ siano identicamente distribuite.

3.3.2 L'indice estremo

L'assunzione di stazionarietà nei processi stocastici introduce due forme di dipendenza temporale tra le osservazioni, quella a breve raggio, anche detta locale, e quella a lungo raggio. La dipendenza locale caratterizza le proprietà delle osservazioni vicine nel tempo e la tendenza dei fenomeni estremi a manifestarsi in cluster, come ad esempio, le giornate di caldo ravvicinate di un'ondata di calore. La dipendenza a lungo raggio, invece, determina le relazioni tra osservazioni distanti nel tempo, per esempio tra giornate di caldo che avvengono in stagioni estive diverse.

Sotto opportune condizioni (che restringono la dipendenza a lungo raggio), si può dimostrare che la teoria dei valori estremi, precedentemente presentata per osservazioni indipendenti e identicamente distribuite, rimane valida anche per un processo stazionario ed in particolare si ottiene che il tipo di distribuzione generalizzata dei valori estremi

utilizzato per descrivere il comportamento del massimo M_n è lo stesso che si avrebbe se le osservazioni fossero indipendenti e identicamente distribuite. In questo caso la distribuzione del massimo dipende da un indice, definito "extremal index". Questa quantità, legata alla presenza di dipendenza a breve raggio, misura il grado con cui le osservazioni tendono a manifestarsi in gruppo: considerando il caso delle ondate di calore, questo quantifica il grado con cui giornate eccezionalmente calde tendono a manifestarsi in maniera ravvicinata.

La condizione che garantisce la validità della convergenza della distribuzione dei massimi per un processo stazionario è la seguente:

Condizione di dipendenza debole $D(u_n)$ (Leadbetter, 1983): Sia $X_1, \dots, X_n, \dots$ una sequenza di osservazioni stazionarie e siano:

- $F_{i_1,\dots,i_n}(x_1, \dots, x_n) = Pr(X_{i_1} \leq x_1, \dots, X_{i_n} \leq x_n)\ \forall n, i_1, \dots, i_n$;
- $F_{i_1,\dots,i_n}(u, \dots, u) = F_{i_1,\dots,i_n}(u)\ \forall n, i_1, \dots, i_n, u$;
- $\{u_n\}$ una sequenza di costanti.

Allora la sequenza di osservazioni $X_1, \dots, X_n, \dots$ soddisfa la condizione $D(u_n)$ se per ogni n, l e ogni scelta di interi $i_1, \dots, i_p, j_1, \dots, j_k$ tali che

$$1 \leq i_1 < i_2, \dots, < i_p < j_1, \dots, < j_k \leq n \quad \text{e} \quad j_1 - i_p \geq l$$

avviene che

$$|F_{i_1,\dots,i_p,j_1,\dots j_k}(u_n) - F_{i_1,\dots,i_p}(u_n) - F_{j_1,\dots j_k}(u_n)| < \alpha_{n,l}$$

dove per $n \to +\infty$ si ha che $\alpha_{n,l_n} \to 0$ per qualche sequenza l_n tale che $l_n = o(n)$.

Nonostante la definizione complicata, la condizione assicura che la dipendenza a lungo raggio tra le osservazioni diminuisce abbastanza rapidamente al crescere della distanza temporale tra di esse: fissata la sequenza di soglie u_n, il cui valore cresce all'aumentare di n, e considerati due gruppi di variabili sufficientemente distanti, questi risultano asintoticamente indipendenti. Tale condizione risulta soddisfatta per tutti i processi markoviani con spazio degli stati continuo e distribuzione di transizione non degenere.

Da questo risultato discende il seguente teorema:

Teorema 3.3: Sia $X_1, \dots, X_n, \dots$ una sequenza stazionaria tale che $M_n = max(X_1, \dots, X_n)$ ha distribuzione generalizzata dei valori estremi non degenere per delle costanti $c_n > 0, d_n \in$

R, allora se la condizione $D(u_n)$ è soddisfatta per ogni sequenza u_n con $u_n = \frac{x}{c_n} + d_n$ e $-\infty \leq x \leq +\infty$, la distribuzione di M_n è nel dominio di attrazione delle distribuzioni dei valori estremi.

In aggiunta, si può dimostrare che sotto opportune condizioni di regolarità la distribuzione del massimo dipende dall'extremal index. Vale infatti il seguente teorema:

Teorema 3.5 (Chernick, 1981): se per ogni $\tau > 0$, $u_n = u_n(\tau)$ e $n\,Pr(X_t > u_n) \rightarrow \tau$ e la condizione $D(u_n)$ è soddisfatta, allora riesce che $Pr(M_n \leq u_n) \rightarrow e^{-\theta\tau}$, $0 \leq \theta \leq 1$.

Si può anche dimostrare che:

Teorema 3.6: Posto:

- $\widehat{M_n} = max(\hat{X}_1, \ldots, \hat{X}_n)$ con $\hat{X}_1, \ldots, \hat{X}_n$ indipendenti e identicamente distribuite;
- F la comune funzione di ripartizione della sequenza stazionaria $X_1, \ldots, X_n$ analoga a quella della sequenza $\hat{X}_1, \ldots, \hat{X}_n$;
- $M_n = max(X_1, \ldots, X_n)$ di $X_1, \ldots, X_n$.

Allora $\forall \tau > 0$, per ogni sequenza $u_n(\tau)$ tale che $n\,Pr(X_t > u_n) \rightarrow \tau$ con $u_n = \frac{x}{a_n} + b_n$ e qualunque sia $0 < \theta \leq 1$ riesce che:

$$Pr(M_n \leq u_n) \rightarrow e^{-\theta\tau} \quad \text{sse} \quad Pr(\widehat{M_n} \leq u_n) \rightarrow e^{-\tau}.$$

Quindi, se $0 < \theta \leq 1$ le medesime costanti di normalizzazione determinano sia la distribuzione limite di $\widehat{M_n}$ che quella di M_n e le due distribuzioni sono dello stesso tipo e differiscono per un termine potenza pari a θ. Si noti che se $\theta = 1$ le distribuzioni di $\widehat{M_n}$ e M_n sono identiche.

Se le osservazioni sono stocasticamente indipendenti, allora $\theta = 1$, mentre non vale il viceversa. Esistono, infatti, processi stocastici che, pur avendo valore dell'extremal index unitario, sono costituiti da osservazioni stocasticamente dipendenti.

Condizione di dipendenza $D^I(u_n)$ (Leadbetter, 1983): Oltre alla condizione $D(u_n)$, Leadbetter introduce una seconda condizione detta $D^I(u_n)$:

$$\limsup_{n\to\infty} n \sum_{j=2}^{\left[\frac{n}{k}\right]} Pr\left(X_1 > u_n, X_j > u_n\right) \quad k \rightarrow \infty$$

per ogni sequenza $u_n = \frac{x}{a_n} + b_n$ con $a_n > 0, b_n \in R$ che limita la dipendenza a breve raggio, ovvero la probabilità che, osservata un'eccedenza della soglia u_n, si osservi un'altra eccedenza in un intervallo di tempo ravvicinato. Allora, se si assumono valide sia la condizione $D(u_n)$ che la $D^I(u_n)$ si può dimostrare che la distribuzione del massimo e le costanti di normalizzazione a_n e b_n sono le stesse che si userebbero se le osservazioni fossero indipendenti e identicamente distribuite. Infatti, vale il seguente teorema:

Teorema 3.4: Sia $X_1, \ldots, X_n, \ldots$ una sequenza stazionaria e $\{u_n\}$ una sequenza di costanti del tipo $u_n = \frac{x}{a_n} + b_n$ con $a_n > 0, b_n \in R$ tali che siano soddisfatte sia la $D(u_n)$ che la $D^I(u_n)$ e $0 \leq \tau \leq \infty$ allora

$$Pr(M_n \leq u_n) \rightarrow e^{-\tau} \quad \text{sse} \quad n\, Pr(X_t > u_n) \rightarrow \tau.$$

Inoltre, la condizione $D^I(u_n)$ implica che $\theta = 1$.

Si noti che per verificare la validità della condizione $n\, Pr(X_t > u_n) \rightarrow \tau$ è sufficiente mostrare che:

$$\frac{1-F(x^-)}{1-F(X)} \xrightarrow{x\to\infty} 1,$$

condizione che risulta essere soddisfatta da tutte le distribuzioni dei valori estremi.

3.3.3 Point process of clusters

I cluster di valori estremi si verificano quando non si pone alcuna restrizione alla dipendenza locale delle osservazioni.

In questo caso si può dimostrare che i cluster di valori estremi si verificano con un processo di Poisson.

Siano

- $\{u_n\}$ una sequenza di costanti tali che $F(u_n) \rightarrow 1$ per $n \rightarrow +\infty$;
- $X_1, \ldots, X_n$ una sequenza di osservazioni stazionarie;
- $I(n, j) = \delta(X_j > u_n)$ con $\delta(\cdot)$ la funzione indicatrice
- $N_n = \left\{\frac{j}{n}: X_j > u_n\right\}$ il processo su [0,1] costituito dagli istanti normalizzati a cui avvengono le eccedenze.

Condizione di dipendenza debole $\Delta(u_n)$ (Hsing et al. 1988): Sia $\{u_n\}$ una sequenza

di costanti e $B_i^j(u_n)$ la sigma-algebra generata dagli eventi $(X_s \leq u_s)$ per ogni n, i, j tali che $1 \leq i \leq j \leq n$ e $i \leq s \leq j$, allora per ogni n e $1 \leq l \leq n-1$ sia

$$\alpha_{n,l} = max(|Pr(A \cap B) - Pr(A)Pr(B)|: A \in B_1^k(u_n), B \in B_{k+l}^n(u_n), 1 \leq k \leq n-1)$$

La sequenza $X_1, \ldots, X_n$ soddisfa la condizione $\Delta(u_n)$ se $\alpha_{n,l_n} \to 0$ per $n \to +\infty$ per qualche sequenza l_n tale che $l_n = o(n)$. Le costanti α_{n,l_n} sono chiamate coefficienti della condizione $\Delta(u_n)$. Condizione che è bene notare è più restrittiva della condizione $D(u_n)$ introdotta da Leadbetter.

Siano ora $\{k_n\}$ e $\{r_n\}$ due sequenze tali che $k_n \to +\infty$, $\frac{l_n k_n}{n} \to 0$, $k_n \alpha_{n,l_n} \to 0$ e r_n è la parte intera di $\frac{n}{k_n}$. Si definisca la distribuzione del cluster come:

$$\pi_n(j) = Pr\left(\sum_{i=1}^{r_n} I(n,i) = j \,\middle|\, \sum_{i=1}^{r_n} I(n,i) > 0\right) \quad j \geq 1$$

dove π_n rappresenta la distribuzione di un cluster basata su un blocco di lunghezza r_n.

Teorema 3.8 (Hsing et al. 1988): Sotto queste condizioni e le ulteriori:

- Per ogni $\tau > 0$ esiste $u_n = u_n(\tau)$ tale che $n\,Pr(X_t > u_n) \to \tau$ per $n \to +\infty$ condizione che risulta verificata se $\frac{1-F(x^-)}{1-F(X)} \xrightarrow{x \to \infty} 1$;
- $Pr(M_n \leq u_n) \to e^{-\theta\tau}$ per $n \to +\infty$
- $\pi_n(j) \to \pi(j)$

si dimostra che il processo N_n converge ad un processo di Poisson composto il cui processo dei centri dei cluster è un processo di Poisson di intensità $\theta\tau$ e gli elementi di ciascun cluster hanno distribuzione $\pi(\cdot)$ indipendente per ogni cluster.

In particolare, il processo dei centri dei cluster può essere definito come segue: fissato r_n, k_n e s_n=1, …, k_n, se si verifica un'eccedenza per almeno un j tale che $(s-1)\,r_n \leq j \leq s\,r_n$, allora il cluster può essere rappresentato da un singolo punto (i.e., il centro del cluster) la cui eccedenza avviene al tempro $t = \frac{s\,r_n}{n}$. I punti centrali così costruiti costituiscono il processo dei centri dei cluster che converge al processo di Poisson di intensità $\theta\tau$ sopradescritto.

Sotto ulteriori condizioni di regolarità, infine, si può dimostrare che l'extremal index è il limite del reciproco della dimensione media del cluster, ovvero che $\frac{1}{\theta} = \sum_{j=1}^{\infty} j\pi(j)$.

3.4 Applicazione alle ondate di calore

Sulla base dei risultati sopra menzionati si assuma assegnato un blocco di lunghezza r_n, una soglia u_n, $X_1, \dots, X_n, \dots$ la serie storica delle temperature massime nella città di Gorizia per il periodo 1989-2022 e un'ondata di calore nel periodo $\{1, \dots, r_n\}$.

Siano $N(u_n, r_n)$ e $\pi(i, u_n)$ rispettivamente:

- il numero di giornate per cui si osservano temperature superiori alla soglia:
$$N(u_n, r_n) = \#\{X_i > u_n;\ i = 1, \dots, r_n\}$$
- la distribuzione della durata dell'ondata di calore, ovvero la probabilità che l'ondata di calore abbia una durata pari a i sapendo che nel periodo si è verificata un'ondata di calore (i.e., $N(u_n, r_n) \geq 1$):
$$\pi(i, u_n) = Pr(N(u_n, r_n) = i \mid N(u_n, r_n) \geq 1) \quad i = 1, \dots, r_n$$

Dal teorema 3.8 segue che sotto queste ipotesi e sotto opportune condizioni di regolarità il processo normalizzato dei tempi di eccedenza su [0,1] converge a un processo di Poisson composto, le cui osservazioni hanno distribuzione di probabilità limite pari a $\pi(\cdot)$ e indice estremo θ pari al reciproco della dimensione media di un cluster.

Un modo alternativo per caratterizzare la distribuzione della durata di un'ondata di calore $\pi(\cdot)$ parte dalla seguente definizione, che considera la probabilità che l'ondata di calore abbia una durata pari a i sapendo che l'ondata è iniziata il primo giorno di calendario del periodo considerato e non in un momento arbitrario:

$$\theta(i, u_n) = Pr(N(u_n, r_n) = i \mid X_1 > u_n) \quad i = 1, \dots, r_n$$

si definisce quindi

$$\theta(i) = \lim_{n \to +\infty} \theta(i, u_n) \quad i = 1, 2, ..$$

e

$$\pi(i) = \frac{\theta(i) - \theta(i+1)}{\theta(1)} \quad i = 1, 2, ..$$

Il vantaggio di questo approccio è che, differentemente dal precedente, la valutazione non richiede di tener conto di tutte le possibili storie antecedenti l'inizio dell'ondata di calore, che qui coincide con l'indice $t = 1$.

Oltre alle quantità sopra menzionate, nel contesto delle ondate di calore è interessante studiare anche la distribuzione delle durate consecutive, ovvero la probabilità che l'ondata di calore inizi al tempo 1 e perduri per i giorni consecutivi, in simboli:

$$\theta_C(i, u_n) = Pr(C(i, u_n, r_n) = i \,|X_1 > u_n) \quad i = 1, \dots, r_n$$

dove

$$C(i, u_n, r_n) = \{X_1 > u_n, \dots, X_i > u_n, X_{i+1} \leq u_n \cap \nexists\, t = 2, \dots, r_n:$$
$$X_t > u_n, \dots, X_{t+i-1} > u_n\}.$$

L'evento $C(i, u_n, r_n)$ richiede che l'ondata di calore abbia una durata consecutiva pari a i e che non ci siano altri momenti all'interno del periodo considerato in cui si verifica un'ondata di calore di quelle dimensioni. Da cui seguono

$$\theta_C(i) = \lim_{n \to +\infty} \theta_C(i, u_n) \quad i = 1, 2, ..$$

$$\pi_C(i) = \frac{\theta_C(i) - \theta_C(i+1)}{\theta_C(1)} \qquad i = 1, 2, ..$$

e, di conseguenza, si può definire l'extremal index come:

$$\theta_C = \frac{1}{\sum_{i=1}^{\infty} i\, \pi_C(i)}$$

Si noti che $\theta_C(1)$ e $\theta(1)$ sono entrambe uguali a θ in quanto l'evento di interesse è il medesimo (i.e., $X_2 \leq u_n, .., X_{r_n} \leq u_n | \, X_1 > u_n$) e dalla caratterizzazione di O'Berin (1987) riesce che:

$$\theta = \lim_{n \to \infty} Pr(X_2 \leq u_n, .., X_{r_n} \leq u_n | \, X_1 > u_n).$$

3.5 Metodi per calcolare l'extremal index

Nei precedenti paragrafi si è dimostrato che, sotto opportune condizioni di regolarità, l'extremal index può essere visto come il reciproco della dimensione media di un cluster. Questa caratterizzazione suggerisce che per stimare l'extremal index si possono identificare cluster di valori estremi nei dati e calcolarne il reciproco della dimensione media. Esistono due schemi che si possono seguire per costruire cluster di valori estremi: il raggruppamento a blocchi o il runs de-clustering.

3.5.1 Approccio a blocchi

Si supponga di avere n osservazioni da una serie stazionaria e sia N_n il numero di osservazioni che eccedono una soglia u_n considerata estrema.

Nell'approccio a blocchi, i dati vengono divisi in k_n blocchi, ciascuno di lunghezza pari a r_n in modo che $n \cong r_n k_n$, dove ciascun blocco è trattato come un cluster. Sia Z_n il numero di cluster, cioè il numero di blocchi che contengono almeno un'osservazione che eccede una soglia, allora l'extremal index può essere stimato mediante:

$$\hat{\theta} = \frac{Z_n}{N_n}$$

Per esempio, si assuma di avere venti osservazioni, di cui quelle in grassetto costituiscono eccedenze oltre la soglia u. Allora, applicando il metodo con $k = 4$ e $r = 5$ si ottengono i seguenti quattro cluster:

$$\underbrace{x_1\ x_2\ \boldsymbol{x_3}\ \boldsymbol{x_4}\ \boldsymbol{x_5}}_{cluster\ 1}\ \underbrace{x_6\ x_7\ x_8\ x_9\ \boldsymbol{x_{10}}}_{cluster\ 2}\ \underbrace{\boldsymbol{x_{11}}\ \boldsymbol{x_{12}}\ \boldsymbol{x_{13}}\ \boldsymbol{x_{14}}\ \boldsymbol{x_{15}}}_{cluster\ 3}\ \underbrace{x_{16}\ x_{17}\ x_{18}\ \boldsymbol{x_{19}}\ x_{20}}_{cluster\ 4}$$

Da cui $\hat{\theta} = 0.4$ essendo $Z_n = 4$ e $N_n = 10$.

3.5.2 Runs de-clustering

L'approccio runs de-clustering prevede che un cluster inizia a formarsi quando un'osservazione eccede la soglia e si conclude quando si osservano m osservazioni consecutive che non eccedono la soglia u. Il cluster successivo inizia quando si osserva una nuova eccedenza. Definito $W_{n,i}$ uguale a 1 se l'i-esima osservazione eccede la soglia u_n e zero altrimenti. Allora

$$N_n = \sum_{i=1}^{n} W_{n,i}$$

$$Z_n = \sum_{i=1}^{n-r_n} W_{n,i}\,(1 - W_{n,i+1}) \ldots\ (1 - W_{n,i+r_n})$$

$$\hat{\theta} = \frac{Z_n}{N_n}.$$

La definizione di Z_n assicura che un'eccedenza in posizione i sia contata se e soltanto se le successive sono inferiori alla soglia, ovvero se è il membro più a destra del cluster secondo la definizione.

Per esempio, si assuma di avere venti osservazioni di cui quelle in grassetto costituiscono eccedenze oltre la soglia u, allora applicando il metodo con $m = 3$ si ottengono i seguenti due cluster, mentre le altre osservazioni sono perse:

$$x_1\ x_2\ \underbrace{\boldsymbol{x_3\ x_4\ x_5}\ x_6\ x_7\ x_8}_{\boldsymbol{cluster\ 1}}\ x_9\ \underbrace{\boldsymbol{x_{10}\ x_{11}\ x_{12}\ x_{13}\ x_{14}}\ x_{15}\ x_{16}\ x_{17}}_{\boldsymbol{cluster\ 2}}\ x_{18}\ \underbrace{\boldsymbol{x_{19}}\ x_{20}}_{\boldsymbol{cluster\ 3}}$$

Da cui $\hat{\theta} = 0.3$ essendo $Z_n = 3$ e $N_n = 10$.

I parametri da cui dipendono i due metodi sono di solito scelti arbitrariamente, e diverse scelte possono avere diversi impatti sulle stime dei cluster che si possono ottenere a partire dai campioni.

Oltre ai metodi di clustering, il comportamento di cluster di valori estremi può essere studiato specificando un modello per descrivere la legge temporale del processo.

3.6 Catena Markoviana

Nel seguito si assumerà che le osservazioni $X_1, \ldots, X_n, \ldots$ non solo siano un processo stazionario ma più specificatamente una catena Markoviana del primo ordine a tempo discreto con spazio degli stati continuo.

Sotto queste ipotesi riesce che la densità della variabile aleatoria X_t condizionata a tutta la storia passata dipende solo dallo stato occupato all'istante più recente e in particolare non dipende dagli istanti di valutazione in quanto la distribuzione si mantiene omogenea nel tempo:

$$f(x_t|x_1, \ldots, x_{t-1}) = f(x_t|x_{t-1}).$$

Più in generale la distribuzione congiunta risulta pari a:

$$f(x_1, \ldots, x_n) = f(x_1) \prod_{t=1}^{n-1} f(x_{t+1}|x_t).$$

L'assunzione di una catena Markoviana con le caratteristiche sopra menzionate semplifica notevolmente la modellazione in quanto richiede di modellare soltanto la distribuzione congiunta di (X_t, X_{t+1}). Nota questa distribuzione, risulta assegnata tutta la legge del processo. In aggiunta, le singole variabili $X_1, \ldots, X_n, \ldots$ sono identicamente distribuite.

La distribuzione congiunta (X_t, X_{t+1}) può avere due strutture: dipendenza o indipendenza asintotica a seconda del valore assunto dal seguente limite:

$$\chi_1 = \lim_{x \to \bar{x}} Pr(X_{t+1} > x | X_t > x) \quad \bar{x} = sup\{x : F(x) < 1\}$$

Se $\chi_1 = 0$ le variabili (X_t, X_{t+1}) sono asintoticamente indipendenti mentre, se $\chi_1 > 0$, queste sono asintoticamente dipendenti. Quando la soglia si avvicina al suo estremo superiore, gli eccessi tendono a raggrupparsi localmente se il processo è asintoticamente dipendente e a verificarsi singolarmente altrimenti.

Nel caso di dipendenza asintotica si può dimostrare che la struttura di dipendenza è invariante rispetto alla soglia e di conseguenza anche la distribuzione della durata $\pi(\cdot)$ è indipendente dalla soglia usata per definire il cluster (i.e., l'ondata di calore). Al contrario, per i processi asintoticamente indipendenti, le caratteristiche di dipendenza e quindi la tendenza a formare cluster variano con il livello critico e si indeboliscono con l'aumentare del livello fino a quando, al limite, non vi è alcun raggruppamento (i.e., le osservazioni estreme tendono a manifestarsi individualmente) e le osservazioni tendono a comportarsi come variabili indipendenti e identicamente distribuite.

Di conseguenza, se (X_t, X_{t+1}) sono asintoticamente dipendenti (per $u \to +\infty$) allora $\theta < 1$ e $\pi(1) < 1$; mentre se (X_t, X_{t+1}) sono asintoticamente indipendenti, allora la dimensione dei cluster (i.e., la durata delle ondate di calore) si riduce al crescere di u fino a che le eccedenze avvengono singolarmente e in questo caso risulta che $\theta \cong 1$ e $\pi(j)$ è uguale a 1 se $j = 1$ e vale zero se $j > 1$.

In sostanza, da tali considerazioni emerge che le caratteristiche delle ondate di calore possono cambiare con il livello critico usato per identificare il cluster, ossia un'ondata di calore definita a partire da una soglia u_1 avrà caratteristiche diverse da una definita a partire da una soglia u_2 se il processo mostra indipendenza asintotica. Pertanto, per modellare correttamente il fenomeno occorre introdurre classi di modelli che consentano al grado di clusterizzazione di decrescere all'aumentare della soglia.

3.7 Approccio semi-parametrico

Siano $X_1, \ldots, X_n, \ldots$ una catena markoviana, F la comune funzione di ripartizione delle osservazioni e u una soglia considerata estrema.

La funzione delle code delle osservazioni, come descritto nel paragrafo 3.2.2.1, può essere ricavata dalla distribuzione degli eccessi oltre la soglia sotto l'assunto che la GPD(u, σ_u, ξ) costituisca un modello adeguato per le eccedenze della variabile X. Da ciò risulta che:

$$Pr(X > x) = \left[1 + \xi \frac{x-u}{\sigma_u}\right]^{-\frac{1}{\xi}} \varsigma_u$$

dove $\varsigma_u = 1 - F(u)$. Sulla base di questo risultato si ottiene che:

$$F(x) = \begin{cases} 1 - \left[1 + \xi \frac{x-u}{\sigma_u}\right]^{-\frac{1}{\xi}} \varsigma_u & x \geq u \\ \widetilde{F(x)} & x < u \end{cases}$$

dove $\varsigma_u = 1 - F(u)$ e $\widetilde{F(x)}$ è la funzione di ripartizione empirica delle osservazioni.

Seguendo l'approccio di Keef et al. (2013), e senza perdita di generalità, le variabili del processo possono essere trasformate in variabili con distribuzioni marginali di Laplace come segue:

$$T(X_t) = \begin{cases} \log\big(2\, F(X_t)\big) & X_t < F^{-1}(0.5) \\ -\log(2\,[1 - F(X_t)]) & X_t \geq F^{-1}(0.5). \end{cases}$$

Da cui segue che

$$Pr(T(X_t) \leq t) = \begin{cases} \frac{e^t}{2} & t < 0 \\ 1 - \frac{e^{-t}}{2} & t \geq 0 \end{cases}$$

e $T(X_t) - u_l | T(X_t) > u_l$ ha distribuzione esponenziale con media unitaria e u_l è la soglia espressa nella scala di Laplace.

Sulla base delle quantità sopra definite Heffernan e Tawn (2004) introducono un approccio semi-parametrico valido per modellare sia i casi di dipendenza $(\chi_1 > 0)$ sia di indipendenza asintotica $(\chi_1 = 0)$.

L'obiettivo del metodo è di modellare la distribuzione congiunta di $(T(X_t), T(X_{t+1}))$ usando la distribuzione condizionata di $T(X_{t+1})$ dato che $T(X_t)$ è grande

$$Pr(T(X_{t+1}) \leq T(x_{t+1}) | T(X_t), = T(x_t))$$

e se tale distribuzione è non degenere per $x \to \bar{x}$, dove $\bar{x} = sup\{x: F(x) < 1\}$, allora vale la seguente condizione di convergenza:

Condizione di convergenza: assegnate le funzioni $a(\cdot)$: $R_+ \to R$, $b(\cdot)$: $R_+ \to R_+$ per $x > 0$ vale che:

$$Pr\left(\frac{T(X_{t+1}) - a\{T(X_t)\}}{b\{T(X_t)\}} \leq z, \frac{X_t - u}{\sigma_u} > x \middle| X_t > u\right) \to G(z)\,[1 + \xi\, x]_+^{-\frac{1}{\xi}}$$

per $u \to \bar{x}$, dove G è una distribuzione non degenere.

La trasformazione delle osservazioni in variabili con distribuzioni marginali di Laplace permette di identificare un'unica classe entro cui scegliere le funzioni $a(\cdot)$ e $b(\cdot)$

$$a(y) = \alpha y$$

$$b(y) = y^{\beta}$$

dove $\alpha \in [-1, 1], \beta \in]-\infty, 1)$. La scelta delle funzioni all'interno della classe sopra descritta non ha impatti sulla struttura di dipendenza e ha il vantaggio di semplificare notevolmente il calcolo. In particolare, se:

- $\alpha = \beta = 0$ e $G(z)$ con distribuzione di Laplace corrisponde al caso di variabili indipendenti;
- $\alpha = 1,\ \beta = 0$ corrisponde al caso di dipendenza asintotica;
- $-1 \leq \alpha \leq 0$ corrisponde al caso di dipendenza negativa;
- $0 < \alpha < 1$ o $\alpha = 0$ e $\beta > 0$ corrisponde al caso di indipendenza asintotica.

L'approccio semi-parametrico stima la forma della struttura di dipendenza come parte della procedura di adattamento. Pertanto, non richiede di scegliere in anticipo la forma della struttura di dipendenza e in aggiunta, consente di tener conto sia della dipendenza che dell'indipendenza asintotica.

Sotto l'assunto che la condizione di convergenza sia soddisfatta per tutti i valori di X_t maggiori della soglia u, è possibile scrivere X_{t+1} dato $X_t > u$ come segue

$$T(X_{t+1}) = \alpha\, T(X_t) + T(X_t)^{\beta}\, Z_t$$

dove Z_t è una variabile aleatoria con distribuzione G indipendente da X_t. In aggiunta, sotto l'assunto che le osservazioni siano una catena markoviana segue che la sequenza $\{Z_t\}$ è formata da variabili indipendenti e identicamente distribuite. Dal momento che la forma funzionale G non è nota, per stimare i parametri α e β si assume in linea con Keef et al. (2013) che $Z_t \sim N(\mu, \sigma^2)$ allora per $y > T(u)$

$$T(X_{t+1})|\, T(X_t) = y \sim N(\alpha y + \mu y^{\beta}, \sigma^2\, y^{2\beta}).$$

I parametri possono quindi essere stimati sulla base del metodo della massima verosimiglianza, la cui funzione di densità è pari a:

$$f(\boldsymbol{x}) = \frac{n}{2} log(2\pi) + \frac{n}{2} log(\sigma^2) + \beta \sum_{i=1}^{n} log(y_i) + \sum_{i=1}^{n} \frac{(x_i - \alpha y_i - \mu y_i^{\beta})^2}{\sigma^2 y_i^{\beta}}.$$

Quindi, la stima non parametrica di G può essere ricavata a partire da

$$\widehat{z_j} = \frac{T(x_{t_j+1}) - \hat{\alpha}\, T(x_{t_j})}{T(x_{t_j})^{\hat{\beta}}} \qquad j = 1, \ldots, n_u$$

dove $\hat{\alpha}, \hat{\beta}$ sono le stime di massima verosimiglianza, $t_1, \ldots, t_{n_u}$ sono gli indici associati alle osservazioni che eccedono la soglia u (i.e., $x_t > u$) e n_u è il numero di osservazioni che eccedono la soglia u. A questo punto le stime di $T(x_{t+j})$ $j = 1, \ldots, h$ si ottengono sostituendo nell'espressione $T(X_{t+1}) = \alpha\, T(X_t) + T(X_t)^{\beta}\, Z_t$ valori simulati dalla distribuzione empirica G e dalla distribuzione di $T(X_t)$ per $T(X_t) > u_l$.

3.8 Approccio non parametrico

L'approccio non parametrico è un caso particolare dell'approccio precedentemente presentato e da esso differisce solamente per la stima della funzione G e per la struttura di dipendenza implicata.

Infatti, partendo dall'approccio semi-parametrico e sotto l'ipotesi che le osservazioni siano asintoticamente dipendenti, cioè $\alpha = 1,\ \beta = 0$, la distribuzione empirica G di Z può essere stimata direttamente dai dati originali mediante la seguente formula:

$$\widehat{z_j} = T(x_{t_j+1}) - T(x_{t_j}) \qquad j = 1, \ldots, n_u$$

3.9 Approccio parametrico

Sia F la funzione di ripartizione congiunta di due variabili consecutive (X_j, X_{j+1}) di una catena Markoviana e F_j la distribuzione marginali di X_j con $j = 1,2$.

Per modellare gli estremi della serie storica, Bortot et al. (1998) e Smith et al. (1997) si concentrano sulla distribuzione congiunta di una coppia consecutiva di osservazioni (X_j, X_{j+1}). La specificazione del modello richiede l'assegnazione delle variabili marginali

F_j per descrivere il comportamento della coda e della struttura di dipendenza, entrambe sulla regione $R = (u, +\infty) \times (u, +\infty)$, dove u è la soglia usata per la modellazione.

La distribuzione marginale F_j è ricavata sotto l'assunto che gli eccessi oltre la soglia u seguano una distribuzione di Pareto generalizzata GPD(u, σ_u, ξ), da cui segue, come descritto nel paragrafo 3.2.2.1, che:

$$F_j(x) = 1 - \left[1 + \xi \frac{x-u}{\sigma_u}\right]^{-\frac{1}{\xi}} \varsigma_u \quad x > u$$

dove $\varsigma_u = 1 - F(u)$.

Per modellare la dipendenza tra le osservazioni si considera la trasformazione delle variabili X_j in

$$Z_j = -\frac{1}{log(F_j(X_j))}$$

così che ciascuna ha distribuzione marginale di Frèchet[13] per $x_1, x_2 > u$. Allora, sotto l'ipotesi che F sia nel dominio di attrazione di una distribuzione multivariata dei valori estremi (che è equivalente a richiedere che la distribuzione congiunta di (Z_1, Z_2) sia nel dominio di attrazione di una distribuzione multivariata dei valori estremi con marginali distribuite come una Frèchet), si può dimostrare che (Ledford et al. 1996, Smith et al. 1997):

$$F(x_1, x_2) = exp(-V(z_1, z_2)) \qquad x_1, x_2 > u$$

dove

$$z_j = -\frac{1}{log\left(1 - \left[1 + \xi \frac{x_j - u}{\sigma_u}\right]^{-\frac{1}{\xi}} \varsigma_u\right)}$$

$$V(z_1, z_2) = \int_0^1 2\, max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right) dH(w)$$

è una funzione omogenea di ordine -1 su R_+^2, $V(z_1, \infty) = {z_1}^{-1}$, $V(\infty, z_2) = {z_2}^{-1}$, $V(1,1)$ è compresa tra $[1,2]$ e il parametro χ_1 è pari a $2 - V(1,1)$.

Invece, $H(\cdot)$ è una distribuzione su [0,1] che soddisfa la seguente condizione:

[13] $Pr(Z_j \leq z) = e^{-\frac{1}{z}}$ per $z > 0$

$$\int_0^1 w\, dH(w) = \frac{1}{2}$$

e la corrispondente probabilità di transizione per $x_t \to \bar{x}$ (dove $\bar{x} = sup\{x: F(x) < 1\}$) è pari a

$$Pr(X_{t+1} \leq x_t + z\,[\sigma_u + \xi\,(x_t - u)]_+ |\, X_t = x_t) \to 2 \int_{\{1+[1+\xi\, z]_+^{\frac{1}{\xi}}\}^{-1}}^1 w\, dH(w).$$

3.9.1 Modello logistico

Un caso particolare del modello sopra descritto si ha quando la funzione $V(\cdot)$ ha la seguente forma funzionale, anche detta dipendenza logistica (Gumbel, 1960) e che nel seguito verrà chiamato "modello parametrico":

$$V_\gamma(z_1, z_2) = \left(z_1^{-\frac{1}{\gamma}} + z_2^{-\frac{1}{\gamma}}\right)^\gamma$$

$$H(w) = \frac{1}{2}\left[\left\{w^{\frac{1-\gamma}{\gamma}} - (1-w)^{\frac{1-\gamma}{\gamma}}\right\}\left\{w^{\frac{1}{\gamma}} + (1-w)^{\frac{1}{\gamma}}\right\}^{\gamma-1} + 1\right]$$

dove $\gamma = 1$ corrisponde al caso di perfetta indipendenza e $\gamma > 0$ a gradi di dipendenza crescenti, $\gamma = 0$ implica perfetta dipendenza e $\chi_1 = 2 - 2^\gamma$.

In generale, questo modello è meno flessibile del modello semi-parametrico perché limita la struttura di dipendenza al caso di dipendenza asintotica o perfetta indipendenza. Infatti, Bortot et al. (1988), dimostrano che per x grande:

$$Pr(X_{t+1} > x | X_t > x) \cong \begin{cases} Pr(X_{t+1} > x) & se\ V(1,1) = 2 \iff \gamma = 1 \\ c > 0 & se\ V(1,1) < 2 \iff \gamma < 1 \end{cases}$$

dove la prima uguaglianza corrisponde esattamente alla definizione di indipendenza stocastica. Si tratta di un modello che è preferibilmente applicato a processi asintoticamente dipendenti in quanto, se questa procedura viene applicata ad un processo asintoticamente indipendente, si ottiene una rappresentazione inadeguata della struttura di dipendenza, con $Pr(X_{t+1} > x | X_t > x)$ approssimata o da un numero $c > 0$, oppure da $Pr(X_{t+1} > x)$ qualunque sia $x > u$. In entrambi i casi c'è un'assunzione di stabilità della struttura di dipendenza rispetto alla soglia che porta ad una rappresentazione non corretta degli eventi sulla coda. Quindi, il modello introdotto copre solo una sottoclasse dei casi di asintotica indipendenza, quelli di perfetta indipendenza.

3.9.2 Copule

La struttura di dipendenza tra variabili aleatorie può essere rappresentata tramite una copula, una funzione di ripartizione congiunta le cui marginali sono uniformi sull'intervallo [0,1]. Infatti, assegnata la funzione di ripartizione congiunta F di (X_1, …, X_d) con distribuzioni marginali $F_1(x),\dots,F_d(x)$, la copula C può essere descritta dalla seguente espressione:

$$F(x_1, \dots, x_d) = C(F_1(x_1),\dots,F_d(x_d)).$$

Un esempio di copula si ha in presenza di osservazioni indipendenti e identicamente distribuite con:

$$C(u_1,\dots,u_d) = \prod_{i=1}^{d} u_i.$$

Nel contesto della teoria multivariata dei valori estremi il modello logistico può essere espresso mediante la seguente copula:

$$C(u_1,\dots,u_d) = exp\left\{-\left[\sum_{i=1}^{d}(-log\ u_i)^{\frac{1}{\gamma}}\right]^{\gamma}\right\}$$

dove γ è il parametro di dipendenza del modello logistico.

3.9.3 Metodo della verosimiglianza censurata

La stima dei parametri del modello logistico è ottenuta con il metodo della verosimiglianza censurata.

Per una soglia elevata u lo spazio viene diviso in quattro quadranti:

- $R_{0,0} = (-\infty,\ u] \times (-\infty,\ u]$
- $R_{1,0} = (\ u,\ +\infty) \times (-\infty,\ u]$
- $R_{0,1} = (-\infty,\ u] \times (u,\ +\infty)$
- $R_{1,1} = (u,\ +\infty) \times\ (u,\ +\infty)$.

Per le osservazioni che non eccedono la soglia u, l'informazione più significativa ai fini della stima è che la soglia non è stata superata, indipendentemente dall'effettivo valore misurato. Quindi, le osservazioni che non eccedono la soglia sono censurate al livello u.

Ne segue che, dato un campione di osservazioni $(x_1, y_1), \ldots, (x_n, y_n)$ la verosimiglianza è pari a:

$$L(\theta; (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{n} \varphi(\theta; (x_i, y_i))$$

dove θ è il vettore di parametri da stimare e

$$\varphi(\theta; (x_i, y_i)) = \begin{cases} F(u, u), & (x_i, y_i) \in R_{0,0} \\ F(x_i, u), & (x_i, y_i) \in R_{1,0} \\ F(u, y_i), & (x_i, y_i) \in R_{0,1} \\ F(x_i, y_i), & (x_i, y_i) \in R_{1,1}. \end{cases}$$

Ottimizzando la funzione rispetto a θ si ottengono le stime dei parametri.

Capitolo 4

4.1 Inferenza sui parametri

Per studiare le ondate di calore nel goriziano – mediante i modelli teorici descritti nel capitolo 3 – e procedere al calcolo delle quantità cardine introdotte nel paragrafo 2.2, è fondamentale procedere con la verifica delle assunzioni alla base del modello, in particolare la validità dell'assunzione di markovianità della catena, la stima dei parametri e l'identificazione della soglia estrema.

4.2 Il processo Markoviano

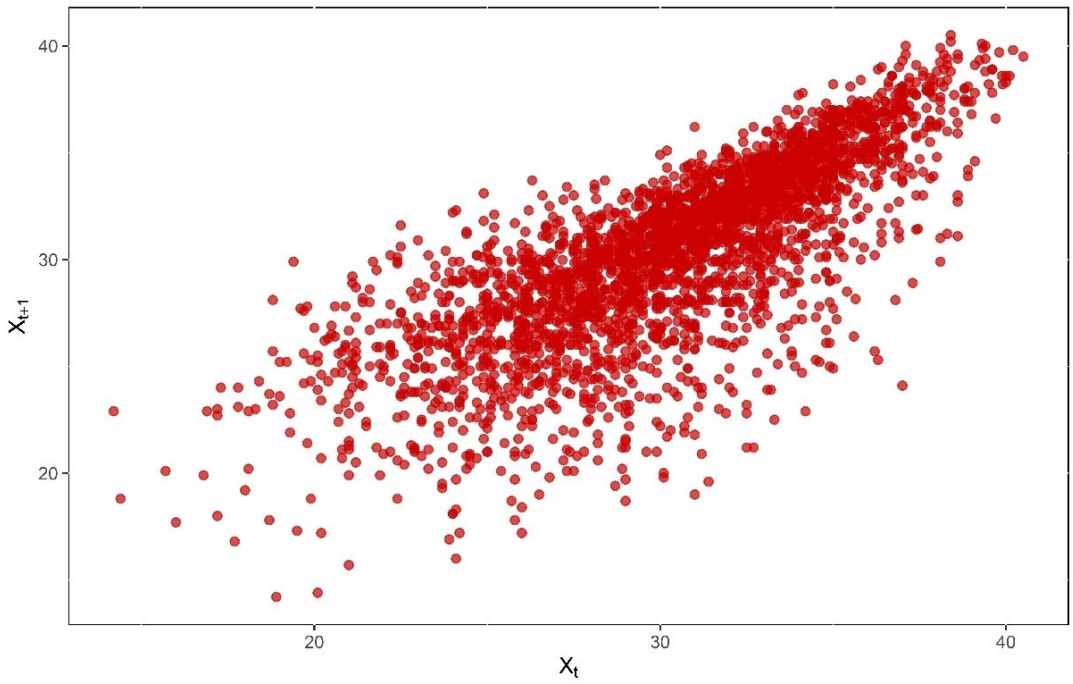
Ciascuno dei modelli presentati nei precedenti capitoli presuppone che i dati, ossia le temperature massime registrate giorno per giorno tra giugno e agosto, seguano un processo di Markov omogeneo del primo ordine con uno spazio degli stati continuo.

Un processo si può definire di Markov del primo ordine se le osservazioni al tempo $t+1$ dipendono solo dall'istante più recente t e non da tutta la storia passata. Per determinare se questa assunzione è soddisfatta dai dati alla base dello studio è possibile condurre alcune analisi grafiche e quantitative, tra cui:

- Grafico a dispersione delle osservazioni (X_t, X_{t+1}): consente di valutare la presenza di correlazione tra le osservazioni con differimento unitario;
- Analisi dell'autocorrelazione (ACF) e dell'autocorrelazione parziale presente nei dati (PACF): queste funzioni misurano il grado di associazione tra i valori della

serie corrente e quelli delle serie passate e consentono di capire quali sono i valori passati più utili per prevedere quelli futuri.

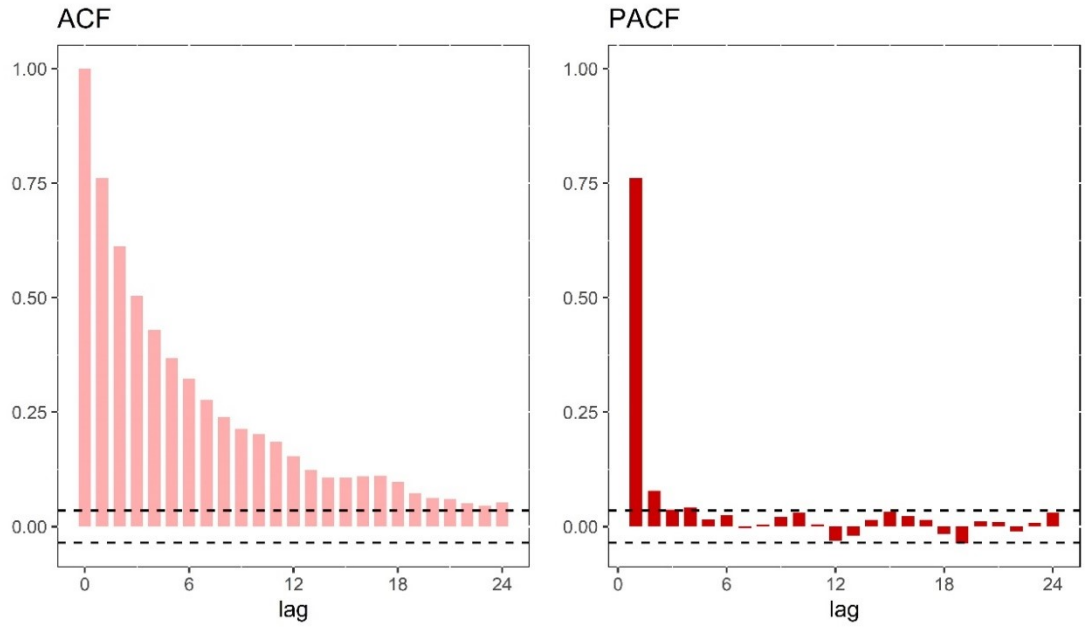
Figura 6: Grafico a dispersione delle osservazioni (X_t, X_{t+1})



Dal grafico a dispersione si osserva che le osservazioni con differimento unitario presentano una correlazione piuttosto elevata (i.e., circa 76%), il che suggerisce la presenza di dipendenza temporale tra (X_t, X_{t+1}). Tuttavia, affinché l'ipotesi che le osservazioni seguano una catena markoviana risulti soddisfatta è necessario che la dipendenza tra le osservazioni con differimento maggiore ad uno decada molto velocemente, per rendere ragionevole l'assunzione che le osservazioni dipendono solo dall'istante più recente e non da tutta la storia passata.

Tale verifica, descritta dall'analisi dell'autocorrelazione e dell'autocorrelazione parziale, come si può osservare in figura 7, evidenzia che la dipendenza temporale tra le osservazioni decade molto rapidamente con l'unica eccezione di quella relativa al differimento unitario. Da questa considerazione, dunque, emerge che i dati possono essere considerati come determinazioni di un processo markoviano del primo ordine e, in altre parole, che l'informazione contenuta in ogni punto è influenzata solamente dall'osservazione precedente – quella più recente – e non da tutti quelli che lo hanno preceduto.

Figura 7: Funzione di autocorrelazione e di autocorrelazione parziale per verificare l'assunzione che le osservazioni siano una catena markoviana.



4.3 Individuazione della soglia

Tipicamente, per la selezione della soglia da utilizzare nei modelli ci si avvale di particolari tecniche, sia quantitative che di osservazione grafica, come ad esempio le tecniche di Mean Residual Life plot e Parameter Stability plot, presentate nel paragrafo 3.2.2.1

Mean Residual Life Plot

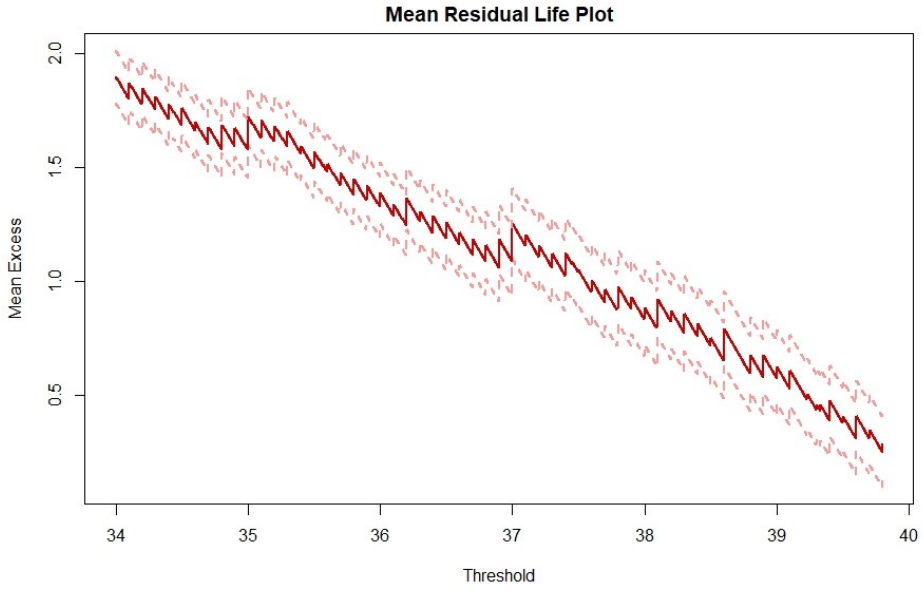
Sotto l'ipotesi che $X_t \sim \text{GPD}(u, \sigma_u, \xi)$, si individua il più piccolo valore u per cui, considerate le k osservazioni che eccedono la soglia, il grafico dello stimatore

$$\widehat{e(u)} = \frac{1}{k}\sum_{j=1}^{k}(x_j - u)$$

è lineare da quel punto in avanti.

Come si osserva dalla figura 8, per i dati del campione in analisi, l'andamento della curva risulta praticamente lineare a partire da una soglia di 34°C e quindi qualsiasi soglia successiva dovrebbe rendere l'assunto che i dati si distribuiscono come una Pareto generalizzata soddisfatto.

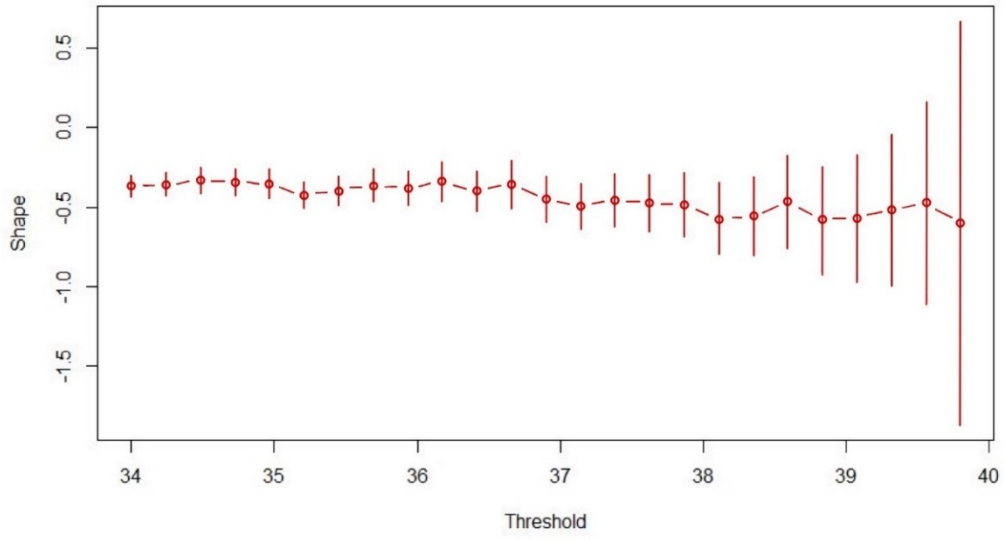
Figura 8: Mean Residual Life Plot ottenuto a partire da una soglia di 34°C, valore a partire dal quale considerate le k osservazioni che eccedono la soglia, il grafico dello stimatore della media risulta approssimativamente lineare.

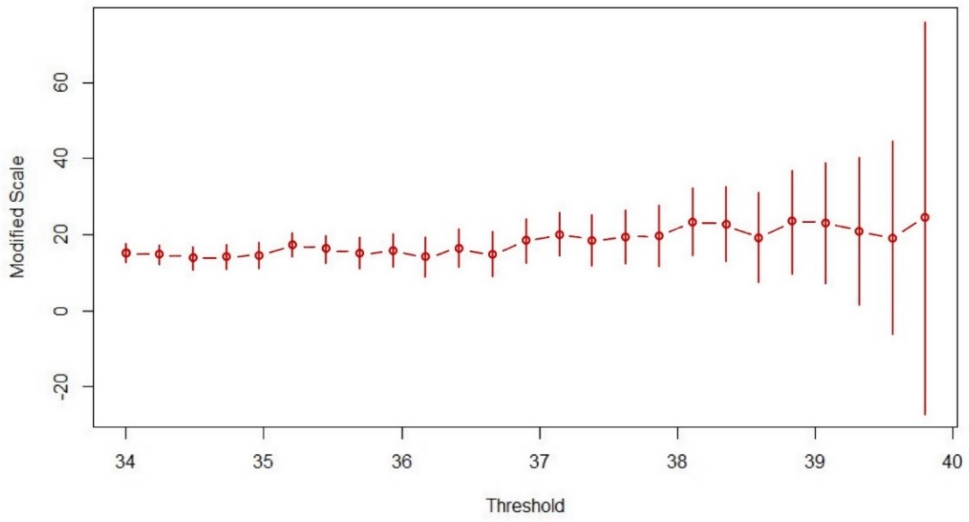


Parameter Stability Plot

Sotto l'ipotesi che $X_t \sim$ GPD(u, σ_u, ξ), e considerata una soglia $v \geq u$, la distribuzione degli eccessi oltre la soglia v risulta una GPD(v, σ_v, ξ) con $\sigma_v = \sigma_u + \xi v$. Dunque, posto $\sigma_{mod} = \sigma_v - \xi v = \sigma_u$, riesce che ξ e σ_{mod} sono costanti $\forall\, v > 0$. Pertanto, la soglia u può essere fissata in modo tale che i parametri modificati siano approssimativamente costanti da lì in avanti.

Figura 9: Parameter Stability plot ottenuto a partire da una soglia di 34°C per il parametro ξ nella prima immagine, e per σ_{mod} nella seconda.





In figura 9, si osserva che, per il campione in analisi, tale condizione può ritenersi soddisfatta per valori superiori ai 34°C.

In conclusione, sulla base delle analisi sin qui condotte, la soglia scelta per la modellazione del fenomeno viene fissata a 34°C, valore corrispondente a circa l'80-esimo percentile della distribuzione.

In letteratura, il valore utilizzato come soglia viene selezionato tra il 90-esimo e il 100-esimo percentile della distribuzione. Tuttavia, è opportuno sottolineare che, nel caso in esame, questa viene utilizzata esclusivamente per modellare le variabili marginali, mentre per determinare le proprietà del cluster si è scelto di adottare una soglia pari al 90-esimo percentile (o superiore), corrispondente a circa 35°C. Tale precisazione risulta necessaria in quanto, se l'assunzione che le variabili siano GPD è soddisfatta, allora non ci si aspetta una variazione nelle stime, indipendentemente dall'utilizzo di una soglia meno estrema dal punto di vista della distribuzione, che risulta ad ogni modo estrema per il fenomeno in esame.

4.4 Stima dei parametri

In seguito all'individuazione della soglia u è possibile procedere alla stima dei parametri che saranno usati per la modellazione.

4.4.1 Approccio semi-parametrico

In caso di utilizzo dell'approccio semi-parametrico, si assume che la distribuzione degli eccessi oltre la soglia sia una GPD(u, σ_u, ξ), da cui risulta che la funzione di ripartizione delle osservazioni è:

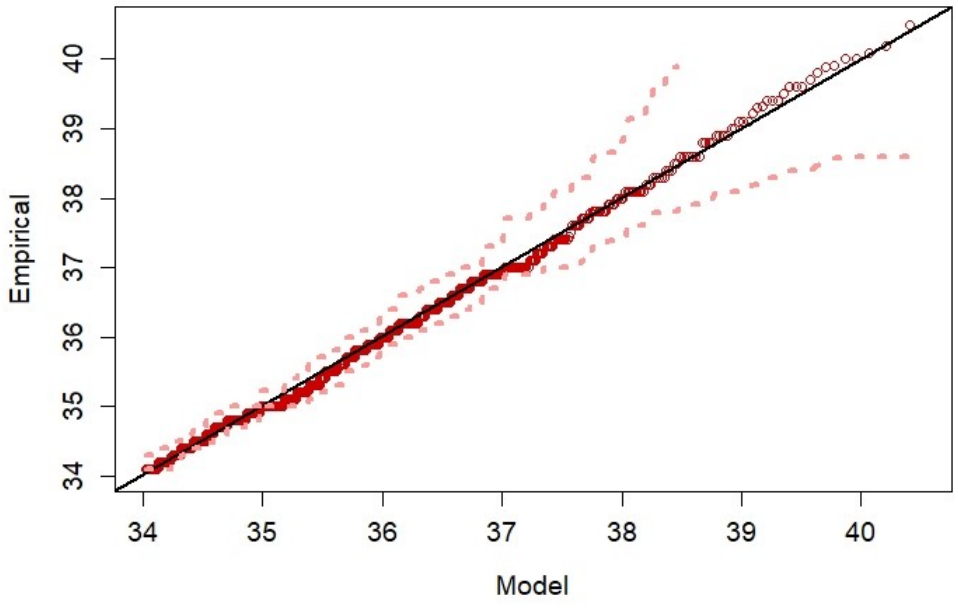
$$F(x) = \begin{cases} 1 - \left[1 + \xi \frac{x-u}{\sigma_u}\right]^{-\frac{1}{\xi}} \varsigma_u & x \geq u \\ \widetilde{F(x)} & x < u \end{cases}$$

dove $\varsigma_u = 1 - F(u)$ e $\widetilde{F(x)}$ è la funzione di ripartizione empirica delle osservazioni. In linea con la metodologia delle copule, i parametri della distribuzione marginale $(\sigma_u, \xi, \varsigma_u)$ sono stimati sotto l'assunto che le osservazioni siano indipendenti mediante il metodo della massima verosimiglianza. Da cui, applicando tale metodo al campione osservato, si ottengono:

- $\hat{\sigma}_u = 2.59$
- $\hat{\xi} = -0.37$
- $\hat{\varsigma}_u = 0.19$

Il QQplot con i parametri stimati è riportato in figura 10.

Figura 10: QQplot della distribuzione delle temperature massime che eccedono la soglia di 34°C, ottenuto sulla base dei parametri stimati con il metodo della massima verosimiglianza (2.59, -0.37, 0.19).



Ottenuti i parametri $(\sigma_u, \xi, \varsigma_u)$, le variabili del processo possono essere trasformate in variabili con distribuzioni marginali di Laplace mediante la seguente trasformazione:

$$T(x_t) = \begin{cases} log\{2\,F(x_t)\} & x_t < F^{-1}(0.5) \\ -log\{2\,[1-F(x_t)]\} & x_t \geq F^{-1}(0.5). \end{cases}$$

Sotto l'assunto che $T(X_{t+1})|\,T(X_t) = y \sim N(\alpha y + \mu y^{\beta},\ \sigma^2\, y^{2\beta})$ per $y > T(u)$, i parametri $(\alpha, \beta, \mu, \sigma^2)$ possono essere stimati sulla base del metodo della massima verosimiglianza. I parametri μ, σ^2 vengono scartati mentre:

- $\hat{\alpha} = 0.80$
- $\hat{\beta} = 0.53.$

Al fine di verificare che i parametri stimati (α, β) siano significativamente diversi da $(1,0)$ è possibile costruire un test basato sul log-rapporto di verosimiglianza confrontando la log verosimiglianza del modello completo, di parametri $(\alpha, \beta, \mu, \sigma^2)$, con il modello ridotto in cui (α, β) sono vincolati rispettivamente a $(1,0)$.

La statistica test $\lambda = -2(log(L(1,0,\mu,\sigma^2)) - log(L(\alpha,\beta,\mu,\sigma^2)) \sim \chi_2^2$ risulta pari a 457.96, il cui p-value è zero. Pertanto, i parametri sono significativamente diversi da $(1,0)$ che si avrebbero se i dati mostrassero asintotica dipendenza.

A questo punto, la stima empirica della funzione di ripartizione di Z può essere ricavata invertendo la seguente espressione

$$\widehat{z_j} = \frac{T(x_{t_j+1}) - \hat{\alpha}\,T(x_{t_j})}{T(x_{t_j})^{\hat{\beta}}} \qquad j = 1, \ldots, n_u$$

dove $\hat{\alpha}, \hat{\beta}$ sono le stime di massima verosimiglianza, $t_1, \ldots, t_{n_u}$ sono gli indici associati alle osservazioni che eccedono la soglia u (i.e., $x_t > u$) e n_u è il numero di osservazioni che eccedono la soglia u.

Nel capitolo successivo, al fine di simulare ondate di calore con certe caratteristiche, sarà necessario seguire un approccio simulativo di tipo backward, ovvero sotto l'ipotesi che $T(X_t)|\,T(X_{t+1}) = y \sim N(\alpha y + \mu y^{\beta},\ \sigma^2\, y^{2\beta})$. Sempre sulla base del metodo della massima verosimiglianza si ottengono:

- $\hat{\alpha}_B = 0.84$
- $\hat{\beta}_B = 0.50.$

4.4.2 Approccio non parametrico

Sulla base delle medesime variabili $T(x_t)$ ottenute nell'approccio semi-parametrico e sotto l'ipotesi che le osservazioni siano asintoticamente dipendenti, cioè $\alpha = 1,\ \beta = 0$, la distribuzione di Z può essere stimata direttamente dai dati originali mediante la seguente formula:

$$\widehat{z_j} = T(x_{t_j+1}) - T(x_{t_j}) \qquad j = 1, \dots, n_u.$$

I parametri di tipo backward si possono semplicemente stimare come

$$\hat{z}_j^B = T(x_{t_j}) - T(x_{t_j+1}) \qquad j = 1, \dots, n_u.$$

4.4.3 Approccio parametrico

I parametri nell'approccio parametrico sono stimati sulla base del metodo della verosimiglianza censurata presentato nel capitolo 3.9 da cui si ottiene:

- $\hat{\gamma} = 0.41$.

Il parametro che misura la dipendenza asintotica del modello χ_1 è positivo e pari a 0.67. Ciò dipende dal fatto che l'unica forma di dipendenza consentita dal modello parametrico è l'asintotica dipendenza, nonostante l'approccio semi-parametrico suggerisca la presenza di indipendenza asintotica.

Capitolo 5

5.1 Simulare le ondate di calore

Per studiare le caratteristiche delle ondate di calore e ottenere stime delle probabilità di accadimento degli eventi esaminati nel capitolo 2 si è scelto di adottare un approccio basato sulla simulazione Montecarlo, che consente di simulare diversi scenari climatici per valutare l'incidenza dei diversi eventi.

Per semplicità di lettura, viene qui di seguito riepilogata la notazione del capitolo 2 e i modelli presentati nel capitolo 3. Siano:

- v la soglia usata per definire l'ondata di calore, con $v \geq u$, dove u è la soglia usata per la modellazione delle variabili marginali;
- η una soglia di temperatura considerata ancora più estrema di v (i.e., $\eta > v$);
- N il numero di giorni che eccedono la soglia v;
- N_C il numero di giorni consecutivi che eccedono la soglia v;
- M la temperatura massima osservata durante l'ondata di calore;
- r_n la lunghezza del blocco (come definita nel paragrafo 3.4) per studiare le caratteristiche dell'ondata di calore.

Considerando un periodo r_n sufficientemente grande, le distribuzioni delle durate $\pi(i) = Pr(N = i \,|N \geq 1)$ e $\pi_C(i) = Pr(N_C = i|\, N_C \geq 1)$ per $i = 1,2,\ldots$ e i relativi extremal index θ e θ_C si possono derivare mediante i seguenti algoritmi (a seconda dell'approccio selezionato).

Algoritmo 1: Approccio Semi-parametrico

L'algoritmo richiede come input i parametri $\hat{\alpha}, \hat{\beta}$, la distribuzione non parametrica di Z e la soglia v trasformata su scala di Laplace (nel seguito chiamata v_l).

Sotto l'ipotesi che l'ondata di calore inizi il primo giorno di calendario del periodo r_n considerato (i.e. $X_1 > v_l$), si simula una determinazione di $T(X_1) - v_l | T(X_1) > v_l$ da una distribuzione esponenziale di media unitaria e a questo valore simulato $T(x_1) - v_l$ si aggiunge la soglia, così da ottenere la simulazione di $T(x_1)$. A questo punto le temperature di ogni giorno di calendario successivo al primo $T(x_{t+1})$ per $t = 2, \ldots, r_n$ si ottengono mediante la regola di aggiornamento $T(x_{t+1}) = \hat{\alpha}\, T(x_t) + T(x_t)^{\hat{\beta}}\, \hat{z}_t$. In sintesi, l'algoritmo è così descritto:

Input: $\hat{\alpha}, \hat{\beta}$, distribuzione di Z
Step 1: si simula $T(x_1) \sim Esp(1) + v_l$
Step 2: Per t in $1: r_n - 1$:
si campiona $\hat{z}_t \sim Z$
si pone $T(x_{t+1}) = \hat{\alpha}\, T(x_t) + T(x_t)^{\hat{\beta}}\, \hat{z}_t$
Output: $T(x_1)$, …, $T(x_{r_n})$ con struttura di dipendenza $\hat{\alpha}, \hat{\beta}$.

In via teorica, le giustificazioni asintotiche a supporto dell'algoritmo risultano verificate solo quando $X_t > u$. Per osservazioni $X_t < u$, l'algoritmo fornisce comunque approssimazioni ragionevoli a meno che $X_t \ll u$. In questi casi la probabilità che le temperature ritornino a superare la soglia u che rende valida l'approssimazione è trascurabile. Questa situazione emerge solo quando le osservazioni, che sono state trasformate su scala di Laplace, assumono valori negativi. Infatti, tali valori corrispondono a determinazioni al di sotto della media e quindi al di fuori della regione in cui si considera valida l'approssimazione. Pertanto, i valori successivi ad un'osservazione negativa sono impostati a zero e non influiscono sulle proprietà del cluster (Winter et al. 2016).

Algoritmo 2: Approccio Non-parametrico

L'algoritmo per l'approccio non parametrico funziona in maniera analoga al precedente, ma in questo caso i parametri (α, β) sono vincolati rispettivamente a $(1,0)$. Quindi:

Input: distribuzione di Z

Step 1: si simula $T(x_1) \sim Esp(1) + v_l$

Step 2: Per t in 1: $r_n - 1$:

si campiona $\hat{z}_t \sim \hat{G}$

si pone $T(x_{t+1}) = T(x_t) + \hat{z}_t \;\; t = 1, \ldots, r_n$

Output: $T(x_1)$, …, $T(x_{r_n})$ con struttura di dipendenza (1,0).

Algoritmo 3: Approccio Parametrico

Il modello parametrico segue uno schema simile a quelli presentati negli algoritmi 1 e 2 dove, al posto dei parametri di dipendenza (α, β) e della distribuzione di Z, si usano il parametro γ e la distribuzione uniforme su (0,1). Quindi:

Input: γ, distribuzione uniforme su (0,1)

Step 1: si simula $x_t \sim \text{GPD}(u, \sigma_v, \xi) + \text{v}$

Step 2: Per t in 1: $r_n - 1$:

si simula $U_t \sim Unif(0,1)$

si calcola $\hat{z}_t = -\frac{\gamma}{\sigma_v} log(U_t^{\frac{1}{\gamma-1}} - 1)$

si pone $x_{t+1} = x_t + \hat{z}_t \, [\sigma_v + \xi(x_t - u)_+] \; t = 1, \ldots, r_n$

Output: x_1, …, x_{r_n} con struttura di dipendenza γ.

A questo punto, le proprietà delle ondate di calore si possono ricavare dalle simulazioni ottenute. Per esempio, la distribuzione delle durate $\pi(i)$ e quella delle durate consecutive $\pi_C(i)$ per $i = 1, 2, ..$ si stimano misurando la durata delle ondate di calore in ogni simulazione, mentre gli extremal index θ e θ_C si calcolano come reciproco della durata media dell'ondata.

Per assicurare che le distribuzioni di probabilità $\pi(\cdot)$ e $\pi_C(\cdot)$ non assumano valori negativi, i tre algoritmi presentati sono combinati con una tecnica detta "Pool adjacent violators algorithm", che assicura che le probabilità ottenute siano sempre positive. Maggiori dettagli sul funzionamento dell'algoritmo sono riportati in appendice.

Per valutare le probabilità che un'ondata di calore abbia una durata di n giorni sapendo

che nel periodo la temperatura ha raggiunto un picco pari a η, cioè $Pr(N = n|\, M = \eta)$ e $Pr(N_C = n|\, M = \eta)$, è necessario adottare un algoritmo misto di tipo forward e backward.

Algoritmo 4: Approccio Semi-parametrico e Non parametrico

L'algoritmo richiede come input i parametri $\hat{\alpha}, \hat{\beta}$ e i parametri backward $\hat{\alpha}_B, \hat{\beta}_B$, che sono stati ricavati nel paragrafo 4.3, la distribuzione non parametrica di Z e la distribuzione non parametrica di Z^B backward, la soglia v_l e il massimo η trasformato su scala di Laplace η_l.

Input: $(\hat{\alpha}, \hat{\beta}, \hat{\alpha}_B, \hat{\beta}_B)$, (Z, Z^B), v_l, η_l

Step 1: Si assume che il picco η_l sia osservato al tempo zero $\eta_l = T(x_0)$

Step 2: Per t in 0: $r_n - 1$:

si campiona $\hat{z}_t \sim Z$

si pone $T(x_{t+1}) = \hat{\alpha}\, T(x_t) + T(x_t)^{\hat{\beta}}\, \hat{z}_t$

se $T(x_{t+1}) > T(x_0)$, l'osservazione non viene considerata e si riparte dallo step 2.

Step 3: Per t in 0: $r_n - 1$:

si campiona $\hat{z}_{-t} \sim Z^B$

si pone $T(x_{-t-1}) = \hat{\alpha}_B\, T(x_{-t}) + T(x_{-t})^{\hat{\beta}_B}\, \hat{z}^B_{-t}$

se $T(x_{-t-1}) > T(x_0)$, l'osservazione non viene considerata e si riparte dallo step 2.

Output: $T(x_{-r_n}), \ldots, T(x_0)$, …, $T(x_{r_n})$ con massimo pari a η_l e struttura di dipendenza $(\hat{\alpha}, \hat{\beta}, \hat{\alpha}_B, \hat{\beta}_B)$.

Per l'approccio non parametrico l'algoritmo è il medesimo con l'aggiunta del vincolo $(\hat{\alpha}, \hat{\beta}, \hat{\alpha}_B, \hat{\beta}_B)$ pari a $(1,0,1,0)$.

Infine, per l'approccio parametrico l'algoritmo può essere adatto come segue:

Input: γ, distribuzione uniforme su $(0,1)$

Step 1: Si assume che il picco η sia osservato al tempo zero $\eta = x_0$

Step 2: Per t in 0: $r_n - 1$:

si simula $U_t \sim Unif(0,1)$

si calcola $\hat{z}_t = -\frac{\gamma}{\sigma_v} log(U_t^{\frac{1}{\gamma-1}} - 1)$

si pone $x_{t+1} = x_t + \hat{z}_t [\sigma_v + \xi(x_t - u)_+]$ $t = 1, \ldots, r_n$

se $x_{t+1} > x_0$, l'osservazione non viene considerata e si riparte dallo step 2.

Step 3: Per t in $0: r_n - 1$:

si simula $U_{-t} \sim Unif(0,1)$

si calcola $\hat{z}_{-t} = -\frac{\gamma}{\sigma_v} log(U_{-t}^{\frac{1}{\gamma-1}} - 1)$

si pone $x_{-t-1} = x_{-t} + \hat{z}_{-t} [\sigma_v + \xi(x_{-t} - u)_+]$ $t = 1, \ldots, r_n$

se $x_{-t-1} > x_0$, l'osservazione non viene considerata e si riparte dallo step 2.

Output: $x_{-r_n}, \ldots, x_0, \ldots, x_{r_n}$ con massimo pari a η e struttura di dipendenza γ.

Sulla base degli algoritmi sopra descritti, la probabilità che un'ondata di calore abbia una durata di n giorni sapendo che la massima temperatura è stata almeno pari a η, cioè $Pr(N = n | M \geq \eta)$ e $Pr(N_C = n | M \geq \eta)$, descritta dal seguente integrale:

$$Pr(N = n | M \geq \eta) = \int_{\eta}^{\infty} Pr(N = n | M = s) \; dF_M(\eta, s)$$

può essere approssimata tramite simulazione Montecarlo: si simulano osservazioni dalla distribuzione del massimo del cluster che può essere approssimata da GPD(η, σ_η, ξ) e in ciascuna si valuta la $Pr(N = n | M = \eta)$. La media sulle simulazioni fornisce un'approssimazione della suddetta probabilità. Sostituendo nell'integrale N_C al posto di N, si ottiene la medesima probabilità per le eccedenze consecutive.

Le proprietà over-cluster possono essere calcolate sotto l'assunto, introdotto nel paragrafo 3.3.3, che le ondate di calore (i.e., i cluster di valori estremi) seguano un processo di Poisson di intensità $\theta\tau$. Allora nel periodo n_T il numero medio di ondate di calore si può approssimare con $\rho = \theta \, n_T \, \varsigma_u$ e con $\rho_C = \theta_C n_T \, \varsigma_u$. Equivalentemente, considerata una soglia $v \geq u$ il numero medio di ondate di calore nel periodo n_T è dato da

$$\rho_v = \theta_v \, n_T \, \varsigma_u \left[1 + \xi \, \frac{v - u}{\sigma_u}\right]_+^{-\frac{1}{\xi}}$$

e da

$$\rho_{C,v} = \theta_{C,v}\, n_T\, \varsigma_u \left[1 + \xi\, \frac{v-u}{\sigma_u}\right]_+^{-\frac{1}{\xi}}$$

dove θ_v e $\theta_{C,v}$ sono gli extremal index calcolati rispetto alla soglia v.

In ultimo, la probabilità di osservare almeno un'ondata di calore nella stagione con certe caratteristiche si può ottenere mediante la seguente formula:

$$1 - exp(-\rho_v\, \Upsilon)$$

dove $\Upsilon = Pr(N \geq k | N \geq 1)$, cioè la probabilità che un'ondata di calore duri k giorni, oppure $\Upsilon = P(N \geq k, M \geq \eta | M > v)$, ovvero la probabilità che un'ondata di calore duri k giorni e abbiano una temperatura massima almeno pari a η. Sostituendo N_C al posto di N, si ottengono le medesime probabilità per le eccedenze consecutive.

5.2 Analisi dei risultati

Come descritto nel paragrafo 2.2, in letteratura sono utilizzati due principali approcci per l'analisi e la previsione delle ondate di calore: l'approccio "within-cluster", che si focalizza sul singolo evento con lo scopo di studiare il comportamento dell'ondata di calore "tipo" che potrebbe interessare una area geografica, e l'approccio "over-cluster" che considera tutte le possibili ondate di calore che potrebbero colpire la stagione estiva. Nel seguito sono dapprima studiate le caratteristiche della singola ondata di calore e poi ricavate le caratteristiche delle ondate che potrebbero interessare, ad esempio, l'estate del 2023.

5.2.1 L'ondata di calore "tipo"

Per studiare le ondate di calore nel goriziano si è scelto di fissare il valore di soglia estrema a 34.9°C, in modo da identificare l'inizio dell'ondata di calore quando le temperature eccedono i 35°C, temperatura che, se superata per più giorni consecutivi, può avere impatti negativi sulla salute umana (Pascal et al. 2013).

L'extremal index θ e θ_C, così come le distribuzioni di probabilità delle durate $\pi(\cdot)$ e $\pi_C(\cdot)$ sono state calcolate, oltre che con i tre modelli descritti nel capitolo 3, anche mediante il metodo di runs de-clustering, sviluppato in Visual Basic for Application all'interno del software Microsoft Excel, così da avere un termine di confronto empirico con cui

comparare i risultati.

Tale metodo richiede di identificare il numero minimo m di osservazioni consecutive che non eccedono i 34.9°C e, dopo un'attenta analisi dei dati disponibili, si è deciso di fissare il parametro pari a 4. In questo modo, se per quattro giornate consecutive le temperature massime sono inferiori ai 34.9°C, l'ondata di calore si considera conclusa.

La scelta di tale parametro è arbitraria e dipende dal particolare dataset studiato (Smith et al., 1994), per esempio, Winter et al. (2016) fissano il parametro pari a 3, mentre Coles (2001) utilizza una soglia di 2 o 4. Considerando il particolare campione in esame, una soglia di 4 permette di rappresentare meglio due ondate di calore più durature, che altrimenti sarebbero state divise in due cluster, consentendo una più accurata rappresentazione del fenomeno in esame. In aggiunta, una scelta diversa, ad esempio 3, avrebbe prodotto delle differenze trascurabili sulla distribuzione empirica dell'ondata di calore.

Sulla base del metodo sopra descritto si ottiene che, nel periodo tra il 1989 e il 2022, la città di Gorizia è stata colpita da 84 ondate di calore (con temperature superiori ai 34.9°C) di lunghezza ed intensità variabile. Nella tabella 2, è riportata la distribuzione delle durate empiriche.

Tabella 3: Ondate di calore nella città di Gorizia: Metodo Runs De-clustering ($m = 4$)

Durata	Frequenza Assoluta	Frequenza Relativa
1	20	23.8%
2	16	19.0%
3	8	9.5%
4	7	8.3%
5	10	11.9%
6	3	3.6%
7	2	2.4%
8	3	3.6%
9	3	3.6%
10	3	3.6%
11	3	3.6%
12	1	1.2%
13	0	0.0%
14	0	0.0%
15	0	0.0%
16	1	1.2%
17	0	0.0%
18	3	3.6%
19	0	0.0%
20	0	0.0%

Durata	Frequenza Assoluta	Frequenza Relativa
21	1	1.2%
Totale	**84**	**100%**

Complessivamente, l'ondata di calore più lunga si è verificata nel 2003, durata dal 3 al 24 agosto, periodo durante il quale le temperature massime giornaliere hanno ecceduto la soglia di 35°C per quasi 21 giorni consecutivi. Si tratta dell'unica estate in cui le temperature massime hanno superato per due volte i 40°C (5 agosto e 11 agosto). La temperatura media massima osservata durante l'ondata è stata di 37.5°C e non è mai scesa sotto i 35.6°C tranne nella giornata del 15 agosto quando la temperatura ha raggiunto un valore di 34°C. Nel medesimo anno si è registrata un'ondata di calore anche nel mese di giugno, la più lunga mai verificatasi in questo mese, durata 11 giorni consecutivi dal 7 al 17 giugno, con temperature che hanno raggiunto un picco di 39.4°C e una temperatura media massima di 36.8°C.

Altre ondate di calore particolarmente lunghe si sono verificate anche nell'estate del 1994, dal 24 luglio al 10 agosto (16 giorni) con temperature massime fino a 39.1°C e durante il 2006, 2013 e 2022. Queste ultime sono durate tutte 18 giorni avvenute rispettivamente nei periodi dal 10 al 31 luglio, dal 20 luglio al 12 agosto e dal 15 luglio al 6 agosto. Le temperature medie massime durante l'ondata sono state inferiori a quelle del 2003 ma comunque molto elevate e rispettivamente pari a 36.3°C nel 1994 e nel 2006, 36.2°C nel 2013 e 36.7°C nel 2022.

Durante l'ondata di calore del 2022 si è verificato un incendio particolarmente intenso e duraturo sul Carso Goriziano, Sloveno e Triestino, causando complessivamente la perdita di più di 3700 ettari di bosco. In aggiunta, sempre durante l'ondata di calore del 2022 le temperature hanno superato il massimo storico di 40.2°C (raggiunto nel corso dell'estate del 2015, 21 luglio 2015) arrivando a 40.5°C nella giornata del 22 luglio.

Vale la pena notare che, seppure non ugualmente protratte nel tempo, ci sono state altre tre ondate di calore degne di nota durante le quali le temperature medie massime sono state superiori a quelle sperimentate nel 2003. In particolare, esse si sono verificate rispettivamente nelle estati del 2007, 2010 e 2015, durate rispettivamente 5, 3 e 10 giorni consecutivi, e hanno raggiunto temperature massime medie rispettivamente di 37.8°C, 37.7°C e 38°C.

5.2.1.1 Extremal Index θ e θ_C

Le quantità cardine introdotte nel paragrafo 2.2 sono ora calcolate sulla base dei tre modelli presentati nel capitolo 3. Il periodo r_n, in linea con Winter et al (2016), è stato fissato pari a 40.

Nella tabella che segue sono riportate le stime dell'extremal index per tutti gli approcci considerati, compresi i risultati ottenuti con il metodo di runs de-clustering, per valori crescenti della soglia al fine di studiare più approfonditamente la tendenza delle osservazioni a eccedere la soglia in maniera ravvicinata.

Se l'assunzione che i dati siano asintoticamente indipendenti è soddisfatta, l'extremal index dovrebbe mostrare un andamento crescente al crescere della soglia e al limite dovrebbe risultare prossimo a uno.

Tabella 4: Extremal Index θ per livelli crescenti della soglia

Soglie	Livelli di ritorno	Extremal Index			
		Metodo Runs	Semi Parametrico	Non Parametrico	Parametrico
34.4	1-in-6	0.19	0.20	0.25	0.24
34.9	1-in-8	0.21	0.21	0.25	0.24
35.5	1-in-10	0.24	0.22	0.25	0.24
36.3	1-in-15	0.29	0.24	0.24	0.24
36.8	1-in-20	0.31	0.25	0.24	0.24
37.1	1-in-25	0.34	0.26	0.23	0.24
37.4	1-in-30	0.34	0.27	0.23	0.24
37.7	1-in-40	0.41	0.28	0.23	0.24
38.0	1-in-50	0.42	0.29	0.24	0.24
38.2	1-in-60	0.42	0.30	0.23	0.24
38.6	1-in-92	0.53	0.32	0.23	0.24

La stima empirica dell'extremal index θ mostra un andamento crescente al crescere del livello di ritorno. Tra i vari approcci considerati, l'unico che riesce a cogliere questo andamento è il metodo semi-parametrico, al contrario degli altri, che hanno un andamento approssimativamente costante e per valori della soglia elevati sottostimano notevolmente il valore del parametro.

Un comportamento analogo si osserva anche per l'extremal index associato a eccedenze consecutive θ_C.

Tabella 5: Extremal Index θ_C per livelli crescenti della soglia

Soglie	Livelli di ritorno	Extremal Index			
		Metodo Runs	Semi Parametrico	Non Parametrico	Parametrico
34.4	6	0.30	0.25	0.29	0.28

Soglie	Livelli di ritorno	Extremal Index			
		Metodo Runs	Semi Parametrico	Non Parametrico	Parametrico
34.9	8	0.30	0.26	0.28	0.29
35.5	10	0.35	0.28	0.28	0.28
36.3	15	0.37	0.30	0.28	0.28
36.8	20	0.40	0.32	0.28	0.28
37.1	25	0.40	0.33	0.28	0.28
37.4	30	0.38	0.34	0.27	0.28
37.7	40	0.44	0.35	0.27	0.28
38.0	50	0.45	0.36	0.28	0.28
38.2	60	0.46	0.37	0.27	0.29
38.6	92	0.63	0.39	0.27	0.28

Analogamente al caso precedente, l'unico metodo che riesce a cogliere l'andamento crescente è quello semi-parametrico.

Questo comportamento si origina in quanto, in situazioni di asintotica indipendenza come nel caso in esame, al crescere della soglia le eccedenze tendono a manifestarsi individualmente e al limite le osservazioni si comportano come variabili indipendenti, ovvero non mostrano alcuna tendenza a clusterizzare.

5.2.1.2 Distribuzione delle durate $\pi(i)$ e $\pi_C(i)$

Nella tabella che segue sono riportate le distribuzioni di probabilità delle durate non consecutive (i.e., $\pi(i) = Pr(N = i \,|N \geq 1)$ per $i = 1,2,\ldots$) per una durata fino a 21 giorni di calendario dove le probabilità assumono valori non trascurabili.

Tabella 6: Distribuzione di probabilità delle durate non consecutive $\pi(\cdot)$

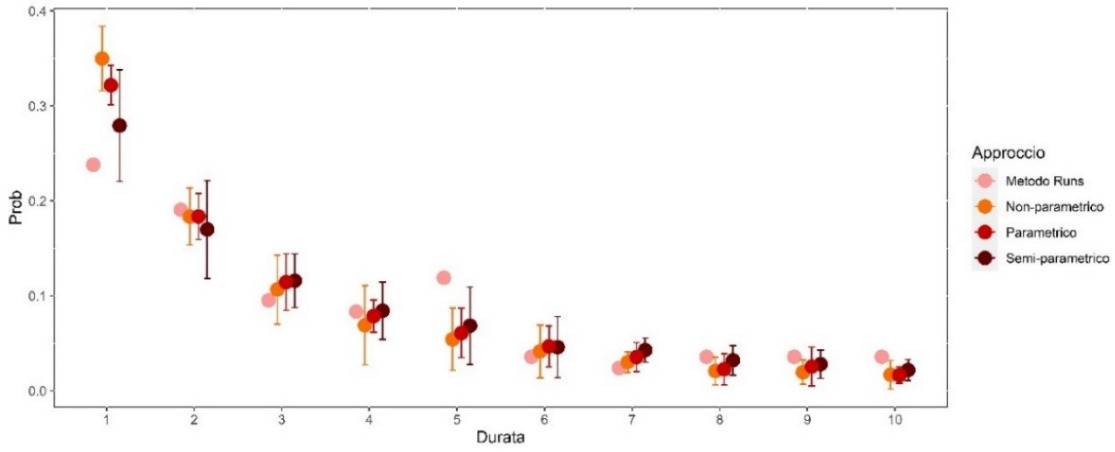
Durata	Distribuzione delle Durate			
	Metodo Runs	Semi Parametrico	Non Parametrico	Parametrico
1	23.8%	27.9%	35.0%	32.2%
2	19.0%	17.0%	18.3%	18.4%
3	9.5%	11.6%	10.7%	11.4%
4	8.3%	8.4%	6.9%	7.9%
5	11.9%	6.9%	5.4%	6.1%
6	3.6%	4.6%	4.2%	4.7%
7	2.4%	4.3%	3.0%	3.6%
8	3.6%	3.2%	2.1%	2.3%
9	3.6%	2.8%	2.0%	2.5%
10	3.6%	2.2%	1.7%	1.6%
11	3.6%	1.8%	1.1%	1.5%
12	1.2%	1.6%	1.2%	1.1%
13	0.0%	1.1%	1.0%	1.0%
14	0.0%	1.0%	0.8%	0.8%
15	0.0%	0.9%	0.8%	0.8%
16	1.2%	0.8%	0.7%	0.7%
17	0.0%	0.5%	0.6%	0.4%
18	3.6%	0.5%	0.3%	0.3%
19	0.0%	0.6%	0.3%	0.4%

Durata	Distribuzione delle Durate			
	Metodo Runs	Semi Parametrico	Non Parametrico	Parametrico
20	0.0%	0.3%	0.3%	0.3%
21	1.2%	0.3%	0.4%	0.3%

L'analisi della tabella evidenzia che, in tutti e tre i metodi considerati, i valori delle probabilità tendono a diminuire esponenzialmente all'aumentare della durata. Sebbene vi sia una certa variabilità nei risultati per $\pi(1)$ e $\pi(5)$, dove il metodo semi-parametrico è quello che più si avvicina al comportamento dei dati, le stime delle probabilità sono simili tra loro e risultano vicine a quelle empiriche, con differenze assolute inferiori al 2.5% ad eccezione di $\pi(1)$ e $\pi(5)$.

Anche dal confronto degli intervalli di confidenza, si nota che tutti e tre i metodi descrivono bene il comportamento dei dati; infatti, la maggior parte dei punti rientra nei relativi intervalli di confidenza. Le maggiori differenze sono osservabili sulla coda per durate maggiori o uguali a 10, dove il metodo empirico basandosi su un numero limitato di punti fornisce stime più incerte.

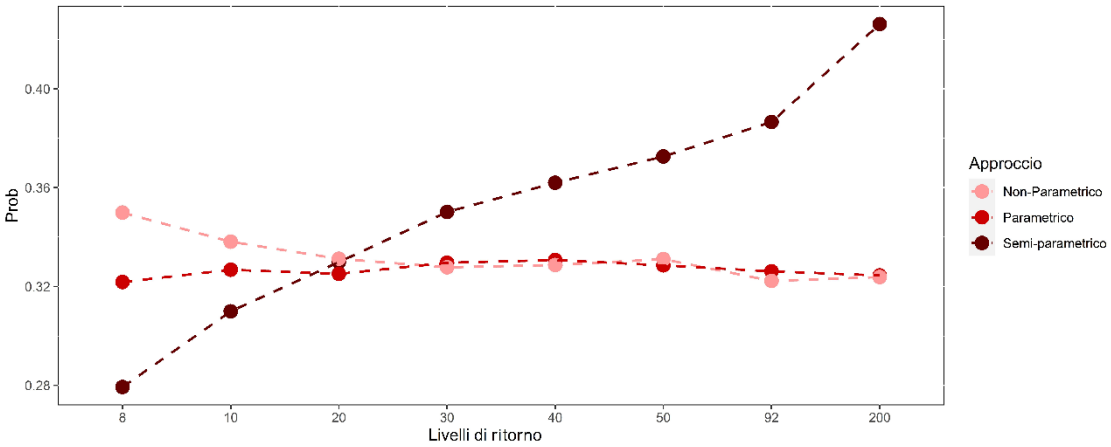
Figura 11: Intervalli di confidenza al 95% e stima della distribuzione delle durate non consecutive $\pi(\cdot)$ fino ad una durata di 10 giorni (per semplicità di visualizzazione). I punti in rosa rappresentano le stime delle probabilità per le diverse durate ottenute con il metodo di runs de-clustering; in arancione, rosso e marrone le medesime probabilità ottenute con i metodi non parametrico, parametrico e semi-parametrico.



La struttura di dipendenza implicata dai tre metodi può essere analizzata osservando il comportamento della probabilità $\pi(1)$ al crescere della soglia. Come si osserva dalla figura, tolta una certa variabilità inziale, all'aumentare della soglia la probabilità rimane asintoticamente costante nel modello parametrico e non parametrico, mentre è crescente nel metodo semi-parametrico.

Ciò accade perché l'approccio semi-parametrico consente l'interazione tra la distribuzione della durata e la soglia usata per definire l'ondate di calore: al crescere della soglia la struttura di dipendenza può cambiare e, in particolare, si indebolisce via via fino tanto che, al limite (i.e., $u \to \infty$), le eccedenze si verificano singolarmente. Al contrario, nell'approccio parametrico e quello non parametrico la struttura di dipendenza rimane costante al variare della soglia così come la tendenza delle osservazioni a manifestarsi in gruppo, e di conseguenza, anche la probabilità che una sola osservazione ecceda la soglia considerata.

Figura 12: Andamento di $\pi(1)$ al variare del livello di ritorno usato per definire l'ondata di calore: in rosa sono riportate le stime di $\pi(1)$ ottenuta sulla base del metodo non-parametrico per diversi livelli di ritorno (i.e., 8, 10, 20, 30, 40, 50, 92 e 200 giorni) in rosso e in marrone le medesime probabilità per il metodo parametrico e semi-parametrico.



Nella tabella che segue è riportata la distribuzione di probabilità delle durate consecutive (i.e., $\pi_C(i) = Pr(N_C = i | N_C \geq 1)$ per $i = 1,2,\ldots$). La distribuzione empirica è ottenuta sempre con il metodo di runs de-clustering, ma il parametro m è fissato pari a 0, così che un'ondata di calore si forma solo quando si verificano eccedenze consecutive della soglia.

Tabella 7: Distribuzione di probabilità delle durate consecutive $\pi_C(\cdot)$

Durata	Distribuzione delle Durate			
	Metodo Runs	Semi Parametrico	Non Parametrico	Parametrico
1	36.1%	32.8%	39.7%	37.4%
2	19.3%	18.9%	18.8%	19.8%
3	10.9%	12.5%	11.2%	11.3%
4	7.6%	8.3%	6.8%	7.9%
5	6.7%	6.3%	4.6%	5.3%
6	3.4%	4.4%	3.7%	4.0%
7	3.4%	3.7%	2.5%	3.0%
8	2.5%	2.6%	2.0%	2.0%
9	2.5%	2.1%	1.9%	1.6%
10	3.4%	1.7%	1.0%	1.5%

Durata	Distribuzione delle Durate			
	Metodo Runs	Semi Parametrico	Non Parametrico	Parametrico
11	3.4%	1.2%	1.0%	1.3%
12	0.8%	1.1%	0.8%	0.8%
13	0.0%	0.8%	0.9%	0.7%
14	0.0%	0.7%	0.6%	0.5%
15	0.0%	0.6%	0.7%	0.6%
16	0.0%	0.5%	0.3%	0.4%
17	0.0%	0.2%	0.4%	0.3%
18	0.0%	0.3%	0.3%	0.2%
19	0.0%	0.2%	0.1%	0.2%
20	0.0%	0.2%	0.3%	0.3%
21	0.0%	0.1%	0.3%	0.2%

Rispetto alla tabella precedente, il valore di $\pi_C(1)$ è diverso, e ciò dipende dalla diversa definizione degli eventi usati per calcolare le quantità. Infatti, come evidenziato nel paragrafo 3.4 pur di considerare r_n sufficientemente ampio la distribuzione di probabilità delle durate non consecutive si ricava come

$$\pi(i) = \frac{\theta(i) - \theta(i+1)}{\theta(1)}$$

dove $\theta(i) = Pr(N(u_n, r_n) = i \,|X_1 > u_n) \;\; i = 1, \ldots, r_n$. Invece, nel caso delle durate consecutive essa è pari a

$$\pi_C(i) = \frac{\theta_C(i) - \theta_C(i+1)}{\theta_C(1)}$$

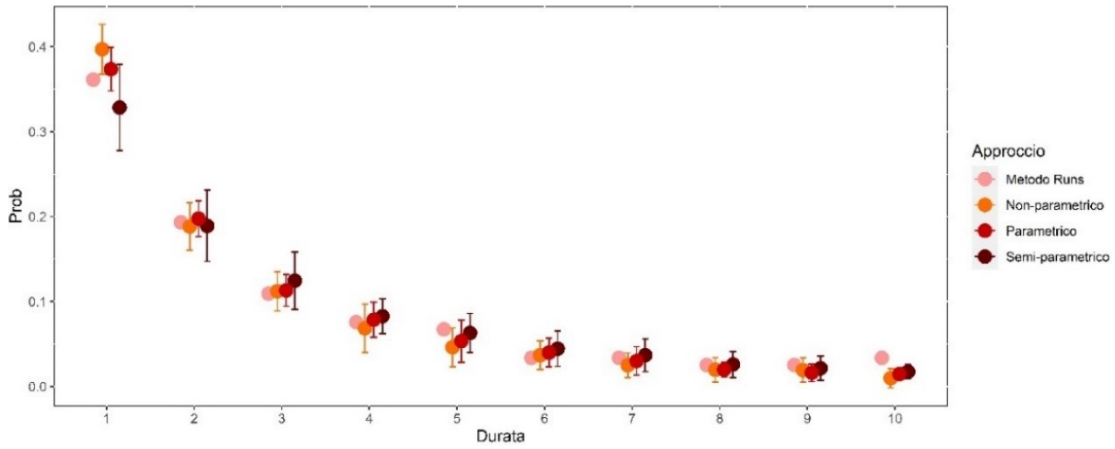
ma $\theta_C(i) = Pr(C(i, u_n, r_n) = i \,|X_1 > u_n) \;\; i = 1, \ldots, r_n$, dove l'evento $C(i, u_n, r_n)$ può essere descritto dalla seguente proposizione logica: l'ondata di calore dura i giorni consecutivi e non ci sono altri momenti all'interno del periodo r_n in cui si verifica un'ondata di calore di tali dimensioni. Quindi, anche se in linea teorica $\pi(1)$ e $\pi_C(1)$ definiscono lo stesso evento e quindi dovrebbero assumere lo stesso valore, la diversa definizione di $\theta(2)$ e $\theta_C(2)$ causa gli scostamenti nelle determinazioni sopra riportate.

Ritornando all'analisi della tabella, in tutti e tre i metodi considerati, i valori delle probabilità tendono a diminuire esponenzialmente all'aumentare della durata. Eccetto $\pi_C(1)$, in tutti i casi le differenze assolute rispetto alle osservazioni empiriche sono inferiori al 2.5%.

Dall'analisi degli intervalli di confidenza in figura 13, si osserva che le stime empiriche sono contenute nei rispettivi intervalli di confidenza tranne che sulla coda dove si osserva

una maggiore variabilità tra i risultati teorici e quelli empirici dovuta al ridotto numero di punti su cui questi ultimi sono basati.

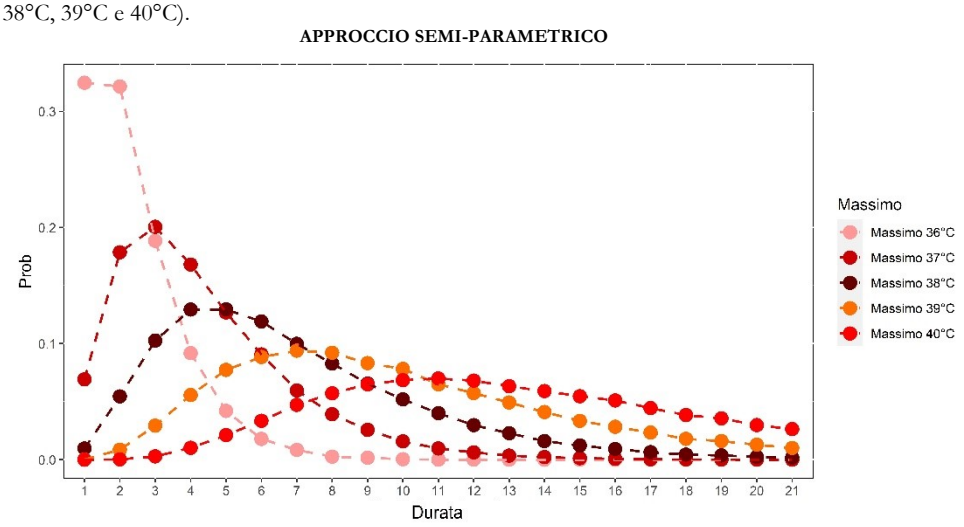
Figura 13: Intervalli di confidenza al 95% e stima della distribuzione delle durate consecutive $\pi_C(\cdot)$ fino ad una durata di 10 giorni (per semplicità di visualizzazione). I punti in rosa rappresentano le stime delle probabilità per le diverse durate ottenute con il metodo di runs de-clustering; in arancione, rosso e marrone le medesime probabilità ottenute con i metodi non parametrico, parametrico e semi-parametrico.

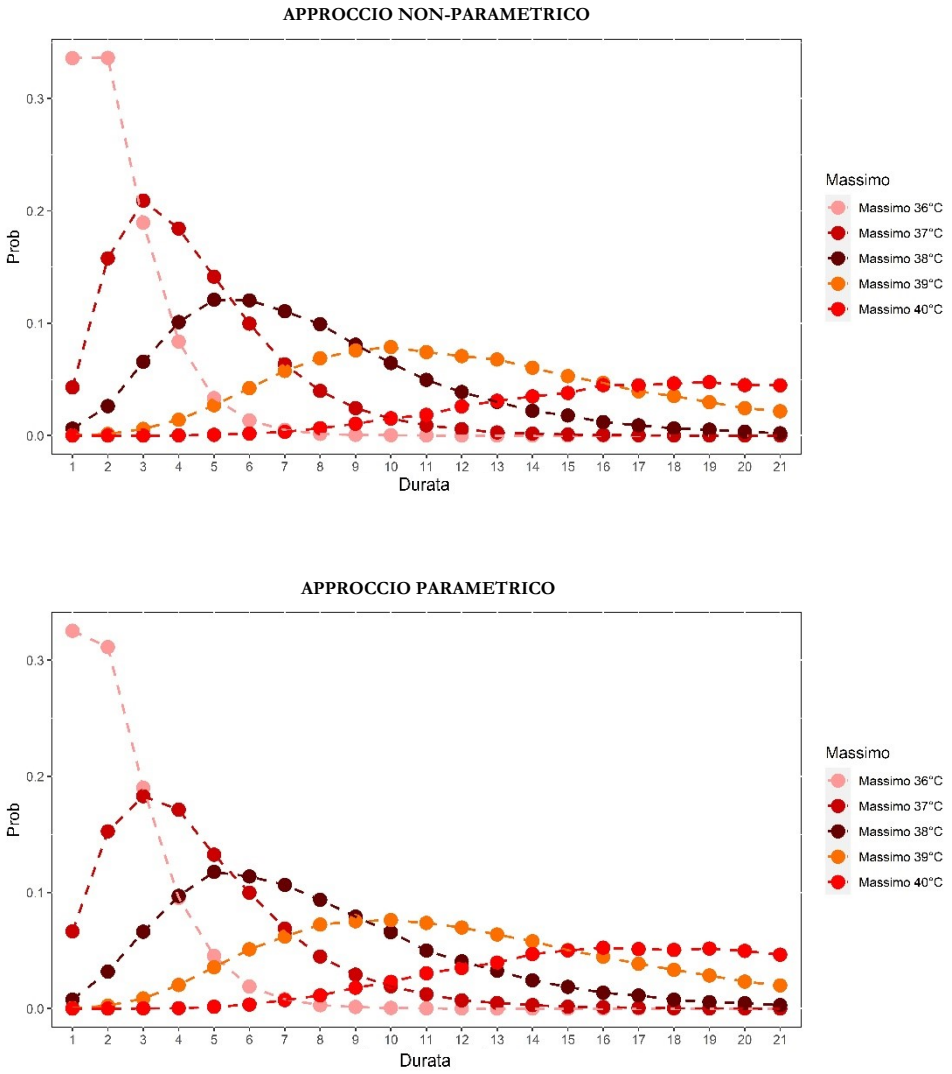


5.2.1.3 Distribuzione delle durate condizionate a η

Nelle figure che seguono sono presentate le distribuzioni di probabilità delle durate condizionate all'informazione che la temperatura massima del periodo ha raggiunto un picco pari a η, dove η è posto pari a 36°C, 37°C, 38°C, 39°C e 40°C rispettivamente.

Figura 14: Distribuzione delle durate non consecutive condizionata all'informazione che il massimo è stato pari a η, fino ad una durata di 21 giorni (per semplicità di visualizzazione) per l'approccio semi-parametrico, non parametrico e parametrico rispettivamente. In ogni immagine è riportata la $Pr(N = n|\, M = \eta)$ per diversi valori di η (i.e., 36°C, 37°C, 38°C, 39°C e 40°C).





In tutti i modelli, si osserva un'asimmetria positiva più pronunciata nella distribuzione condizionata della durata al picco quando il picco è vicino alla soglia usata per definire l'ondata di calore, mentre questa tende a diminuire man mano che il valore massimo considerato aumenta (Winter et al 2016).

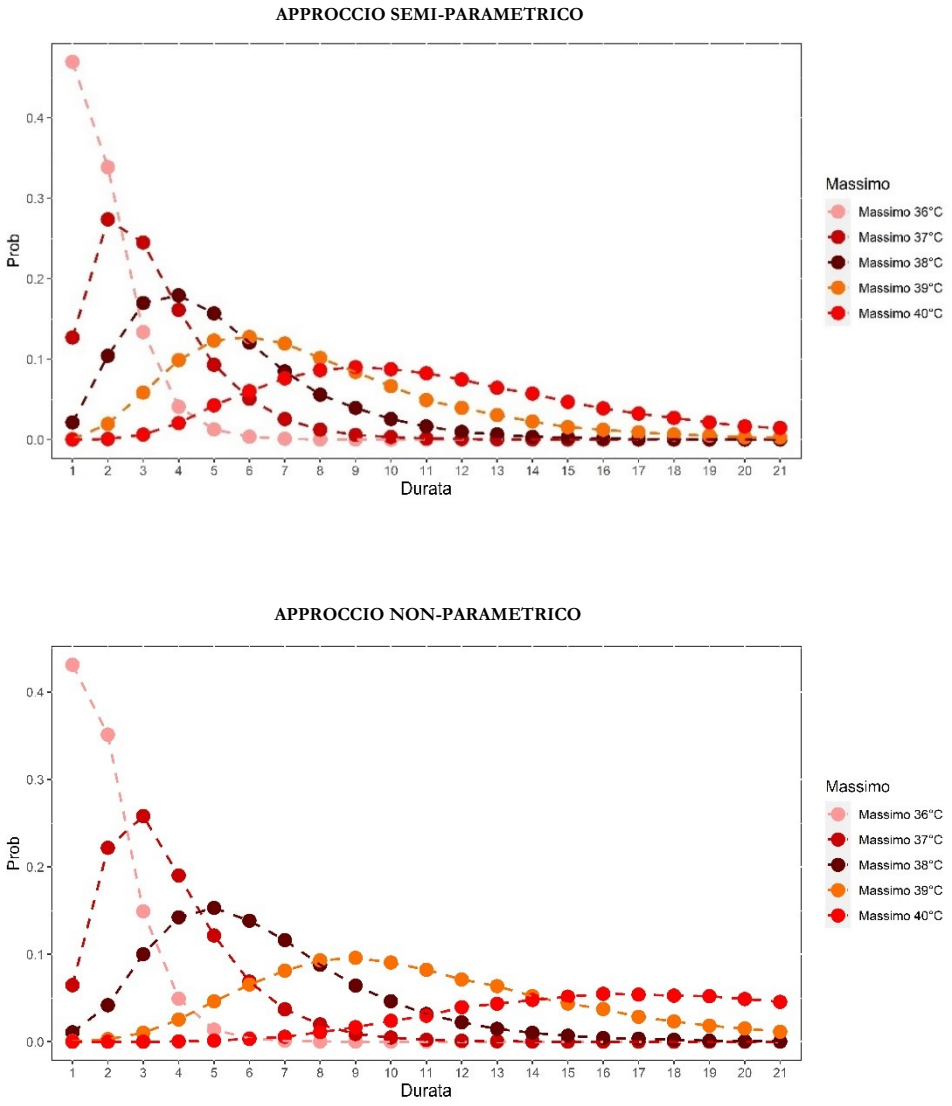
Ciò è dovuto al fatto che, durante un'ondata di calore le temperature tendono ad aumentare fino al picco per poi diminuire e non eccedere più la soglia, secondo uno schema a campana. Per questa stessa ragione, se la temperatura picco è prossima alla soglia, è ragionevole aspettarsi che la durata dell'ondata di calore sia limitata nel tempo.

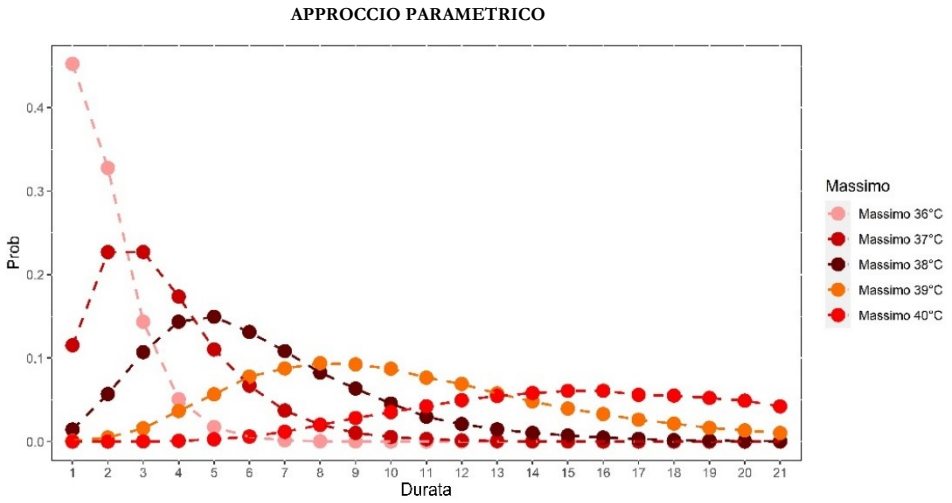
Osservando le distribuzioni delle durate di eventi condizionate ai massimi, si può notare che, in generale, l'approccio semi-parametrico è quello le cui code tendono a zero più rapidamente. In particolare, le differenze maggiori si vedono per temperature massime tra

i 39°C e 40°C, dove gli altri due approcci assegnano una probabilità maggiore di sperimentare eventi anche molto lunghi, mentre per picchi di 36°C, 37°C e 38°C le distribuzioni selezionate con i vari approcci forniscono risultati più simili.

Come emerge dalle figure sotto riportate, un comportamento analogo si osserva per la distribuzione condizionate al massimo delle eccedenze consecutive. Infatti, anche in questo caso i maggiori scostamenti sono visibili per temperature massime tra i 39°C e 40°C, valori per cui il metodo parametrico e non parametrico mostrano code più pronunciate.

Figura 15: Distribuzione delle durate consecutive condizionata all'informazione che il massimo è stato pari a η, fino ad una durata di 21 giorni (per semplicità di visualizzazione) per l'approccio semi-parametrico, non parametrico e parametrico rispettivamente. In ogni immagine è riportata la $Pr(N_C = n | M = \eta)$ per diversi valori di η (i.e., 36°C, 37°C, 38°C, 39°C e 40°C).





5.2.1.4 Distribuzione delle durate condizionate maggiori di η

Le probabilità che un'ondata di calore duri n giorni e abbia un massimo pari o superiore a η, $Pr(N = n|\, M \geq \eta)$ e $Pr(N_C = n|\, M \geq \eta)$, si possono ricavare mediante l'approccio Montecarlo come descritto nel paragrafo 5.2.

Per esempio, per l'approccio semi parametrico, la probabilità che l'ondata di calore duri almeno 21 giorni sapendo che il massimo è stato pari o superiore a 40°C risulta pari a 24.9% (sd: 0.02).

Anche in questo caso il risultato è in linea con le aspettative in quanto, empiricamente risulta che picchi di temperature così estreme si sono verificatiti nel contesto di quattro ondate rispettivamente del 2003, 2013, 2015 e 2022 e di queste solo una quella del 2003 è durata 21 giorni (corrispondente ad una probabilità condizionata ad un massimo di 40°C del 25%).

5.2.2 Le ondate lungo la stagione estiva

Una volta descritte le caratteristiche di un'ondata di calore "tipo" si può procedere a prevedere le caratteristiche delle ondate di calore che potrebbero interessare una stagione estiva, quale, ad esempio, quella del 2023.

Tutti e tre i metodi considerati concordano che il numero di ondate di calore che è ragionevole aspettarsi è tra 2.5 e 3, numero in linea con le aspettative considerato che

mediamente nel periodo considerato (i.e., 1989-2022) la città è stata colpita da 2.5 ondate di calore ogni estate.

Se si aumenta la soglia usata per definire l'ondata di calore, ad esempio, a 35.6°C si ottiene che il numero atteso di ondate di calore è attorno 1.3 mentre, con una soglia di 37.4°C è tra 0.7 e 0.8 a seconda del modello considerato.

In particolare, in tabella 8 e tabella 9 sono riportate le probabilità che si verifichi almeno un'ondata di calore le cui temperature superano i 35°C per un fissato numero di giorni, la prima, e le cui temperature eccedono consecutivamente la soglia dei 35°C, la seconda.

Tabella 8: Probabilità che si verifichi almeno un'ondata di calore con una fissata durata

Giorni	Semi-parametrico		Non Parametrico		Parametrico	
	Prob.	Sd	Prob.	Sd	Prob.	Sd
5	58.6%	0.001	56.3%	0.005	57.3%	0.004
8	38.8%	0.001	36.3%	0.007	36.4%	0.004
11	24.7%	0.004	23.5%	0.004	22.7%	0.003
18	8.0%	0.001	7.8%	0.003	7.4%	0.002
21	5.0%	0.001	4.3%	0.002	4.5%	0.002

Tabella 9: Probabilità che si verifichi almeno un'ondata di calore con una fissata durata consecutiva

Giorni	Semi-parametrico		Non Parametrico		Parametrico	
	Prob.	Sd	Prob.	Sd	Prob.	Sd
5	57.5%	0.004	54.1%	0.004	55.2%	0.004
8	34.1%	0.002	33.0%	0.003	32.3%	0.003
11	18.9%	0.002	20.4%	0.004	18.6%	0.004
18	4.5%	0.002	6.4%	0.003	5.5%	0.001
21	2.5%	0.001	3.8%	0.002	3.0%	0.001

I tre modelli forniscono risultati tra loro simili e in linea con le aspettative, considerate le numerose ondate di calore che hanno colpito la città.

Se si aumenta la soglia usata per definire l'ondata di calore, la probabilità che si verifichino eventi lunghi è molto minore nell'approccio semi-parametrico rispetto agli altri metodi, in quanto all'aumentare della soglia gli eventi estremi tendono a verificarsi in modo sempre meno raggruppato. Questo si nota, per esempio, per le eccedenze consecutive già a partire da una soglia di 35°C, dove gli eventi prolungati hanno una minor probabilità di accadimento, ma non per le eccedenze non consecutive (tabella 8). Tale considerazione dipende dal fatto che il risultato sopra menzionato è asintotico e in questo caso la soglia considerata non è così estrema da rendere valida l'approssimazione, ma se, ad esempio, si considera la probabilità di avere un'ondata di calore di almeno 21 giorni con una soglia di

35.6°C (i.e., il 95-esimo percentile della distribuzione), allora l'andamento sopra descritto risulta soddisfatto anche per le probabilità non consecutive.

Tabella 10: Probabilità che si verifichi almeno un'ondata di calore a partire da una soglia di 35.6°C

Giorni	Semi-parametrico		Non Parametrico		Parametrico	
	Prob.	Sd	Prob.	Sd	Prob.	Sd
21 gg non consecutivi	1.4%	0.0003	2.3%	0.0005	2.0%	0.0005
21 gg consecutivi	0.6%	0.0003	1.8%	0.0010	1.4%	0.0003

Questo comportamento viene poi amplificato al crescere della soglia, per esempio per una soglia di 37.4°C (i.e., il 97-esimo percentile della distribuzione) le medesime probabilità risultano di:

Tabella 11: Probabilità che si verifichi almeno un'ondata di calore a partire da una soglia di 37.4°C

Giorni	Semi-parametrico		Non Parametrico		Parametrico	
	Prob.	Sd	Prob.	Sd	Prob.	Sd
21 gg non consecutivi	0.6%	0.0002	1.4%	0.0006	1.2%	0.0003
21 gg consecutivi	0.2%	0.0002	1.0%	0.0004	0.8%	0.0001

Fino ad ora non è stato posto nessun vincolo circa l'intensità del fenomeno. Tuttavia, per studiare la probabilità che si verifichi un'ondata di calore almeno altrettanto lunga ed intensa quanto quella del 2003 è necessario calcolare la probabilità congiunta che l'ondata di calore abbia una durata almeno pari a n e che il massimo sia almeno pari a η, ovvero la probabilità $P(N \geq k,\ M \geq \eta | M > v)$. Quest'ultima può essere ricavata dalla seguente scomposizione:

$$P(N \geq k \mid M \geq \eta)\ Pr(M \geq \eta \mid M > v)$$

dove la prima è ottenuta come descritto nel paragrafo 5.2.1.4 mentre la seconda è ottenuta a partire da una GPD(v, σ_v, ξ).

Concentrandosi solo sull'approccio semi-parametrico, la probabilità che, in un'estate, si verifichi almeno un'ondata di calore della durata di 21 giorni con temperature superiori ai 35°C e un picco superiore ai 40°C è dello 0.7% (sd: 0.0005).

Si tratta di un risultato in linea con le attese, in quanto, empiricamente nel corso dei 34 anni considerati, nonostante le temperature estremamente elevate solo in 4 estati sono state superate le temperature di 40°C rispettivamente nel 2003, 2013, 2015 e 2022 e tra queste solo quella del 2003 presenta le caratteristiche analizzate (i.e., durata di 21 giorni e un picco superiore ai 40°C), ossia un'osservazione su 84 che, come mostrato nella tabella

3, corrisponde ad una probabilità dell'1.2%.

5.3 Conclusioni ed estensioni

Le analisi condotte nel paragrafo precedente mostrano che per le soglie considerate i risultati forniti dai tre approcci sono simili e vicini a quelli empirici. Tuttavia, l'approccio semi-parametrico è l'unico in cui la distribuzione delle durate può variare all'aumentare della soglia usata per definire le ondate di calore. In questo caso, le caratteristiche di dipendenza e la tendenza a formare ondate di calore variano con il livello critico e, in particolare, si indeboliscono all'aumentare di esso fino a quando, al limite, non vi è alcun raggruppamento (i.e., le osservazioni estreme tendono a manifestarsi individualmente) e le osservazioni tendono a comportarsi come variabili indipendenti e identicamente distribuite. Al contrario, gli altri approcci, che modellano solo i casi di dipendenza asintotica, mostrano una struttura di dipendenza invariante rispetto alla soglia e, di conseguenza, anche la distribuzione della durata $\pi(\cdot)$ risulta indipendente dalla soglia usata per definire l'ondate di calore. Dunque, studiando il comportamento dei valori per $u \to \infty$, questi metodi forniranno probabilità più elevate di osservare eventi lunghi rispetto a quello semi-parametrico.

L'assunto alla base di tutti i modelli considerati è che le osservazioni si possono assimilare ad un processo stazionario. Esso presuppone che le osservazioni future abbiano una distribuzione costante al variare del tempo e, di conseguenza, che anche la soglia v, usata per definire le eccedenze, rimanga stabile nel tempo. Tuttavia, in un contesto di cambiamenti climatici come quello attuale, questa ipotesi potrebbe essere problematica in quanto dopo un certo numero di anni si potrebbero osservare più eccedenze di quelle attese a causa dell'aumento delle temperature non pareggiato da un pari aumento della soglia.

Per questa ragione è bene tenere presente che i risultati sopra riportati e le stime ottenute si possono considerare affidabili se guardati per un orizzonte temporale limitato, in cui si può assumere che gli effetti dei cambiamenti climatici sulla distribuzione delle temperature siano limitati.

Un modello alternativo che tiene esplicitamente conto del cambiamento climatico a lungo termine è proposto da Winter et al. (2016). Nel loro modello le osservazioni vengono

trasformate in una serie stazionaria standardizzata mediante una trasformazione di Box-Cox che tiene esplicitamente conto dell'evoluzione della temperatura media globale. Sulla serie standardizzata viene poi stimata una Pareto Generalizzata la cui volatilità viene fatta dipendere linearmente dalla temperatura media globale al fine di cogliere tendenze residue nella coda della distribuzione. La serie così ottenuta, invariante rispetto al tempo e alla temperatura media globale, viene poi modellata sulla base dell'approccio semi-parametrico esposto in questo documento. È importante sottolineare che il modello implementato elimina la stazionarietà nella distribuzione marginale, ma non nella struttura di dipendenza; quindi, permette di modellare più adeguatamente l'intensità dell'ondata di calore, ma non la loro persistenza che si modifica solo tenendo esplicitamente conto della non stazionarietà anche nelle caratteristiche di dipendenza.

Appendice

6.1 Pool adjacent violators algorithm

Nel paragrafo 3.4 si è visto che la distribuzione della durata di un'ondata di calore $\pi(\cdot)$ può essere calcolata a partire dalla seguente formula:

$$\theta(i) = Pr(N(u_n, r_n) = i \mid X_1 > u_n) \quad i = 1, \ldots, r_n$$

ovvero, la probabilità che l'ondata di calore abbia una durata pari a i sapendo che è iniziata il primo giorno del periodo considerato e non in un momento arbitrario e quindi pur di considerare un periodo r_n sufficientemente ampio si può definire

$$\pi(i) = \frac{\theta(i) - \theta(i+1)}{\theta(1)} \quad i = 1, 2, \ldots .$$

Tuttavia, dal momento che l'algoritmo simula un numero finito di catene markoviane, potrebbe accadere che per valori grandi di i, dove sono generate solo poche osservazioni, $\theta(i) < \theta(i+1)$ e di conseguenza $\pi(i) < 0$. Analoghe considerazioni valgono per

$$\theta_C(i) = Pr(C(i, u_n, r_n) = i \mid X_1 > u_n) \quad i = 1, \ldots, r_n$$

la cui distribuzione delle durate consecutive risulta pari a

$$\pi_C(i) = \frac{\theta_C(i) - \theta_C(i+1)}{\theta_C(1)}.$$

Una soluzione a questo problema è utilizzare l'algoritmo di Pool adjacent violators (Robertson et al., 1988), che impone una condizione di monotonicità a $\theta(i)$ e $\theta_C(i)$ in modo da garantire che le rispettive probabilità risultino non negative.

L'algoritmo prevede che al generico passo i si verifichi se $\theta(i) < \theta(i+1)$ (o equivalentemente se $\theta_C(i) < \theta_C(i+1)$ e in caso contrario si ponga la seguente condizione di pooling:

$$\theta(i)^{New} = \frac{\theta(i+1) + \theta(i)}{2}.$$

Se $\theta(i-1) \geq \theta(i)^{New}$, l'algoritmo procede, altrimenti si applica nuovamente la condizione di pooling fino a che non risulta $\theta(i-1) \geq \theta(i)^{New}$. Sostituendo θ con θ_C si ottiene la condizione di pooling per l'extremal index associato ad eccedenze consecutive.

L'algoritmo può portare a valori di $\pi(i)$ o $\pi_C(i)$ uguali a zero, ma evita che questi possano essere negativi.

Bibliografia

2003 Assessment of the Impact of the Heat Wave and Drought of theSummer 2003 on Agriculture and Forestry Report (Brussels: COPA-COGECA) p 15 COPA-COGECA (Committee of Professional Agricultural Organisations in the European Union–General Confederation of Agricultural Co-Operatives in the European Union

Abaurrea, J., Asín, J., Cebrián, A. C., & Centelles, A. (2007). Modeling and forecasting extreme hot events in the central Ebro valley, a continental-Mediterranean area. Global and Planetary Change, 57(1-2), 43-58.

Alexander, L. V., & Arblaster, J. M. (2017). Historical and projected trends in temperature and precipitation extremes in Australia in observations and CMIP5. Weather and climate extremes, 15, 34-56.

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., & García-Herrera, R. (2011). The hot summer of 2010: redrawing the temperature record map of Europe. Science, 332(6026), 220-224.

Bom, C. A. (2015). Climate Change in Australia Information for Australia's Natural Resource Management Regions: Technical Report, 216

Bortot, P., & Tawn, J. A. (1998). Models for the extremes of Markov chains. *Biometrika*, *85*(4), 851-867.

Brás, T. A., Seixas, J., Carvalhais, N., & Jägermeyr, J. (2021). Severity of drought and heatwave crop losses tripled over the last five decades in Europe. Environmental Research Letters, 16(6), 065012.

Ceccherini, G., Russo, S., Ameztoy, I., Romero, C. P., & Carmona-Moreno, C. (2016). Magnitude and frequency of heat and cold waves in recent decades: the case of South America. *Natural Hazards and Earth System Sciences*, *16*(3), 821-831.

Ceccherini, G., Russo, S., Ameztoy, I., Marchese, A. F., & Carmona-Moreno, C. (2017). Heat waves in Africa 1981–2015, observations and reanalysis. *Natural Hazards and Earth System Sciences*, *17*(1), 115-125.

Chase, T. N., Wolter, K., Pielke Sr, R. A., & Rasool, I. (2008). Reply to comment by WM Connolley on "Was the 2003 European summer heat wave unusual in a global context?". Geophysical Research Letters, 35(2).

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., ... & Valentini, R. (2005). Europe-wide reduction in primary productivity caused by the heat and drought in 2003. Nature, 437(7058), 529-533.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208, p. 208). London: Springer.

Davin, E. L., Seneviratne, S. I., Ciais, P., Olioso, A., & Wang, T. (2014). Preferential cooling of hot extremes from cropland albedo management. *Proceedings of the National Academy of Sciences*, *111*(27), 9757-9761.

D'Ippoliti, D., Michelozzi, P., Marino, C., de'Donato, F., Menne, B., Katsouyanni, K., ... & Perucci, C. A. (2010). The impact of heat waves on mortality in 9 European cities: results from the EuroHEAT project. Environmental Health, 9(1), 1-9.

Dong, B., Sutton, R. T., & Shaffrey, L. (2017). Understanding the rapid summer warming and changes in temperature extremes since the mid-1990s over Western Europe. *Climate Dynamics*, *48*, 1537-1554.

Domeisen, D. I., Eltahir, E. A., Fischer, E. M., Knutti, R., Perkins-Kirkpatrick, S. E., Schär, C., ... & Wernli, H. (2023). Prediction and projection of heatwaves. Nature Reviews Earth & Environment, 4(1), 36-50.

Erdenebat, E., & Sato, T. (2016). Recent increase in heat wave frequency around Mongolia: role of atmospheric forcing and possible influence of soil moisture deficit. *Atmospheric science letters*, *17*(2), 135-140.

Faivre, N, Vallejo Calzada, V. R., Cardoso Castro Rego, F. M., Moreno Rodríguez, J. M., & Xanthopoulos, G. (2018). Forest fires. Sparking firesmart policies in the EU.

Ferreira, H., & Ferreira, M. (2018). Estimating the extremal index through local dependence.

FitzGerald, G., Xu, Z., Guo, Y., Jalaludin, B., & Tong, S. (2016). Impact of heatwave on mortality under different heatwave definitions: a systematic review and meta-analysis. Environment international, 89, 193-203.

Fontaine, B., Janicot, S., & Monerie, P. A. (2013). Recent changes in air temperature, heat waves occurrences, and atmospheric circulation in Northern Africa. *Journal of Geophysical Research: Atmospheres*, *118*(15), 8536-8552.

Forzieri, G., Feyen, L., Russo, S., Vousdoukas, M., Alfieri, L., Outten, S., ... & Cid, A. (2016). Multi-hazard assessment in Europe under climate change. *Climatic Change*, *137*, 105-119.

Fraga, H., Molitor, D., Leolini, L., & Santos, J. A. (2020). What is the impact of heatwaves on European viticulture? A modelling assessment. Applied Sciences, 10(9), 3030.

Grieshaber, K. (2018). Animals, crops suffering as Europe's heatwave hits new highs. The News & Observer, Retrieved from https://www.voanews.com/europe/animals-crops-suffering-europes-heatwave-hits-new-highs

Heffernan, J. E., & Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(3), 497-546

Hsing, T., Hüsler, J., & Leadbetter, M. R. (1988). On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields*, *78*(1), 97-112.

Hsing, T. (1988). On the extreme order statistics for a stationary sequence. *Stochastic processes and their applications*, *29*(1), 155-169.

Keef, C., Papastathopoulos, I., & Tawn, J. A. (2013). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, *115*, 396-404.

Koppe, C., Kovats, S., Jendritzky, G., & Menne, B. (2004). *Heat-waves: risks and responses* (No. EUR/03/5036810). World Health Organization. Regional Office for Europe.

Ledford, A. W., & Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, *83*(1), 169-187.

Ledford, A. W., & Tawn, J. A. (2003). Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 521-543.

Leadbetter, M. R. (1982). *Extremes and Local Dependence in Stationary Sequences*. North Carolina univ at chapel hill dept of statistics.

Létard, V., Flandre, H., & Lepeltier, S. (2004). La France et les Français face à la canicule: les leçons d'une crise. *Rapport d'information du Sénat*, (195).

Leistung, E. (2018). The first nuclear power plants reduce production. Retrieved from https://www.world-nuclear.org/information-library/country-profiles/countries-g-n/germany.aspx

Liu, X., He, B., Guo, L., Huang, L., & Chen, D. (2020). Similarities and differences in the mechanisms causing the European summer heatwaves in 2003, 2010, and 2018. Earth's Future, 8(4), e2019EF001386.

Moron, V., Oueslati, B., Pohl, B., Rome, S., & Janicot, S. (2016). Trends of mean temperatures and warm extremes in northern tropical Africa (1961–2014) from observed and PPCA-reconstructed time series. *Journal of Geophysical Research: Atmospheres*, *121*(10), 5298-5319.

Mouhamed, L., Traore, S. B., Alhassane, A., & Sarr, B. (2013). Evolution of some observed climate extremes in the West African Sahel. *Weather and climate extremes*, *1*, 19-25.

O'Brien, G. L. (1987). Extreme values for stationary and Markov sequences. *The Annals of Probability*, 281-291.

Pascal, M., Wagner, V., Le Tertre, A., Laaidi, K., Honoré, C., Bénichou, F., & Beaudeau, P. (2013). Definition of temperature thresholds: the example of the French heat wave warning system. *International journal of biometeorology*, *57*, 21-29.

Perfekt, R. (1997). Extreme value theory for a class of Markov chains with values in ℝd. *Advances in Applied Probability*, *29*(1), 138-164.

Peters, W., Krol, M. C., Van Der Werf, G. R., Houweling, S., Jones, C. D., Hughes, J., ... & Tans, P. P. (2010). Seven years of recent European net terrestrial carbon dioxide exchange constrained by atmospheric observations. Global Change Biology, 16(4), 1317-1337.

Rahmstorf, S., & Coumou, D. (2011). Increase of extreme events in a warming world. *Proceedings of the National Academy of Sciences*, *108*(44), 17905-17909.

Ratnam, J. V., Behera, S. K., Ratna, S. B., Rajeevan, M., & Yamagata, T. (2016). Anatomy of Indian heatwaves. *Scientific reports*, *6*(1), 24395.

Reich, B. J., Shaby, B. A., & Cooley, D. (2014). A hierarchical model for serially-dependent extremes: A study of heat waves in the western US. Journal of Agricultural, Biological, and Environmental Statistics, 19, 119-135.

Robertson, T. Wright. FT and Dykstra, RL (1988) Order Restricted Statistical Inference.

Robine, J. M., Cheung, S. L. K., Le Roy, S., Van Oyen, H., Griffiths, C., Michel, J. P., & Herrmann, F. R. (2008). Death toll exceeded 70,000 in Europe during the summer of 2003. *Comptes rendus biologies*, *331*(2), 171-178.

Rohini, P., Rajeevan, M., & Srivastava, A. K. (2016). On the variability and increasing trends of heat waves over India. *Scientific reports*, *6*(1), 1-9.

Rootzén, H. (1988). Maxima and exceedances of stationary Markov chains. *Advances in applied probability*, *20*(2), 371-390.

Russo, S., Sillmann, J., & Fischer, E. M. (2015). Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, *10*(12), 124003.

Russo, S., Marchese, A. F., Sillmann, J., & Immé, G. (2016). When will unusual heat waves become normal in a warming Africa?. *Environmental Research Letters*, *11*(5), 054016.

Sánchez-Benítez, A., Barriopedro, D., & García-Herrera, R. (2020). Tracking Iberian heatwaves from a new perspective. *Weather and Climate Extremes*, *28*, 100238.

Sénat, 2004. Information report no. 195 – France and the French face the canicule: the lessons of a crisis. Appendix to the minutes of the session of February 3, 2004, 59-62

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., ... & Zhou, B. (2021). 11 Chapter 11: Weather and climate extreme events in a changing climate.

Sisson, S., & Coles, S. (2003). Modelling dependence uncertainty in the extremes of Markov chains. *Extremes*, *6*, 283-300.

Smith, R. L. (1992). The extremal index for a Markov chain. *Journal of applied probability*, *29*(1), 37-45.

Smith, R. L., & Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(3), 515-528

Smith, R. L., Tawn, J. A., & Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, *84*(2), 249-268.

Somini, (2018). 2018 will be fourth-hottest year on record, climate scientists predict. The Independent. Retrieved from https://www.theguardian.com/environment/2019/feb/06/global-temperatures-2018-record-climate-change-global-warming

Stéfanon, M., Drobinski, P., D'Andrea, F., Lebeaupin-Brossier, C., & Bastin, S. (2014). Soil moisture-temperature feedbacks at meso-scale during summer heat waves over Western Europe. Climate dynamics, 42, 1309-1324.Perkins, S.E. and P.B Gibson, 2015: Increased Risk of the 2014 Australian May Heatwave Due to Anthropogenic Activity. Bulletin of the American Meteorological Society, 96(12), S154–S157

Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, *75*(3), 397-415.

Trenberth, K.E., P.D. Jones, P.G. Ambenje, R. Bojariu, D.R. Easterling, A.M.G. Klein Tank, D.E. Parker, J.A. Renwick and Coauthors, 2007: Observations: surface and atmospheric climate change. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller, Eds., Cambridge University Press, Cambridge, 235-336.

Trenberth, K. E., & Fasullo, J. T. (2012). Climate extremes and climate change: The Russian heat wave and other climate extremes of 2010. *Journal of Geophysical Research: Atmospheres*, *117*(D17).

Vogel, M. M., Zscheischler, J., Wartenburger, R., Dee, D., & Seneviratne, S. I. (2019). Concurrent 2018 hot extremes across Northern Hemisphere due to human-induced climate change. Earth's future, 7(7), 692-703.

Xie, W., Zhou, B., You, Q., Zhang, Y., & Ullah, S. (2020). Observed changes in heat waves with different severities in China during 1961–2015. *Theoretical and Applied Climatology*, *141*, 1529-1540.

Wadsworth, J. L., Tawn, J. A., Davison, A. C., & Elton, D. (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 149-175.

Wild, M., Gilgen, H., Roesch, A., Ohmura, A., Long, C. N., Dutton, E. G., ... & Tsvetkov, A. (2005). From dimming to brightening: Decadal changes in solar radiation at Earth's surface. *Science*, *308*(5723), 847-850.

Winter, H. C., Brown, S. J., & Tawn, J. A. (2016). Characterising the changing behaviour of heatwaves with climate change. *Dynamics and Statistics of the Climate System*, *1*(1).

Winter, H. C., & Tawn, J. A. (2016). Modelling heatwaves in central France: a case-study in extremal dependence. Journal of the Royal Statistical Society: Series C: Applied Statistics, 345-365.

Winter, H. C., & Tawn, J. A. (2017). k th-order Markov extremal models for assessing heatwave risks. Extremes, 20, 393-415.

You, Q., Jiang, Z., Kong, L., Wu, Z., Bao, Y., Kang, S., & Pepin, N. (2017). A comparison of heat wave climatologies and trends in China based on multiple definitions. *Climate Dynamics*, *48*, 3975-3989.

https://www.ilpost.it/2022/06/28/estate-2003-caldo/

https://www.copernicus.eu/en/news/news/observer-wrap-europes-summer-2022-heatwave

https://www.salute.gov.it/portale/caldo/homeCaldo.jsp