Valutazione dei Rischi Climatici e della Vulnerabilità del Sistema AQP

Convenzione Operativa n° 2

Report finale dell'attività CG01.43.3

CONSUMI IDROPOTABILI – Analisi della correlazione tra consumi idropotabili e temperature per la valutazione degli effetti del cambiamento climatico sui fabbisogni idropotabili

Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici
Acquedotto Pugliese S.p.A.

cmcc

acquedotto pugliese
l'acqua, bene comune

Autori

Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici	Mohamed Azhar Paola Mercogliano Roberta Padulano
Acquedotto Pugliese S.p.A.	Gianluigi Fiori Vincenzo Patruno Davide Ritossa Luciano Venditti Gerardo Ventafridda

In copertina: Le Sorgenti AQP (Fonte: https://www.aqp.it/scopri-acquedotto/gli-impianti/sorgenti)

Sommario

Premessa

Il presente Report è riferito all'attività CG01.43.3: Correlazione tra consumi idropotabili e temperature (target "Consumi idropotabili"). L'attività nasce dalla considerazione che il cambiamento climatico potrebbe avere effetti non soltanto in termini di disponibilità della risorsa idrica (con adeguati parametri di quantità e qualità) ma anche sui fabbisogni idrici. È infatti possibile, in linea generale, che l'aumento delle temperature, come anche la diversa concentrazione degli eventi di pioggia, possa tradursi in un aumento dei consumi domestici. In proiezione, ciò, accompagnato da una generale diminuzione della risorsa idrica disponibile, potrebbe comportare una condizione (o un aggravamento) delle condizioni di scarsità idrica, con il conseguente mancato soddisfacimento, *tout court* o in solo in alcune parti dell'anno, dei fabbisogni.

La presente attività si differenzia dalle altre attività della Convenzione per alcuni aspetti. Innanzitutto, essa si configura come attività di ricerca, e pertanto non vuole e non può essere né esaustiva né definitiva: la si può considerare invece un'analisi esplorativa, all'interno della quale saranno proposti molteplici risultati ed evidenziati quelli più utili ai fini della Convenzione. In secondo luogo, l'attività prende le mosse da evidenze osservative interne al sistema acquedottistico, che però non sono mai state quantificate rigorosamente: pertanto, non è detto né scontato che tali evidenze si possano effettivamente tradurre in "numeri" affidabili o utili agli scopi della Convenzione. A riprova di ciò, verrà proposta una breve analisi di letteratura, dalla quale, al netto di poche eccezioni, non si riscontra in generale un forte effetto dei fattori climatici sui consumi idropotabili. In ultimo luogo, l'attività nasce con un "peccato originale": partendo dall'assunzione, comunque da verificare, che i consumi seguano un pattern stagionale concorde con il pattern climatico (dunque consumi più alti durante la stagione estiva), larga parte del territorio analizzato presenta una forte vocazione turistica e dunque una notevole fluttuazione della popolazione residente durante la stagione estiva, con un conseguente importante aumento dei consumi che potrebbe sovrapporsi, mascherandolo, all'effetto del clima. Purtroppo, all'interno della Convenzione non si dispone di tali informazioni, e dunque non è possibile discriminare le due concause.

1. Analisi di letteratura

Il diritto umano fondamentale di accedere ad acqua potabile sicura, economica e affidabile è sempre più minacciato. Mentre il mondo si confronta con la diminuzione delle risorse idriche e la crescente domanda, i sistemi di approvvigionamento idrico municipali sono sottoposti a uno stress eccezionalmente elevato. Una crescente disparità tra l'offerta e la domanda di acqua rende questa situazione già difficile molto più complicata, a causa di una convergenza di fattori demografici, socioeconomici ed idrologici. Un aspetto cruciale dell'uso dell'acqua riguarda il consumo domestico, che è soggetto a una vasta gamma di fattori, da quelli climatici e idrologici a quelli tecnici e socioeconomici. Schleich & Hillenbrand (2008) hanno classificato queste determinanti in tre categorie principali: economici, sociali ed ambientali. I fattori climatici sono un sottogruppo significativo all'interno della categoria ambientale che influisce sui meccanismi di utilizzo dell'acqua.

La stragrande maggioranza delle ricerche in questo campo si è concentrata sulla previsione della domanda residenziale di acqua per gestire attentamente le fonti idriche esistenti e pianificare strategicamente per le esigenze imminenti. Tipicamente, la domanda di acqua è divisa in "uso base" e "uso stagionale", o, in altre parole "indoor" e "outdoor". Mentre l'uso indoor viene tradizionalmente considerato indipendente dalle variabili climatiche, l'uso all'aperto è innegabilmente influenzato dalle condizioni meteorologiche, quali la temperatura, l'evapotraspirazione e la pioggia. Tuttavia, come evidenziato da Gato et al. (2007), è cruciale non trascurare l'impatto potenziale del clima sull'uso indoor, specialmente in contesti climatici specifici. In un'epoca in cui l'aumento delle temperature, i periodi di siccità prolungati e la diminuzione delle piogge stanno diventando la nuova norma, comprendere il legame tra clima e consumo d'acqua è di primaria importanza. Le previste conseguenze del cambiamento climatico – caratterizzate in generale da temperature più elevate, piogge ridotte ed estremi più frequenti – richiedono un'analisi completa di questa relazione. Come specificato da Zhoua et al. (2000), svelare queste complesse relazioni è essenziale per prevedere con precisione la futura domanda d'acqua e dunque i fabbisogni da soddisfare.

Numerose ricerche hanno indagato sull'influenza potenziale delle variabili climatiche sui modelli di consumo domestico dell'acqua. Alshaikhli et al. (2021), mediante uno studio in Qatar, hanno dimostrato che sembra essere soprattutto la temperatura, tra le variabili meteorologiche, ad influire sulla quantità di acqua utilizzata dalle persone. Ma hanno anche affermato che sono necessarie ulteriori ricerche, in particolare riguardo alla popolazione transitoria, poiché quest'ultimo aspetto potrebbe avere un impatto ancora maggiore.

Balling et al. (2006) hanno condotto uno studio per analizzare come le variabili climatiche influenzino il consumo annuale di acqua a Phoenix, USA. Lo studio ha affermato che temperatura e pioggia cambiano effettivamente la quantità di acqua utilizzata dalle persone. Ma in grandi città come Phoenix, sebbene la maggior parte dell'acqua sia utilizzata all'aperto, questi cambiamenti sono piccoli: ciò accade, nello specifico, perché la fonte di approvvigionamento della città è così abbondante da rendere la città totalmente resiliente rispetto al clima rispetto ai consumi idrici. Invece, in centri più piccoli con fonti di approvvigionamento più locali anche un piccolo aumento della temperatura può causare un significativo aumento del consumo idrico.

Slavíková et al. (2013) valutano gli impatti delle variabili climatiche sul consumo domestico nella Repubblica Ceca. Lo studio si basa sull'analisi di un'unica serie temporale di consumo complessivo, e dimostra che, nel caso in esame, non si riscontrano forti correlazioni con le condizioni meteorologiche. Lo studio conclude, in ogni caso, ribadendo la necessità di ulteriori verifiche.

In un articolo di ricerca scritto da Syme et al. (2004) è stata investigata l'influenza del giardinaggio sul consumo idrico. Lo studio ha rivelato che le persone con una propensione per il giardinaggio e le attività all'aperto tendono a consumare più acqua.

Timotewos et al. (2022) hanno analizzato il consumo mensile di acqua in tre città etiopi per comprendere l'influenza dei fattori idrologici sull'uso domestico dell'acqua. Utilizzando modelli di regressione multipla, hanno cercato di identificare quali delle variabili climatiche – temperatura media, precipitazione o umidità relativa – influissero prevalentemente sul consumo d'acqua. I loro risultati hanno indicato che un aumento di 1°C della temperatura ha comportato un aumento del 5.28% nell'uso dell'acqua nella prima città. Tuttavia, nella seconda città, tutti i fattori climatici hanno mostrato una correlazione minima con i consumi. La terza città ha evidenziato l'umidità relativa come un leggero deterrente all'uso dell'acqua, mentre gli altri fattori sembrano avere scarso effetto. In sostanza, lo studio ha rivelato che la temperatura ha un certo impatto in regioni specifiche, ma nel complesso le influenze climatiche sono inconsistenti o deboli.

2. Dati utilizzati

Nella presente attività sono essenzialmente utilizzate tre tipologie di dati:

- Dataset di osservazioni climatiche (precipitazione e temperatura) sul periodo di riferimento 1981-2010. Si sceglie in particolare di utilizzare E-OBS, avente risoluzione orizzontale di circa 10 km e temporale giornaliera. Tale dataset è frutto di un ensemble di esperimenti di interpolazione di osservazioni raccolte in corrispondenza delle stazioni termo-pluviometriche monitorate dai singoli stati europei e appartenenti ad alcuni network (Cornes et al., 2018). Ai fini dell'analisi i valori giornalieri E-OBS sono spazialmente aggregati in modo da ottenere un'unica serie riferita a ciascun areale di riferimento (definito nel seguito), eventualmente aggregato a scala mensile.
- Dataset di variabili climatiche (precipitazione e temperatura) restituite da catene di simulazione climatica sul periodo presente (1981-2010) e sull'orizzonte futuro 2021-2050. Si sceglie in particolare di utilizzare le 14 catene modellistiche EURO-CORDEX illustrate in Tabella 1, sotto gli scenari di concentrazione RCP 2.6, 4.5 e 8.5. Anche in questo caso si andrà a mediare le serie di valori giornalieri su ciascuno degli areali di riferimento, e si considererà anche l'aggregazione mensile.
- Osservazioni di volumi recapitati all'utenza. Si tratta in particolare di serie temporali di volumi giornalieri misurati in corrispondenza dei "punti di consegna", ovvero punti strategici all'interno del sistema acquedottistico situati a monte delle reti di distribuzione urbana, in cui vengono effettuate misurazioni utili all'ottimizzazione della gestione dell'intera rete. L'ubicazione dei punti di consegna, i consumi misurati e i relativi metadati (tra cui in particolare i Comuni serviti) sono stati trasmessi da AQP per un totale di 117 punti, nel seguito indicati come "sensori". Le serie di consumo sono giornaliere e si estendono dal 1 settembre 2012 al 30 settembre 2022; tuttavia, ne verrà utilizzato solo un subset, in modo da avere un numero finito di anni monitorati. Il dataset è descritto e raffigurato in maniera più estesa nel seguito.

Tabella 1: Lista delle simulazioni climatiche adottate in questo studio.

Global Climate Model (Institution)	Regional Climate Model (Institution)	Realizzazione*
EC-EARTH (ICHEC, Ireland)	RCA4 (SMHI)	r12i1p1
	CLM4-8-17 (CLMcom)	r12i1p1
	RACMO22E (KNMI)	r12i1p1
	HIRHAM5 (DMI)	r3i1p1
HadGEM2-ES (UK Met Office UK)	RCA4 (SMHI)	r1i1p1
	RACMO22E (KNMI)	r1i1p1
	HIRHAM5 (DMI)	r1i1p1
MPI-ESM-LR (MPI, Germany)	RCA4 (SMHI, Sweden)	r1i1p1
	CSC-REMO2009 (MPI)	r2i1p1
	CSC-REMO2009 (MPI)	r1i1p1
M-CM5 (CNRM-CERFACS-CM5)	RACMO22E (KNMI)	r1i1p1
	ALADIN63 (CNRM)	r1i1p1
NorESM1-M (NCC)	RCA4 (SMHI)	r1i1p1
	REMO2015 (GERICS)	r1i1p1

** La realizzazione r<N>i<N>p<N> viene utilizzata per distinguere simulazioni strettamente correlate che differiscono, ad esempio, per condizioni iniziali o parametrizzazioni fisiche.*

3. Nota metodologica

La Figura 1 mostra il workflow delle analisi condotte, che può essere ricondotto essenzialmente a quattro fasi:

1. "Comprensione dei dati". In questa fase viene condotta un'analisi esplorativa del database fornito per verificare la consistenza delle informazioni e la coerenza con gli obiettivi dell'attività. Tra le principali operazioni figurano la rimozione dei sensori duplicati, la verifica delle coordinate spaziale dei sensori, l'associazione con le serie temporali di consumo idropotabile, l'aggregazione per "macro-area" (vedi seguito).
2. "Pre-processing dei dati". Questa fase consiste nella "pulizia" del dataset in modo tale che questo possa essere utilizzato con successo come input delle analisi. Tra le principali operazioni figurano l'individuazione e la rimozione di anomalie e dati inaffidabili.
3. "Clustering". Questa fase consiste nella ripartizione delle serie temporali di consumo in un numero non noto a priori di possibili gruppi, con lo scopo di individuare possibili diversi "comportamenti" ed eventualmente tenerne conto nelle analisi (ad esempio, effettuando le operazioni separatamente per ciascun gruppo).
4. "Analisi delle correlazioni". Si tratta del *core* della presente attività, giacché vengono ricercate e individuate possibili correlazioni tra i consumi idropotabili e una selezione di variabili esogene di natura climatica, ipotizzate rilevanti in base alla letteratura vigente.
5. "Valutazione degli effetti del cambiamento climatico sui consumi". Sulla scorta di quanto ritrovato nella fase precedente, le eventuali correlazioni vengono utilizzate per inferire circa i valori di consumo da attendersi in futuro, tenendo conto degli effetti del cambiamento climatico sulle variabili atmosferiche risultate rilevanti.

Per ciascuna di queste fasi verranno qui presentate le principali operazioni effettuate, e nel seguito verranno presentati i risultati.

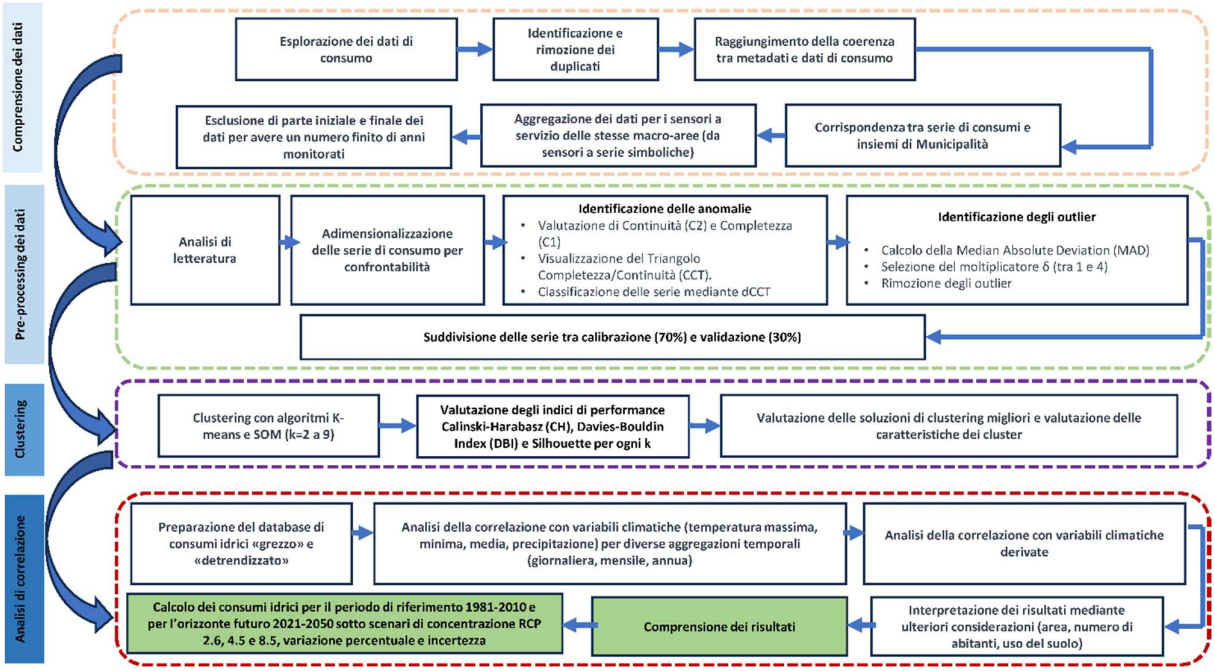

Figura 1. Work flow della metodologia.

3.1 Comprensione dei dati

In questa fase, è stata effettuata un'attenta disamina dell'insieme di dati sul consumo idrico forniti. I dati, raccolti tramite sensori remoti posizionati in vari punti di misurazione, detti "punti di consegna", sono stati trasmessi all'interno di due file Excel, per un totale di 186 punti di misurazione, o "sensori". Un passaggio cruciale in questa fase è stato affrontare le incongruenze all'interno del database. In primo luogo, con il supporto del personale AQP, sono stati identificate e rimosse le voci duplicate nel file dei metadati, mentre il file dei consumi è risultato privo di duplicati. Per alcuni sensori duplicati, invece, è stata rilevata e soddisfatta la necessità di unire le informazioni, in particolare l'elenco dei Comuni serviti.

Dal punto di vista spaziale, per ciascun sensore l'insieme dei Comuni serviti è stato trasformato in un'unica zona geografica, denominata "macro-area". Per ciascuna macro-area (tante quante sono i sensori) è stato estratto il dato di popolazione complessiva da fonte ISTAT. Infine, si è rilevato che alcuni sensori risultano servire la stessa macro-area: in questo caso, è stata associata a quella macro-area una serie temporale dei consumi pari alla somma delle due o più serie misurate. Ciò ha permesso da un lato di risolvere alcune criticità (serie misurate ai sensori con valori nulli nella prima o nell'ultima parte del periodo osservato) e dall'altro di ridurre il numero di serie temporali analizzate. Per distinguere le serie temporali misurate ai punti di consegna (sensori) da quelle "totali" associate a una macro-area (coincidenti con le prime quando la macro-area è associata nei metadati a un solo sensore), si farà nel seguito riferimento alle seconde come "symbolic points" o "symbolic time series". L'intera trattazione presentata nel seguito riguarderà, appunto, le serie "simboliche".

Come ultima operazione, è stata effettuata la rimozione dei dati dal 1 settembre al 31 dicembre 2012 e dal 1 gennaio al 30 settembre 2022, al fine di avere serie estese su un numero finito di anni, nel caso in esame pari a 11.

3.2 Pre-processing dei dati

La fase di pre-processing consiste essenzialmente nell'individuazione di anomalie e *outlier* nel database analizzato (serie simboliche). Le anomalie consistono in particolare nella presenza di un numero sospetto di dati assenti (*missing data*) o nulli (*zero data*), mentre gli *outlier* sono valori non nulli ma molto "diversi" dai valori "normali". Data l'eterogeneità del dataset (in particolare, la significativa differenza nel numero di abitanti serviti, che determina valori medi di consumo fortemente diversi) si è preferito, come prima operazione, normalizzare i dati di consumo dividendo ciascun dato giornaliero per la media dei dati. In altre parole, per ciascuna serie simbolica giornaliera è stato valutato il valore medio (media annua dei volumi giornalieri consumati, crescente al crescere del numero di abitanti) ed è quindi stato diviso ciascun dato della serie per la media della stessa serie. I dati così trasformati vengono definiti "adimensionali": l'adimensionalizzazione permette di trattare il database come un *unicum* per la fase di pre-processing.

3.2.1 Individuazione delle anomalie

Per l'individuazione delle anomalie è stata utilizzata la procedura proposta da Padulano & Del Giudice (2020), che si basa sulla valutazione delle due caratteristiche principali complessive di una serie temporale di dati: la Completezza e la Continuità.

- La Completezza (C_1) è definita come il rapporto tra il numero N_{val} di dati "validi" e il numero massimo possibile N di dati all'interno di una serie temporale (Eq. 1). La "validità" di un dato è stabilita dall'operatore in base allo scopo delle analisi: dati non validi possono essere, ad esempio, i dati mancanti o i dati nulli, o entrambi. Una serie in cui tutti i dati sono validi presenta un valore di C_1 pari a 1, mentre una serie in cui nessun dato è valido presenta un valore di C_1 pari a zero.
- La Continuità (C_2) misura il modo in cui i dati validi e non validi sono organizzati all'interno della serie temporale. Essa è il complemento all'unità del doppio del rapporto tra il numero n_{inv} di intervalli di dati non validi e il numero N massimo possibile di dati nella serie (Eq. 2). Un valore di C_2 pari a 1 indica una serie in cui tutti i dati sono validi e sono distribuiti in un unico, lungo intervallo; un valore di C_2 pari a 0 indica una serie formata da un unico intervallo di dati non validi.

$$C_1 = \frac{N_{val}}{N} \tag{1}$$

$$C_2 = 1 - 2\frac{n_{inv}}{N} \tag{2}$$

Continuità e completezza non variano in modo indipendente bensì congiunto. Quando la completezza è pari al 50% (metà dei dati sono validi, l'altra metà è costituita da dati non validi) la continuità minima possibile è pari a 0 e corrisponde ad una serie formata da una continua alternanza di dati validi e non validi. Il "triangolo di continuità e completezza" (CCT) mostrato in Figura 2 mostra i limiti di variazione di C_1 e C_2. Una serie temporale può essere rappresentata in questo grafico da un singolo punto di coordinate pari alla continuità e alla completezza della serie, e, secondo le Eq. 1 e 2, tale punto deve essere compreso nel triangolo. Una serie temporale ottimale presenta solo dati validi organizzati in un unico, lungo intervallo: essa presenterà continuità e completezza unitarie ed è rappresentata in Figura 2 dal punto B.

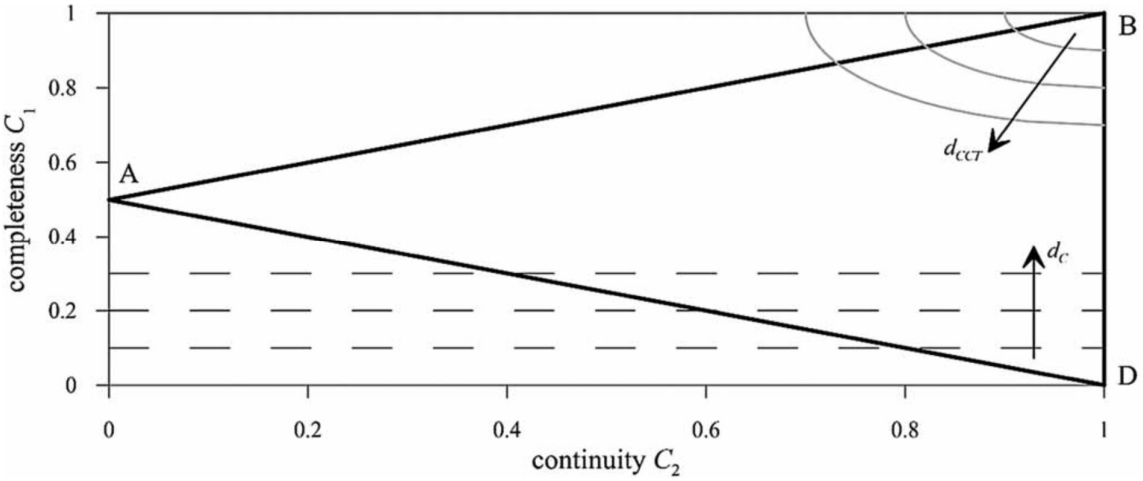

Figura 2. Triangolo di continuità e completezza.

Solitamente, il controllo della qualità di una serie temporale di dati viene effettuato fissando una soglia sulla completezza minima (d_C): come rappresentato in Figura 2, maggiore è la soglia maggiore è la qualità (cioè la completezza) che si vuole garantire. Dato un database di serie temporali con diverse caratteristiche, maggiore è la soglia inferiore sarà il numero di serie temporali che passeranno il controllo di qualità, e, analogamente, maggiore sarà il numero di serie "scartate".

L'approccio proposto da Padulano & Del Giudice (2020) considera invece una soglia congiunta d_{CCT}, definita come distanza geometrica dal punto B che rappresenta le condizioni ottimali (Eq. 3). Come verrà presentato nel seguito, sono stati esplorati diversi valori di d_{CCT}, pervenendo infine a un valore ottimale per gli scopi del lavoro.

$$d_{CCT} = \sqrt{(1 - C_1)^2 + (1 - C_2)^2} \tag{3}$$

3.2.2 Individuazione degli *outlier*

Per l'individuazione degli *outlier* è stato utilizzato il criterio della mediana, secondo quanto proposto da Padulano & Del Giudice (2020). Il metodo della mediana consiste nell'andare a misurare la distanza tra ciascun punto della serie e la mediana della serie stessa, valutata però su dati opportunamente normalizzati per tener conto della eventuale asimmetria nella distribuzione dei dati (sicuramente esistente nel caso in esame poiché i dati sono vincolati ad essere non negativi). I dati normalizzati che distano "troppo" dalla mediana così valutata vengono ritenuti *outlier* e rimossi dal database.

La normalizzazione dei dati di consumo (già precedentemente adimensionalizzati) viene effettuata mediante l'algoritmo proposto nell'Eq. 4. Quindi, viene calcolata la Median Absolute Deviation (MAD) che rappresenta la distanza di ciascun dato (adimensionalizzato e normalizzato) dalla mediana (Eq. 5). Quindi, viene fissata una soglia di accettabilità del dato rispetto alla MAD (Eq. 6).

$$x_{iT} = \sqrt{\frac{x_i - x_{min}}{x_{max} - x_{min}}} \tag{4}$$

$$MAD = median\left(x_{iT} - median(x_T)\right) \tag{5}$$

$$x_{iT} - median(x_T) > |\delta \cdot \alpha \cdot MAD| \quad (6)$$

dove x_i è il generico dato adimensionalizzato mentre x_{iT} è il dato adimensionalizzato e normalizzato. Nell'Eq. 4, x_{min} e x_{max} rappresentano rispettivamente il minimo e il massimo dell'insieme dei valori di consumo adimensionalizzati tra tutte le serie simboliche. Nelle Eq. 5 e 6, x_T rappresenta l'insieme dei valori di consumo adimensionalizzati e normalizzati tra tutte le serie simboliche. Nell'Eq. 6, α è un fattore di scala posto pari a 1.4826, che presenta una precisa interpretazione solo se i dati sono normalmente distribuiti; δ è invece un fattore correttivo ("moltiplicatore") di α, che tiene conto della non-normalità. Sulla base di quanto proposto da Miller (1991), sono stati esplorati diversi valori di δ tra 1 e 4.

Utilizzando l'Eq. 6, vengono classificati come *outlier*, e rimossi dal database (ovvero sostituiti con *missing data*) quei valori di consumo per i quali i corrispondenti valori adimensionalizzati e normalizzati hanno una distanza dalla loro mediana superiore a una certa soglia, costituita dal valore assoluto del prodotto $\delta \cdot \alpha \cdot MAD$. Essendo però i dati di consumo non negativi, si considererà soltanto il limite superiore (ciò equivale a considerare l'Eq. 6 priva del simbolo di valore assoluto).

3.2.3 Aggregazione e ripartizione del dataset

Successivamente alla rimozione degli *outlier*, può essere opportuno, in caso il numero di *outlier* trovati sia alto, operare una successiva scrematura andando ad eliminare quelle serie simboliche per le quali la percentuale di outlier supera una soglia definita (nella presente analisi ciò non è stato necessario). Infine, la rimozione degli *outlier* (giornalieri) può avere delle ripercussioni sull'aggregazione a scala mensile: tale aggregazione è stata condotta con il vincolo che, se in un mese vi sono meno di 20 giorni validi, il valore mensile viene ritenuto *missing data*, poiché la sua stima non sarebbe affidabile.

L'ultima operazione della fase di pre-processing è la suddivisione dell'intero database, così "pulito", in due subset:

- Un subset di calibrazione, per il quale verranno effettuate le analisi di correlazione;
- Un subset di validazione, per il quale verranno verificati i risultati.

Per ottimizzare la ripartizione del database, si è deciso di far confluire tutte le serie "scartate" nelle fasi di identificazione di anomalie e *outlier* nel database di validazione; quindi, ulteriori serie sono state estratte, con campionamento casuale, dalle serie "accettate" e aggiunte al dataset di validazione in modo tale che la ripartizione calibrazione/validazione raggiunga la quota 70%/30%.

3.3 Clustering

Il clustering è una tecnica di *data mining* che consiste nel ripartire un set multidimensionale di dati in distinti subset, o cluster (Padulano & Del Giudice, 2018). Ogni cluster raggruppa oggetti con caratteristiche simili, con l'obiettivo di svelare pattern o comportamenti non evidenti nel dataset (Zhou et al., 2013; Sancho-Asensio et al., 2014). L'efficienza di un algoritmo di clustering può essere quantificata misurando la distanza tra oggetti appartenenti allo stesso cluster (distanza *within-cluster*), che deve essere minima, e la distanza tra cluster diversi (distanza *between-cluster*), che deve essere massima (López et al., 2011; Avni et al., 2015). La distanza tra due oggetti multidimensionali è in genere espressa mediante la norma Euclidea (Keogh et al., 2001; Popivanov & Miller, 2002).

Il concetto di cluster non presenta un'interpretazione univoca in letteratura; ciò ha portato alla nascita di molteplici algoritmi (Rokach & Maimon, 2005) caratterizzati da criteri di ottimizzazione diversi (Fraley & Raftery, 1998). Un tentativo di classificazione degli algoritmi di clustering è rappresentato dalla distinzione tra algoritmi "di partizione" (quali il k-means), "gerarchici" (quali il dendrogramma) e "model-based" (quali la Self-Organizing Map) (Zhou et al., 2013).

L'algoritmo k-means è forse il più semplice e diffuso metodo di clustering: per la sua applicazione, va prima fissato il numero *k* di cluster a priori; quindi, l'algoritmo suddivide gli oggetti tra i vari cluster minimizzando una funzione obiettivo che misura la distanza tra i centroidi dei cluster e la distanza degli oggetti appartenenti a un cluster dal centroide di quel cluster. Il dendrogramma mira a costruire i cluster partendo dall'insieme degli oggetti e dividendoli progressivamente fino al raggiungimento di determinate condizioni (*top-down*), oppure partendo dai singoli oggetti e unendoli progressivamente fino al raggiungimento di determinate condizioni (*bottom-up*). La Self-Organizing Map (SOM) è un algoritmo di clustering che sfrutta il concetto di rete neurale: esso richiede l'identificazione a priori di un numero massimo possibile di cluster, e riempie man mano i cluster di oggetti conservando le rappresentazioni topologiche (cluster che occupano posizioni vicine presentano centroidi simili tra loro). Padulano & Del Giudice (2018) hanno dimostrato che gli algoritmi di clustering si possono anche sfruttare in modo combinato, ad esempio effettuando dapprima una SOM, e quindi applicando il metodo del k-means ai centroidi risultanti per affinare ulteriormente il risultato (si parla in questo caso di *mixed strategy*). La letteratura vigente non è in grado di stabilire quale sia l'algoritmo migliore; piuttosto, la scelta va vagliata a seconda dello specifico obiettivo, giacché ogni algoritmo presenta vantaggi e svantaggi (Räsänen et al., 2010; López et al., 2011; Zhou et al., 2013).

Nel presente lavoro, ci si è focalizzati sull'applicazione dei metodi SOM e k-means, per i quali va stabilito a priori il numero di cluster da esplorare: di conseguenza, l'analisi è stata eseguita per un numero variabile di cluster (tra 2 e 9) e scegliendo poi la soluzione ottimale per gli scopi del progetto. Per effettuare tale scelta, i risultati ottenuti sono stati confrontati attraverso opportuni indicatori di performance o *Clustering Validity Index* (CVI), che quantificano l'efficienza dell'algoritmo misurando la distanza *within-cluster* e quella *between-cluster* (Everitt, 1980). Tra i numerosi CVI esistenti sono stati considerati i seguenti:

- l'indice di Calinski-Harabasz (Calinski & Harabasz, 1974), che misura il rapporto tra la varianza *between-cluster* e quella *within-cluster*. Valori alti di CH indicano elevate performance.
- L'indice di Davies-Bouldin (Davies & Bouldin, 1979), che misura la somiglianza media tra ciascun cluster e il cluster più simile. Valori alti di DB indicano basse performance.
- La *silhouette* (Rousseeuw, 1987), che misura la distanza di ciascun oggetto in un cluster da ciascun oggetto in un cluster simile. Valori alti di S indicano cluster ben delineati, quindi alte performance (S può variare tra -1 e 1).

Si noti che l'esplorazione dei risultati del clustering mediante CVI non dà risultati esatti, quanto piuttosto essi possono essere considerati un supporto alle decisioni, da affiancare all'ispezione visiva dei risultati e da interpretare a seconda degli scopi dell'analisi.

3.4 Analisi di correlazione

Lo scopo di questa attività è determinare se, e in che misura, le serie simboliche di consumo siano correlabili a variabili esogene, cioè esterne al dataset, e nella fattispecie variabili climatiche. A tal fine, è stata effettuata un'analisi esplorativa di correlazione considerando, per i consumi idrici, diversi possibili accorgimenti: il database tal quale, inteso come "pulito" e riportato alle sue unità di misura native, cioè i litri ("database grezzo"); il database "detrendizzato", in cui ogni dato di ciascuna serie simbolica viene diviso per il valore

medio annuo dell'anno in cui il dato viene misurato; l'aggregazione nativa, cioè quella giornaliera; l'aggregazione mensile.

La scelta di analizzare sia il database grezzo sia quello detrendizzato è stata fatta per considerare diversi aspetti. Quando confrontiamo l'uso giornaliero dell'acqua con le variabili climatiche giornaliere, stiamo cercando di capire come piccole variazioni nelle variabili climatiche (temperatura/precipitazioni) ogni giorno possano influenzare la quantità di acqua utilizzata dalle persone. Tuttavia, a volte (e ciò accade nel presente database) vi sono grandi variazioni nell'uso dell'acqua da un anno all'altro che non possono essere spiegate solo dal clima. La detrendizzazione elimina queste grandi variazioni e consente di focalizzarsi solo sul pattern day-by-day (stesse considerazioni possono essere fatte sulle serie mensili). Naturalmente, l'analisi di correlazione delle serie detrendizzate deve essere accompagnata dall'analisi dei valori medi annui, che ci aiuta a capire se, in generale, gli anni più caldi o più freddi influenzano l'uso dell'acqua.

Per quanto concerne le variabili esogene, la Tabella 2 mostra l'elenco di quelle considerate nell'analisi. Alcune sono variabili generiche (temperatura media ed estrema, pioggia, giorni piovosi), altre sono state introdotte sulla base della letteratura vigente e di osservazioni ad hoc (differenziale di temperatura, soglie, condizioni specifiche). Solo per le variabili più semplici sono state esplorate diverse risoluzioni temporali; considerando i risultati di tale analisi "esplorativa", si è preferito considerare, per altre variabili, solo la risoluzione nativa. L'intera analisi è stata applicata sia al database grezzo sia a quello detrendizzato.

Tabella 2. Lista delle variabili esogene considerate nelle analisi di correlazione, e risoluzione temporale considerata.

Variabili	Risoluzione temporale considerate nell'analisi di correlazione
Temperatura media giornaliera	Giornaliera, settimanale, mensile, stagionale, annuale
Temperatura massima giornaliera	Giornaliera, settimanale, mensile, stagionale, annuale
Temperatura minima giornaliera	Giornaliera, settimanale, mensile, stagionale, annuale
Precipitazione	Giornaliera, settimanale, mensile, stagionale, annuale
Differenza di tra la temperatura media giornaliera corrente e quella del giorno prima (con lag tra 0 e 3 giorni)	Giornaliera (lag > 0 solo per regressione multipla)
Differenza di tra la temperatura minima giornaliera corrente e quella del giorno prima (con lag tra 0 e 3 giorni)	Giornaliera (lag > 0 solo per regressione multipla)
Differenza di tra la temperatura massima giornaliera corrente e quella del giorno prima (con lag tra 0 e 3 giorni)	Giornaliera (lag > 0 solo per regressione multipla)
Temperatura media giornaliera nei giorni asciutti	Giornaliera
Temperatura massima giornaliera nei giorni asciutti	Giornaliera
Temperatura minima giornaliera nei giorni asciutti	Giornaliera
Numero di giorni piovosi	Mensile (solo per regressione multipla)
Numero di giorni secchi dall'ultimo giorno piovoso	Giornaliera
Temperatura media giornaliera nei giorni asciutti o poco piovosi (<10mm)	Giornaliera
Temperatura massima giornaliera nei giorni asciutti o poco piovosi (<10mm)	Giornaliera
Temperatura media giornaliera nei giorni asciutti o poco piovosi (<10mm)	Giornaliera
Temperatura massima giornaliera nei giorni asciutti o poco piovosi (<10mm)	Giornaliera
Precipitazione nei giorni asciutti o poco piovosi (<10mm)	Giornaliera
Temperatura media giornaliera (Inverno)	Giornaliera
Temperatura massima giornaliera (Inverno)	Giornaliera
Precipitazione (Inverno)	Giornaliera
Temperatura media giornaliera (Estate)	Giornaliera
Maximum temperature(Estate)	Giornaliera

Precipitazione (Estate)	Giornaliera
Temperatura media giornaliera (Autunno)	Giornaliera
Temperatura massima giornaliera (Autunno)	Giornaliera
Precipitazione (Autunno)	Giornaliera
Temperatura media giornaliera (Primavera)	Giornaliera
Temperatura massima giornaliera (Primavera)	Giornaliera
Precipitazione (Primavera)	Giornaliera
Temperatura media giornaliera (Tutte le stagioni tranne l'inverno)	Giornaliera
Temperatura massima giornaliera (Tutte le stagioni tranne l'inverno)	Giornaliera
Precipitazione (Tutte le stagioni tranne l'inverno)	Giornaliera
Temperatura media giornaliera per temperature > di una soglia (25°C e 30°C)	Giornaliera
Temperatura massima giornaliera per temperature > di una soglia (25°C e 30°C)	Giornaliera
Precipitazione per temperature > di una soglia (25°C e 30°C)	Giornaliera

Per quantificare l'entità della correlazione tra due variabili è stato utilizzato il ben noto coefficiente di Pearson ρ (Eq. 7). Il quadrato del coefficiente di Pearson è il ben noto coefficiente di determinazione R^2, definito come la percentuale della varianza totale della variabile target (nel nostro caso i consumi) che riesce ad essere spiegata dalla variabile esogena. Maggiore è il valore di R^2, maggiore è l'entità della correlazione, maggiore è l'efficienza della regressione; valori prossimi a zero indicano assenza di correlazione. Mentre R^2 può variare tra 0 e 1, ρ può variare tra -1 e 1: il valore assoluto rappresenta l'entità della correlazione, il segno rappresenta una dipendenza diretta (+) o inversa (-) tra le due variabili.

$$\rho = \frac{1}{n} \cdot \frac{\sum_i (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sigma_X \cdot \sigma_Y} \tag{7}$$

dove X_i e Y_i sono i valori delle due variabili che caratterizzando il singolo dato, $\bar{X}$ e $\bar{Y}$ sono le medie sull'intero dataset delle due variabili, e σ_X e σ_Y ne sono le deviazioni standard; n è invece il numero di dati. Infine, seguendo le indicazioni di Guilford (1965), è possibile classificare il grado di correlazione in base al valore del coefficiente di Pearson (Tabella 3).

Tabella 3. Classificazione di Guilford.

Coefficiente di correlazione ρ	Classe di correlazione
$\rho = 0.0$	Nessuna correlazione
$0.0 < \rho \leq 0.1$	Lieve correlazione
$0.1 < \rho \leq 0.3$	Debole correlazione
$0.3 < \rho \leq 0.5$	Correlazione moderata
$0.5 < \rho \leq 0.7$	Alta correlazione
$0.7 < \rho \leq 0.9$	Correlazione molto alta
$0.9 < \rho < 1.0$	Correlazione quasi perfetta
$\rho = 1$	Correlazione perfetta

4. Risultati

In questa sezione sono presentati i risultati di ciascuna fase illustrata nella Nota Metodologica.

4.1 Comprensione dei dati

I consumi idropotabili oggetto del presente lavoro sono, come detto, misurati in corrispondenza dei "punti di consegna", o "sensori" nel testo: essi sono dislocati sull'intero territorio della regione Puglia (Figura 3). I consumi si riferiscono, nel complesso, a oltre 4 milioni di abitanti, e si riferiscono ad acqua approvvigionata da varie fonti (sorgenti naturali, invasi superficiali, acque sotterranee). I dati sono forniti in formato Excel per un totale di 186 sensori distribuiti in due file. Ogni sensore è univocamente identificato mediante un codice, ed è associato a una serie temporale di volumi idrici giornalieri misurata in litri. Sono inoltre associati ad ogni sensore dei metadati, quali ad esempio le coordinate nel sistema di riferimento WGS84, e il nome delle Municipalità (una o più) servite dal sensore.

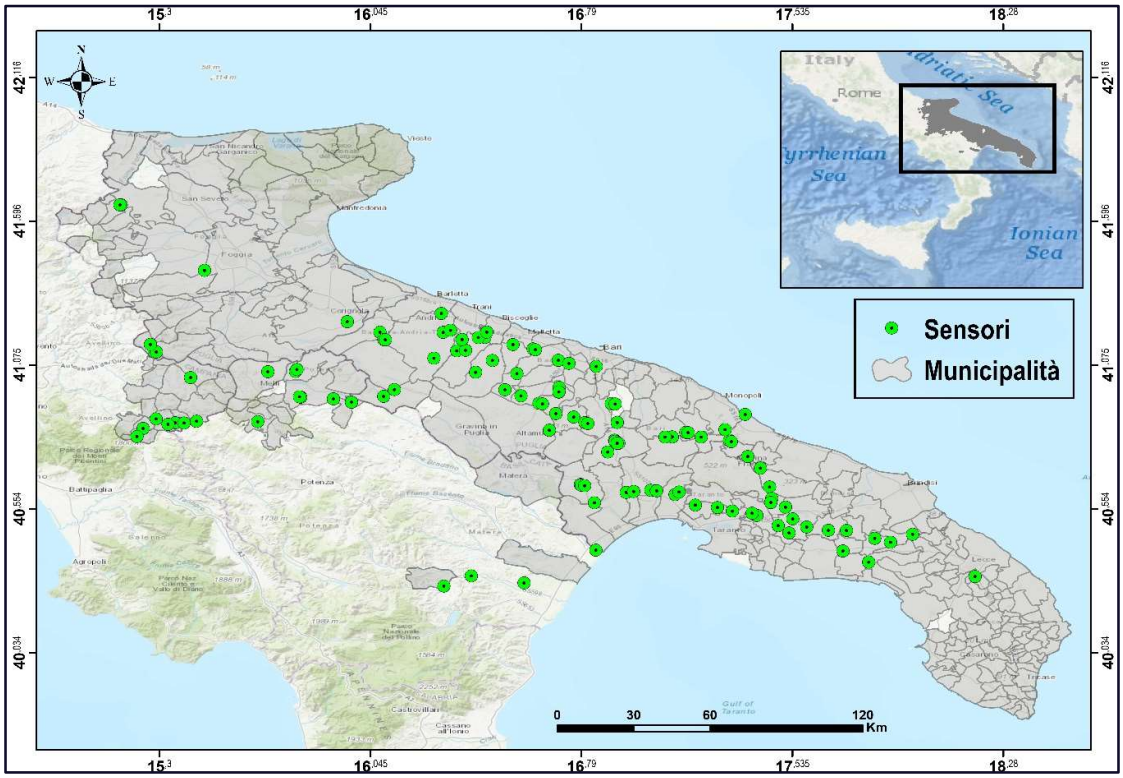

Figura 3. Ubicazione delle Municipalità e dei sensori nell'area analizzata.

Da un controllo preliminare di consistenza, è emerso che 16 sensori presentavano voci perfettamente duplicate nel database dei metadati; inoltre, alcuni sensori presentavano voci duplicate ma differivano per il nome dei Comuni serviti. Per queste ultime, con il supporto del personale AQP, è stato possibile unire i Comuni in un'unica voce. Sensori diversi ma associati alle stesse coordinate sono inoltre possibili (essi non sono quindi distinguibili in Figura 3) in corrispondenza di serbatoi con diverse uscite, tutte monitorate.

Dal punto di vista spaziale, è stato necessario risolvere una inconsistenza legata a leggere differenze nel nominativo di alcuni Comuni tra il database dei metadati e il database ISTAT, usato come riferimento per una stima del numero di abitanti serviti. La Figura 4 mostra il numero di sensori a servizio della stessa Municipalità, mentre la Figura 6 mostra il numero di Municipalità servite da ciascun sensore. Dalla Figura 4 si evince che la maggior parte (114) dei punti di misura sono associati a una municipalità; 24 punti di misura sono associati a 2 municipalità; al massimo, vi sono quattro punti di misura associati a 34 municipalità. Dalla Figura 5 si evince invece che il Comune di Bari è monitorato da 14 punti di misura, mentre circa metà dei Comuni è monitorata da 1 solo punto di misura.

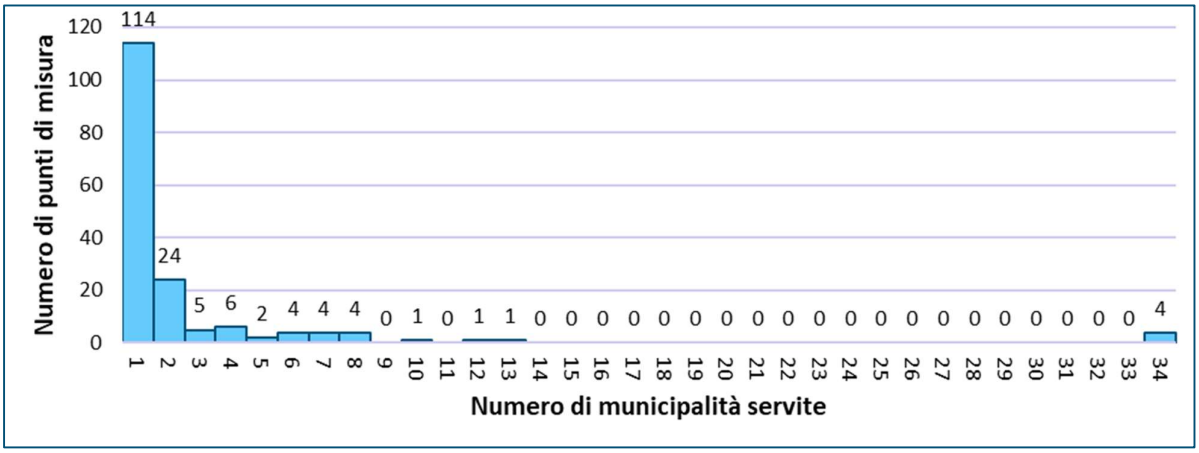

Figura 4. Numero di Municipalità servite da ciascun sensore.

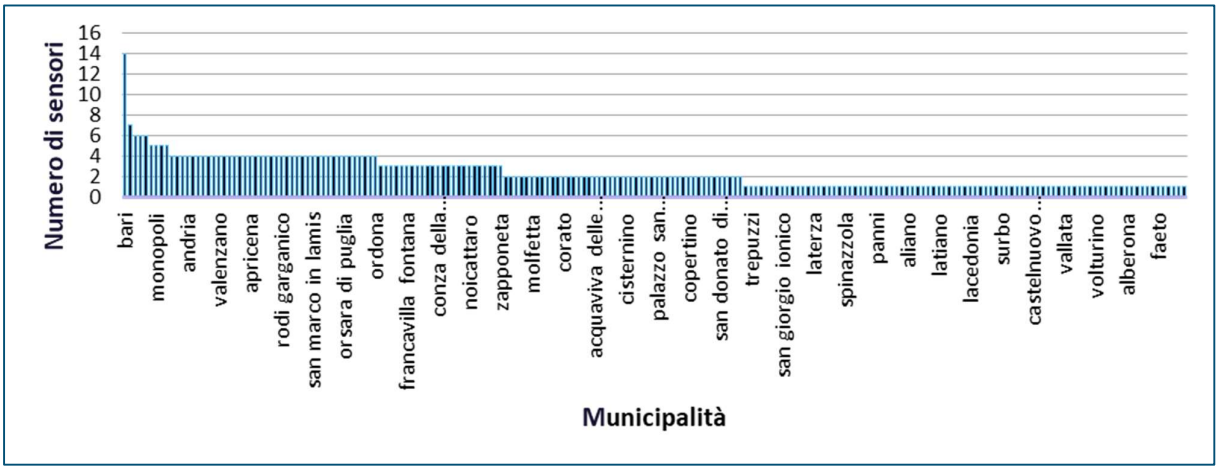

Figura 5. Numero di sensori a servizio di ciascuna Municipalità.

I Comuni serviti da ciascun sensore sono stati uniti all'interno di "macro-aree", e le serie temporali a servizio della stessa macro-area sono state sommate, ottenendo un numero più ridotto di serie simboliche. Il numero totale delle serie simboliche è 106, cui corrispondono le macro-aree mostrate in Figura 6. L'associazione tra i sensori e le Municipalità servite è presentata in Appendice I.

Infine, l'ultima operazione di questa fase è stata la rimozione dei dati compresi tra il 1 settembre e il 31 dicembre 2012 e tra il 1 gennaio e il 30 settembre 2022, affinché tutte le serie si estendano su un numero di anni finito, che risulta essere pari a 11. La Figura 7 mostra un istogramma dell'insieme dei dati contenuti nel database, dove si evidenzia la rilevante asimmetria dei dati, nettamente schiacciati verso i valori più bassi, distribuzione tipica dei dati non-negativi. La Figura 8 mostra invece il pattern medio annuo di consumo per ogni serie simbolica, dove cioè, per ogni mese, si riporta il valore medio tra gli anni di osservazione del consumo aggregato (la somma, cioè, dei valori giornalieri di ciascun mese). Come si evince dalla Figura 8, esistono alcune macro-aree con consumi medi nettamente più alti: si tratta delle macro-aree a servizio delle Municipalità più popolose. La Figura 9 mostra, per ciascuna macro-area, la correlazione tra la media annua dei pattern in Figura 8 e il numero di abitanti serviti, dove si nota una correlazione generalmente positiva, sebbene non perfetta, tra le due quantità.

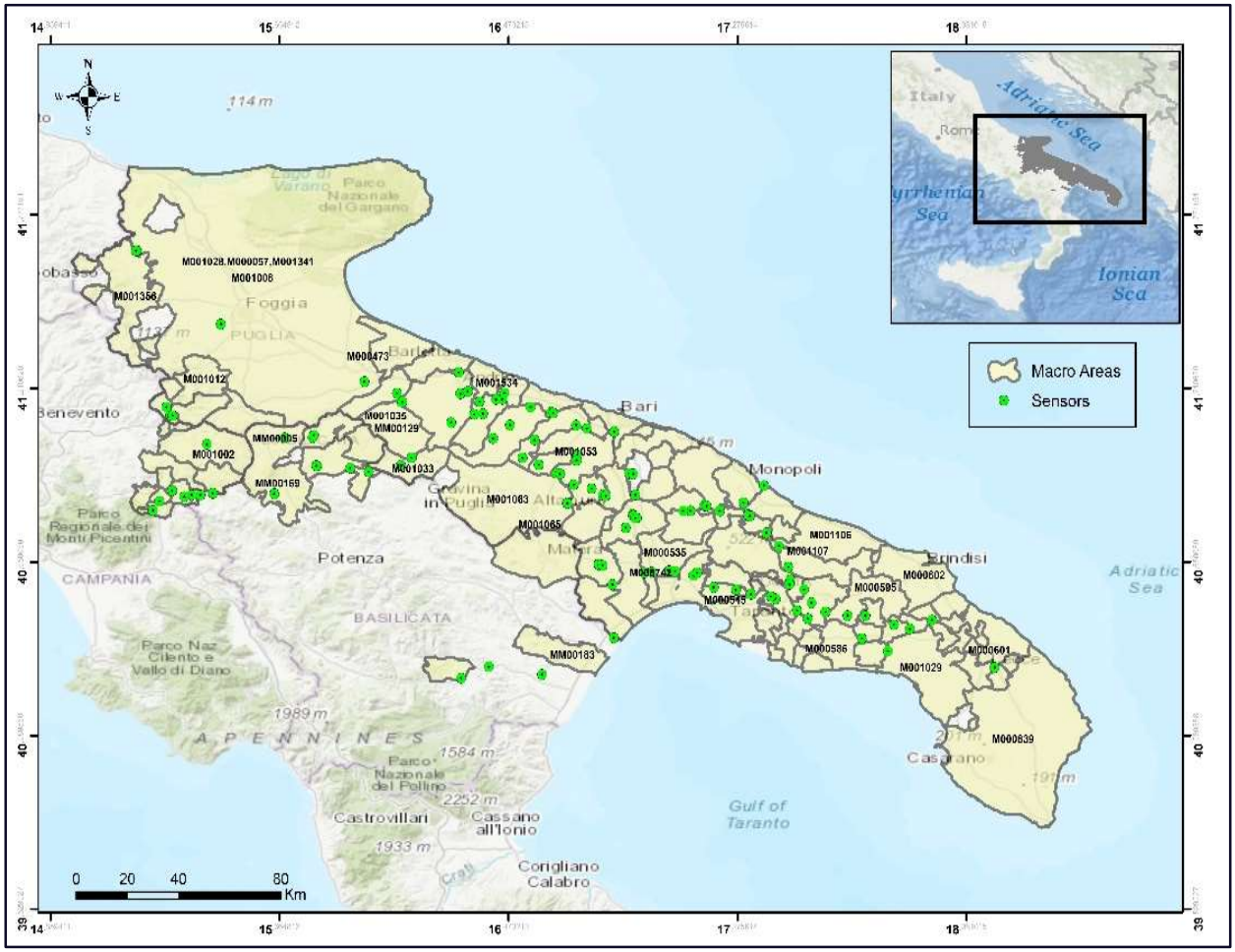

Figura 6. Macro-aree e relative sensori associate.

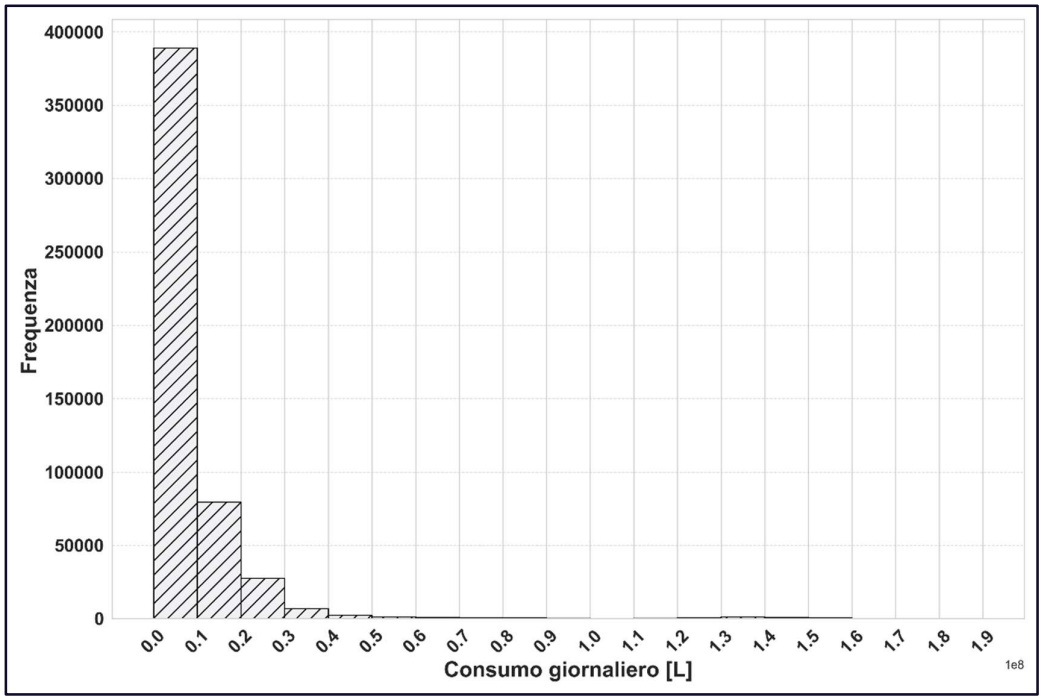

Figura 7. Istogramma complessivo dei valori di consumo giornaliero del database.

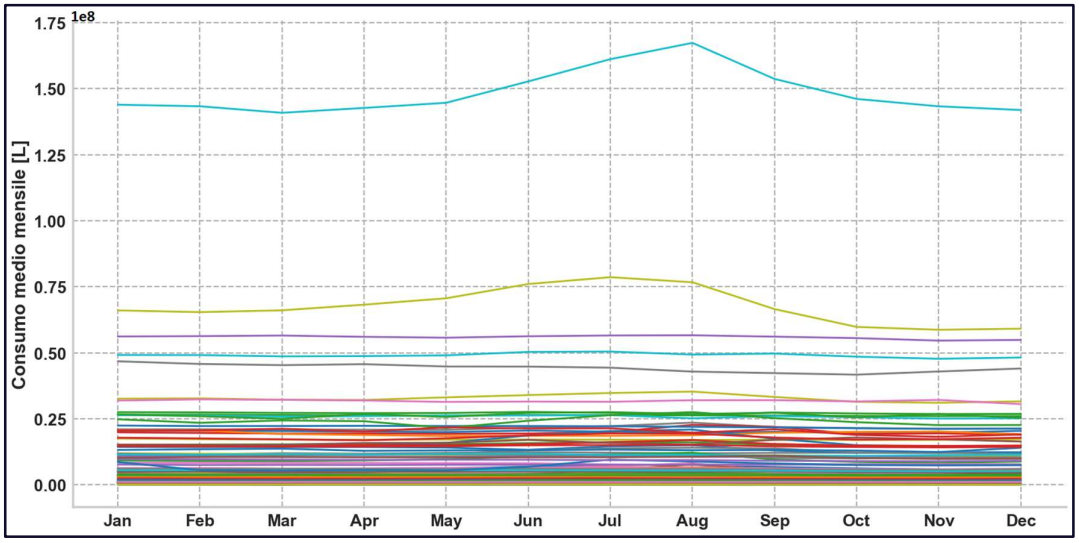

Figura 8. Pattern di valori medi mensili di consumo per tutte le macro-aree.

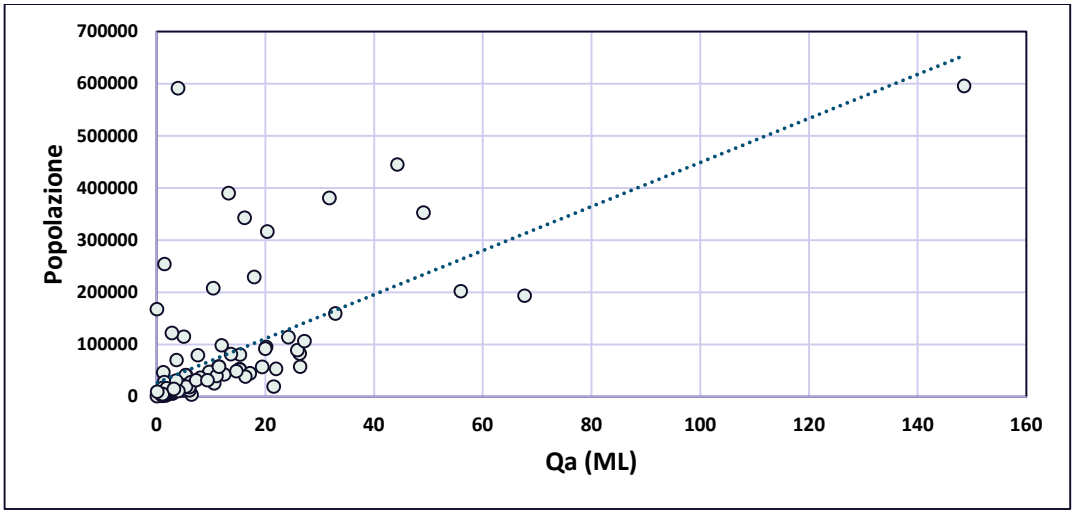

Figura 9. Correlazione tra il consumo medio annuo e il numero di abitanti per ciascuna macro-area.

4.2 Pre-processing dei dati

4.2.1 Identificazione delle anomalie

La *heat map* in Figura 10 mostra, per ciascuna serie simbolica, tutti I valori giornalieri di consumo: in particolare, la coordinata verticale identifica il codice della macro-area, la coordinata orizzontale rappresenta il tempo, il colore rappresenta il dato di consumo, in una gradazione giallo/blu dove il giallo rappresenta i valori massimi, il blu rappresenta i valori minimi (il minimo assoluto è zero), il bianco rappresenta i valori mancanti. Proprio in relazione ai valori mancanti, la Figura 10 già permette di individuare alcune serie simboliche caratterizzate da una importante presenza di *missing data*: per queste serie la striscia di dati è quasi totalmente bianca; per altre serie, vi sono solo brevi intervalli di dati mancanti; per altre infine, non ve ne sono affatto. Inoltre, analizzando i valori massimi risaltano immediatamente quelle serie (una, in particolare, nella parte alta della figura) che presentano valori nettamente più elevati delle altre.

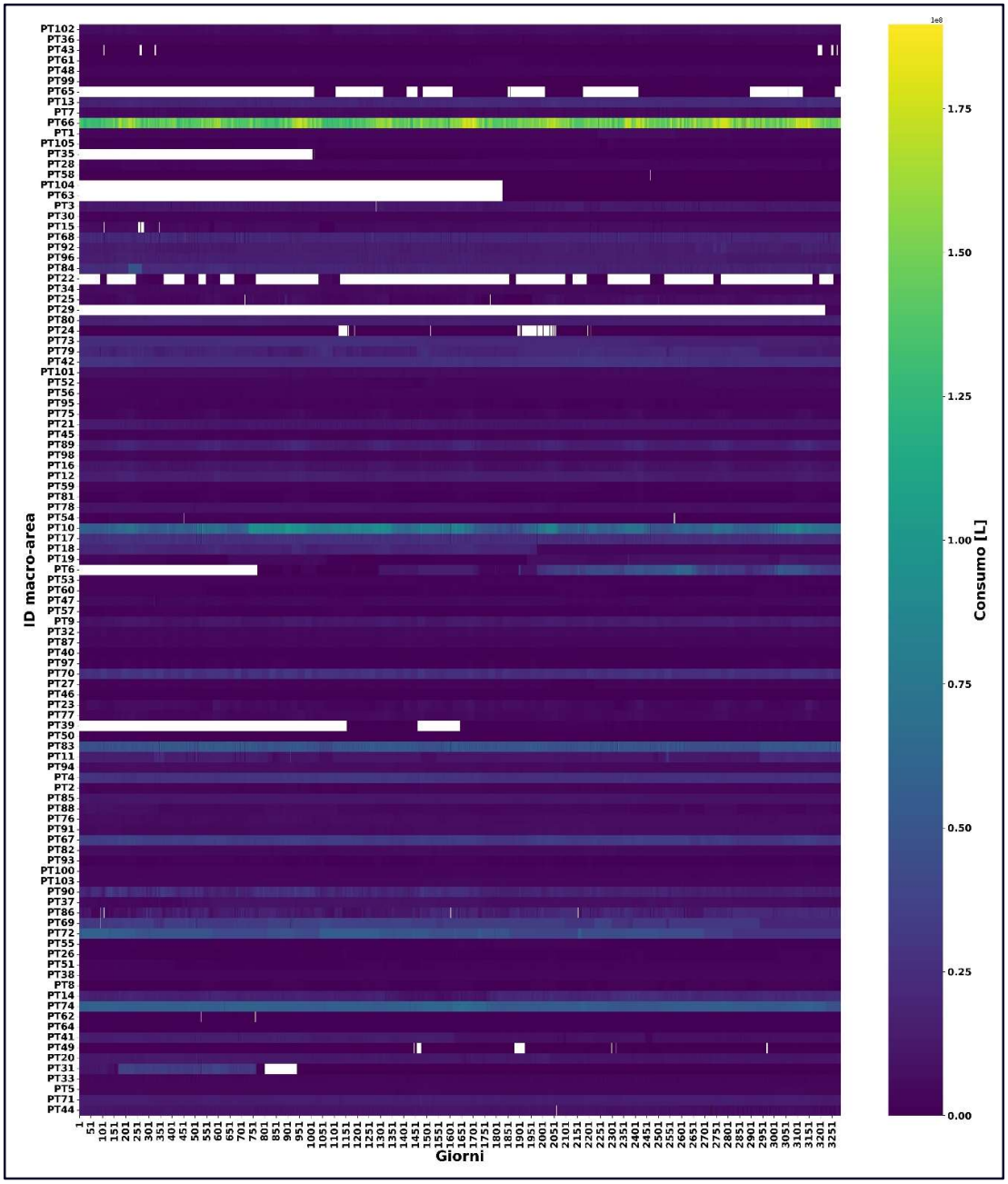

Figura 10. Heat map dei valori di consumo giornaliero [L] per ciascuna macro-area nel tempo.

Seguendo quanto illustrato nella Nota Metodologica, sul database in Figura 10 è stata condotta la procedura di identificazione delle anomalie valutando la posizione di ciascuna serie nel triangolo di continuità e completezza. Inoltre, giacché ciascuna serie serve un numero non ridotto di abitanti, e considerando anche l'esistenza di perdite nella rete di distribuzione, è altamente improbabile che si realizzino valori pari a zero. Di conseguenza, si è deciso di sostituire ai *zero data* l'etichetta di *missing data*, trattandoli cioè come anomalie. In altre parole, a vantaggio di sicurezza, si tratta come "anomala" la presenza sia di troppi dati mancanti, sia di troppi dati nulli.

La Figura 11 mostra il posizionamento delle serie simboliche all'interno del triangolo di continuità e completezza. A testimonianza dell'ottima qualità del database, tutti i punti risultano caratterizzati da una

continuità superiore al 90%, il che indica che, anche per diversi livelli di completezza, tutte le serie presentano i dati validi organizzati in pochi, lunghi intervalli, mentre i dati non validi sono organizzati in pochi, brevi intervalli. Sulla scorta di quanto proposto da Padulano & Del Giudice (2020), le serie sono state raggruppate in cinque classi di qualità a seconda della loro distanza *dCCT* dal punto ottimale, di coordinate (1,1). Le cinque classi sono:

- Classe 1 (*dCCT* <=0.2): in blu in Figura 11 e 13, raggruppa le serie più vicine allo scenario ottimale;
- Classe 2 (0.2 < *dCCT* <= 0.4): in verde in Figura 11 e 13, presenta le serie un po' meno vicine allo scenario ottimale;
- Classe 3 (0.4 < *dCCT* <=0.6): in giallo in Figura 11 e 13, presenta le serie a distanza intermedia tra lo scenario ottimale e lo scenario peggiore;
- Classe 4 (0.6 < *dCCT* <=0.8): in rosso in Figura X11 e 13 presenta le serie alquanto vicine allo scenario peggiore;
- Classe 5 (dCCT > 0.8): in viola in Figura 11 e 13, raggruppa le serie peggiori, più lontane dallo scenario ottimale.

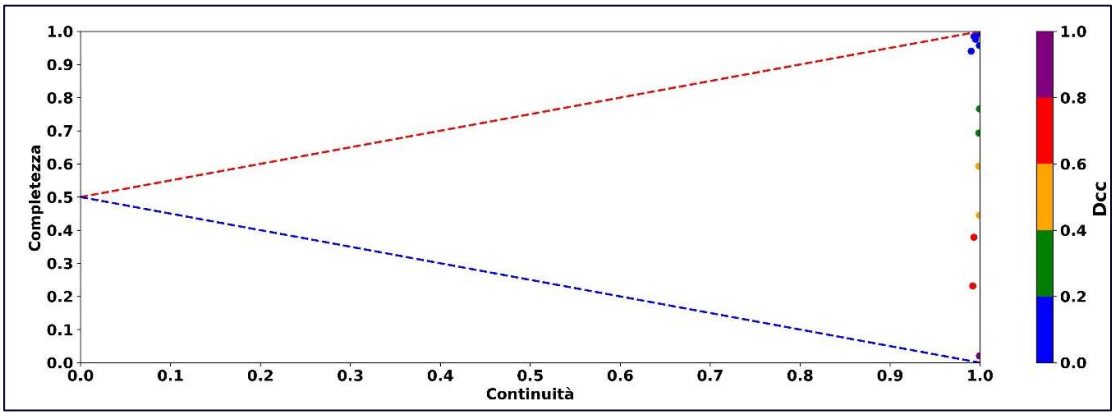

Figura 11. Rappresentazione del database nel triangolo di continuità e completezza.

Delle 106 serie simboliche totali, 81 presentano caratteristiche ideali; di conseguenza, la soglia sul valore di *dCCT* potrebbe essere tranquillamente settata sullo zero, portando all'accettazione delle 81 serie "perfette" e al rigetto delle 25 serie rimanenti. Tuttavia, si è in questo caso deciso di aumentare la soglia al valore di 0.2, accettando cioè tutte le serie ricadenti nella classe 1, per un totale di 98 serie accettate e 8 serie rigettate, pari al 7.5% del database. La Figura 12 mostra il numero di serie simboliche accettate (a sinistra) e il numero di serie rigettate (a destra) al variare del valore soglia assegnato alla distanza *dCCT* dall'ottimo.

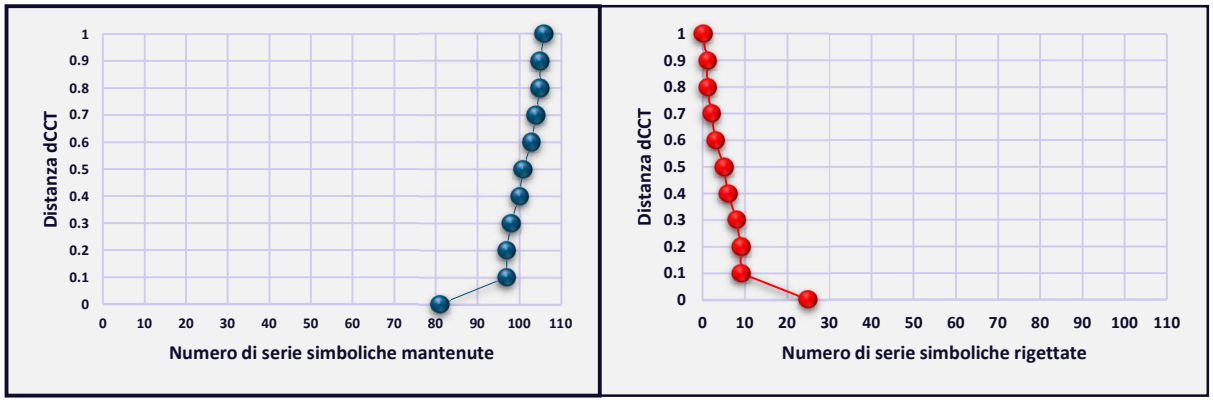

Figura 12. Numero di serie simboliche che passano il test di qualità (a sinistra) e che non lo passano (a destra) al variare della distanza complessiva dCCT dall'ottimo.

Infine, la Figura 13 mostra, per ognuna delle cinque classi, una serie simbolica di esempio. Come si evince dalla Figura 13, anche le classi vicine alla prima (ad esempio, la classe 2) possono presentare andamenti a prima vista poco plausibili: infatti, il triangolo consente di individuare le anomalie, in questo caso, in termini di valori mancanti e valori nulli troppo frequenti, ma non permette di esprimersi circa il valore assunto dai dati ritenuti validi.

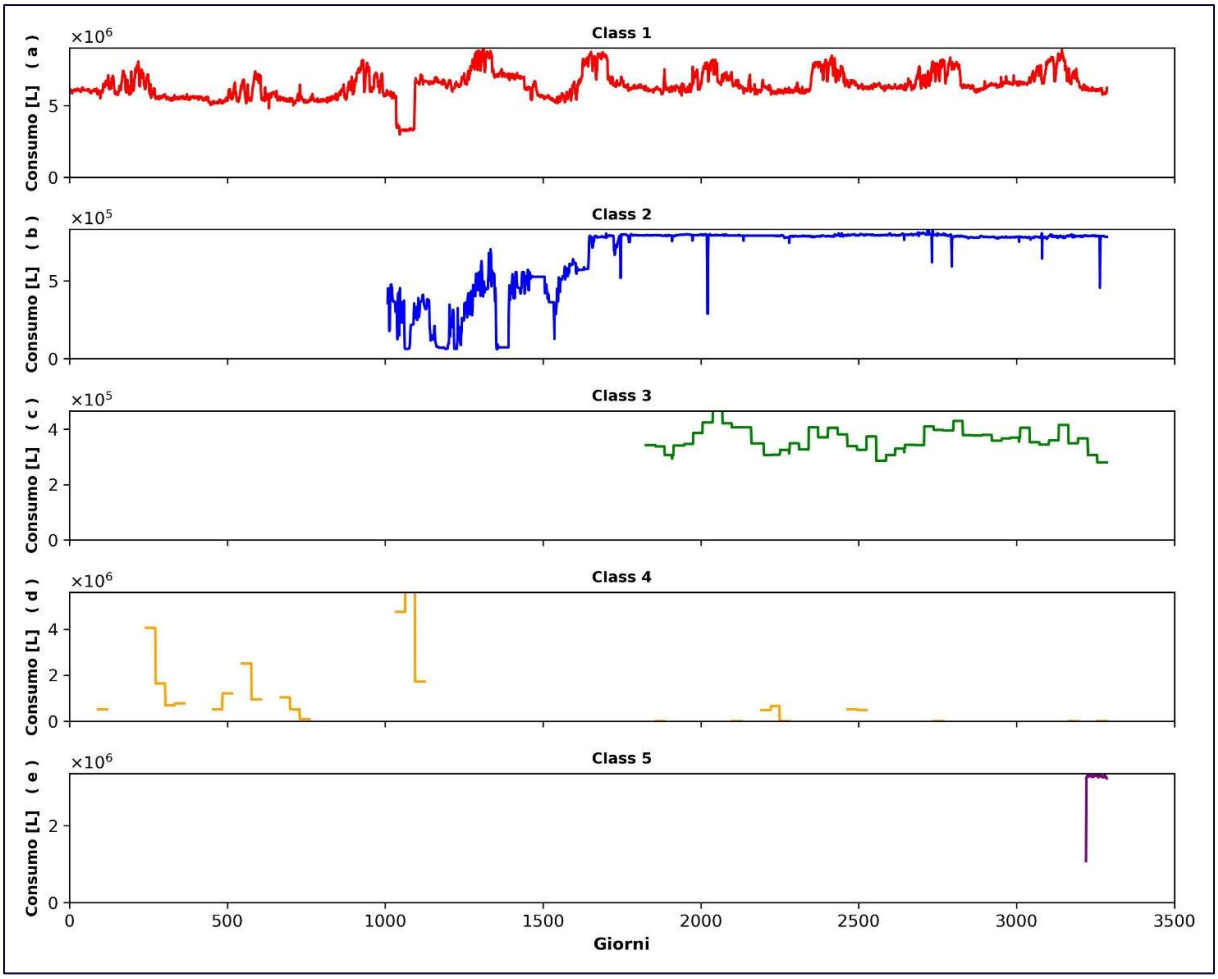

Figura 13. Serie simbolica "tipo" estratta casualmente da ciascuna delle cinque classi di distanza complessiva dCCT dall'ottimo.

4.2.2 Identificazione degli *outlier*

Come illustrato nella Nota Metodologica, con riferimento alle Eq. 4, 5 e 6 è stato settato un valore di α pari a 1.4826 e sono stati testati valori di δ tra 1 e 4 con passo 0.5. Valori più bassi di δ comportano una maggiore sensitività del dataset rispetto agli *outlier*, nel senso che valori anche non troppo distanti dalla mediana vengono classificati come *outlier*, con la certezza di scartare valori effettivamente troppo alti ma con la possibilità di scartare anche valori "normali". Invece, valori alti comportano una bassa sensitività, nel senso che vengono scartati soltanto i valori estremamente lontani dalla mediana, con la possibilità di mantenere nel database valori particolarmente alti.

La procedura di identificazione degli *outlier* è stata condotta sull'intero dataset adimensionalizzato (si noti che la fase di identificazione delle anomalie non prende in considerazione i valori assunti dai dati, pertanto l'adimensionalizzazione è irrilevante), comprensivo delle 8 serie simboliche rigettate in seguito all'analisi delle anomalie. La Tabella 4 mostra i risultati della procedura al variare del valore di δ: sono in particolare mostrati la mediana e la Mean Absolute Deviation dei dati adimensionalizzati, il valore di α, quello di δ e il

valore di consumo adimensionale che si ricava come soglia per l'identificazione degli outlier. Il valore soglia di consumo, in litri, può essere ottenuto, per ciascuna serie simbolica, moltiplicando il valore in Tabella 4 per la media dei consumi utilizzata per l'adimensionalizzazione (Figura 9, un istogramma generale è presentato in Figura 14).

Tabella 4. Soglie per l'outlier detection al variare del moltiplicatore δ.

MAD	Mediana	α	δ	Threshold
0.010	0.185	1.4826	1.000	1.159
0.010	0.185	1.4826	1.500	1.247
0.010	0.185	1.4826	2.000	1.339
0.010	0.185	1.4826	2.500	1.434
0.010	0.185	1.4826	3.000	1.532
0.010	0.185	1.4826	3.500	1.634
0.010	0.185	1.4826	4.000	1.738

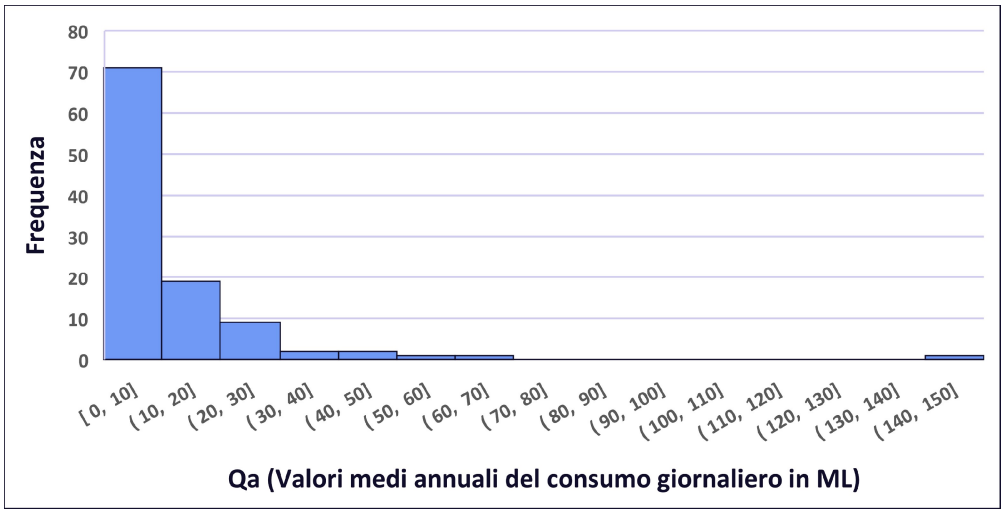

Figura 14. Istogramma complessivo dei valori di consumo idrico totale annuo nel database.

La Figura 15 presenta la frequenza del consumo giornaliero adimensionale (*Cd*) e i valori soglia ottenuti al variare di δ. Si noti che l'istogramma si estende fino ad un valore di *Cd* pari a 30: tuttavia, per facilitare la visualizzazione dei risultati, se ne rappresenta il particolare per *Cd*<=5. La Tabella 5 riassume invece la percentuale di dati classificati come *outlier*, rispetto al totale dei dati, tra le varie serie simboliche e al variare di δ. Si nota che, ad esempio, scegliendo δ = 1 quasi tutte le serie simboliche risultano interessate dalla presenza di *outlier*, e in questo caso gli outlier risultano il 17% del totale dei dati. Invece, scegliendo δ = 4 circa il 34% delle serie simboliche risultano interessate dalla presenza di *outlier*, i quali risultano solo il 2% del totale. Al fine di raggiungere un bilanciamento tra la rimozione di valori inaffidabili e la necessità di preservare la numerosità delle serie si è optato per un valore di δ pari a 3. Per tale valore, 58 serie su 106 risultano interessate dalla presenza di *outlier*, e il numero di *outlier* si aggira intorno al 2% del totale dei dati.

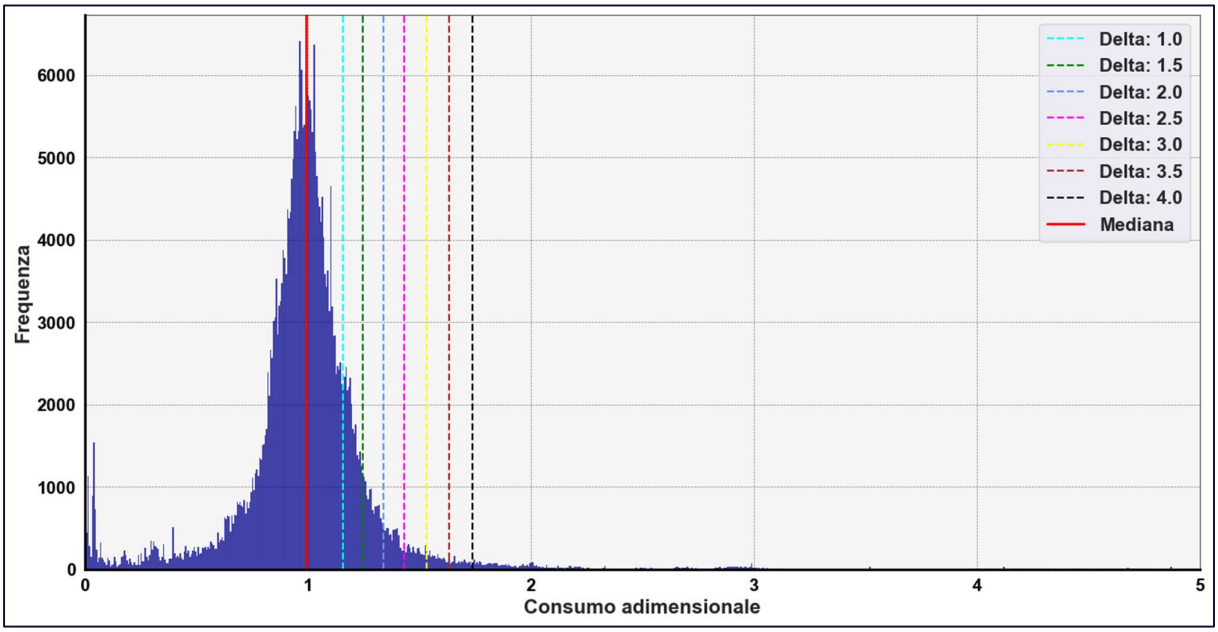

Figura 15. Istogramma complessivo dei valori di consumo idrico adimensionale Cd e soglie per l'identificazione degli outlier al variare del moltiplicatore δ. Si noti che il valore massimo dell'asse delle ascisse è stato limitato a 5 per migliorare la visualizzazione dei dati, ma il massimo valore attinto supera 30.

Tabella 5. Percentuale di outlier al variare del moltiplicatore δ.

δ	Numero totale di valori nel database	Numero di *outlier*	Percentuale di *outlier*	Numero di serie simboliche in cui è rilevata presenza di almeno un *outlier*	Numero di serie simboliche prive di *outlier*
1	333327	58262	17%	103	3
1.5	333327	32259	10%	95	11
2	333327	20114	6%	82	24
2.5	333327	13903	4%	68	38
3	333327	10238	3%	58	48
3.5	333327	7866	2%	42	64
4	333327	6338	2%	36	70

La Tabella 5 mostra, per ciascun valore del moltiplicatore δ, la percentuale complessiva di *outlier* nel database: si vede chiaramente come valori alti δ facciano sì che un numero molto ridotto di valori sia identificato come *outlier*, mentre per valori bassi vengono probabilmente eliminati anche valori non eccessivamente elevati. È da notare, inoltre, che gli *outlier* non sono uniformemente distribuiti all'interno del database, bensì risultano concentrati soprattutto in alcune serie simboliche. Ad esempio, la serie simbolica identificata con il codice PT6, associata ad un'area pari a 114'294 kmq, presenta una percentuale di *outlier* pari al 27%. Tuttavia, non vi è proporzionalità tra l'estensione (come anche il numero di abitanti) e il numero di *outlier*, poiché anche serie simboliche rappresentative di piccole aree presentano percentuali di *outlier* particolarmente elevate.

4.2.3 Aggregazione e ripartizione delle serie

Le Figure 16 e 17 mostrano i pattern medi annui di consumo mensile separatamente per i due dataset di calibrazione (73 serie simboliche) e validazione (33 serie simboliche). Come si vede, nel complesso (parte

bassa delle figure) i valori di consumo si muovono nello stesso range per entrambi i database, che risultano dunque sufficientemente omogenei.

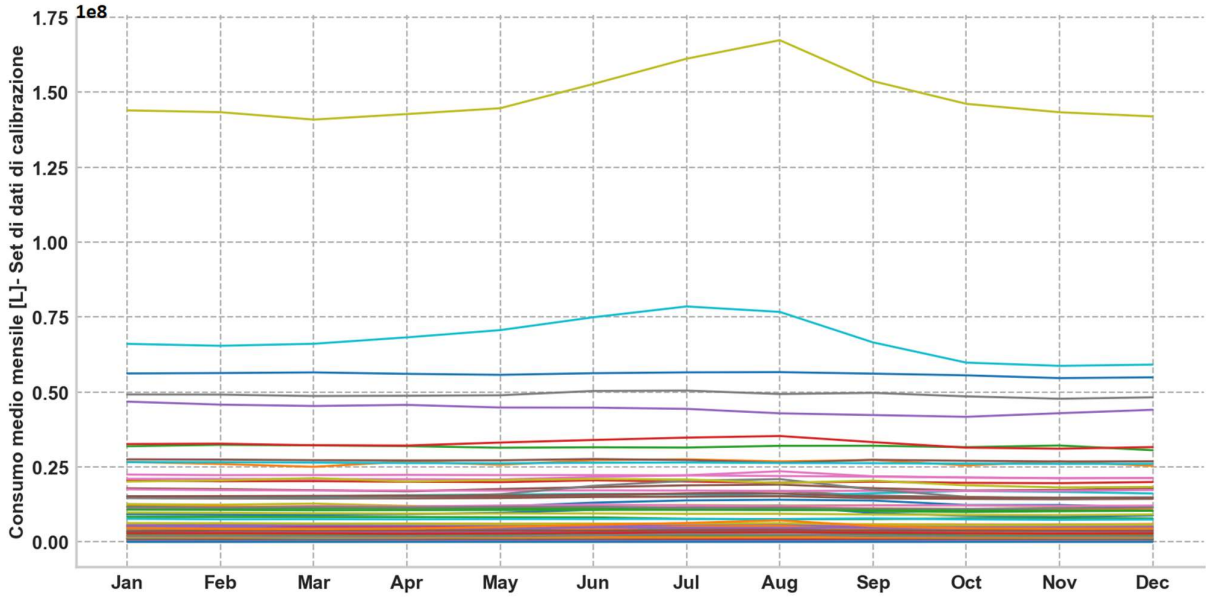

Figura 16. Pattern medi annui di consumo mensile per le serie simboliche del dataset di calibrazione.

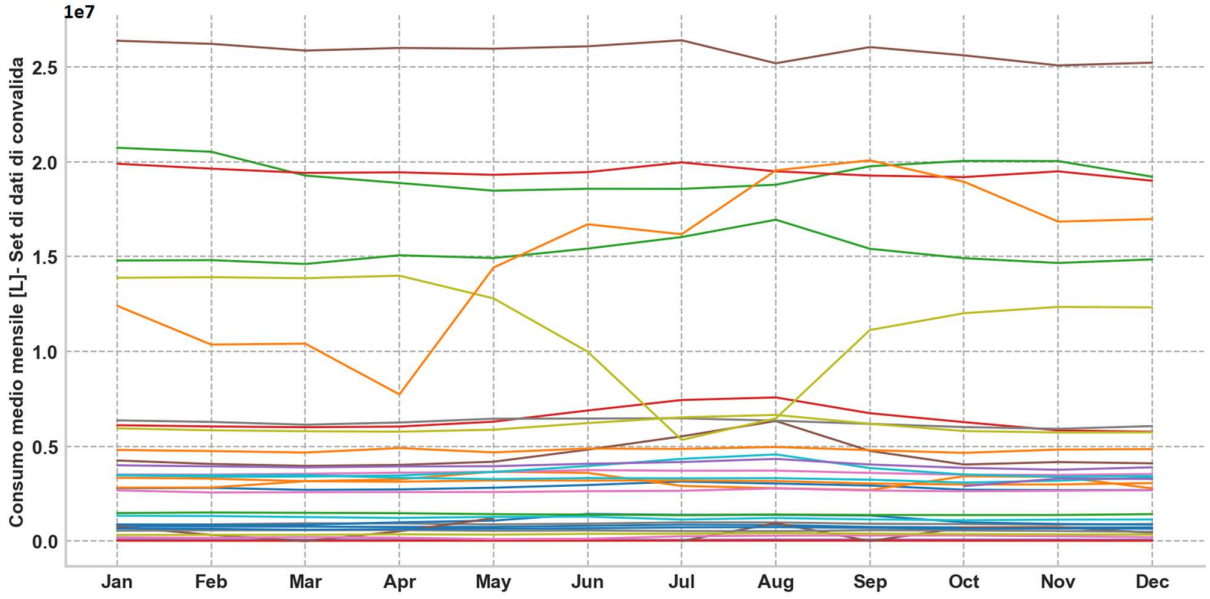

Figura 17. Pattern medi annui di consumo mensile per le serie simboliche del dataset di validazione.

4.3 Clustering

L'obiettivo principale dell'analisi di clustering è quello di individuare possibili pattern di consumo annuale utilizzando serie temporali aggregate mensilmente. La preferenza per l'aggregazione mensile rispetto a quella giornaliera è giustificata dalla sua resistenza ai comportamenti giornalieri, volatili, dei consumatori. Nello specifico, l'oggetto di cui ricercare eventuali similitudini tra le serie simboliche è il pattern medio annuo (quali quelli rappresentati nelle Figure 8, 16 e 17). Durante il calcolo dei pattern medi annui da utilizzare in input al clustering, ci si è resi conto che per alcune serie simboliche il mese di Agosto risultava spesso assente, in quanto ritenuto *outlier*; di conseguenza, per garantire omogeneità e comparabilità – presupponendo inoltre che i consumi estivi (e soprattutto di Agosto) siano fortemente influenzati dalla fluttuazione,

incognita, della popolazione per effetto del turismo – si è deciso di rimuovere, ai fini del clustering, il mese di Agosto.

Durante alcuni test preliminari, sono stati testati molteplici algoritmi di cluster, quali la Self-Organizing Map, il k-means, il dendrogramma, il metodo Gaussiano, quello spettrale e molti altri. A seguito di questa valutazione, si è deciso di concentrarsi sui due che sono risultati i più stabili, ovvero il k-means e la SOM. Quindi, questi due algoritmi sono stati eseguiti facendo variare il numero di cluster, da stabilirsi a priori, tra 2 e 9; infine, si sono rappresentati i risultati mediante tre diversi indicatori di performance (CH, DB, S), i quali sono stati interpretati in modo euristico. Dalla Figura 18, che mostra gli andamenti dei tre indici per i due algoritmi, si evince che i risultati di k-means e SOM sono tra loro simili, soprattutto per un numero di cluster inferiore a 5. Inoltre, sono emerse le seguenti considerazioni:

- Silhouette: questo indice presenta un Massimo per un numero di cluster pari a 2-3.
- Calinski-Harabasz: il trend di questo indicatore è descrescente, facendo convergere la decisione verso un numero di cluster ridotto.
- Davies-Bouldin: questo indicatore sembra essere abbastanza stabile al variare del numero di cluster, non fornendo informazioni che puntino ad una specifica soluzione. Si può però notare che sussiste un minimo (indice di un buon clustering) locale che si ritrova per SOM in corrispondenza di un numero di cluster pari a 5, per k-means pari a 3.

La Figura 19 confronta i risultati ottenuti da k-means per un numero di cluster pari a 2 (a sinistra) e a 3 (destra), mentre la Figura 20 propone lo stesso confronto per SOM. In tutte le figure, i centroidi (cioè le medie) di ciascun cluster sono confrontati con il pattern medio annuo di consumo adimensionale, rappresentato da una linea continua nera. È possibile notare che i risultati dei due algoritmi sono alquanto simili. Inoltre, i diversi cluster si distinguono tra loro non per la forma del pattern, ma soltanto per l'entità della fluttuazione stagionale. In altre parole, nel caso di un numero di cluster pari a 2, si ha un centroide con una spiccata variabilità stagionale (cluster 1, in blu) e uno con una minore variabilità (cluster 2, in arancione). Nel caso di un numero di cluster pari a 3, vi è anche un centroide intermedio.

Ai fini del presente lavoro, non si ritiene che, in definitiva, i comportamenti individuati dai diversi cluster indichino comportamenti di consumo diversi. Di conseguenza, nelle analisi successive, non si ritiene necessario effettuare le operazioni separatamente per i diversi cluster. Tuttavia, i risultati del clustering verranno presi in considerazione, nell'analisi dei risultati della correlazione, per provare a spiegare l'eterogeneità dei risultati.

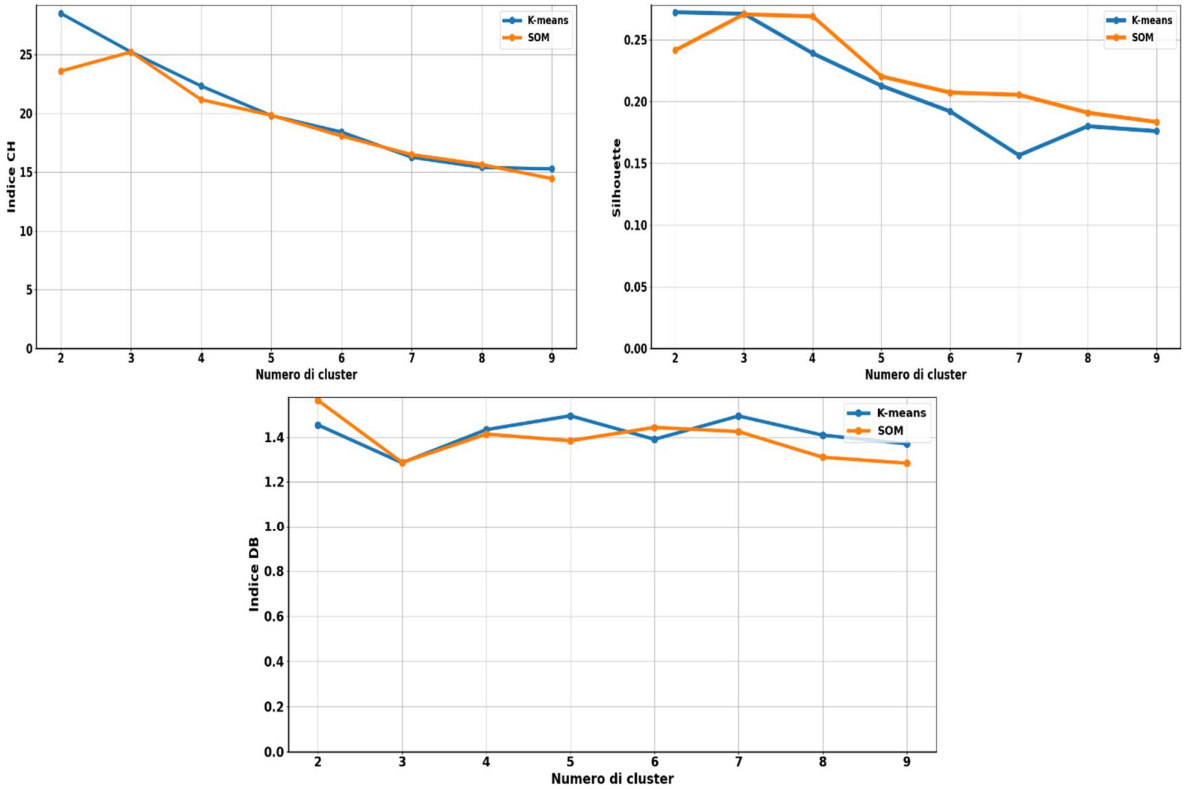

Figura 18. Indici di performance del clustering al variare del numero di cluster, per due algoritmi (k-means, in blu, e SOM, in arancione).

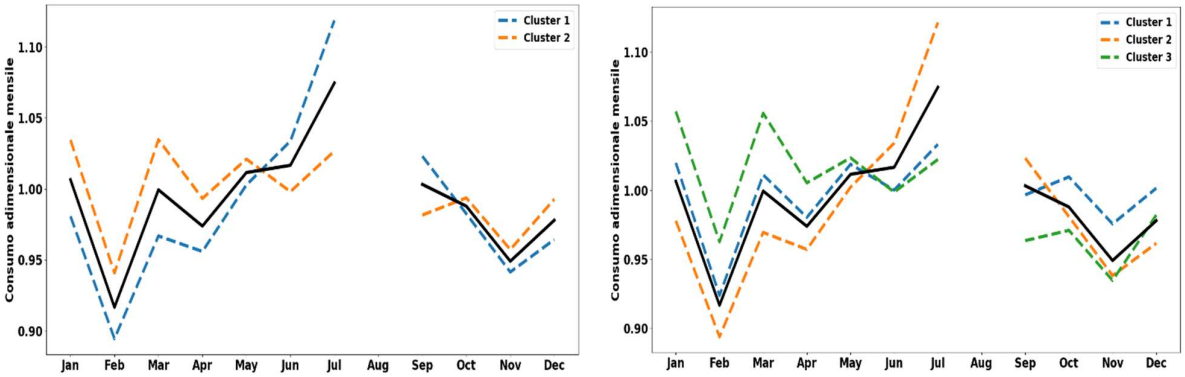

Figura 19. Pattern di consumo mensile adimensionale per k=2 (a sinistra) e k=3 (a destra), e confronto con il pattern medio dell'intero database. Risultati ottenuti mediante k-means.

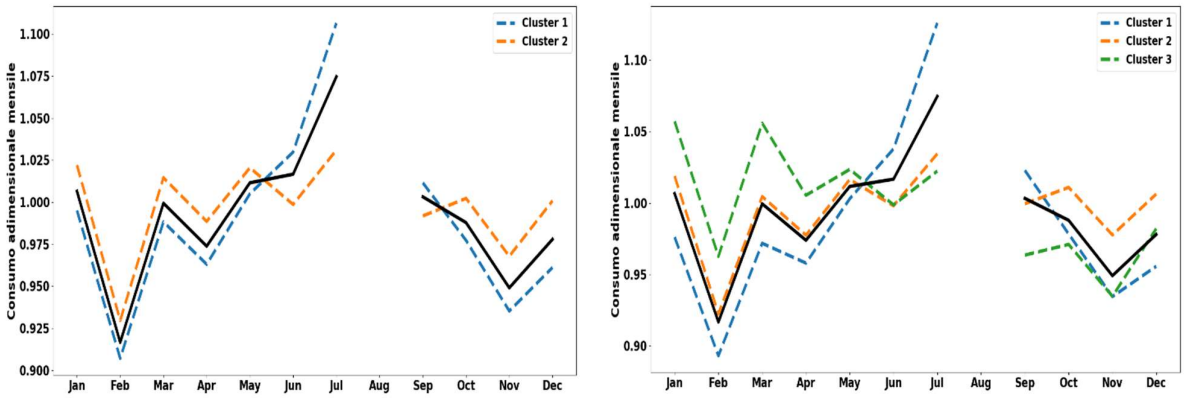

Figura 20. Pattern di consumo mensile adimensionale per k=2 (a sinistra) e k=3 (a destra), e confronto con il pattern medio dell'intero database. Risultati ottenuti mediante SOM.

4.4 Analisi di correlazione

Seguendo quanto riportato nella Nota Metodologica, l'obiettivo principale dell'analisi di correlazione è quello di individuare pattern di consumo utilizzando diverse ipotesi di dataset (grezzo e detrendizzato, giornaliero, settimanale, mensile, stagionale, annuo). Per ciascuna di queste combinazioni, e per ciascuna delle variabili esogene, in particolare climatiche, ritenute plausibili predittori, per ciascuna delle serie simboliche è separatamente calcolato il coefficiente di Pearson ρ. Non tutti i risultati sono riportati nel presente documento, poiché per un gran numero di variabili esogene le correlazioni sono risultate insignificanti; inoltre, alcune scale, quali quella settimanale, non hanno rivelato alcun valore aggiunto rispetto alla risoluzione nativa o quella più usuale mensile. Tutte le correlazioni sono comunque riportate in Appendice al presente documento. Nel seguito sono quindi proposte soltanto le correlazioni con le variabili principali (temperatura media, massima e minima giornaliera, precipitazione giornaliera) alle scale più significative. Le correlazioni proposte sono inoltre valutate secondo le indicazioni di Guilford (Tabella 4).

4.4.1 Dataset grezzo

La Figura 21 riassume i risultati dell'analisi di correlazione del dataset grezzo, alla risoluzione nativa (giornaliera) con le principali variabili climatiche (media, massimo e minimo di temperatura giornaliera, precipitazione). In generale, si manifestano correlazioni debolmente positive con la temperatura e debolmente negative con la precipitazione, in accordo con quanto ritrovato dall'analisi di letteratura. Passando all'aggregazione mensile (Figura 22) si manifesta una più spiccata correlazione positiva con la temperatura; correlazioni di entità maggiore rispetto alla scala giornaliera, ma negative, si manifestano anche per la precipitazione. In generale, ciò rivela che l'aggregazione mensile risulta più consistente rispetto al clima, mentre, probabilmente, l'aggregazione giornaliera è fin troppo influenzata dalle fluttuazioni day-by-day e dalla volatilità dei comportamenti degli utenti, mascherando i pattern sottostanti. Muovendoci infine alla scala annua (Figura 23), si rivela una contraddittoria dipendenza negativa dei consumi dalla temperatura, suggerendo che le alte temperature possano portare ad un aumento dei consumi nell'immediato, ma che abbiano effetti discordanti sul lungo periodo. Di contro, i consumi annui esibiscono ancora una correlazione negativa con la temperatura.

Le correlazioni medie (in altre parole, il valore medio del coefficiente di correlazione ρ) dell'intero database rispetto alle quattro variabili climatiche considerate nelle Figure da 21 a 23 sono mostrate in Tabella 6. Confrontando tali valori con quelli mostrati nelle figure, si nota come tali medie siano il frutto di un comportamento estremamente disomogeneo, con alcune serie che esibiscono correlazioni molto alte e molte altre serie che esibiscono correlazioni scarse; per la precipitazione, le correlazioni medie sono addirittura negative.

Tabella 6. Valore medio del coefficiente di regressione ρ (adimensionale) relativo alla regressione lineare dei consumi idrici (database grezzo) con le quattro variabili climatiche fondamentali.

Aggregazione	Temperatura massima	Temperatura media	Temperatura minima	Pioggia
Giornaliera	0.14	0.14	0.14	-0.05
Mensile	0.20	0.20	0.19	-0.13
Annuale	-0.01	-0.02	-0.02	-0.05

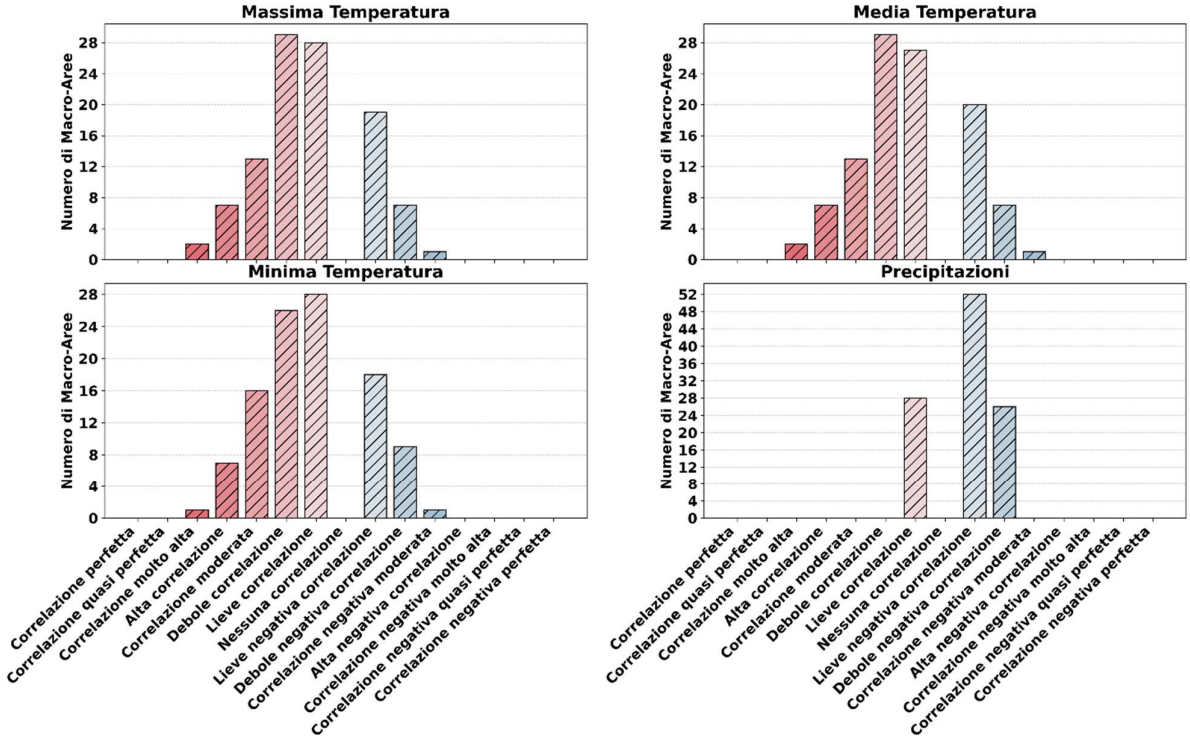

Figura 21. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database grezzo).

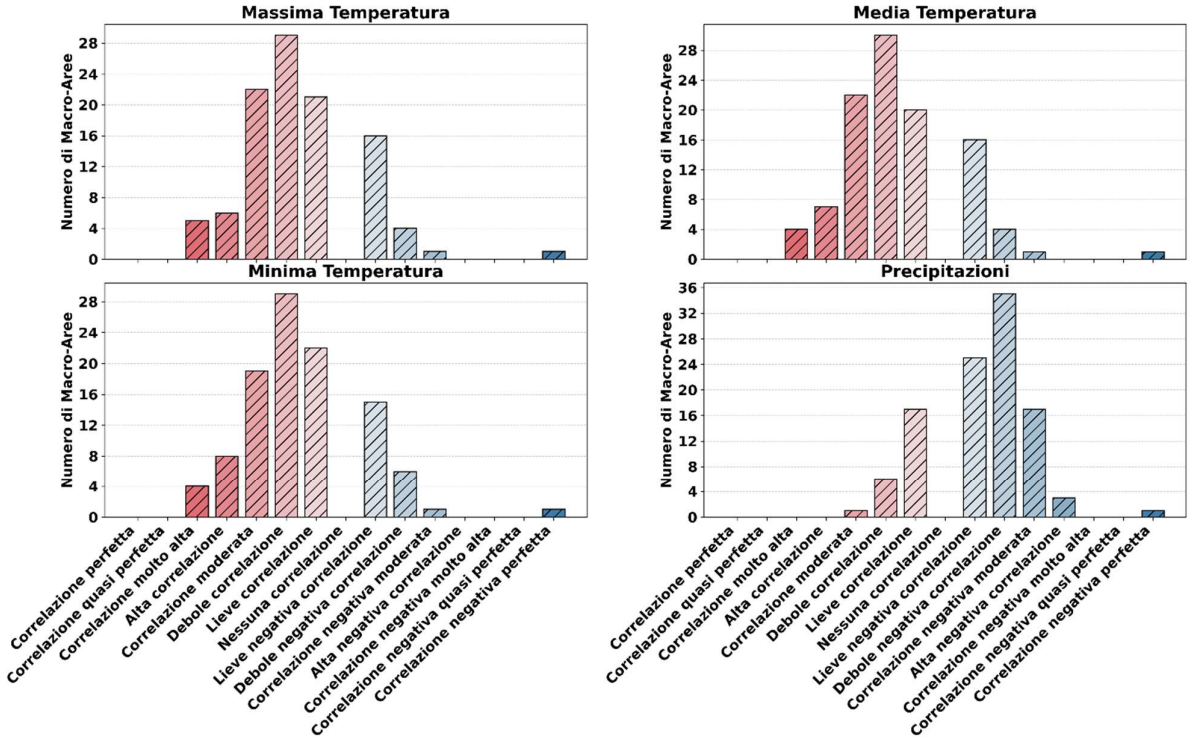

Figura 22. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione mensile, database grezzo).

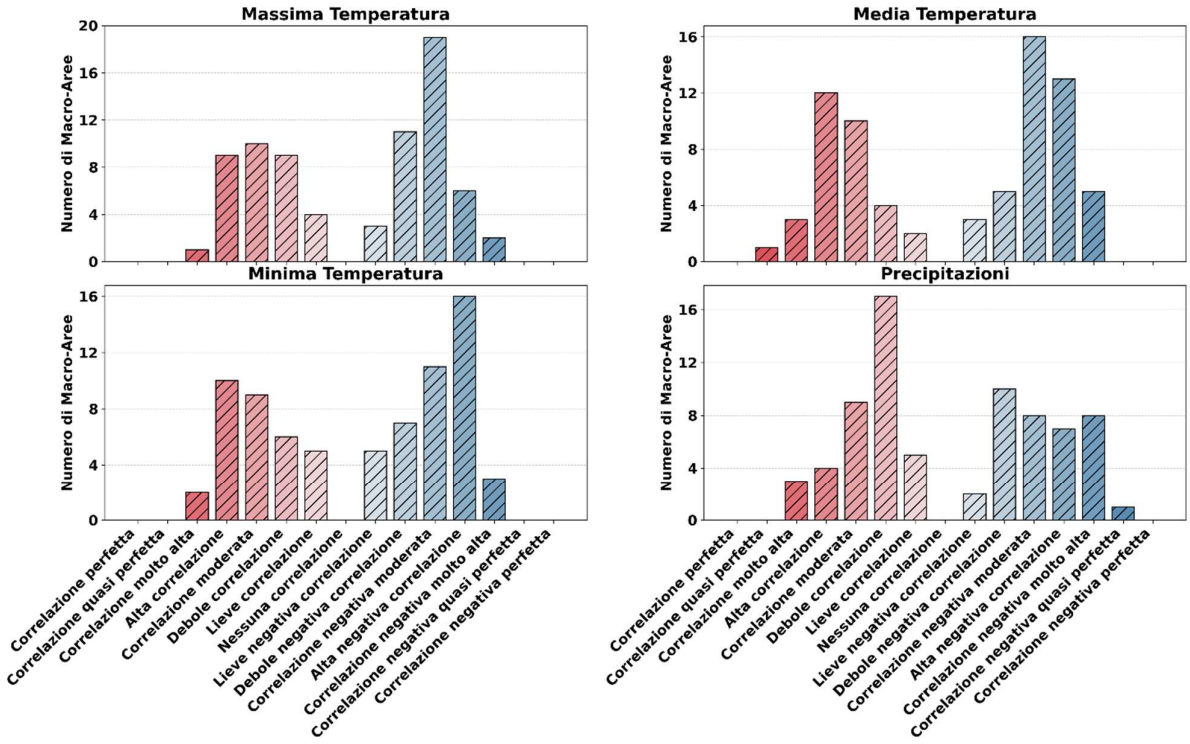

Figura 23. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione annuale, database grezzo).

Questi risultati mettono in evidenza, nell'insieme, l'importanza dell'aggregazione temporale nella comprensione delle tendenze del consumo idrico. Mentre i dati giornalieri forniscono informazioni sulle reazioni immediate alle variabili climatiche, le aggregazioni mensili e annuali offrono una prospettiva più ampia, evidenziando tendenze di lungo periodo. La costante correlazione negativa delle precipitazioni a tutti i livelli temporali e l'interessante cambiamento nelle correlazioni delle temperature da scale giornaliere e mensili a scale annuali richiedono un'approfondita indagine per elucidarne le cause.

Successivamente, sono state investigate correlazioni con variabili climatiche derivate da quelle fondamentali (Tabella 3), di cui qui si riportano soltanto alcune considerazioni. Una prima osservazione è stata l'assenza di una chiara correlazione tra il consumo giornaliero e la differenza tra la temperatura giornaliera corrente e quella del giorno precedente. Approfondendo invece l'analisi di correlazione tra consumo e temperatura media, è emersa una moderata correlazione positiva (ρ = 0.167 in media nell'intero database) durante la primavera, l'estate e l'autunno, mentre la correlazione durante l'inverno è risultata irrilevante. Ciò è potenzialmente attribuibile a un aumento delle attività all'aperto e al maggiore utilizzo di acqua durante i periodi più caldi, mentre, in inverno, i consumi sono più che altro dettati dalle abitudini di utilizzo piuttosto che dal clima. Nei giorni più secchi o con piogge minime (meno di 10 mm), è stata osservata una tendenza positiva simile, con valori medi di correlazione di ρ = 0.150 e ρ = 0.144 rispettivamente, indicando una maggiore domanda d'acqua potenzialmente per attività all'aperto.

L'analisi stagionale ha rivelato correlazioni positive durante l'autunno (ρ = 0.130) e l'estate (ρ = 0.099), in linea con un aumento del consumo durante i mesi più caldi. In modo significativo, per temperature estremamente elevate, sopra i 25°C e i 30°C, il consumo aumenta ma a tassi decrescenti (ρ = 0.097 e ρ = 0.066 rispettivamente), indicando possibilmente comportamenti adattivi nell'uso dell'acqua. Al contrario,

l'inverno ha mostrato una leggera correlazione negativa (ρ = -0.052), suggerendo una riduzione delle attività che richiedono molta acqua durante i mesi più freddi. La primavera ha presentato una leggera correlazione positiva (ρ = 0.040).

Infine, è stata valutata la relazione tra il consumo d'acqua e i giorni consecutivi senza pioggia, per la quale è stata notata una modesta correlazione media positiva (ρ = 0.102). Dei punti studiati, 69 hanno mostrato una correlazione positiva, implicando un aumento del consumo con periodi secchi prolungati. Al contrario, 37 punti hanno indicato una diminuzione del consumo in presenza di periodi prolungati senza pioggia.

4.4.2 Dataset detrendizzato

Le Figure 24 e 25 riassumono i risultati dell'analisi di correlazione del dataset detrendizzato, alla risoluzione nativa (giornaliera) e a quella mensile. Ancora una volta, i valori di consumo giornaliero (Figura 24) risultano correlati per lo più debolmente, ma positivamente, con le variabili di temperatura, e negativamente, sempre debolmente, con la precipitazione, confermando quanto emerso dall'analisi del database grezzo. Spostandosi alle scale mensili (Figura 25), la correlazione positiva con la temperatura aumenta di entità rispetto sia alla scala giornaliera, sia alla scala mensile nel caso del database grezzo. Nel complesso quindi, il dataset detrendizzato sembra catturare meglio il pattern stagionale, epurandolo delle fluttuazioni year-by-year.

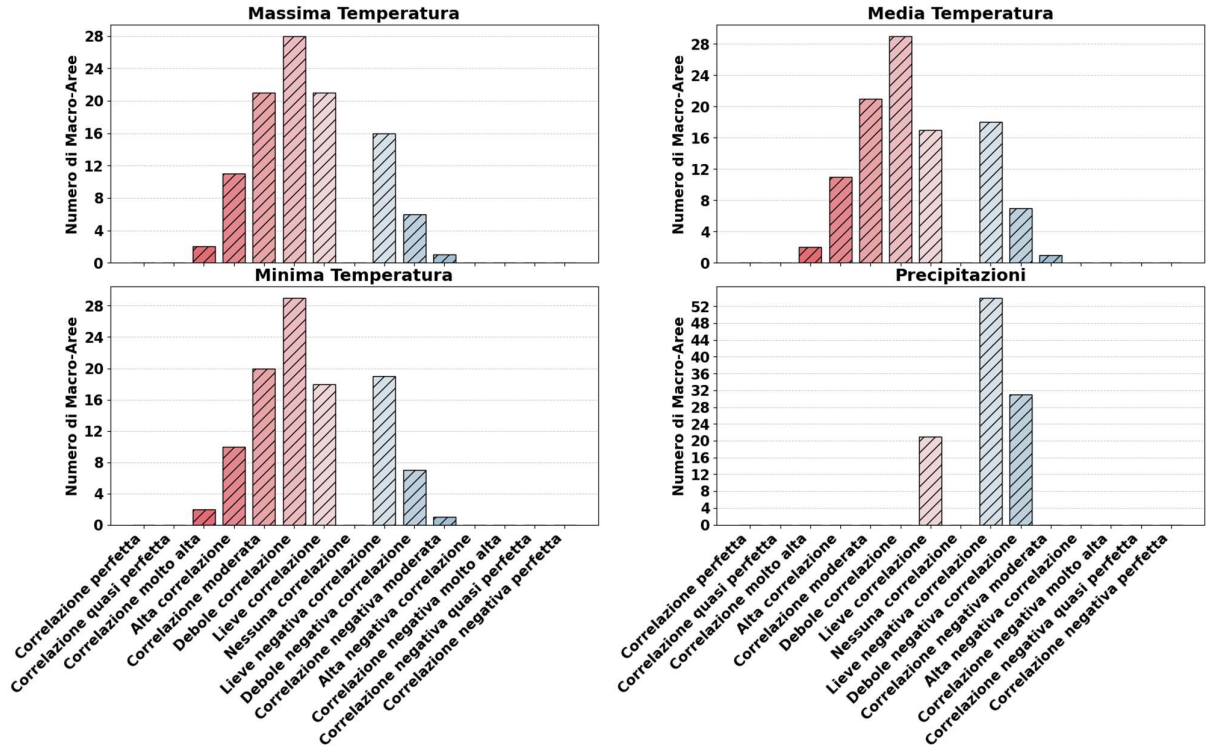

Figura 24. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato).

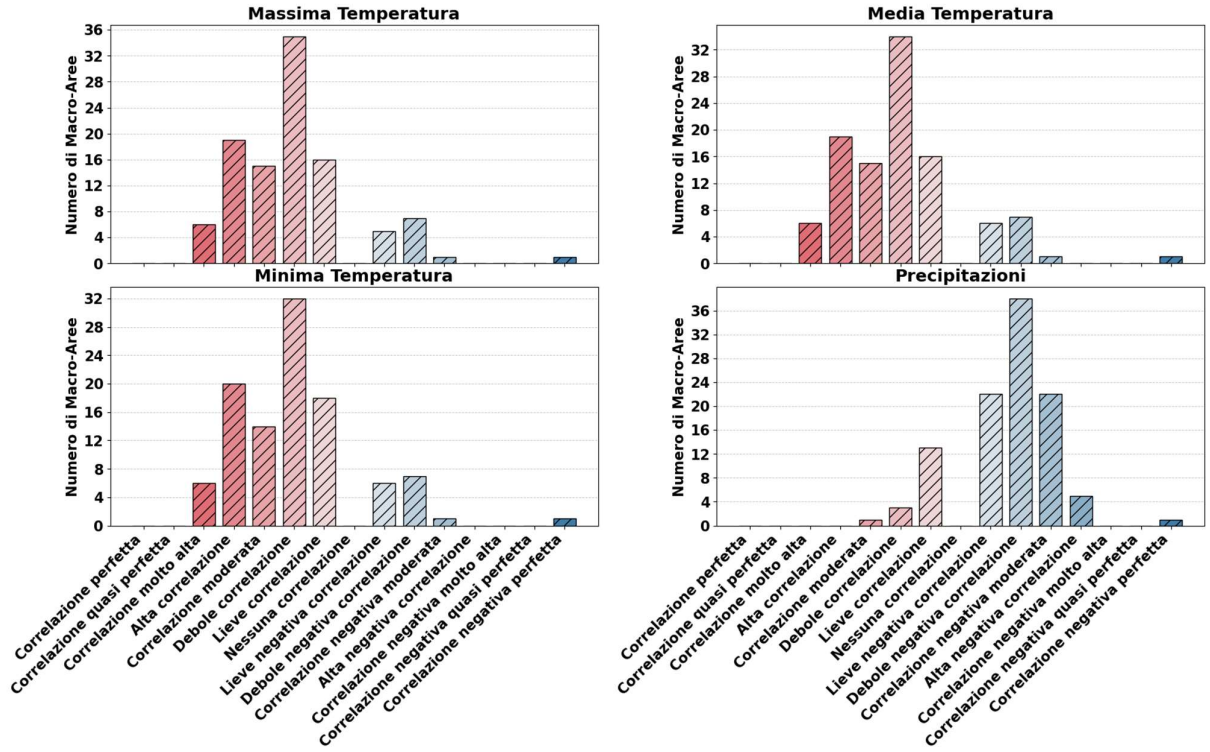

Figura 25. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione mensile, database detrendizzato).

Tabella 7. Valore medio del coefficiente di regressione ρ (adimensionale) relativo alla regressione lineare dei consumi idrici (database detrendizzato) con le quattro variabili climatiche fondamentali.

Aggregazione	Temperatura massima	Temperatura media	Temperatura minima	Pioggia
Giornaliero	0.19	0.19	0.18	-0.06
Mensile	0.27	0.27	0.27	-0.18

I valori di correlazione media sono riportati in Tabella 7, da cui si evince, come già rilevato dalle Figure 24 e 25, che i valori medi di ρ sono il frutto di una distribuzione estremamente variegata dei coefficienti di correlazione, con alcune serie che esibiscono valori elevati, molte altre serie che esibiscono valori modesti, e, per la precipitazione, la maggior parte dei valori risultano negativi.

Per quanto riguarda le variabili climatiche più complesse, alla scala giornaliera non si nota, come per il database grezzo, alcuna significativa correlazione con la differenza di temperatura tra il giorno corrente e il giorno precedente. Per quanto riguarda invece la correlazione con la temperatura media alla scala stagionale, una correlazione lievemente superiore è emersa in primavera, estate e autunno, con valori del coefficiente di correlazioni maggiori rispetto al database grezzo. Una correlazione più significativa emerge inoltre con i giorni asciutti, a conferma dei maggiori consumi per le attività all'aperto rispetto ai giorni piovosi.

Si nota infine, sebbene non visibile dalle figure presentate, che le serie simboliche (in altre parole, le macro-aree) che esibiscono i valori massimi di correlazione con la temperatura nel caso del database detrendizzato sono le stesse che esibivano correlazione massima considerando il database grezzo. La Figura 26 mostra una mappa delle macro-aree classificate, considerando tutte le classi di Guilford (Tabella 4) per ρ positivi, e raggruppando tutti i valori di ρ negativi in un'unica classe.

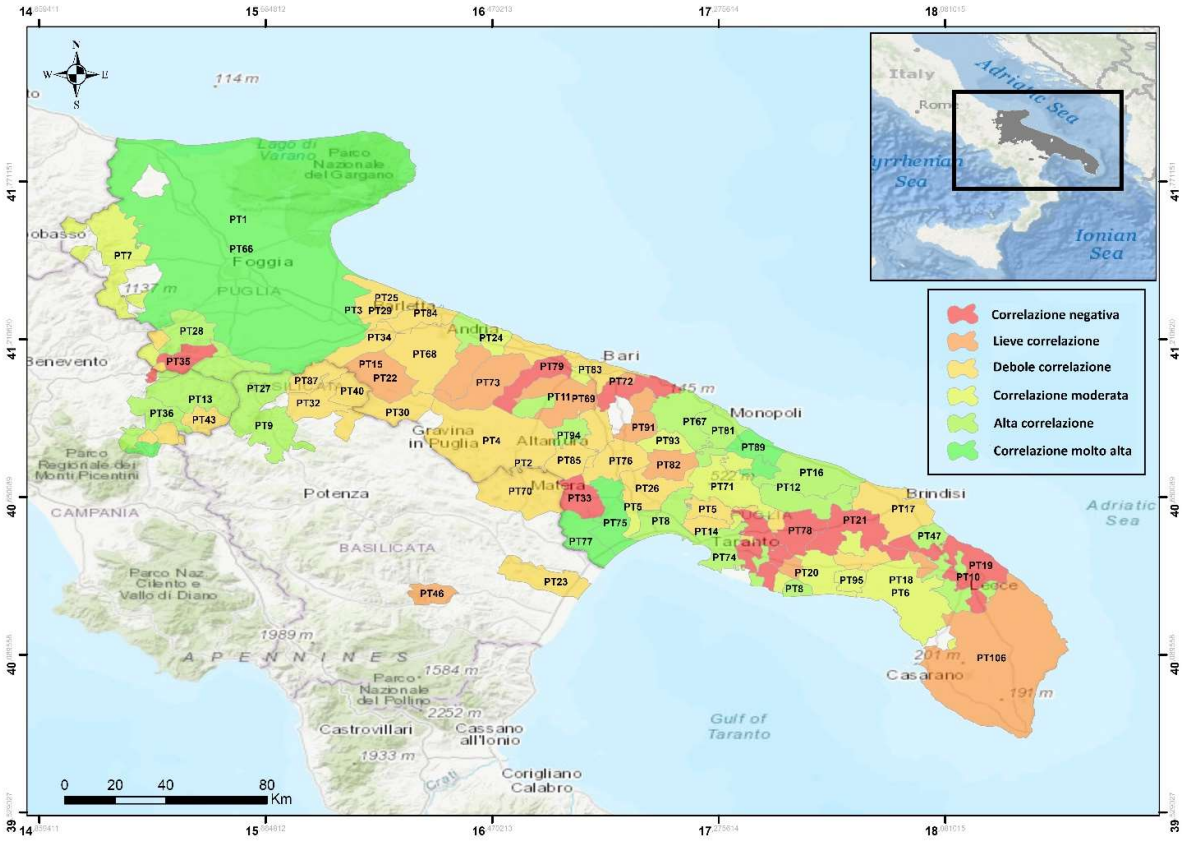

Figura 26. Classificazione delle macro-aree in base al coefficiente di correlazione (classificazione di Guilford, Tabella 4).

4.4.3 Correlazione vs. clustering

L'applicazione del clustering ha messo in evidenza essenzialmente comportamenti analoghi in termini di pattern annuo di valori mensili, con differenziazioni dipendenti dall'entità della fluttuazione stagionale, con un cluster di valori più omogenei, e un cluster in cui i valori invernali sono nettamente più bassi dei valori estivi (prendendo in considerazione la soluzione a doppio cluster ricavata da k-means). Ai fini dell'analisi di correlazione, non è stato necessario prendere in considerazione l'esistenza di cluster separati, poiché l'analisi è stata condotta serie per serie. Tuttavia, il clustering potrebbe essere utile provare a spiegare l'eterogeneità dei risultati.

La Figura 27 mostra la classificazione delle macro-aree in base all'entità e al segno della correlazione (riferita al database detrendizzato alla scala mensile con la temperatura media giornaliera) e, contemporaneamente, i due cluster individuati. In effetti, si ritrova che il cluster 1 presenta, mediamente tra le macro-aree in esso ricadenti, un coefficiente di correlazione pari a 0.43, mentre la media dei coefficienti di correlazione del cluster 2 è pari a 0.08, quindi notevolmente più bassa. Si ricorda che il cluster 1 conta 56 elementi, il cluster 2 49, dunque essi hanno in sostanza la stessa numerosità. Ciò porta a concludere che le serie che esibiscono la maggiore correlazione con il clima sono quelle per le quali le fluttuazioni stagionali sono più pronunciate; per queste aree, tuttavia, non è possibile discriminare quale quota parte della correlazione sia dovuta a un effetto dell'incremento di popolazione nel periodo turistico.

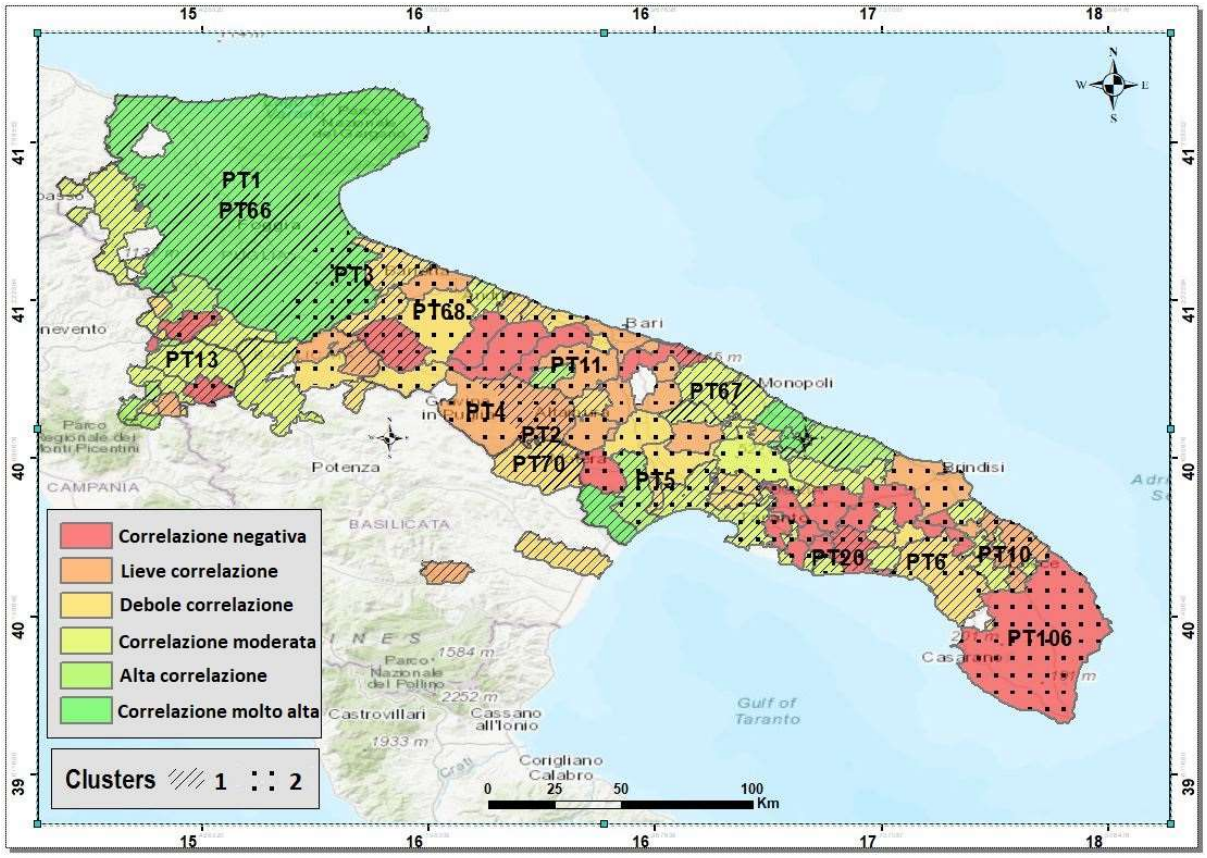

Figura 27. Corrispondenza tra classi di correlazione e cluster.

5. Impatto dei cambiamenti climatici

La forte eterogeneità dei risultati, se da un lato conferma quanto ipotizzato dall'analisi della letteratura recente in materia, non è particolarmente utile agli scopi dell'attività nell'ambito della Convenzione. Partendo dal presupposto che l'influenza del clima sui consumi idropotabili possa essere mascherata da effetti locali, si è quindi scelto, per concretizzare le analisi, di campionare dall'intero dataset solo quelle serie simboliche per le quali il grado di correlazione con il clima risulta significativo. Si proverà quindi a quantificare gli impatti del cambiamento climatico solo sulle macro-aree corrispondenti a queste serie simboliche.

5.1 Analisi di correlazione per le serie simboliche a correlazione massima

La Tabella 9 mostra le 8 serie simboliche, su un totale di 106[1], per le quali risulta massima la correlazione tra i consumi idrici (database grezzo) e la temperatura media giornaliera per una scala di aggregazione mensile. Come si vede dalla tabella, le caratteristiche delle macro-aree associate a tali consumi sono piuttosto eterogenee, con valori importanti di correlazione associati tanto a macro-aree estese e popolose, tanto ad aree piccole. Le macro-aree in Tabella 9 sono inoltre rappresentate in Figura 28, dove, per le sole macro-aree a servizio di un solo Comune, si è anche indicato il nome; nella stessa figura è inoltre indicato con una colorazione diversificata il numero di Municipalità servite.

[1] Si noti che la prima ipotesi è stata di analizzare le "migliori" 10 serie. Tuttavia, già per la nona e la decima le correlazioni risultavano non più significative per gli scopi della Convenzione.

Tabella 9. Serie simboliche che presentano i valori più alti del coefficiente di correlazione (database detrendizzato, aggregazione mensile, temperatura media giornaliera).

ID serie simbolica	Numero di abitanti	Area (kmq)	Numero di Municipalità servite	Municipalità servite	Coefficiente di correlazione
PT63	1435	15.3	1	Scampitella	0.66
PT102	3797	41.0	1	Caposele	0.69
PT98	12078	54.3	1	Cisternino	0.81
PT89	38667	129.2	1	Fasano	0.91
PT77	22146	188.2	1	Ginosa	0.77
PT75	17393	240.9	1	Castellaneta	0.72
PT16	47861	329.5	2	Ostuni, Carovigno	0.64
PT66	595370	5564.4	34	Territori nel foggiano	0.75

La serie simbolica che esibisce il maggiore coefficiente di correlazione è quella identificata dal codice PT89, a servizio del Comune di Fasano. Considerazioni effettuate su tale serie, che risulta anche avere ottime caratteristiche di qualità (in termini di continuità e completezza), hanno guidato le scelte riguardanti il modello di regressione, che sono state quindi applicate alle altre serie in Tabella 9.

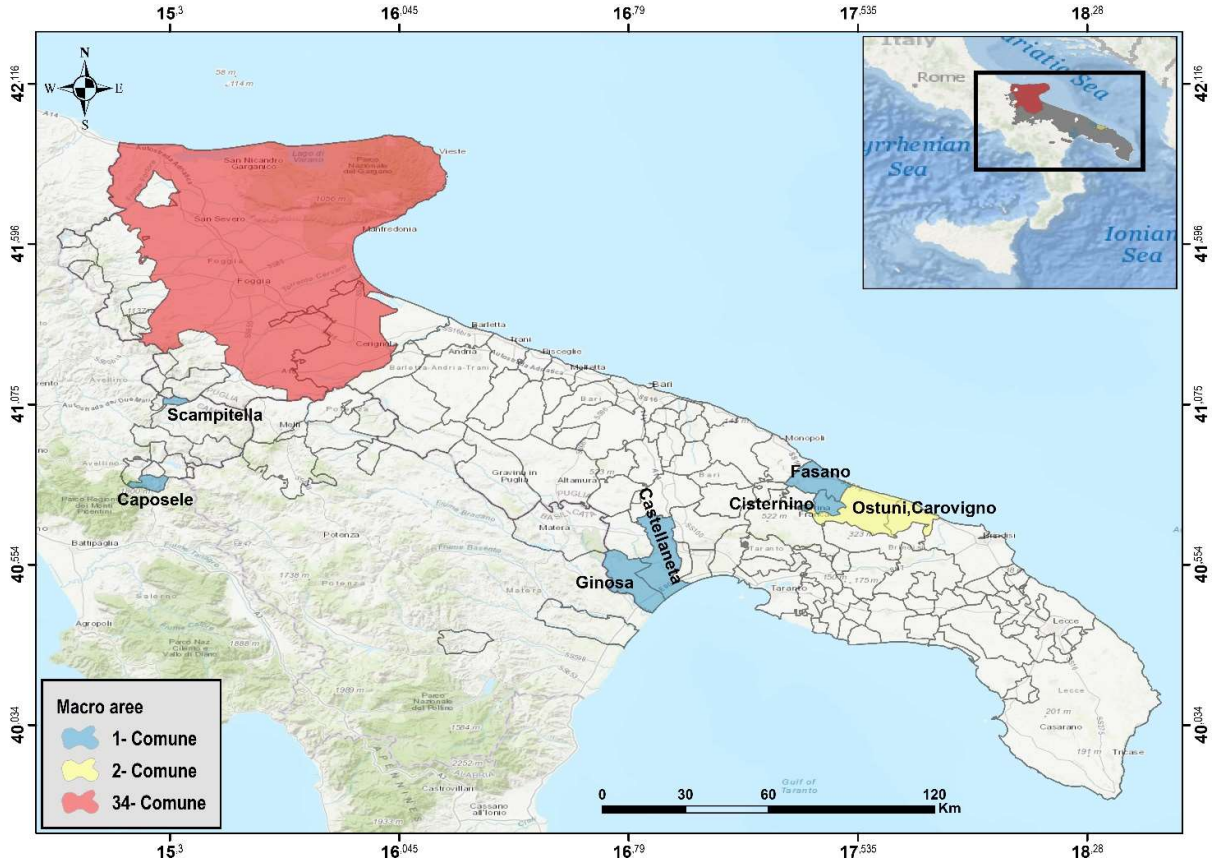

Figura 28. Mappa delle macro-aree corrispondenti alle serie simboliche in Tabella 7, con indicazione del numero di Municipalità servite.

5.2 Analisi di correlazione per la serie simbolica "migliore"

La Figura 29 mostra la serie temporale dei dati per questa macro-area, che corrisponde al Comune di Fasano. Si può notare, in particolare, come tale serie presenti un grado di completezza e continuità molto alto, senza "vuoti" all'interno della serie; inoltre, risulta particolarmente spiccata la corrispondenza stagionale con la fluttuazione delle temperature. La Figura 30 mostra invece la spiccata relazione tra i consumi mensili (in

milioni di Litri, per facilitare la visualizzazione) e la media mensile delle temperature medie giornaliere: si nota, in particolare, che la correlazione tra le due variabili è spiegabile mediante una relazione lineare soprattutto per valori di temperatura superiori ai 10°C. Ciò potrebbe spiegarsi, come già menzionato in precedenza, con il fatto che nella stagione invernale i consumi sono per lo più dettati dalle abitudini sociali, e non dal clima, che invece mostra un maggiore effetto nelle stagioni più calde. Limitando dunque la correlazione ai valori di temperatura maggiori o uguali a 10°C, si può stabilire una relazione del tipo:

$$Q_m = a + b \cdot T_m \tag{8}$$

dove Q_m è il consumo aggregato alla scala mensile, in milioni di L, e T_m è la media mensile della temperatura media giornaliera, e deve essere $T_m \geq 10$°C; a e b sono invece i coefficienti della regressione lineare, da ricavarsi mediante, ad esempio, l'algoritmo dei minimi quadrati o *Ordinary Least Squares* (sono comunque stati testati altri algoritmi, quali *Polynomial*, *Decision Tree Regressor*, *RandomForest*, XGBOOST, FBPROPHET e LSTM, senza rilevare particolari cambiamenti). OLS ha fornito un valore di a pari a 114.67 Milioni di L, e un valore di b pari a 30.83 Milioni di L/°C, con un valore del coefficiente di determinazione R^2 pari a 0.825, indice di una correlazione molto elevata. La retta di regressione è parimenti rappresentata in Figura 30.

Si noti che, sebbene non riportato nel documento, è stata tentata una regressione lineare multivariata aggiungendo come seconda variabile esogena il numero di giorni piovosi nel mese. Ciò ha consentito di aumentare il coefficiente di determinazione da 0.825 a 0.836, quindi con un miglioramento del tutto trascurabile.

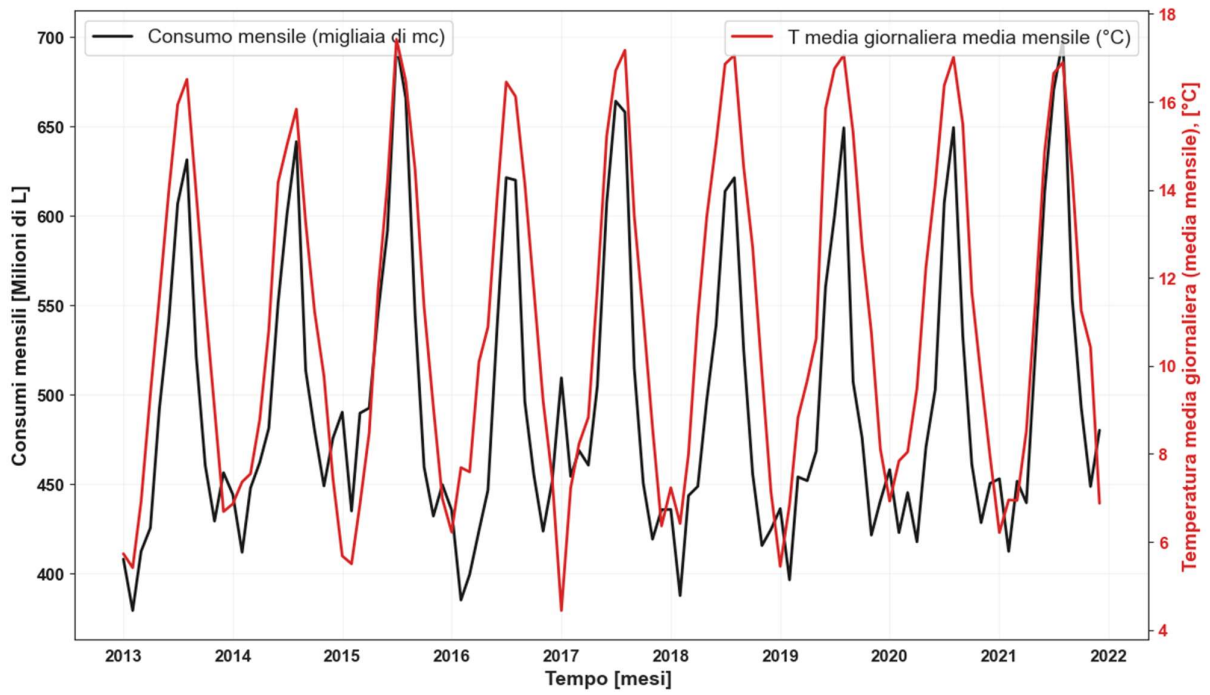

Figura 29. Andamento dei consumi mensili e delle temperature mensili nel periodo monitorato per la serie "migliore".

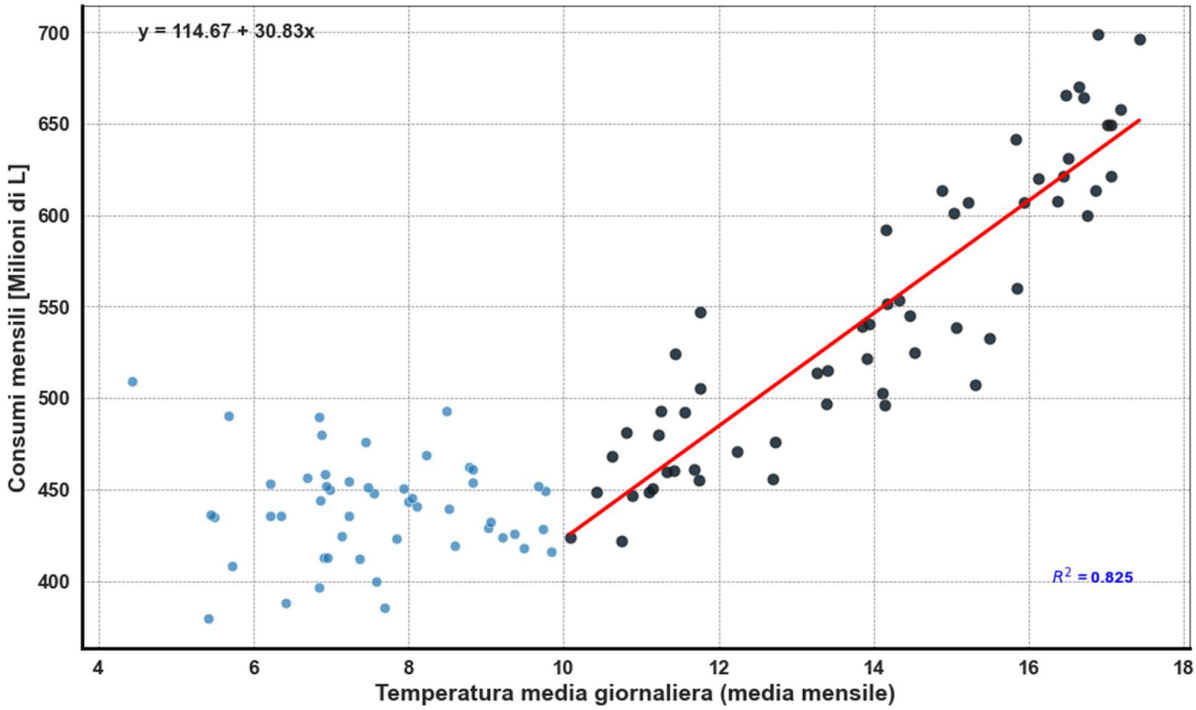

Figura 30. Correlazione tra consumi mensili e temperature mensili nel periodo monitorato per la serie "migliore" e caratteristiche della regressione lineare.

5.3 Analisi di correlazione per tutte le serie simboliche

La Tabella 10 riporta, per ciascuna delle serie simboliche in Tabella 9, corrispondenti alle macro-aree in Figura 28, i valori di *a*, *b* ed R^2 della regressione lineare. Come si vede dalla tabella, i valori del coefficiente di determinazione indicano una regressione lineare piuttosto efficiente; tuttavia, ricordando che si tratta delle serie che già esibivano la correlazione migliore con la temperatura media giornaliera, è eclatante notare come soltanto le prime 5 presentino un valore del coefficiente di determinazione maggiore del 50%.

Tabella 10. Valori notevoli della regressione lineare in Eq. 8 per le serie simboliche in Tabella 9.

ID serie simbolica	*a* [Milioni di L]	*b* [Milioni di L/°C]	R^2
PT89	30.83	114.67	0.83
PT98	2.10	75.76	0.66
PT75	3.45	74.40	0.59
PT77	52.35	3657.16	0.56
PT102	3.39	44.66	0.52
PT63	3.90	132.61	0.48
PT66	0.16	8.48	0.43
PT16	6.58	183.65	0.40

5.4 Analisi degli effetti del cambiamento climatico

Mediante l'Eq. 8, calibrata per ogni serie, i cui risultati sono riportati in Tabella 10, è possibile calcolare per ognuna delle otto macro-aree un valore notevole di consumo, ovvero il consumo medio annuo (come media dei consumi mensili tra tutti gli anni di osservazione) e il consumo massimo annuo (come media del massimo annuale di consumo mensile tra tutti gli anni di osservazione). Essi verranno nel seguito indicati come *QM* e *QMAX*. È inoltre utile calcolare *QMIN*, come media del minimo annuale di consumo mensile tra tutti gli anni di osservazione). Nel seguito vengono elencati tutti gli scenari in cui è possibile valutare, per ciascuna serie

simbolica, QM e QMAX: naturalmente, ci si riferisce alle possibilità di valutazione della variabile esplicativa T_m in Eq. 8:

1. Per il periodo di osservazione 2013-2021, la temperatura è fornita da E-OBS. QM e QMAX sono calcolati direttamente a partire dai dati di consumo osservati.
2. Per il periodo di riferimento 1981-2010, la temperatura è fornita da E-OBS. Per ogni mese di ogni anno del periodo di riferimento, l'Eq. 8 permette di passare dalla serie di temperatura alla serie di consumo mensile, con il vincolo che, quando la temperatura mensile è inferiore a 10°C, il consumo viene posto pari a QMIN. Quindi QM e QMAX vengono calcolati a partire da questa nuova serie mensile di consumo.
3. Per il periodo di riferimento 1981-2010, la temperatura è fornita da ciascuna delle 14 catene di simulazione climatica in Tabella 1. Per ogni mese di ogni anno del periodo di riferimento, l'Eq. 8 permette di passare dalla serie di temperatura alla serie di consumo mensile, con il vincolo che, quando la temperatura mensile è inferiore a 10°C il consumo viene posto pari a QMIN. Quindi QM e QMAX vengono calcolati a partire da questa nuova serie mensile di consumo.
4. Per l'orizzonte futuro 2021-2050, la temperatura è fornita da ciascuna delle 14 catene di simulazione climatica in Tabella 1. Per ogni mese di ogni anno dell'orizzonte futuro, l'Eq. 8 permette di passare dalla serie di temperatura alla serie di consumo mensile, con il vincolo che se la temperatura mensile è inferiore a 10°C il consumo viene posto pari a QMIN. Quindi QM e QMAX vengono calcolati a partire da questa nuova serie mensile di consumo.

È possibile passare da (1) a (2) ipotizzando che la regressione calibrata sul periodo di osservazione rimanga valida anche nel periodo di riferimento.

Al punto (3), dati i 14 risultati diversi per QM e QMAX, è possibile calcolare l'ensemble mean EM, la deviazione standard DS e il coefficiente di variazione CV, come rapporto tra DS e EM. La discrepanza (variazione percentuale) tra EM e il valore individuato al punto (2) rappresenta il bias medio dell'ensemble di modelli, ovvero l'errore complessivo commesso dall'insieme di catene modellistiche considerate nella rappresentazione dei consumi osservati. Similmente, il bias di ciascuna catena modellistica può essere calcolato confrontando ciascun valore di QM o QMAX con i valori stimati al punto (2). Il valore di DS rappresenta invece la variabilità data dall'insieme dei modelli.

Al punto (4), dati i 14 risultati diversi per QM e QMAX, è possibile calcolare per ciascuna delle 14 catene modellistiche la variazione percentuale rispetto al valore stimato al punto (3), e, di queste 14 variazioni percentuali, è possibile calcolare l'ensemble mean EM, la deviazione standard DS e il coefficiente di variazione CV, come rapporto tra DS e EM. Il valore di EM rappresenta la variazione media attesa in futuro sui consumi idropotabili per effetto del cambiamento climatico.

La Tabella 11 mostra il confronto, per ognuna delle serie simboliche in Tabella 9, tra i valori di consumo "osservati", cioè direttamente estrapolati dai dati forniti nel periodo 2013-2021, e quelli "di riferimento", ricavati dall' Eq. 8 mediante le temperature fornite dal dataset E-OBS sul periodo di riferimento 1981-2010, sia per QM sia per QMAX. Lo scarto percentuale tra le due misure, mostrato in Tabella 11, è dovuto sia all'errore indotto dall'aver utilizzato il modello di regressione per inferire sui consumi, sia al diverso intervallo temporale cui i due valori si riferiscono. Lo scarto percentuale mostra una distribuzione alquanto eterogenea, con valori mediamente più alti (scarto medio tra le serie pari a 6.3%) per QM e mediamente più bassi (scarto medio tra le serie pari a 4.1%) per QMAX. Gli scarti su QM raggiungono valori positivi localmente più elevati, pari a +21.8% per la serie PT102, mentre gli scarti su QMAX attingono valori localmente più elevati, pari a -

25.8% per la serie PT89 (tale serie attinge valori di scarto negativi anche per QM). In generale, i valori di riferimento sovrastimano tra circa il +3% e circa il +22% i consumi medi annui rispetto ai valori osservati, e tra circa il +5% e circa il +19% i consumi massimi annuali rispetto ai valori osservati. Ciò accade per tutte le serie tranne che per PT102, dove invece i valori di riferimento sottostimano quelli osservati del circa -14% per QM e -26% per QMAX, e per la serie PT77, dove i valori di riferimento sottostimano quelli osservati di QMAX di circa il -4%. Si noti, infine, che in Tabella 11 le macro-aree sono elencate in conformità con la Tabella 10 dalla serie con correlazione maggiore a quella con correlazione minore.

Tabella 11. Valori osservati e di riferimento, in milioni di L, e scarto percentuale tra le due misure, per le serie simboliche in Tabella 9.

Serie simbolica	QM			QMAX		
	Osservazione (2013-2021)	Riferimento (1981-2010)	scarto %	Osservazione (2013-2021)	Riferimento (1981-2010)	scarto %
PT89	496.5	575.9	-13.8	652.6	879.2	-25.8
PT98	112.7	105.7	6.6	142.0	127.9	11.0
PT75	95.2	91.2	4.4	125.9	131.7	-4.4
PT77	137.6	123.5	11.5	196.3	164.8	19.1
PT102	195.4	160.4	21.8	238.6	213.6	11.7
PT63	11.0	9.9	11.8	13.2	12.1	9.1
PT66	4519.2	4379.0	3.2	5191.0	4953.7	4.8
PT16	294.1	280.9	4.7	375.1	348.8	7.5

Analizzando i valori sia osservati, sia di riferimento, si nota come la serie simbolica PT66 presenti un consumo sia medio sia massimo molto maggiore delle altre serie; ciò è dovuto certamente al fatto che essa si riferisce ad una macro-area molto estesa e molto più popolosa rispetto alle altre. Allo stesso modo, la serie PT63 è associata al numero di abitanti minore, e, di conseguenza, a consumi idrici di minima entità. Le Figure 31 e 32 mostrano la correlazione di QM e QMAX con il numero di abitanti serviti e la superficie della macro-area (Tabella 9). In entrambe le figure gli assi sono rappresentati in scala logaritmica per facilitare la visualizzazione, resa difficoltosa dalla presenza della serie PT66 con valori notevolmente più alti sia di consumo, sia di estensione, sia di abitanti. Si nota una correlazione generalmente positiva, sebbene non perfetta, tra le varie grandezze.

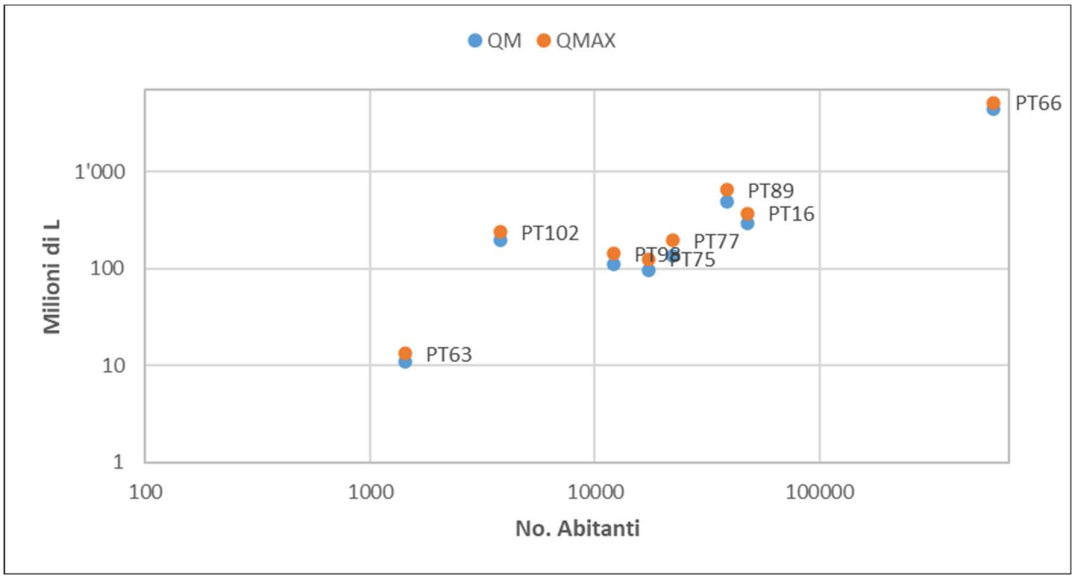

Figura 31. Correlazione tra QM, QMAX e numero di abitanti serviti per le serie simboliche in Tabella 9.

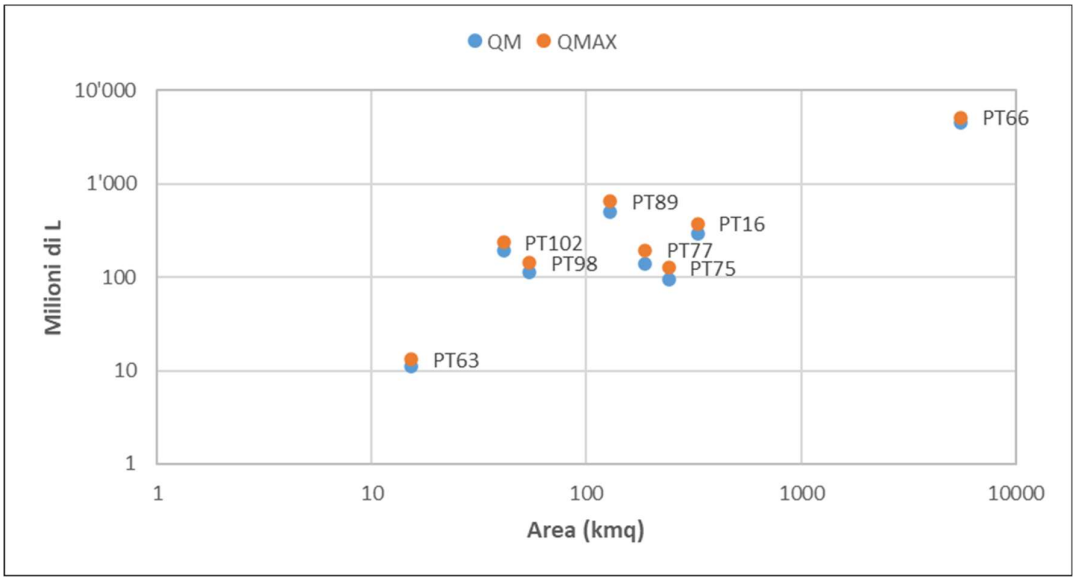

Figura 32. Correlazione tra QM, QMAX e estensione della macro-area per le serie simboliche in Tabella 9.

I risultati della procedura per la valutazione degli impatti del cambiamento climatico sono mostrati nella Tabella 12 per QM, nella Tabella 13 per QMAX. In entrambi i casi, le serie simboliche sono elencate analogamente alle Tabelle 10 e 11, dalla serie con correlazione migliore alla serie con correlazione peggiore: di conseguenza, i risultati sono da considerarsi sempre meno robusti dall'alto verso il basso nelle tabelle. È interessante notare come le serie con i valori maggiori di consumo siano quelle (relativamente) meno robuste, mentre la "miglior serie" è quella più piccola e meno abitata.

Tabella 12. Valori di riferimento, in milioni di L, e variazioni percentuali EM attese (con, inoltre, indicazione della deviazione standard DS, in %, e del coefficiente di variazione CV, adimensionale) per il consumo idropotabile medio annuo QM per i tre scenari di concentrazione, per ciascuna serie simbolica in Tabella 9.

Serie simbolica	Riferimento (2013-2021)	RCP2.6 (2021-2050)			RCP4.5 (2021-2050)			RCP8.5 (2021-2050)		
		EM (%)	±DS (%)	CV	EM (%)	±DS (%)	CV	EM (%)	±DS (%)	CV
PT89	496.5	5.0	2.0	0.4	5.8	1.3	0.2	7.2	1.3	0.2
PT98	112.7	2.3	0.8	0.3	2.9	0.8	0.3	3.5	0.9	0.3
PT75	95.2	5.2	1.7	0.3	5.4	1.8	0.3	6.7	1.9	0.3
PT77	137.6	3.9	1.3	0.3	4.0	1.3	0.3	5.0	1.3	0.3
PT102	195.4	4.1	1.7	0.4	4.2	1.7	0.4	5.2	1.9	0.4
PT63	11.0	2.5	1.1	0.4	2.6	1.2	0.4	3.4	1.3	0.4
PT66	4519.2	1.1	0.5	0.5	1.2	0.6	0.5	1.7	0.2	0.1
PT16	294.1	3.1	1.1	0.4	3.5	0.8	0.2	4.1	1.1	0.3

Tabella 13. Valori di riferimento, in milioni di L, e variazioni percentuali EM attese (con, inoltre, indicazione della deviazione standard DS, in %, e del coefficiente di variazione CV, adimensionale) per il consumo idropotabile massimo annuale QM per i tre scenari di concentrazione, per ciascuna serie simbolica in Tabella 9.

Serie simbolica	Riferimento (2013-2021)	RCP2.6 (2021-2050)			RCP4.5 (2021-2050)			RCP8.5 (2021-2050)		
		EM (%)	±DS (%)	CV	EM (%)	±DS (%)	CV	EM (%)	±DS (%)	CV
PT89	652.6	4.8	1.5	0.3	5.1	1.2	0.2	5.7	1.2	0.2
PT98	142.0	2.3	0.7	0.3	2.4	0.6	0.3	2.7	0.6	0.2
PT75	125.9	3.7	1.1	0.3	3.9	0.9	0.2	4.3	0.9	0.2
PT77	196.3	3.0	0.9	0.3	3.2	0.7	0.2	3.5	0.8	0.2
PT102	238.6	2.6	0.8	0.3	2.9	0.7	0.2	3.1	0.7	0.2
PT63	13.2	2.0	0.6	0.3	2.2	0.5	0.2	2.4	0.5	0.2
PT66	5191.0	1.5	0.4	0.3	1.6	0.4	0.2	1.7	0.4	0.2
PT16	375.1	2.6	0.8	0.3	2.8	0.7	0.2	3.1	0.7	0.2

Per quanto concerne le variazioni attese, si nota immediatamente che tutti i valori di EM sono molto bassi (mai superiori all'8% per QM, al 6% per QMAX) e sono caratterizzati da una incertezza molto ridotta, con un coefficiente di variazione CV mai superiore a 0.6 per QM, a 0.4 per QMAX. Mediamente tra le serie (non mostrato in tabella), la variazione media di QM e QMAX aumenta spostandosi da RCP 2.6 (+3.4% per QM, +2.8% per QMAX) a RCP 4.5 (+3.7% per QM, +3.0% per QMAX) e infine a RCP 8.5 (+4.6% per QM, +3.3% per QMAX). Localmente, i valori maggiori di variazione per QM sono attinti da PT89 (+7.2%) e da PT75 (+6.7%) sotto RCP 8.5, e le stesse serie attingono anche i valori maggiori di variazione per QMAX (+5.7% per PT89 e +4.3% per PT75) sempre sotto RCP 8.5. I valori minimi sono invece attinti dalla serie PT66 sotto RCP 4.5 (+1.1% per QM, +1.5% per QMAX); la stessa serie presenta i valori più bassi anche per gli altri scenari.

Si noti che, nelle Tabelle 12 e 13, si è preferito utilizzare come valore di riferimento quello fornito direttamente dalle osservazioni, che si riferisce a un periodo di riferimento diverso a quello di default della Convenzione (2013-2021 vs. 1981-2010). Sebbene più breve e dunque i valori di riferimento medi meno stabili, si è preferito tale arco temporale poiché gli scarti mostrati nella Tabella 11 sono talvolta rilevanti (e, per quanto già menzionato, dovuti in larga parte alla diversa finestra temporale), e necessitano di ulteriori approfondimenti.

6. Discussione

L'analisi di correlazione finora presentata evidenzia che non esiste una chiara e netta dipendenza dei consumi idropotabili dai fattori climatici, e soprattutto che esiste una forte eterogeneità dei risultati in funzione, presumibilmente, di una pletora di variabili di tipo non solo climatico, ma anche sociale ed economico. Inoltre, anche in presenza di una correlazione significativa, le fluttuazioni stagionali potrebbero essere profondamente esasperate dall'aumento di popolazione residente nella stagione estiva, a causa della profonda vocazione turistica di gran parte della regione. Alcuni approfondimenti proposti hanno portato all'individuazione di due cluster di comportamento, di cui nel primo, dove le fluttuazioni stagionali sono spiccate, le correlazioni sono mediamente maggiori, mentre nel secondo, dove le fluttuazioni sono più contenute, le correlazioni sono mediamente minori. La corrispondenza tra correlazione è cluster non è però perfetta e probabilmente poco rappresentativa. Infatti, se si esaminano le otto "migliori" serie, quelle cioè che esibiscono il maggiore livello di correlazione, si nota che esse interessano in parte zone turisticamente rilevanti (ad esempio Fasano, Cisternino, Ostuni, Castellaneta e Ginosa), in parte zone poco interessate dal turismo estivo (Scampitella, Caposele), in parte aree molto ampie e variegate dal punto di vista dei flussi turistici (la provincia di Foggia).

L'analisi di correlazione è stata preceduta da una fase di preparazione del database che ha richiesto un notevole sforzo di tempo e risorse. Il database ha esibito, in definitiva, una buona qualità, con ottimi livelli di completezza e continuità per quasi tutte le serie, e un numero non elevato di *outlier*. Non sarebbe stato possibile apprezzare ciò senza la predisposizione di un controllo di qualità, che ha occupato una parte considerevole del tempo dedicato a questa attività. Similmente, la procedura di clustering, per la quale si sono tentate diverse strade, ha rivelato un comportamento alquanto omogeneo tra le serie, senza particolari differenze, se non in termini di entità delle fluttuazioni stagionali.

L'analisi di correlazione è stata condotta effettuando diverse ipotesi per il database dei consumi (grezzo e detrendizzato), per le variabili esogene (alcune più semplici, altre più complesse), e per la risoluzione temporale (da quella nativa a quella annua). Invece di valutare le correlazioni *tout court*, si è provato ad iniziare con quelle più semplici, facendo leva sui risultati per intraprendere strade successive. Ad esempio, l'aggregazione settimanale presentava risultati del tutto analoghi a quella giornaliera, per cui essa è stata in

breve tempo abbandonata. Ad ogni modo, l'utilizzo di risoluzioni temporali più aggregate di quella giornaliera non ha svelato correlazioni differenti, ma ha soltanto, in alcuni casi, enfatizzato correlazioni già rilevate alla scala nativa.

Durante l'analisi sono emersi livelli di correlazione tra i consumi e il clima, rappresentato da diverse variabili, mediamente bassi, fatta eccezione per un manipolo di territori. Si segnala, da questo punto di vista, che osservare una coesistenza tra consumi alti e temperature elevate, e viceversa (in altre parole, un andamento come quello riprodotto in Figura 32), non basta a confermare che esista una dipendenza tra le due variabili. Ciò è particolarmente evidente per i consumi minimi, che si fermano, grossomodo, ad un valore costante indipendentemente dal valore di temperatura, purché bassa.

Le variabili maggiormente correlate ai consumi sono risultate essere quelle direttamente legate alla temperatura, senza particolari differenze tra temperatura media, massima e minima giornaliera. Altre variabili, legate alla precipitazione (ad esempio, il numero di giorni piovosi), hanno mostrato dei livelli di correlazione che, in quanto al segno, sono fisicamente interpretabili, ma, relativamente all'entità, non sono utili per modelli di regressione. Ad ogni modo, si rileva che l'analisi presentata non vuole e non può essere esaustiva. Altre correlazioni possono essere esplorate, eventualmente mediante diverse tecniche non statistiche (ad esempio, mediante approcci di *Machine Learning*). Tra gli elementi che sarebbe utile integrare figura la fluttuazione della popolazione per effetto del turismo estivo: quantificare tale fattore potrebbe infatti essere utile a isolare l'effetto del solo clima.

La necessità di ritrovare un risultato che fosse di utilizzo pratico per gli scopi della Convenzione ha fatto convergere verso la selezione di un numero limitato di serie di dati di consumo che manifestano il maggiore livello di correlazione con variabili climatiche dell'intero database. L'ipotesi è che tali serie possano esibire correlazioni che risultano invece mascherate, per le altre serie temporali, da altri effetti, tra cui il turismo estivo ma anche le anomalie nell'andamento delle serie che sono sfuggite al controllo qualità. Ad esempio, vi sono serie temporali che, ad un controllo puntuale, sembrano essere utilizzabili solo su un periodo temporale più ristretto. Gli esiti della regressione sulle serie maggiormente correlate sono mostrati in Tabella 10. Si noti che i coefficienti *a* e *b* della regressione risentono del dataset utilizzato per le temperature, che nel caso in esame è E-OBS: per applicare le stesse relazioni ad un altro dataset, o a osservazioni puntuali, bisognerebbe prima quantificare l'accuratezza di E-OBS, che, nei luoghi di interesse, è noto sottostimare le temperature a causa dell'esiguità delle stazioni di misura alla base dell'interpolazione.

Tra le otto serie per le quali si è osservata una correlazione significativa, è particolarmente rilevante quella a servizio della provincia di Foggia. Poiché essa racchiude una popolazione che è pari a circa il 20% degli abitanti serviti nel complesso da Acquedotto Pugliese S.p.A., i relativi risultati sono caratterizzati da una buona rappresentatività per l'intera utenza. Per quelle aree che hanno esibito maggiori correlazioni tra consumi e variabili climatiche, ed in particolare la temperatura media giornaliera, la valutazione degli impatti del cambiamento climatico ha rivelato variazioni attese nei consumi medi e massimi annuali molto modeste, mai superiori al +10% sebbene sempre positive, ma caratterizzate da una ridotta variabilità inter-modello. La ridotta entità delle variazioni attese è senz'altro spiegata osservando che, nei territori in esame, l'aumento atteso delle temperature è comunque ridotto – sebbene con conseguenze potenzialmente molto gravi – a pochi gradi o frazioni di grado. La ridotta incertezza invece è spiegabile osservando che i modelli climatici esibiscono tipicamente risultati molto concordi tra loro, mentre sono le variabili climatiche legate alla precipitazione quelle rispetto alle quali i modelli climatici divergono, creando una forte variabilità inter-modello.

Dall'incrocio dei risultati con la conoscenza del sistema infrastrutturale AQP è stato possibile dare una più dettagliata spiegazione circa l'esistenza di macro-aree con una correlazione più spiccata. Il motivo è da ricercarsi nel fatto che, mentre tutti gli altri punti di misurazione sono localizzati lungo le arterie principali della rete di adduzione, a monte di importanti volumi di invaso nonché a monte di intersezioni tra diverse linee di approvvigionamento, per le otto serie in questione i punti di misura sono posti o immediatamente a valle dei serbatoi o a monte di serbatoi di volume molto piccolo, e comunque le macro-aree in questione sono servite da un'unica linea di approvvigionamento. In altre parole, per le otto serie il punto di misura è collocato in un luogo che risente in modo spiccato delle fluttuazioni stagionali e day-by-day, mentre negli altri casi la fluttuazione dei consumi è totalmente laminata dalla presenza dell'invaso.

Le considerazioni summenzionate consentono anche di capire che la posizione della stragrande maggioranza dei punti di consegna non è adeguata per condurre un'analisi dei consumi e delle loro variazioni. Eventuali futuri approfondimenti dovranno tener conto di ciò, considerando la necessità di utilizzare altri sensori localizzati in punti più opportuni, ad esempio lungo le condotte di avvicinamento o comunque a valle dei serbatoi a servizio delle reti urbane di distribuzione.

7. Conclusioni e messaggi chiave

I consumi idrici costituiscono, in relazione alla disponibilità di risorsa idropotabile, la controparte nell'ambito della definizione di "scarsità idrica". Maggiori sono infatti i fabbisogni idropotabili, maggiore deve essere la quantità di risorsa idrica da destinare allo scopo di soddisfarli. In definitiva, **una variazione positiva (aumento) dei consumi idropotabili rappresenta un aumento del pericolo di mancato soddisfacimento dei fabbisogni per effetto del cambiamento climatico**, mentre una variazione negativa (diminuzione) rappresenta una diminuzione del pericolo di mancato soddisfacimento dei fabbisogni per effetto del cambiamento climatico. Nella fattispecie, in questo documento si va a quantificare la variazione attesa su due particolari valori di consumo: il valore mensile medio annuo QM e il valore mensile massimo annuale QMAX, nell'ipotesi che essi manifestino una sensitività diversa rispetto al clima. Entrambi i valori sono mediati su un periodo di trent'anni. Naturalmente, l'intero impalcato del lavoro poggia sulla considerazione che i consumi idrici possano essere un efficace *proxy* dei fabbisogni.

Le analisi di correlazione tra consumi e variabili climatiche hanno mostrato risultati molto eterogenei nel database oggetto di studio. Mediamente, tutte le correlazioni investigate si sono rivelate statisticamente poco significative, e dunque poco utili per costruire un modello di impatto dei cambiamenti climatici generalizzato. Ciò può essere spiegato nell'inadeguatezza dei punti di misura a cogliere la fluttuazione dei consumi per via della loro localizzazione a monte di grandi invasi e di interconnessioni tra linee di approvvigionamento plurime. Dunque, l'attenzione si è spostata su **un manipolo di territori per i quali i consumi mensili si sono rivelati altamente correlati con la media mensile delle temperature medie giornaliere, purché superiori a 10°C**. Per questi territori, è stata modellata una regressione lineare tra le due variabili, che ha concesso di inferire circa gli effetti del cambiamento climatico sui consumi idrici**. L'esistenza di tali correlazioni per queste specifiche macro-aree è spiegata dal fatto che i relativi punti di misura sono ben posizionati** (a valle di serbatoi, o a monte di serbatoi di volume ridotto) e dall'esistenza di un'unica linea di approvvigionamento. In altre parole, per tali macro-aree le variazioni di consumo sono più visibili rispetto alle altre macro-aree; di contro, nelle macro-aree rimanenti non è detto che tali variazioni non esistano, ma esse non sono colte dalle misurazioni analizzate.

Dalle analisi su tale selezione di territori risulta che **i consumi idrici sono attesi aumentare per tutti gli scenari di concentrazione, con una ridotta incertezza associata all'utilizzo di un ensemble di proiezioni climatiche. Tali aumenti sono comunque alquanto ridotti, e mai superiori al 10%**.

In media sul territorio, i consumi medi annui sono attesi aumentare del +3.4% sotto RCP 2.6, del +3.7% sotto RCP 4.5 e **del +4.6% sotto RCP 8.5**. **I consumi massimi annuali sono attesi aumentare** con tasso lievemente minori, pari a +2.8% sotto RCP 2.6, +3.0% sotto RCP 4.5 e **+3.3% sotto RCP 8.5**.

Localmente, le serie che esibiscono i valori maggiori di variazione attesa sono PT89 (Comune di Fasano) e PT75 (Comune di Castellaneta), che presentano rispettivamente +7.2% e +6.7% per i consumi medi annui e +5.7% e +4.3% per i consumi massimi annuali sotto RCP 8.5. Tuttavia, tali variazioni potrebbero rivelarsi non così critiche poiché associate a valori di riferimento dei consumi non alti. Al contrario, la serie che presenta il maggior valore di riferimento è PT66, associata a un insieme **di 34 Comuni nel foggiano, per cui è attesa una variazione massima sotto RCP 8.5 che è la più modesta, pari a +1.7%** sia per i consumi medi annui sia per i massimi annuali. **Tale informazione è alquanto rilevante dal momento che l'area interessata comprende circa il 20% dell'utenza complessivamente servita da AQP**. Si noti che i suoi consumi medi annui di riferimento sono maggiori di circa 8 volte rispetto a PT89, e di circa 15 volte rispetto a PT75.

Tutte le variazioni sono associate a incertezze molto modeste, conseguenza dell'aver considerato come variabile climatica la temperatura, le cui proiezioni sono in genere alquanto omogenee tra i diversi modelli climatici disponibile. **Permane, naturalmente, l'incertezza associata al modello regressivo, che è comunque molto bassa soprattutto per la serie PT89 (Comune di Fasano)**.

Si ricorda, a conclusione, che l'analisi presentata non vuole e non può essere esaustiva, ma consiste soltanto in una prima esplorazione dei dati disponibili. Uno dei principali vulnus è costituito dal non aver considerato l'aumento della popolazione residente per effetto del turismo estivo, che, in larga parte del territorio analizzato, è molto significativo. **Tutte le limitazioni e le assunzioni dell'attività sono trattate nella sezione di discussione**.

Appendice I: Composizione delle macro-aree

Serie simboliche	Sensori associati	Abitanti	Area (kmq)	Numero di Municipalità servite	Municipalità servite
PT1	M001008	591309	5667.93	34	Rocchetta Sant'Antonio, Candela, Torremaggiore, San Severo, Serracapriola, Chieuti, Lesina, Poggio Imperiale, Apricena, Sannicandro Garganico, Cagnano Varano, Carpino, Ischitella, Vico del Gargano, Rodi Garganico, Peschici, Vieste, Mattinata, Monte Sant'Angelo, San Giovanni Rotondo, San Marco in Lamis, Rignano Garganico, Manfredonia, Foggia, Lucera, Troia, Orsara di Puglia, Castelluccio dei Sauri, Ascoli Satriano, Cerignola, Stornara, Stornarella, Orta Nova, Ordona,
PT2	M001065	121952	817.19	2	Matera, Altamura
PT3	M000473	98760	812.81	4	Margherita di Savoia, San Ferdinando di Puglia, Trinitapoli, Cerignola
PT4	M001063	106321	808.58	2	Altamura, Gravina in Puglia
PT5	M000742	70239	675.75	5	Mottola, Castellaneta, Palagianello, Palagiano, Crispiano
PT6	M001029	114294	525.79	10	Porto Cesareo, Copertino, Leverano, Veglie, Carmiano, Salice Salentino, Guagnano, Nardò, Seclì,
PT7	M001356	19285	505.20	12	Casalnuovo Monterotaro, Casalvecchio di Puglia, Castelnuovo della Daunia, Pietramontecorvino, Volturino, Roseto Valfortore, Alberona, Castelluccio Valmaggiore, Faeto, Motta Montecorvino, San Marco la Catola, Carlantino
PT8	MM00131	254157	492.80	4	Palagiano, Massafra, Taranto, Maruggio
PT9	MM00169	42297	488.25	7	Atella, Barile, Filiano, Ginestra, Melfi, Rionero in Vulture, Ripacandida
PT10	M000601	193489	475.32	13	Trepuzzi, Novoli, Surbo, Monteroni di Lecce, Arnesano, Carmiano, Lecce, Lequile, San Pietro in Lama, San Cesario di Lecce, Cavallino, Lizzanello, San Donato di Lecce
PT11	M001053	390014	474.13	8	Toritto, Grumo Appula, Binetto, Bitetto, Palo del Colle, Sannicandro di Bari, Bitritto, Bari
PT12	M001107	80589	448.10	4	Ceglie Messapica, San Michele Salentino, San Vito dei Normanni, Ostuni
PT13	M001002	19348	424.58	6	Calitri, Lacedonia, Bisaccia, Vallata, Monteverde, Aquilonia
PT14	M000545	229591	424.35	3	Taranto, Statte, Crispiano
PT15	M001035	41658	404.01	2	Minervino Murge, Canosa di Puglia
PT16	M001106	47861	329.53	2	Ostuni, Carovigno
PT17	M000602	89081	327.73	1	Brindisi
PT18	M000838	81865	324.53	7	Porto Cesareo, Copertino, Leverano, Veglie, Carmiano, Salice Salentino, Guagnano
PT19	M000840	114997	312.56	4	Squinzano, Lecce, Cavallino, San Donato di Lecce
PT20	M000586	57378	298.16	4	Manduria, Sava, Torricella, Maruggio
PT21	M000595	58167	259.31	3	Oria, Latiano, Mesagne
PT22	MM00129	10213	254.56	1	Minervino Murge
PT23	MM00183	17811	229.55	1	Pisticci
PT24	M001534	167403	229.19	3	Trani, Bisceglie, Molfetta
PT25	M000474	41394	224.15	3	San Ferdinando di Puglia, Margherita di Savoia, Trinitapoli

PT26	M000535	16575	211.19	1	Mottola
PT27	MM00005	16110	203.54	1	Melfi
PT28	M001012	10810	189.42	3	Accadia, Deliceto, Bovino
PT29	M005545	28809	188.68	2	San Ferdinando di Puglia, Trinitapoli
PT30	M001033	7362	181.30	1	Spinazzola
PT31	M000593	31747	178.86	1	Manduria
PT32	MM00170	12148	168.26	1	Venosa
PT33	M002951	14996	159.40	1	Laterza
PT34	M001195	31445	149.45	1	Canosa di Puglia
PT35	M001241	3297	148.49	2	Panni, Sant'Agata di Puglia
PT36	M000997	6143	143.17	4	Cairano, Andretta, Morra de Sanctis, Guardia Lombardi
PT37	M001155	79701	141.03	5	Cellamare, Valenzano, Capurso, Rutigliano, Mola di Bari
PT38	M000542	30923	127.86	1	Massafra
PT39	M004588	2321	115.84	1	Sant'Agata di Puglia
PT40	MM00173	2000	113.93	1	Montemilone
PT41	M000582	39558	111.41	6	San Giorgio Ionico, Carosino, Faggiano, Lizzano, Roccaforzata, Monteparano
PT42	M001141	82846	101.82	2	Molfetta, Giovinazzo
PT43	M001006	5843	99.72	1	Calitri
PT44	M000800	31894	99.45	1	Grottaglie
PT45	M000597	19354	99.24	2	Torre Santa Susanna, Erchie
PT46	MM00009	1284	96.84	1	Aliano
PT47	M000928	20131	78.16	2	San Pietro Vernotico, Torchiarolo
PT48	M000995	7518	76.18	2	Morra de Sanctis, Lioni
PT49	M000584	9468	75.92	2	Torricella, Maruggio
PT50	M001051	8916	75.59	1	Toritto
PT51	M000539	15815	69.47	1	Palagiano
PT52	M001168	27532	68.79	1	Terlizzi
PT53	M000868	10551	55.95	1	San Pancrazio Salentino
PT54	M000829	8740	44.79	1	Erchie
PT55	M000533	7483	43.51	1	Palagianello
PT56	M000826	14469	41.55	2	San Marzano di San Giuseppe, Fragagnano
PT57	M000924	6818	37.43	1	Cellino San Marco
PT58	M001013	1413	36.63	1	Monteleone di Puglia
PT59	M000743	8635	34.80	1	Villa castelli
PT60	M000869	7117	33.56	1	San Donaci
PT61	M001075	1573	23.19	1	Teora
PT62	M000570	4277	16.58	1	Montemesola
PT63	MM00306	1435	15.34	1	Scampitella
PT64	MM00102	5199	8.91	1	Monteiasi
PT65	M001235	1930	7.08	1	Sant'Andrea di Conza
PT66	M001028, M000057, M001341	595370	5564.43	34	Torremaggiore, San Severo, Serracapriola, Chieuti, Lesina, Poggio Imperiale, Apricena, Sannicandro Garganico, Cagnano Varano, Carpino, Ischitella, Vico del Gargano, Rodi Garganico, Peschici, Vieste, Mattinata, Monte Sant'Angelo, San Giovanni Rotondo, San Marco in Lamis, Rignano Garganico, Manfredonia, Foggia, Lucera, Troia, Orsara di Puglia, Castelluccio dei Sauri, Ascoli Satriano, Cerignola, Stornara, Stornarella, Orta Nova, Ordona, Carapelle, Zapponeta,
PT67	M001077, M001078	159517	566.42	6	Putignano, Castellana Grotte, Conversano, Mola di Bari, Monopoli, Polignano a Mare
PT68	M001038, M001502, M002593, MM00104	95653	401.00	1	Andria

PT69	M000723, M000724	381098	398.53	7	Sannicandro di Bari, Bitritto, Grumo Appula, Binetto, Bitetto, Palo del Colle, Bari
PT70	MM00176, MM00184, MM00185	57785	389.76	1	Matera
PT71	M001096, M001095	48756	296.11	1	MARTINA FRANCA
PT72	M000779, M000780, M000781	444917	293.82	8	Triggiano, Adelfia, Noicattaro, Mola di Bari, Cellamare, Valenzano, Capurso, Bari
PT73	M001049, M001048	53273	289.43	2	Ruvo di Puglia, Terlizzi
PT74	M000559, M000347, M000348, M000575	202033	246.95	1	Taranto
PT75	M000530, M000531	17393	240.87	1	Castellaneta
PT76	M003395, M000671, M003396, M000685	27655	206.87	1	Gioia del Colle
PT77	MM00186, M000674, M000675	22146	188.18	1	Ginosa
PT78	M000795, M000823, M000822	36274	176.95	1	Francavilla Fontana
PT79	M001050, M001159, M001166	56929	174.32	2	Bitonto, Bitonto
PT80	M001579, M001574	44971	168.23	1	Corato
PT81	M002394, M002417, MM00181	46708	156.15	1	Monopoli
PT82	M001079, M001080	19564	148.61	1	Noci
PT83	M001052, M001151	352512	147.55	2	Modugno, Bari
PT84	M003123, M003124, M005535, M001586	92094	147.41	1	Barletta
PT85	M001062, M001060	26050	143.69	1	Santeramo in Colle
PT86	M000706, M000707, M000708, M000709	342844	135.86	2	Triggiano, Bari
PT87	MM00171, MM00172	13247	133.03	1	Lavello
PT88	M001067, M000741	21613	131.40	1	Acquaviva delle Fonti
PT89	M001091, M001090	38667	129.20	1	Fasano
PT90	M001147, M001134	316532	115.57	1	Bari
PT91	M000673, M000694	18284	104.51	2	Sammichele di Bari, Turi
PT92	M001536, M003140, M003141	53139	102.29	1	Trani
PT93	M001083, MM00132, M001084	28176	99.42	1	Putignano
PT94	M001059, M000668, M000739	11958	89.25	1	Cassano delle Murge
PT95	M000830, M000831, M000832	7303	73.50	1	AVETRANA
PT96	M001535, M001139	51718	68.96	1	Bisceglie
PT97	MM00174, MM00175	5184	62.47	1	Palazzo San Gervasio
PT98	M001097, M001098	12078	54.30	1	Cisternino
PT99	MM00001, MM00002, MM00003	1457	51.21	1	Conza della Campania
PT100	M001085, M001094, M001093	13928	48.62	1	Locorotondo
PT101	M001142, M001144	20300	43.87	1	Giovinazzo
PT102	M001342, M001076, M004867, M001074	3797	41.05	1	Caposele
PT103	M001089, M001087	10859	40.26	1	Alberobello
PT104	MM00304, MM00305	1486	14.18	1	Vallesaccarda
PT105	MM00127, M001009	2239	10.89	1	Anzano di Puglia
PT106	M000839	208069	1919.00	84	San Pancrazio Salentino, San Donaci, San Pietro Vernotico, Cellino San Marco, Acquarica del Capo, Alessano, Alezio, Alliste, Andrano, Aradeo, Bagnolo del Salento, Botrugno, Calimera, Campi Salentina, Cannole, Caprarica di Lecce, Carpignano Salentino, Casarano, Castri

	di Lecce, Castrignano de' Greci, Castrignano del Capo, Castro, Collepasso, Corigliano d'Otranto, Corsano, Cursi, Cutrofiano, Diso, Gagliano del Capo, Galatina, Gallipoli, Giuggianello, Giurdignano, Maglie, Martano, Martignano, Matino, Melendugno, Melissano, Melpignano, Miggiano, Minervino di Lecce, Montesano Salentino, Morciano di Leuca, Muro Leccese, Neviano, Nociglia, Ortelle, Otranto, Palmariggi, Parabita, Patù, Poggiardo, Presicce, Racale, Ruffano, Salve, San Cassiano, San Donato di Lecce, Sanarica, Sannicola, Santa Cesarea Terme, Scorrano, Seclì, Sogliano Cavour, Soleto, Specchia, Spongano, Sternatia, Supersano, Surano, Taurisano, Taviano, Tiggiano, Tricase, Tuglie, Ugento, Uggiano la Chiesa, Vernole, Zollino

Appendice II: Correlazioni analizzate per il database grezzo

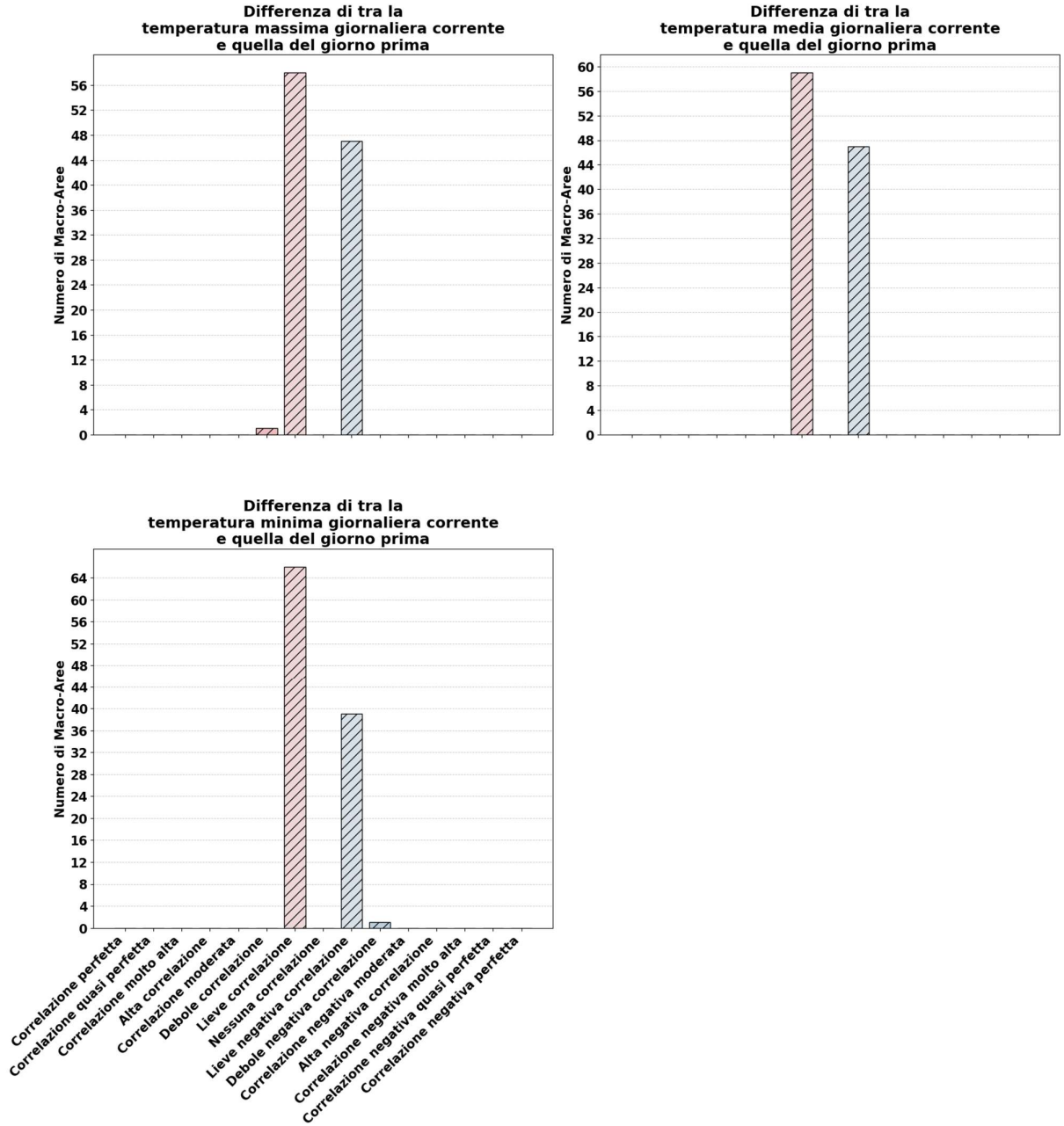

Figura A1. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

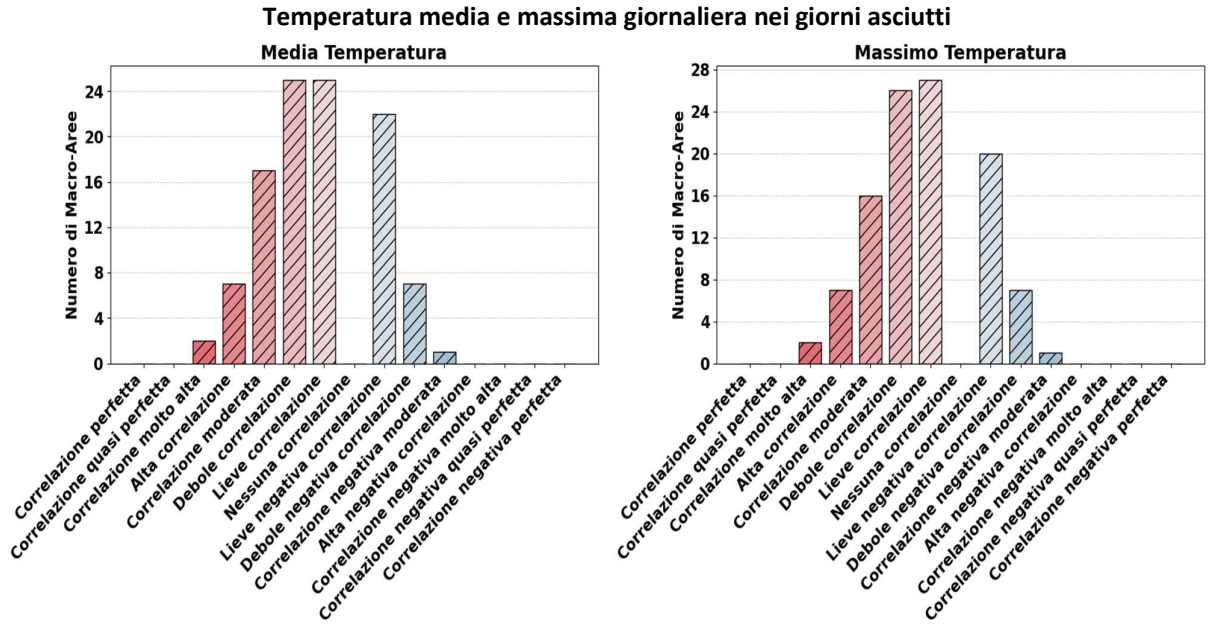

Figura A2. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

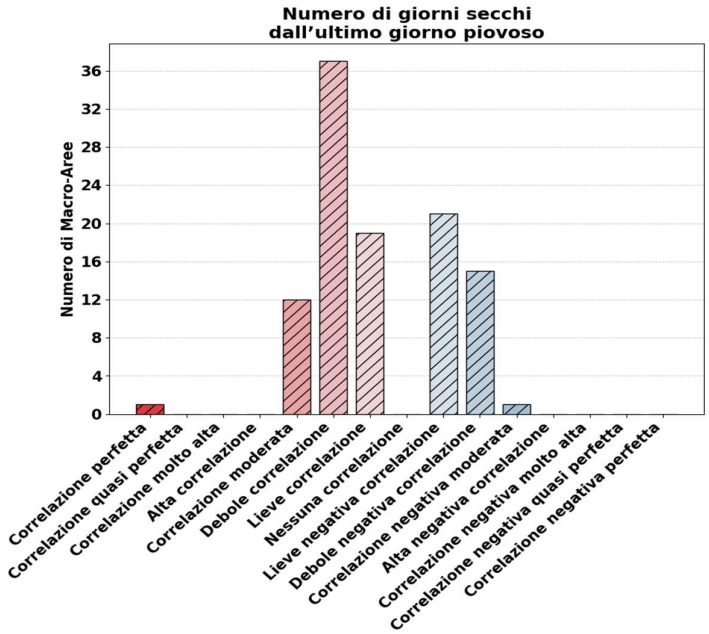

Figura A3. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

Temperatura media e massima giornaliera nei giorni poco piovosi (<10mm), precipitazione nei giorni poco piovosi (<10 mm)

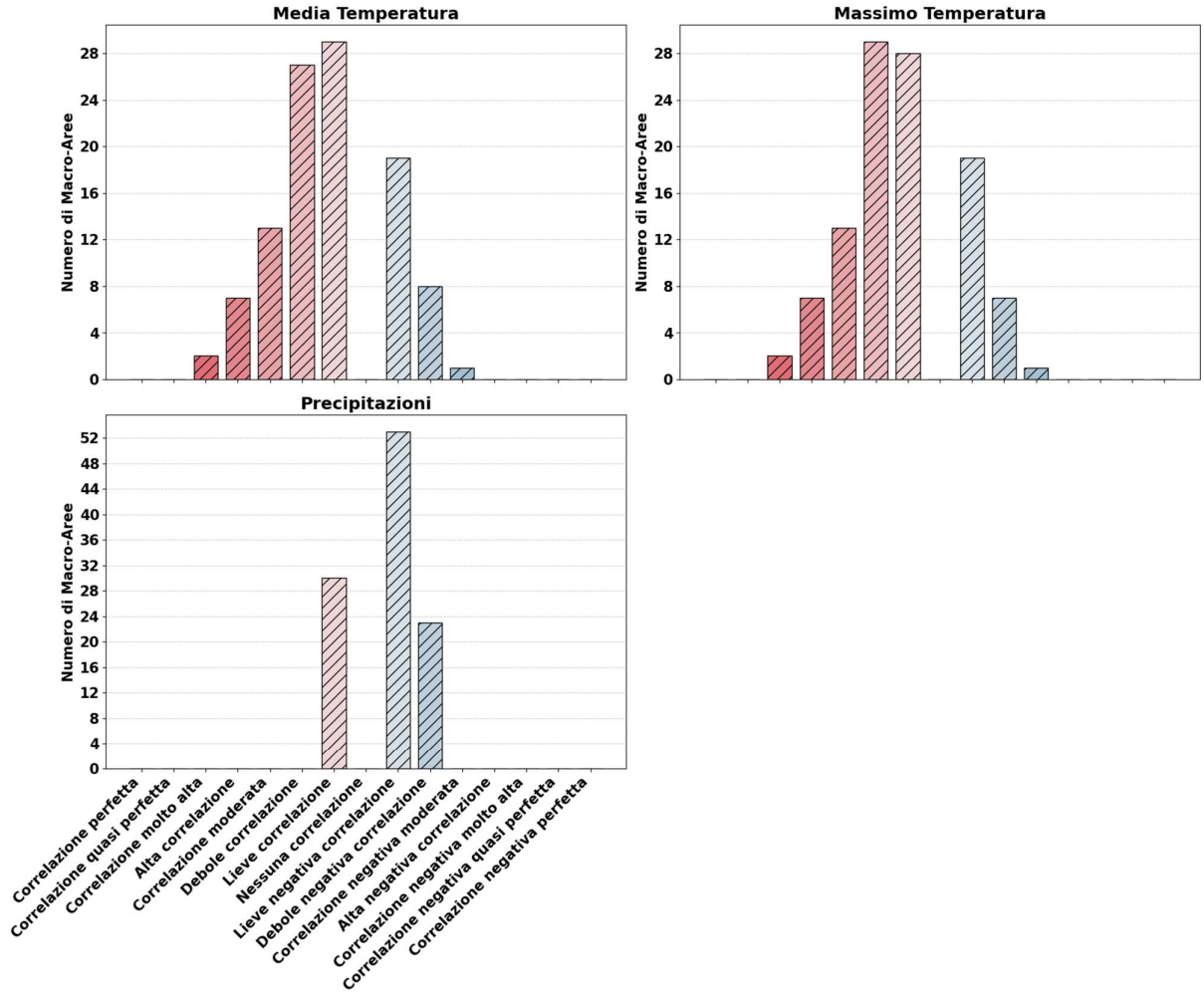

Figura A4. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

Temperatura media e massima giornaliera invernali, precipitazione invernale

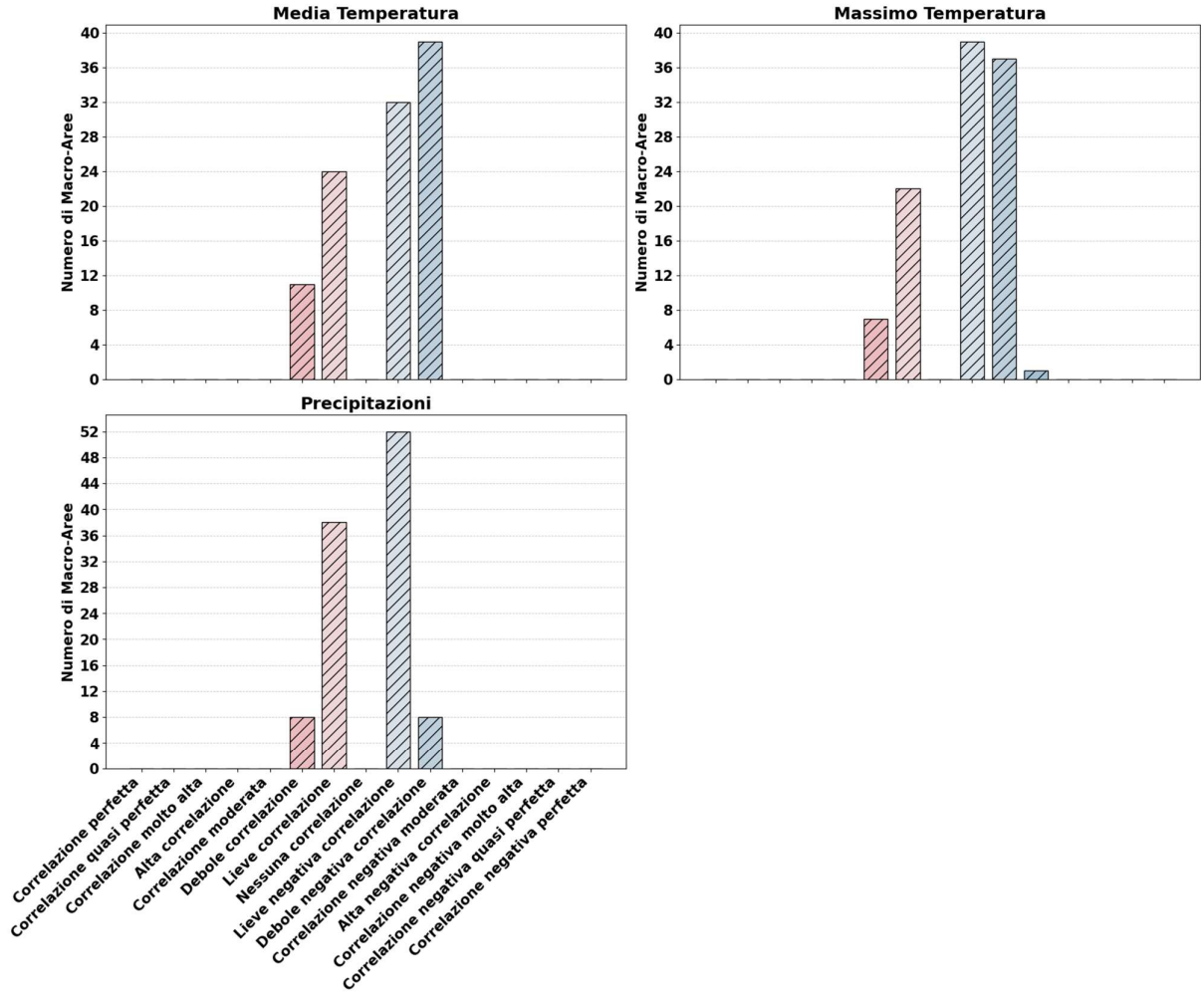

Figura A5. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

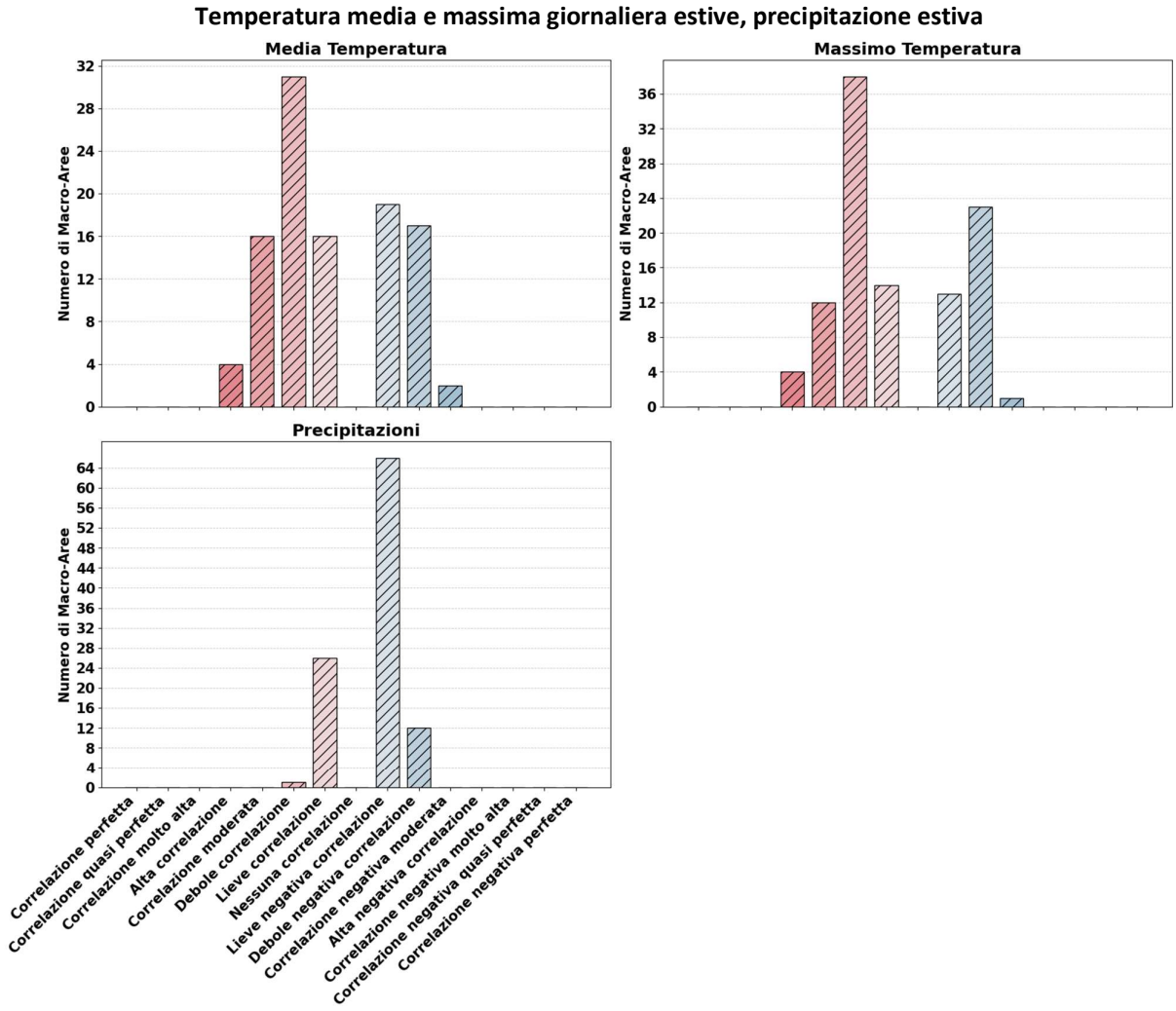

Figura A6. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

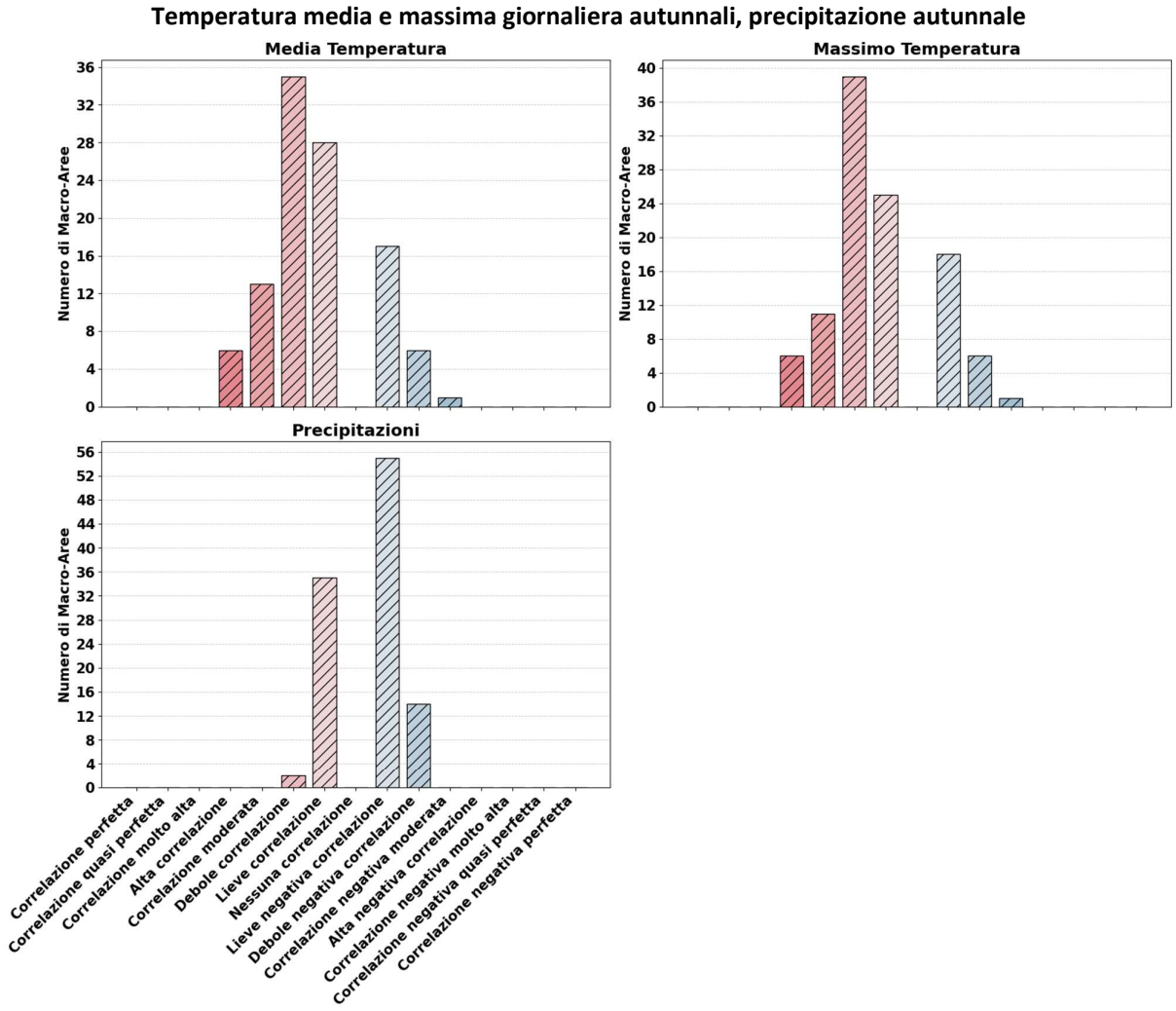

Figura A7. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

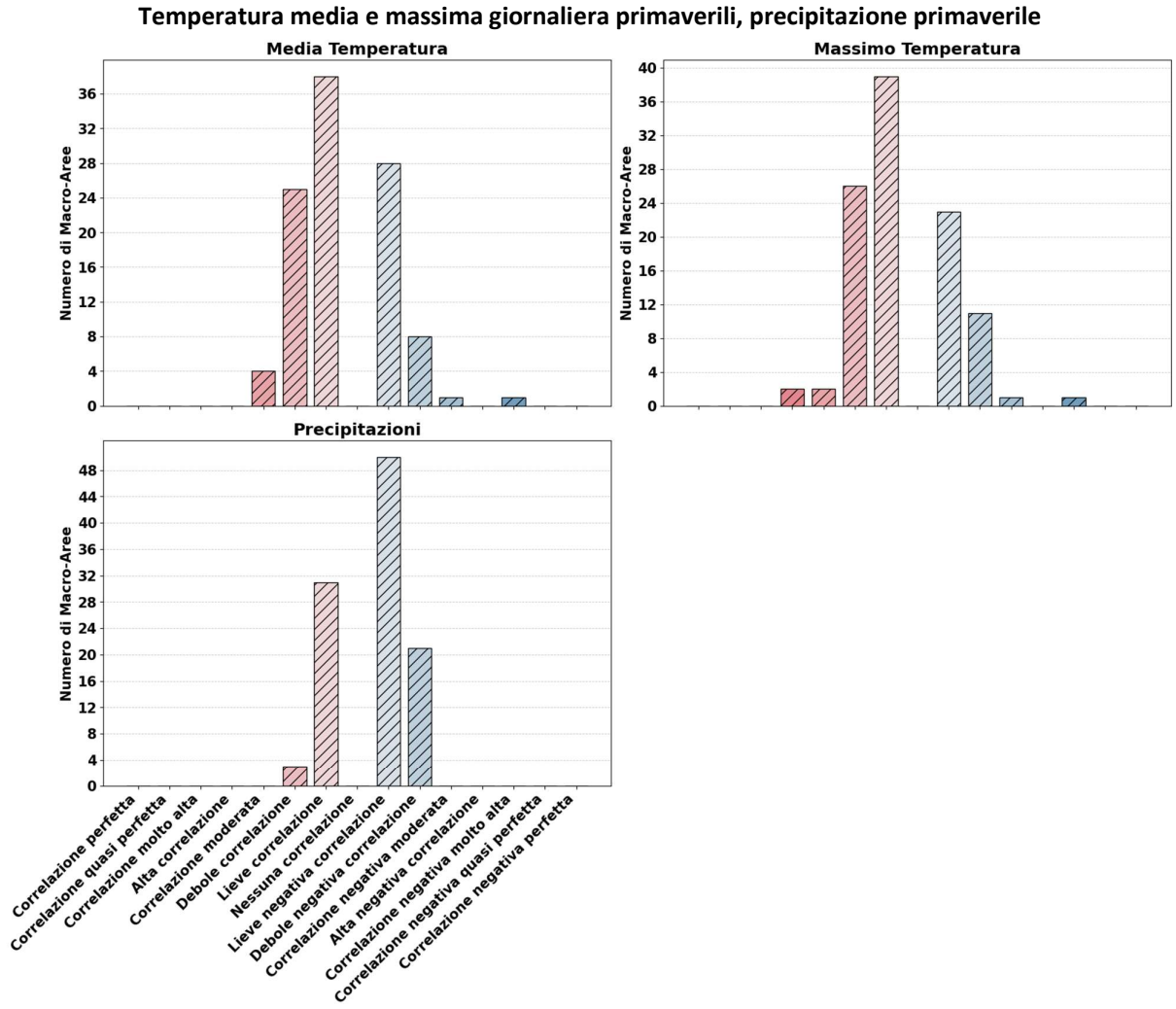

Figura A8. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

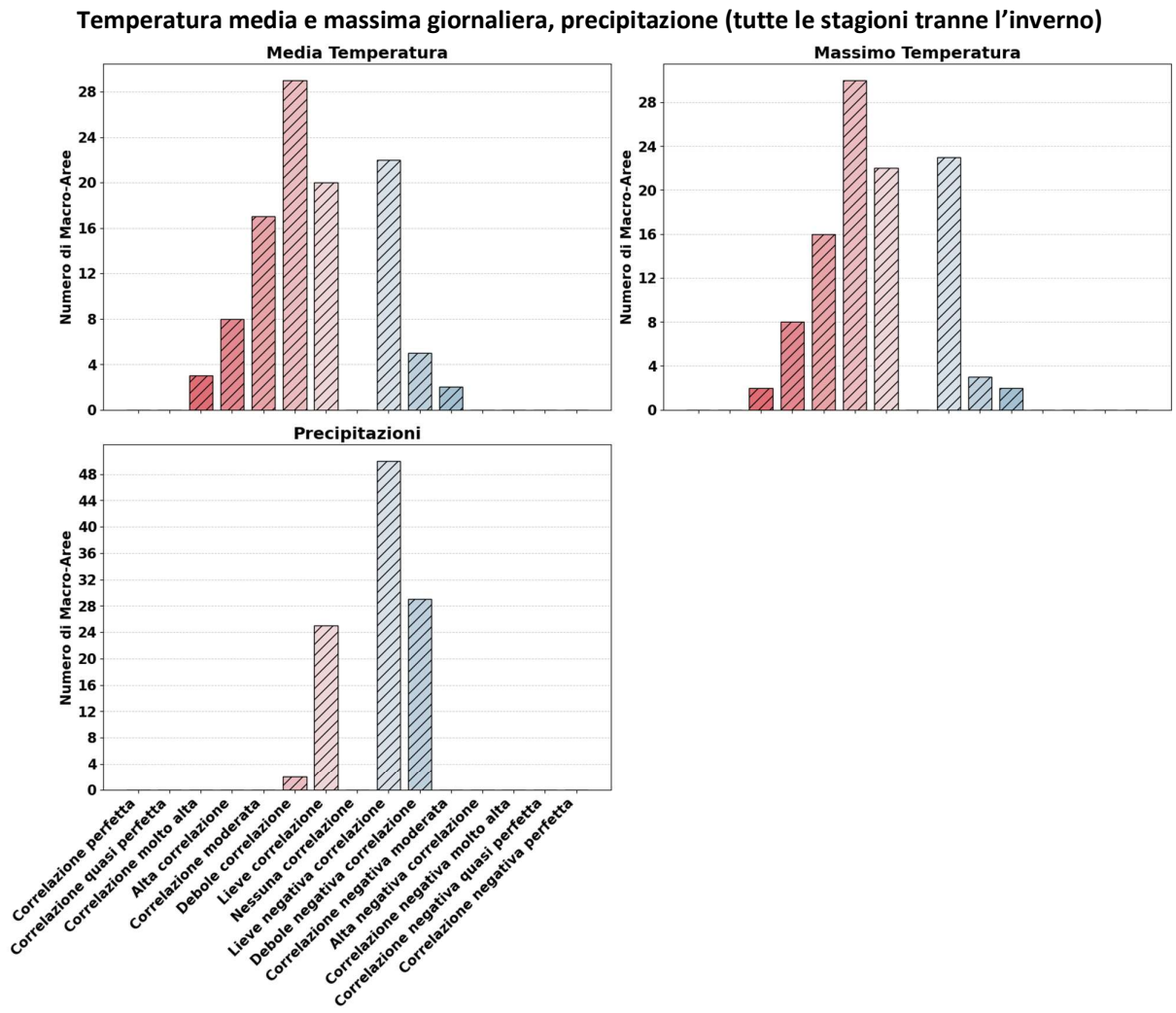

Figura A9. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

Temperatura media e massima giornaliera, precipitazione (solo nei giorni in cui la temperatura massima è >25°C)

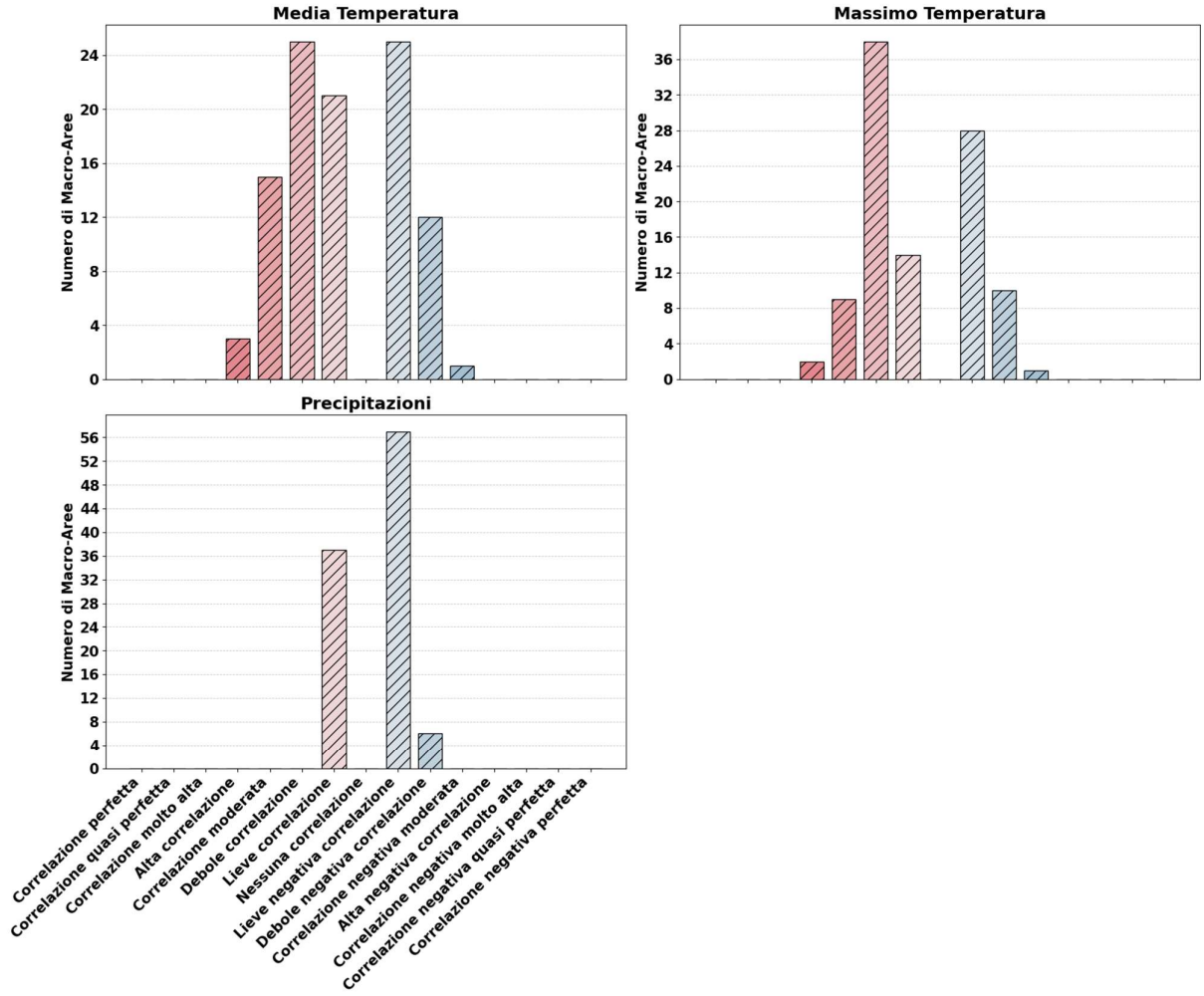

Figura A10. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

Temperatura media e massima giornaliera, precipitazione (solo nei giorni in cui la temperatura massima è >30°C)

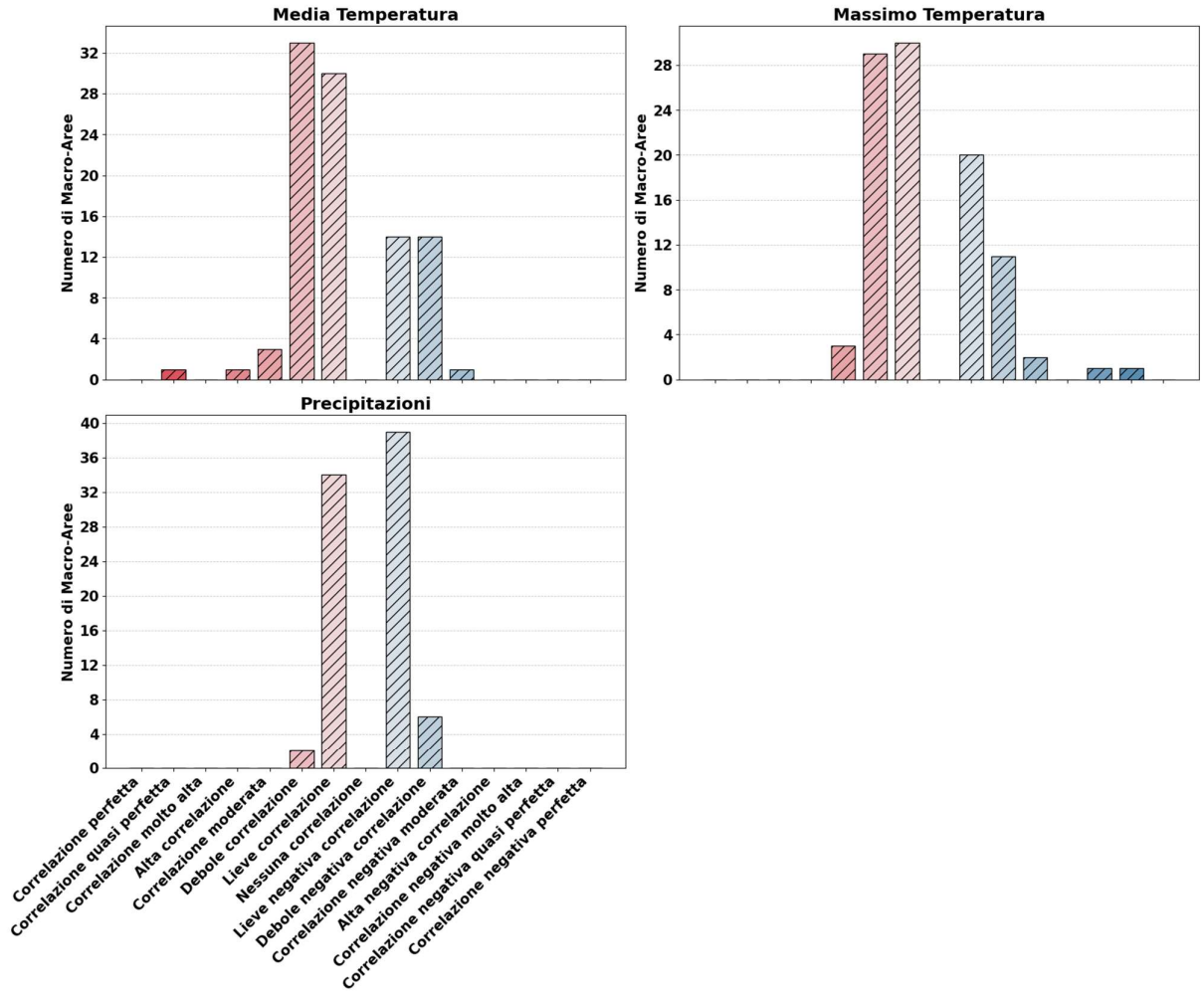

Figura A11. Distribuzione delle macro-aree per classi di correlazione (database grezzo, aggregazione giornaliera, variabile climatica indicata nel titolo del grafico.

Appendice III: Correlazioni analizzate per il database detrendizzato

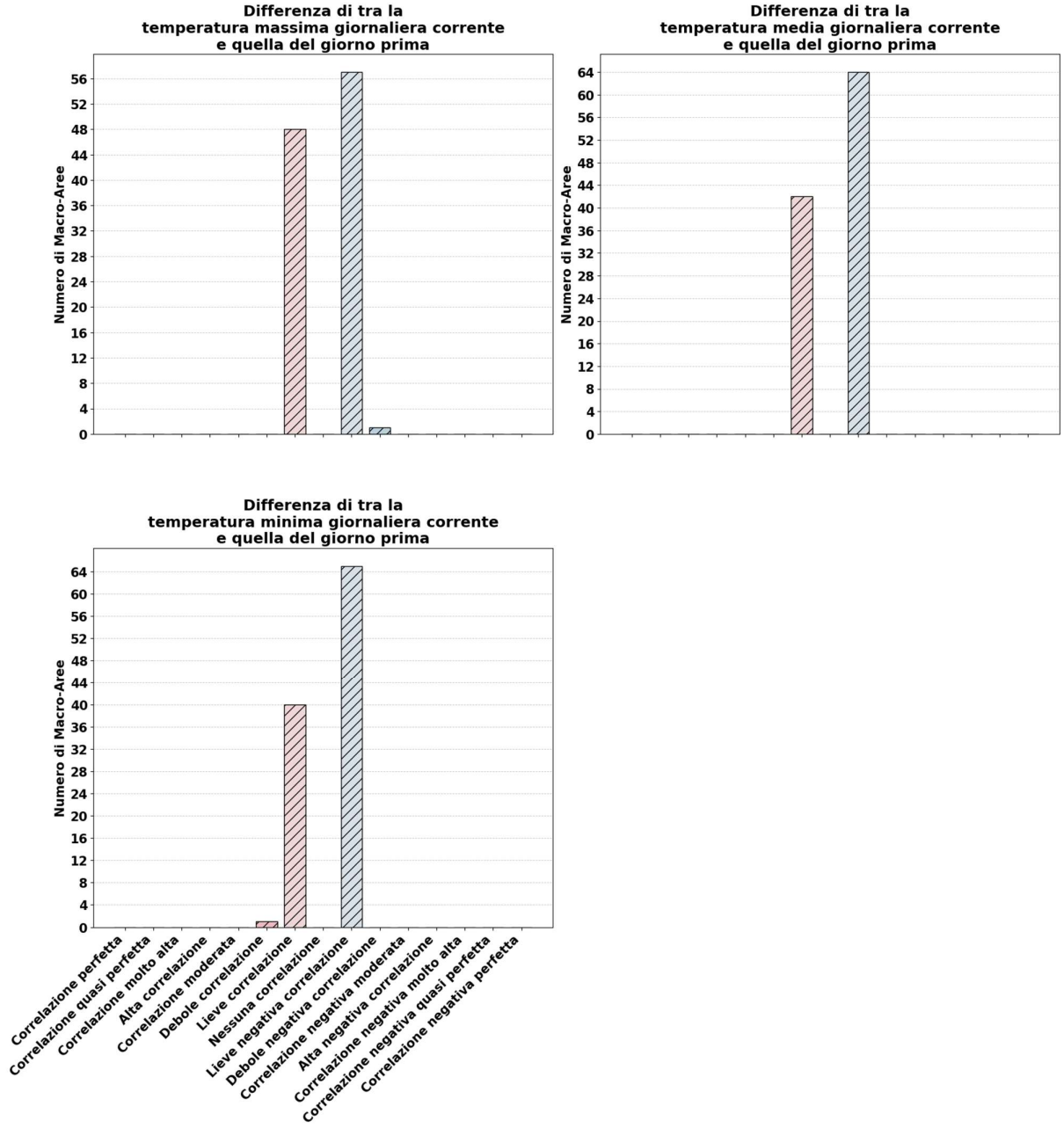

Figura A12. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

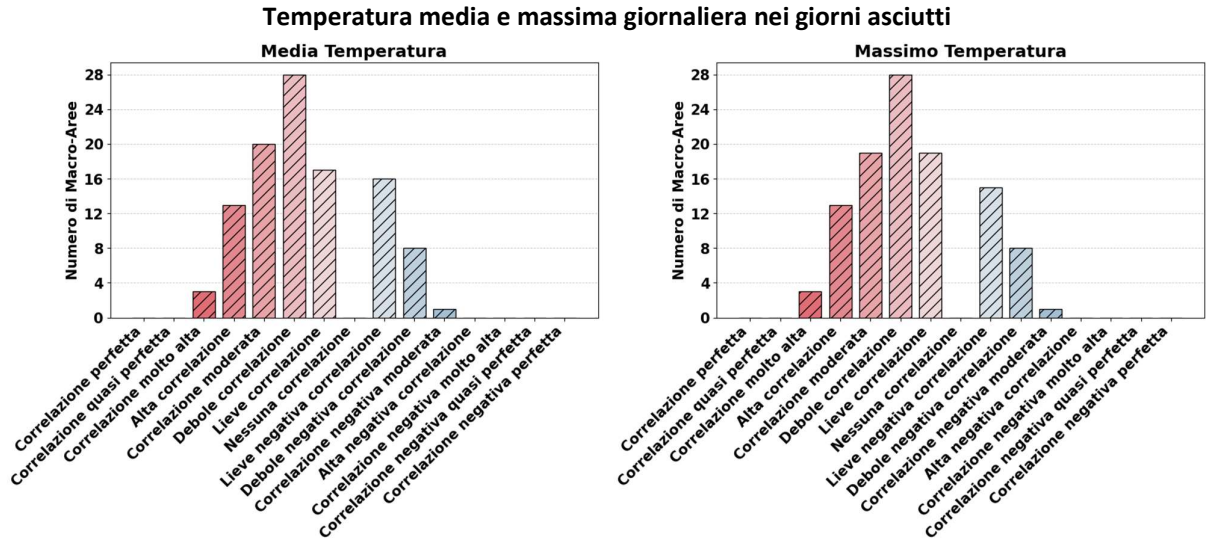

Figura A13. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

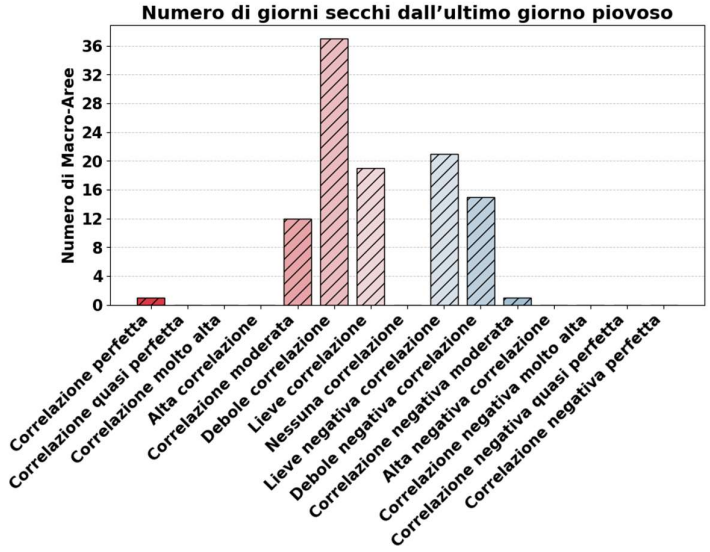

Figura A14. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

Temperatura media e massima giornaliera nei giorni poco piovosi (<10mm), precipitazione nei giorni poco piovosi (<10mm)

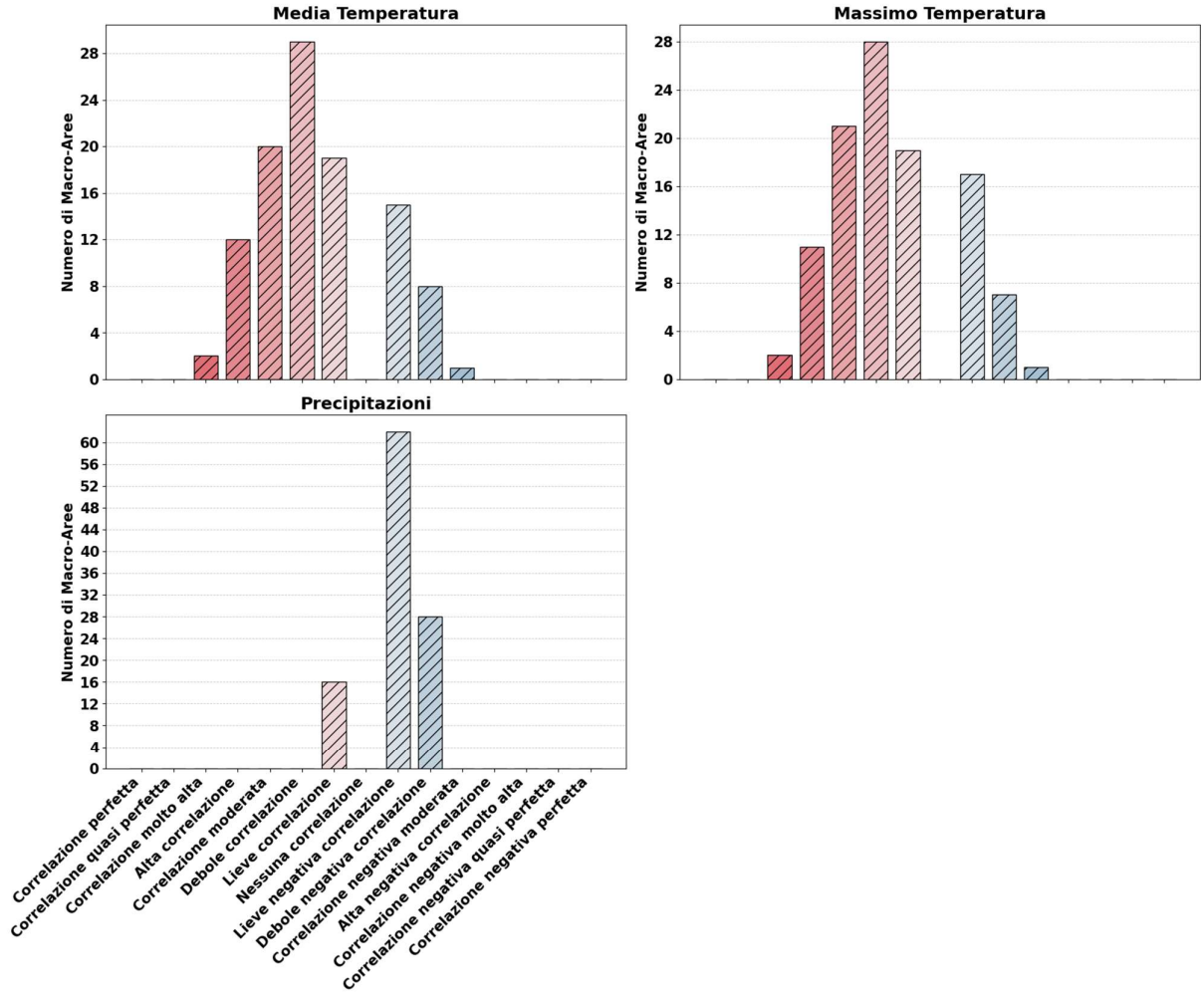

Figura A15. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

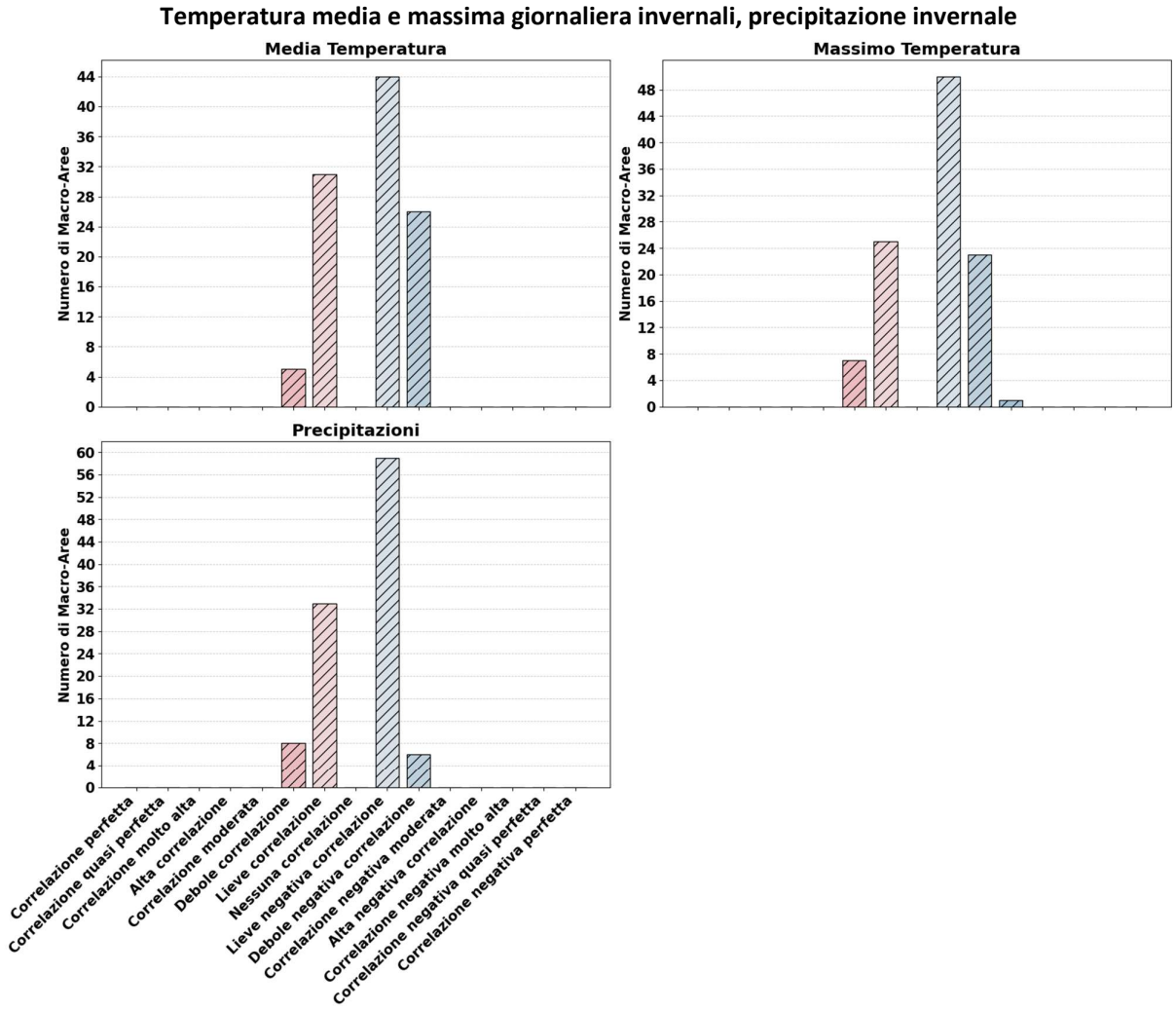

Figura A16. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

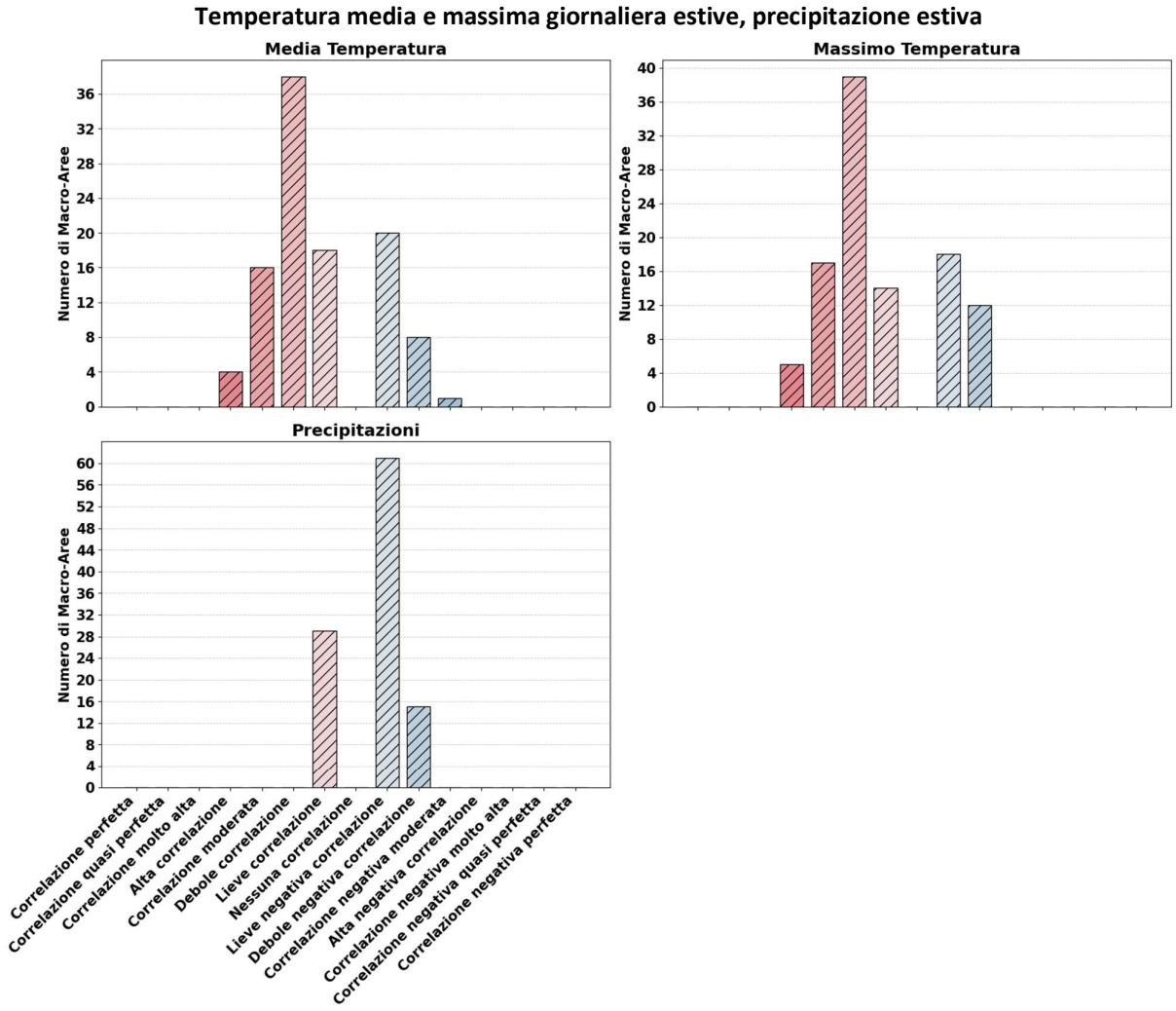

Figura A17. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

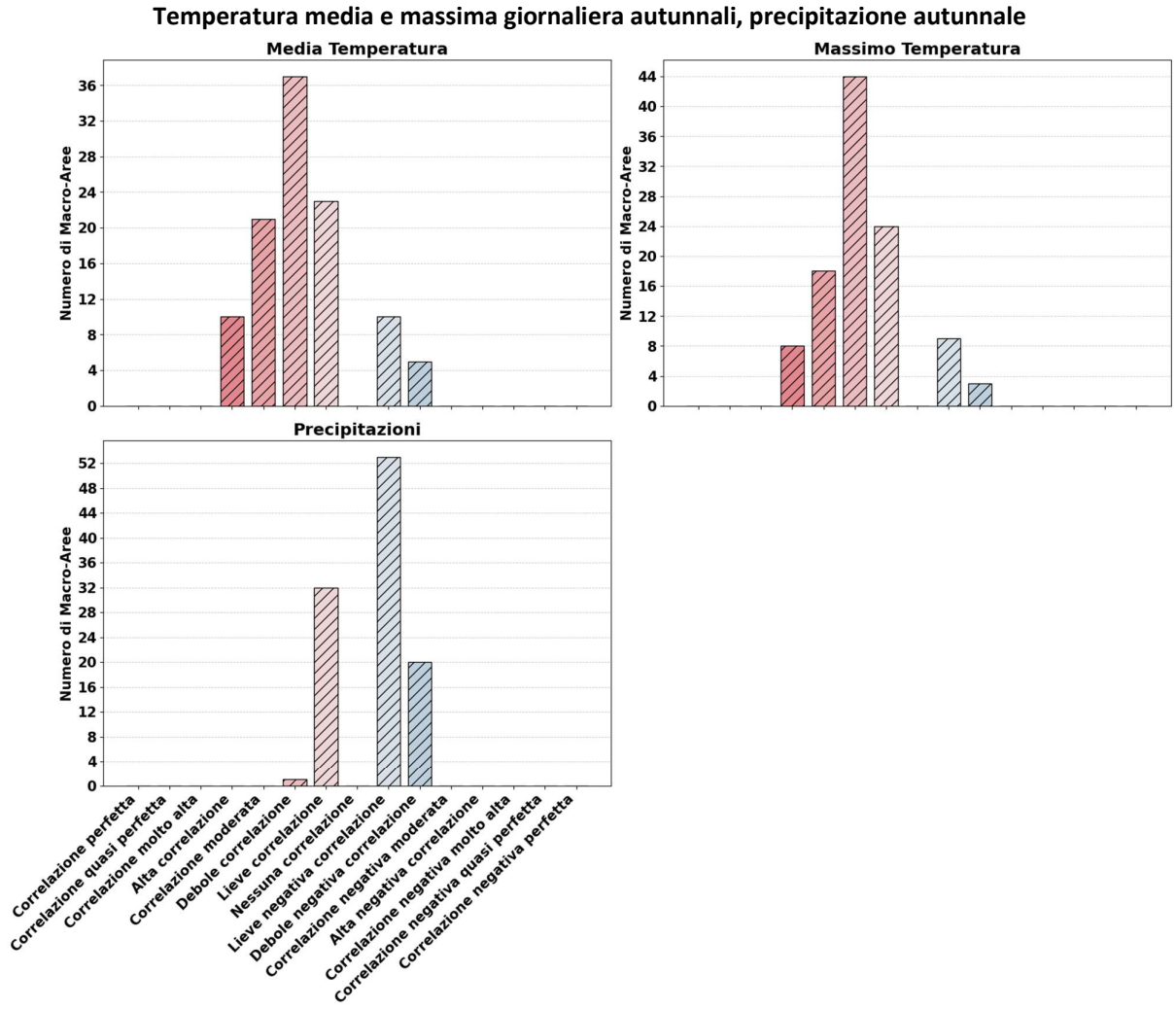

Figura A18. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

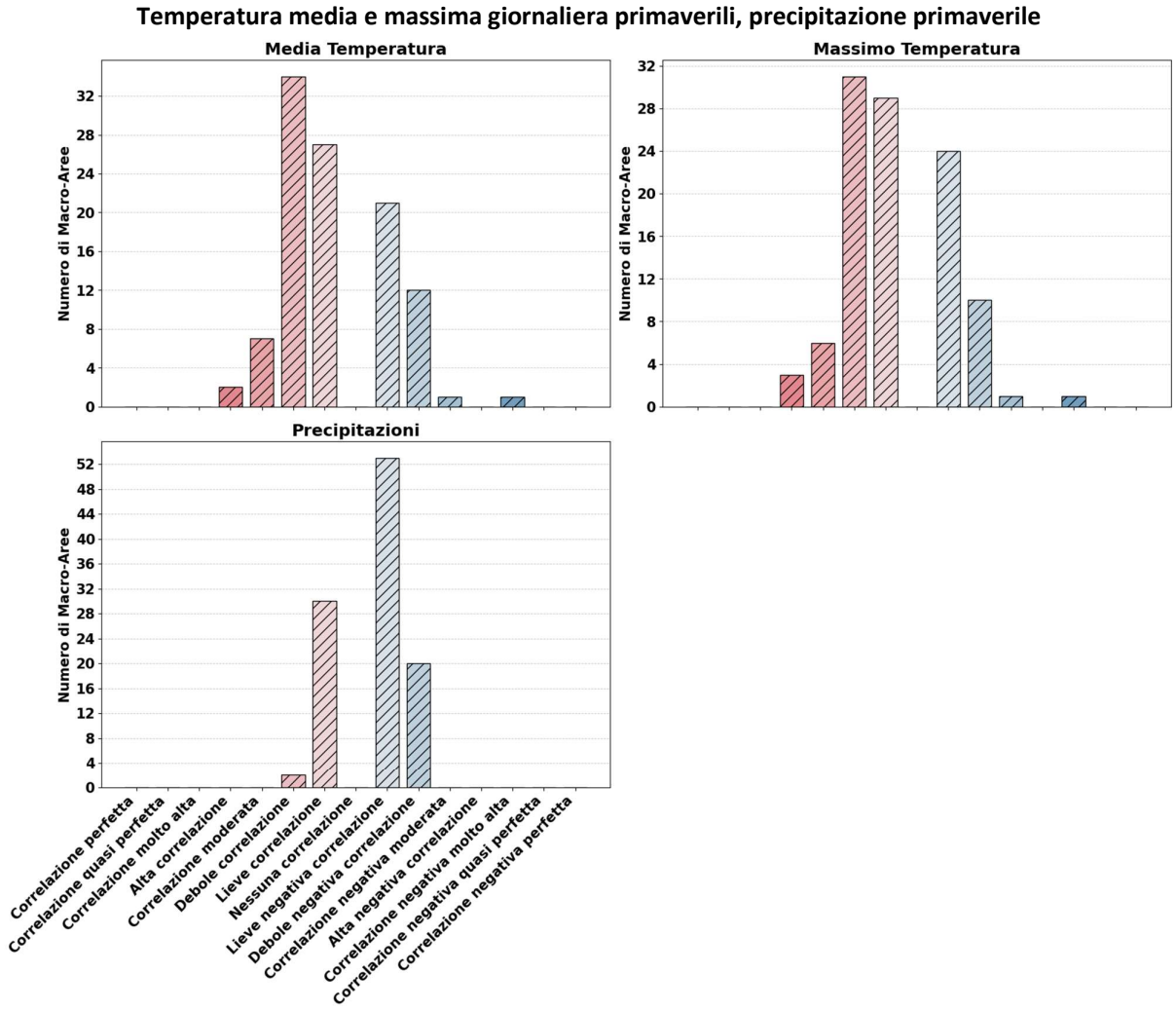

Figura A19. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

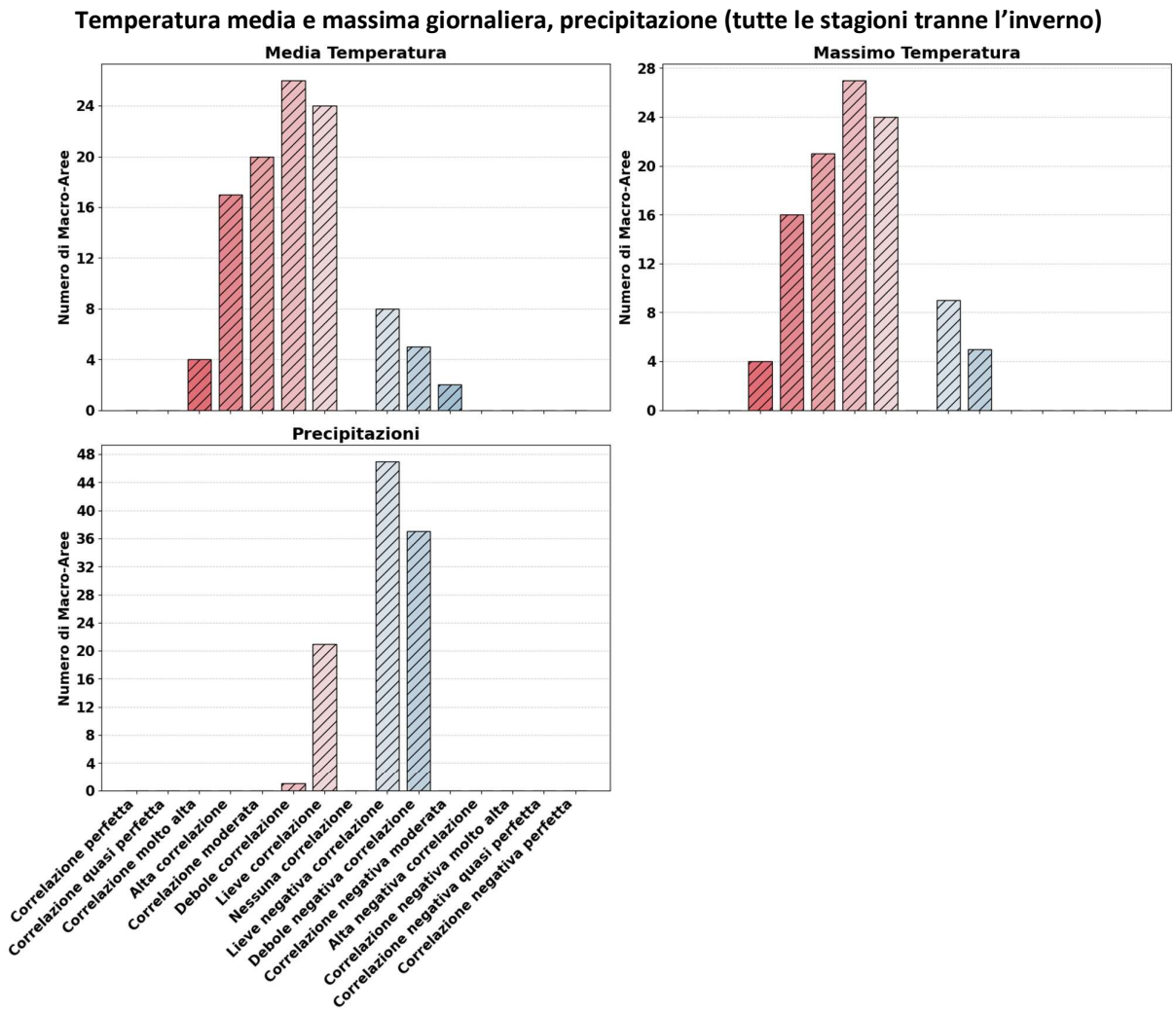

Figura A20. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

Temperatura media e massima giornaliera, precipitazione (solo nei giorni con temperatura massima >25°C)

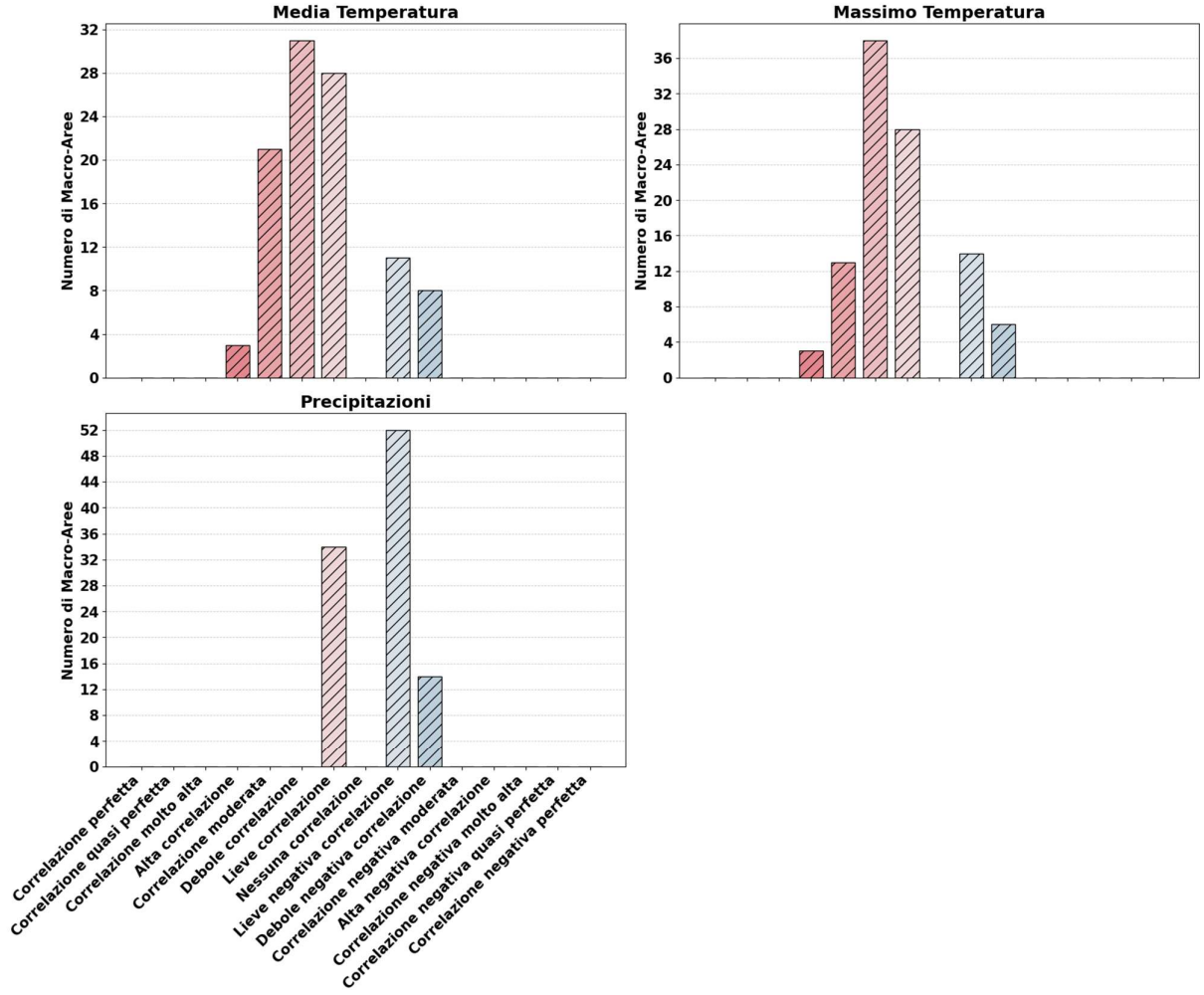

Figura A21. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

Temperatura media e massima giornaliera, precipitazione (solo nei giorni con temperatura massima >30°C)

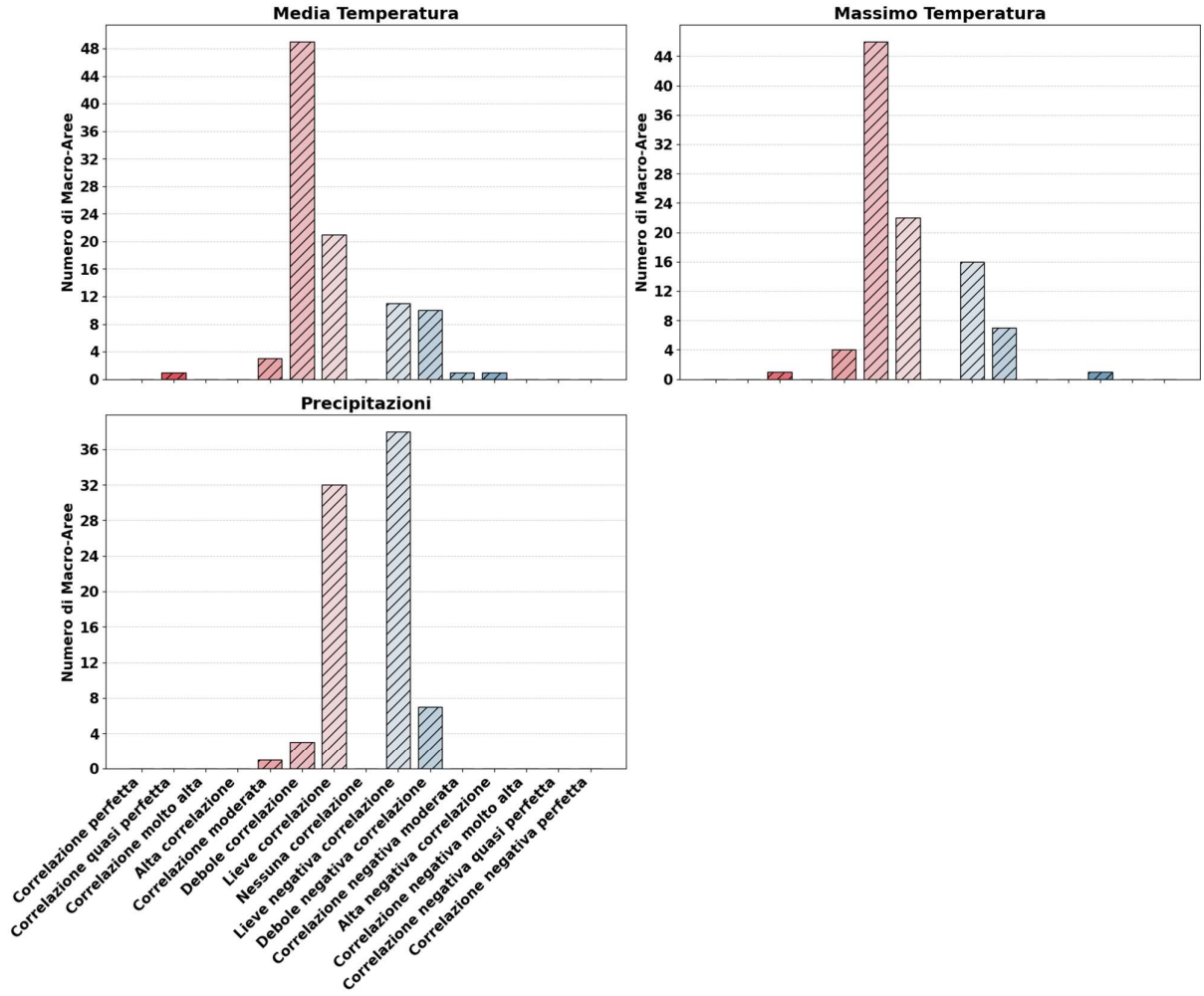

Figura A22. Distribuzione delle serie simboliche tra le diverse classi di correlazione (aggregazione giornaliera, database detrendizzato) considerando la variabile climatica riportata nel titolo dei grafici.

Bibliografia

Alshaikhli, M., Aqeel, S., Valdeolmillos, N., Fathima, F., & Choe, P. (2021). A Multi-Linear Regression Model to Predict the Factors Affecting Water Consumption in Qatar. IOP Conference Series: Earth and Environmental Science, 691(1).

Avni, N., Fishbain, B., & Shamir U. (2015). Water consumption patterns as a basis for water demand modeling. Water Resources Research, 51(10), 8165-8181.

Balling, R.C., & Gober, P. (2007). Climate variability and residential water use in the city of Phoenix, Arizona. Journal of Applied Meteorology and Climatology, 46(7), 1130-1137.

Calinski, T. & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis: Communications in Statistics. Theory and Methods, 3, 1-27.

Cornes, R., van der Schrier, G., van den Besselaar, E.J.M., & Jones, P.D. (2018). An Ensemble Version of the E-OBS Temperature and Precipitation Datasets. Journal of Geophysical Research: Atmospheres, 123(17), 9391-9409.

Davies, D.L. & Bouldin, D.W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1, 224-227.

Edossa, D.C., Babel, M.S., & Das Gupta, A. (2010). Drought analysis in the Awash river basin, Ethiopia. Water Resources Management, 24(7), 1441-1460.

Everitt, B. (1980). Cluster analysis, Quality & Quantity: International Journal of Methodology, 14, (1), 75-100.

Fleig, A.K., Tallaksen, L.M., Hisdal, H., & Demuth, S. (2006). A global evaluation of streamflow drought characteristics. Hydrology and Earth System Sciences, 10(4), 535-552.

Fraley, C. & Raftery, A. (1998). How Many Clusters? Which clustering method? Answers via Model-Based Cluster Analysis. The Computer Journal, 41, 578-588.

Gato, S., Jayasuriya, N., & Roberts, P. (2007). Forecasting residential water demand: Case study. Journal of Water Resources Planning and Management, 133(4), 309-319.

Guttman, N.B. (1999). Accepting the standardized precipitation index: a calculation algorithm. Journal of the American Water Resources Association, 35(2), 311-322.

IPCC, 2021. Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. In Press.

Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Locally adaptive dimensionality reduction for indexing large time series database. In: Proceedings of the 2001 ACM SIGMOD international conference on Management of data (pp. 151-162).

Lopez, J.J., Aguado, J.A., Martın, F., Munoz, F., Rodrıguez, A., & Ruiz, J.E. (2011). Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers. Electric Power Systems Research, 81(2), 716-724.

Martínez-Fernández, J., Ruiz-Benito, P., Bonet, A., & Gómez, C. (2019). Methodological variations in the production of CORINE land cover and consequences for long-term land cover change studies. The case of Spain. International Journal of Remote Sensing, 40(23), 8914–8932.

McKee, T.B., Doesken, N.J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scale. In: Proceedings of the Eighth Conference on Applied Climatology, Anaheim, California, 17-22 January 1993. Boston, American Meteorological Society, 179-184.

Miller, J. (1991). Reaction time analysis with outlier exclusion: bias varies with sample size. The Quarterly Journal of Experimental Psychology Section A, 43(4), 907-912.

Padulano, R., & Del Giudice, G. (2018). A Mixed Strategy Based on Self-Organizing Map for Water Demand Pattern Profiling of Large-Size Smart Water Grid Data. Water Resources Management, 32(11), 3671–3685.

Padulano, R., & Del Giudice, G. (2020). A nonparametric framework for water consumption data cleansing: An application to a smart water network in Naples (Italy). Journal of Hydroinformatics, 22(4), 666–680.

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., & Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Applied Energy, 87(11), 3538–3545.

Rokach, L. & Maimon, O. (2005). Clustering Methods. In: Maimon, O. & Rokach, L., Eds., Data Mining and Knowledge Discovery Handbook, Springer, Berlin, 321-352.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In Journal of Computational and Applied Mathematics (Vol. 20).

Sancho-Asensio, A., Navarro, J., Arrieta-Salinas, I., Armendáriz-Iñigo, J.E., Jiménez-Ruano, V., Zaballos, A., & Golobardes, E. (2014). Improving data partition schemes in smart grids via clustering data streams. Expert Systems with Applications, 41(13), 5832-5842.

Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., et al. (2020). Predicting standardized streamflow index for hydrological drought using machine learning models. Engineering Applications of Computational Fluid Mechanics, 14(1), 339-350.

Shukla, S. & Wood, A.W. (2008). Use of a standardized runoff index for characterizing hydrologic drought. Geophysical Research Letters, 35(2).

Slavíková, L., Malý, V., Rost, M., Petružela, L., & Vojáček, O. (2013). Impacts of Climate Variables on Residential Water Consumption in the Czech Republic. Water Resources Management, 27(2), 365–379.

Stagge, J.A., Tallaksen, L.M., Gudmundsson, L., Van Loon, A.F., & Stahle, K. (2015). Candidate distributions for climatological drought indices (SPI and SPEI). International Journal of Climatology 35(13), 4027-4040.

Syme, G. J., Shao, Q., Po, M., & Campbell, E. (2004). Predicting and understanding home garden water use. Landscape and Urban Planning, 68(1), 121–128.

Timotewos, M.T., Barjenbruch, M., & Behailu, B.M. (2022). The Assessment of Climate Variables and Geographical Distribution on Residential Drinking Water Demand in Ethiopia. Water, 14(11), 1722.

Van Loon, A.F. & Van Lanen, H.A. (2013). Making the distinction between water scarcity and drought using an observation-modeling framework. Water Resources Research, 49(3), 1483-1502.

Vicente-Serrano, S.M., Beguería, S., & López-Moreno, J.I. (2010). A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. Journal of Climate, 23, 1696-1718.

World Meteorological Organization (2012). Standardized Precipitation Index User Guide (M. Svoboda, M. Hayes and D. Wood). (WMO-No. 1090), Geneva.

Zhou, K.l., Yang, S.l., & Shen, C. (2013). A review of electric load classification in smart grid environment. Renewable and sustainable energy reviews, 24, 103-110.