Università degli Studi della Basilicata

SCUOLA DI INGENGERIA


CORSO DI DOTTORATO IN INGEGNERIA PER L'INNOVAZIONE E LO SVILUPPO SOSTENIBILIE

XXXVI ciclo

# Studio delle potenzialità delle tecniche di machine learning per valutare l'impatto delle variazioni del clima sull'ambiente e la salute umana

**Relatore**
Prof. Vito Telesca

**Coordinatrice**
Prof.ssa Aurelia Sole

**Laureando**
Gianfranco Castronuovo
Matr. 62649

ANNO ACCADEMICO 2022/2023

# Abstract

The thesis delves into the integration of machine learning with hydrology and public health to assess climate impact. It is structured in two main parts, each addressing a distinct aspect of this intersection. The first part focuses on improving precipitation simulation through a combined approach of statistical downscaling and machine learning. It specifically targets the shortcomings in predicting intermediate seasons and extreme precipitation events. This is achieved by enhancing Non-homogeneous Hidden Markov Models (NHMMs) with machine learning techniques, demonstrated in the case study of the Agro-Pontino Plain. Here, the model incorporates pre-processed atmospheric predictors to address NHMMs' limitations in representing extreme precipitation and seasonal variations, thereby providing a more accurate and robust simulation of rainfall patterns. In the second part, the thesis shifts its focus to the health impacts of climate change, particularly on cardiovascular diseases (CVD). Utilizing data from the Giovanni XIII Polyclinic in Bari, Italy, and incorporating weather and air quality data, a Random Forest machine learning model was developed to simulate trends in hospital admissions for CVD. The model's performance was rigorously evaluated, and the SHapley Additive exPlanations (SHAP) method was applied to ascertain the importance of various features. The study found that atmospheric pressure, minimum temperature, and carbon monoxide levels are critical factors influencing CVD-related hospitalizations, with atmospheric pressure being the most significant contributor. This research underscores the significance of integrating machine learning into the study of climate change's impacts on both environmental and health aspects. It highlights the critical role of climate variables in public health and provides a comprehensive framework for policymakers and healthcare professionals to mitigate the adverse effects of climate change on cardiovascular health. The thesis thus offers a multidimensional perspective on the climate crisis, combining advanced machine learning techniques with practical applications in environmental and health policy.

# Sommario

# Premessa: Importanza del Machine Learning e dell'Impatto Climatico in Ambiti Multidisciplinari – Idrologia e salute umana

La convergenza tra i due segmenti di ricerca presentati in questa tesi non si limita all'uso innovativo delle tecniche di machine learning per affrontare problemi distinti, ma si estende all'intersezione più ampia dell'impatto del cambiamento climatico nella nostra società. Il primo studio, incentrato sulla previsione degli eventi idrologici, dimostra come l'intelligenza artificiale possa essere impiegata per interpretare e prevedere fenomeni naturali complessi, offrendo importanti strumenti per gestire gli effetti diretti delle variazioni climatiche sull'ambiente circostante. D'altra parte, l'analisi dell'impatto delle variabili climatiche sulle patologie cardiovascolari svela una dimensione meno immediata, ma altrettanto critica, del cambiamento climatico: la sua incidenza sulla salute umana. La capacità di utilizzare il machine learning per identificare e quantificare questi collegamenti rappresenta un passo avanti significativo nella nostra capacità di prevenire e rispondere ai rischi sanitari emergenti.

Se considerati singolarmente, entrambi gli studi possono offrire un contributo rilevante ai rispettivi campi di applicazione. Tuttavia, quando vengono visualizzati insieme, emerge una narrazione più grande. Questa narrazione sottolinea l'urgenza e la necessità di una comprensione multidisciplinare dell'impatto del cambiamento climatico, dove le tecniche di machine learning servono come ponte, unificando e amplificando le nostre risposte sia a livello ambientale che sanitario. In questo contesto, la versatilità e la trasversalità delle tecniche di machine learning si rivelano non solo come strumenti di ricerca, ma come alleati essenziali nella lotta contro le conseguenze trasversali del cambiamento climatico.

# Parte I: Un modello combinato di downscaling statistico e machine learning per la simulazione delle precipitazioni

## Introduzione NHMM

Influenzando gli estremi delle precipitazioni, la stagionalità e quindi alluvioni e siccità, l'impatto del cambiamento climatico colpisce le comunità locali, minando la crescita economica, la sicurezza alimentare e aumentando le disparità e le ineguaglianze sociali. A tal proposito, le Nazioni Unite hanno sviluppato l'Agenda 2030 per lo Sviluppo Sostenibile. Al suo centro ci sono gli Obiettivi di Sviluppo Sostenibile (SDG) che intendono affrontare - tra gli altri - crescente povertà, insicurezza alimentare, problemi di salute, scarsità d'acqua, infrastrutture danneggiate, che sono tutti potenzialmente influenzati dal cambiamento climatico [1].

Eventi estremi come forti piogge, alluvioni e siccità sono le principali cause di perdita di vite umane e danni economici [2]. Il danno causato dalle alluvioni è particolarmente alto nelle aree costiere dove l'alta densità di popolazione e il conseguente alto livello di urbanizzazione e l'uso intensivo del terreno per attività agricole e industriali rendono l'area particolarmente vulnerabile a tali eventi estremi [3]. Le siccità prolungate riducono la disponibilità d'acqua e, in contesti particolarmente vulnerabili, come quelli di alcune regioni mediterranee, causano l'aumento delle malattie, la fame e la destabilizzazione delle strutture sociali e politiche. C'è quindi un forte interesse nello sviluppo di metodi che consentano una previsione futura della frequenza, intensità e conseguenze di tali eventi. Negli ultimi anni, la crescente consapevolezza che l'emissione di gas serra nell'atmosfera potrebbe portare ad un aumento dell'intensità e della frequenza degli eventi estremi ha messo in discussione le ipotesi di casualità, stazionarietà e analisi del processo limitato alla scala del bacino, che sono alla base degli studi idrologici/idraulici. Pertanto, è necessario ampliare l'approccio metodologico che include processi deterministici su scala globale, come l'evoluzione temporale e spaziale delle strutture di circolazione atmosferica e i processi di trasporto dell'umidità su scala planetaria [4]. Questo ampliamento dell'approccio metodologico, che cerca di dedurre i processi su scala regionale o di bacino su una scala planetaria, pone una serie di sfide. Le relazioni e la loro formalizzazione tra processi globali e locali presentano ancora molte limitazioni, incertezze e problemi, in particolare riguardo agli eventi estremi. In questo contesto, la Commissione Europea sottolinea l'importanza dello sviluppo di metodi e strumenti predittivi nel Libro Bianco sull'Adattamento

ai Cambiamenti Climatici. In accordo con [5], la metodologia centrale utilizzata nella valutazione delle conseguenze dei cambiamenti climatici sui processi idrologici consiste nei seguenti passi:

1) valutare le prestazioni dei Modelli di Circolazione Generale (GCM) confrontando le simulazioni di questi con i dati generati dai modelli di rianalisi per selezionare il più adatto al nostro caso di studio
2) implementare tecniche di downscaling che associano gli output dei GCM identificati ai regimi di pioggia osservati a scala locale
3) sviluppare modelli che simulano a scala di bacino gli effetti dei cambiamenti climatici.

La simulazione della distribuzione spaziale delle altezze di pioggia è essenziale per valutare l'impatto dei cambiamenti climatici sulle attività agricole, sulla gestione del rischio di alluvioni, sui processi di deflusso delle piogge e sulla gestione delle acque superficiali e sotterranee. Le probabilità di pioggia giornaliere, la persistenza dei regimi umidi e secchi e altre statistiche sulle piogge possono variare sostanzialmente nel tempo e nello spazio in modo sistematico. Le fluttuazioni climatiche sono il principale motore dei cambiamenti periodici nelle probabilità di pioggia e influenzano le statistiche delle precipitazioni nello spazio e nel tempo.

La necessità di previsioni basate su modelli fisici piuttosto che su modelli puramente statistici ha portato allo sviluppo dei Modelli di Circolazione Generale (GCM). I GCM si comportano abbastanza bene per le condizioni meteo annuali e stagionali su una vasta scala spaziale e rappresentano una delle principali fonti per le proiezioni climatiche. I risultati dei GCM costituiscono la base della maggior parte delle valutazioni dell'impatto climatico; sebbene tendano a sovrastimare la frequenza e a sottostimare l'intensità delle precipitazioni giornaliere rispetto ai record storici a livello locale [6]. Le restrizioni associate ai modelli di circolazione generale (GCM), che operano su griglie di calcolo orizzontale estese tra i 250 e i 600 km [7], hanno catalizzato lo sviluppo di metodi statistici per modellare le precipitazioni su scala locale in relazione ai GCM. In questo ambito, si sono affermate le tecniche di downscaling statistico per colmare il divario di scala spaziale tra le griglie dei GCM e le osservazioni meteorologiche puntuali. È rilevante notare che, al di sotto dei 200 km, i GCM non riescono a fornire dati affidabili per variabili essenziali, come quelle necessarie per una modellazione accurata dei processi idrogeologici [8]. Il downscaling, pertanto, viene definito come il processo di adattamento tra i vasti campi spaziali dei GCM e le variabili idrologiche rilevanti a livello locale, quali le precipitazioni giornaliere misurate con pluviometri [9]. Nei modelli di downscaling statistico, gli effetti locali sono catturati associando i parametri statistici delle distribuzioni di probabilità della pioggia alle variabili climatiche di grande scala (come temperatura, geopotenziale e venti) per catturare la frequenza e l'intensità degli eventi idrologici [10]. Le

equazioni del moto e dell'energia nei modelli dinamici vengono integrate in modo simile ai modelli globali ma con una risoluzione maggiore e una regione limitata. Le condizioni al contorno dei modelli dinamici sono ottenute dalle simulazioni effettuate con i modelli globali. I modelli di downscaling dinamici e statistici vengono calibrati confrontando le simulazioni con le osservazioni storiche. In generale, i modelli statistici sono preferiti a quelli dinamici a causa del costo computazionale significativamente inferiore e della maggiore precisione. Alcune ricerche [11] [12] identificano diversi meccanismi indotti dai cambiamenti nei gradienti di temperatura globale, che influenzano la circolazione atmosferica e, quindi, di conseguenza, i flussi di umidità, che spiegano i cambiamenti osservati e simulati (da GCM) nello spazio e nei modelli di pioggia nel tempo. È ragionevole ipotizzare che queste variazioni nei gradienti di temperatura possano influenzare anche la variabilità stagionale delle precipitazioni locali, causando, ad esempio, spostamenti stagionali che, altrimenti, non potrebbero essere catturati attraverso una demarcazione "a priori" delle stagioni.

Molti ricercatori hanno sviluppato modelli di downscaling basati su modelli Markov non omogenei con stati nascosti (NHMM). L'NHMM è un modello doppiamente stocastico in cui vengono generate serie temporali multivariate, utilizzando una specifica distribuzione, condizionalmente a uno stato nascosto identificato, gli stati nascosti del modello subiscono transizioni markoviane. Tale metodo può fornire informazioni utili sulle modalità di circolazione e sui meccanismi di precipitazione associati. L'NHMM è stato inizialmente sviluppato per modellare i modelli di occorrenza delle piogge [13] ed è stato esteso per includere le quantità [14]. Hughes et al. 1994 ha utilizzato l'NMM per modellare una sequenza di dati invernali di 15 anni per 30 stazioni pluviometriche nel sud-ovest dell'Australia. I risultati hanno mostrato che il modello riproduce accuratamente la probabilità di pioggia osservata e fornisce alcune informazioni utili sul processo di precipitazione nel sud-ovest dell'Australia [15]. Sempre in Australia è stato utilizzato l'NHMM per monitorare i cambiamenti climatici, scoprendo che può essere un utile modello di downscaling [16]. Charles et al. 1999 ha esteso l'NHMM alle altezze delle pioggia, simulando accuratamente le curve di intensità della durata, la distribuzione delle accumulazioni giornaliere e le correlazioni tra diversi siti per le altezze delle piogge giornaliere [17]. Bellone et al. 2000 ha utilizzato l'NHMM per modellare le quantità di pioggia indipendentemente in ogni stazione pluviometrica, verificando che il modello risponda ai cambiamenti della circolazione atmosferica a seconda della posizione [18]. Robertson et al. 2007 ha combinato l'NHMM con un modello di coltura per studiare la disaggregazione spaziale e temporale delle piogge stagionali, simulando accuratamente la resa del mais in dieci stazioni nel sud-est degli Stati Uniti [19]. Robertson et al. 2009 ha utilizzato la stessa tecnicha applicata negli

Stati Uniti ad una rete di 17 stazioni nel distretto di Indramayu, in Indonesia, ottenendo risultati con livelli accurati di varianza interannuale nelle precipitazioni [20], dimostrando l'accuratezza e l'affidabilità del metodo per regioni geografiche con diverse caratteristiche climatiche. Tan et al. 2013 ha applicato un NHMM a 6 stati ai dati di precipitazione giornaliera di 20 stazioni nella penisola malese durante una sequenza di dati di 33 anni nella stagione dei monsoni [21]. La prestazione del modello è stata valutata confrontando la correlazione e il grafico quantile-quantile tra osservazione e simulazione delle precipitazioni. Ancora una volta, l'NHMM sembra riprodurre ragionevolmente la distribuzione generale delle altezze delle precipitazioni.

L'applicazione del modello NHMM al caso delle forti piogge nella Piana dell'Agro-Pontino in Italia [22] ha identificato significative relazioni statistiche tra le piogge su scala locale e i predittori atmosferici. Sono stati analizzati dati atmosferici provenienti dall'archivio NCEP/NCAR e registrazioni di 56 anni (1951-2004) di misurazioni quotidiane delle precipitazioni provenienti da 7 stazioni nella Piana dell'Agro-Pontino. Sono stati effettuati diversi test di convalida per identificare il miglior set di predittori atmosferici per modellare le precipitazioni locali e valutare le prestazioni del modello nella cattura della variabilità inter-annuale e stagionale delle precipitazioni, nonché nei modelli di pioggia media ed estrema. In [22] e [23], l'NHMM è stato costruito senza alcuna demarcazione "a priori" delle stagioni; infatti, la variabilità delle precipitazioni è stata considerata solo come funzione della variazione temporale dei predittori atmosferici. Hanno ottenuto buoni risultati, che sottolineano anche la non linearità tra i predittori atmosferici e le precipitazioni.

Nonostante i risultati positivi ottenuti, è stato notato che il modello NHMM ha difficoltà nel riconoscere le stagioni. Come mostrato in [22] quando si applica l'NHMM a regimi temperati:

1) sottostima l'intensità delle piogge nella stagione autunnale e sovrastimare quelle della stagione primaverile,
2) presenta difficoltà nel rappresentare gli estremi delle precipitazioni,
3) necessita di procedure euristiche per identificare l'ottimo set di predittori atmosferici, che sono spesso in grande numero.

[23] ha risolto quest'ultimo problema dimostrando che un piccolo numero di predittori atmosferici può ottenere prestazioni paragonabili o addirittura migliori rispetto agli studi precedenti. In [23], un modello NHMM è stato applicato in Florida, mostrando la robustezza e l'efficacia del metodo nell'identificare la stagionalità e gli estremi idrologici utilizzando solo due predittori atmosferici: il campo dell'Altezza Geopotenziale a 850 hPa (GPH) e il Trasporto di Vapore Integrato (IVT).

L'obiettivo della ricerca è il superamento dei limiti precedentemente descritti.

Una via per aumentare l'accuratezza del modello, sia nel cogliere le stagioni intermedie, sia nel valutare gli estremi di precipitazione, è rappresentata dall'utilizzo di modelli i cui obiettivi siano quelli di risolvere le dipendenze non lineari che legano le piogge alle variabili atmosferiche e le dipendenze temporali tra i vari stati idrologici. Al momento non esistono singoli modelli in grado di risolvere entrambi i passaggi ma, in tale contesto, una combinazione di modelli, ottenuta attraverso ensemble models, sembra adatta per superare tali limiti. Le diverse configurazioni della combinazione di modelli può dare vita a tre tipi di ensamble models:

1) *Bagging models:* consiste in tanti modelli uguali (ad esempio tanti modelli ad albero) calibrati su diverse parti di un dataset (il risultato sarà un modello "foresta"),
2) *Boosting models:* sono modelli che si adattano per iterazioni dando più peso alle zone del dataset in cui l'errore dei modelli delle iterazioni precedenti è maggiore,
3) *Stacking models:* sono modelli diversi, disposti in successione a formare dei layer. L'input del primo layer sono le variabili in ingresso trattate, il suo output è l'input del layer successivo, fino a che l'ultimo layer restituisce come output il risultato atteso del modello.

In particolare, con il presente lavoro di ricerca si indagherà l'efficacia dei modelli di Stacking per migliorare l'approccio già utilizzato in [23]. L'idea consiste nello sviluppo di un modello di classificazione a monte dell'NHMM, secondo schemi in grado di catturare la non linearità intrinseca delle dipendenze tra le precipitazioni e i predittori atmosferici. Tale approccio propone di aumentare sia l'efficacia nel riconoscere le fluttuazioni stagionali sia la capacità di catturare accuratamente gli estremi delle precipitazioni. Questo modello produrrà una o più variabili lineari che diventano l'input del modello successivo, un modello NHMM.

Lo scopo di questo lavoro di tesi è contribuire allo sviluppo di uno strumento per l'analisi e la previsione degli eventi idrologici estremi e non estremi a livello locale che tiene conto dell'impatto dei cambiamenti climatici e della non stazionarietà delle serie naturali. La regione del Lazio, situata in Italia, è stata scelta come area di studio, data la sua esposizione ai cambiamenti climatici, l'economia agroalimentare sviluppata, la ricca biodiversità e la presenza di numerose zone di bonifica costiere. L'obiettivo primario è migliorare l'efficacia delle previsioni delle precipitazioni, un elemento cruciale per la gestione delle risorse idriche, la prevenzione delle inondazioni, la pianificazione agricola e la sicurezza ambientale. Affrontare l'incertezza crescente associata ai cambiamenti climatici e alle variazioni temporali richiede l'adozione di nuovi approcci avanzati e metodologie innovative. Nella ricerca, è stata data priorità all'applicazione di tali approcci avanzati nella regione del Lazio, dove le sfide connesse ai cambiamenti climatici assumono particolare rilevanza. L'obiettivo consiste nel fornire

strumenti decisionali basati su previsioni precise delle precipitazioni agli stakeholder locali, tra cui autorità, operatori agricoli e gestori delle risorse idriche, al fine di contribuire alla mitigazione degli impatti delle precipitazioni estreme e al miglioramento della gestione delle risorse idriche, promuovendo così la sicurezza e la sostenibilità ambientale nella regione. La ricerca rappresenta un avanzamento significativo nell'ambito della comprensione delle dinamiche delle precipitazioni e delle loro interazioni con il clima. Campi atmosferici di Integrated Vapor Transport (IVT) e Geopotential Height (GPH) a 850 hPa sono stati adottati come predittori atmosferici, basandosi su risultati promettenti precedentemente identificati in studi precedenti [23]. Per la formazione e la validazione del modello, è stato utilizzato un database di 55 anni (1951-2004) di dati pluviometrici giornalieri provenienti da 35 stazioni meteorologiche situate nel territorio del Lazio.

# Materiali e Metodi

## Dati

Per la costruzione del modello, sono stati utilizzati dati atmosferici e pluviometrici. È stata considerata la serie storica relativa ai campi di variabili atmosferiche, con una risoluzione spaziale di 0,28125° x 0,28125°, considerando latitudini tra 80N e 20S e longitudini tra -90W e 70E. Questi dati provengono dal dataset di rianalisi "ERA5" dell'ECMWF, disponibile per il download presso il Copernicus Climate Change Service (C3S) nel Climate Data Store [24]. In dettaglio, le variabili atmosferiche considerate includono:

1) Altezza di geopotenziale a 850hPa (GPH850);
2) Venti zonali e meridionali e umidità specifica misurata tra i livelli di 1000hPa e 300hPa.

Il punto 2 è stato essenziale per calcolare il Trasporto Integrato del Vapore (IVT). Il calcolo dell'IVT ha permesso di quantificare il trasporto orizzontale di umidità nella colonna d'aria tra 1000 e 300 hPa.

Per questo studio è utilizzato un archivio di 55 anni (1951-2005) di rilevazioni giornaliere di precipitazione, raccolte da 36 stazioni relative a differenti bacini della regione del Lazio.

## Metodologia

È stata effettuata una comparazione tra due metodi per la simulazione delle altezze delle precipitazioni. Il primo (Fig. 1) è il metodo classico utilizzato in precedenti lavori di ricerca che impiega direttamente predittori atmosferici nel NHMM. Il secondo (Fig. 2), invece, rappresenta un nuovo metodo che prevede l'introduzione di un modello di machine learning di tipo Stacking

che combina altri modelli di machine learning e NHMM, utilizzando l'abilità di modelli diversi per essere performante su aspetti diversi del problema. Per la simulazione delle precipitazioni, questo nuovo modello, non sfrutta più direttamente i predittori atmosferici, ma le probabilità giornaliere di pioggia relative a sei percentili derivanti dall'output dello Stacking. L'obiettivo principale nell'adottare il nuovo modello è quello di utilizzare predittori più interpretabili e con caratteristiche meno non lineari per il modello NHMM. Questo si ottiene introducendo questi predittori in modo indiretto, anziché inserire direttamente i predittori atmosferici all'interno dell'NHMM.

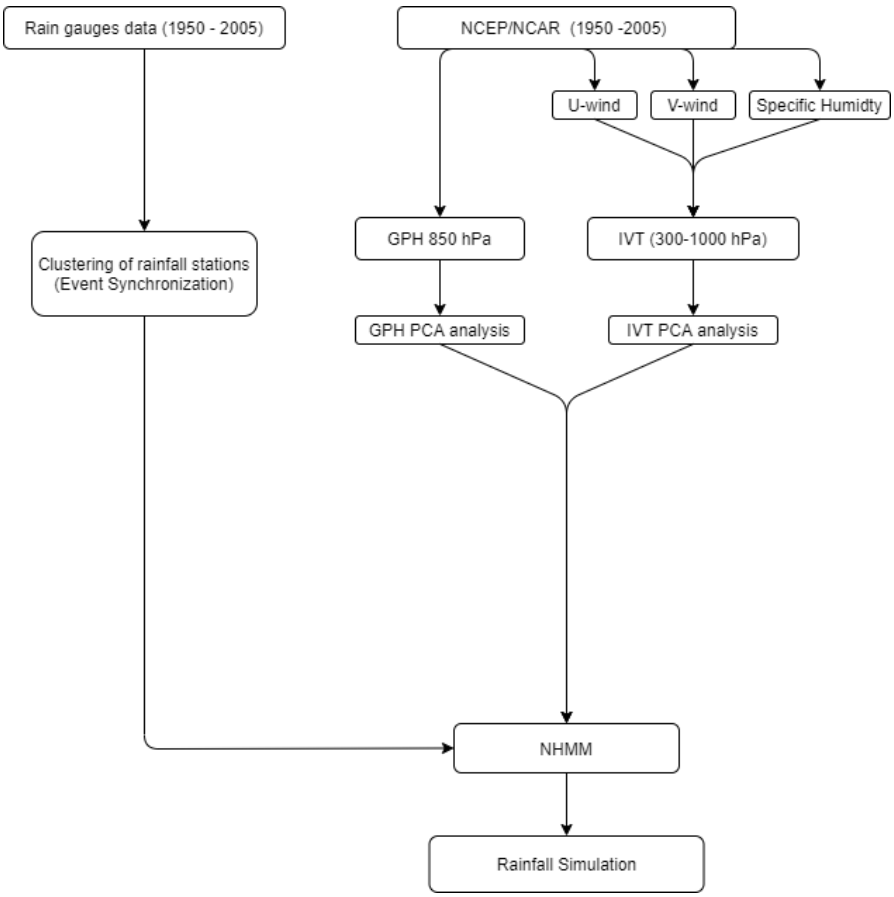


*Figura 1. Schema metodologico per la simulazione delle precipitazioni proposto da Conticello et al.2018 [23]*

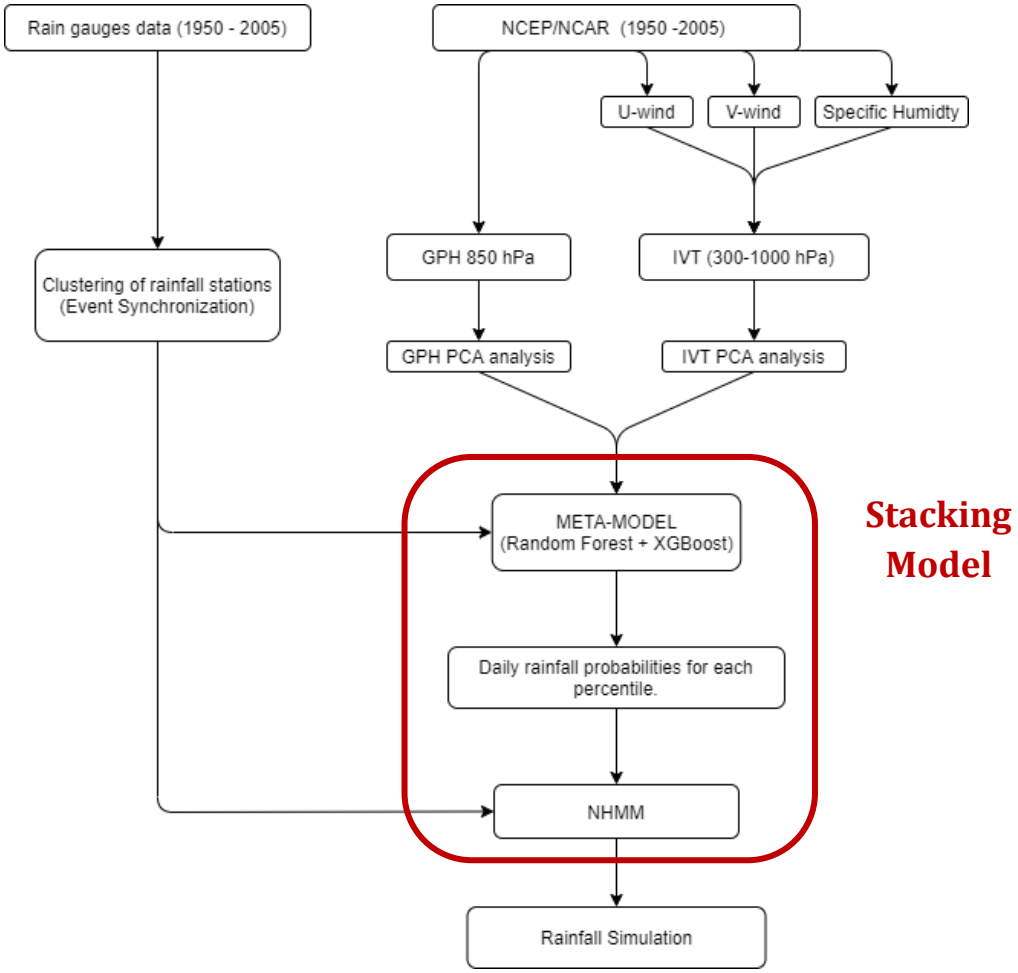


*Figura 2. Schema metodologico per la simulazione delle precipitazioni implementato con l'introduzione di un modello di machine learning di tipo Stacking*

## Data preprocessing

### Clustering

A causa della vasta estensione territoriale della regione Lazio, l'intero dataset delle precipitazioni (35 stazioni) è stato suddiviso in cluster di stazioni con condizioni meteorologiche locali simili, utilizzando le tecniche di Event Synchronization e Modularity. Questa metodologia è stata applicata alle serie temporali con una finestra temporale pari a $\pm 1$ giorno. Questa finestra temporale garantisce che gli HPEs, registrati in diversi pluviometri e che mostrano un alto grado di sincronizzazione tra loro, siano causati dalle stesse caratteristiche meteorologiche.

L'Event Synchronization è un metodo che misura la correlazione tra due serie temporali binarie in una data finestra temporale. La versione di Event Synchronization utilizzata in questo studio deriva da una combinazione di [25] e [26] ed è descritta nella sua versione finale in [23]. La formula principale che descrive l'approccio è:

$$Q^{\tau} = \frac{s^{\tau}(x|y)+s^{\tau}(y|x)}{m_{x+}m_{y}} \qquad (1)$$

dove $Q$ rappresenta la sincronizzazione tra due serie temporali binarie, denominate $x$ e $y$, in una finestra temporale $\tau$; $s(x|y)$ rappresenta il numero di volte in cui gli eventi della serie temporale $y$ hanno influenzato gli eventi nella serie temporale $x$; $s(y|x)$ rappresenta il numero di volte in cui gli eventi della serie temporale $x$ hanno influenzato gli eventi nella serie temporale $y$; $m_x$ e $m_y$ sono il numero di eventi nelle serie temporali $x$ e $y$, rispettivamente.

Classificazione delle precipitazioni

Gli accumuli giornalieri sono stati trasformati in sei classi di precipitazione relative a 6 percentili, elaborati per ogni cluster di stazioni omogenee. La Classe 0 rappresenta valori NON-piovosi o altezze di precipitazione giornaliera inferiori a 1 mm. Le altre cinque classi di precipitazione rappresentano rispettivamente: 20° percentile, 40° percentile, 60° percentile, 80° percentile e 90° percentile.

Analisi PCA

Inoltre, i dati dei predittori atmosferici sono stati analizzati tramite la tecnica PCA per ridurre le dimensioni dei campi di input atmosferici. Ridurre i gradi di libertà del problema garantisce un addestramento più semplice e una strutturazione del modello. Un modello con troppi gradi di libertà tende ad adattarsi troppo al set di dati di addestramento, fornendo previsioni inaccurate quando applicato a nuovi dati atmosferici. L'uso della tecnica PCA (Principal Component Analysis) rappresenta un pilastro fondamentale nell'ambito dell'analisi dei dati, permettendo di navigare e interpretare con efficienza grandi volumi di informazioni. In primo luogo, grazie alla PCA, si consegue un'efficienza computazionale notevole. La capacità di ridurre la dimensionalità dei dati consente di accelerare le analisi, rendendo le elaborazioni non solo più veloci, ma anche più gestibili, soprattutto quando ci si confronta con dataset di grandi dimensioni. Questo si lega strettamente al secondo vantaggio, ovvero la facilitazione della visualizzazione dei dati. La trasformazione dei dati in spazi a due o tre dimensioni rende possibile una rappresentazione grafica intuitiva, permettendo di evidenziare con chiarezza le relazioni e le strutture sottostanti. Tuttavia, non si tratta solo di una questione di semplificazione. La PCA gioca un ruolo cruciale nella mitigazione della multicollinearità. Le applicazioni analitiche, come la regressione lineare, possono essere fortemente influenzate dalla presenza di variabili fortemente correlate tra loro. La PCA, tramite la creazione di componenti principali ortogonali, elimina queste correlazioni, garantendo stabilità e chiarezza interpretativa dei risultati. A ciò si aggiunge la capacità della PCA di rilevare pattern e tendenze dominanti nei dati, offrendo una visione chiara delle informazioni cruciali e aiutando i ricercatori a discernere le dinamiche fondamentali che guidano il dataset. Infine, ma non meno importante, è la riduzione del rumore. Ogni dataset porta con sé variazioni che possono non contribuire significativamente all'informazione principale, agendo come

semplice "rumore" di fondo. La PCA, focalizzandosi sulle componenti che rappresentano le informazioni fondamentali, consente di isolare e rimuovere tali distorsioni, garantendo un'analisi più pulita e precisa. In sintesi, l'adozione della PCA è essenziale per fornire una base solida e informativa nell'analisi dei dati, consentendo di distillare informazioni complesse in insight chiari e gestibili.

## Modello Stacking

I campi atmosferici ridotti grazie alla tecnica PCA costituiranno l'input del modello Stacking.
Lo Stacking è un algoritmo di apprendimento automatico basato su ensemble che apprende come combinare al meglio le previsioni provenienti da diversi modelli di apprendimento automatico che offrono buone prestazioni. Nei metodi statistici e di apprendimento automatico, i metodi di ensemble utilizzano diversi algoritmi di apprendimento per ottenere una performance predittiva migliore rispetto a quella che potrebbe essere ottenuta da ciascuno dei singoli algoritmi di apprendimento. A differenza di un ensemble statistico nella meccanica statistica, che è solitamente infinito, un ensemble di apprendimento automatico è composto solo da un insieme finito e concreto di modelli alternativi, ma consente tipicamente una struttura molto più flessibile tra queste alternative. Lo Stacking è progettato per migliorare le prestazioni di modellazione, in particolare quando esistono relazioni complesse e fortemente non lineari tra gli input e l'output.
L'architettura di un modello di Stacking (Fig. 3) coinvolge due o più modelli di base, spesso denominati modelli di livello-0, e un meta-modello che combina le previsioni dei modelli di base, noto come modello di livello-1.

1) Modelli di Livello-1 (o Modelli di Base): Modelli adattati sui dati di addestramento e le cui previsioni vengono compilate.
2) Modello di Livello-2 (o Meta-Modello): Modello che apprende come combinare al meglio le previsioni dei modelli di base.

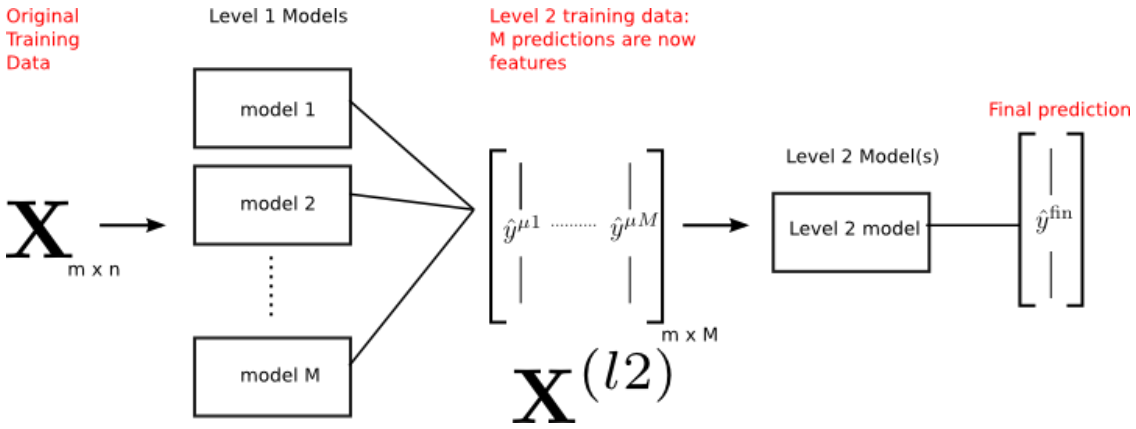


*Figura 3. Struttura di un modello di tipo Stacking*

I dati di addestramento iniziali *X* hanno m osservazioni e n campi. Ci sono M modelli diversi addestrati su *X*. Ogni modello fornisce previsioni per l'outcome (*y*) che vengono poi inserite in un secondo set di dati di addestramento $X^{l2}$. Le M previsioni diventano caratteristiche per questi dati di secondo livello. Un modello di secondo livello (o modelli) può quindi essere addestrato su questi dati per produrre gli outcome finali che saranno utilizzati per le previsioni.

*Modelli di Livello-1: Random Forest e Extreme Gradient Boosting*

Lo Stacking è appropriato quando diversi modelli di machine learning sono efficaci su un dataset ma lo sono in modi diversi. Un altro modo di dirlo è che le previsioni fatte dai modelli o gli errori nelle previsioni fatte dai modelli sono non correlati o hanno una bassa correlazione. È spesso una buona idea utilizzare una gamma di modelli che fanno ipotesi molto diverse su come risolvere il compito di modellazione predittiva. Qui, come modello di base, è stata utilizzata una combinazione di modelli bagging e boosting. Abbiamo utilizzato il bagging per ridurre la varianza delle previsioni e l'approccio boosting per produrre un modello robusto meno incline agli errori dei suoi componenti. Random Forest è stato scelto come modello di tipo bagging, e Extreme Gradient Boosting come modello di tipo boosting. Questi modelli sono stati addestrati con predittori atmosferici e classi percentili di precipitazioni giornaliere divise per cluster di stazioni. I modelli di Livello-0 hanno prodotto un output di tipo etichetta di classe. L'output consiste in probabilità di pioggia giornaliere, divise in sei classi, ognuna associata a un percentile specifico:

1) classe 0: No pioggia
2) classe 1: 20° percentile
3) classe 2: 40° percentile
4) classe 3: 60° percentile

5) classe 4: 80° percentile
6) classe 5: 90° percentile;

in modo che la somma delle probabilità di pioggia giornaliera di tutte e sei le classi sia uguale a 1. Questo output è stato utilizzato come input per il modello di Livello-1: NHMM.

*Modello di Livello-2: NHMM*

L'NHMM è un doppio processo stocastico costituito da: stati nascosti e una sequenza di stati osservata. Il modello scompone il campo di precipitazione giornaliera su una rete di stazioni in stati nascosti discreti, modellati come una catena di Markov di primo ordine che progredisce nel tempo. Ogni stato nascosto è associato a un regime di circolazione atmosferica distinto. La transizione degli stati nascosti è inevitabilmente influenzata da predittori su larga scala. Gli stati nascosti di qualsiasi giorno nella sequenza temporale sono determinati congiuntamente da quelli del giorno precedente e dai predittori su larga scala del giorno corrente. In tal modo, l'intera sequenza di precipitazione viene simulata stocasticamente. Senza i predittori esterni su larga scala, l'NHMM diventerebbe un semplice modello di Markov nascosto.
L'NHMM è definito con due ipotesi [15] [27].

$$P(R_t|S_{1:T}, R_{1:t-1}, X_{1:T}) = P(R_t|S_t) \qquad (2)$$
$$P(S_t|S_{1:t-1}, X_{1:T}) = P(S_t|S_{t-1}, X_t) \qquad (3)$$

in cui Rt denota un vettore multivariato che fornisce le precipitazioni osservate al tempo t, con t che è la sequenza temporale in giorni e varia da t = 1, 2 ... T. St presenta lo stato meteorologico (nascosto) al tempo t, il numero di stati nascosti è indicato con K, e quindi S = $\{q_1, q_2 \dots q_i \dots q_K\}$. Sia $R_t$ che $S_t$ sono definiti in una rete di W stazioni nell'area di studio. $X_t$ è un vettore che rappresenta il campo di circolazione atmosferica nel giorno t e quindi $X_{1:T}$ = ($X_1$, $X_2$ ... $X_T$) rappresenta la sequenza dei dati atmosferici da t=1 a t=T(e lo stesso per $R_t$ e $S_t$ ).
La prima ipotesi (1) è che la precipitazione multivariata $R_t$ al tempo t sia indipendente da tutte le altre variabili, dati gli stati nascosti St al tempo t. La seconda ipotesi (2) indica che gli stati nascosti St nel giorno t dipendono solo dal vettore predittore $X_t$ per il giorno t e dagli stati nascosti St−1 nel giorno t-1. Pertanto, la prima ipotesi afferma che il processo di pioggia è condizionalmente indipendente dato lo stato meteorologico corrente e la seconda ipotesi afferma che il processo di pioggia dipende solo dallo stato meteorologico precedente e dai dati atmosferici attuali [14].
Per la precipitazione $P(R_t|S_t)$ (condizionale agli stati nascosti St), stabiliamo una funzione δ e una funzione esponenziale per descrivere le probabilità di non-pioggia e di pioggia.

$$P(R_t|S_t = q_i) = \prod_{w=1}^{W} P\,/R_t^w = r|S_t = q_i) = \prod_{w=1}^{W} a_{iw} \qquad (4)$$

$$a_{iw} = \begin{cases} P_{iw0} & r = 0 \\ \sum_{c=1}^{2} P_{iwc}\,\gamma_{iwc} e^{-\gamma_{iwc} r} & r > 0 \end{cases}$$

dove "r" è la precipitazione osservata alla stazione "w" nel giorno t, "$q_i$" è lo stato nascosto nel giorno t, con w = 1, 2 ... W e i = 1, 2 ... K. Con "c" si indica il numero di esponenziali, "$P_{iwc}$" si riferisce al peso e "$\gamma_{iwc}$" è il parametro della funzione esponenziale. Per calcolare la matrice di transizione $P(S_t|S_{t-1}, X_t)$ può essere usata la teoria della probabilità condizionale di Bayes, per scomporla in un prodotto della matrice di transizione γji ($P(St = q_i|St\text{-}1 = q_j)$ e una funzione dei predittori atmosferici $P(Xt|St\text{-}1 = q_j, St = q_i)$. Considerando che i predittori atmosferici rilevanti $X_t$ sono solitamente variabili derivate da campi atmosferici ad alta dimensione, si può assumere che le $X_t$ siano multi-variate e normalmente distribuite. Questo porta al seguente modello per $P(S_t|S_{t-1}, X_t)$:

$$P(S_t = q_i|S_{t-1} = q_j, X_t) \propto P(S_t = q_i|S_{t-1} = q_j)P(X_t|S_{t-1} = q_j, S_t = q_i)$$

$$= \gamma_{ji} exp[-\frac{1}{2}(X_t - \mu_{ij}) \sum^{-1} (X_t - \mu_{ji})' \qquad (5)$$

Qui, $\mu_{ji}$ è la media dei predittori atmosferici associati alle transizioni dallo stato $q_j$ al giorno t-1 allo stato $q_i$ al giorno t. ∑-1 è la matrice di covarianza dei predittori atmosferici. Per garantire l'identificabilità dei parametri, vengono imposti i vincoli $\sum i\gamma_{ji} = 1$ e $\sum i\mu_{ji} = \mu_j = 0$. La stima dei parametri si ottiene con la metodologia della massima verosimiglianza. Denotando con Θ i parametri del modello, la probabilità può essere scritta come:

$$L(\Theta) = P(R_{1:T}|X_{1:T}) = \sum_{S_{1:T}} P\,(R_{1:T}, S_{1:T}|X_{1:T}$$

$$= \sum_{S_{1:T}} P\,(S_1|X_1) \prod_{t=2}^{T} P\,(S_t|S_{t-1}, X_t) \prod_{t=1}^{T} P\,(R_t|S_t) \qquad (6)$$

L'insieme di parametri Θ che massimizzano L(Θ) può essere ottenuto con l'algoritmo Baum-Welch [28], per ottenere stime dei parametri di massima verosimiglianza per modelli con variabili nascoste e/o dati mancanti. Dopo la calibrazione dell'NHMM, i parametri del modello inclusi gli stati nascosti $S = \{q_1, q_2, ... q_i ... q_K\}$, la matrice di transizione $P(S_t|S_{t-1}, X_t)$ e la precipitazione

PDF (condizionata dagli stati nascosti $S_t$) $P(R_t|S_t)$ sono completamente determinati. Rimangono invariati per tutte le simulazioni di pioggia, comprese le condizioni climatiche future.

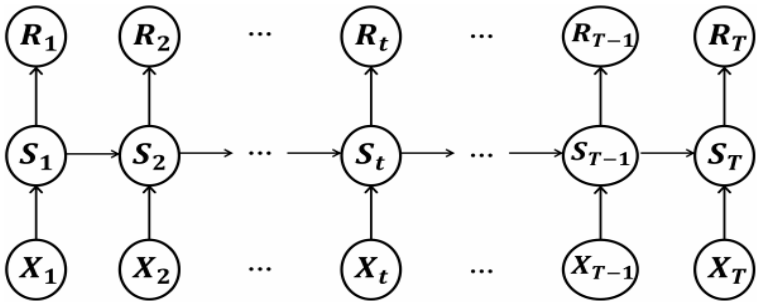


*Figura 4. Struttura di un modello NHMM. $X_{1,T}$ - predittori esterni; $S_{1,T}$ - stati nascosti; $R_{1,T}$ – altezze di pioggia*

Come mostrato dal diagramma di flusso tecnico (Fig. 4), il primo passo nell'utilizzo di un NHMM è generare una catena di Markov di stati nascosti, $S_1$, $S_2$, ..., $S_T$, basata su una sequenza di predittori atmosferici giornalieri e la matrice di transizione $P(S_t|S_{t-1}, X_t)$. Il passaggio successivo consiste nel simulare la precipitazione giornaliera "r" in base alle probabilità $P(R_{tw} = r|S_t = q_i)$. Nel caso del riscaldamento climatico futuro, i predittori atmosferici corrispondenti al riscaldamento globale porterebbero a modifiche nella frequenza degli stati nascosti con quantità di precipitazioni diverse. Alla fine, possono essere generate più (o meno) precipitazioni. Pertanto, il caso del cambiamento climatico per uno scenario futuro è esplicitamente preso in considerazione nella frequenza di ciascuno stato nascosto. Il numero di stati nascosti K è ottimizzato dal criterio dell'informazione bayesiana (BIC). Il punteggio BIC con stati K è definito come [29]:

$$BIC_K = 2L(\Theta_K^*) - plogTlogT \quad (7)$$

dove Θ*K è il vettore del parametro di massima verosimiglianza stimato, come si ottiene mediante EM applicato ai dati di addestramento per un modello con K stati, L(Θ*K è la probabilità del modello valutato a Θ*K e "p" è il numero di parametri nel modello a K stati. Il punteggio BIC minimo corrisponde a un modello ottimale che esplora i dati di addestramento [29]. L'NHMM, applicato al downscaling climatico, necessita di predittori adeguatamente selezionati. Il principio alla base della selezione dei predittori è che devono avere una buona correlazione con la precipitazione e un chiaro significato fisico.

In questa applicazione idro-climatica, uno stato nascosto rappresenta una particolare configurazione meteorologica che influenza la probabilità e la quantità di precipitazione in un certo numero di luoghi simultaneamente. In un modello NHMM, tuttavia, ogni stato nascosto è associato a un distinto regime di circolazione atmosferica. Predittori su larga scala influenzano

inevitabilmente la transizione degli stati nascosti. Gli stati nascosti di ogni giorno nel cronogramma sono determinati da quelli del giorno precedente e dai predittori su larga scala del giorno corrente. L'NHMM è stato costruito senza alcuna demarcazione "a priori" delle stagioni. La variabilità delle precipitazioni è stata considerata solo come funzione delle variazioni temporali dei predittori atmosferici. Alcune pubblicazioni [11]identificano diversi meccanismi indotti da cambiamenti nei gradienti di temperatura globale, che influenzano la circolazione atmosferica e, quindi, di conseguenza, i flussi di umidità, che spiegano i cambiamenti osservati e simulati (dai GCM) nei pattern di pioggia nello spazio e nel tempo. È ragionevole ipotizzare che queste variazioni nei gradienti di temperatura possano anche influenzare la variabilità stagionale delle precipitazioni locali, causando, ad esempio, spostamenti stagionali che, altrimenti, non potrebbero essere catturati attraverso una demarcazione "a priori" delle stagioni. Qui, sono stati addestrati molti modelli diversi per determinare le migliori variabili di input e i loro parametri (ad esempio, il numero di stati nascosti). È stato costruito un modello per ogni cluster di stazioni meteorologiche. Ogni modello identifica cinque stati nascosti. Sulla base del confronto tra i valori della Log-Likelihood e del criterio di informazione bayesiano per i diversi modelli, sono stati selezionati come ottimali cinque stati nascosti. Infine, l'algoritmo di Viterbi identifica la sequenza più probabile di stati nascosti associati alla sequenza di osservazioni [30]. È stato effettuato un processo di identificazione e caratterizzazione degli stati nascosti. Si sottolinea che la probabilità di occorrenza e le funzioni di densità di probabilità delle precipitazioni giornaliere e i corrispondenti campi compositi delle variabili atmosferiche sono associati a ciascuno stato nascosto. Sono stati identificati gli stati che presentavano le caratteristiche di un evento meteorico estremo. Per ogni cluster, verranno elaborate cento simulazioni di pioggia utilizzando i modelli creati in questo processo. Il codice utilizzato è quello proposto da [31] e scaricabile da: (http://www.sergeykirshner.com/software/mvnhmm). I risultati ottenuti sono stati poi confrontati con l'uso diretto dei predittori atmosferici all'interno dell'NHMM senza applicare lo Stacking.

## Applicazione e risultati per la Regione Lazio

La figura sottostante mostra i tre cluster di pluviometri identificati tramite l'approccio della *modularity*. I cluster di stazioni pluviometriche (Fig. 5) sono situati nelle aree costiere settentrionali e meridionali e nella parte montuosa orientale della regione. Questi gruppi di pluviometri mostrano una chiara omogeneità spaziale (che è coerente con la distribuzione spaziale/temporale attesa delle precipitazioni all'interno dello stesso cluster), nonostante il fatto che la prossimità o altre caratteristiche spaziali non vengano prese in considerazione nella

creazione dei cluster. Inoltre, l'approccio proposto identifica senza ambiguità la suddivisione ottimale per una data rete, poiché il metodo della *modularity* si basa su un processo di ottimizzazione che trova il numero ideale di cluster (o comunità) massimizzando il valore della modularità [23].

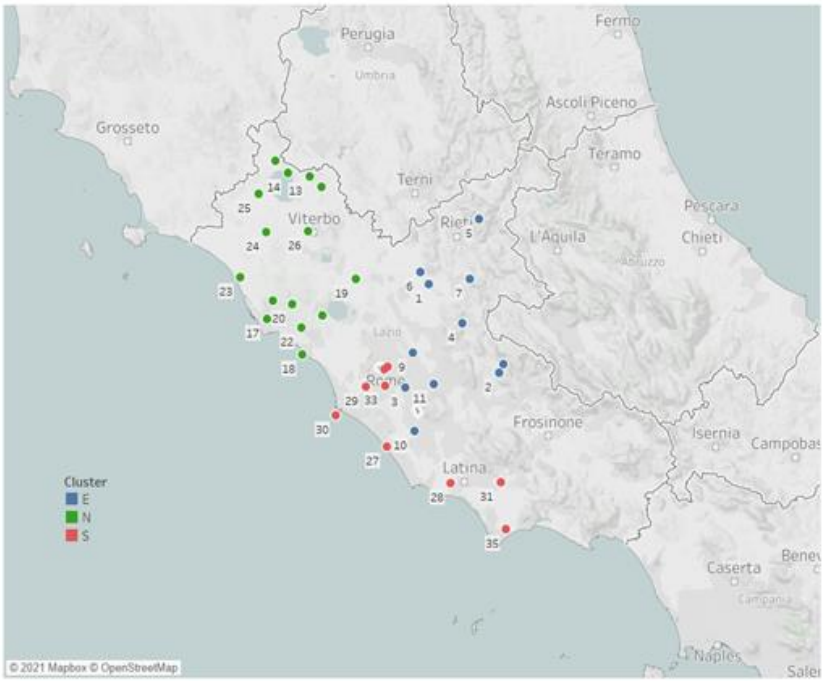


*Figura 5. Rappresentazione spaziale dei 3 cluster di stazioni pluviometriche con caratteristiche di pioggia omogenee: Cluster Est, Cluster Nord e Cluster Sud*

La tecnica PCA ha permesso di ottenere 20 componenti del GPH e 13 dell'IVT. Le 33 PCAs così elaborate, che rappresentano il 90% della varianza totale del campione, costituiscono i predittori utilizzati all'interno del modello Stacking.

## Identificazione degli stati nascosti

I risultati preliminari del modello NHMM mostrano che cinque stati nascosti erano ottimali per tutti e tre i cluster. Come si può vedere dai due grafici sottostanti, l'incremento percentuale delle prestazioni del modello (la pendenza della curva) diminuisce all'aumentare del numero di stati nascosti. Maggiore è il numero di stati nascosti scelti per l'addestramento del modello migliori saranno le performance dello stesso in fase di simulazione dei dati reali di pioggia. In questo caso, è stato scelto di utilizzare cinque stati nascosti a causa del compromesso tra prestazioni e tempo di computazionale di simulazione. Infatti, come mostrato dall'analisi delle metriche di Likelihood (Fig. 6) e Bayesian Information Criteria (BIC) (Fig. 7).

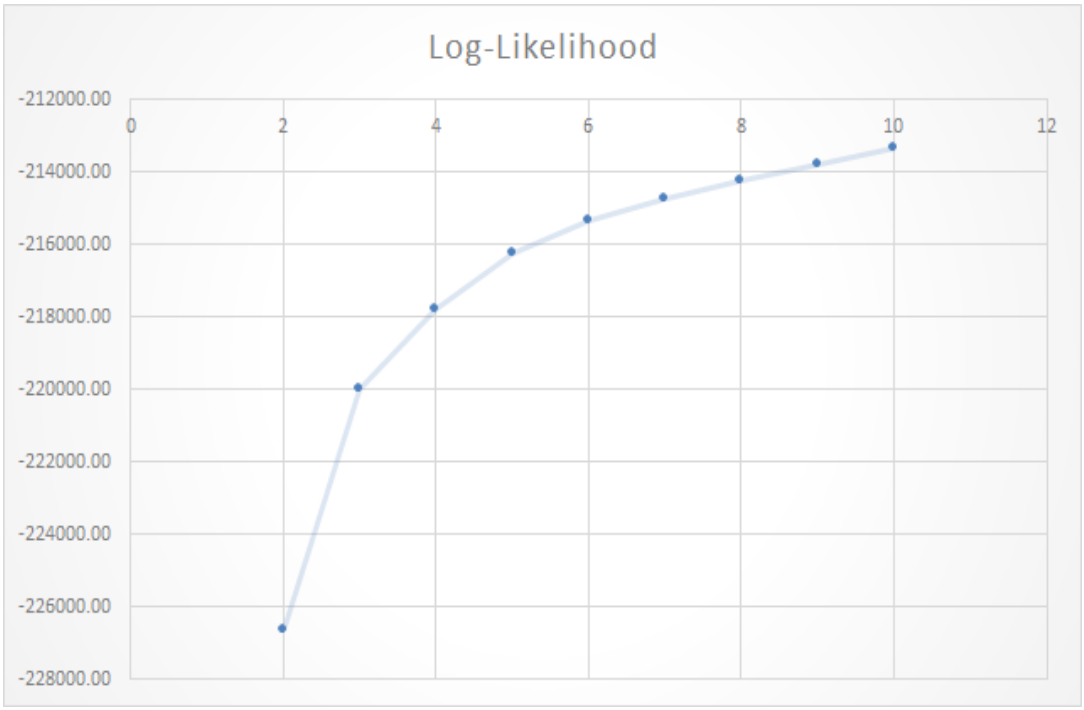


*Figura 6. Analisi delle performance del numero di stati nascosti del modello NHMM con l'ausilio della metrica Likelihood rappresentata su scala logaritmica*

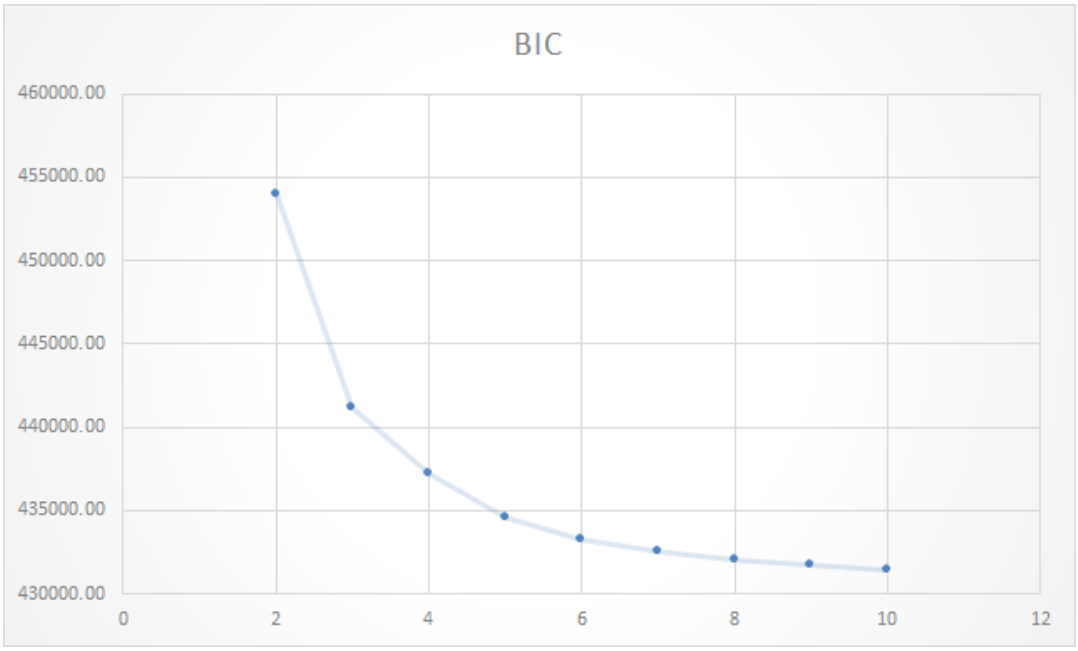


*Figura 7. Analisi delle performance del numero di stati nascosti del modello NHMM con l'ausilio della metrica Bayesian Information Criteria (BIC)*

L'addestramento dell'NHMM con cinque stati nascosti è stato effettuato per ottenere i modelli relativi ai tre gruppi di stazioni. Una volta definiti i parametri del modello, la sequenza di stati di pioggia più probabile per l'intera serie di dati storici è stata ottenuta attraverso l'algoritmo di Viterbi. Le mappe delle anomalie di GPH e IVT sono state poi confrontate con gli stati di pioggia corrispondenti elaborati da Viterbi (Fig. 8-10), al fine di caratterizzare qualitativamente

i cinque stati di pioggia prodotti dall'NHMM e associarli a un determinato stato di precipitazione.

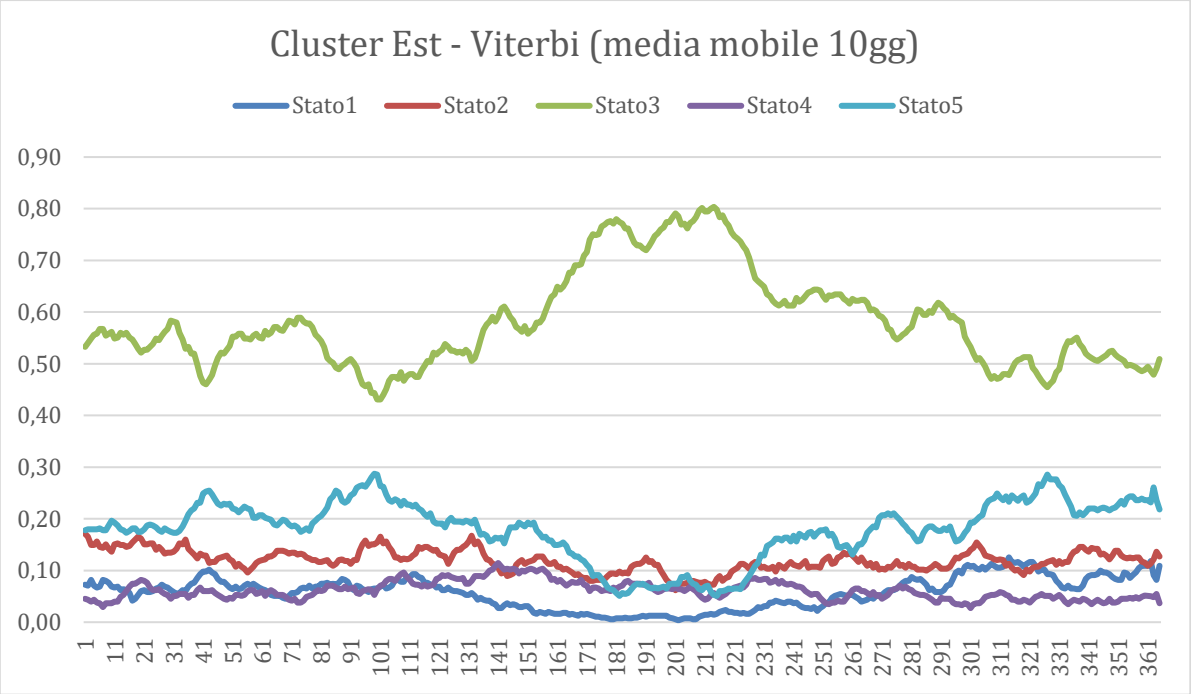


*Figura 8. Rappresentazione della probabilità degli Stati nascosti di pioggia per i giorni dell'anno, mediati sull'intera serie storica – Cluster Est*

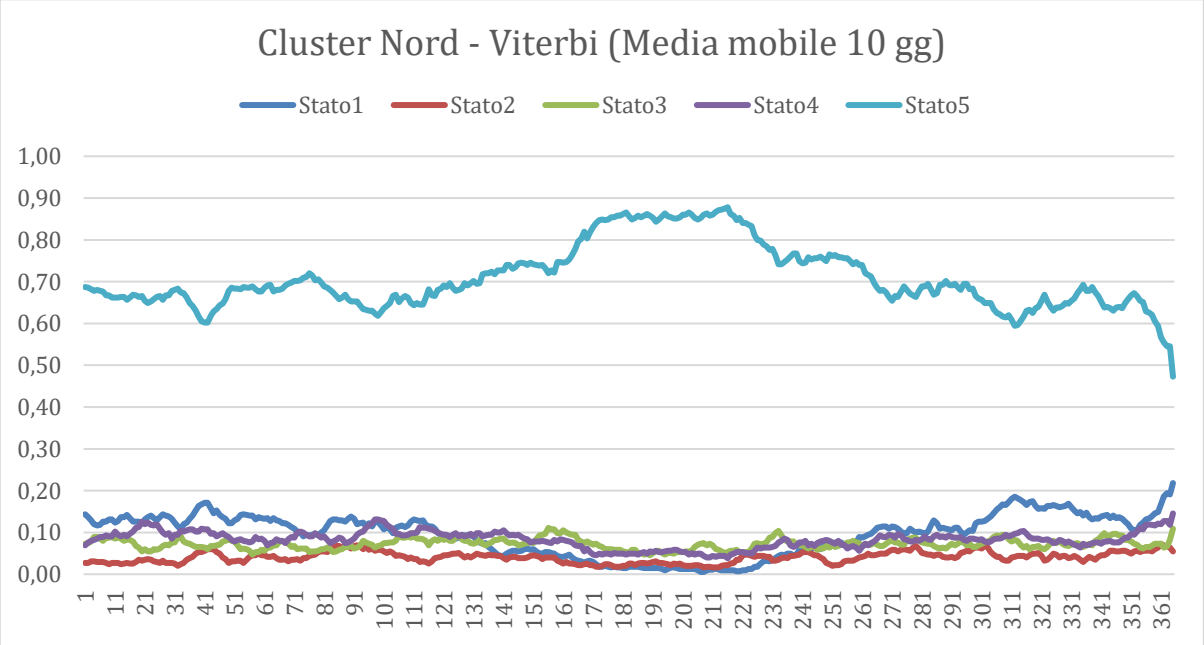


*Figura 9. Rappresentazione della probabilità degli Stati nascosti di pioggia per i giorni dell'anno, mediati sull'intera serie storica – Cluster Nord*

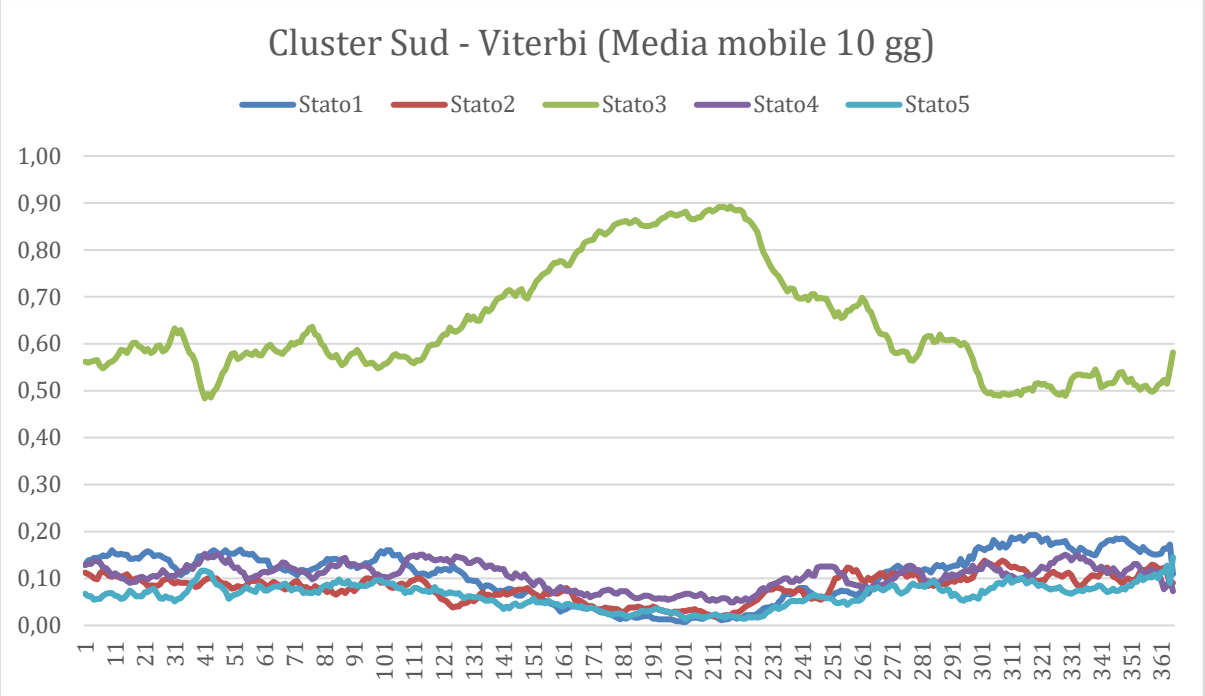


*Figura 10. Rappresentazione della probabilità degli Stati nascosti di pioggia per i giorni dell'anno, mediati sull'intera serie storica – Cluster Sud*

Tutti i cluster sono caratterizzati da uno stato di pioggia dominante rispetto a tutti gli altri durante tutto l'anno. Quest'ultimi (stato n. 3 per i Cluster Est e Sud, stato n. 5 per il Cluster Nord) hanno una maggiore frequenza di accadimento nella stagione estiva rispetto al resto dell'anno, motivo per cui può essere identificato come lo stato di tempo asciutto del cluster est. Il trend opposto si è verificato per gli stati, che si prefigurano come stati tipicamente piovosi poiché hanno una maggiore frequenza di occorrenza in corrispondenza del periodo invernale e autunnale della serie storica. Questi risultati sono confermati anche analizzando la probabilità di pioggia e gli accumuli medi giornalieri calcolati rispetto a ogni stato piovoso.

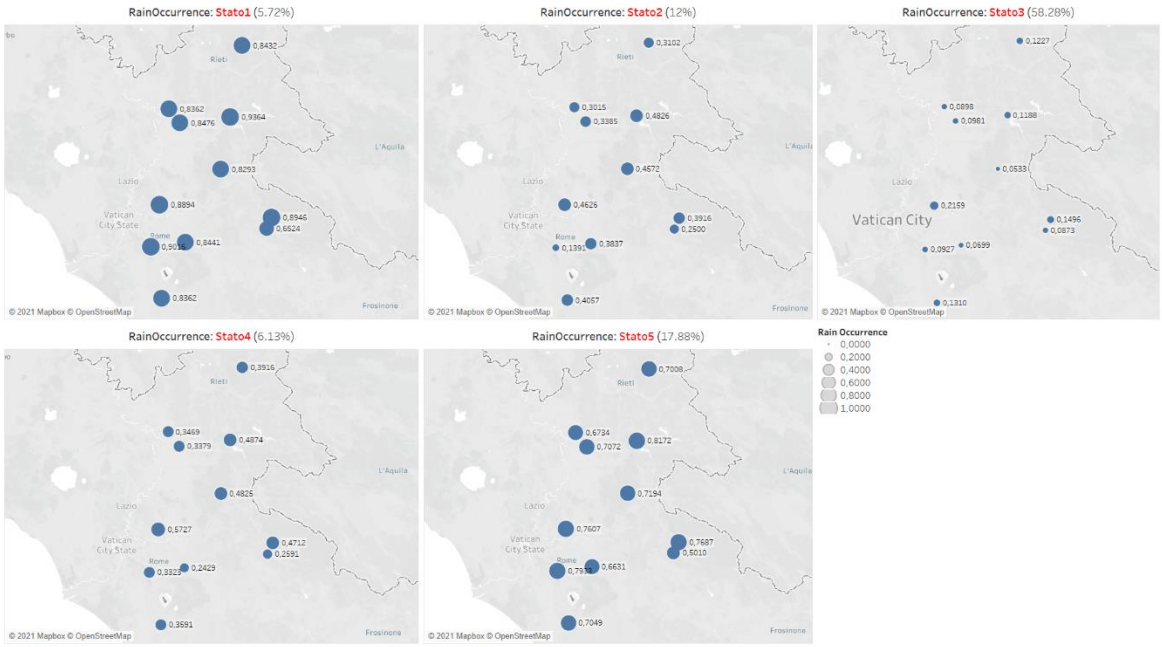


*Figure 11. Probabilità di precipitazione - Cluster Est*

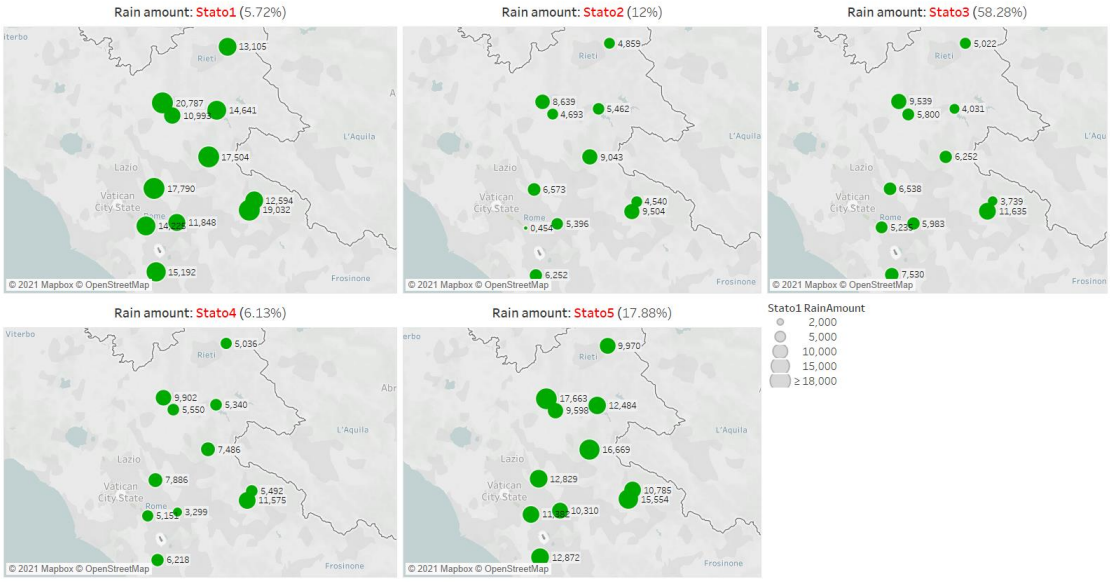


*Figure 12. Altezze di pioggia - Cluster Est*

Le figure precedenti (Fig. 11,12), relative al cluster Est della Regione Lazio, illustrano come l'evento di pioggia estrema possa essere associato allo stato piovoso numero 1. Questo presenta le classiche caratteristiche di un evento di pioggia estrema: forte precipitazione, alta probabilità di pioggia quando si verifica, e bassa probabilità di accadimento all'interno della serie storica (5,72%).

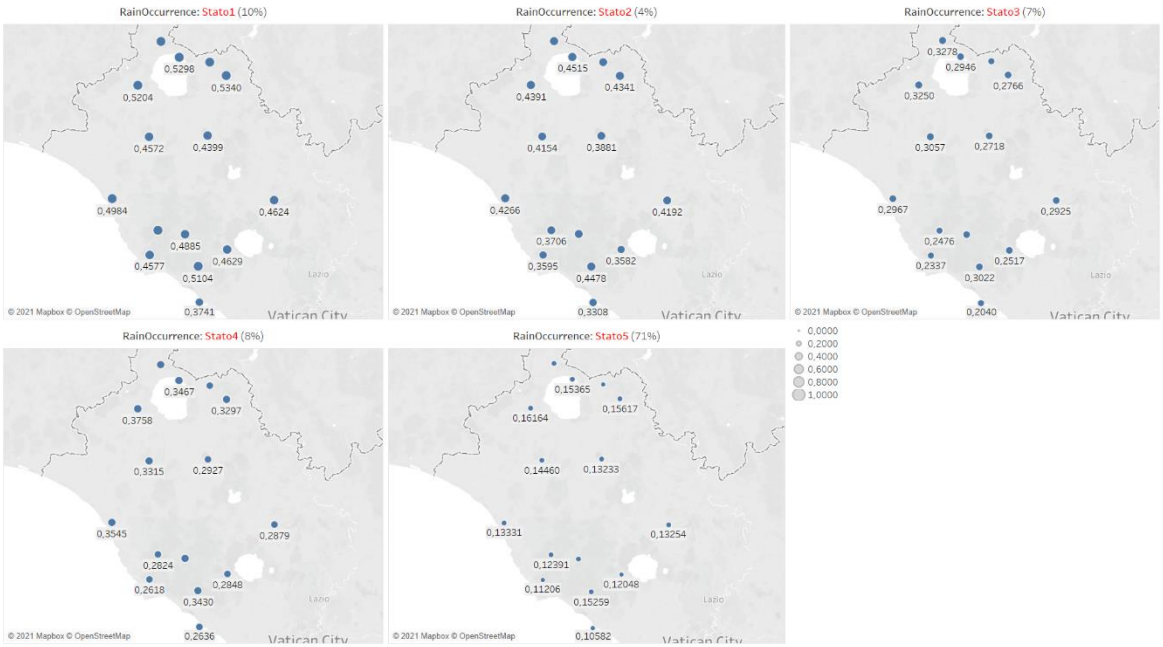


*Figure 13. Probabilità di precipitazione - Cluster Nord*

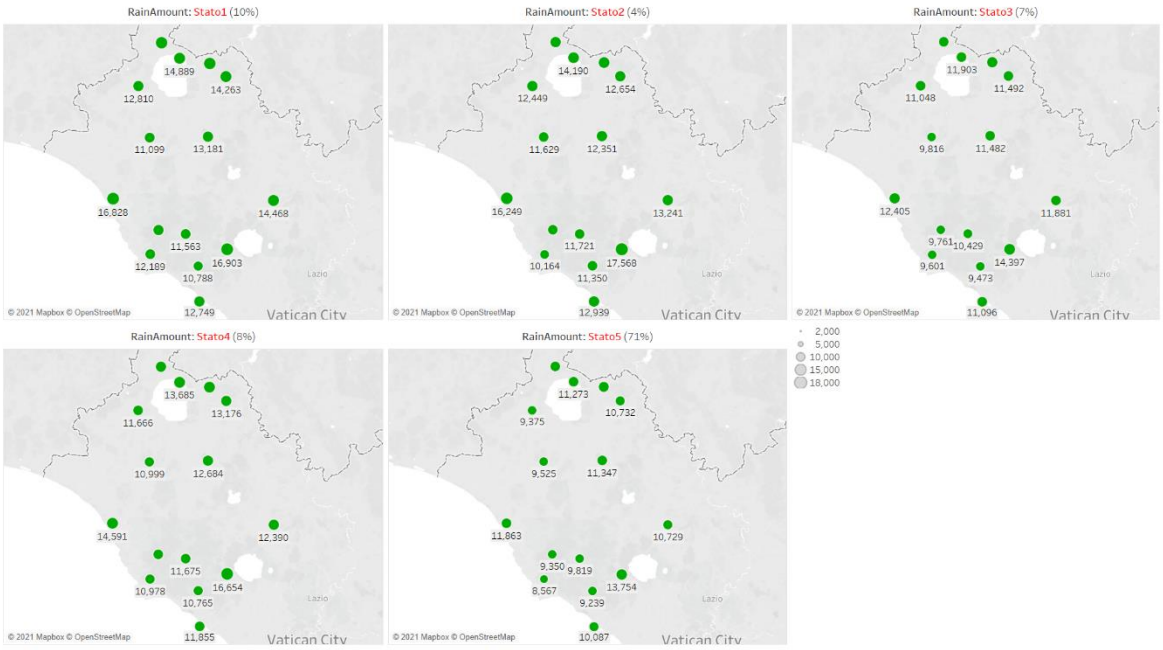


*Figure 14. Altezze di pioggia - Cluster Nord*

Le figure precedenti (Fig. 13,14), relative al cluster Nord della Regione Lazio, illustrano come l'evento di pioggia estrema possa essere associato agli stati piovoso numero 1 e 2, anche se non in maniera così netta come per il cluster Est. Questi ultimi presentano le classiche caratteristiche di un evento di pioggia estrema: forte precipitazione, alta probabilità di pioggia quando si

verifica, e bassa probabilità di accadimento all'interno della serie storica (4%).

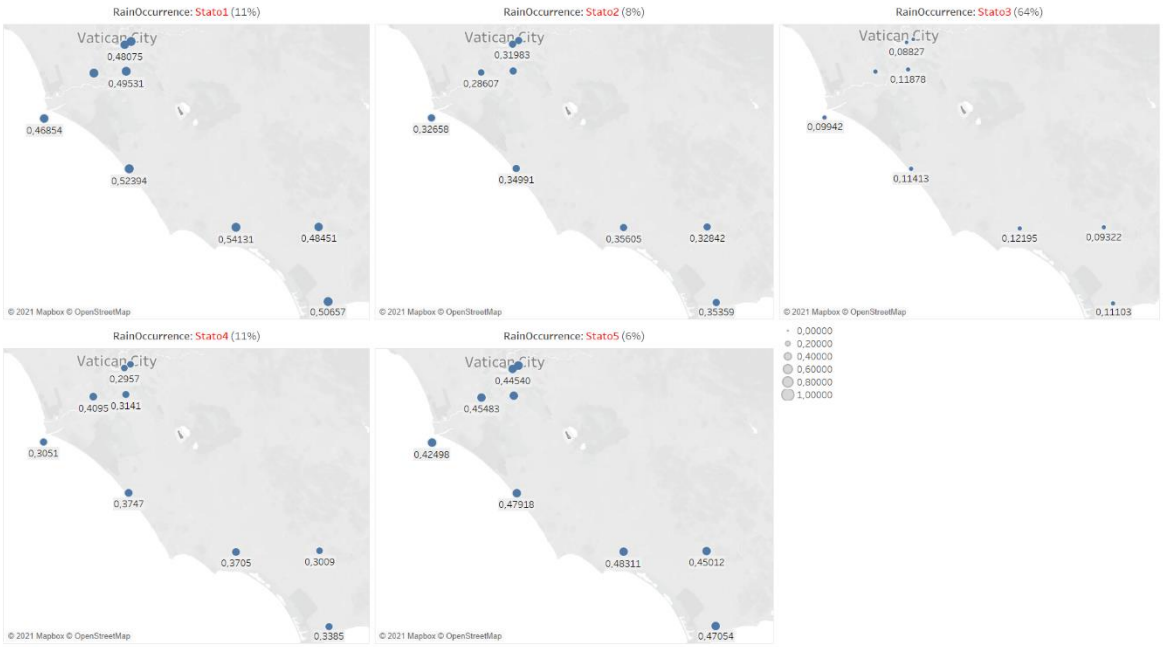


*Figure15. Probabilità di precipitazione - Cluster Sud*

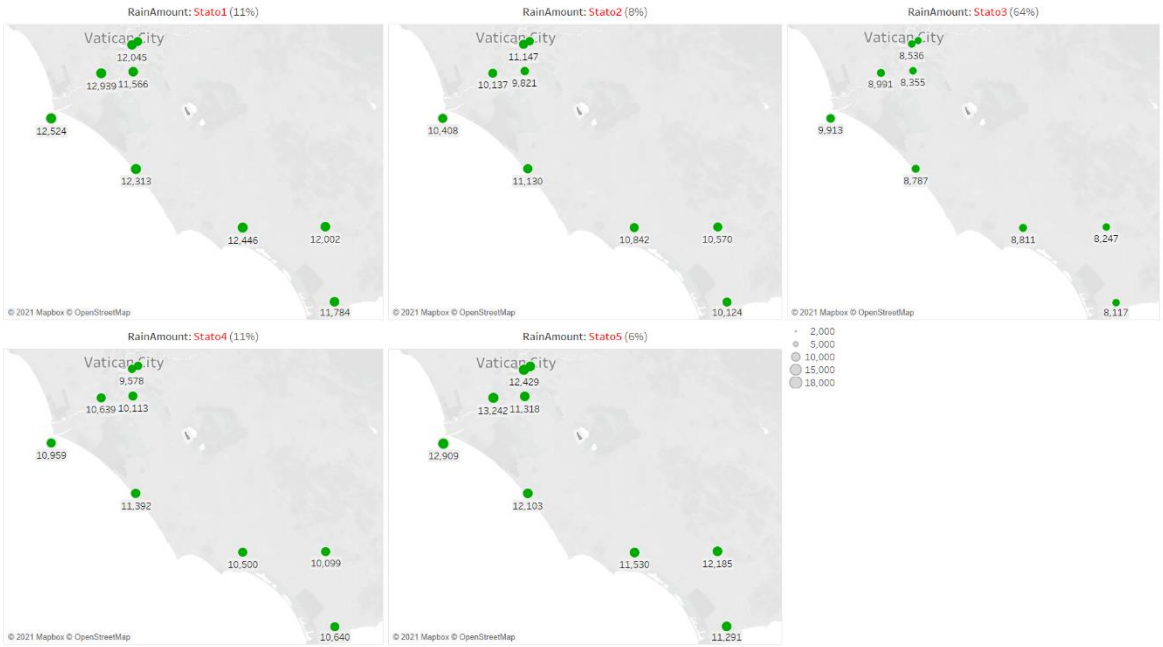


*Figure 16. Altezze di pioggia - Cluster Sud*

Infine, relativamente al cluster Sud della Regione Lazio, le immagini sopra riportate (Fig. 15,16) illustrano come l'evento di pioggia estrema possa essere associato allo stato piovoso numero 5, tuttavia non in maniera così netta come per il cluster Est. Questo presenta le classiche caratteristiche di un evento di pioggia estrema: forte precipitazione, alta probabilità di pioggia quando si verifica, e bassa probabilità di accadimento all'interno della serie storica (6%).

La validità della procedura di clusterizzazione delle stazioni pluviometriche precedentemente discussa viene anche confermata dall'analisi grafica della frequenza e dell'intensità delle precipitazioni. Infatti, da quest'ultima analisi grafica, emerge che ciascun cluster conserva caratteristiche meteorologiche omogenee.

## Simulazioni

Attraverso le simulazioni, è stato possibile mettere a confronto i risultati prodotti dai modelli di previsione - sia con l'utilizzo dello stacking che senza - rispetto ai dati effettivi di precipitazioni rilevati dalle stazioni pluviometriche.

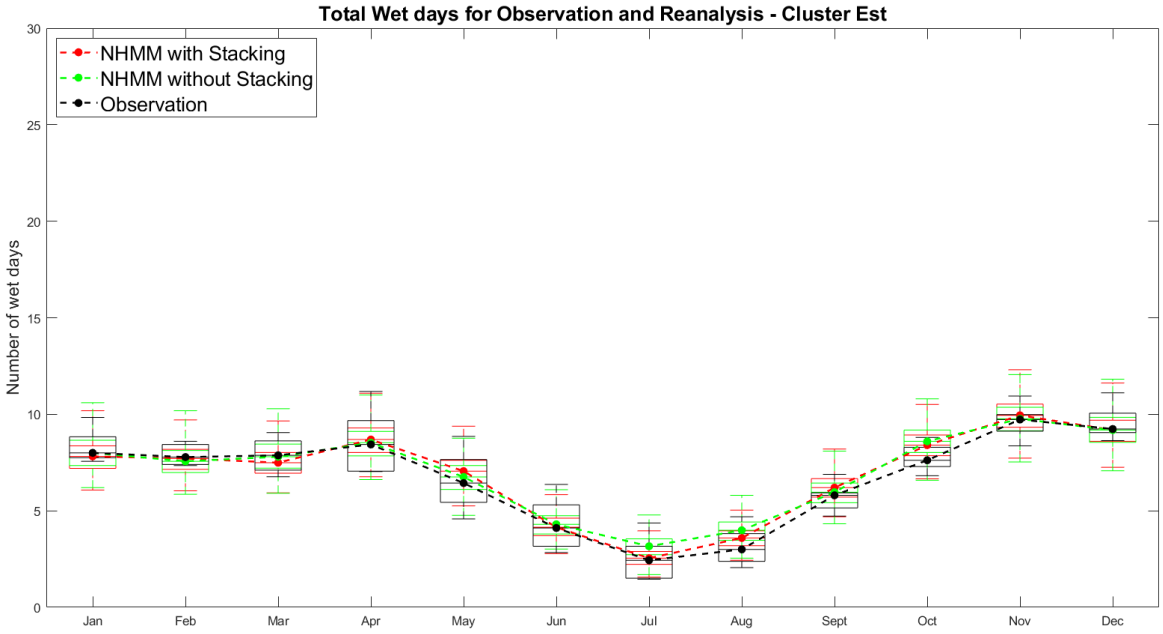


*Figura 17. Confronto tra numero di giorni di pioggia osservati (reali), simulati con il modello Stacking, simulati senza l'utilizzo dello Stakcing - Cluster Est*

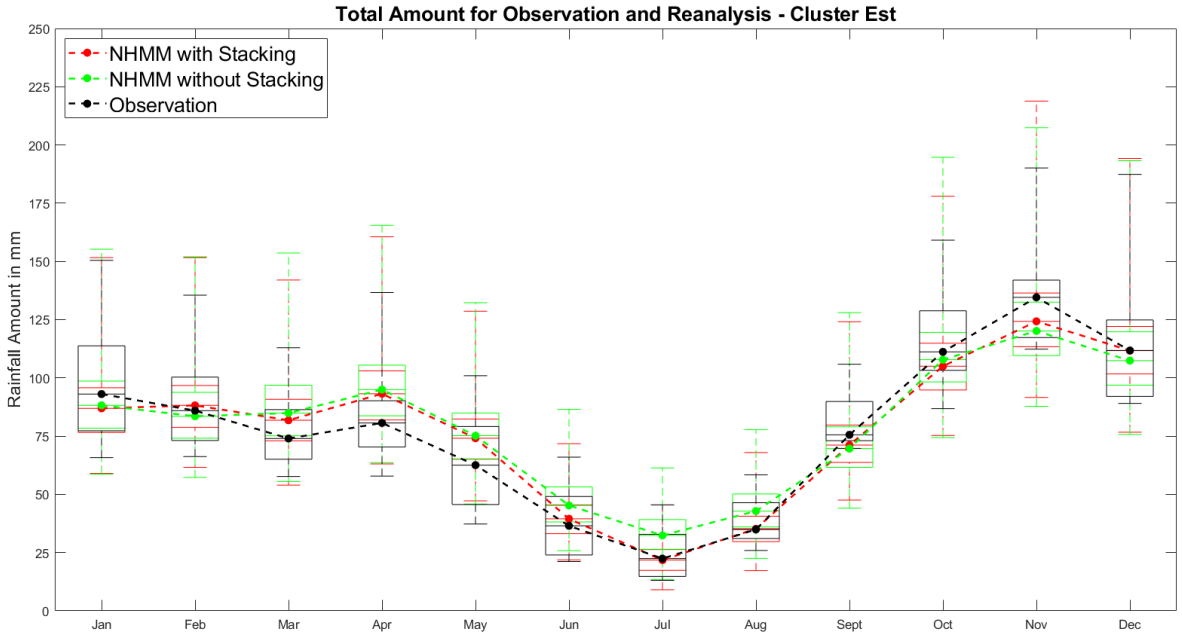


*Figura 18. Confronto tra altezze di pioggia osservate (reali), simulate con il modello Stacking, simulate senza l'utilizzo dello Stakcing - Cluster Est*

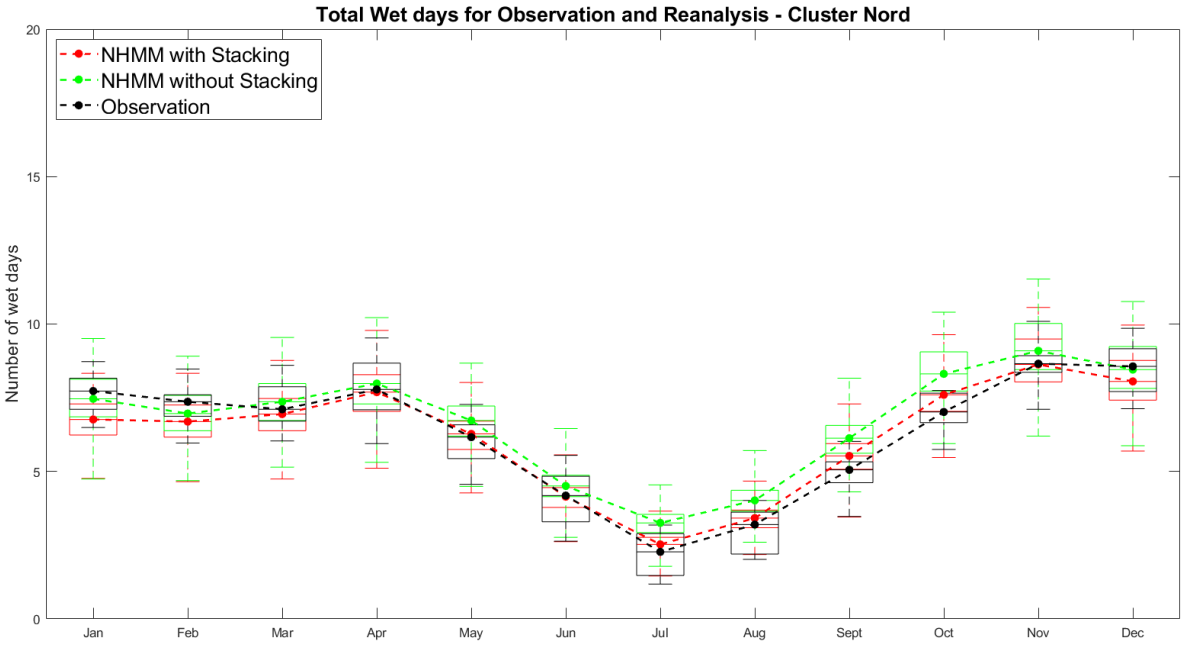


*Figura 19. Confronto tra numero di giorni di pioggia osservati (reali), simulati con il modello Stacking, simulati senza l'utilizzo dello Stakcing - Cluster Nord*

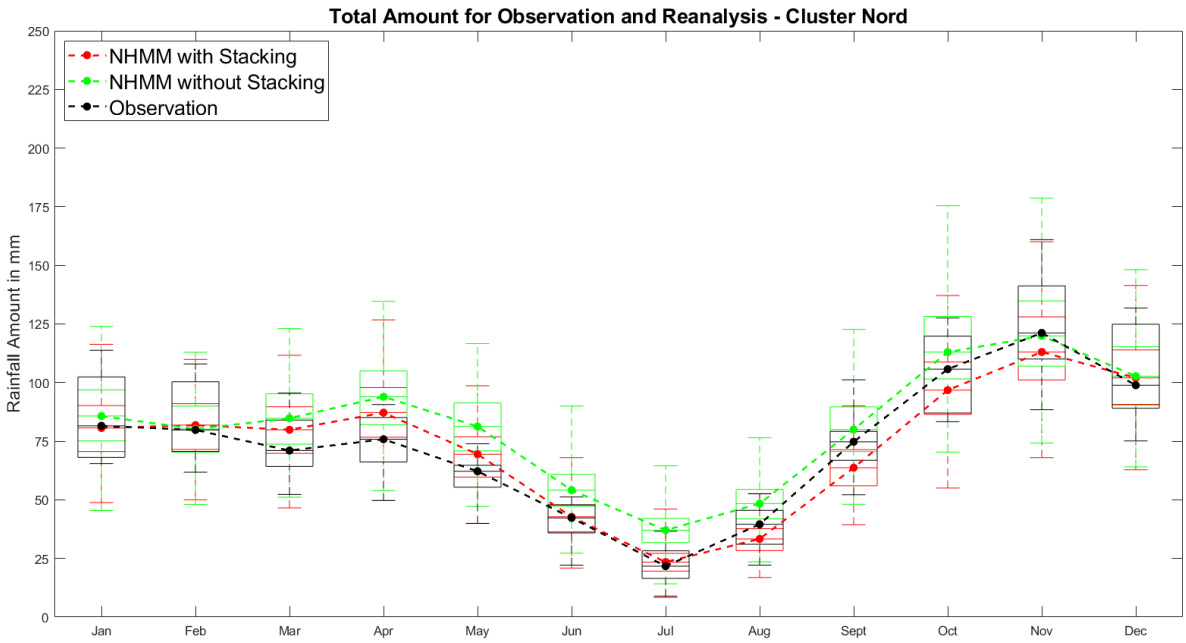


*Figura 20. Confronto tra altezze di pioggia osservate (reali), simulate con il modello Stacking, simulate senza l'utilizzo dello Stakcing - Cluster Nord*

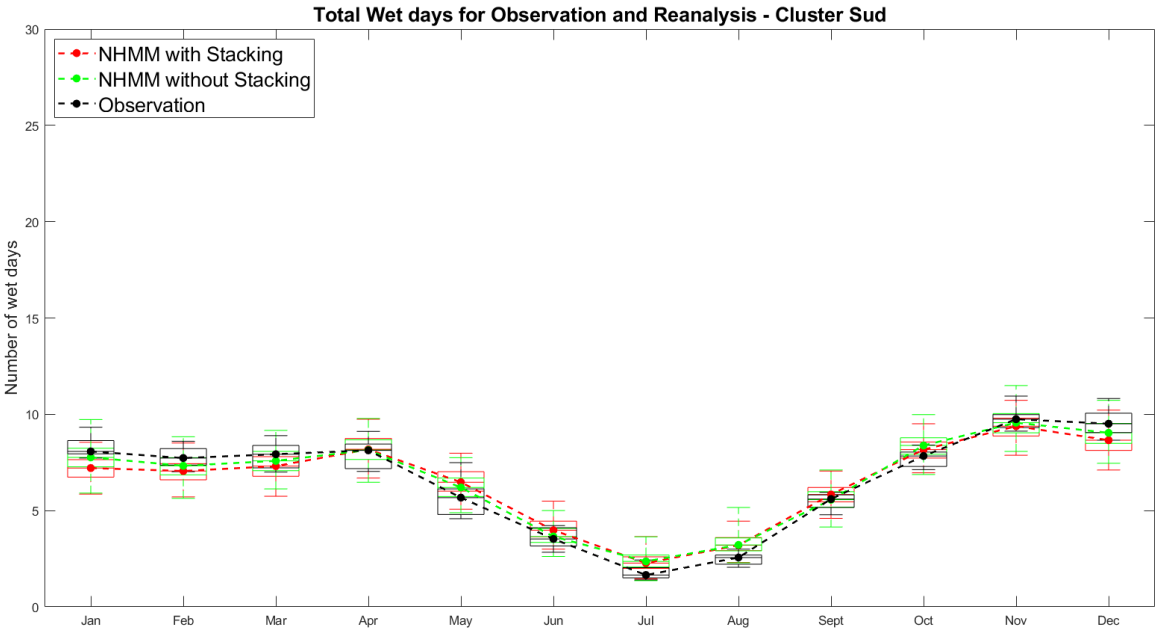


*Figura 21. Confronto tra numero di giorni di pioggia osservati (reali), simulati con il modello Stacking, simulati senza l'utilizzo dello Stakcing - Cluster Sud*

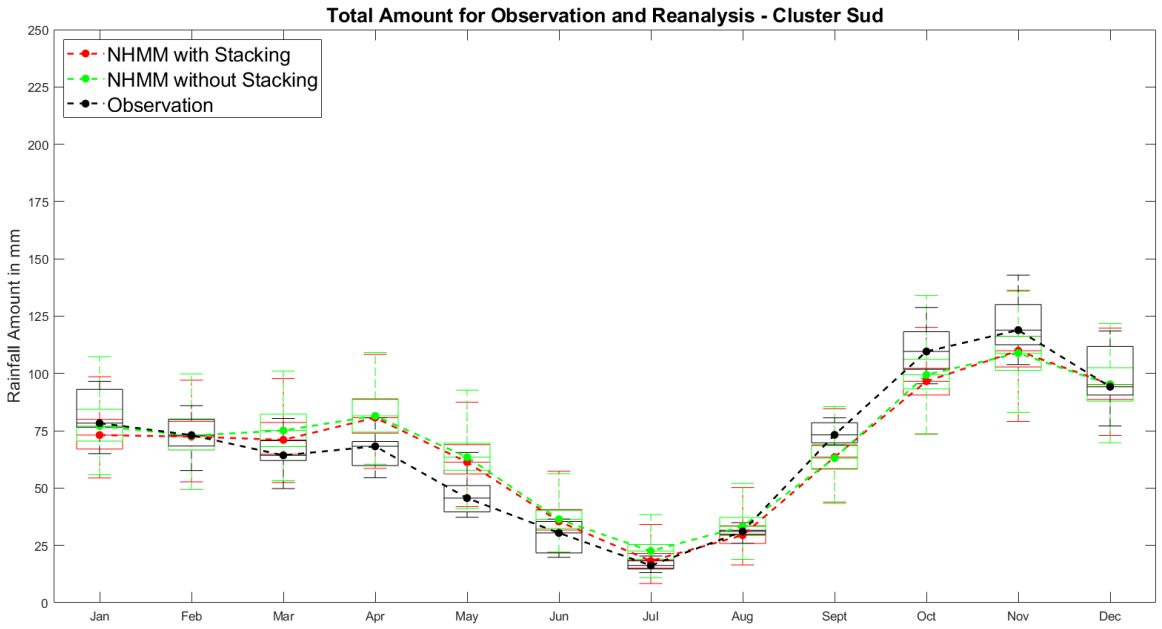


*Figura 22. Confronto tra altezze di pioggia osservate (reali), simulate con il modello Stacking, simulate senza l'utilizzo dello Stakcing - Cluster Sud*

Per tutti i cluster, il modello con stacking migliora le prestazioni di previsione rispetto al modello senza stacking. Come si può osservare dai grafici (Fig. 17-22), sia nelle simulazioni del numero medio di giorni piovosi che nel caso delle altezze di pioggia, calcolate su scala mensile, i valori con stacking (in rosso) risultano più vicini ai valori osservati (in nero). Questa differenza è particolarmente evidente nella stagione estiva, dove il modello con stacking ha dimostrato una maggiore capacità nel prevedere l'andamento stagionale delle precipitazioni. Tali risultati mostrano come l'introduzione del modello di Stacking possa essere particolarmente utile per una accurata simulazione dei periodi di siccità idrologica. I risultati presentati nei grafici sopra riportati sono stati ulteriormente dettagliati attraverso il calcolo dell'Errore Medio Assoluto (MAE) tra valori osservati e valori simulati, sintetizzando in un'unica tabella i risultati ottenuti dai tre cluster di stazioni omogenee.

*Tabella 1. Analisi delle performance di simulazione rispetto ai dati reali utilizzando la metrica MAE (Mean Absolute Error) tra modello con Stacking e senza Stacking - Cluster Est, Nord e Sud*

**Tot Wet Days**

| | *Cluster Est* | | *Cluster Nord* | | *Cluster Sud* | |
|---|---|---|---|---|---|---|
| | Stacking | No-Stacking | Stacking | No-Stacking | Stacking | No-Stacking |
| **MAE** | 0.31 | 0.32 | 5.85 | 9.10 | 0.54 | 0.36 |

**Rainfall Amount**

| | *Cluster Est* | | *Cluster Nord* | | *Cluster Sud* | |
|---|---|---|---|---|---|---|
| | Stacking | No-Stacking | Stacking | No-Stacking | Stacking | No-Stacking |
| **MAE** | 5.43 | 8.31 | 5.85 | 9.10 | 6.85 | 7.44 |

Analizzando i dati presentati (Tab. 1), si osserva che nel Cluster Est e Sud il numero di giorni piovosi delle stazioni pluviometriche presenta un MAE inferiore all'unità, mostrando un errore inferiore ad uno, evento di pioggia, rispetto al dato reale. In questo caso, tuttavia, la nuova implementazione del modello con Stacking non presenta vantaggi significativi rispetto al modello classico (senza lo Stacking). Per il Cluster Nord, invece, il modello con Stacking si distingue nettamente, presentando un MAE molto più basso (5.85) rispetto al modello senza Stacking (9.10), e ottenendo pertanto un significativo aumento delle performance di simulazione. Questo indica che il modello con Stacking ha una maggiore precisione nella previsione del numero totale di giorni di pioggia in questa area geografica. Analizzando i risultati inerenti i valori di altezza di pioggia, si nota che nel Cluster Est e Nord, il modello con Stacking ha un MAE più basso rispetto al modello senza Stacking, segnalando ancora una volta una maggiore precisione nella previsione della quantità di pioggia. Al contrario, nel Cluster Sud, il MAE è leggermente più alto per il modello con Stacking rispetto a quello senza Stacking. In generale, da questa analisi emerge che l'utilizzo del Stacking tende a migliorare la precisione dei modelli di previsione, con un impatto particolarmente positivo osservato nel Cluster Nord. Nonostante ciò, è importante sottolineare che i risultati possono variare significativamente a seconda della metrica considerata e dell'area geografica analizzata.

## Simulazione degli estremi di precipitazione

Nell'ambito dell'analisi delle performance dei due modelli di previsione delle precipitazioni, è stata valutata anche la loro capacità di simulare gli estremi di precipitazione in diverse aree geografiche: Cluster Est, Cluster Nord e Cluster Sud. Questa analisi è essenziale per comprendere quanto accuratamente i modelli possano prevedere eventi di precipitazione estrema, che rivestono un'importanza cruciale per la pianificazione e la gestione delle risorse

idriche e per la mitigazione dei rischi associati agli eventi meteorologici estremi. I grafici (Fig. 23-28) e le tabelle offrono un quadro dettagliato delle performance dei due modelli. Questo viene fatto sia graficamente, confrontando, attraverso l'uso di bande di errore, i valori simulati con i valori osservati, sia in termini numerici, esprimendo le performance attraverso l'Errore Assoluto Medio (MAE) e il coefficiente di determinazione ($R^2$).

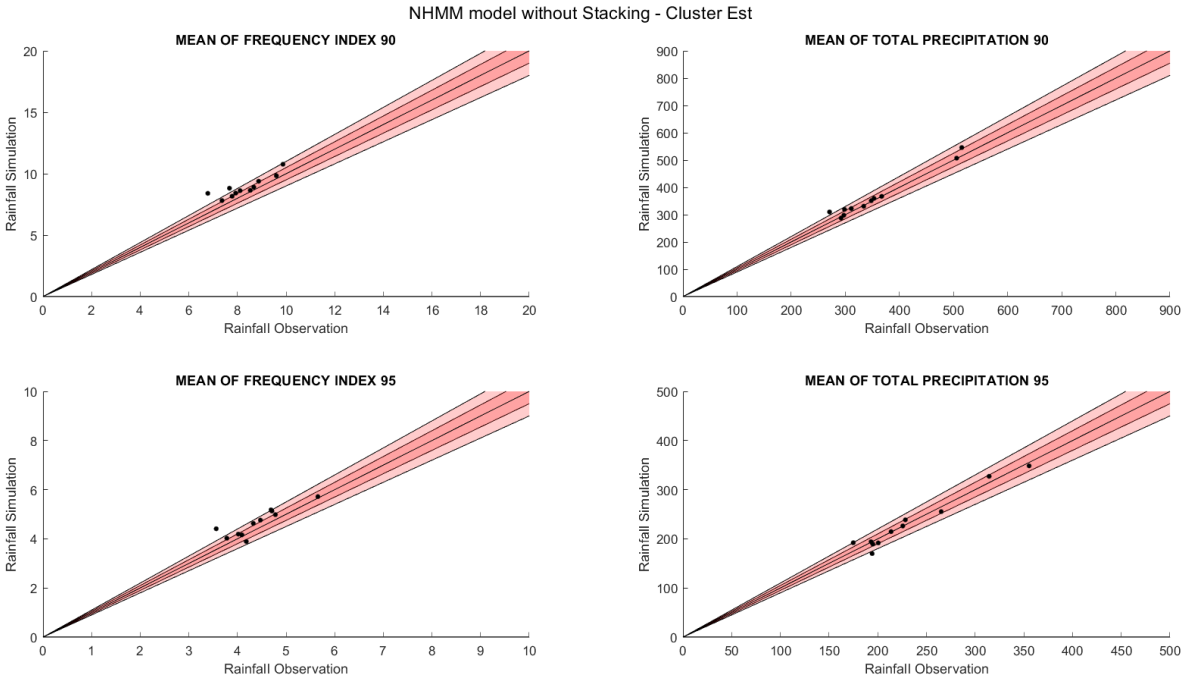


*Figura 23. Analisi degli estremi di precipitazione, modello senza lo Stacking - Cluster Est*

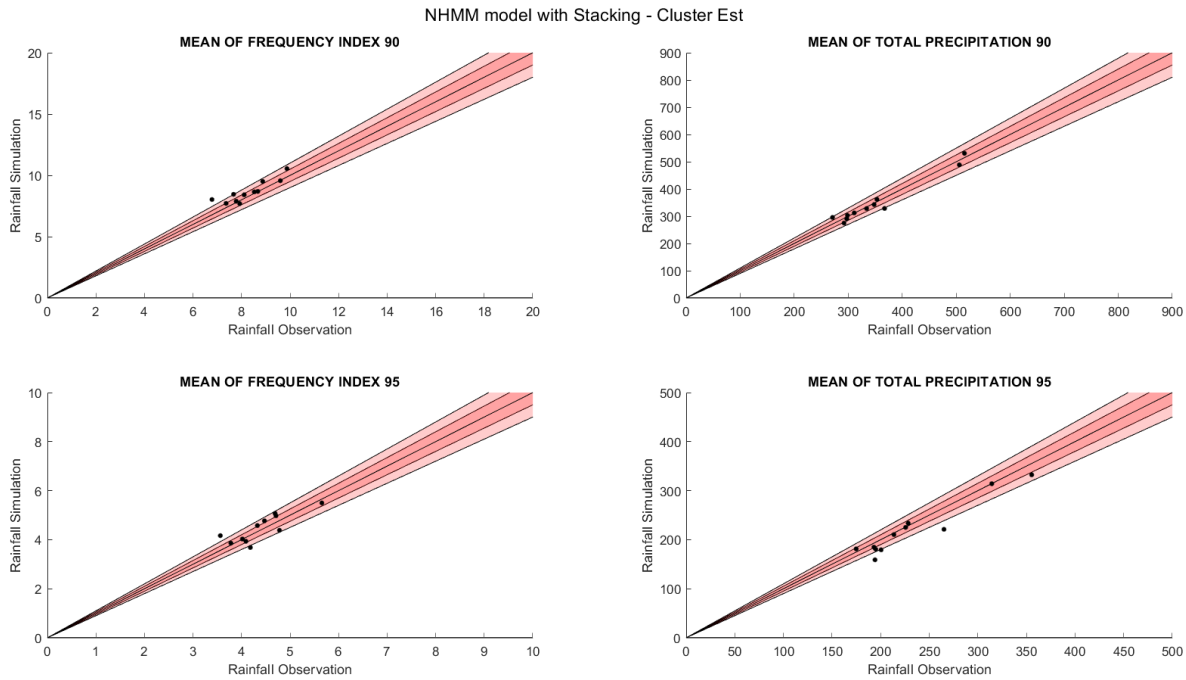


*Figura 24. Analisi degli estremi di precipitazione, modello con lo Stacking - Cluster Est*

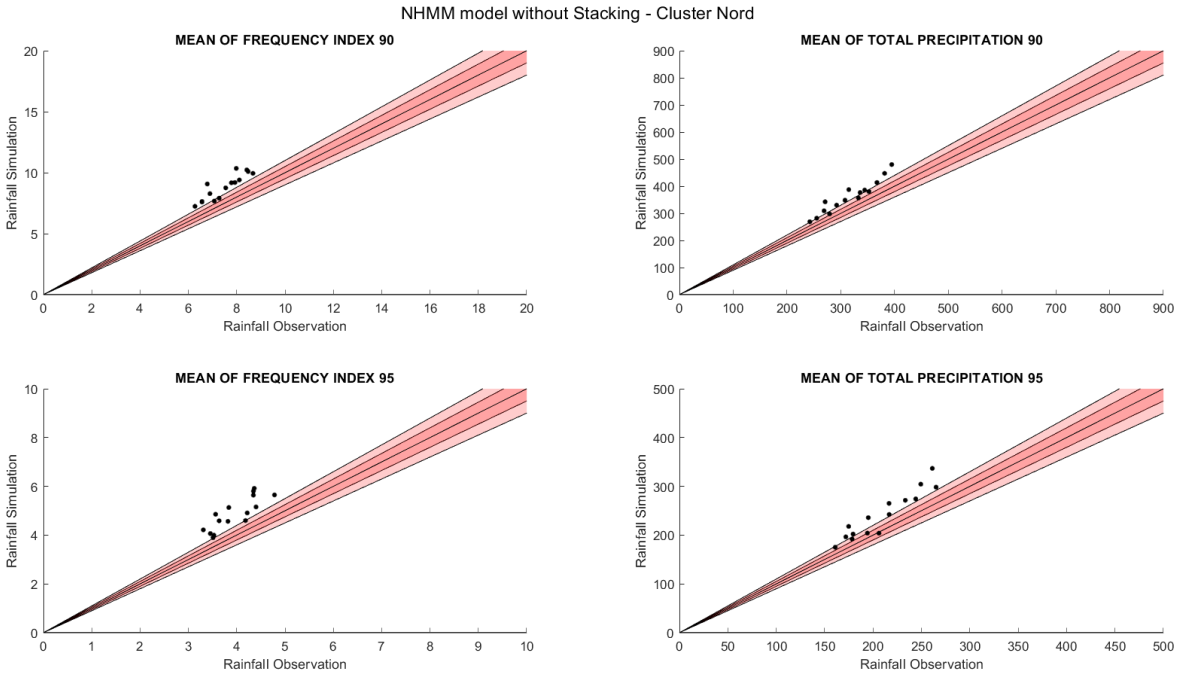


*Figura 25. Analisi degli estremi di precipitazione, modello senza lo Stacking - Cluster Nord*

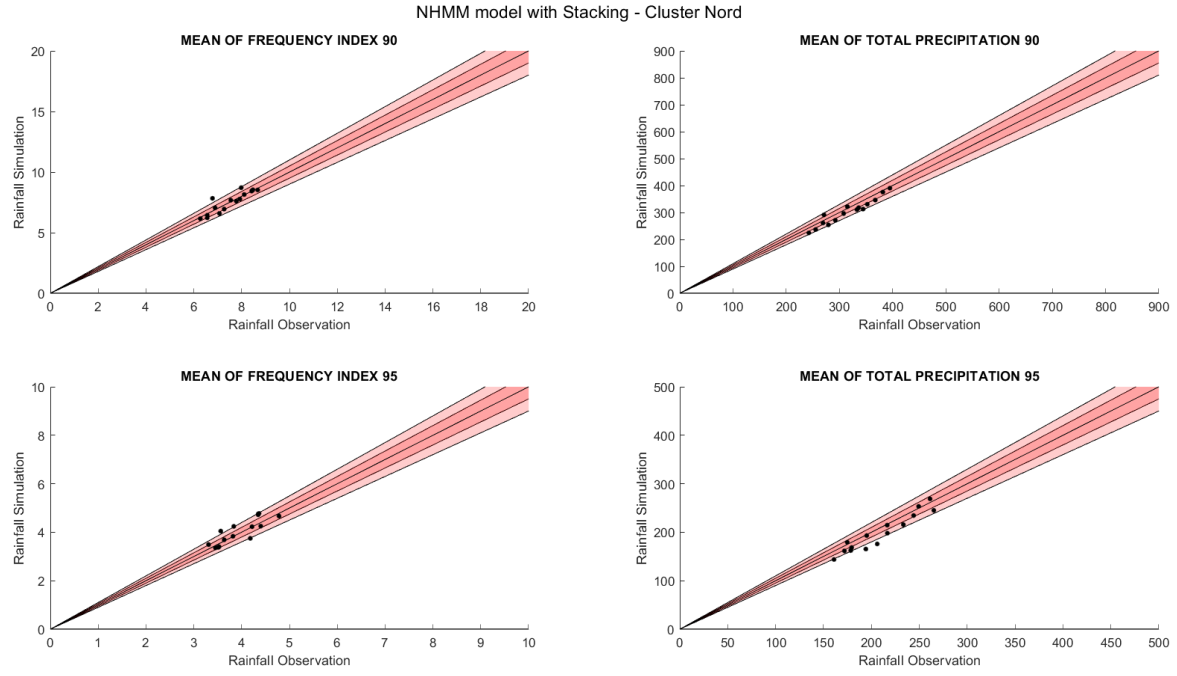


*Figura 26. Analisi degli estremi di precipitazione, modello con lo Stacking - Cluster Nord*

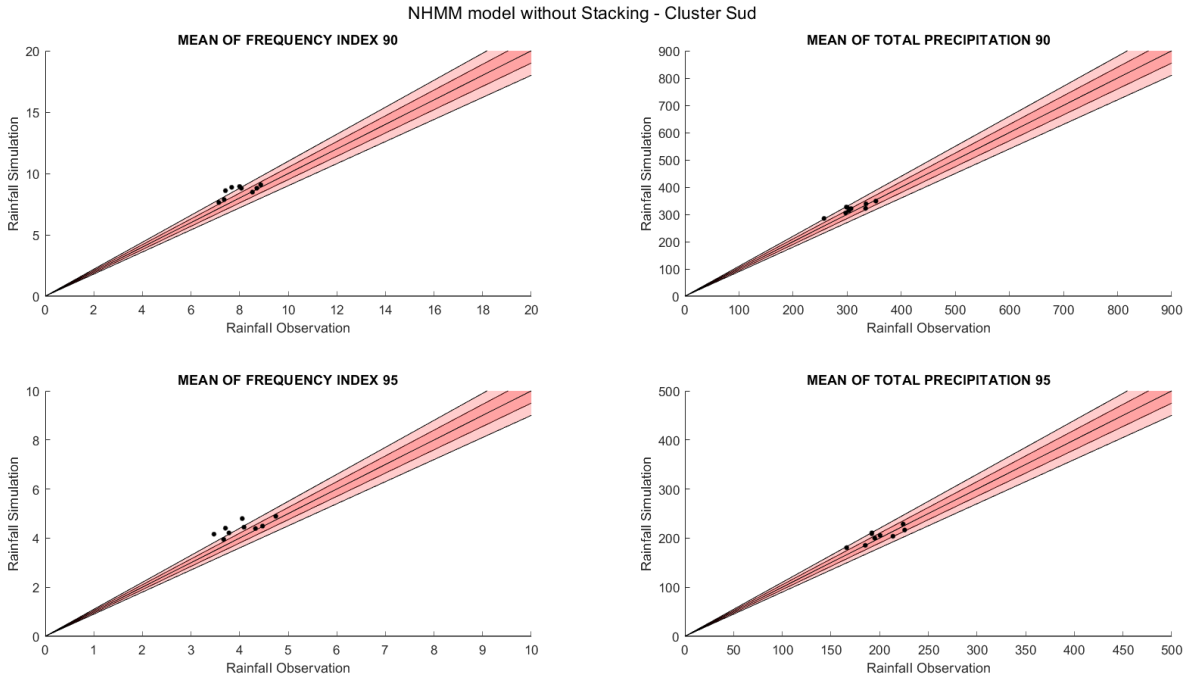


*Figura 27. Analisi degli estremi di precipitazione, modello senza lo Stacking - Cluster Sud*

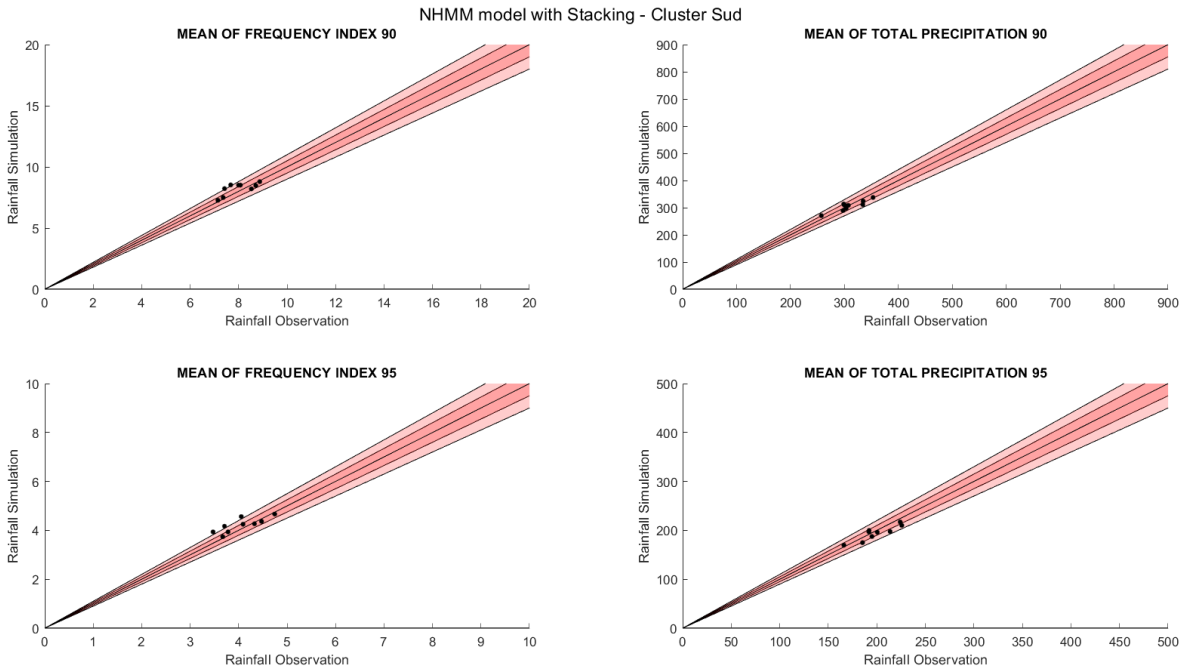


*Figura 28. Analisi degli estremi di precipitazione, modello con lo Stacking - Cluster Sud*

Per identificare e analizzare gli eventi di precipitazione estrema è stato utilizzato il metodo proposto in [32]. Quest'ultimo permette di definire due indici per identificare un evento estremo di precipitazione: frequenza e precipitazione totale. Il primo rappresenta il numero di eventi all'anno la cui quantità di pioggia giornaliera supera la soglia; il secondo è la somma delle quantità di pioggia giornaliere di tali eventi. Sono state tracciate due bande di confidenza della bisettrice,

una con un errore del 5% e l'altra con un errore del 10%. Per ogni pluviometro, viene definita una soglia in termini di percentile fisso della serie di quantità di pioggia giornaliera, considerando solo i giorni con pioggia non nulla; successivamente, viene identificato il numero di eventi di precipitazione non nulli durante ogni anno e un percentile prefissato (90° e 95°) di questa serie stimato per ogni anno. La mediana di questi percentili attraverso tutti gli anni viene scelta come soglia. Come possiamo vedere dalle immagini sopra riportate (Fig. 23-28), i valori del modello Stacking sono tutti all'interno delle bande di errore del 5 e 10%, a differenza dei valori del modello senza stacking.

**Mean frequency Index 90**

| | *Cluster Est* | | *Cluster Nord* | | *Cluster Sud* | |
|---|---|---|---|---|---|---|
| | No-Stacking | Stacking | No-Stacking | Stacking | No-Stacking | Stacking |
| **MAE** | 0,41 | 0,60 | 0,31 | 1,42 | 0,32 | 0,49 |
| **R2** | 0,89 | 0,88 | 0,89 | 0,89 | 0,73 | 0,69 |

**Mean frequency Index 95**

| | *Cluster Est* | | *Cluster Nord* | | *Cluster Sud* | |
|---|---|---|---|---|---|---|
| | No-Stacking | Stacking | No-Stacking | Stacking | No-Stacking | Stacking |
| **MAE** | 0,28 | 0,31 | 0,26 | 0,93 | 0,19 | 0,31 |
| **R2** | 0,83 | 0,88 | 0,85 | 0,84 | 0,81 | 0,74 |

**Mean Total Precipitation 90**

| | *Cluster Est* | | *Cluster Nord* | | *Cluster Sud* | |
|---|---|---|---|---|---|---|
| | No-Stacking | Stacking | No-Stacking | Stacking | No-Stacking | Stacking |
| **MAE** | 13,22 | 10,98 | 17,80 | 46,88 | 9,21 | 12,23 |
| **R2** | 0,98 | 0,99 | 0,96 | 0,95 | 0,91 | 0,89 |

**Mean Total Precipitation 95**

| | *Cluster Est* | | *Cluster Nord* | | *Cluster Sud* | |
|---|---|---|---|---|---|---|
| | No-Stacking | Stacking | No-Stacking | Stacking | No-Stacking | Stacking |
| **MAE** | 14,77 | 8,91 | 14,42 | 32,48 | 7,04 | 7,64 |
| **R2** | 0,96 | 0,98 | 0,95 | 0,94 | 0,90 | 0,84 |

La tabella sopra riportata (Tab. 2) presenta i risultati ottenuti dall'addestramento di due modelli, uno con Stacking e uno senza Stacking, analizzando le precipitazioni in tre diverse aree geografiche: Cluster Est, Cluster Nord e Cluster Sud. I dati sono espressi attraverso due metriche fondamentali: l'Errore Assoluto Medio (MAE) e il coefficiente di determinazione (R^2), e sono riferiti a quattro diversi indici: Mean frequency Index 90 e 95, e Mean Total Precipitation 90 e 95.

Nel caso del Mean Frequency Index 90, nel Cluster Est, il modello senza Stacking mostra un MAE di 0.41 e un R^2 di 0.89, mentre il modello con Stacking ha un MAE di 0.60 e un R^2 di 0.88. Questo suggerisce che, nonostante un errore leggermente maggiore, il modello con Stacking ha una capacità simile di spiegare la variabilità dei dati rispetto al modello senza Stacking in questa regione.
Per il Cluster Nord, il modello senza Stacking ha un MAE di 0.31 e un R^2 di 0.89, mentre il modello con Stacking mostra un aumento significativo del MAE a 1.42, mantenendo però un R^2 simile di 0.89. Nel Cluster Sud, il modello senza Stacking ha un MAE di 0.32 e un R^2 di 0.73, mentre il modello con Stacking ha un MAE di 0.49 e un R^2 di 0.69, indicando una leggera diminuzione della capacità di spiegare la variabilità dei dati.
Analogamente, per il Mean Frequency Index 95 e gli altri indici, si osservano variazioni nei valori di MAE e R^2 tra i modelli con e senza Stacking nelle diverse regioni. Ad esempio, per il Mean Total Precipitation 90, nel Cluster Est, il modello senza Stacking ha un MAE di 13.22 e un R^2 di 0.98, mentre il modello con Stacking ha un MAE ridotto a 10.98 e un R^2 leggermente superiore di 0.99.
In sintesi, la tabella 2 offre una visione dettagliata delle performance dei due modelli nelle diverse regioni e per diversi indici, evidenziando le differenze in termini di errore assoluto medio e capacità di spiegare la variabilità dei dati di precipitazione.
Il risultato del confronto delle piogge estreme mostra una notevole differenza nella previsione delle precipitazioni estreme tra il modello con stacking e il modello senza stacking. Questa differenza è particolarmente evidente nelle precipitazioni estreme oltre il novantesimo percentile, dove il modello senza stacking mostra un errore più che doppio rispetto al modello con stacking.

## Standard Precipitation Index (SPI)

Lo Standard Precipitation Index (SPI) è l'indicatore più comunemente utilizzato per rilevare e descrivere le siccità meteorologiche. L'indicatore SPI mostra le anomalie di pioggia totale osservata per una determinata località e periodo. Lo SPI può essere calcolato su diversi periodi di accumulo delle precipitazioni (tipicamente varia da 1 a 48 mesi). I diversi indicatori SPI risultanti consentono di stimare i differenti impatti potenziali della siccità meteorologica. Da 1 a 3 mesi, lo SPI può essere utilizzato come indicatore per impatti immediati, come la diminuzione dell'umidità del suolo. Questa scala è utile per monitorare la siccità agricola. Una siccità su queste scale può ridurre la disponibilità d'acqua per le colture e influenzare negativamente le rese agricole. Da 3 a 12 mesi, lo SPI da indicazioni sulla siccità idrologica

fornendo utili informazioni per il livello delle riserve d'acqua nei bacini idrografici e nei serbatoi. I valori dello SPI possono variare da fortemente negativi a fortemente positivi. Valori negativi indicano condizioni di siccità, mentre valori positivi indicano condizioni di maggiore umidità. Ad esempio:

- SPI = 0: Precipitazioni medie
- SPI < -2: Siccità estrema
- -1.5 < SPI < -2: Siccità grave
- -1 < SPI < -1.5: Siccità moderata
- -1 < SPI < 1: Condizioni normali
- SPI > 2: Precipitazioni significativamente superiori alla media

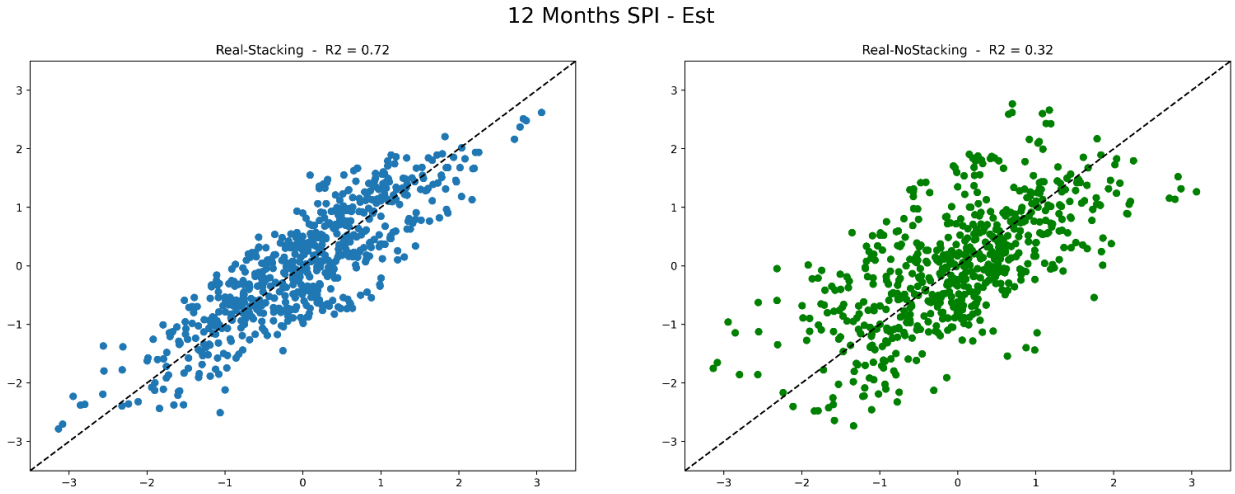


*Figura 19. Confronto dei valori di Standard Precipitation Index (SPI) a 12 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Est*

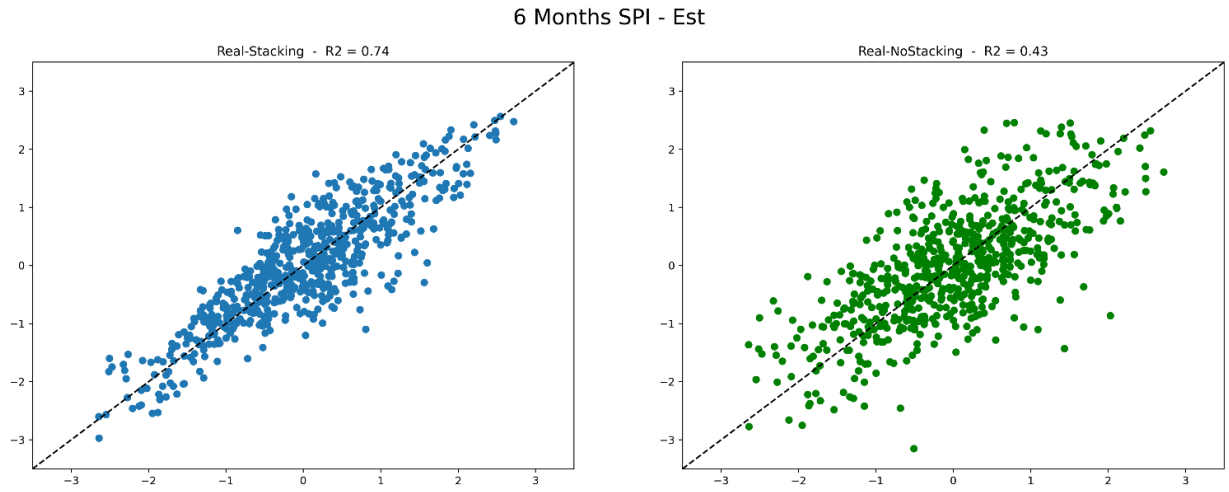


*Figura 30. Confronto dei valori di Standard Precipitation Index (SPI) a 6 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Est*

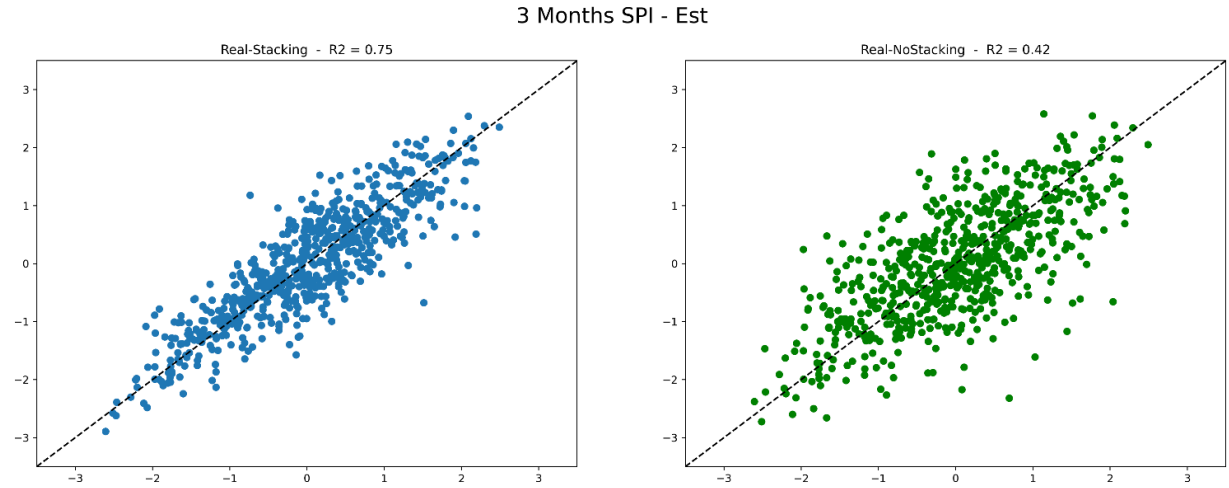


*Figura 31. Confronto dei valori di Standard Precipitation Index (SPI) a 3 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Est*

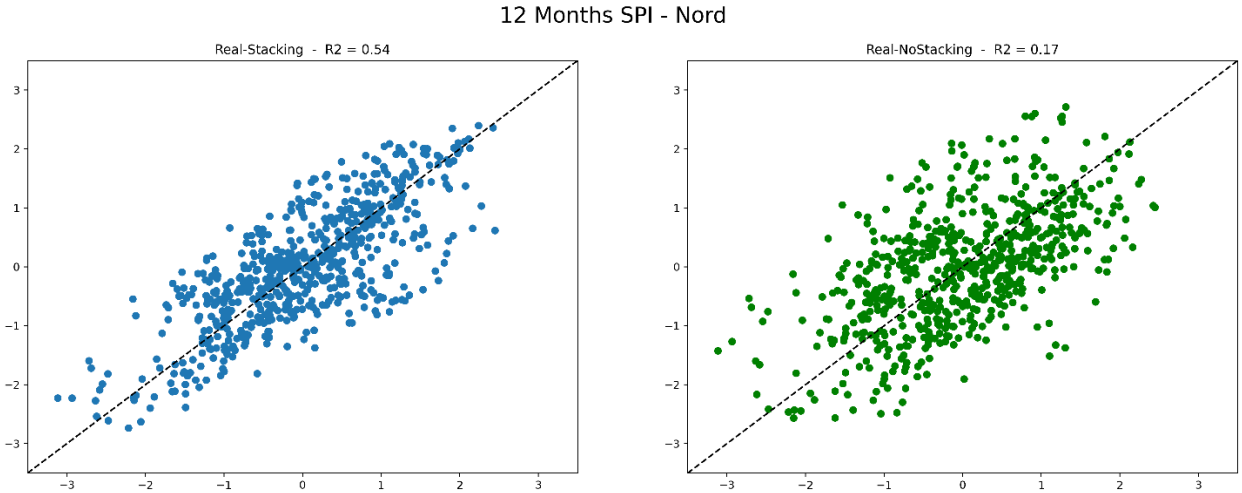


*Figura 32. Confronto dei valori di Standard Precipitation Index (SPI) a 12 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Nord*

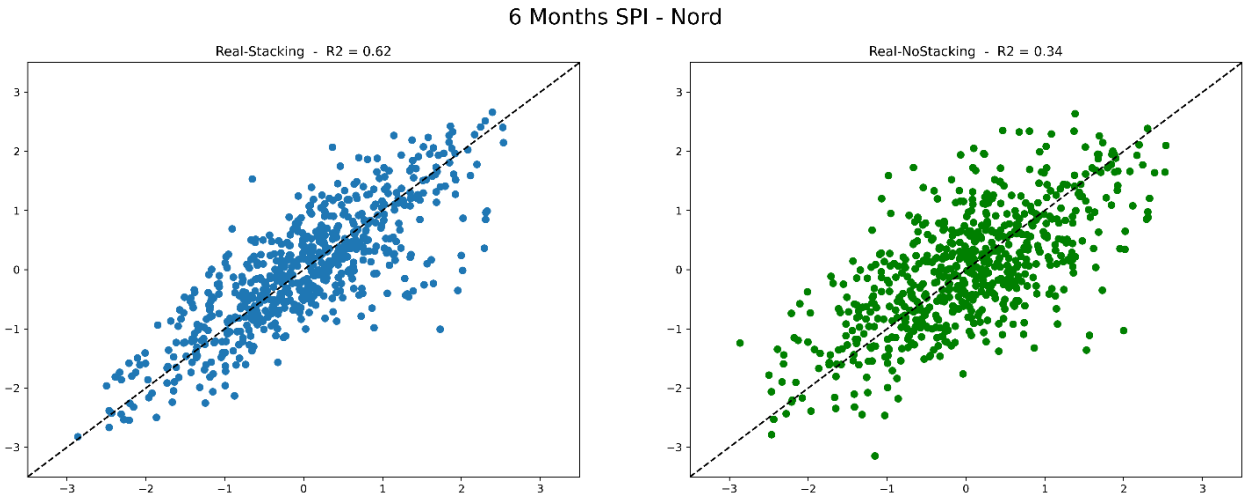


*Figura 33. Confronto dei valori di Standard Precipitation Index (SPI) a 6 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Nord*

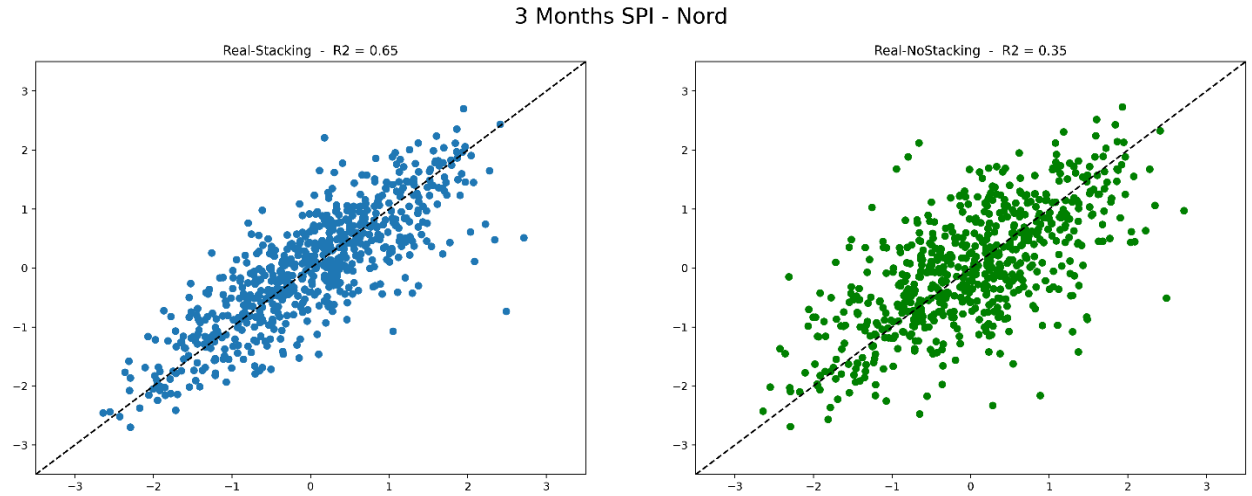


*Figura 34. Confronto dei valori di Standard Precipitation Index (SPI) a 3 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Nord*

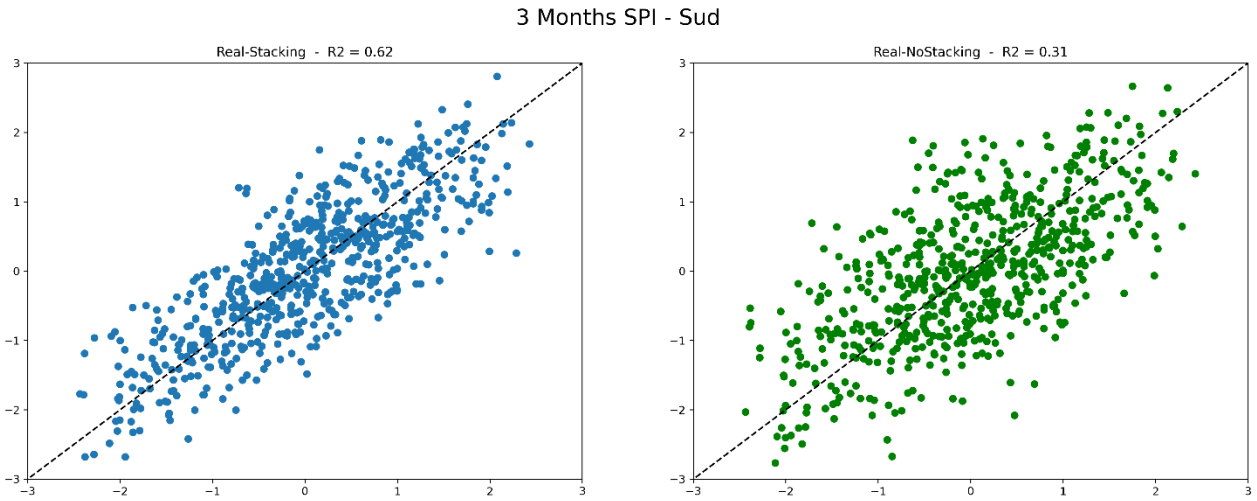


*Figura 35. Confronto dei valori di Standard Precipitation Index (SPI) a 3 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Sud*

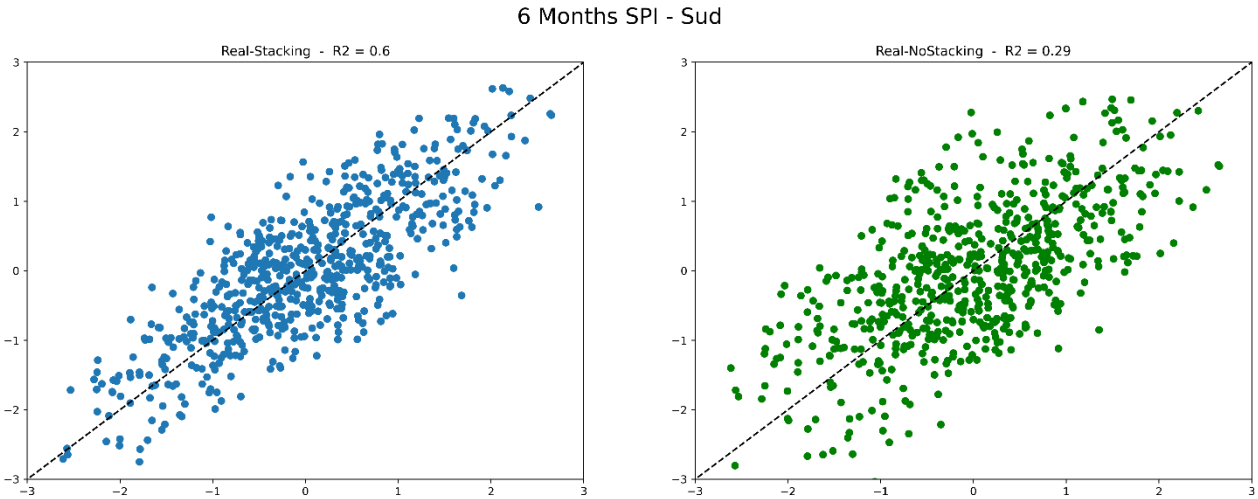


*Figura 36. Confronto dei valori di Standard Precipitation Index (SPI) a 6 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Sud*

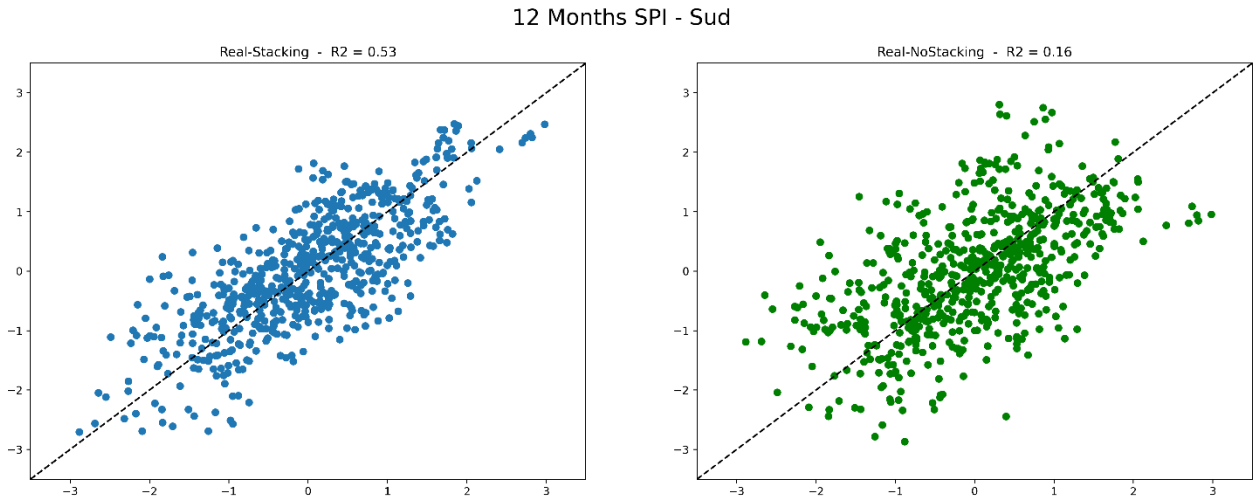


*Figura 37. Confronto dei valori di Standard Precipitation Index (SPI) a 12 mesi, tra valori reali e simulati (con e senza Stacking). Migliori performance si ottengono per valori più vicini alla bisettrice degli assi - Cluster Sud*

I grafici sopra riportati (Fig. 29-37) mostrano una buona correlazione tra il modello di Osservazione e quello di Stacking con valori di R-quadrato compresi tra 0,72 e 0,75. D'altra parte, il modello senza stacking non mostra risultati altrettanto buoni come i precedenti, come dimostrato dai valori di R-quadrato tra 0,32 e 0,43. Per tutti gli intervalli temporali di calcolo dello SPI, il nuovo modello di stacking riduce la varianza dei valori rispetto alla linea bisettrice, dimostrando di essere il modello con le migliori prestazioni di simulazione delle precipitazioni.

## Applicazione e risultati per la Regione Basilicata

Per l'addestramento e la validazione del modello, è stato impiegato un archivio di precipitazioni giornaliere, relativo a un periodo di 50 anni (1951-2000), raccolto da 20 stazioni pluviometriche. Dall'analisi dei parametri BIC e Likelihood, è emerso che, anche per la Regione Basilicata, 5 stati nascosti rappresentano il bilanciamento ideale tra prestazioni di simulazione e tempo di elaborazione. Diversamente da quanto avvenuto per la Regione Lazio, le stazioni pluviometriche non sono state aggregate in cluster omogenei a causa del loro numero limitato.

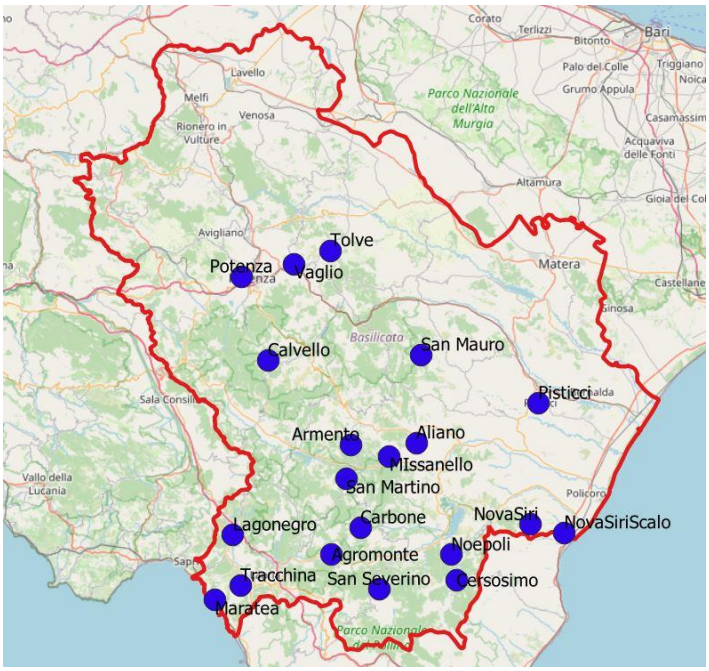


*Figura 38. Stazioni pluviometriche considerate per l'applicazione dei modelli ai dati della Regione Basilicata*

Attraverso le simulazioni, sono stati confrontati i risultati forniti dal modello Stacking con le misurazioni effettive di precipitazioni rilevate dalle stazioni. Il modello ha mostrato un'ottima aderenza ai dati reali sia nelle altezze di pioggia che nel conteggio dei giorni piovosi. Infatti, i valori simulati presentano un MAE di 3,5 mm di pioggia rispetto ai dati osservati, mostrando un errore contenuto. Le discrepanze più significative, in termini di altezze di pioggia, sono state rilevate nei mesi di maggio e giugno. Queste potrebbero derivare dalla variabilità spaziale dei

dati durante l'estate, accentuata dalle differenze climatiche tra la costa Jonica (Est) e l'area del Lagonegrese (Ovest) della regione. Tale ipotesi è corroborata dai boxplot presenti nelle figure allegate (Fig. 39,40), in cui si evidenzia una varianza dei dati sensibilmente maggiore rispetto a quella delle stazioni del Lazio, categorizzate in cluster con caratteristiche di precipitazione omogenee.

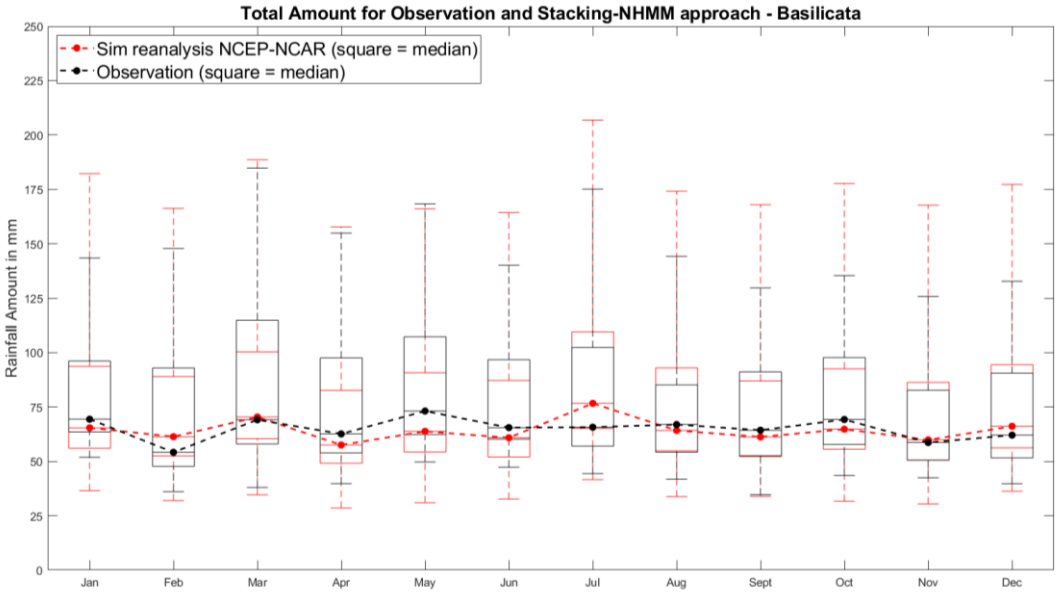


*Figura 29. Confronto tra altezze di pioggia osservate (reali) e simulate con il modello Stacking*

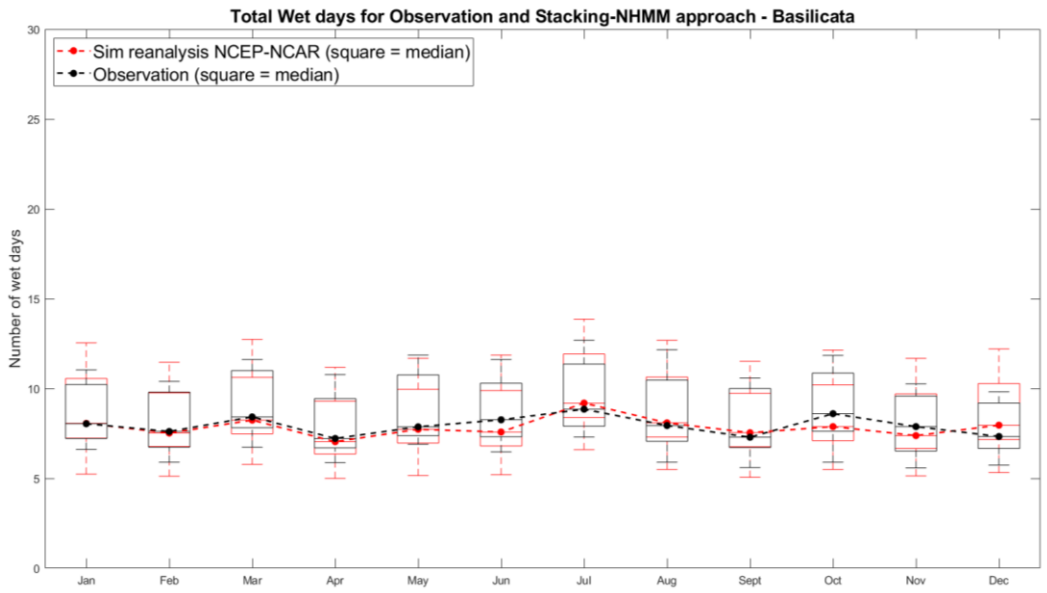


*Figura 40. Confronto tra numero di giorni piovosi osservati (reali) e simulati con il modello Stacking*

# Conclusioni e prospettive future

L'indagine condotta sulle regioni del Lazio e della Basilicata ha evidenziato l'efficacia e l'importanza delle tecniche di clustering e modellazione delle precipitazioni in ambito scientifico. Per la Regione Lazio, la metodologia basata sulla modularity ha dimostrato una capacità notevole nell'identificare cluster omogenei di stazioni pluviometriche, focalizzandosi esclusivamente sull'omogeneità climatica senza essere influenzata da altre variabili potenzialmente non pertinenti.

L'utilizzo della tecnica PCA per i predittori GPH e IVT ha evidenziato una notevole competenza nella riduzione dimensionale, garantendo la conservazione di una parte significativa della varianza intrinseca dei dati. La combinazione di questa metodica con l'adozione del modello NHMM, impostato con cinque stati nascosti, ha offerto un equilibrio ideale tra l'accuratezza delle previsioni e l'efficienza computazionale, suggerendo una direzione promettente per ricerche future.

L'analisi comparativa delle prestazioni dei modelli con e senza Stacking ha consolidato la validità dell'approccio proposto. La significativa riduzione del MAE per la Regione Lazio, unitamente ai risultati ottenuti per la Basilicata, sottolinea l'importanza dell'adozione di tecniche di modellazione avanzate nel dominio delle previsioni meteorologiche, in particolar modo in contesti caratterizzati da una pronunciata variabilità climatica.

Alla luce di questi risultati, si ritiene imperativo proseguire nell'esplorazione e nell'approfondimento di tecniche innovative di modellazione, potenzialmente integrando approcci basati su intelligenza artificiale o algoritmi di machine learning, per affinare ulteriormente la capacità predittiva. Inoltre, l'espansione di tali modelli ad altre regioni potrebbe fornire insights ulteriori sulle loro applicazioni e potenziali limiti.

In una prospettiva futura, la raffinatezza di tali modelli avrà un ruolo determinante non soltanto nella gestione delle risorse idriche, ma anche nella formulazione di strategie adattive in risposta ai mutamenti climatici. Data l'importanza cruciale delle precipitazioni nel bilancio idrologico di una regione, una previsione meticolosa e accurata potrebbe orientare decisioni strategiche in ambito infrastrutturale, agronomico e di protezione civile.

In conclusione, le tecniche e i modelli presentati in questa ricerca rappresentano un passo avanti significativo nella modellazione delle precipitazioni e delineano un percorso ricco di potenziali sviluppi. La loro implementazione, affiancata da ulteriori innovazioni, potrà contribuire a una gestione informata del territorio e a una comprensione sempre più approfondita dei fenomeni climatici, elementi essenziali in un contesto di crescenti sfide ambientali e climatiche.

In particolare la Regione Basilicata, pur mostrando risultati promettenti nella presente ricerca, offre ulteriori spunti di approfondimento e sviluppo metodologico. Una delle principali linee di intervento riguarda la clusterizzazione delle stazioni pluviometriche. Nonostante l'attuale studio non abbia implementato una categorizzazione in cluster omogenei a causa del numero limitato di stazioni, si ritiene che un futuro lavoro in questa direzione potrebbe portare a ulteriori ottimizzazioni. Infatti, la creazione di cluster basati sull'omogeneità climatica potrebbe ridurre ulteriormente la discrepanza tra le precipitazioni osservate e quelle simulate, assicurando una maggiore precisione delle previsioni.

Un altro aspetto che potrebbe arricchire ulteriormente il modello riguarda l'introduzione di un terzo predittore a grande scala, specificamente la temperatura dell'aria. La correlazione tra le variazioni di temperatura e i pattern di precipitazione è ben documentata nella letteratura scientifica, e l'incorporazione di questa variabile potrebbe rafforzare la robustezza del modello e affinare ulteriormente la sua accuratezza predittiva.

Infine, per assicurare una copertura più completa del territorio regionale, si intende fare riferimento al dataset ISPRA – SCIA. Questo dataset, caratterizzato da una griglia regolare di 10 km x 10 km, fornirà una risoluzione spaziale dettagliata delle precipitazioni, potenziando così la rappresentatività del modello sull'intero territorio regionale. L'utilizzo di tale dataset consentirà non solo di comprendere meglio la distribuzione delle precipitazioni a livello locale, ma anche di integrare la modellazione su aree precedentemente non coperte o meno rappresentate.

Le prospettive future per la Regione Basilicata sono orientate verso una sempre maggiore precisione e completezza della modellazione delle precipitazioni. Attraverso l'implementazione di nuovi predittori, l'adozione di tecniche avanzate di clusterizzazione e l'utilizzo di dataset ad alta risoluzione, si mira a fornire un quadro sempre più dettagliato e accurato del regime pluviometrico della regione, con impatti diretti sulla gestione delle risorse e sulle politiche territoriali.

# Parte II: Analisi delle interazioni tra parametri ambientali e malattie cardiovascolari attraverso tecniche di machine learning

## Introduzione

Le malattie cardiovascolari (CVD) costituiscono la principale causa di mortalità globale, superando tutte le altre patologie [33]. Tra queste, primeggiano la malattia cardiaca ischemica e la malattia cerebrovascolare. Secondo l'IPCC (Intergovernmental Panel on Climate Change), il cambiamento climatico è probabile che influenzi la salute umana in modo diretto, attraverso le fluttuazioni di temperatura, e in modo indiretto, attraverso modifiche nei vettori di malattia, come le zanzare, e altri fattori [34, p. 2]. È stato osservato che le ondate di freddo e di calore sono aumentate a causa del cambiamento climatico [35] [36] [37]. La comprensione dell'interazione tra il cambiamento climatico e le malattie cardiovascolari è fondamentale per sviluppare strategie di prevenzione e mitigazione che possano ridurre il carico di malattie cardiovascolari legate al clima.

Una revisione completa della letteratura scientifica esistente ha rivelato che l'aumento delle temperature porterà molto probabilmente a un incremento della morbilità e della mortalità legate alle condizioni meteorologiche, con una parte significativa dei decessi correlati a eventi cardiovascolari [38] [39] [40]. Numerosi studi condotti in diverse parti del mondo hanno confermato che le temperature estreme aumentano il rischio di mortalità da CVD [41] [42] [43] [44] [45]. E' stato osservato come le ondate di calore possano causare un aumento della mortalità dovuta a patologie cardiovascolari (dal 13% al 90%), e cerebrovascolari (dal 6 al 52%) e respiratorie (fino al 14%)[46].

Negli Stati Uniti, si sono verificati circa 5600 decessi legati al calore ogni anno dal 1997 al 2006 in 297 contee [47]. Episodi come l'ondata di calore del Luglio 2006 avvenuta in California confermano la notevole incidenza di accessi al pronto soccorso per patologie cardiovascolari [48], registrando circa 140 morti il 15 Luglio e l'1 Agosto dello stesso anno. Studi condotti in 9 città statunitensi hanno identificato un aumento del 1,8% nella mortalità associato ad incrementi della temperatura apparente [49]. Allo stesso modo, in Nord America, un aumento di 4,7 °C nella temperatura media giornaliera è stato correlato a un aumento del 2,6% nella mortalità cardiovascolare [20]. Inoltre, nelle regioni dove la temperatura nei mesi più caldi supera i 30 °C, ogni grado di aumento è associato a un incremento del 3% nella mortalità [50].

Alcuni studi dimostrano come il rischio di mortalità da CVD aumenti sia durante i giorni caldi che freddi. Associazioni simili tra temperatura e mortalità sono state osservate anche in Cina, dove si verifica un aumento del rischio sia a basse che alte temperature [51]. Ad esempio, un'analisi sugli anziani (>65 anni) ha dimostrato che un aumento di 1 °C nella temperatura ha portato a un aumento del 3,44% nella mortalità cardiovascolare, mentre una diminuzione di 1 °C ha comportato un aumento del 1,66% nella mortalità cardiovascolare [52].

In generale, le analisi dei tassi di mortalità giornalieri hanno evidenziato che sia le basse che le alte temperature sono associate a un aumento della mortalità da CVD [53]. L'esposizione cronica al freddo o al calore può compromettere la funzione cardiovascolare, portando a una maggiore suscettibilità a infarti, aritmie cardiache maligne, malattie tromboemboliche e sepsi indotte dal calore come lo shock [54]. Le variazioni della temperatura ambiente contribuiscono all'aumento della pressione sanguigna, della viscosità del sangue e della frequenza cardiaca, aumentando così il rischio cardiovascolare [54]. La maggior parte dei decessi durante le ondate di calore si verifica in individui con CVD croniche preesistenti [54]. Di fatto, relativamente a CVD, è stato dimostrato come esista una forte correlazione positiva tra temperatura massima e mortalità ($r = 0,83$, $p < 0,01$), ma anche una correlazione negativa significativa tra le temperature minime e la mortalità [55].

Le fluttuazioni stagionali nell'incidenza delle CVD sono motivo di notevole preoccupazione. Si è osservato un incremento delle ospedalizzazioni e degli eventi fatali in specifici periodi dell'anno, in particolar modo durante la stagione invernale [56]. Tale fenomeno risulta particolarmente accentuato nelle popolazioni residenti in climi miti, le quali potrebbero essere meno adattate a variazioni climatiche estreme nel corso dell'anno [56]. La stagionalità ha un impatto significativo sull'incidenza di numerosi sottotipi di CVD [56]. Numerosi studi hanno evidenziato come il periodo invernale sia correlato a un marcato aumento delle patologie cardiovascolari e dei decessi ad esse correlati, specialmente nelle regioni dell'emisfero settentrionale caratterizzate da temperature particolarmente rigide [44] [57]. In particolare, i tassi giornalieri di eventi cardiovascolari aumentano con la diminuzione della temperatura media dell'aria, con una diminuzione di 10 °C associata a un aumento del 19% nei tassi giornalieri di eventi cardiovascolari per individui sopra i 65 anni [58].

Va sottolineato che il cambiamento climatico non influenza solo le temperature, ma ha anche effetti avversi su altre condizioni ambientali, in particolare sull'inquinamento dell'aria [59]. Secondo uno studio sul Global Burden of Disease, l'inquinamento dell'aria è stato responsabile di almeno 9 milioni di decessi globali nel 2019 [60]. I dati dell'OMS indicano che quasi l'intera

popolazione globale respira aria che supera i limiti guida dell'OMS e contiene alti livelli di inquinanti [61]. Nelle aree urbane, l'impatto è ancora più significativo, poiché il cambiamento climatico influisce sull'inquinamento dell'aria esterna, legato ai modelli locali di temperatura, vento e precipitazioni [62]. Questi cambiamenti ambientali stanno già causando eventi acuti di CVD quantificabili ed evitabili e dovrebbero essere integrati nei nostri sforzi per prevenire e trattare le CVD [63].

In un'epoca caratterizzata da un crescente accumulo di dati clinici, biometrici e di biomarcatori, i medici si trovano a gestire una mole sempre maggiore di informazioni. Nell'era dei "big data", si sta diffondendo l'idea che la risposta a tutte le domande cliniche e scientifiche possa essere trovata nei "big data", e che questi dati possano trasformare la medicina tradizionale in una medicina di precisione. Tuttavia, con l'introduzione recente della scienza dei dati nel settore sanitario, emerge l'opportunità di riconsiderare questa visione centrata sui dati. Una comprensione approfondita della scienza dei dati e delle sue applicazioni è essenziale per interpretare i risultati e per tradurre le nuove scoperte in pratica clinica [64].

L'impiego dell'intelligenza artificiale (IA) e del machine learning nel campo dell'assistenza sanitaria apre nuove frontiere di ricerca e applicazione. Queste tecnologie, con la loro capacità di analizzare grandi quantità di dati, hanno trovato impiego in diversi ambiti, dalla sorveglianza delle malattie più precisa all'interpretazione avanzata delle immagini e alla gestione ottimizzata delle operazioni sanitarie. Inoltre, considerando il carattere spesso empirico della pratica medica, l'intelligenza artificiale si rivela uno strumento aggiuntivo estremamente utile, che va ad affiancarsi agli strumenti esistenti, fornendo un supporto significativo nella decisione clinica e nella personalizzazione delle cure.

L'indagine condotta presso il Policlinico Giovanni XXIII di Bari [65] ha non solo messo in luce l'abilità dell'IA nel modellare le correlazioni tra condizioni climatiche e l'incidenza delle malattie cardiovascolari (CVD), ma ha anche evidenziato il potenziale delle tecniche di IA nell'utilizzare dati climatici per simulare le CVD. Mediante tecniche di importanza delle caratteristiche derivate dall'algoritmo Random Forest, variabili meteorologiche come la temperatura media, massima e apparente, insieme all'umidità relativa, sono state identificate come indicatori chiave delle ospedalizzazioni correlate alle CVD.

Dal punto di vista economico e operativo, i benefici dell'IA nell'ambito sanitario sono tangibili sia in termini di miglioramenti dei risultati nell'assistenza sia nell'assistenza ai professionisti sanitari. Il mercato dell'IA nel settore sanitario evidenzia una crescita impressionante, con un tasso di crescita annuale composto globale del 28% [66]. Questo non solo promette un elevato potenziale di mercato, ma dimostra che l'implementazione dell'IA può concretamente condurre

a una riduzione dei costi sanitari e al contempo migliorare l'esito della salute per la popolazione [66]. L'attuale epoca è caratterizzata dalla raccolta di dati massicci riguardanti inquinanti, salute pubblica e fattori ambientali. Questi dati, pur rappresentando opportunità notevoli, sfidano i metodi epidemiologici tradizionali. Per affrontare queste sfide, si è orientati sempre più verso algoritmi di data mining e apprendimento automatico, specialmente nell'epidemiologia dell'inquinamento atmosferico [67].

L'analisi della letteratura recente ha confermato l'efficacia delle tecniche di machine learning nell'analisi dei dati clinici, offrendo intuizioni preziose per la diagnosi precoce e la gestione delle malattie [68] [69] [70]. Questi dati suggeriscono che il machine learning, quando applicato in contesti clinici specifici, può superare i metodi tradizionali, offrendo modelli previsionali di maggiore precisione. Numerosi studi [68] [69] [70] [71] [72] [73] [74] hanno analizzato l'applicazione del machine learning nelle diagnosi e previsioni mediche, dalle malattie cardiovascolari al deterioramento clinico dei pazienti. Questi studi hanno impiegato una varietà di algoritmi, tra cui reti neurali, SVM, metodi Boosting come il Random Forest, e combinazioni di tecniche come CNN e LSTM, raggiungendo performance previsionali elevate (AUC di 0,809 [75]).

Infine, studi incentrati sui determinanti sociali e sui dati nutrizionali hanno mostrato che l'integrazione di queste informazioni in modelli basati su machine learning può migliorare significativamente la previsione del rischio cardiovascolare [33] [76] [77] [75]. Queste ricerche sottolineano l'importanza crescente delle tecniche di machine learning nella gestione avanzata delle patologie e nella previsione dei rischi associati.

In conclusione, l'evoluzione dell'IA e del machine learning rappresenta una frontiera cruciale nel panorama dell'assistenza sanitaria contemporanea, con promesse di innovazioni significative e profonde per il futuro. Gli studi che utilizzano l'intelligenza artificiale nel contesto climatico e delle patologie rappresentano una promettente area di studio, in cui variabili meteorologiche possono essere utilizzate per prevedere e gestire meglio le condizioni di salute della popolazione. Questi approcci innovativi possono contribuire a mitigare gli effetti negativi del cambiamento climatico sulla salute umana.

L'uso di metodi di machine learning ha dimostrato di avere un potenziale significativo nell'analisi e nella previsione delle malattie cardiovascolari. Questi metodi offrono nuove opportunità per migliorare la previsione del rischio, la decisione clinica e l'identificazione dei pazienti a rischio di futuri eventi avversi. Inoltre, le evidenze scientifiche convergono nel mostrare come il cambiamento climatico abbia un impatto significativo sulla salute umana, in particolare sulla salute cardiovascolare. Le variazioni delle temperature, le ondate di calore e le basse temperature

possono aumentare in modo sostanziale il rischio di mortalità da CVD. È importante riconoscere che questo fenomeno non riguarda solo le regioni con climi estremamente freddi, ma anche le aree con climi più temperati. Inoltre, il cambiamento climatico non si limita alle temperature, ma ha anche conseguenze negative sull'inquinamento dell'aria, il che contribuisce ulteriormente al rischio cardiovascolare.

La previsione e la gestione delle patologie cardiovascolari è di fondamentale importanza per garantire un'assistenza sanitaria ottimale. Con l'avvento della digitalizzazione e l'accumulo di dati sanitari, il machine learning emerge come uno strumento potente per affrontare queste sfide, aprendo una nuova prospettiva per simulare e prevedere tali patologie basandosi su parametri climatici. Questo lavoro si concentra sull'analisi e l'integrazione di vari studi che hanno esplorato l'applicazione di algoritmi di machine learning ai dati sanitari per migliorare la previsione del rischio e la decisione clinica per queste malattie.

L'obiettivo primario di questo studio è investigare l'impatto delle variabili climatiche sulle patologie cardiovascolari, al fine di utilizzare queste variabili per sviluppare un quadro di intervento preventivo per salvaguardare la salute umana. Per affrontare questa sfida, proponiamo uno schema prescritto che sfrutta i metodi dell'IA per simulare e comprendere la relazione tra le condizioni climatiche e l'insorgenza delle patologie cardiovascolari. L'approccio proposto riconosce la natura complessa di questa questione e prende in considerazione le variabili meteorologiche più pertinenti, incluse la temperatura media, la temperatura massima, la temperatura apparente e l'umidità relativa. Impiegando tecniche di importanza delle caratteristiche basate sull'algoritmo Random Forest, il nostro studio mira ad identificare i parametri climatici che contribuiscono alle ospedalizzazioni dovute a patologie cardiovascolari. Questa comprensione globale dell'interazione tra variabili clima-meteo e patologie cardiovascolari ci permetterà di identificare le popolazioni vulnerabili e formulare strategie di intervento preventivo mirate. In sostanza, questo nuovo approccio proposto offre un quadro pratico per i decisori politici e i professionisti sanitari per mitigare gli effetti nocivi del cambiamento climatico sulla salute cardiovascolare. Attraverso l'integrazione dei metodi IA e dei dati climatici, questo studio contribuisce a una comprensione migliorata dei meccanismi sottostanti e facilita lo sviluppo di misure preventive efficaci.

# Materiali e Metodi

## Dati ospedalieri

Lo studio ha analizzato i dati giornalieri degli accessi al pronto soccorso dell'Ospedale Policlino di Bari tra il 2013 e il 2021. Nel campo "main problem" del database sono indicate le patologie dei pazienti al momento dell'arrivo. Tali patologie sono codificate in 33 categorie, elencate nella Tabella 3.

*Tabella 2. Principali patologie di accesso al pronto soccorso e relativo codice identificativo*

| CODE | MAIN PROBLEMS/SYMPTOMATOLOGY | CODE | MAIN PROBLEMS/SYMPTOMATOLOGY |
|---|---|---|---|
| 1 | COMA | 18 | OTORHINO LARYNGEAL SYMPTOMS OR DISORDERS |
| 2 | ACUTE NEUROLOGICAL SYNDROME | 19 | OBSTETRIC-GYNECOLOGICAL SYMPTOMS OR DISORDERS |
| 3 | OTHER NERVOUS SYSTEM SYMPTOMS | 20 | DERMATOLOGICAL SYMPTOMS OR DISORDERS |
| 4 | ABDOMINAL PAIN | 21 | ODONTOSTOMATOLOGICAL SYMPTOMS OR DISORDERS |
| 5 | CHEST PAIN | 22 | UROLOGICAL SYMPTOMS OR DISORDERS |
| 6 | DYSPNEA | 23 | OTHER SYMPTOMS OR DISORDERS |
| 7 | PRE CORDIAL PAIN | 24 | LEGAL-MEDICAL INVESTIGATIONS |
| 8 | SHOCK | 25 | SOCIAL PROBLEM |
| 9 | NON-TRAUMATIC HEMORRAHAGE | 26 | FALL FROM HIGH |
| 10 | TRAUMA | 27 | SCALDING |
| 11 | INTOXICATION | 28 | PSYCHIATRIC |
| 12 | FEVER | 29 | PNEUMOLOGY-RESPIRATORY PATHOLOGY |
| 13 | ALLERGIC REACTION | 30 | VIOLENCE FROM OTHER |
| 14 | CHANGES IN RHYTHM | 31 | SELF-HARM |
| 15 | HYPERTENSION | 98 | DEHYDRATION |
| 16 | PSYCHOMOTOR AGITATION | 99 | ANIMAL BITE |
| 17 | EYE SYMPTOMS OR DISORDERS | | |

Nella fase iniziale di elaborazione, i dati sono stati organizzati anno per anno, dal 2013 al 2021. Questi dati evidenziano le tendenze e le distribuzioni degli accessi al pronto soccorso, suddivisi per sesso. Per un dettaglio completo degli accessi per ciascun anno, incluse le suddivisioni per sesso e i dati mancanti, si rimanda alla Tabella 4 e alla Figura 41.

*Tabella 3. Accessi al pronto soccorso per genere, dal 2013 al 2021*

| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|---|
| **Total admissions in the ER** | 75,927 | 80,690 | 75,334 | 71,550 | 65,984 | 65,641 | 68,052 | 43,729 | 48,489 |
| **Male** | 40,265 | 42,554 | 40,091 | 38,007 | 35,130 | 34,798 | 36,034 | 24,517 | 26,904 |
| **Female** | 35,032 | 37,127 | 34,327 | 32,914 | 30,343 | 30,544 | 31,442 | 18,960 | 21,355 |
| **Undeclared gender** | 630 | 1009 | 916 | 629 | 511 | 299 | 576 | 252 | 230 |

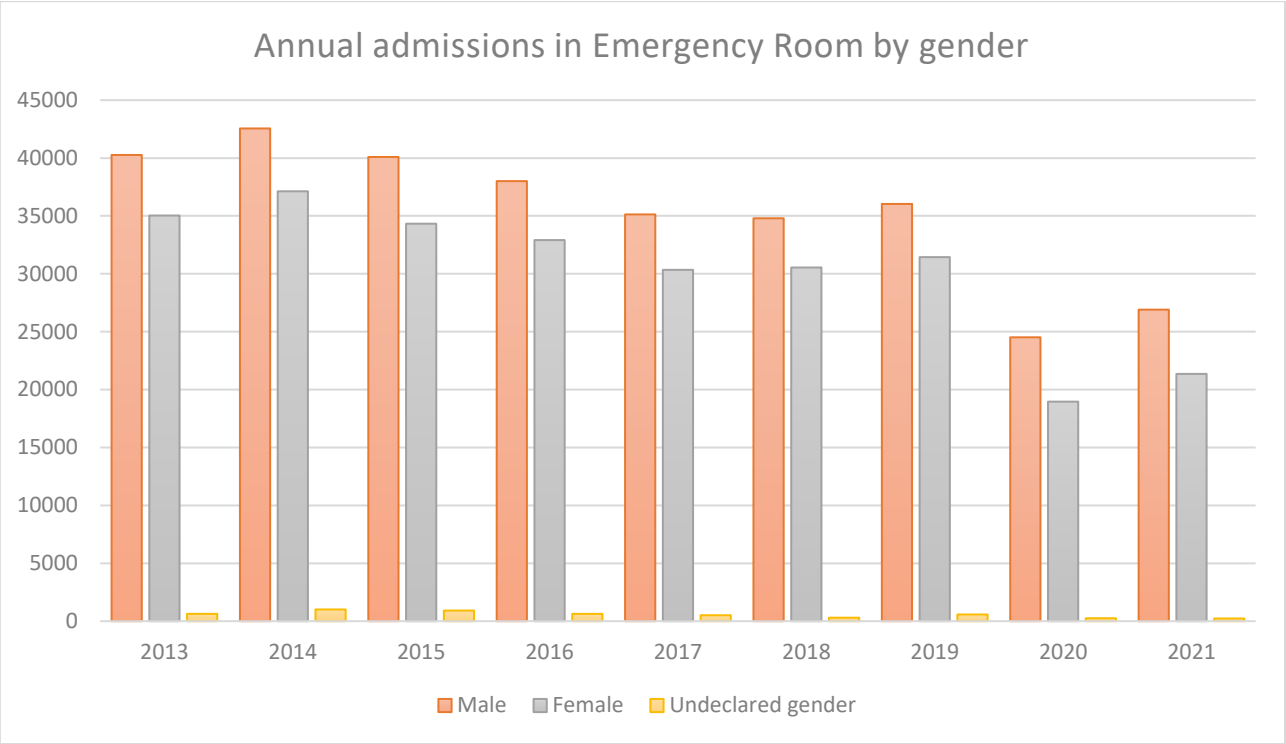


*Figura 41. Accessi al pronto soccorso per genere, dal 2013 al 2021*

Per lo scopo di questo studio, sono stati considerati solo dati e statistiche relativi a patologie cardiovascolari e fortemente correlati con fattori meteorologici (Tabella 5).

*Tabella 4. Selezione delle patologie di tipo cardiovascolare*

| Code | Specific Problem | Classification |
|---|---|---|
| 5 | CHEST PAIN | CARDIOVASCULAR DISEASES |
| 7 | PRECORDIAL PAIN | |
| 14 | CHANGES IN RHYTHM | |
| 15 | HYPERTENSION | |

L'analisi degli effetti delle temperature ambientali calde e fredde sulla mortalità e morbilità degli anziani (oltre 65 anni) ha evidenziato che un aumento di 1 °C della temperatura media stagionale porta a un incremento della mortalità cardiovascolare del 3,44%, mentre una diminuzione di 1 °C la aumenta del 1,66% [78]. Pertanto, tra le 33 patologie descritte nella Tabella 3, sono state

selezionate solo quelle relative alle malattie cardiovascolari (MCV) secondo la classificazione riportata nella Tabella 5. Queste sono state analizzate per gli anni in esame, come mostrato nelle Figure 41-42. È importante notare che le variazioni climatiche possono influenzare diversamente ciascun individuo. Un primo passo per identificare eventuali differenze significative tra gli individui si basa sull'analisi del genere. La Tabella 6 e la Fig. 41 illustrano gli accessi annuali al pronto soccorso per patologie cardiovascolari suddivisi per genere dal 2013 al 2021. Un ulteriore criterio distintivo importante è l'età del paziente. La distribuzione degli accessi cardiovascolari al pronto soccorso in base all'età è illustrata nelle Tabelle 7-8 e nella Fig. 43, come percentuale del totale degli accessi.

*Tabella 5. Accessi al pronto soccorso per CVD, classificati per genere, dal 2013 al 2021*

| | | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Cardiovascular diseases** | **Admissions in the ER** | 6854 | 6252 | 5728 | 5319 | 4284 | 4558 | 4615 | 2268 | 2040 |
| | **Male** | 3762 | 6422 | 3143 | 2893 | 2396 | 2586 | 2548 | 1353 | 1256 |
| | **Female** | 3040 | 2781 | 2540 | 2393 | 1873 | 1955 | 2050 | 908 | 778 |
| | **-** | 52 | 49 | 45 | 33 | 15 | 17 | 17 | 7 | 6 |

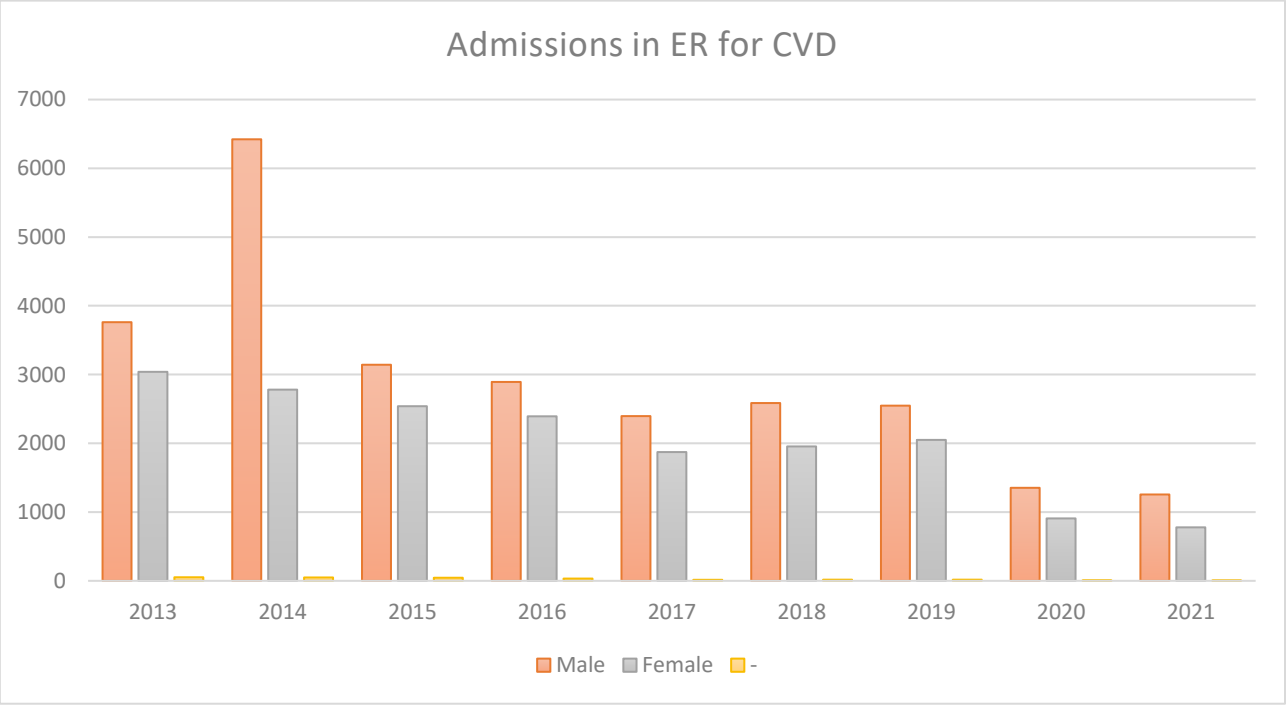


*Figura 42. Accessi al pronto soccorso per CVD, classificati per genere, dal 2013 al 2021*

*Tabella 6. Accessi al pronto soccorso per CVD, classificati per range di età*

| | Admissions in Emergency Room for cardiovascular diseases | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Year** | **under 20** | **20–29** | **30–39** | **40–54** | **55–64** | **65–75** | **over 75** | **tot** | **tot admission** |
| **2013** | 92 | 447 | 688 | 1617 | 1236 | 1440 | 1326 | 6846 | 75,927 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **2014** | 95 | 401 | 597 | 1532 | 1122 | 1251 | 1250 | 6248 | 80,690 |
| **2015** | 71 | 348 | 545 | 1456 | 1035 | 1218 | 1053 | 5726 | 75,334 |
| **2016** | 88 | 338 | 464 | 1320 | 1020 | 1073 | 1016 | 5319 | 71,550 |
| **2017** | 35 | 238 | 388 | 994 | 865 | 819 | 894 | 4233 | 65,985 |
| **2018** | 63 | 288 | 394 | 1232 | 905 | 895 | 778 | 4555 | 65,641 |
| **2019** | 86 | 307 | 379 | 1183 | 942 | 918 | 797 | 4612 | 68,052 |
| **2020** | 31 | 129 | 165 | 589 | 470 | 469 | 413 | 2266 | 43,729 |
| **2021** | 32 | 161 | 178 | 482 | 467 | 381 | 338 | 2040 | 48,489 |

*Tabella 7. Accessi al pronto soccorso per CVD, classificati per range di età (%)*

| | Emergency Room admissions (%) for cardiovascular diseases | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **year** | **under 20** | **20–29** | **30–39** | **40–54** | **55–64** | **65–75** | **over 75** | **tot** |
| **2013** | 1.34 | 6.53 | 10.05 | 23.62 | 18.05 | 21.03 | 19.37 | 9.02 |
| **2014** | 1.52 | 6.42 | 9.56 | 24.52 | 17.96 | 20.02 | 20.01 | 7.74 |
| **2015** | 1.24 | 6.08 | 9.52 | 25.43 | 18.08 | 21.27 | 18.39 | 7.60 |
| **2016** | 1.65 | 6.35 | 8.72 | 24.82 | 19.18 | 20.17 | 19.10 | 7.43 |
| **2017** | 0.83 | 5.62 | 9.17 | 23.48 | 20.43 | 19.35 | 21.12 | 6.42 |
| **2018** | 1.38 | 6.32 | 8.65 | 27.05 | 19.87 | 19.65 | 17.08 | 6.94 |
| **2019** | 1.86 | 6.66 | 8.22 | 25.65 | 20.42 | 19.90 | 17.28 | 6.78 |
| **2020** | 1.37 | 5.69 | 7.28 | 25.99 | 20.74 | 20.70 | 18.23 | 5.18 |
| **2021** | 1.57 | 7.89 | 8.73 | 23.63 | 22.89 | 18.68 | 16.57 | 4.21 |

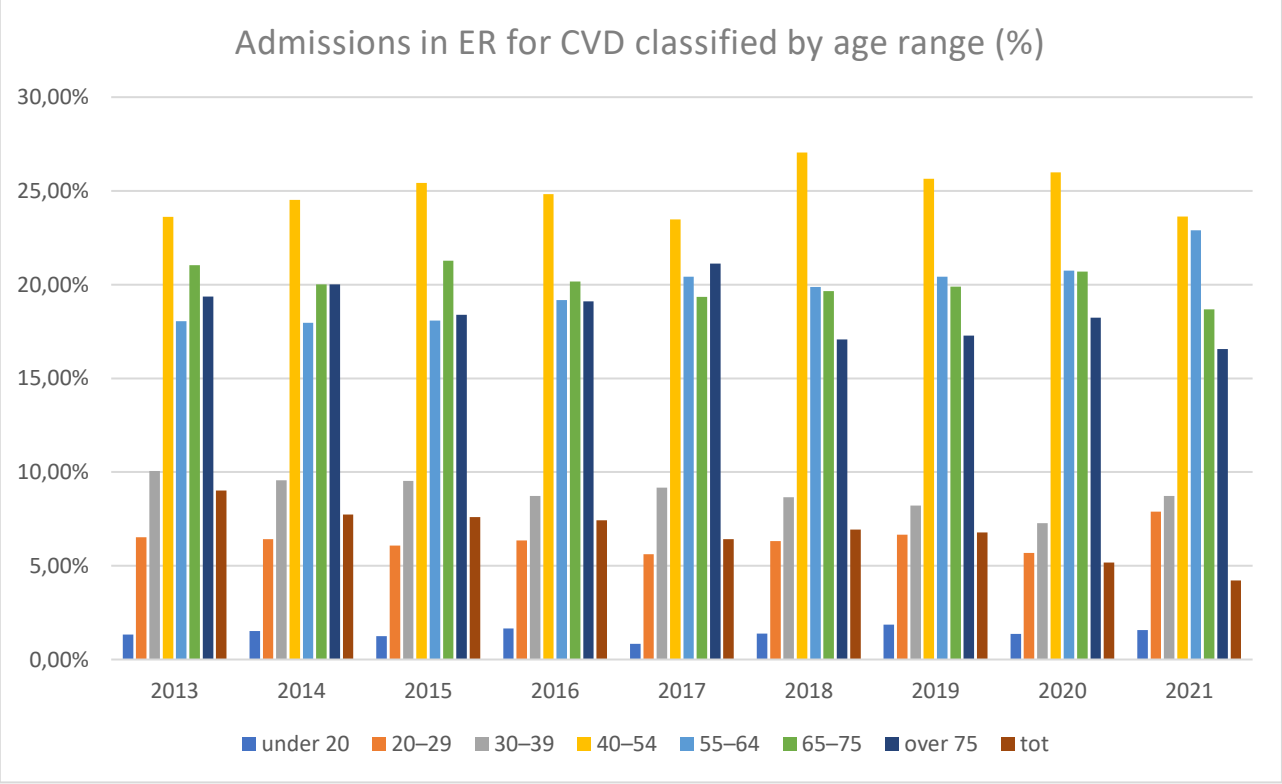


*Figura 43. Accessi al pronto soccorso per CVD, classificati per range di età (%)*

Dal 2013 al 2019, si osserva un aumento costante degli accessi al pronto soccorso per le CVD, con un picco evidente nelle fasce d'età comprese tra i 40 e i 54 anni e tra i 65 e i 75 anni. Questo

indica che le malattie cardiovascolari sembrano essere più comuni in queste fasce d'età. Tuttavia, a partire dal 2020, si verifica una diminuzione significativa del numero di accessi in tutti i gruppi di età, che è continuata anche nel 2021. Questo calo potrebbe essere influenzato da vari fattori, inclusi gli effetti della pandemia da COVID-19.

Interessante notare che, nonostante il calo nei numeri assoluti di accessi alle CVD nel 2020 e nel 2021, le percentuali relative rimangono relativamente costanti o addirittura in aumento, specialmente in alcune fasce d'età. Ciò suggerisce che, sebbene meno persone abbiano visitato gli ospedali, la proporzione di coloro che lo hanno fatto a causa di malattie cardiovascolari è rimasta significativa. In particolare, la fascia d'età 55-64 sembra mostrare un aumento costante delle percentuali di accessi alle CVD dal 2013 al 2021. Tuttavia, va notato che la percentuale totale di accessi alle CVD rispetto al totale degli accessi in pronto soccorso è diminuita negli ultimi due anni. Questo potrebbe indicare che altre emergenze o patologie, come la pandemia da COVID-19, abbiano avuto un impatto maggiore sulle visite ospedaliere complessive.

## Parametri climatici

I parametri meteo-climatici considerati in questo studio includono: temperatura media minima giornaliera (Tmin), temperatura media giornaliera (Tmean), temperatura massima media giornaliera (Tmax), temperatura media giornaliera del punto di rugiada (Dewp), temperatura apparente media giornaliera (Tapp), pressione atmosferica media giornaliera (Patm) e umidità relativa media giornaliera (RH). I parametri relativi alla qualità dell'aria considerati includono CO (monossido di carbonio), O3 (ozono), PM10 (materiale particolato), SO2 (biossido di zolfo) e NO2 (biossido di azoto) (Tabella 9). I dati meteorologici per la città di Bari nel periodo di riferimento 2013-2021 sono stati ottenuti dal sito web di Arpa Puglia e dalla rete di misurazione di Meteonetwork. Arpa Puglia gestisce due reti per le attività di monitoraggio:

- Una rete dedicata: 5 stazioni automatiche situate presso le sue sedi provinciali (Bari, Brindisi, Foggia, Lecce e Taranto).
- Una rete meteorologica che supporta la rete di monitoraggio della qualità dell'aria, attualmente composta da 19 stazioni.

I dati sulla qualità dell'aria per la città di Bari dal 2013 al 2021 sono stati ottenuti dal sito web di Arpa Puglia [79], tramite le stazioni di monitoraggio di Bari-Caldarola, Bari-CUS, Bari-Kennedy, Bari-Carbonara e il laboratorio mobile. I dati meteorologici vengono registrati con frequenza semioraria per le stazioni meteorologiche di Arpa Puglia e con frequenza giornaliera di cinque minuti e un'ora per la rete di misurazione di Meteonetwork. La frequenza oraria riguarda solo

gli anni 2020 e 2021. I dati sulla qualità dell'aria sono registrati su base giornaliera. La temperatura apparente e il SO2 (biossido di zolfo) sono stati esclusi rispettivamente dai parametri meteorologici e dalla qualità dell'aria a causa della mancanza di dati rappresentativi, che sono insufficienti per un'elaborazione completa ed accurata degli stessi.

*Tabella 8. Statistica descrittiva dei parametri climatici*

| | min | avg | max | std | 75th | 50th | 25th |
|---|---|---|---|---|---|---|---|
| **Tmin (°C)** | – 0.17 | 17.08 | 32.86 | 6.40 | 22.50 | 16.65 | 11.80 |
| **Tmean (°C)** | – 0.11 | 17.78 | 33.59 | 6.40 | 23.34 | 17.20 | 12.40 |
| **Tmax (°C)** | – 0.04 | 18.51 | 41.60 | 6.70 | 23.92 | 17.90 | 13.12 |
| **Dewp (°C)** | – 6.38 | 11.92 | 26.00 | 5.52 | 16.49 | 12.00 | 7.82 |
| **Tapp (°C)** | 2.96 | 23.32 | 52.78 | 8.74 | 30.25 | 22.19 | 15.80 |
| **P_atm (hPa)** | 976.60 | 1010.93 | 1039.35 | 8.25 | 1016.38 | 1011.55 | 1005.30 |
| **RH (%)** | 25.49 | 70.22 | 99.00 | 10.89 | 78.00 | 71.00 | 62.95 |
| **CO (ppm)** | 0.10 | 0.77 | 3.00 | 0.41 | 1.00 | 0.70 | 0.50 |
| **O3 (µg/m³)** | 13.00 | 83.51 | 154.00 | 21.09 | 99.00 | 83.00 | 68.00 |
| **PM10 (µg/m³)** | 1.00 | 22.54 | 117.00 | 11.30 | 27.00 | 21.00 | 15.00 |
| **SO2 (µg/m³)** | 0.00 | 17.41 | 104.00 | 21.41 | 26.90 | 6.90 | 3.10 |
| **NO2 (µg/m³)** | 5.00 | 52.96 | 157.00 | 25.65 | 69.00 | 50.00 | 33.00 |
| **CVD (ER visit count)** | 0.00 | 12.90 | 37.00 | 6.30 | 17.00 | 13.00 | 8.00 |

*Legenda: Tmin, Minimum Temperature; Tmean, Mean Temperature; Tmax, Maximum Temperature; Dewp, Dew Point; Tapp, Apparent Temperature; P_atm, Atmospheric Pressure; RH, Relative Humidity; CO, Carbon monoxide; O3, Ozone; PM10, Particulate Matter smaller than about 10 micrometers; SO2, Sulfur dioxide; NO2, Nitrogen dioxide; CVD, Cardiovascular Diseases; avg, average; std, standard deviation; 25% 50% and 75% 25th 50th and 75th percentile respectively, min–max = range).*

L'analisi descrittiva (Tab. 9) rivela una serie di tendenze e variazioni chiave nei parametri ambientali considerati. Le temperature, rappresentate come Tmin, Tmean e Tmax, spaziano da valori sub-zero a oltre 41.60°C, sottolineando una notevole diversità termica. La Tmean si attesta su una media di 17.78°C con una deviazione standard di 6.40°C, rivelando una consistente variabilità durante il periodo di osservazione. Il punto di rugiada, Dewp, oscilla tra -6.38°C e 26.00°C, offrendo una panoramica significativa sull'umidità atmosferica. Parallelamente, la temperatura apparente, Tapp, evidenzia estremi tra 2.96°C e 52.78°C. La

pressione atmosferica, P_atm, mostra una fluttuazione moderata con una deviazione standard di 8.25 hPa, mentre l'umidità relativa, RH, indica un'atmosfera prevalentemente umida con una mediana del 71%. Concentrandosi sui gas atmosferici, il CO presenta una media di 0.77, mentre l'O3 ha una variazione sostanziale con una media di 83.51. Le sostanze SO2 e NO2 evidenziano ampie dispersioni attorno alle loro medie, rispettivamente 17.41 e 52.96. Il particolato PM10, cruciale per la valutazione della qualità dell'aria, ha una concentrazione media di 22.54. In relazione alla salute, le malattie cardiovascolari (CVD) variano in un range di 0-37, con una media di 12.90, sottolineando l'importanza di monitorare l'interazione tra condizioni ambientali e impatti sanitari.

## Metodologia

L'obiettivo principale dello studio è determinare quali variabili meteorologiche e di qualità dell'aria hanno il maggior impatto sulle ammissioni al pronto soccorso per le malattie cardiovascolari (CVD). Il primo passo è consiste nell'individuare le correlazioni tra le variabili meteorologiche e le ammissioni al pronto soccorso per le CVD attraverso un'analisi di correlazione, calcolando il coefficiente di Pearson "r" e il valore p. Successivamente, utilizzando il machine learning, verrà addestrato un modello tramite l'algoritmo Random Forest e calcolate le relative metriche di performance per valutare la capacità del modello di simulare i dati reali. Nel caso in cui l'analisi produca risultati accettabili (considerando le metriche MAE ed R2), sarà elaborata un'ulteriore analisi per determinare le variabili meteorologiche più significative. Nel caso in cui i risultati non saranno accettabili, sarà applicato un modello di decomposizione dei dati per estrarre la componente di trend. Successivamente, verrà ripetuta l'analisi di correlazione. Tale processo iterativo porterà a definire le variabili più significative per il modello di simulazione.

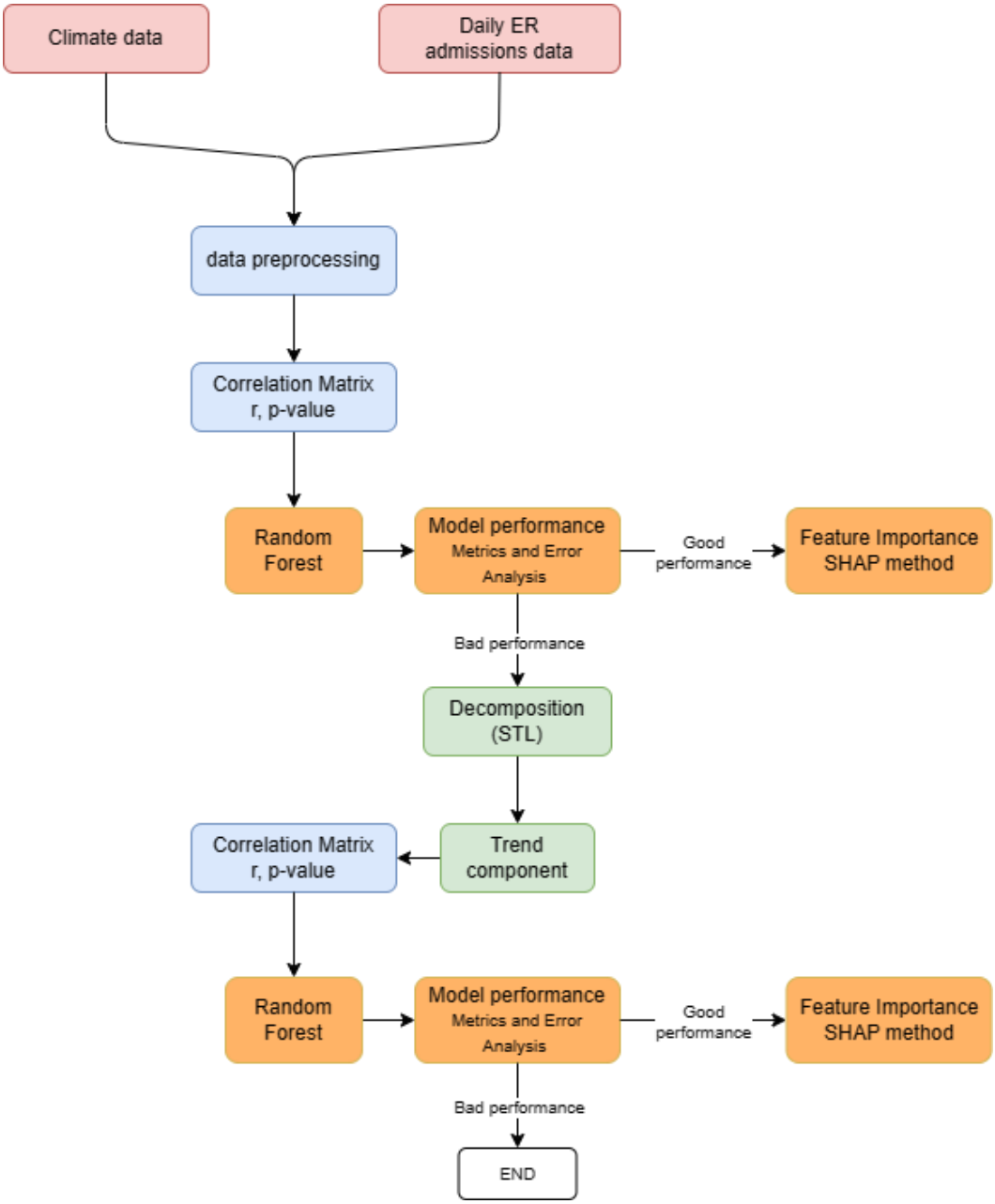


*Figura 44. Experimental design; ER, Emergency Room; r, Pearson correlation coefficient; SHAP, SHAP, SHapley Additive exPlanations; STL, Seasonal and Trend decomposition using Loess.*

### Aspetti Teorici dei modelli Random Forest

Il Random Forest è un algoritmo di tipo ensemble che si basa su alberi decisionali, strutture gerarchiche utilizzate per prendere decisioni sequenziali nell'assegnazione di punti dati a classi o nella previsione di valori continui in compiti di regressione. Combinando le previsioni di più modelli, il Random Forest ottiene previsioni accurate per la regressione. Ciascun modello nell'ensemble rappresenta un albero decisionale, e insieme sfruttano la conoscenza collettiva per migliorare la capacità predittiva complessiva. Per costruire gli alberi decisionali, l'algoritmo utilizza l'aggregazione bootstrap, nota anche come bagging. Questa tecnica coinvolge la creazione di molteplici campioni bootstrap dai dati di addestramento originali. Ciascun

campione viene ottenuto selezionando casualmente i punti dati con la possibilità di ripetizione, formando sottoinsiemi utilizzati per addestrare alberi decisionali individuali. Oltre al bagging, il Random Forest introduce casualità nella selezione delle features durante la costruzione dell'albero decisionale. Invece di considerare tutte le variabili in ogni nodo, viene scelto un sottoinsieme casuale di variabili per la divisione. Questo approccio garantisce che ciascun albero decisionale nell'ensemble apprenda e faccia previsioni basate su aspetti diversi dei dati, portando a un ensemble diversificato e robusto. Nella generazione di previsioni con un modello di regressione Random Forest, l'ensemble combina le previsioni dei singoli alberi decisionali. Un approccio comunemente utilizzato consiste nel calcolare la media dei valori previsti in tutti gli alberi. Questa tecnica di media aiuta a mitigare l'impatto degli outlier e del rumore, producendo previsioni più affidabili. Random Forest fornisce un meccanismo conveniente per stimare le prestazioni del modello senza richiedere un set di convalida separato. Questo meccanismo si basa su campioni out-of-bag (OOB), che sono punti dati lasciati fuori in ogni campione bootstrap. Confrontando le previsioni di questi campioni OOB con i loro valori target effettivi, è possibile valutare l'accuratezza del modello. Inoltre, Random Forest offre una panoramica sull'importanza relativa delle diverse variabili nei problemi di regressione. Analizzando l'effetto delle divisioni delle features nell'ensemble di alberi decisionali, il modello calcola punteggi di importanza delle variabili. Questi punteggi indicano quali features contribuiscono di più alla capacità predittiva del modello, fornendo interpretazione e comprensione delle relazioni sottostanti nei dati. In breve, ci sono vari vantaggi teorici associati all'uso delle tecniche Random Forest per i modelli di regressione. Questi includono l'apprendimento ensemble utilizzando gli alberi decisionali, il bagging per una costruzione robusta degli alberi, la selezione casuale delle features per garantire la diversità e la capacità di stimare le prestazioni del modello e valutare l'importanza delle variabili. Collettivamente, questi aspetti contribuiscono all'efficacia complessiva, all'accuratezza e all'interpretabilità di Random Forest come potente tecnica di modellizzazione per la regressione.

I principali vantaggi che hanno portato alla scelta dell'algoritmo Random Forest per lo sviluppo del modello sono:

1. **Riduzione del rischio di sovra-adattamento (overfitting**): Gli alberi decisionali possono sovrastimare le informazioni, adattandosi eccessivamente ai campioni di addestramento. Tuttavia, grazie alla presenza di numerosi alberi nella Random Forest, il classificatore non tende a sovra-adattarsi, poiché la mediazione di alberi non correlati riduce la varianza complessiva e l'errore di previsione.

2. **Flessibilità:** L'algoritmo è versatile, poiché può gestire compiti sia di classificazione che di regressione con alta precisione. La selezione casuale delle features rende inoltre efficace la Random Forest nella stima di valori mancanti, mantenendo l'accuratezza anche in presenza di dati incompleti.
3. **Facilità nella determinazione dell'importanza delle variabili:** Con la Random Forest, è semplice valutare l'importanza o il contributo di ogni variabile al modello. Esistono diversi metodi per questo, come l'importanza di Gini e la diminuzione media dell'impurità. Tuttavia, un altro indicatore è la diminuzione media dell'accuratezza, che misura quanto l'accuratezza del modello diminuisce quando i valori di una specifica caratteristica vengono modificati casualmente.

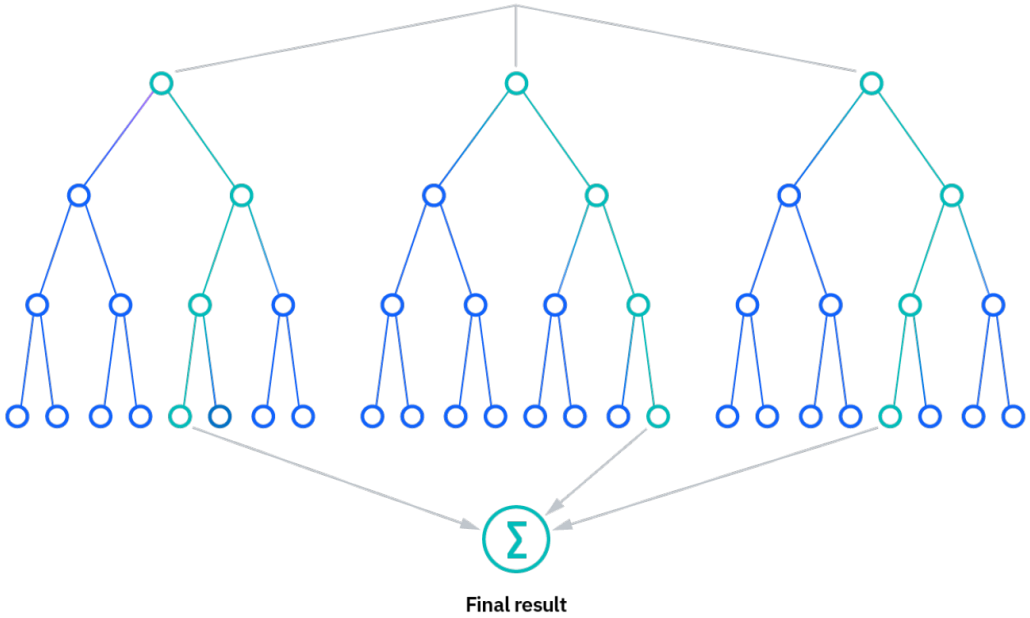


*Figura 45. Struttura di un algoritmo Random Forest [https://www.ibm.com/topics/random-forest]*

### Seasonal Trend Decomposition using LOESS

La tecnica Seasonal Trend decomposition using LOESS (STL) è un approccio ampiamente utilizzato nell'analisi delle serie temporali per scomporre una serie temporale in componenti distinte al fine di comprenderne meglio la struttura sottostante e identificare pattern o tendenze rilevanti. Questa tecnica si basa su due principi fondamentali:

1. **Decomposizione Additiva:** La tecnica STL decompone una serie temporale in tre componenti principali - stagionale, di tendenza e residua - attraverso un processo di decomposizione additiva. Questo significa che la serie temporale originale viene suddivisa nella somma delle tre componenti, ciascuna delle quali offre informazioni specifiche.

2. **Utilizzo del LOESS:** La componente di tendenza e la componente stagionale vengono stimate utilizzando il metodo LOESS (Local Regression). Il LOESS è un metodo di regressione non parametrica che calcola una media pesata dei punti dati circostanti per stimare il comportamento locale della serie temporale. Questo consente di rilevare le tendenze locali e le variazioni stagionali, adattandosi alla complessità dei dati.

L'approccio Seasonal Trend decomposition using LOESS (STL) è basato sull'equazione fondamentale:

$$Y_v = T_v + S_v + R_v \quad (1)$$

Dove:

- $Y_v$ rappresenta il valore osservato della serie temporale al momento v.
- $T_v$ è la componente di tendenza al momento v, che rappresenta l'andamento generale della serie temporale.
- $S_v$ è la componente stagionale al momento v, che cattura le variazioni cicliche o regolari a intervalli regolari nel tempo.
- $R_v$ è la componente residua al momento v, che rappresenta la variazione non spiegata dalle componenti di tendenza e stagionale.

Più in particolare:

1. **la componente stagionale** cattura le fluttuazioni cicliche e regolari all'interno dei dati, consentendo di identificare modelli ricorrenti che si ripetono a intervalli regolari. Questa componente è fondamentale per comprendere le variazioni stagionali in fenomeni come vendite al dettaglio, dati climatici o dati finanziari con stagionalità evidente.
2. **La componente di tendenza** rappresenta l'andamento generale della serie temporale nel lungo termine, evidenziando le direzioni di crescita o decrescita. Questa componente è utile per identificare trend a lungo termine che possono influenzare il comportamento dei dati nel tempo.
3. **La componente residua** rappresenta le variazioni non spiegate dalle componenti di tendenza e stagionale. Questa componente può contenere informazioni su eventi eccezionali, anomalie o variabilità non prevista nei dati.

Questa scomposizione additiva consente di analizzare separatamente le diverse influenze che contribuiscono ai dati della serie temporale, rendendo più chiara l'interpretazione delle variazioni nel tempo. Nel contesto dell'analisi delle serie temporali, l'uso di STL fornisce una struttura chiara per esaminare le variazioni nei dati, separando le influenze stagionali, le tendenze di fondo

e le deviazioni impreviste. Ciò facilita la comprensione delle dinamiche sottostanti e può essere prezioso per scopi predittivi, analisi dei dati e formulazione di ipotesi di ricerca.

### SHAP feature importance

La tecnica SHAP (SHapley Additive exPlanations) Feature Importance rappresenta un approccio fondamentale all'interno del panorama delle tecniche di Explainable Artificial Intelligence (XAI) [80]. Questo metodo si basa sui valori SHAP, che costituiscono un'approccio esplicativo model-centrico in cui ciascuna previsione del modello è spiegata attraverso il contributo delle feature presenti nel dataset rispetto all'output del modello. Più specificamente, SHAP si avvale della concettualizzazione dei valori Shapley, derivata dalla teoria dei giochi [81], e che rappresenta la soluzione al problema computazionale di calcolare il contributo di ciascun sottoinsieme di feature all'interno di un dataset con m feature, per predire il risultato di un modello. Calcolare in modo esatto i valori Shapley risulterebbe impraticabile a causa della natura esponenziale del problema. Tuttavia, SHAP approssima questa soluzione attraverso una speciale regressione lineare pesata [80], adattabile a qualsiasi tipo di modello, oppure mediante diverse assunzioni sulla dipendenza tra feature in modelli di alberi di decisione ensemble [82]. Nei modelli di regressione lineare, i coefficienti utilizzati per ponderare le feature sono utilizzati per spiegare le previsioni per tutti i punti dati, ma non tengono conto dell'eterogeneità delle singole osservazioni. Spesso, l'effetto di una feature su un punto dati può differire da un altro punto dati, il che è coerente con il fatto che le spiegazioni locali risultano più accurate di quelle globali. Questo approccio richiama l'idea di approssimare similitudini globali attraverso una serie di similitudini locali, come fatto dalle tecniche di riduzione della dimensionalità non lineari. SHAP sfrutta la proprietà dell'esplicabilità locale per costruire modelli surrogati per modelli di machine learning di tipo "black-box". In questo caso, SHAP apporta piccole variazioni agli input e testa le variazioni nelle previsioni: se la previsione del modello non cambia significativamente, allora la feature per quel particolare punto dati potrebbe non essere un predittore importante. La somma dei contributi, o valori SHAP, di ciascuna feature, è uguale alla previsione finale. Un valore SHAP non rappresenta la differenza tra la previsione con e senza una feature, ma è il contributo di una feature alla differenza tra la previsione effettiva e la previsione media. In questo modo, la tecnica SHAP si configura come un potente strumento che sfrutta i valori Shapley e l'esplicabilità locale per consentire una più profonda comprensione dei modelli di machine learning, promuovendo la trasparenza e la fiducia nelle applicazioni dell'Intelligenza Artificiale.

## Preprocessing dei Dati

Nella fase iniziale, sono state eliminati i record con dati mancanti dal database per ottenere un database completo ed utilizzabile ai fini dell'addestramento con le tecniche di machine learning. Successivamente, sono stati analizzati i trend degli accessi ospedalieri tra il 2013 e il 2021 per identificare possibili interferenze dovute a variazioni anomale nelle visite al Pronto Soccorso durante gli anni pandemici del 2020 e 2021 dovuti alle infezioni da Covid-19. L'esclusione dei dati del 2020 e 2021 è stata determinata da un'analisi che ha evidenziato marcate differenze nelle visite al Pronto Soccorso in questi anni, come rappresentato in Figura 46.

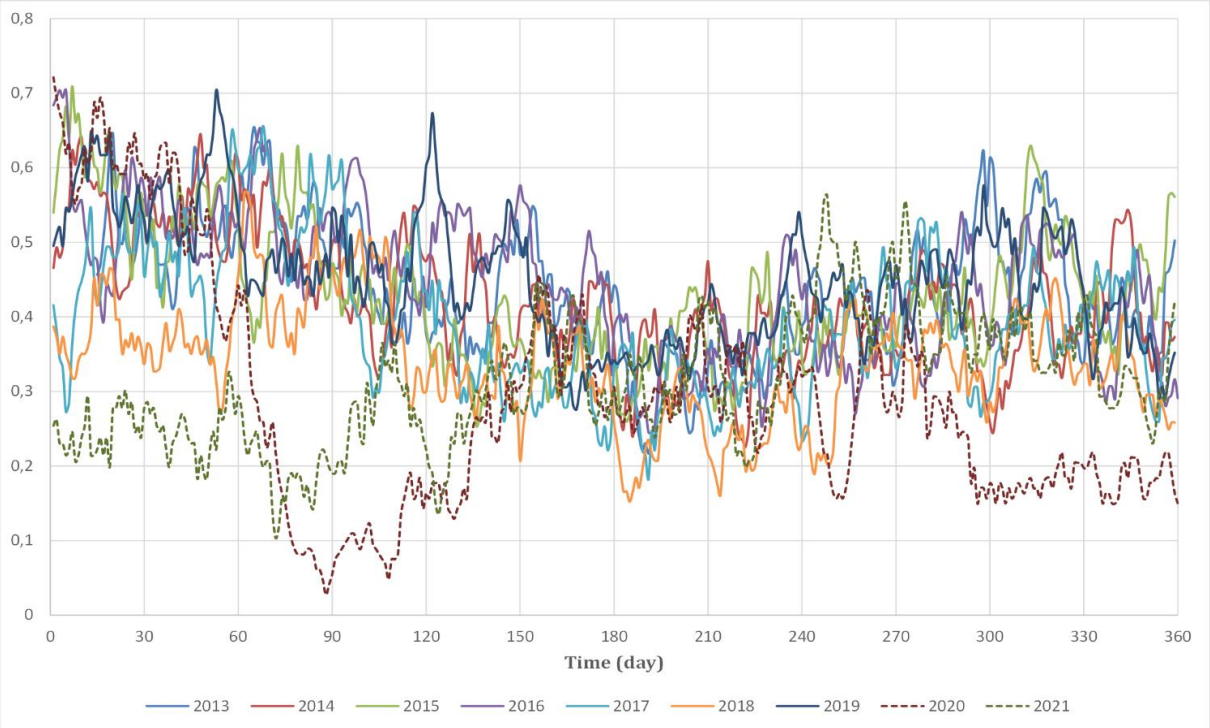


*Figura 46. Medie mobili su 7 giorni degli accessi al Pronto Soccorso, anno per anno, dal 2013 al 2021*

La figura (Fig. 46) mostra le medie mobili su 7 giorni, normalizzate attraverso il metodo min/max, per assicurare una comparabilità tra i diversi anni. Pertanto, sono stati considerati solo i dati relativi agli anni dal 2013 al 2019 per le ulteriori analisi.

## Creazione del modello

### Analisi di correlazione

L'analisi di correlazione è una tecnica statistica bivariata che valuta la forza della relazione lineare tra due variabili e quantifica questa relazione. Per esprimere quantitativamente l'intensità del legame tra due variabili, è essenziale calcolare un coefficiente di correlazione. Esistono diversi tipi di coefficienti di correlazione, ma tutti presentano alcune caratteristiche comuni, come il valore che oscilla tra -1 e +1, rappresentando una relazione perfetta tra le variabili. Un valore di

0 indica invece l'assenza di una relazione. Per condurre questa analisi, è stata utilizzata una matrice di correlazione. Tale matrice è una tabella in cui ogni cella mostra la correlazione tra due variabili. La matrice calcola il coefficiente di correlazione di Pearson "r", uno dei coefficienti di correlazione più utilizzati, per ogni coppia di variabili.

L'analisi di correlazione è stata effettuata utilizzando il valore p come strumento per testare la significatività dell'ipotesi nulla. Il valore p rappresenta la probabilità di ottenere un determinato insieme di valori osservati supponendo che l'ipotesi nulla sia vera, indicando la correttezza della nostra affermazione con un errore minimo. Un valore p significativamente basso indica che un risultato osservato estremo sarebbe altamente improbabile secondo l'ipotesi nulla. Questa misura statistica stabilisce l'affidabilità dei valori di correlazione ottenuti. Le ipotesi accettabili per l'insieme di variabili di input sono state definite come quelle aventi valori p inferiori a 0,01 e valori r maggiori o uguali a 0,45. Nel contesto considerato, nessuna delle variabili di input del modello soddisfa le condizioni sopra menzionate, come dimostrato dalla matrice di correlazione (Tab. 10). Pertanto, sarà utilizzato il modello di decomposizione.

*Tabella 9. Analisi di correlazione tra features e CVD*

| Features | r |
|---|---|
| CO | 0.28 |
| P_atm | 0.26 |
| rh | 0.21 |
| pm10 | 0.041 |
| NO2 | -0.021 |
| Tdewp | -0.16 |
| Tmax | -0.22 |
| o3 | -0.22 |
| Tmean | -0.23 |
| Tmin | -0.25 |

## Modello di decomposizione

Molti metodi di previsione si basano sull'idea che, se esiste un pattern sistematico, esso può essere identificato e separato da qualsiasi fluttuazione casuale attraverso metodi di smoothing delle serie storiche. L'effetto smoothing consiste nella rimozione di disturbi casuali; una volta conosciuto il modello, questo può essere proiettato nel futuro a scopi di previsione. I modelli di decomposizione sono principalmente suddivisi in modelli di serie temporali additivi e modelli di serie temporali moltiplicativi. Il modello additivo presuppone che gli effetti di ciascun componente siano indipendenti l'uno dall'altro e che ciascun componente sia espresso in termini assoluti. Un modello additivo è appropriato quando l'ampiezza dell'oscillazione stagionale non

varia con il livello della serie. L'errore può assumere valori positivi o negativi, mentre il valore neutro è espresso con il valore 0, il che significa che non influisce sulla serie. Il modello moltiplicativo presuppone che gli effetti di ciascun componente sull'evoluzione del fenomeno siano interrelati sulla base della grandezza assoluta del componente di tendenza e che gli altri componenti siano espressi proporzionalmente. Un modello moltiplicativo è adatto quando la fluttuazione stagionale varia, aumenta o diminuisce proporzionalmente con la variazione del livello della serie. L'errore può assumere solo valori non negativi ed ha un valore neutro di 1. STL è un metodo statistico che permette la decomposizione delle serie temporali in tre componenti: tendenza, stagionalità e residuo. La componente di tendenza riflette la variazione a lungo termine della serie. Esiste una tendenza quando vi è una direzione persistentemente crescente o decrescente nei dati. La componente stagionale riflette la stagionalità, cioè la variazione dei dati che si verifica a specifici intervalli regolari di un anno o meno. La componente residuale descrive influenze casuali o irregolari. Rappresenta i residui della serie dopo che le altre componenti sono state rimosse [83].

Siano $Y_v$, $T_v$, $S_v$ e $R_v$, per v = 1 a n: la componente di trend, la componente stagionale e la componente residuuale (o rumore) dei dati della serie storia da analizzare.

$$Y_v = T_v + S_v + R_v \quad (1)$$

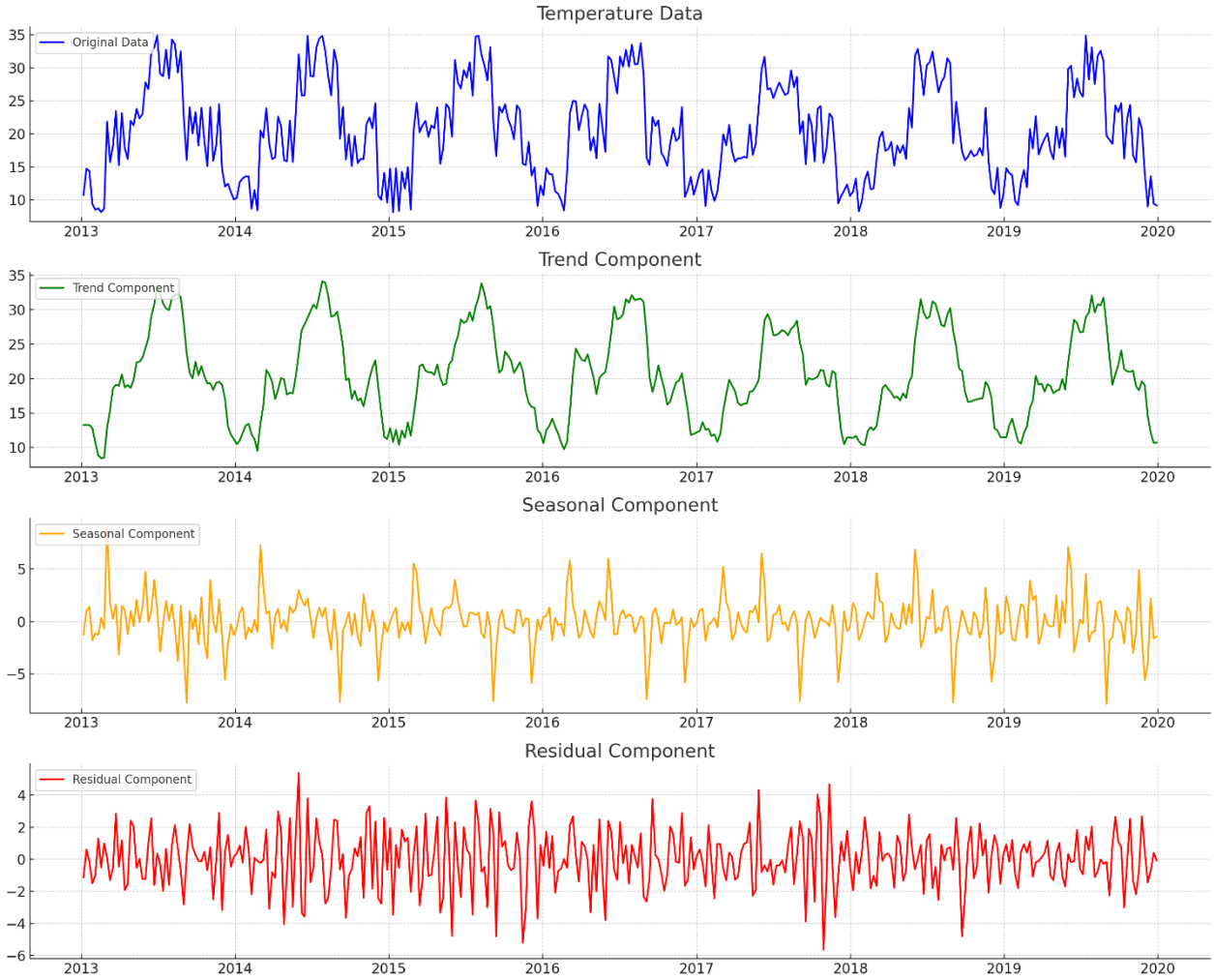


*Figura 47. Esempio di decomposizione STL della temperatura media giornaliera dal 2013 al 2021 nelle tre componenti: trend, stagionalità e residuale*

Nell'analisi il modello di decomposizione STL è stato applicato ai dati originali dopo aver verificato la stazionarietà della serie utilizzando il metodo Dickey-Fuller [84]. Il test di Dickey-Fuller è un metodo statistico impiegato per determinare se una serie temporale sia stazionaria o meno, ossia se le sue proprietà statistiche rimangano invariate nel tempo. Una serie temporale è stazionaria se le sue proprietà statistiche (media, varianza, covarianza) non cambiano nel tempo. Il test di Dickey-Fuller si basa sull'idea che una serie temporale non stazionaria possa essere trasformata in una serie stazionaria attraverso la differenziazione. Quindi, il test confronta la serie originale con la sua differenza lag-1 per determinare se esiste una relazione lineare tra i valori successivi della serie.

In seguito, è stata applicata la metodologia STL all'intera serie di dati, comprendente sia le variabili ambientali sia il numero di accessi al pronto soccorso per CVD. Questo procedimento ha portato alla creazione di un nuovo database, nel quale i valori originali sono stati rielaborati per enfatizzare unicamente la componente di trend rilevata dall'analisi STL. Attraverso questa

trasformazione, sono state rimosse sia la componente stagionale sia le irregolarità (o 'rumore') dai dati. Il risultato è una versione più uniforme dei dati originali, che ha prodotto una rappresentazione più chiara e diretta dell'andamento storico dei dati, depurato dalle distorsioni stagionali e dalle fluttuazioni casuali. Tale raffinata elaborazione dei dati fornisce una visione più accurata dei pattern emergenti, fondamentale per un'analisi approfondita e per dedurre conclusioni più precise. Ricalcolando poi la matrice di correlazione, si sono ottenuti valori del coefficiente di Pearson significativamente maggiori rispetto alla serie di dati originali (Tab. 11). Questi risultati permettono l'applicazione di modelli di apprendimento automatico per produrre un modello di simulazione efficace degli accessi ospedalieri nel tempo.

*Tabella 10. Analisi di correlazione tra features e CVD dopo l'applicazione del modello di decomposizione*

| Features | r |
|---|---|
| rh | 0.5 |
| o3 | 0.48 |
| pm10 | 0.076 |
| NO2 | -0.22 |
| Tmean | -0.28 |
| Tmin | -0.36 |
| P_atm | -0.38 |
| Tdewp | -0.38 |
| CO | -0.41 |
| Tmax | -0.5 |

## Applicazione del modello di Machine Learning

Il Random Forest è l'algoritmo di intelligenza artificiale utilizzato per creare il modello di simulazione del trend di accessi al pronto soccorso per CVD e per stimare le variabili meteorologiche di maggiore influenza. In seguito alla fase di training, la robustezza del modello è stata validata attraverso la tecnica della convalida incrociata. Tra le diverse applicazioni di convalida incrociata, in questo lavoro, è stata utilizzata la tecnica K-fold cross validation. Il K-fold divide i dati in k differenti sottoinsiemi e k-1 vengono utilizzati per la fase di addestramento e l'ultimo sottoinsieme rimanente per la fase di test. L'errore viene poi calcolato sulle osservazioni dei sottoinsiemi esclusi. Questa procedura viene ripetuta k volte scegliendo un diverso sottoinsieme e ottenendo k stime dell'errore di test. La stima finale sarà una media di questi valori [85].

## Analisi delle performance del modello

La Tabella 12 riporta i valori delle metriche MAE (Mean Absolute Error) e R2 per il campione estratto durante le fasi di addestramento e test, nonché le stesse metriche nel valore medio dei sottoinsiemi di convalida incrociata.

*Tabella 11. Performance del modello calcolate usando le metriche MAE (Mean Absolute Error) ed R2. La valutazione delle performance è stata effettuata sui dati del modello e sui dati della valutazione incrociata per effettuarne una comparazione*

| | **R2 train** | **R2 test** | **MAE train** | **MAE test** |
|---|---|---|---|---|
| **CVD with STL** | 0.996 | 0.969 | 0.13 | 0.36 |
| **cross validation (mean values)** | 0.996 | 0.973 | 0.12 | 0.33 |

I valori ottenuti attraverso la convalida incrociata sono analoghi a quelli determinati per il campione casuale, evidenziando la capacità del modello di evitare fenomeni di overfitting, dimostrando così la precisione del modello. Successivamente, nella Fig. 48 è mostrato come il trend dei dati relativi alle patologie cardiovascolari predetti dal modello Random Forest, descritto in blu, presenti un andamento identico ai dati effettivi.

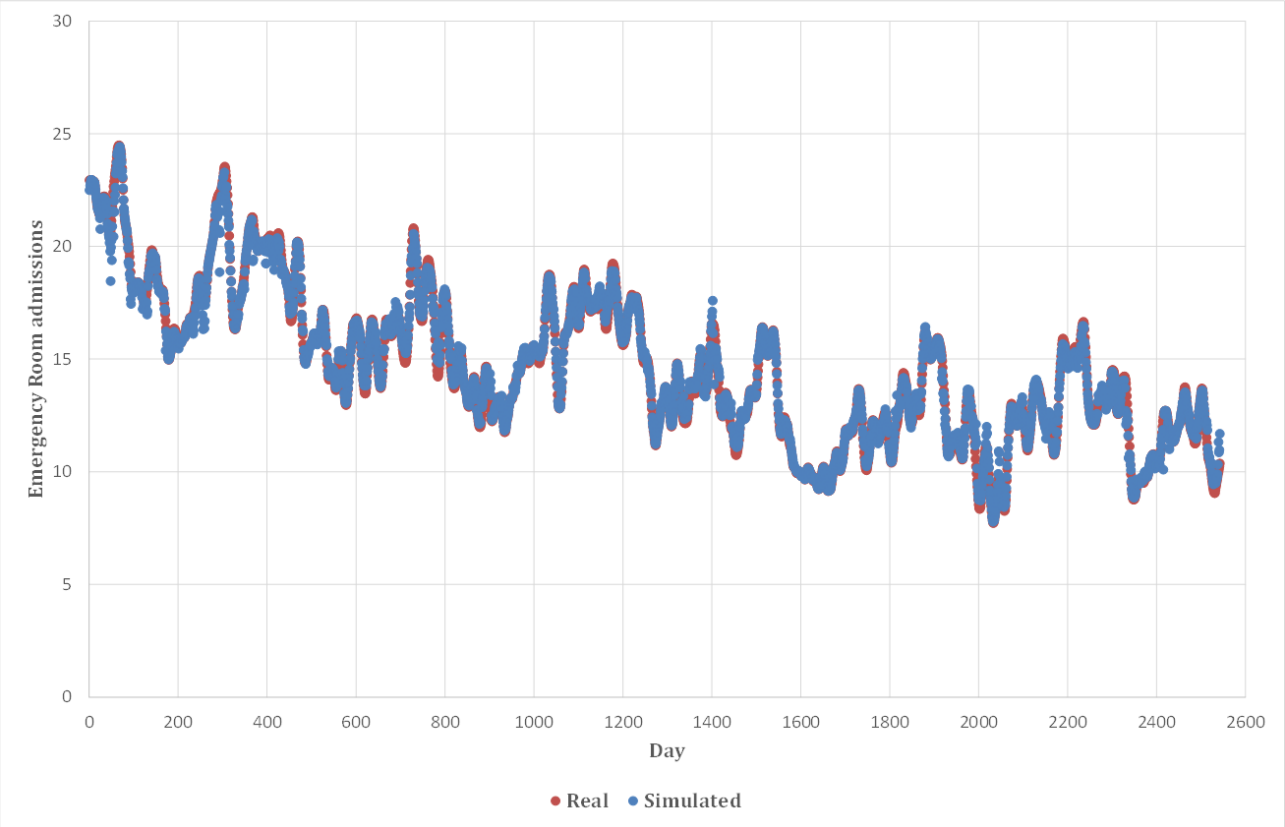


*Figura 48. Confronto tra dati reali e simulati ottenuti dal modello, per l'intera serie storica*

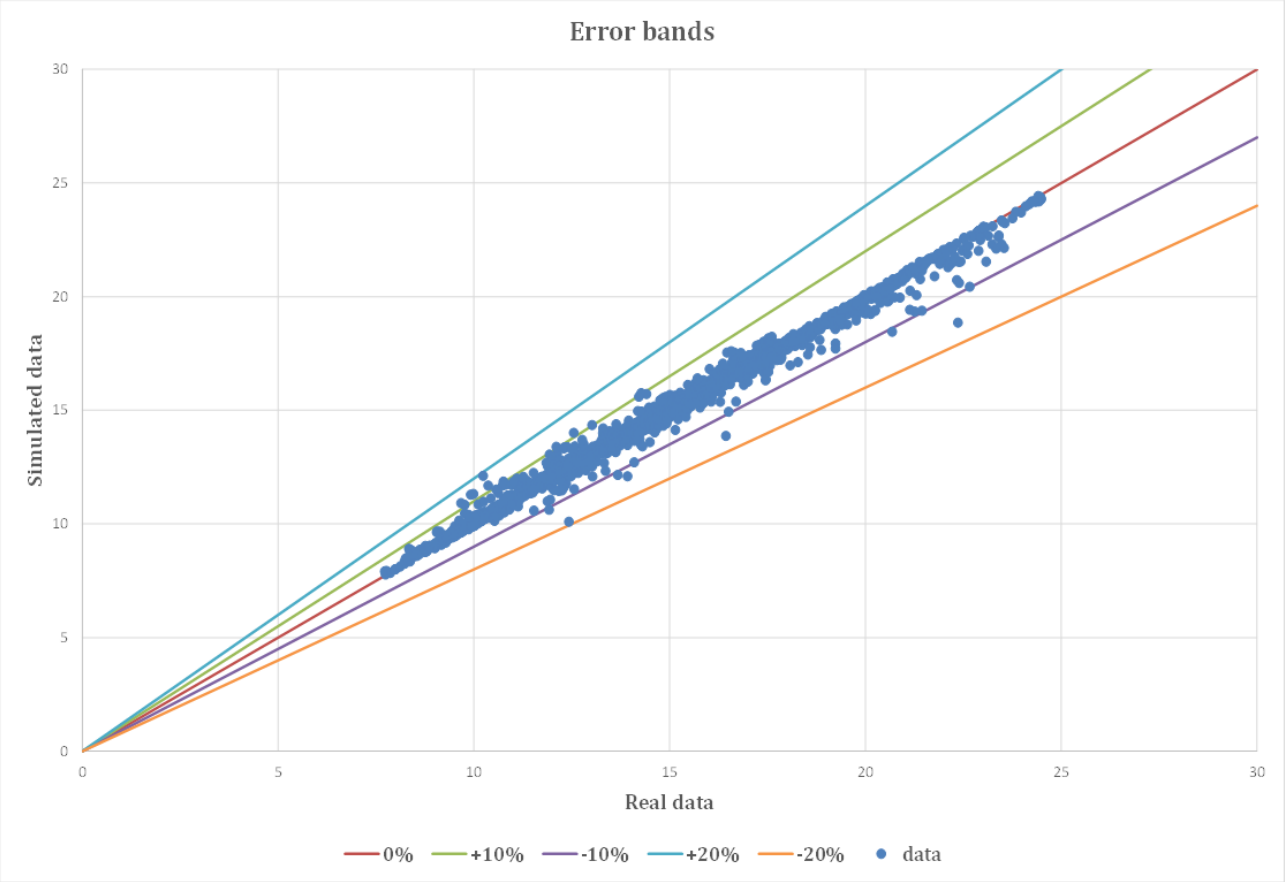


*Figura 49. Analisi dell'errore tra dati reali e simulati*

La precisione del modello è evidenziata nella Figura 49. I punti blu rappresentano sia i dati reali che simulati, e per la maggior parte, questi rientrano all'interno delle fasce di errore (Fig. 49). La maggior parte dei punti dati rientra entro il ± 10% delle fasce di errore e solo pochi valori sono compresi entro il ± 20% delle fasce di errore. In particolare, 2443 dei 2542 punti dati (numero totale dei punti dati, ovvero il 96,10% del totale delle ammissioni al pronto soccorso), rientrano nella fascia di errore del ± 5% (Tabella 13).

*Tabella 12.Data number between error bands*

| **Error X (%)** | **Data numbers and relative %** |
|---|---|
| X <–20 | 1 (0.04) |
| –20 < X < –15 | 3 (0.12) |
| –15 < X < –10 | 7 (0.28) |
| –10 < X < –5 | 34 (1.34) |
| –5 < X < 0 | 1189 (46.77) |
| 0 < X < +5 | 1254 (49.33) |
| +5 < X < +10 | 48 (1.89) |
| +10 < X < +15 | 5 (0.20) |

| | |
|---|---|
| +15 < X < +20 | 1 (0.04) |
| X >+20 | 0 |

# Risultati

Dal momento che il modello di simulazione basato su Machine Learning ha dimostrato di fornire risultati accurati, possiamo procedere con l'utilizzo delle tecniche di feature importance attraverso il metodo SHAP per identificare le variabili più influenti nel determinare il trend delle ammissioni giornaliere in ospedale per CVD, secondo il diagramma a blocchi mostrato nella metodologia (Fig. 44). La Feature Importance (Fig. 50) mostra come le variabili più caratteristiche per i CVD siano la pressione atmosferica con una percentuale del 37%, la temperatura minima con il 19% e il monossido di carbonio (CO) con una percentuale del 18%. Queste tre variabili rappresentano circa il 74% della prevedibilità del modello, con la pressione atmosferica che risulta predominante

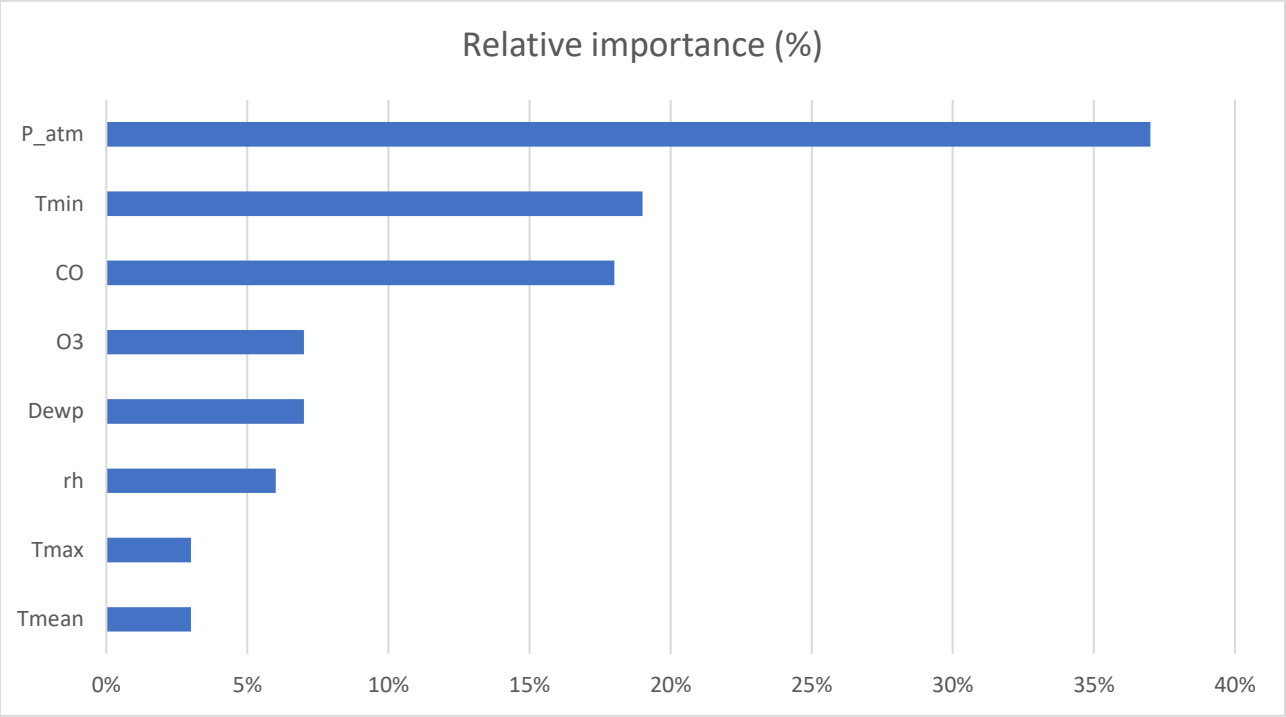


*Figura 50. Feature importance per CVD*

Il modello SHAP ha integrato e analizzato set di dati cumulativi derivati dai modelli precedenti in questo studio. Questi dati sono stati visualizzati tramite diagrammi bee swarm, come illustrato nella (Fig. 51). Le features sono classificate in base al loro peso contributivo alla previsione, tenendo conto dell'intervallo e del valore (cioè, ±) nel modello. Il metodo SHAP fornisce intuizioni sia sulle contribuzioni globali sia locali di ogni variabile, e i risultati sono presentati attraverso una classifica d'importanza visualizzata tramite un diagramma a dispersione bee

swarm. In questa rappresentazione, l'asse orizzontale rappresenta il valore SHAP, mentre il colore del punto indica l'intensità dei valori (blu per valori bassi, rosso per valori alti) di ogni variabile. Il modello SHAP ha identificato le variabili i con l'impatto più pronunciato sulle ammissioni ospedaliere. Queste variabili sono organizzate in ordine decrescente d'importanza come segue: Pressione Atmosferica (P_atm), Temperatura Minima (Tmin) e concentrazione di Monossido di Carbonio (CO). La correlazione tra valori bassi di P_atm e Tmin e la loro corrispondenza con valori SHAP elevati (indicando significatività) è facilmente intuibile. Questa rappresentazione grafica sottolinea il loro ruolo sostanziale nel potenziare l'efficacia del modello di simulazione. Al contrario, il comportamento del CO mostra il pattern opposto, aumentando di importanza per valori più alti.

Le variabili rimanenti si sono rivelate avere una significatività minore. Queste includono: Punto di Rugiada (Tdewp), Umidità Relativa (rh), Ozono (O3) e Temperatura Media (Tmean). D'altra parte, Tmax fornisce un contributo praticamente trascurabile, indipendentemente dai suoi valori. Questa analisi complessiva permette una comprensione degli impatti di ogni singola variabile sugli accessi ospedalieri per CVD.

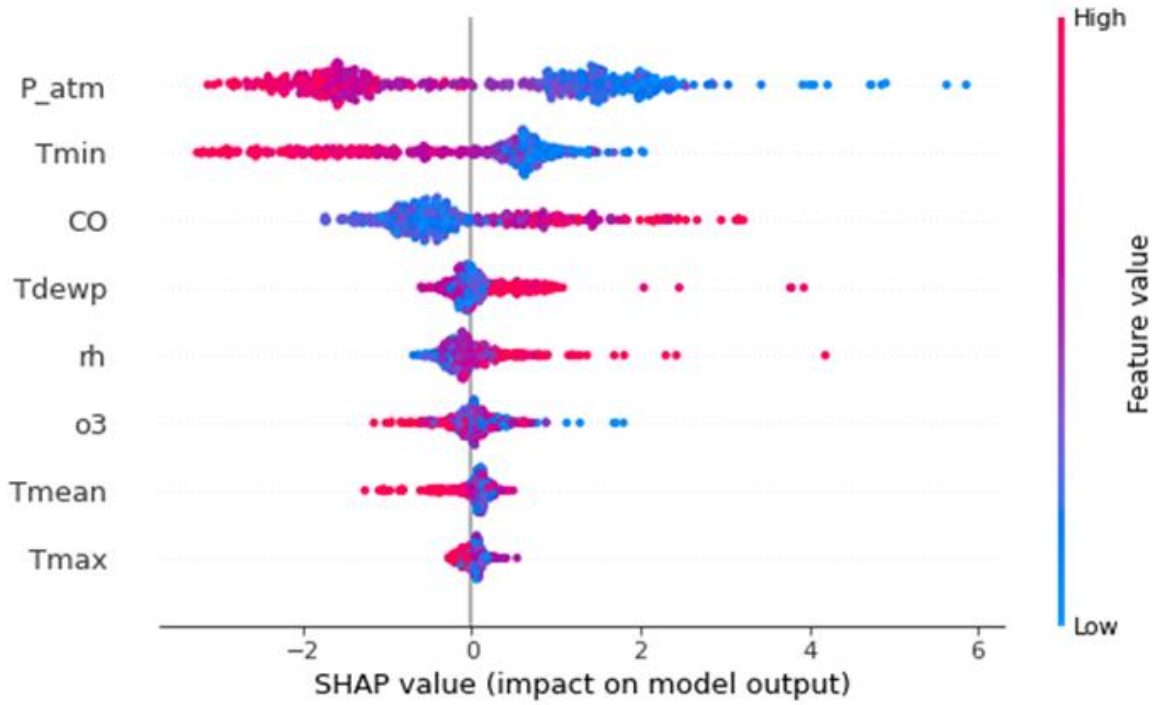


*Figura 51. Analisi dei valori di SHAP per i parametri ambientali utilizzati come predittori nell'addestramento del modello*

# Conclusioni e prospettive future

L'obiettivo di questo studio era esplorare le correlazioni tra fattori meteorologici e qualità dell'aria con le ammissioni al pronto soccorso per CVD. Per raggiungere questo obiettivo, è stato utilizzato il modello AI Random Forest per evidenziare le features più importanti. I dati iniziali ottenuti dalle matrici di correlazione hanno fornito risultati insoddisfacenti. Di conseguenza, è stato utilizzato un modello di decomposizione STL per ridurre l'errore nella previsione degli accessi ospedalieri. L'utilizzo di questo modello ha notevolmente migliorato i risultati quando i dati sono stati sottoposti a ulteriori analisi tramite il modello Random Forest. Abbiamo determinato che sia i dati reali sia quelli predetti dell'intervallo di errore cadevano principalmente all'interno del range del ±10%, ad eccezione di pochi valori che superavano leggermente il margine del ±20%. Quest'ultimi valori, analizzati attraverso la procedura successiva, hanno mostrato un'accuratezza notevolmente alta, sottolineando così l'importanza di specifiche variabili nella generazione delle previsioni. Le variabili più significative per le CVD sono state identificate come la pressione atmosferica (che contribuisce per oltre il 37%), la temperatura minima (19%) e il monossido di carbonio (CO) (18%). Queste tre variabili rappresentano collettivamente il 74% della prevedibilità del modello, con la pressione atmosferica che gioca un ruolo preminente.

## Ruolo della Temperatura, della Pressione Atmosferica e della Concentrazione di Monossido di Carbonio nella Simulazione delle Ammissioni per CVD

I risultati di questo studio offrono preziose intuizioni sulla relazione tra fattori ambientali e salute cardiovascolare. Le nostre scoperte suggeriscono che la pressione atmosferica, la temperatura minima e il monossido di carbonio siano i fattori più rilevanti nella previsione del numero di ricoveri ospedalieri per CVD. Questi risultati sono coerenti con studi analizzati in precedenzai che hanno identificato questi fattori come fattori di rischio per le CVD. Studi precedenti hanno dimostrato in modo coerente che le temperature estreme aumentano la mortalità da CVD. Infatti, l'esposizione cronica al freddo o al caldo può compromettere la funzione cardiovascolare, aumentando di conseguenza il rischio di infarto, aritmie, malattie tromboemboliche e sepsi indotta dal calore [78]. Inoltre, le variazioni della temperatura ambiente sono state mostrate come contribuenti alla mortalità cardiovascolare causando un aumento della pressione sanguigna, della viscosità del sangue e della frequenza cardiaca [78]. Le variazioni stagionali nelle CVD rappresentano una sfida sanitaria significativa [56], in particolare per le popolazioni che risiedono in climi più miti, che potrebbero essere meno adattate o preparate a estreme variazioni climatiche. La maggior parte degli studi ha riportato un aumento

invernale di anomalie cardiovascolari e morte cardiaca nell'emisfero settentrionale, dove le temperature sono estremamente fredde [44] [57]. Alla luce di queste prove, è evidente che la temperatura svolge un ruolo fondamentale nella prevalenza delle CVD e dovrebbe essere considerata un parametro vitale nello sviluppo di modelli predittivi per i ricoveri ospedalieri legati a queste condizioni.

I nostri risultati non solo enfatizzano l'importanza della temperatura, ma evidenziano anche il rilievo della pressione atmosferica e della concentrazione di monossido di carbonio nella previsione dei ricoveri ospedalieri per CVD. Queste scoperte sono coerenti con la letteratura esistente. È stata dimostrata una connessione tra piccoli cambiamenti in bassi livelli di concentrazioni ambientali di monossido di carbonio e ricoveri ospedalieri cardiovascolari [86]. L'esposizione al monossido di carbonio a concentrazioni riscontrate nei forti fumatori o negli individui con significativa esposizione occupazionale è stata mostrata come un fattore nel contribuire alla patogenesi delle CVD [87]. Una breve esposizione al monossido di carbonio ambientale è stata associata a un aumento del rischio di ricoveri ospedalieri per CVD [86], nonché a tassi aumentati di malattie cardiovascolari e coronariche nelle principali città cinesi [88]. I livelli ambientali di monossido di carbonio sono stati anche positivamente correlati ai ricoveri ospedalieri per CVD [89]. Inoltre, si è dimostrato che l'esposizione al monossido di carbonio ha effetti negativi sulla performance fisica in soggetti con malattia delle arterie coronarie, evidenziando ulteriormente il suo impatto sull'ischemia miocardica [90]. Oltre al monossido di carbonio, alte pressioni atmosferiche sono state associate a un aumento del numero di ricoveri per arresto cardiaco [91]. Queste evidenze suggeriscono che sia la pressione atmosferica che la concentrazione di monossido di carbonio sono fattori cruciali da considerare nello sviluppo di modelli predittivi per i ricoveri ospedalieri legati alle CVD. Alla luce della crescente mole di ricerche, è fondamentale incorporare temperatura, pressione atmosferica e concentrazione di monossido di carbonio nei modelli predittivi per i ricoveri dovuti alle CVD, poiché questi fattori giocano un ruolo significativo nell'insorgenza e nell'esacerbazione delle condizioni cardiovascolari.

## Creazione di un Framework di assistenza ai servizi sanitari per la cura delle CVD basato sull'analisi dei parametri ambientali

La recente letteratura scientifica ha offerto avanzamenti significativi nella comprensione e mitigazione delle ripercussioni dei cambiamenti climatici sulla patologia cardiovascolare (CVD). In particolare, lo studio [92] presenta un framework strutturato per la proposta personalizzata di interventi sanitari basato su un modello di profilo utente avanzato. Tale strategia potrebbe

aprire la strada a interventi mirati nel campo delle CVD, guidati da dati relativi alle condizioni meteorologiche e alla qualità dell'aria. Parallelamente, alcune ricerche [93] [94] hanno introdotto metodologie basate su reti neurali convoluzionali e modelli di classificazione ResNet per identificare patologie, ad esempio il COVID-19, attraverso l'analisi di immagini radiologiche. Si ipotizza che simili approcci possano essere modulati per la diagnosi precoce delle CVD, correlando specifiche manifestazioni radiologiche. Inoltre, l'indagine condotta da [95] ha delineato una metodologia rivolta alla comprensione delle dinamiche causali tra patologie e sintomi. Questo framework potrebbe svelare connessioni causali tra variabili meteorologiche, qualità dell'aria e CVD. Complessivamente, questi studi evidenziano la necessità di una comprensione integrata delle relazioni tra fattori ambientali e salute cardiovascolare.

Le future direzioni di ricerca in questo ambito prevedono la definizione di soglie d'allerta basate su parametri ambientali. Si auspica una progressiva espansione del modello per incorporare un più ampio spettro di patologie e per considerare diversi contesti geografici. L'ottimizzazione della capacità predittiva del modello richiederà rigorose procedure di convalida e potenziali revisioni. Una delle strategie contempla l'analisi di dati raccolti da diverse strutture ospedaliere distribuite in vari contesti geografici e climatici. Un obiettivo chiave è l'accesso al database completo degli ospedali della regione Puglia per testare la generalizzabilità del modello, consolidandone l'affidabilità e l'ambito di applicazione. In conclusione, esiste una forte relazione tra le ammissioni giornaliere al pronto soccorso e la variazione di alcuni parametri climatici e di qualità dell'aria. I futuri studi dovrebbero essere indirizzati verso l'identificazione e la definizione di soglie di allerta basate su questi parametri.

# Conclusioni generali

La convergenza delle discipline presentate in questa tesi mette in evidenza l'importanza fondamentale di un approccio multidisciplinare per affrontare le sfide emergenti del XXI secolo. Mentre l'impulso potrebbe inizialmente sembrare focalizzato sull'interazione tra Machine Learning e le varie discipline, emerge chiaramente che la vera protagonista è l'intersezione tra cambiamenti climatici, idrologia e salute umana.

L'analisi dell'influenza di variabili climatiche sulle patologie cardiovascolari ha rivelato l'entità e la profondità dell'impatto del cambiamento climatico sulla salute umana. Questa non è solo una questione di statistica o di modellazione, ma una dimostrazione di come le variazioni ambientali possano avere un effetto diretto e tangibile sulla vita quotidiana delle persone. Inoltre, il fatto che specifiche variabili, come la pressione atmosferica, abbiano un'influenza preponderante sulle patologie cardiovascolari sottolinea la complessità dei sistemi biologici e la loro interazione con l'ambiente. D'altro canto, l'indagine sulle precipitazioni nelle regioni del Lazio e della Basilicata ha messo in evidenza non solo l'importanza della modellazione avanzata nella previsione meteorologica, ma anche il ruolo cruciale che tali previsioni possono avere nella gestione delle risorse idriche e nella pianificazione delle risposte ai mutamenti climatici. La raffinatezza dei modelli presentati nella ricerca rappresenta un passo significativo verso una migliore comprensione dei fenomeni climatici. Le innovazioni nei metodi di clusterizzazione e nella modellazione delle precipitazioni delineano un futuro in cui la previsione meteorologica potrebbe diventare sempre più precisa, consentendo decisioni strategiche più informate in vari settori, dallo sviluppo infrastrutturale alla protezione civile. E, mentre le tecniche e i modelli presentati in questa tesi rappresentano sviluppi significativi, vi è ancora ampio spazio per ulteriori progressi, in particolare attraverso l'introduzione di altre variabili e metodi di modellazione. Riunendo le due aree di studio, si può affermare che l'impatto del cambiamento climatico sulla società è vasto e interconnesso. Dal punto di vista ambientale, le risorse idriche sono direttamente influenzate dai cambiamenti climatici, il che a sua volta influisce sul benessere umano. Dal punto di vista sanitario, l'impatto si estende ben oltre le risorse idriche, influenzando la salute cardiovascolare delle persone. L'importanza di un approccio multidisciplinare, come dimostrato in questa tesi, non può essere sottovalutata. Questa ricerca sottolinea non solo l'urgenza di rispondere alle sfide poste dal cambiamento climatico, ma anche la necessità di utilizzare strumenti avanzati come il machine learning per migliorare la comprensione e la gestione di tali sfide.

# Bibliografia


[1] «Climate change and health». Consultato: 26 settembre 2023. [Online]. Disponibile su: https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

[2] C. Butsch *et al.*, «Health impacts of extreme weather events – Cascading risks in a changing climate», set. 2023, Consultato: 26 settembre 2023. [Online]. Disponibile su: https://edoc.rki.de/handle/176904/11269

[3] S. N. Jonkman, «Global perspectives on loss of human life caused by floods», *Nat. Hazards*, vol. 34, fasc. 2, pp. 151–175, 2005.

[4] B. Merz *et al.*, «Floods and climate: emerging perspectives for flood risk assessment and management», *Nat. Hazards Earth Syst. Sci.*, vol. 14, fasc. 7, pp. 1921–1942, lug. 2014, doi: 10.5194/nhess-14-1921-2014.

[5] C. Xu, E. Widén, e S. Halldin, «Modelling hydrological consequences of climate change—Progress and challenges», *Adv. Atmospheric Sci.*, vol. 22, fasc. 6, pp. 789–797, nov. 2005, doi: 10.1007/BF02918679.

[6] L. O. Mearns, F. Giorgi, L. McDaniel, e C. Shields, «Analysis of daily variability of precipitation in a nested regional climate model: comparison with observations and doubled CO2 results», *Glob. Planet. Change*, vol. 10, fasc. 1, pp. 55–78, apr. 1995, doi: 10.1016/0921-8181(94)00020-E.

[7] J. Olsson, W. Yang, e T. Bosshard, «Climate model precipitation in hydrological impact studies : limitations and possibilities», *VATTEN – J Water Manage Res*, vol. 69, pp. 221–230, gen. 2013.

[8] U. Ehret, E. Zehe, V. Wulfmeyer, K. Warrach-Sagi, e J. Liebert, «HESS Opinions "Should we apply bias correction to global and regional climate model data?"», *Hydrol. Earth Syst. Sci.*, vol. 16, fasc. 9, pp. 3391–3404, set. 2012, doi: 10.5194/hess-16-3391-2012.

[9] A. Bronstert, D. Niehoff, e G. Bürger, «Effects of climate and land-use change on storm runoff generation: present knowledge and modelling capabilities: EFFECTS OF CLIMATE AND LAND-USE CHANGE ON STORM RUNOFF GENERATION», *Hydrol. Process.*, vol. 16, fasc. 2, pp. 509–529, feb. 2002, doi: 10.1002/hyp.326.

[10] «Downscaling general circulation model output: a review of methods and limitations - R.L. Wilby, T.M.L. Wigley, 1997». Consultato: 9 settembre 2023. [Online]. Disponibile su: https://journals.sagepub.com/doi/abs/10.1177/030913339702100403

[11] C. Karamperidou, F. Cioffi, e U. Lall, «Surface Temperature Gradients as Diagnostic Indicators of Midlatitude Circulation Dynamics», *J. Clim.*, vol. 25, fasc. 12, pp. 4154–4171, giu. 2012, doi: 10.1175/JCLI-D-11-00067.1.

[12] «The Response of Precipitation Minus Evapotranspiration to Climate Warming: Why the "Wet-Get-Wetter, Dry-Get-Drier" Scaling Does Not Hold over Land in: Journal of Climate Volume 28 Issue 20 (2015)». Consultato: 9 settembre 2023. [Online]. Disponibile su: https://journals.ametsoc.org/view/journals/clim/28/20/jcli-d-15-0369.1.xml?tab_body=fulltext-display

[13] «Non-Homogeneous Hidden Markov Model for Precipitation Occurrence | Journal of the Royal Statistical Society Series C: Applied Statistics | Oxford Academic». Consultato: 9 settembre 2023. [Online]. Disponibile su: https://academic.oup.com/jrsssc/article/48/1/15/6990642

[14] S. P. Charles, B. C. Bates, I. N. Smith, e J. P. Hughes, «Statistical downscaling of daily precipitation from observed and modelled atmospheric fields», *Hydrol. Process.*, vol. 18, fasc. 8, pp. 1373–1394, 2004, doi: 10.1002/hyp.1418.

[15] J. P. Hughes e P. Guttorp, «Incorporating Spatial Dependence and Atmospheric Data in a Model of Precipitation», *J. Appl. Meteorol. Climatol.*, vol. 33, fasc. 12, pp. 1503–1515, dic. 1994, doi: 10.1175/1520-0450(1994)033<1503:ISDAAD>2.0.CO;2.

[16] «A spatiotemporal model for downscaling precipitation occurrence and amounts», *J. Geophys. Res. Atmospheres*, vol. 104, fasc. D24, pp. 31657–31669, 1999, doi: 10.1029/1999JD900119.
[17] S. P. Charles, B. C. Bates, P. H. Whetton, e J. P. Hughes, «Validation of downscaling models for changed climate conditions: case study of southwestern Australia», *Clim. Res.*, vol. 12, fasc. 1, pp. 1–14, giu. 1999, doi: 10.3354/cr012001.
[18] E. Bellone, J. P. Hughes, e P. Guttorp, «A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts», *Clim. Res.*, vol. 15, fasc. 1, pp. 1–12, mag. 2000, doi: 10.3354/cr015001.
[19] A. W. Robertson, A. V. M. Ines, e J. W. Hansen, «Downscaling of Seasonal Precipitation for Crop Simulation», *J. Appl. Meteorol. Climatol.*, vol. 46, fasc. 6, pp. 677–693, giu. 2007, doi: 10.1175/JAM2495.1.
[20] A. W. Robertson, V. Moron, e Y. Swarinoto, «Seasonal predictability of daily rainfall statistics over Indramayu district, Indonesia», *Int. J. Climatol.*, vol. 29, fasc. 10, pp. 1449–1462, 2009, doi: 10.1002/joc.1816.
[21] W. L. Tan, F. Yusof, e Z. Yusop, «Non-Homogeneous Hidden Markov Model for Daily Rainfall Amount in Peninsular Malaysia», *J. Teknol.*, vol. 63, fasc. 2, giu. 2013, doi: 10.11113/jt.v63.1916.
[22] F. Cioffi, F. Conticello, U. Lall, L. Marotta, e V. Telesca, «Large scale climate and rainfall seasonality in a Mediterranean Area: Insights from a non-homogeneous Markov model applied to the Agro-Pontino plain», *Hydrol. Process.*, vol. 31, fasc. 3, pp. 668–686, 2017, doi: 10.1002/hyp.11061.
[23] F. Conticello, F. Cioffi, B. Merz, e U. Lall, «An event synchronization method to link heavy rainfall events and large-scale atmospheric circulation features», *Int. J. Climatol.*, vol. 38, fasc. 3, pp. 1421–1437, 2018, doi: 10.1002/joc.5255.
[24] «cds.climate.copernicus». Consultato: 9 settembre 2023. [Online]. Disponibile su: https://cds.climate.copernicus.eu/#!/home
[25] R. Quian Quiroga, T. Kreuz, e P. Grassberger, «Event synchronization: A simple and fast method to measure synchronicity and time delay patterns», *Phys. Rev. E*, vol. 66, fasc. 4, p. 041904, ott. 2002, doi: 10.1103/PhysRevE.66.041904.
[26] T. Kreuz, N. Bozanic, e M. Mulansky, «SPIKE-Synchronization: a parameter-free and time-resolved coincidence detector with an intuitive multivariate extension», *BMC Neurosci.*, vol. 16, fasc. 1, p. P170, dic. 2015, doi: 10.1186/1471-2202-16-S1-P170.
[27] J. P. Hughes, P. Guttorp, e S. P. Charles, «A Non-Homogeneous Hidden Markov Model for Precipitation Occurrence», *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 48, fasc. 1, pp. 15–30, mar. 1999, doi: 10.1111/1467-9876.00136.
[28] B. . -H. Juang and L. R. Rabiner, «The segmental K-means algorithm for estimating parameters of hidden Markov models», *Segmental K-Means Algorithm Estim. Parameters Hidden Markov Models*, vol. 38, fasc. 9, pp. 1639–1641, set. 1990, doi: doi: 10.1109/29.60082.
[29] L. Guo, Z. Jiang, M. Ding, W. Chen, e L. Li, «Downscaling and projection of summer rainfall in Eastern China using a nonhomogeneous hidden Markov model», *Int. J. Climatol.*, vol. 39, fasc. 3, pp. 1319–1330, 2019, doi: 10.1002/joc.5882.
[30] A. Viterbi, «Error bounds for convolutional codes and an asymptotically optimum decoding algorithm», *IEEE Trans. Inf. Theory*, vol. 13, fasc. 2, pp. 260–269, apr. 1967, doi: 10.1109/TIT.1967.1054010.
[31] «Modeling of multivariate time series using hidden Markov models - ProQuest». Consultato: 10 settembre 2023. [Online]. Disponibile su: https://www.proquest.com/openview/0aea4bf7444405b23a7bf1a069c746cb/1?pq-origsite=gscholar&cbl=18750&diss=y

[32] F. Cioffi, U. Lall, E. Rus, e C. K. B. Krishnamurthy, «Space-time structure of extreme precipitation in Europe over the last century», *Int. J. Climatol.*, vol. 35, fasc. 8, pp. 1749–1760, 2015, doi: 10.1002/joc.4116.
[33] Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, e R. Chunara, «Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review», *Am. J. Prev. Med.*, vol. 61, fasc. 4, pp. 596–605, ott. 2021, doi: 10.1016/j.amepre.2021.04.016.
[34] I. P. on C. Change e I. P. on C. C. W. G. I, *Climate Change 2007 - The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Cambridge University Press, 2007.
[35] L. Filleul *et al.*, «The Relation Between Temperature, Ozone, and Mortality in Nine French Cities During the Heat Wave of 2003», *Environ. Health Perspect.*, vol. 114, fasc. 9, pp. 1344–1347, set. 2006, doi: 10.1289/ehp.8328.
[36] P. Zhai, A. Sun, F. Ren, X. Liu, B. Gao, e Q. Zhang, «Changes of Climate Extremes in China», in *Weather and Climate Extremes: Changes, Variations and a Perspective from the Insurance Industry*, T. R. Karl, N. Nicholls, e A. Ghazi, A c. di, Dordrecht: Springer Netherlands, 1999, pp. 203–218. doi: 10.1007/978-94-015-9265-9_13.
[37] J. Huang, J. Wang, e W. Yu, «The Lag Effects and Vulnerabilities of Temperature Effects on Cardiovascular Disease Mortality in a Subtropical Climate Zone in China», *Int. J. Environ. Res. Public. Health*, vol. 11, fasc. 4, Art. fasc. 4, apr. 2014, doi: 10.3390/ijerph110403982.
[38] J. A. Patz, D. Campbell-Lendrum, T. Holloway, e J. A. Foley, «Impact of regional climate change on human health», *Nature*, vol. 438, fasc. 7066, Art. fasc. 7066, nov. 2005, doi: 10.1038/nature04188.
[39] P. R. Epstein, «Climate Change and Human Health», *N. Engl. J. Med.*, vol. 353, fasc. 14, pp. 1433–1436, ott. 2005, doi: 10.1056/NEJMp058079.
[40] A. Haines e J. A. Patz, «Health Effects of Climate Change», *JAMA*, vol. 291, fasc. 1, pp. 99–103, gen. 2004, doi: 10.1001/jama.291.1.99.
[41] T. W. Anderson e W. H. Le Riche, «COLD WEATHER AND MYOCARDIAL INFARCTION», *The Lancet*, vol. 295, fasc. 7641, pp. 291–296, feb. 1970, doi: 10.1016/S0140-6736(70)90651-3.
[42] B. Marchant, K. Ranjadayalan, R. Stevenson, P. Wilkinson, e A. D. Timmis, «Circadian and seasonal factors in the pathogenesis of acute myocardial infarction: the influence of environmental temperature.», *Heart*, vol. 69, fasc. 5, pp. 385–387, mag. 1993, doi: 10.1136/hrt.69.5.385.
[43] F. A. Spencer, R. J. Goldberg, R. C. Becker, J. M. Gore, e null null, «Seasonal Distribution of Acute Myocardial Infarction in the Second National Registry of Myocardial Infarction 1», *J. Am. Coll. Cardiol.*, vol. 31, fasc. 6, pp. 1226–1233, mag. 1998, doi: 10.1016/S0735-1097(98)00098-9.
[44] D. A. Lawlor, G. D. Smith, R. Mitchell, e S. Ebrahim, «Temperature at birth, coronary heart disease, and insulin resistance: cross sectional analyses of the British women's heart and health study», *Heart*, vol. 90, fasc. 4, pp. 381–388, apr. 2004, doi: 10.1136/hrt.2002.009548.
[45] W.-H. Pan, L.-A. Li, e M.-J. Tsai, «Temperature extremes and mortality from coronary heart disease and cerebral infarction in elderly Chinese», *The Lancet*, vol. 345, fasc. 8946, pp. 353–355, feb. 1995, doi: 10.1016/S0140-6736(95)90341-0.
[46] E. M. Kilbourne, «The spectrum of illness during heat waves», *Am. J. Prev. Med.*, vol. 16, fasc. 4, pp. 359–360, mag. 1999, doi: 10.1016/s0749-3797(99)00016-1.
[47] «The Impact of Climate Change on Our Health and Health Systems». Consultato: 11 ottobre 2023. [Online]. Disponibile su:

https://www.commonwealthfund.org/publications/explainer/2022/may/impact-climate-change-our-health-and-health-systems
[48] K. Knowlton *et al.*, «The 2006 California Heat Wave: Impacts on Hospitalizations and Emergency Department Visits», *Environ. Health Perspect.*, vol. 117, fasc. 1, pp. 61–67, gen. 2009, doi: 10.1289/ehp.11594.
[49] A. Zanobetti e J. Schwartz, «Temperature and Mortality in Nine US Cities», *Epidemiol. Camb. Mass*, vol. 19, fasc. 4, pp. 563–570, lug. 2008, doi: 10.1097/EDE.0b013e31816d652d.
[50] A. J. McMichael, R. E. Woodruff, e S. Hales, «Climate change and human health: present and future risks», *The Lancet*, vol. 367, fasc. 9513, pp. 859–869, mar. 2006, doi: 10.1016/S0140-6736(06)68079-3.
[51] W. Ma *et al.*, «The temperature–mortality relationship in China: An analysis from 66 Chinese communities», *Environ. Res.*, vol. 137, pp. 72–77, feb. 2015, doi: 10.1016/j.envres.2014.11.016.
[52] A. Bunker *et al.*, «Effects of Air Temperature on Climate-Sensitive Mortality and Morbidity Outcomes in the Elderly; a Systematic Review and Meta-analysis of Epidemiological Evidence», *eBioMedicine*, vol. 6, pp. 258–268, apr. 2016, doi: 10.1016/j.ebiom.2016.02.034.
[53] N. Watts *et al.*, «The 2019 report of The Lancet Countdown on health and climate change: ensuring that the health of a child born today is not defined by a changing climate», *The Lancet*, vol. 394, fasc. 10211, pp. 1836–1878, nov. 2019, doi: 10.1016/S0140-6736(19)32596-6.
[54] M. Gostimirovic *et al.*, «The influence of climate change on human cardiovascular function», *Arch. Environ. Occup. Health*, vol. 75, fasc. 7, pp. 406–414, ott. 2020, doi: 10.1080/19338244.2020.1742079.
[55] M. BAAGHIDEH e F. MAYVANEH, «Climate Change and Simulation of Cardiovascular Disease Mortality: A Case Study of Mashhad, Iran», *Iran. J. Public Health*, vol. 46, fasc. 3, pp. 396–407, mar. 2017.
[56] S. Stewart, A. K. Keates, A. Redfern, e J. J. V. McMurray, «Seasonal variations in cardiovascular disease», *Nat. Rev. Cardiol.*, vol. 14, fasc. 11, Art. fasc. 11, nov. 2017, doi: 10.1038/nrcardio.2017.76.
[57] D. M. Beyerbach, R. J. Kovacs, A. A. Dmitrienko, D. M. Rebhun, e D. P. Zipes, «Heart rate–corrected QT interval in men increases during winter months», *Heart Rhythm*, vol. 4, fasc. 3, pp. 277–281, mar. 2007, doi: 10.1016/j.hrthm.2006.11.008.
[58] M. Morabito *et al.*, «Relationships between weather and myocardial infarction: A biometeorological approach», *Int. J. Cardiol.*, vol. 105, fasc. 3, pp. 288–293, dic. 2005, doi: 10.1016/j.ijcard.2004.12.047.
[59] M. Scortichini, M. De Sario, F. K. De'Donato, M. Davoli, P. Michelozzi, e M. Stafoggia, «Short-Term Effects of Heat on Mortality and Effect Modification by Air Pollution in 25 Italian Cities», *Int. J. Environ. Res. Public. Health*, vol. 15, fasc. 8, Art. fasc. 8, ago. 2018, doi: 10.3390/ijerph15081771.
[60] B. Frist, «Climate Change And Health: A Heart Specialist's Notes On How A Warming Planet Impacts Our Health And Wellbeing», Forbes. Consultato: 11 ottobre 2023. [Online]. Disponibile su: https://www.forbes.com/sites/billfrist/2021/12/17/climate-change-and-health-a-heart-specialists-notes-on-how-a-warming-planet-impacts-our-health-and-wellbeing/
[61] «Air pollution». Consultato: 11 ottobre 2023. [Online]. Disponibile su: https://www.who.int/health-topics/air-pollution
[62] G. D'Amato *et al.*, «Climate Change and Air Pollution: Effects on Respiratory Allergy», *Allergy Asthma Immunol. Res.*, vol. 8, fasc. 5, pp. 391–395, mar. 2016, doi: 10.4168/aair.2016.8.5.391.

[63] A. Peters e A. Schneider, «Cardiovascular risks of climate change», *Nat. Rev. Cardiol.*, vol. 18, fasc. 1, Art. fasc. 1, gen. 2021, doi: 10.1038/s41569-020-00473-5.
[64] Y. Raita, C. A. Camargo, L. Liang, e K. Hasegawa, «Big Data, Data Science, and Causal Inference: A Primer for Clinicians», *Front. Med.*, vol. 8, 2021, Consultato: 7 ottobre 2023. [Online]. Disponibile su: https://www.frontiersin.org/articles/10.3389/fmed.2021.678047
[65] V. Telesca, G. Castronuovo, G. Favia, C. Marranchelli, V. A. Pizzulli, e M. Ragosta, «Effects of Meteo-Climatic Factors on Hospital Admissions for Cardiovascular Diseases in the City of Bari, Southern Italy», *Healthcare*, vol. 11, fasc. 5, Art. fasc. 5, gen. 2023, doi: 10.3390/healthcare11050690.
[66] A. Väänänen, K. Haataja, K. Vehviläinen-Julkunen, e P. Toivanen, «AI in healthcare: A narrative review». F1000Research, 8 ottobre 2021. doi: 10.12688/f1000research.26997.2.
[67] C. Bellinger, M. S. M. Jabbar, O. R. Zaiane, e A. Osornio-Vargas, «A systematic review of data mining and machine learning for air pollution epidemiology», *BMC Public Health*, vol. 17, 2017, doi: 10.1186/s12889-017-4914-3.
[68] S. Cho *et al.*, «Pre-existing and machine learning-based models for cardiovascular risk prediction», *Sci. Rep.*, vol. 11, 2021, doi: 10.1038/s41598-021-88257-w.
[69] S. Weng, J. Reps, J. Kai, J. Garibaldi, e N. Qureshi, «Can machine-learning improve cardiovascular risk prediction using routine clinical data?», *PLoS ONE*, vol. 12, 2017, doi: 10.1371/journal.pone.0174944.
[70] C. Krittanawong *et al.*, «Machine learning prediction in cardiovascular diseases: a meta-analysis», *Sci. Rep.*, vol. 10, 2020, doi: 10.1038/s41598-020-72685-1.
[71] M. Akel, K. Carey, C. Winslow, M. Churpek, e D. Edelson, «Less is More: Detecting Clinical Deterioration in the Hospital with Machine Learning Using Only Age, Heart Rate, and Respiratory Rate.», *Resuscitation*, 2021, doi: 10.1016/j.resuscitation.2021.08.024.
[72] E. Ross, K. Jung, J. Dudley, L. Li, N. Leeper, e N. Shah, «Predicting Future Cardiovascular Events in Patients With Peripheral Artery Disease Using Electronic Health Record Data», *Circ. Cardiovasc. Qual. Outcomes*, vol. 12, 2019, doi: 10.1161/CIRCOUTCOMES.118.004741.
[73] V. Dominic, D. Gupta, e S. Khare, «An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease», *Appl. Med. Inform.*, vol. 36, 2015, Consultato: 8 ottobre 2023. [Online]. Disponibile su: https://consensus.app/details/applying-machine-learning-techniques-data-specifically-dominic/d74c11f443015de5aaf0a9c3216c63e5/
[74] J. Peng, C. Chen, M. Zhou, X. Xie, Y. Zhou, e C.-H. Luo, «Peak Outpatient and Emergency Department Visit Forecasting for Patients With Chronic Respiratory Diseases Using Machine Learning Methods: Retrospective Cohort Study», *JMIR Med. Inform.*, vol. 8, fasc. 3, p. e13075, mar. 2020, doi: 10.2196/13075.
[75] L. Li *et al.*, «Prediction and Diagnosis of Respiratory Disease by Combining Convolutional Neural Network and Bi-directional Long Short-Term Memory Methods», *Front. Public Health*, vol. 10, 2022, Consultato: 12 ottobre 2023. [Online]. Disponibile su: https://www.frontiersin.org/articles/10.3389/fpubh.2022.881234
[76] J. Rigdon e S. Basu, «Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the USA using nationally randomly sampled data», *BMJ Open*, vol. 9, 2019, doi: 10.1136/bmjopen-2019-032703.
[77] J. L. M. Amaral, A. J. Lopes, J. M. Jansen, A. C. D. Faria, e P. L. Melo, «An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms», *Comput. Methods Programs Biomed.*, vol. 112, fasc. 3, pp. 441–454, dic. 2013, doi: 10.1016/j.cmpb.2013.08.004.

[78] M. Gostimirovic *et al.*, «The influence of climate change on human cardiovascular function», *Arch. Environ. Occup. Health*, vol. 75, fasc. 7, pp. 406–414, ott. 2020, doi: 10.1080/19338244.2020.1742079.
[79] «Agenzia Regionale per la Prevenzione e la Protezione dell'Ambiente - Meteo». Consultato: 14 ottobre 2023. [Online]. Disponibile su: https://www.arpa.puglia.it/pagina2839_meteo.html
[80] W. E. Marcílio e D. M. Eler, «From explanations to feature selection: assessing SHAP values as feature selection mechanism», in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, nov. 2020, pp. 340–347. doi: 10.1109/SIBGRAPI51738.2020.00053.
[81] «Contributions to the Theory of Games (AM-28), Volume II - Google Libri». Consultato: 18 dicembre 2023. [Online]. Disponibile su: https://books.google.it/books?hl=it&lr=&id=Pd3TCwAAQBAJ&oi=fnd&pg=PA307&dq=L.+S.+S.+Shapley,+%E2%80%9CA+valuue+for+n-person+games,%E2%80%9D+in+Contributions+to+the+Theory+of+Games,+1953,+pp.+307%E2%80%93317.&ots=gunVC87nxU&sig=5n3Lsfky-eZEwAySD9DY7upPaUI&redir_esc=y#v=onepage&q&f=false
[82] «From local explanations to global understanding with explainable AI for trees | Nature Machine Intelligence». Consultato: 18 dicembre 2023. [Online]. Disponibile su: https://www.nature.com/articles/s42256-019-0138-9
[83] M. Theodosiou, «Forecasting monthly and quarterly time series using STL decomposition», *Int. J. Forecast.*, vol. 27, fasc. 4, pp. 1178–1195, ott. 2011, doi: 10.1016/j.ijforecast.2010.11.002.
[84] D. A. Dickey e W. A. Fuller, «Distribution of the Estimators for Autoregressive Time Series with a Unit Root», *J. Am. Stat. Assoc.*, vol. 74, fasc. 366a, pp. 427–431, giu. 1979, doi: 10.1080/01621459.1979.10482531.
[85] «Train/Test Split and Cross Validation in Python | by Adi Bronshtein | Towards Data Science». Consultato: 14 ottobre 2023. [Online]. Disponibile su: https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6
[86] M. L. Bell, R. D. Peng, F. Dominici, e J. M. Samet, «Emergency Hospital Admissions for Cardiovascular Diseases and Ambient Levels of Carbon Monoxide», *Circulation*, vol. 120, fasc. 11, pp. 949–955, set. 2009, doi: 10.1161/CIRCULATIONAHA.109.851113.
[87] W. S. Aronow, «Effect of carbon monoxide on cardiovascular disease», *Prev. Med.*, vol. 8, fasc. 3, pp. 271–278, mag. 1979, doi: 10.1016/0091-7435(79)90003-3.
[88] C. Liu *et al.*, «Ambient carbon monoxide and cardiovascular mortality: a nationwide time-series analysis in 272 cities in China», *Lancet Planet. Health*, vol. 2, fasc. 1, pp. e12–e18, gen. 2018, doi: 10.1016/S2542-5196(17)30181-X.
[89] null Wei Yang Brian L. Jennison Stanley T. Omaye, «Cardiovascular Disease Hospitalization and Ambient Levels of Carbon Monoxide», *J. Toxicol. Environ. Health A*, vol. 55, fasc. 3, pp. 185–196, ott. 1998, doi: 10.1080/009841098158485.
[90] E. N. Allred *et al.*, «Short-Term Effects of Carbon Monoxide Exposure on the Exercise Performance of Subjects with Coronary Artery Disease», *N. Engl. J. Med.*, vol. 321, fasc. 21, pp. 1426–1432, nov. 1989, doi: 10.1056/NEJM198911233212102.
[91] Y. Borghei, M. T. Moghadamnia, A. E. Sigaroudi, e A. Ghanbari, «Association between climate variables (cold and hot weathers, humidity, atmospheric pressures) with out-of-hospital cardiac arrests in Rasht, Iran», *J. Therm. Biol.*, vol. 93, p. 102702, ott. 2020, doi: 10.1016/j.jtherbio.2020.102702.
[92] J. Xiao, X. Liu, J. Zeng, Y. Cao, e Z. Feng, «Recommendation of Healthcare Services Based on an Embedded User Profile Model», *Int. J. Semantic Web Inf. Syst. IJSWIS*, vol. 18, fasc. 1, pp. 1–21, gen. 2022, doi: 10.4018/IJSWIS.313198.

[93] G. N. Nguyen, N. H. L. Viet, M. Elhoseny, K. Shankar, B. B. Gupta, e A. A. A. El-Latif, «Secure blockchain enabled Cyber–physical systems in healthcare using deep belief network with ResNet model», *J. Parallel Distrib. Comput.*, vol. 153, pp. 150–160, lug. 2021, doi: 10.1016/j.jpdc.2021.03.011.
[94] K. Shankar, E. Perumal, M. Elhoseny, F. Taher, B. B. Gupta, e A. A. A. El-Latif, «Synergic Deep Learning for Smart Health Diagnosis of COVID-19 for Connected Living and Smart Cities», *ACM Trans. Internet Technol.*, vol. 22, fasc. 3, p. 61:1-61:14, nov. 2021, doi: 10.1145/3453168.
[95] H. Q. Yu e S. Reiff-Marganiec, «Learning Disease Causality Knowledge From the Web of Health Data», *Int. J. Semantic Web Inf. Syst. IJSWIS*, vol. 18, fasc. 1, pp. 1–19, gen. 2022, doi: 10.4018/IJSWIS.297145.