Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Energetica e Nucleare

Tesi di Laurea Magistrale

Applicazione di tecniche di data analytics per la caratterizzazione spazio temporale della qualità dell'ambiente interno negli edifici

Relatore

Prof. Alfonso Capozzoli

Dr. Marco Savino Piscitelli

Ing. Roberto Chiosa

Candidata

QICHAO YE

Anno Accademico 2022/2023

Sommario

Gli edifici utilizzano circa il 40% dell'energia totale consumata a livello globale e generano il 36% delle emissioni totali di anidride carbonica. Al loro interno, i sistemi di riscaldamento, ventilazione e condizionamento dell'aria (HVAC) possono arrivare a costituire la quota più rilevante di consumo energetico fino a circa il 40%. I sistemi HVAC sono essenziali per il controllo della temperatura interna e garantire un adeguato livello di comfort termico e di qualità dell'aria interna. Tra tutti i parametri ambientali la temperatura dell'aria è sicuramente tra i più importanti. L'approfondita caratterizzazione sia spaziale che temporale di tale parametro ha un grande impatto sulla definizione di strategie di controllo del sistema HVAC consentendo di ottimizzare i consumi energetici dell'edificio minimizzando le violazioni dei livelli prescrizionali di comfort interno.

In questo elaborato è stata effettuata una caratterizzazione spazio-temporale dei trend di temperatura interna di un edificio ad uso ufficio, con il fine di proporre un approccio al problema predittivo che tenesse conto sia di una ampia distribuzione spaziale dei punti di misura che dell'evoluzione temporale di ciascuna serie storica misurata. In particolare, il processo di analisi è stato suddiviso in due fasi. la prima ha riguardato la caratterizzazione dei profili temporali giornalieri della temperatura associati a 63 sensori istallati in campo attraverso la tecnica di clustering e la loro classificazione tramite un albero decisionale con lo scopo di estrarre conoscenza riguardo le principali relazioni spaziali che sussistono tra i profili di temperatura tipologici e la posizione dei sensori. La seconda fase ha riguardato lo sviluppo di un modello di predizione in grado di prevedere per alcuni timestep in avanti l'andamento della temperatura interna nelle medesime 63 posizioni sfruttando le relazioni spaziali precedentemente ricavate. A tal fine è stato utilizzato un algoritmo di recente concezione di Spatial–Temporal Graph Neural Networks (STGNN) le cui prestazioni sono state comparate con un modello predittivo di baseline di tipo Long Short-Term Memory neural network. I risultati ottenuti dimostrano come nonostante le prestazioni dei due approcci siano globalmente comparabili, il modello GNN è in grado di trarre vantaggio dalle informazioni spaziali per raggiungere un risultato più accurato in alcune zone. In aggiunta il modello GNN ha consentito

di sviluppare un unico modello predittivo multi-output generando un significativo vantaggio dal punto di vista implementativo.

Indice

Indice di Tabelle

Indice delle Figure

Indice degli Acronimi

ANN Artificial Neural Network

CART Classification and Regression Tree

ConvGNN Convolutional Graph Neural Networks

CNN Convolutional Neural Network

GNN Graph Neural Network

IEQ Indoor Environment Quality

KNN K-Nearest Neighbors

LSTM Long Short-Term Memory Neural Network

MAPE Mean Average Percentage Error

MAE Mean Absolut Error

MSE Mean Square Error

NA Not Available/ Missing Value

PAA Piecewice Aggregate Approximation

ReLu Rectified Linear Unit

RecGNN Recurrent Graph Neural Networks

RNN Recurrent Neural Network

STGNN Spatial–Temporal Graph Neural Networks

Capitolo 1

Introduzione

Gli edifici utilizzano circa il 40% dell'energia totale consumata a livello globale e generano il 36% delle emissioni totali di anidride carbonica. Al loro interno, i sistemi di riscaldamento, ventilazione e condizionamento dell'aria (HVAC) possono arrivare a costituire la quota più rilevante di consumo energetico fino a circa il 40%. I sistemi HVAC sono essenziali per il controllo della temperatura interna e garantire un adeguato livello di comfort termico e di qualità dell'aria interna [1].

D'altra parte, sono in aumento le richieste per la qualità dell'ambiente interno negli spazi degli edifici. La condizione dell'ambiente interna, infatti, ha un impatto sostanziale sul benessere e sulla produttività degli occupanti degli edifici [2].

Nel caso di IEQ degli uffici è in relazione al benessere mentale, alla produttività, la concentrazione mentale, funzione cognitiva e assenteismo, infatti un basso IEQ può causare un peggioramento di salute psicologica, di umore, diminuzione di motivazione e aumento d distrazione e quindi un minore ritorno finanziario [3].

Nello studio del Allen et.al [4] ha determinato una diretta connessione tra la variazione di temperatura interna dell'edificio e la produttività, infatti viene diminuito di 2% di produttività per ogni 4°F allontanati dalla temperatura ottimale di 72°F. Il recupero di 2% di produttività per un'azienda che lavora in questo ambiente si trasforma un aumento di 9% di ricavo netto.

Pertanto, la riduzione del consumo energetico e l'aumento del comfort umano sono due motivi fondamentali per sostenere economicamente e socialmente lo sviluppo degli edifici intelligenti. A tal fine, la comprensione approfondita degli ambienti degli edifici è

essenziale per una gestione efficiente del consumo dell'energia negli edifici.

Tra tutti i parametri ambientali la temperatura dell'aria è sicuramente tra i più importanti. L'approfondita caratterizzazione sia spaziale che temporale di tale parametro ha un grande impatto sulla definizione di strategie di controllo del sistema HVAC consentendo di ottimizzare i consumi energetici dell'edificio minimizzando le violazioni dei livelli prescrizionali di comfort interno.

1.1 Literature review

Sono presenti numerosissime ricerche riguardo lo studio dei parametri dell'IEQ che tengono in considerazione sia le caratteristiche spaziali che temporali.

L'ispirazione dell'elaborato viene data dalla ricerca del Wickramasinghe et al.[5], riguardo allo studio delle serie temporali di temperatura interna rilevata da un elevato numero di sensori posizionati in diversi punti dell'edificio, nella ricerca di una relazione sia spaziale che temporale di temperatura. Questi autori hanno studiato un caso di un edificio commerciale selezionando i tre piani considerati più problematici dal proprietario. Obbiettivo di tale studio è la ricerca di un intervallo accettabile di numeri di sistemi di controllo e delle posizioni dei termostati affinché migliorasse il controllo di temperatura all'interno dell'edificio. È stato fatto un confronto di 2 tecniche di clustering per time series data, considerando varie combinazioni delle variabili dell'ambiente interno: la temperatura dell'aria, l'umidità relativa e la pressione dell'ambiente. Sono stati applicati k-means sulla media giornalieri e 2 metodi gerarchici sui profili temporali dei variabili. Notano che i risultati migliori si ha con il metodo Ward e la distanza Euclidea. In seguito, hanno compiuto un confronto dei cluster all'interno di questi piani divisi in zone termiche e osservano che la somiglianza tra clustering e zoning sono mediamente bassi.

Sahu et al. [6] hanno analizzato la concentrazione di particelle all'interno di una biblioteca universitaria per verificare che le concetrazioni delle particelle fossero sotto il limite imposto. Si tratta di un edificio a 4 piani e hanno collezionato i valori della concentrazione in 4-5 punti per piano e i parametri esogeni. Dall'analisi statistica è emerso che la variazione temporale delle concentrazioni di PM hanno un andamento simile, indipendentemente dalle loro dimensioni delle frazioni, e una maggiore concentrazione di PM è principalmente dovuto al movimento delle persone. Invece dai risultati del test

di Kruskal-Willis si afferma che esiste una variabilità statisticamente significativa tra i piani. La variazione spaziale della concentrazione degli inquinanti può essere dovuto alla differenza di occupazione, alla ventilazione e alle caratteristiche dei rispettivi piani. Inoltre i risultati confermano la forte correlazione tra la concentrazione di CO2 e occupazione.

Differentemente, Pollard et al. [3] hanno indagato la continua esposizione alla qualità dell'ambiente interno degli occupanti basando sui loro dati spazio-temporali. hanno creato creare un modello che fonde i dati di localizzazione degli occupanti in tempo reale ad alta risoluzione e di parametri di IEQ in modo da poter calcolare l'esposizione degli occupanti al IEQ. Attraverso spline cubica sono riusciti ad stimare i valori delle variabili di IEQ in ogni posizione del piano, utilizzando le misure dei sensori. In seguito, queste stime sono state combinate con i dati della posizione dei partecipanti, rilevati attraverso il sistema RTLS, in modo da poter valutare la qualità ambientale che ciascun partecipante sono esposti in qualunque posizione e istante di tempo. Mediate tale metodo sono riusciti a trarre le distribuzioni dei valori dei parametri IEQ a cui sono esposti gli occupanti.

Troncos et al. [7] hanno ricercato un modello di apprendimento automatico in grado di fornire informazioni dettagliate della condizione ambientale di un intero volume di un edificio basando solamente su i dati di monitoraggio in alcuni punti fissi acquisiti da un sistema automatizzato a basso costo. Viene svolto all'interno di in grande stanza multizonale open space, al piano terra del centro di ricerca CINTECX, a nord-ovest della Spagna. Oltre ai dati collezionati attraverso dei sensori installati a parete è stato utilizzato anche un dispositivo mobile per raccogliere dati in vari punti della stanza. Tra i tre modelli, MultiLayer Perceptron(MLP) neural network, Random Forest(RF) e Support Vector Regression(SVF), utilizzando solamente i dati dei sensori fissi, risulta migliore il modello RF. Nel secondo passo, aggiungendo anche i valori rilevati dal dispositivo mobile e la sua posizione come input al modello RF, i risultati si dimostrano di essere migliori rispetto a quelli senza.

Le seguenti sezioni contengono una breve trattazione dei due approcci utilizzati per la caratterizzazione spazio-temporale.

1.1.1 Spatio-temporal Clustering

Prima di parlare subito del spatio-temporal clustering, è necessario conoscere le diverse tipologie di dati spazio-temporali, secondo Kisilevieh et al[8] ci sono 5 tipi di dati:

1. ST event: a cui è assocciato solamente informazioni riguardo lo spazio e il tempo, è costituito da una tripletta di <latitudine, longitudine, timestamp>, che spesso utilizzato per segnalare accadimento di un evento,a esempio sisma o una malattia;

2. Geo-referenced variable: descrive l'osservazione di evoluzione nel tempo di un fenomeno avvenuto in un punto fisso, un esempio tipico è una stazione meteorologica che mostra la temperatura dell'aria più recente;

3. Geo-referenced variable: descrive l'osservazione di evoluzione nel tempo di un fenomeno avvenuto in un punto fisso, un esempio tipico è una stazione meteorologica che mostra la temperatura dell'aria più recente; Geo-referenced time series racconta tutta la storia di un oggetto evolvente in un arco di tempo, un esempio è la misurazione di temperatura oraria di un edificio per 3 anni;

4. Moving object: cambia la sua posizione nel tempo, non conservando la storia del movimento, mentre trajectories è proprio la sequenza di tutti i punti del moving object.

Due tipici algoritmi utilizzati per lo studio di dati spazio-temporali sono ST-DBSCAN [9] e ST-OPTICS [10], entrambi appartenenti alla famiglia density based clustering DBSCAN e OPTICS, richiedono la distanza massima e il numero minimo affinché un oggetto possa essere classificato come noise.

ST-DBSCAN, introdotto da Birant and Kut [9], richiede 2 parametri di distanza per supportare i dati spaziali bidimensionali, Eps1 utilizzato per definire la vicinanza di due punti geografici, Eps2 utile per misurare la somiglianza tra gli attributi non spaziali di 2 punti. Ad esempio di 2 punti A(x_1,y_1,t_1,t_2) e B(x_2,y_2,t_3,t_4), dove x e y sono attributi spaziali, t1 e t3 sono la temperatura giornaliera e t2 e t4 sono le temperature notturne. Eps1 e Eps2 sono calcolate in seguente modo:

$Esp_1 = sqrt((x_1 - x_2)^2 + (y_1 - y_2)^2)$

$Esp_2 = sqrt((t_1 - t_3)^2 + (t_2 - t_4)^2)$

Ovviamente prima di agire sulla parte spaziale i dati spatio temporali devono essere filtrati conservando solo i vicini temporali e loro corrispondenti valori spaziali.

ST-OPTICS, sviluppato da Agrawal et al [10] , si articola in 2 fasi, la prima consiste nel clusterizzare i dati ST mettendo come input all'algoritmo ST-OPTICS per ottenere raggruppamenti di oggetti con caratteristiche simili e passati all'algoritmo Extract -

STDBSCAN estraendo dei micro-clusters. Per una migliore analisi, visualizzazione e interpretazione dei cluster viene fatta la fase di agglomerazione utilizzando 2 categorie diverse di clustering, cioè density e hierarchical based algorithm.

Nel caso studio si tratta di avere geo-referenced time series, e secondo l'autore Kisilevieh et al[8] geo-referenced time series clustering ha obbiettivo di comparare l'evoluzione di serie temporali degli oggetti in relazione alle loro posizioni spaziali.

Izakian et al. [11] rivisitarono l'algoritmo di Fuzzy C-Means, rendendolo applicabile ai geo-referenced time series. La rivisitazione sta nel modificare la funzione obbiettivo, hanno introdotto una parte temporale nella funzione distanza:

$d_(\lambda)^2(v_i, x_k) = |v_i(s) - x_k(s)|^2 + \lambda|v_i(t) - x_k(t)|^2 \qquad \lambda > 0$

Dove $v_i(s)$ e $x_k(s)$ è la parte spaziale, $v_i(t)$ e $x_k(t)$ è la parte temporale e λ permette di controllare gli effetti di ciascuna parte per il calcolo della distanza Euclidea, infatti con λ=0, la parte temporale viene ignorata e rimane la classica funzione della distanza Euclidea.

Husch et al.[12] propongono l'algoritmo Correlation based Clustering of Big Spatiotemporal Datasets (CorClustST), partendo dallo spunto preso da ST-DBSCAN e ST-OPTICS. Viene determinato il numero di vicini per tutti i punti spaziali calcolando la correlazione di Pearson, che deve essere superiore a certo valore, entro un certo limite di distanza definiti precedentemente. Questi numeri di vicini vengono disposti in ordine decrescente. I clusters sono risultato di un processo iterativo a partire da questa lista del numero di vicini. Gli autori hanno testato questo algoritmo con successo al caso di cluster analysis di errori di previsione dell'energia eolica in Europa.

Per quanto riguardo analisi di clustering spazio-temporali riguardo i parametri IEQ sono ancora molto scarsi. Nei seguenti 2 articoli non viene applicato un metodo specifico per gestire i dati spazio-temporali, ma gli autori hanno fatto prima un time series clustering e successivamente attraverso un metodo di non clustering hanno analizzato i dati riguardo lo spazio.

Geng et al [13] studiano 2 casi studi i cui obbiettivo è l'individuazione dei pattern tipici dai profili giornalieri dei variabili IEQ. Dividono i profili giornalieri in diversi subsequences di poche ore e applicano k-means per determinare cluster a cui appartengono i vari sub-sequence. Infine ricombinando questi sub-sequence si ottiene il cluster finale dei profili giornalieri. Successivamente gli autori fanno un CART con i profili dominanti

mettendo come input: zona dell'ufficio, giorni della settimana, temperatura e umidità relativa esterna e ottengono informazioni spaziali riguardo di questi cluster.

Un altro metodo utilizzato per condurre analisi di tipo spazio-temporale è costituito dagli algoritmi di graph neural network.

1.1.2 Graph Neural Network

Il grafo è una struttura dati che descrive le relazioni tra un insieme di oggetti, consentendo di elaborare dati non limitati al dominio Euclideo. Grazie alla sua capacità di modellare una vasta gamma di sistemi, dalle scienze sociali alle scienze naturali e oltre, i grafi sono diventati strumenti molto potenti per molte ricerche. In particolare, l'algoritmo di Graph Neural Network (GNN), una tecnica di deep learning che sfrutta i grafi, è diventato molto popolare per l'analisi di diversi aspetti, come la chimica, il ragionamento di senso compiuto, l'elaborazione del linguaggio naturale, i social network e la previsione dei flussi di traffico.

Nel campo dei grafi, i GNN possono svolgere tre tipi di compiti: il primo riguarda la classificazione, la regressione e il clustering dei nodi del grafo stesso; il secondo consiste nella previsione dei collegamenti mancanti tra i nodi e nella classificazione dei collegamenti esistenti; infine, il terzo compito riguarda la classificazione o regressione dell'intero grafo, ottenendo una rappresentazione dell'insieme di nodi e collegamenti che lo compongono.

Secondo Wu et al.[14] sono presenti 4 tipologie di GNN: 1)Recurrent Graph Neural Networks, 2) Convolutional Graph Neural Networks, 3) graph auto-encoders e Spatial–Temporal Graph Neural Networks (STGNN).

Lo STGNN è generalmente composto da 2 modelli, in quanto tale algoritmo è acconsente di tenere in considerazione in contemporaneo sia le relazioni spaziali che le dipendenze temporali dei dati. Generalmente è una combinazione di ConvGNN che elabora la parte temporale e Recurrent Neural Network o CNN per quanto riguarda la dipendenza spaziale.

Un'applicazione molto comune di GGNN è lo studio del traffico in città, come dimostrato dallo studio condotto da Yu et al. [14], i quali hanno introdotto un modello di STGNN per la previsione del traffico. Il modello consiste in due blocchi convoluzionali spazio-temporali e un layer di output fully-connected. I blocchi sono a loro volta costituiti da due gated convolutional layer, che si occupano della dimensione temporale, e un graph convolutional layer per la dimensione spaziale. Il modello viene applicato al grafo ottenuto dalla matrice di adiacenza spaziale basata sulla distanza tra le stazioni di traffico, e i

risultati ottenuti sono poi confrontati con quelli dei baseline, che includono: 1) Historical Average(Ha); 2) Linear Support Victor Regression (LSVR); 3) Auto-Regressive Integrated Moving Average (ARIMA); 4) Feed-Forward Neural Network (FNN); 5) Full-Connected LSTM (FC-LSTM); 6) Graph Convolutional GRU (GCGRU). Il modello proposto ottiene le migliori performance per tutte le metriche di valutazione utilizzate (MAE, MAPE, RMSE).

Jeon et al. [15] ha proposto un nuovo nuovo modello di previsione dell'irraggiamento solare chiamato attribute-augmented spatiotemporal GCN (AST-GCN). Questo modello è un espansione dei modelli esistenti di Spazio-Temporal Graph Convolutional Neural Network (ST-GCN), a cui vengono aggiunti attributi dinamici. Nel caso studio, questi attributi dinamici sono i parametri atmosferici osservati da più stazioni meteorologiche, rappresentate come nodi del grafo. Quindi, il grafo ha una struttura statica e attributi dinamici. Attraverso questo modello, gli autori hanno analizzato le correlazioni spaziotemporali tra più variabili meteorologiche, considerando l'adiacenza spaziale delle stazioni, i cambiamenti temporali delle variabili meteorologiche e la varietà di variabili che influenzano le prestazioni di previsione. I risultati sperimentali hanno mostrato la proprietà sinergica di queste tre caratteristiche e hanno dimostrato che è difficile stabilire le correlazioni studiando singoli aspetti.

1.2 Obbiettivo dell'elaborato

In questo elaborato è stata effettuata un'analisi su un insieme di serie temporali di temperatura interna raccolti in 63 diversi punti, situati in 2 piani di un edificio ad uso ufficio con sede a Berkeley, in California.

L'analisi si è articolata in due fasi, nella prima fase si è indagato sulla caratterizzazione dei profili temporali giornalieri di temperatura interna attraverso il metodo del Clustering e nella seconda è stata effettuata un'analisi sulla previsione della temperatura nei suddetti punti dell'edifico attraverso il metodo del Graph Neural Network (GNN).

Lo studio di clustering è importante per l'analisi della seconda fase in quanto è stato appresso delle informazioni rilevanti, cioè l'influenza delle caratteristiche spaziali per la previsione della temperatura.

Infatti la maggior parte delle tecniche basati sulla Deep Neural Network (DNN) si focalizza sull'estrazione delle caratteristiche e delle relazioni temporali riguardo i valori rilevati dai sensori, mentre le dipendenze spaziali tra i sensori non vengono considerate nell'allenamento del modello di predizione. Molti ricercatori hanno utilizzato CNN affinché il modello apprenda le caratteristiche spaziali, eppure, per mezzo del suo peculiare meccanismo di funzionamento, non è possibile specificare le relazioni tra i sensori riguardo le caratteristiche spaziali estratte. A differenza, il GNN è in grado di assimila le relazione tra i sensori attraverso il grafo attraverso i collegamenti che sussistono tra i nodi (i sensori), in tal modo il GNN in condizione di propagare le informazioni nei nodi mediante i collegamenti e di apprendere rappresentazione promettente dei nodi o del grafo.

La tesi si compone dei seguenti capitoli: il Capitolo 2, nel quale viene esposta la metodologia e vengono spiegati i metodi utilizzati per effettuare l'analisi; il Capitolo 3, nel quale viene presentato il caso studio con una breve descrizione dell'edificio e del dataset; il Capitolo 4, nel quale vengono discussi i risultati relativi alle due fasi dell'analisi e infine il Capitolo 5, nella quale vengono esposta una conclusione relativa ai fattori salienti emersi nella ricerca.

Capitolo 2

Metodologia

Nel seguente capitolo sono stati esposti in modo dettagliato i procedimenti e i metodi che sono stati utilizzati per compiere l'analisi del caso studio.

In un primo momento sono stati visionati la dati, ovvero i valori misurati dai vari sensori raccolti ed elaborati da terzi.

Dopo la prima fase si procede al pre-processing dei dati, principalmente rivolto all'eliminazione dei dati incoerenti con il resto del dataset e alla sostituzione degli eventuali valori mancanti. Si tratta di un passaggio fondamentale prima della vera analisi, poiché in questa fase si effettua la preparazione dei dati per poi essere analizzati, in quanto la qualità e la bontà dei dati influenza enormemente i risultati finali delle varie tecniche di analisi.

Successivamente, si passa alla caratterizzazione dei profili temporali giornalieri di temperatura interna in diversi punti dell'edificio attraverso Clustering e Classification and Regression Tree (CART), per determinare i profili caratteristici dominanti e i parametri da cui dipende.

Infine, si effettua un'analisi predittiva dei valori di temperatura di qualche ora in avanti in tutti punti studiati dell'edificio tramite una tecnica di deep learning chiamato Graph Neural Network (GNN), i cui risultati saranno paragonati a quelli di un altro modello strutturalmente più semplice, chiamato Long Short-Term Memory Neural Network (LSTM).

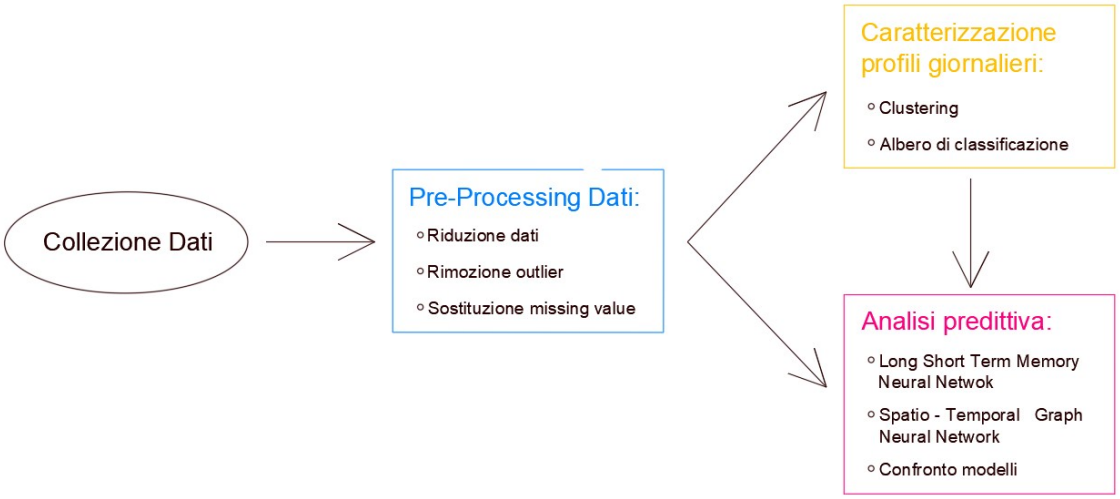


Figura 2.1: Pipeline della metodologia

2.1 Pre-processing

I dati reali, raccolti attraverso i sensori e inviati al sistema di raccolta e di elaborazione, presentano spesso dei valori errati oppure valori mancanti per diversi motivi, possono essere dovuti a un malfunzionamento dei sensori o del sistema, a un'erronea misurazione dei sensori oppure a un calo temporaneo di tensione che interrompe l'invio del segnale causando un possibile accumulo di valori che vengono inviati come la somma e di conseguenza un risultato errato. Perciò il dataset, prima di essere sottoposto a un'attenta analisi, deve subire un pre-processing affinché il dataset pronto all'elaborazione sia pulito e accurato. Il pre-processing è costituito da tre fasi differenti:

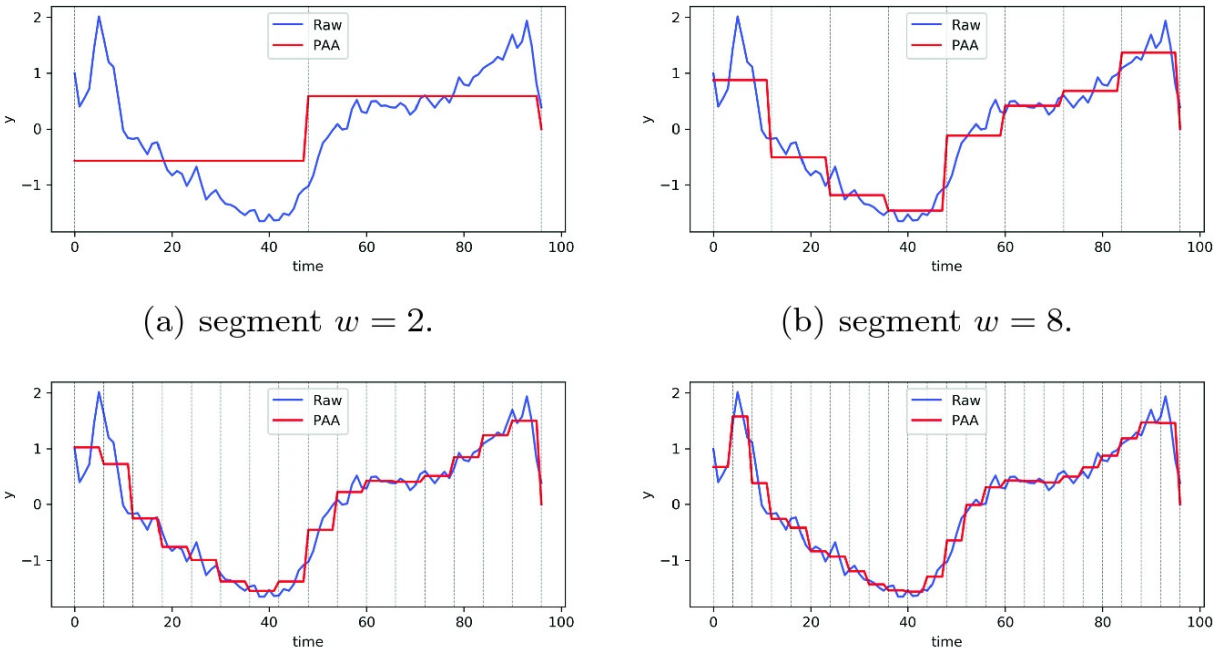
1. Data Reduction;
2. Inviduamento ed eliminazione degli outlier;
3. Sostituzione dei missing value.

2.1.1 Riduzione Dati

Come si vedrà nel capitolo 3 che descrive il dataset, la frequenza del campionamento delle misure è molto piccola e diversa tra le grandezze di misura, dunque, al fine di rendere

uniforme e ridurre la dimensione del dataset su cui lavorare, è stata effettuata una riduzione dei dati, è stata fatta una riduzione dei dati utilizzando il metodo Piecewice Aggregate Approximation (PAA). Questo algoritmo consiste nel dividere una serie temporale in intervalli di stessa lunghezza e non sovrapposti nel tempo e sostituirli con la media di tutti i punti dell'intervallo [16].

Quindi una serie temporale $C = \{c_1, c_2, ..., c_n\}$ dopo l'applicazione del metodo è descritto in $\bar{C} = \{\bar{c_1}, \bar{c_2}, ..., \bar{c_n}\}(1 \leq m \leq N)$, dove $\bar{c_i}$ è la media dell'intervallo i-esimo.



(a) segment $w = 2$.

(b) segment $w = 8$.

(c) segment $w = 16$.

(d) segment $w = 24$.

Figura 2.2: Esempio di applicazione di PAA. Fonte da [17]

Nella figura 2.2 viene riportato un esempio di applicazione di tale metodo.

Adottare una tecnica di riduzione di dati comporta ovviamente una perdita di alcune informazioni, però con il vantaggio di avere un dataset meno ingombrante e quindi un costo computazionale minore.

2.1.2 Individuazione Outlier

La definizione di outlier secondo [18] è un osservazione (o una serie di osservazioni) che risulta significativamente differente dal resto del dataset.

Vengono definiti secondo i seguenti 2 assunzioni:

- sono outliers se differiscono dai valori normali per le loro proprietà;
- rispetto ai valori normali gli outliers sono infrequenti all'interno del dataset.

È possibile avere 2 tipologie di outlier: gli outliers puntuali e sequenze di outlier. Dopo che sono stati individuati, questi valori vengono eliminati e trattati come missing value.

Nel caso studio è stato adottato il metodo Z-score, che utilizza 2 indicatori per determinare gli outlier, la media e la deviazione standard:

$$Z_i = \frac{x_i - \bar{x}}{sd} \tag{2.1}$$

dove $\bar{x}$ è la media e sd è la deviazione standard dei valori. Si è calcolato la media e la deviazione standard per i valori raggruppati per il giorno e la zona di appartenenza, dopodiché quelli che hanno il valore assoluto del Z-score $|Z_i|$ superiore a 6.5 sono stati considerati outliers.

2.1.3 Sostituzione Missing value

I metodi utilizzati per trattare i missing value sono differenti a secondo della tipologia:

- se il missing value deriva dall'eliminazione di un outlier puntale è stato fatto uso dell'interpolazione lineare;
- se si tratta di una sequenza di missing value è stato adottato il metodo K-Nearest Neighbors (KNN). Grazie alla sua facilità di utilizzo, KNN è uno dei metodi più impiegati per l'imputazione dei NA. Il valore di sostituzione è ottenuto facendo la media dei k punti più vicini, trovati attraverso la distanza Euclidea tra attributi del missing value e quelli dell'intero dataset [19]:

$$d_{(x,y)} = sqrt(\sum_{j=1}^{s}(x_j - y_j)^2) \tag{2.2}$$

Dove:

- $d_{(x,y)}$ è la distanza Euclidea;
- s è la dimensione del dataset;
- j è attributo dei dati, con j=1,2,3...;
- x_{aj} valore dal j-attributo contenente missing value;
- y_{bj} valore dal j-attributo contenente intero dataset.

Il valore di k influenza l'accuratezza del metodo, infatti se viene scelto un k troppo piccolo può causare rumori, invece un k troppo grande può limitare il valore e di conseguenza peggiorare l'accuratezza[20].

In questo caso gli iper-parametri sono stati impostati nella seguente maniera per la ricostruzione delle serie temporali:

- k pari a 40;
- attributi: mese, giorno della settimana, orario e tre parametri esterni (la temperatura, l'umidità relativa e la radiazione solare).

2.2 Caratterizzazione Profili temporali giornalieri

Questa sezione si occupa dell'analisi di profili temporali giornalieri di temperatura. In particolare viene eseguito un clustering di questi profili temporali, da cui sono state estratte le informazioni riguardanti i profili tipici di temperatura interna che si verificano nell'edificio e i punti spaziali in cui si presentano.

In seguito viene applicato una tecnica dell'albero di classificazione, con lo scopo di comprendere le variabili indipendenti dei profili giornalieri dominanti e quindi apprendere le occasioni in cui si verificano.

Le conoscenze che sono ottenute attraverso queste tecniche di studio sono fondamentali per la successiva analisi predittiva, poiché decidono la qualità dei risultati di tale analisi.

2.2.1 Clustering

Il clustering permette di creare gruppi di elementi molto simili tra di loro, e nel caso studio gli elementi di similarità sono delle sotto-sequenze dei profili giornalieri di temperatura interna, analizzati attraverso il metodo di sotto sequenze delle serie temporali e l'algoritmo di k-means e la distanza euclidea.

Il Clustering è uno dei metodi più utilizzati nell'analisi statistica e nel data mining. Consiste nel raggruppare gli elementi più simili in un cluster attraverso dei criteri di confronto, perciò il compito del clustering è creare dei clusters con la massima similarità all'interno del cluster e la massima diversità tra i clusters.

I metodi di clustering sono classificati in 5 principali gruppi:

1. Clustering gerarchico: crea una gerarchia annidata di clusters in relazione alla matrice di distanza tra coppie di oggetti;

2. Clustering partitivo: indicando k, il numero di gruppi da ripartire gli oggetti, si creano clusters attraverso un processo iterativo, tra questi il più conosciuto è k-means ;

3. Density based clustering: è impiegato per identificare clusters distintivi sulla base del fatto che un cluster viene considerato una regione se la densità degli oggetti supera una certa soglia. DBSCAN e OPTICS sono due metodi tipici di questo tipo;

4. Model based clustering: questo tipo di clustering presuppone che i dati siano generati da un modello matematico e cerca di ritrovare tale modello. Il clustering segue principalmente 2 approcci, l'approccio statistico e l'approccio di neural network;

5. Grid based clustering: impiega una struttura di dati con griglia a multi-risoluzione, cioè una struttura a griglia costruita considerando lo spazio come un numero finito di celle. Attraverso tale struttura vengono eseguite tutte le operazioni di clustering.

Inoltre, secondo [21] il clustering delle serie temporali può essere raggruppato in 3 categorie:

1. Clustering delle serie temporali intere: il clustering viene applicato direttamente alle serie temporali nella sua interezza, come un oggetto discreto, calcolando la similarità;

2. Clustering delle serie temporali in sotto sequenze: attraverso una finestra scorrevole si attua il clustering al tratto considerato;

3. Clustering dei punti temporali: consiste nel considerare la prossimità nel tempo dei punti temporali e la similarità dei valori corrispondenti.

Nel caso studio analizzato è stato fatto un clustering delle sotto-sequenze con il metodo k-means e la distanza Euclidea.

Innanzitutto la similarità tra gli oggetti può essere determinata attraverso varie tecniche, le più utilizzate sono la distanza Euclidea e il coefficiente di correlazione di Pearson. La scelta della metrica utilizzata può avere un grande impatto sul risultato, in quanto condiziona quali oggetti sono simili tra di loro. In questo caso la distanza Euclidea si calcola con l'equazione 2.2.

Mentre l'algoritmo K-means appartiene al gruppo di clustering partitivo con l'obiettivo di raggruppare gli oggetti di un dataset in k gruppi (clusters).

Il processo dell'algoritmo si articola in diversi steps:

1. viene scelto un numero ottimale di clusters: k;

2. vengono selezionati casualmente k oggetti come centroidi;

3. tutti gli altri punti vengono assegnati ai clusters in relazione alla distanza Euclidea tra oggetto da assegnare e i centroidi;

4. si calcola nuovamente il centroide o la media dei punti di ciascun cluster;

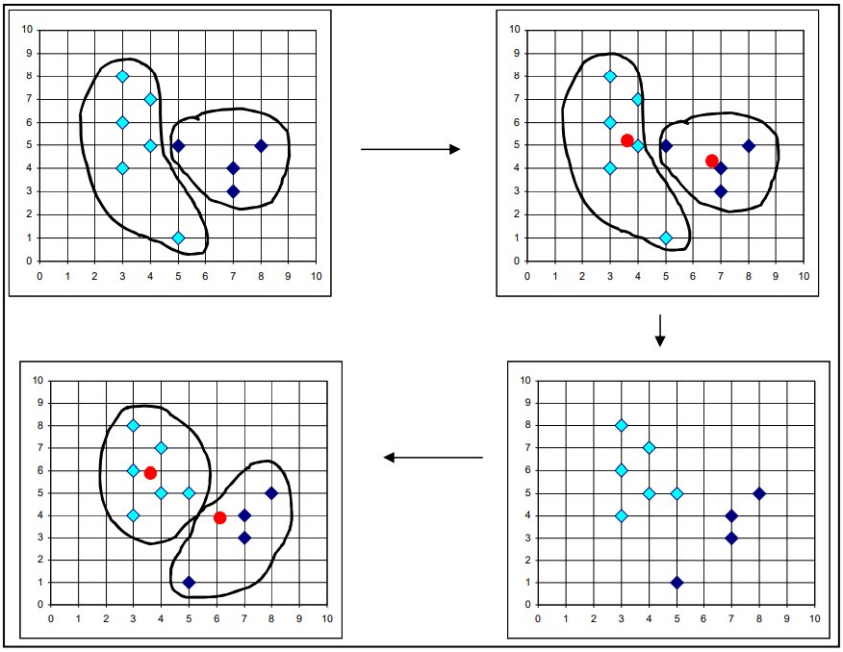


Figura 2.3: Esempio di K-means. Fonte da [22]

5. si ripete il punto 3 e 4 finché i centroidi non si spostano più.

Nel caso studio i profili temporali giornalieri vengono divisi in 3 sotto sequenze da 8 ore imponendo k pari a 3.

Dopo un primo clustering, per aumentare la similarità all'interno dei clusters dominanti ed eliminare i profili più distanti è stato utilizzato il clustering gerarchico con il single linkage.

Il clustering gerarchico può essere realizzato con 2 approcci diversi: partire da oggetti singoli per poi formare un cluster unico (approccio agglomerativo), oppure partire da un cluster unico per poi essere suddiviso in diversi sotto-clusters (approccio divisivo). Oltre al metodo di calcolo di similarità anche il criterio di collegamento è fondamentale per il clustering gerarchico, dato che influenza la forma del clustering:

- complete linkage:

$$max\{d(a,b) : a \in A, b \in B\}$$

- single linkage:

$$min\{d(a,b) : a \in A, b \in B\}$$

- average linkage:

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$$

2.2.2 Classification and Regression Tree

L'albero di classificazione serve per comprendere sia le variabili indipendenti che caratterizzano i clusters, sia le circostanze in cui accadono. Per questo motivo è stato fatto un CART mettendo le variabili indipendenti del tipo temporale, spaziale ed esogeno, e poi calcolo delle metriche della variabile categorica per la qualità dell'albero ottenuto.

Il Classification and Regression Tree (CART) è un algoritmo di machine learning progettato per riconoscere determinate caratteristiche di un insieme di elementi eterogenei e per costruire un modello di previsione.

L'albero di classificazione è designato per variabili dipendenti categorici mentre l'albero di regressione è progettato per variabili dipendenti numerici.

Il modello si costruisce facendo la ripartizione binaria ricorsiva partendo da un nodo padre che contiene tutti gli elementi, una volta ripartizionato si ottengono 2 nodi figli. L'obbiettivo della ripartizione è creare partizioni più pure possibili, l'impurità viene misurata con indice di Gini [23, 24]:

$$GI(D) = 1 - \sum_{i=1}^{n} P_i^2 \tag{2.3}$$

dove, $P_i = \frac{[S_i]}{[S]}$ è il rapporto tra il numero di tuples esistente di una certa classe rispetto al numero totale di tuple esistente in D.

L'indice di Gini di ripartizione binaria per una variabile indipendente t è calcolato con la seguente formula:

$$GI_t(D) = \sum_{i=1}^{2} \frac{[D_i]}{[D]} GI(D_i) \tag{2.4}$$

dove, D_i è l'indice di Gini per una suddivisione. Infine il calcolo di riduzione d'impurità per una suddivisione binaria per una variabile indipendente t è:

$$GI_{red}(t) = GI(D) - GI_t(D) \tag{2.5}$$

Il valore di Gini è compreso tra 0.5 e 0.1, la variabile predittiva che ha minimo valore di Gini viene scelto per la ripartizione.

Il risultato grafico del modello è un albero di decisione con struttura gerarchica, come mostrato nella figura 2.4. Il compito di questo albero è creare un modello di descrizione di

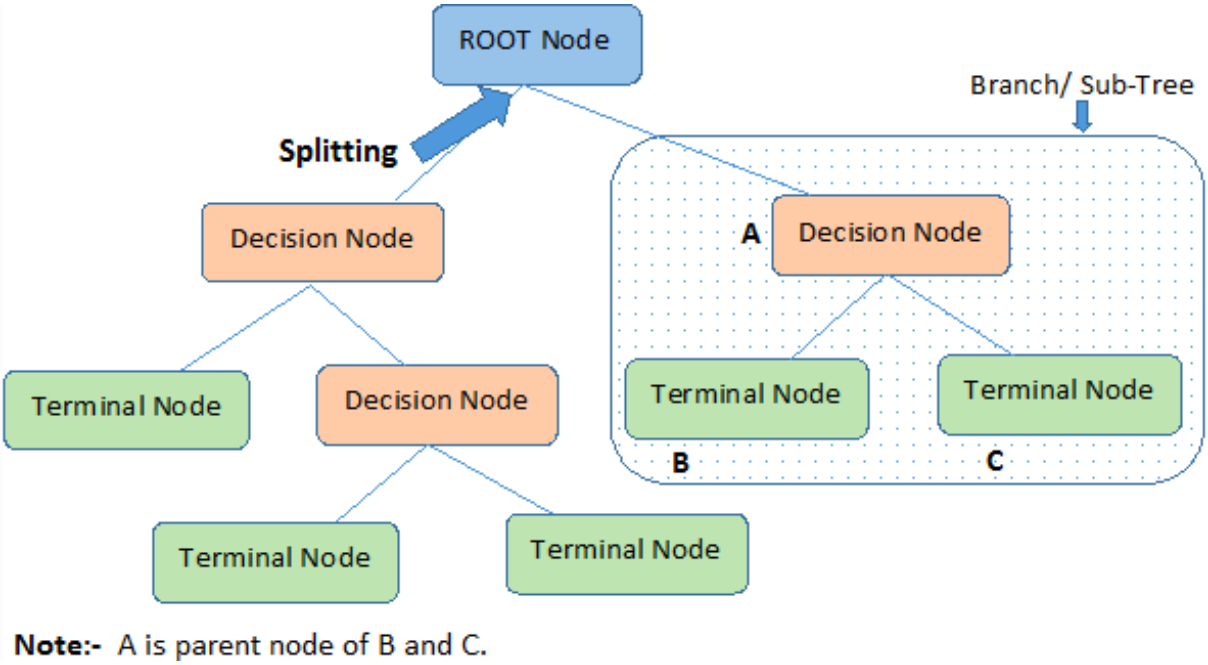


Figura 2.4: Esempio di un albero di decisione. Fonte da [25]

variabili indipendenti utili in funzione del cluster di appartenenza avendo come input la temperatura media, massima, minima giornaliera, l'umidità relativa media giornaliera e la radiazione solare media giornaliera dell'ambiente esterno, il mese, il giorno della settimana e la posizione della zona.

Per rendere il modello creato più robusto, è stata scelta la convalida incrociata come il metodo di validazione. La cosiddetta k-fold consiste nel suddividere il dataset in k parti uguali, dopodiché a ogni iterazione viene fatto l'addestramento sul dataset esclusa dalla k-esima parte, e la convalidazione sulla k-esima parte del dataset.

Nel caso studio è stato utilizzato k pari a 10.

Le metriche utilizzate per valutare la qualità dei risultati ottenuti dalla previsione di tipo classificativo sono:

- Accuratezza indica la frazione di valori predetti correttamente sul numero totale di predizione

$$\text{Accuratezza} = \frac{\text{veri positivi} + \text{veri negativi}}{\text{veri pos.} + \text{veri neg.} + \text{falsi pos.} + \text{falsi neg.}} \tag{2.6}$$

- Precisione specifica la frazione di valori realmente positivi su tutti i valori predetti positivi.

$$\text{Precisione} = \frac{\text{veri positivi}}{\text{veri positivi} + \text{falsi positivi}} \tag{2.7}$$

- Richiamo determina la frazione di valori predetti correttamente positivi su tutti valori realmente positivi.

$$\text{Richiamo} = \frac{\text{veri positivi}}{\text{veri positivi} + \text{falsi negativi}} \tag{2.8}$$

Le variabili che sono contenute all'interno dei nodi decisionali dell'albero di classificazione hanno permesso di capire che la temperatura interna è fortemente influenzata dallo spazio, quindi ci permette di sfruttare le informazioni spaziali per la costruzione della matrice di adiacenza e di conseguenza del grafo, elemento fondamentale del GNN.

2.3 Analisi predittiva

Questo paragrafo si concentra sull'analisi predittiva di temperatura interna di 3 passi in avanti attraverso 2 algoritmi di Artificial Neural Network (ANN), Long Short-Term Memory Neural Network e Spatial–Temporal Graph Neural Networks (STGNN). Il dataset di studio comprende la temperatura interna di diverse posizioni, in particolare si hanno di 63 sensori, localizzati in 2 piani come già accennato nel capitolo dell'introduzione. Quindi per effettuare la previsione della temperatura a 3 passi in avanti in tutti questi punti è stato molto utile l'algoritmo GNN che esegue in contemporanea un elevato numero di modelli di previsione.

Infatti è stato utilizzato lo STGNN [14] che considera simultaneamente sia la dipendenza temporale che quella spaziale. La relazione legata allo spazio viene introdotta attraverso la matrice di adiacenza. Dopo aver studio con la matrice spaziale, è state create anche le matrici di adiacenza attraverso approccio di similarità.

I risultati del GNN sono poi confrontati con quelli di un LSTM, per dimostrare il vantaggio di adoperare con GNN.

Il Artificial Neural Network è una famiglia di algoritmi della disciplina di Machine Learning, costituito dai neuroni artificiali. Un neurone artificiale riceve dei input e genera un singolo output per poi essere eventualmente passato a un altro neurone per una ulteriore elaborazione:

$$y = g(wx + b) \tag{2.9}$$

Dove y e x sono l'output e l'input del neurone, w è il peso dell'input calcolato da una funzione di attivazione, b è il termine bias.

L'obbiettivo è apprendere la relazione nascosta tra l'input e l'output, in modo da poter essere usufruito per la previsione dell'output avendo solamente l'input. Il processo quindi si divide in 2 passaggi, la fase di allenamento/training, quella di apprendimento, portata avanti con il compito di minimizzare una funzione di errore, e la fase di testing, che verifica la qualità del modello allenato attraverso delle metriche come MSE, RMSE, MAPE, MAE. L'intento di questa analisi è poter predire i valori di temperatura interna dell'edificio. A tal fine è stato adoperato un dataset di 2 anni, registrato dal 2019 al 2020, che viene diviso in 2 parti, la porzione del dataset raccolto nel 2019 viene adottato per la fase di training e la seconda parte per il testing.

Nella fase di training è imposto RMSProp come algoritmo di ottimizzazione e MSE come funzione di errore da minimizzare.

Per quanto riguarda le metriche adottate sono:

- MSE è un indicatore che misura la media dei quadrati di errore:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2 \tag{2.10}$$

- MAE è la media degli errori in valore assoluto:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.11}$$

- MAPE è la media degli errori relativi in valore assoluto:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2.12}$$

Dove y è il valore predetto e $\hat{y}$ è il valore reale.

2.3.1 Baseline - Long Short Term Memory Network

LSTM è un algoritmo di Artificial Neural Network costruito sulla base di RNN risolvendo il problema dell'esplosione/scomparsa del gradiente durante il processo di apprendimento della dipendenza a lungo termine, rendendolo adatto anche per i problemi con il lag temporale molto lungo [26, 27]. Tale problema viene superato aggiungendo constant error carousel (CEC), che produce un segnale di errore in ogni unità di cella.

Un'unità vanilla LSTM è costituita da una cella, un input gate, un output gate e un forget gate, i 3 gate regolano il flusso di informazione che entra nella cella per essere ricordato per un intervallo di tempo arbitrario.

Il CEC viene esteso all'input gate e output gate, formando il blocco di memoria, ciò permette di evitare possibili conflitti di aggiornamento dei pesi. La struttura di un LSTM è composta da diversi blocchi di memoria.

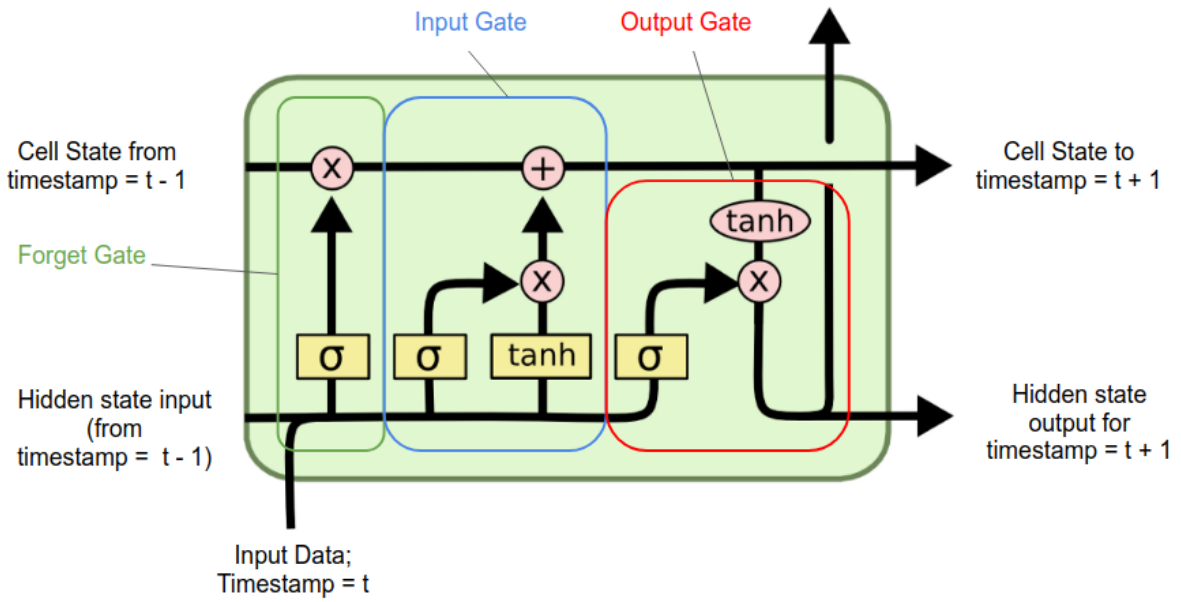


Figura 2.5: La struttura di un LSTM. Fonte da [28]

2.3.2 Graph Neural Network

Graph Neural Network (GNN) è una classe del Artificial Neural Network che opera sui dati raffigurati come grafo. L'algoritmo modella sull'input e/o output attraverso le informazioni contenute all'interno del grafo in relazione ai collegamenti tra i nodi.

Il grafo, illustrato in figura 2.6, è una rappresentazione di un insieme di entità, chiamati nodi (evidenziati in giallo in figura 2.6), collegati tra di loro a seconda della relazione che intercorre tra essi. Tale collegamento è chiamato edge, mentre con master node si intende l'insieme delle informazioni relative a tutti nodi.

Il grafo è l'elemento più importante del GNN, infatti a differenza di altre tipologie di ANN, il grafo permette sia di creare contemporaneamente diversi modelli di previsione garantendo dei risultati ottimali, sia il passaggio di informazioni tra i nodi che condividono caratteristiche simili.

Inoltre, viene definito come un potente strumento, perché non solo può modellare oggetti relativamente semplici come i social network ma anche testi e immagini. Il grafo è creato per mezzo della matrice di adiacenza, che in generale è una matrice binaria quadrata. Le

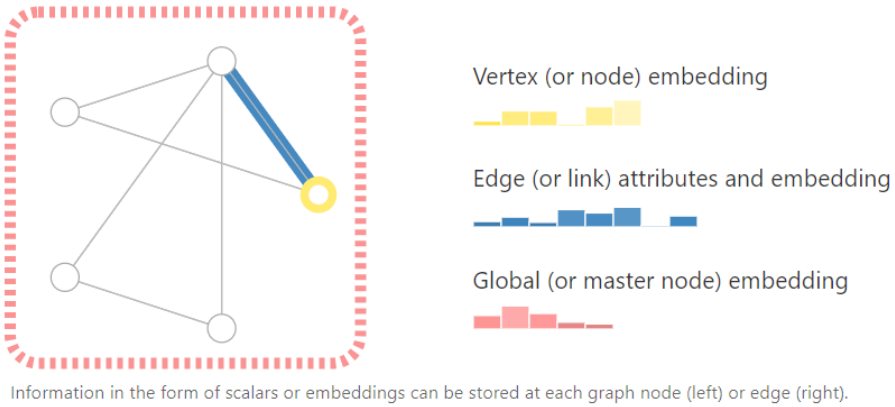


Figura 2.6: Elementi di un grafo. Fonte da [29]

indici delle righe e delle colonne della matrice sono riferite ai nomi dei nodi del grafo e nella cella (i, j) viene inserito 1 se esiste collegamento tra i nodi i e j, oppure 0 se tale collegamento non sussiste.

Nel caso studio sono state utilizzate 2 approcci per la costruzione della matrice di adiacenza:

- approccio spaziale: le matrici di adiacenza spaziali sono costruiti in base alla vicinanza spaziale dei sensori attraverso criteri di vicinanza.

  Non avendo la distanza tra i vari sensori, ma solo la mappa di sensori, è fatto compilata la matrice manualmente, sensore per sensore.

- approccio di similarità: le matrici sono realizzate attraverso l'algoritmo KNN e il metodo di Pearson correlation in base alla similarità delle serie temporali dei sensori. In particolare è stato deciso di realizzare le matrici attraverso il KNN che analizza le serie temporali e trova un certo numero imposto di punti più vicini a ciascun sensore.

  Sulla base della matrice creata con KNN vengono aggiunti altri punti con gli indici di correlazione di Pearson superiore al valore imposto.

  La approccio consente di creare la matrice di adiacenza in modo automatico impostando semplicemente i parametri degli algoritmi.

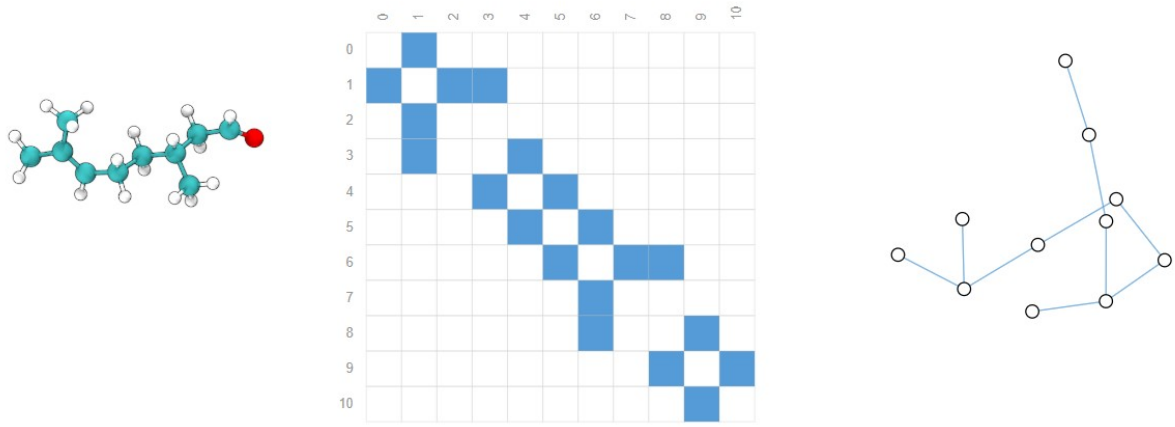


Figura 2.7: Esempio di una matrice di adiacenza(in centro) e del grafo (a sinistra) a partire da una struttura molecolare. Fonte da [29]

Inoltre un grafo può essere direzionale o non orientato, ciò dipende dalla direzionalità degli edges, infatti nel primo caso gli edges sono puntati da un nodo all'altro, mentre nel secondo caso gli edges sono bidirezionali.

In base alla tipologia dell'output che si vuole ottenere dal modello si distinguono 3 tipi di attività di previsione:

1. a livello di nodi: l'obbiettivo è la previsione di una certa proprietà per ciascun nodo oppure la classificazione dei nodi in diverse classi;
2. a livello di edge: il compito è classificare la tipologia degli edges oppure la previsione dell'esistenza dell'edge tra i nodi;
3. a livello di grafo: l'intento è la previsione di una caratteristica del grafo.

I GNN comunemente contengono i seguenti moduli computazionali[30], non necessariamente in contemporanea:

- Modulo di propagazione: il modulo ha la funzione di propagare informazioni aggregati ai nodi. Nel modulo di propagazione l'operazione skip connection raccoglie le informazioni storiche dei nodi e attenua il problema di over-smoothing, in seguito le informazioni che riguardano le caratteristiche e le informazioni topologiche dei nodi vicini vengono aggregati dall'operatore di convoluzione oppure dall'operatore ricorrente.

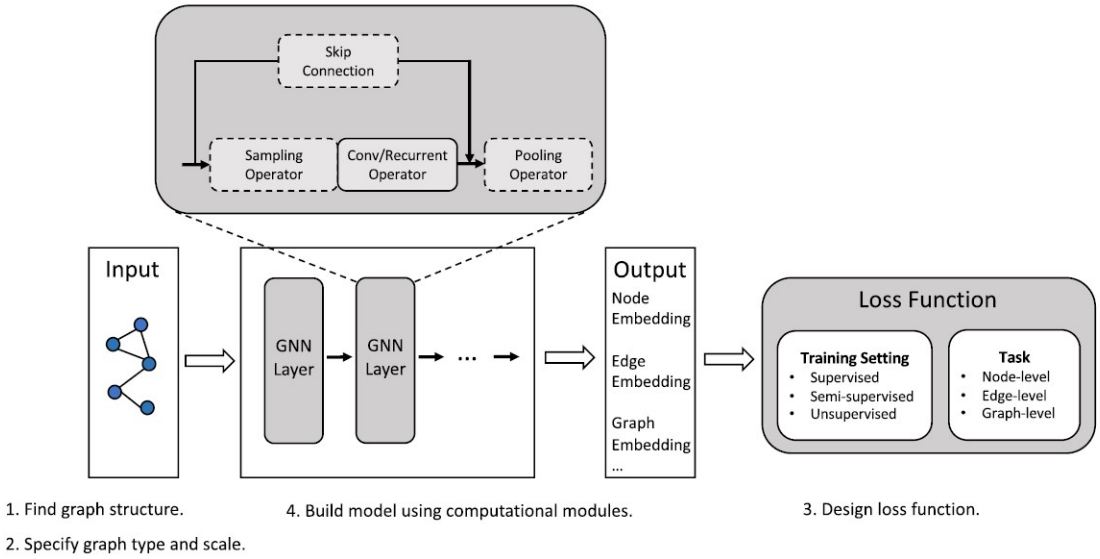


Figura 2.8: Pipeline di un GNN generico. Fonte da [30]

- Modulo di campionamento: è spesso combinato con il modulo di propagazione, in quanto solitamente gestisce la propagazione, in particolare quando il grafo è gigante.
- Modulo di pooling: utile per estrarre l'informazione dai nodi quando si desidera di ottenere rappresentazioni di alto-livello dei sotto-grafi o dei grafi.

Quindi riassumendo, il passaggio di informazioni all'interno del GNN avviene in 3 stadi:

1. Preparazione dati: le caratteristiche dei nodi input vengono elaborati per produrre un messaggio, tale elaborazione può essere semplicemente una trasformazione lineare.
2. Aggregazione di informazioni: le informazioni dei nodi che sono collegati al nodo di riferimento vengono aggregati rispetto a un peso attraverso un'operazione di pooling, per preparare un messaggio aggregato da passare al nodo di riferimento. Tale passaggio viene compiuto per ciascun nodo in contemporanea.
3. Aggiornamento: vengono elaborate le rappresentazioni dei nodi e il messaggio aggregato per creare una nuova caratteristica dei nodi.

Secondo [14] GNN si divide in 4 categorie:

- Recurrent Graph Neural Networks (RecGNN): ha una struttura neurale ricorrente, perciò i nodi del grafo scambiano continuamente le informazioni finché non viene raggiunto un equilibrio solido;

- Convolutional Graph Neural Networks (ConvGNN): un modello derivante da CNN, infatti ConvGNN applica operazione di convoluzione sui dati di un grafo invece dei dati di una griglia. A differenza di RecGNN, ConvGNN si serve di tanti layers convoluzionali per ricavare delle rappresentazioni di alto livello dei nodi;

- Graph Autoencoders: è un modello di apprendimento non supervisionato con la fine di ricostruire il grafo con informazioni derivanti dalla codifica dei nodi o del grafo;

- Spatial–Temporal Graph Neural Networks: l'obbiettivo del modello è considerare la dipendenza spaziale e temporale in contemporanea e apprendere il pattern nascosto. Spesso si tratta d'integrare ConvGNN che tiene conto della dipendenza spaziale con RNN o CNN che considera la dipendenza temporale.

Spatio-Temporal Graph Neural Network

Il modello utilizzato nello studio si basa sul lavoro compiuto da [31] e in particolare il codice pubblicato online.

Si tratta di una combinazione di 2 modelli, ossia il Convolutional Graph Neural Networks (ConvGNN) e il Long Short-Term Memory Neural Network (LSTM). Infatti, come caratteristica del STGNN, il LSTM analizza le caratteristiche temporali, mentre il ConvGNN tiene conto degli attributi spaziali.

In un primo momento viene applicato ConvGNN al grafo ricavato dalla matrice di adiacenza. Il risultato del ConvGNN viene successivamente passato all'algoritmo LSTM per ottenere l'output finale.

Nonostante la stretta relazione tra i ConvGNN e i RecGNN, i primi si distinguono per la struttura dei layer, infatti sono costituiti con un numero finito di layer, ciascuno con pesi differenti.

I ConvGNN si dividono in 2 categorie:

- Spectral based: introduce il filtro nel processo di elaborazione dei segnali e opera nel dominio dei spettri;

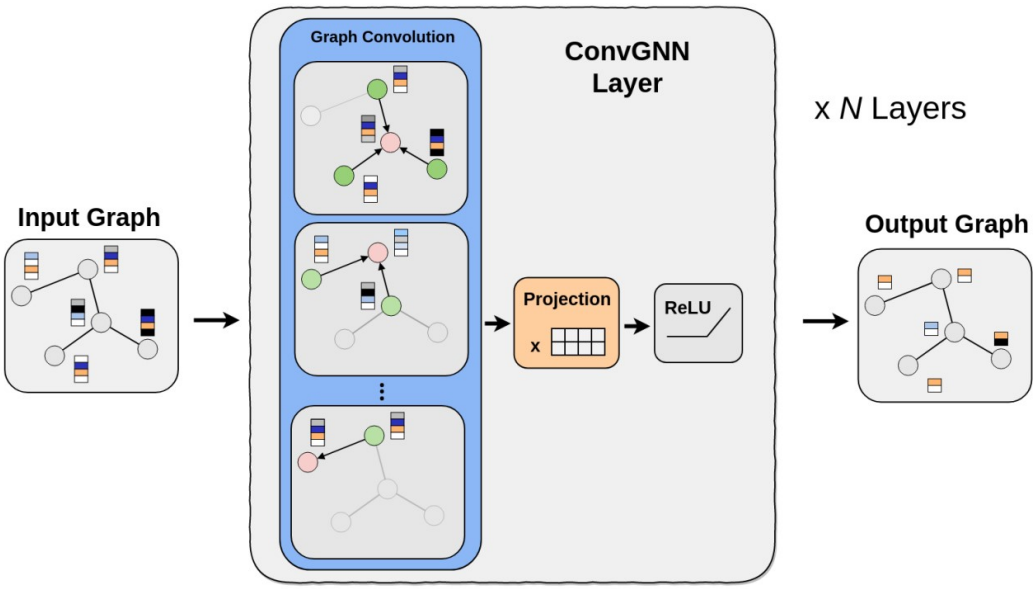


Figura 2.9: Struttura tipica di un ConvGNN. Fonte da [32]

- Spatial based: applica direttamente la convoluzione sul grafo senza un'ulteriore trasformazione.

Il modello di studio nello specifico è costituito da un convolutional layer e due lstm layer. Il primo è costituito da 16 unità, mentre il layer di LSTM è composto da 64 unità e ReLu come funzione di attivazione. Tale modello riceve come input una sequenza di 12 timestep producendo a sua volta una sequenza di valore di previsione.

Capitolo 3

Caso studio

Il dataset utilizzato in questo studio è stato fornito da Luo et al.[33], che hanno riportato il dataset nel loro articolo della rivista intitolato "A three-year dataset supporting research on building energy management and occupancy analytics". Si tratta di un dataset contenente il monitoraggio triennale, cioè dal 2018 al 2020, dell'edificio a uso ufficio costruito nel 2015 a Berkeley, in California.

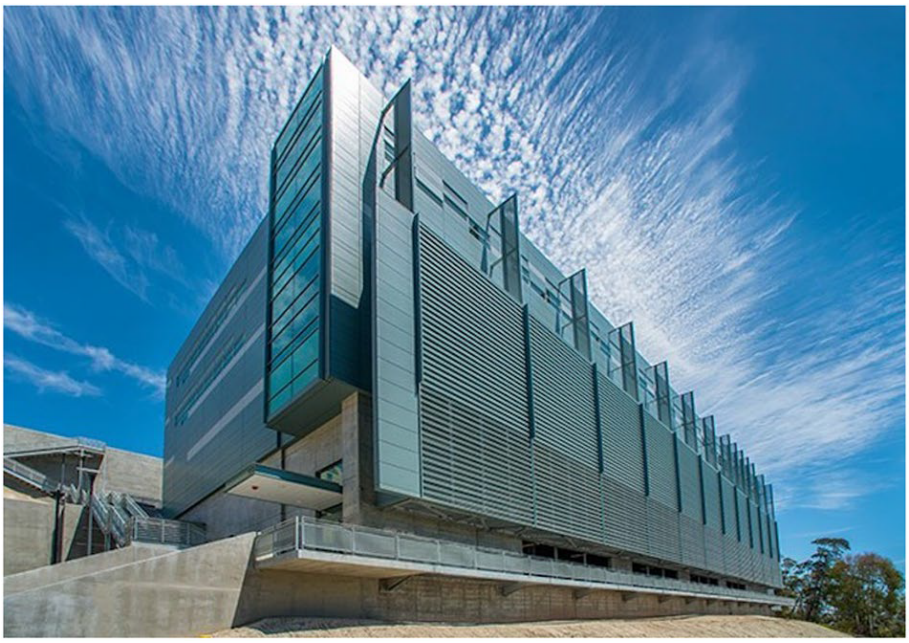


Figura 3.1: L'edificio ad uso ufficio a Berkely,California. Fonte da [33]

3.1 Descrizione dell'edificio

L'edificio, collocato all'interno del campus Lawrence Berkeley National Laboratory (Berkeley Lab), ha un superficie totale di 10.400 m^2, la quale è distribuita equamente su 4 piani.

Nonostante l'edificio è ad uso ufficio, solo il terzo e il quarto piano vengono utilizzati a tale finalità, poiché al piano terra sono presenti dei sistemi meccanici e al secondo piano è operativo il National Energy Research Scientific Computing Center (NERSC). I due piani ad uso ufficio, ciascuno con una superficie di 2.325 m^2, sono strutturati in maniera diversa, infatti, mentre al piano inferiore ci sono principalmente degli uffici chiusi, al piano superiore gli uffici sono principalmente open space.

Il moderno edificio, costruito nel 2015, ha una struttura in telaio ad acciaio integrato da un sistema di facciate esterne con finestre e isolante in schiuma inserito nell'intercapedine. La struttura è riparata dalle radiazioni solari attraverso delle frangisole verticali posizionate all'esterno dell'edificio.

La copertura dell'edificio è di cemento con degli strati di isolante e una membrana mono strato di PVS su un pannello di copertura da 1/2".

Per isolare termicamente il primo piano ad uso ufficio dall'area di calcolo ad alte prestazioni è stato aggiunto uno strato di isolante R30 nel solaio tra i 2 piani, a differenza, tra i due piani a uso ufficio è stato lasciato uno spazio dedicato al plenum per il sistema di riscaldamento e di raffrescamento.

Infatti nelle aree adibite a uffici, è stato realizzato un pavimento rialzato nel punto in cui termina la moquette per dare spazio al plenum per il sistema di distribuzione dell'aria a pavimento.

Gli autori hanno chiamato in modo diverso le zone che utilizzano strumenti di misura differenti, infatti le zone esterne sono quelle che hanno pareti esterne e finestre, con un totale di 57 zone in 2 piani, mentre le restanti sono cosiddette zone interne.

3.2 Descrizione del sistema di riscaldamento e raffrescamento

Il riscaldamento e il raffrescamento degli uffici è garantito da un sistema di distribuzione ad aria a pavimento (UFAD) attraverso quattro unità di roof-top (RTU) poste al tetto dell'edificio.

L'aria di mandata della RTU viene inviata alle zone interne ed esterne attraverso i diffusori a pavimento, mediante i plenum posti al livello di pavimento.

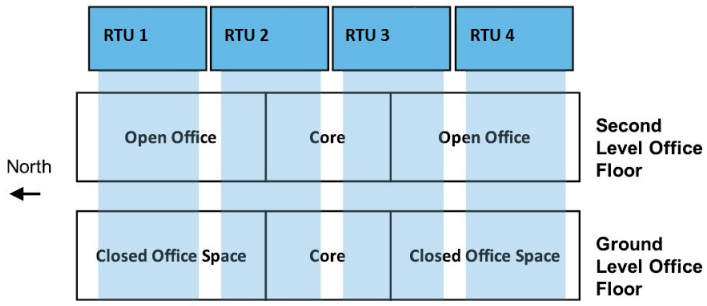


Figura 3.2: L'area di servizio di ciascun RTU

Ciascun RTU presta il servizio a uno quarto dell'area di entrambi piani, come mostrato nella figura 3.2, questi 4 RTU avviano i loro ventilatori di alimentazione alla stessa velocità, piuttosto che riferirsi ai sensori e al set-point delle zone. I ventilatori di mandata e di ritorno sono dotati di azionamenti a velocità variabile.

Per il riscaldamento dell'ambiente, oltre a RTU sono presenti 50 unità di terminali azionati dai ventilatori (UTF) con una batteria di riscaldamento ad acqua, che nel caso di necessità riscalda aria nella zona perimetrale adiacente all'esterno. L'aria riscaldata dei UTF sono prodotta da una pompa di calore supportata da 2 pompe di potenza minore a frequenza variabile. Riguardo al raffrescamento, ciascun RTU è dotato di 2 compressori scoll con requisiti per il controllo della velocità variabile dal 10% al 100%.

L'edificio è equipaggiato di un sistema di gestione degli edifici (BMS) WebCTRL di Automated Logic (ALC) (Automated Logic 2017) con una vasta gamma di sensori che gestisce anche i sistemi di HVAC. I sensori e i controllori del BMS sono collegati in rete a un server protetto da firewall.

3.3 Descrizione di sensori

In ciascuna zona esterna sono installati dei sensori a parete che fanno parte del sistema di automazione dell'edificio (BAS) e misurano la temperatura di tale ambiente. Invece, nelle zone interne sono installati dei sensori di temperatura costituiti da Raspberry Pi Zero W e DS18B20 Digital Temperature Sensors, aggiunti dal team di ricerca allo scopo di raccogliere i dati. Sono in tutto 16 sensori installati a livello di scrivania e il più vicino possibile agli occupanti.

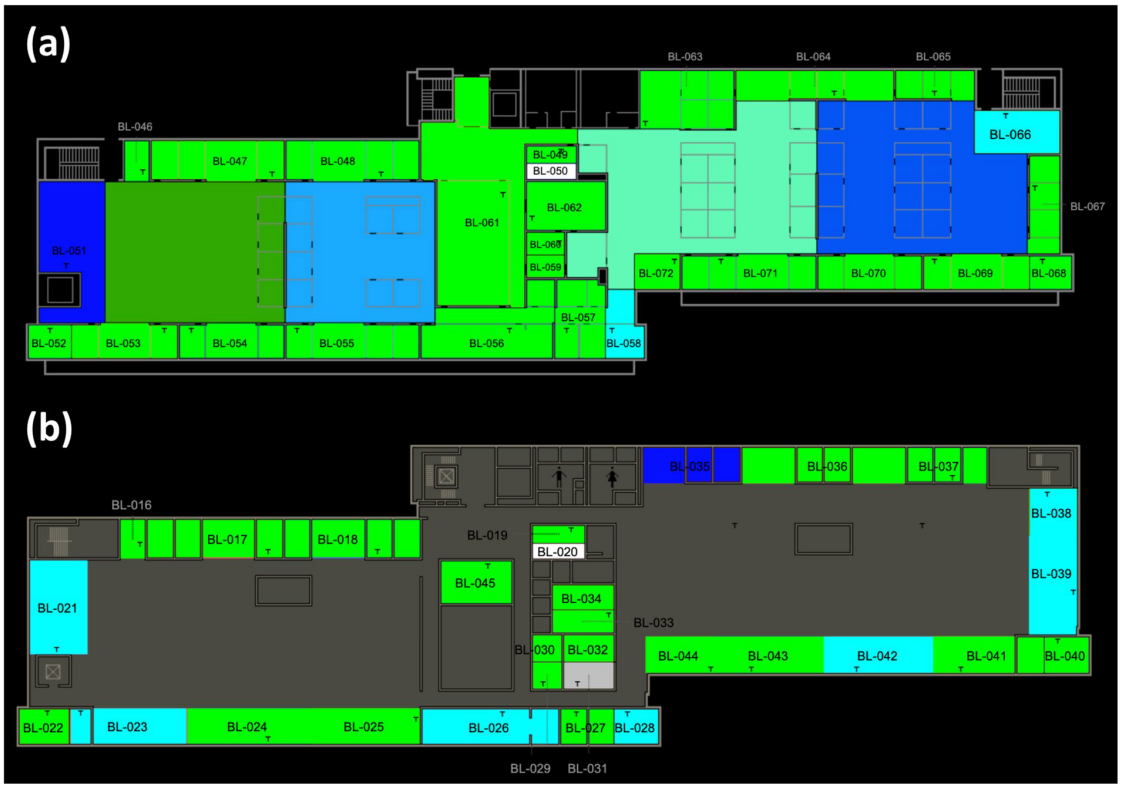


Figura 3.3: Posizione delle zone termiche esterne in due piani a uso ufficio: (a) Primo piano e (b) Secondo piano. Fonte da [33]

Oltre ai sensori di temperatura hanno aggiunto 6 sensori del TRAF-SYS posti negli ingressi dell'ala sud dell'edificio per misurare il numero degli occupanti.

I dati climatici raccolti per l'edificio sono stati collezionati attraverso Synoptical-Labs(MesoWest and SynopticLabs 2017) da una stazione meteorologica all'interno del campus di Berkeley Lab, con una distanza di circa 300 m dall'edificio.

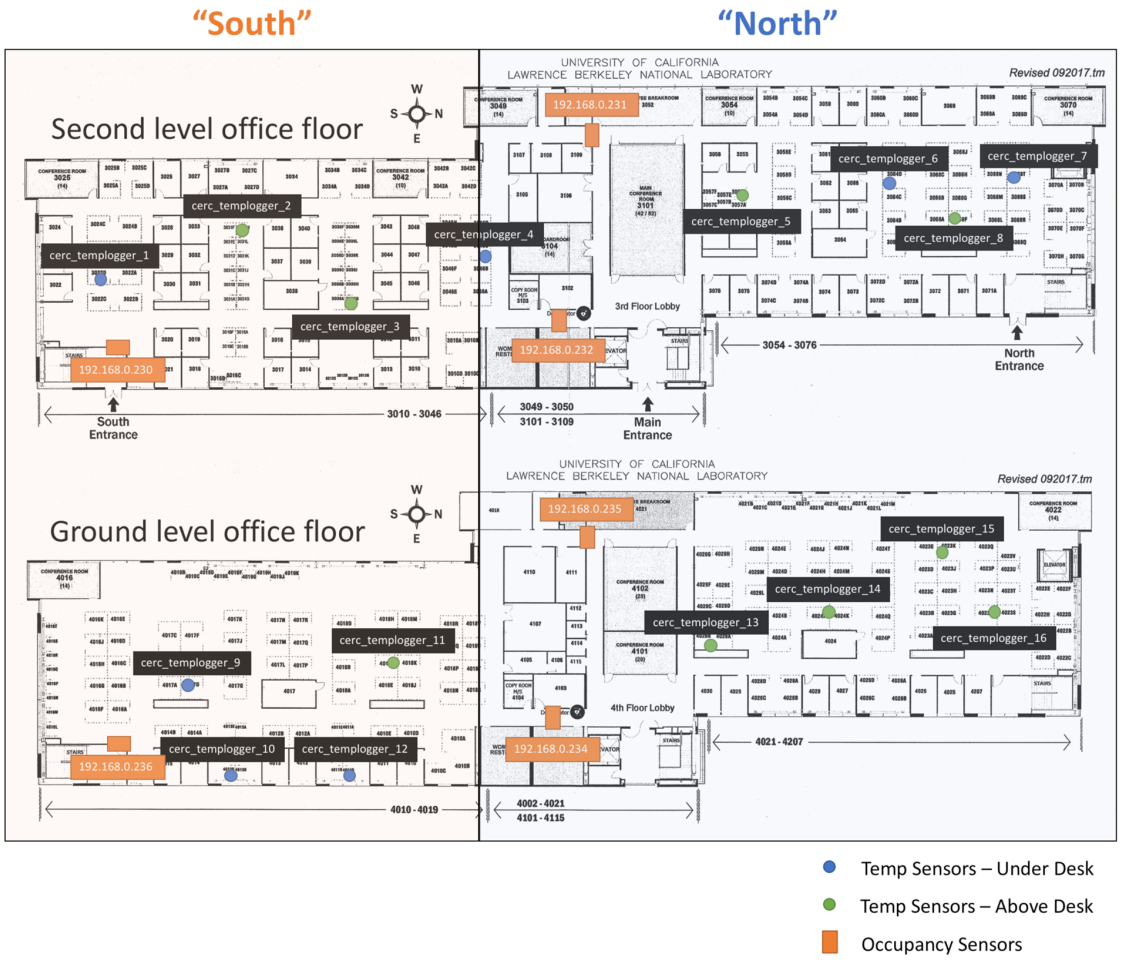


Figura 3.4: Posizione dei sensori di temperatura nelle zone interne e sensori di occupazione. Fonte da [33]

Inoltre, sono presenti molti sensori sul sistema di riscaldamento e raffrescamento.

Infatti, si tratta di un dataset ampissimo che comprende diverse tipologie di dati: il consumo energetico dell'intero edificio e dell'uso finale, le condizioni operative del HVAC, i parametri ambientali interni ed esterni (es. temperatura, umidità relativa, concentrazione del CO_2 ecc.) e il conteggio dell'occupante.

3.4 Descrizione del dataset

In questo studio sono stati utilizzati solo alcuni dati che riguardano l'ambiente interno ed esterno dell'edificio. Per i dati dell'ambiente esterno è stata fornita una cartella csv contenente la temperatura dell'aria esterna da sensori 1 e 2, la temperatura di rugiada dell'aria, l'umidità relativa dell'aria e la radiazione solare provenienti dal sensore 1. Tutti questi dati sono collezionati al passo di 15 minuti.

Mentre, per quanto concerne l'ambiente interno dell'edificio, sono presenti diversi file csv, tra cui 2 file che contengono la temperatura di set point delle zone esterne di riscaldamento e di raffrescamento, 2 file con le temperature delle zone interne ed esterne e infine un file contenente la concentrazione di anidride carbonica in alcune zone esterne.

Tutti i dati riguardo all'ambiente interno hanno la frequenza di misurazione diversa.

La Tabella 3.1 mostra informazioni dettagliate di queste grandezze misurate. L'istante di tempo con cui i valori vengono registrati nei file sono in Tempo coordinato universale (UTC). Nell'analisi il tempo è riportato a quello locale, che risulta importante per analisi di clustering di profili temporali giornalieri.

Leggendo la tabella si può notare che c'è una differenza tra il numero di zone termiche esterne esistenti (57) e il numero di zone termiche esterne realmente monitorate (51).

In questa analisi sono state utilizzate solo le temperature delle zone termiche/sensori riguardo all'ambiente interno e i valori di varie grandezze misurate dal sensore 1 per l'ambiente esterno.

Dopo un'attenta valutazione, è stato deciso di utilizzare solo il dataset raccolti nell'anno 2019 e nell'anno 2020, poiché il dataset riguardo alla temperatura interna del 2018 presenta diversi valori mancanti, che rappresentano circa un decimo del totale.

Nel 2020 si è diffusa un'epidemia a livello globale, chiamata Covid-19, per cui molti lavoratori hanno iniziato lo smart working da fine marzo, perciò si tratta di una situazione inusuale. Siccome il dataset del 2020 viene utilizzato solo nella fase di testing, è giustificabile se le metriche dell'analisi predittiva sono leggermente superiori.

Dati	Descrizione	Numero di punti raccolti	Unità di misura	Frequenza di campionamento
Dati ambiente esterno	Temperatura esterna da sensore 1	1	°C	15 min
	Temperatura esterna da sensore 2	1	°C	15 min
	Temperatura di rugiada da sensore 1	1	°C	15 min
	Umidità relativa da sensore 1	1	%	15 min
	Radiazione solare da sensore 1	1	W/m^2	15 min
Dati ambiente interno	Temperatura di setpoint di raffrescamento	41	°F	5 min
	Temperatura di setpoint di riscaldamento	41	°F	5 min
	Temperatura zonale di zona interna	16	°C	10 min
	Temperatura zonale di zona esterna	51	°F	1 min
	Concentrazione di CO2 di ciascun zona	13	ppm	1 min

Tabella 3.1: Grandezze fornite dagli autori riguardo all'ambiente interno ed esterno

Capitolo 4

Risultati

In questo paragrafo sono state mostrate le osservazioni rispetto al dataset, ai risultati delle tecniche di analisi e alle discussioni riguardo ai risultati; in particolare, con la volontà di capire se gli algoritmi citati in precedenza sono stati efficaci nel caso studio.

4.1 Prima visualizzazione dei dati

Attraverso il metodo della data reduction, tutti i dati necessari allo studio, quindi la temperatura interna di 63 sensori, i parametri esogeni(la temperatura dell'aria, l'umidità relativa e la radiazione solare) sono ridotti allo stesso passo temporale, cioè di 30 minuti. Prima di procedere con l'eliminazione degli outliers e con la sostituzione dei missing values, per avere un quadro chiaro della situazione sono stati eseguiti dei carpetplot per tutti i sensori. Questi hanno dato modo di comprendere i range di temperatura che sono stati misurati da ciascun sensore, la possibile variazione della temperatura interna in base al periodo stagionale e la presenza dei sopracitati outliers e missing values.

Sull'asse x del grafico sono segnalate le date dell'inizio di ciascun mese, mentre sull'asse y sono riportate le ore dell'arco di una giornata. In questo modo, il colore di una piccola cella rappresenta la temperatura rilevata, quelli di una colonna illustrano la variazione della temperatura nell'arco di una giornata, e le celle bianche corrispondono a dei missing values. Ciascun sensore presenta un range di temperatura diversa, il quale viene mostrato nella legenda a destra del grafico.

Visionando tutti i carpetplot si può osservare che nei dati del 2020 non sono presenti missing values.

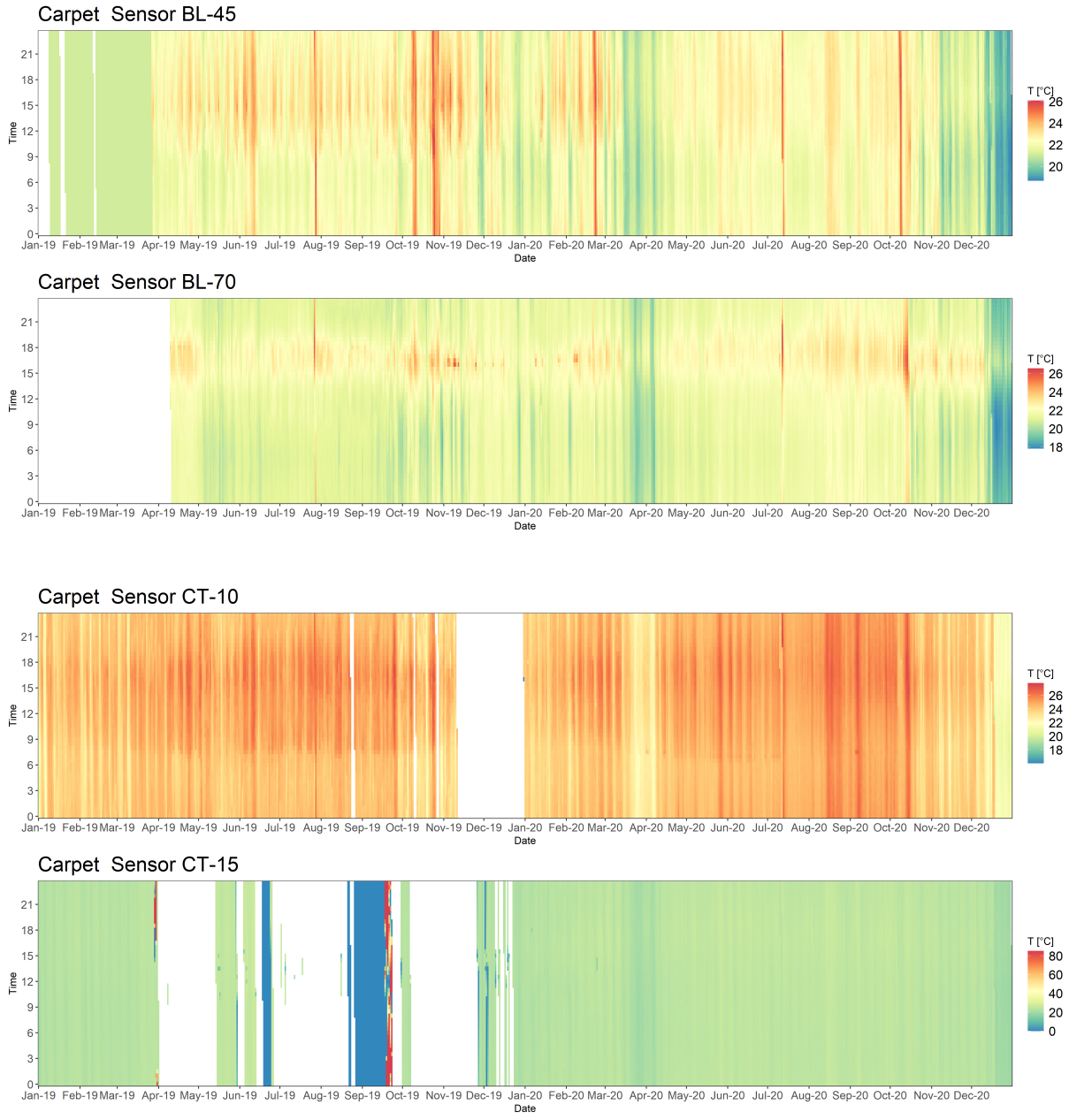


Figura 4.1: Carpetplot dei sensori non considerati: BL-45, BL-70, CT-10, CT-15.

In seguito, nell'osservazione complessiva dei carpetplot sono saltati all'occhio i dati anomali rilevati da 4 sensori, rispettivamente il BL-45, il BL-70, il CT-10 e il CT-15. Infatti, il carpetplot del sensore BL-45 mostra una temperatura costante per circa i primi 3 mesi del 2019, ciò constando un malfunzionamento del sensore poiché risulta impossibile mantenere una temperatura costante in un determinato ambiente per un periodo così lungo. Questi dati potrebbero essere trattati come outliers ed essere sostituiti attraverso il

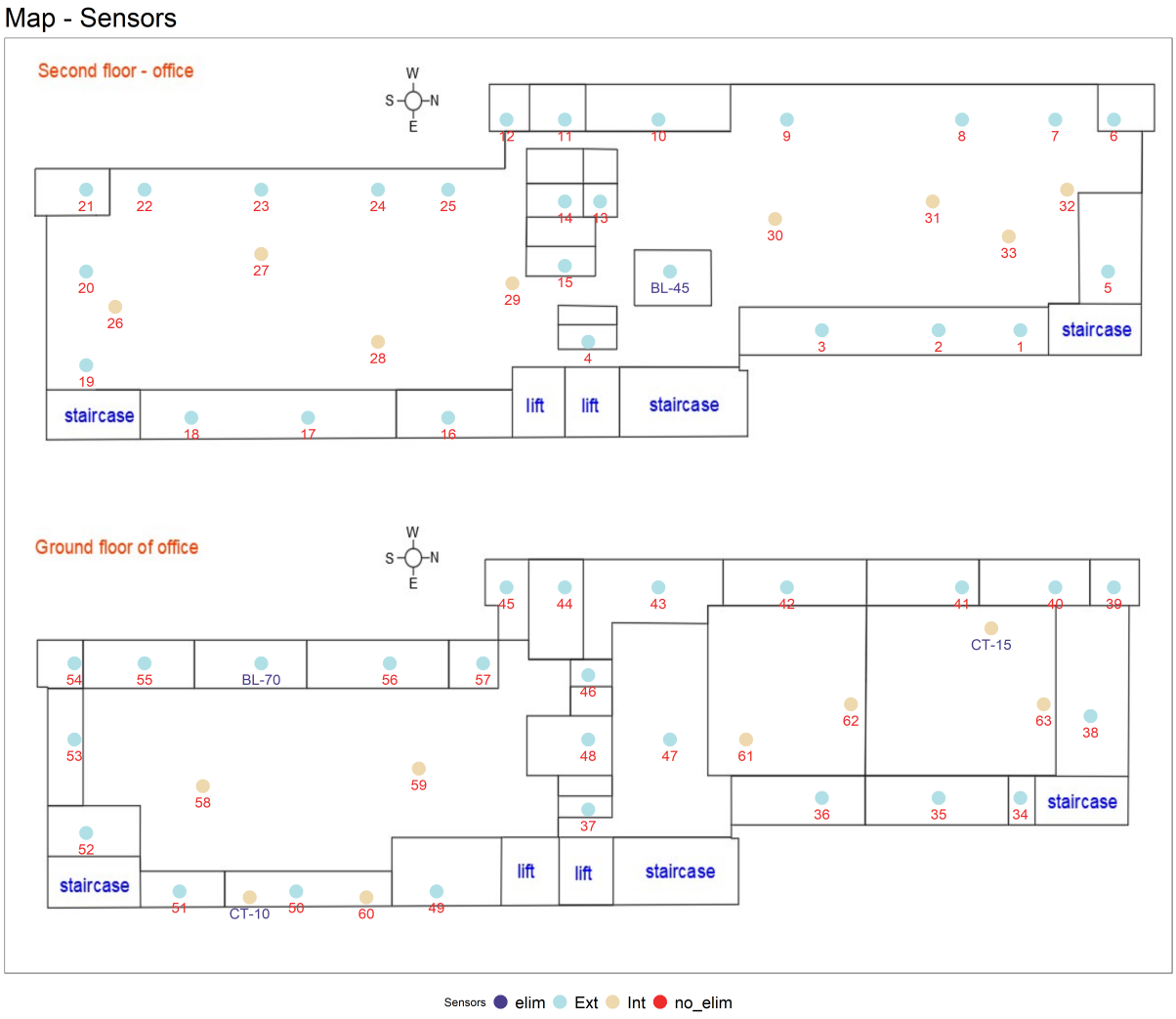


Figura 4.2: Mappa dei due piani a uso ufficio con la posizione dei sensori

metodo di imputazione dei missing values, eppure così facendo il dataset del training di questo sensore sarebbe costituito da circa un quarto dei dati non realmente misurati, per

questo si è evitato di considerarli. Allo stesso modo i carpetplot dei sensori BL-70 e CT-10 presentano lo stesso problema, ossia l'assenza dei dati per un intervallo di tempo maggiore di un mese, nello specifico per più di 4 mesi nel sensore BL-70 e per più di un mese nel sensore CT-10. Il caso più discutibile è sicuramente il carpetplot del sensore CT-15 poiché il dataset utilizzabile per il training del modello di previsione è costituito dai dati di soli 4 mesi a causa del range di temperatura che va da 0 °C a 80°C e dei numerosi lunghi periodi di missing values e di outliers.

Visto che il modello di training opera sulle serie temporali è necessario che tali serie siano continue dal punto di vista temporale, di conseguenza i dati di questi 4 sensori non sono stati impiegati nell'analisi poiché i periodi di interruzione dei dati sono troppo ampi.

I restanti sensori delle zone interne ed esterne sono stati rinominati per facilitare il loro utilizzo come mostrato nella figura 4.2. Le posizioni dei sensori delle zone esterne sono indicative poiché non sono state specificate dagli autori, a differenza le posizioni dei sensori delle zone interne sono segnalate nella figura 3.4. Nella figura 4.2 il colore dei diversi punti contraddistingue la tipologia della zona a cui appartengono i sensori, mentre il colore del loro nome ne identifica l'inclusione o meno nell'analisi.

Quindi i sensori numerati dall' 1 al 33 sono posizionati al secondo piano (di questi gli ultimi otto sono della zona interna), mentre i sensori con la numerazione dal 34 al 63 sono localizzati al primo piano (di questi gli ultimi sei sono installati nella cosiddetta zona interna).

4.2 Pre-processioning

Dai carpetplots dei sensori si osserva chiaramente che alcuni dati dei sensori 26,28,30,31,63 sono degli outliers poiché essi assumono valori oltre i 35°C o inferiori ai 10 °C nonostante sono circondati da valori che in media si aggirano sui 22-25°C. È chiaro che in un ambiente open space così grande è impossibile raggiungere una temperatura così elevata e poi avere un raffreddato così rapido, poiché il sistema di riscaldamento e raffreddamento non è in grado di agire così prontamente a una variazione di temperatura così ampia. Ciò è stato confermato dal fatto che tale fenomeno non è percepito da altri sensori vicini. Nella figura

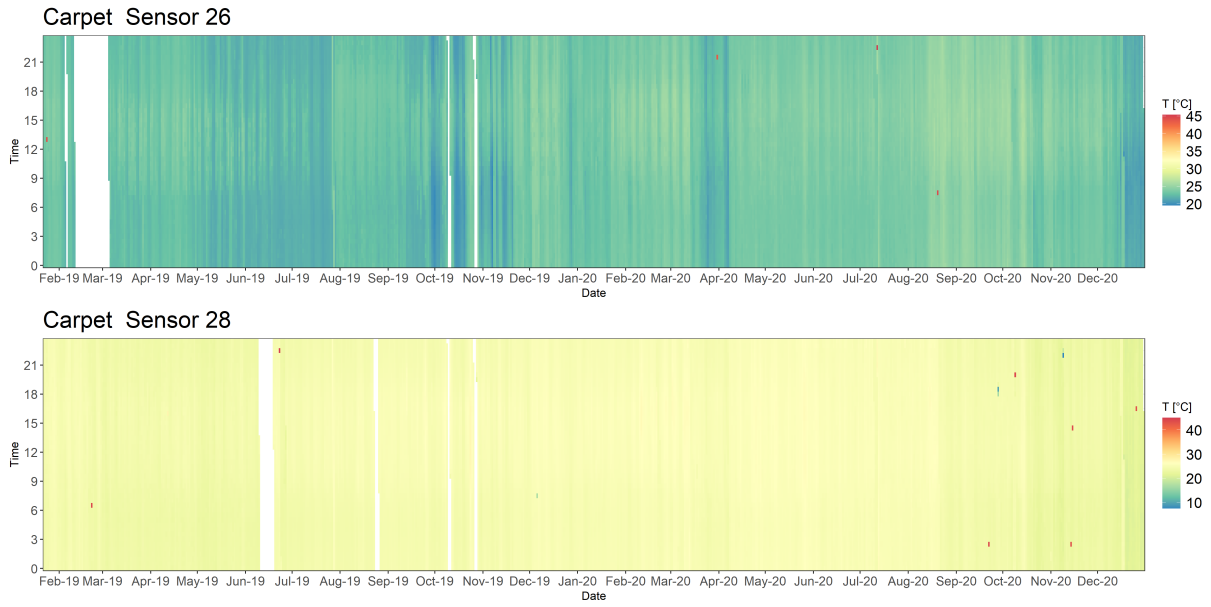


Figura 4.3: Prima di pre-processing: Carpetplot dei sensori di 26,28, dall'alto verso il basso.

4.3 è possibile visualizzare i carpetplot dei sensori 26,28.

Oltre agli outliers, dai carpetplots è possibile osservare anche gli intervalli dei missing values, che come già detto sono assenti all'interno del dataset del 2020. I missing values del 2019 possono essere divisi in due gruppi, nel primo gruppo sono racchiusi quelli delle zone esterne rilevati nei primi due mesi dell'anno, mentre nel secondo gruppo sono presenti quelli delle zone interne misurati ad agosto e ad ottobre. Inoltre nel medesimo dataset sono presenti ulteriori intervalli di missing value al di fuori dei periodi descritti, i cui

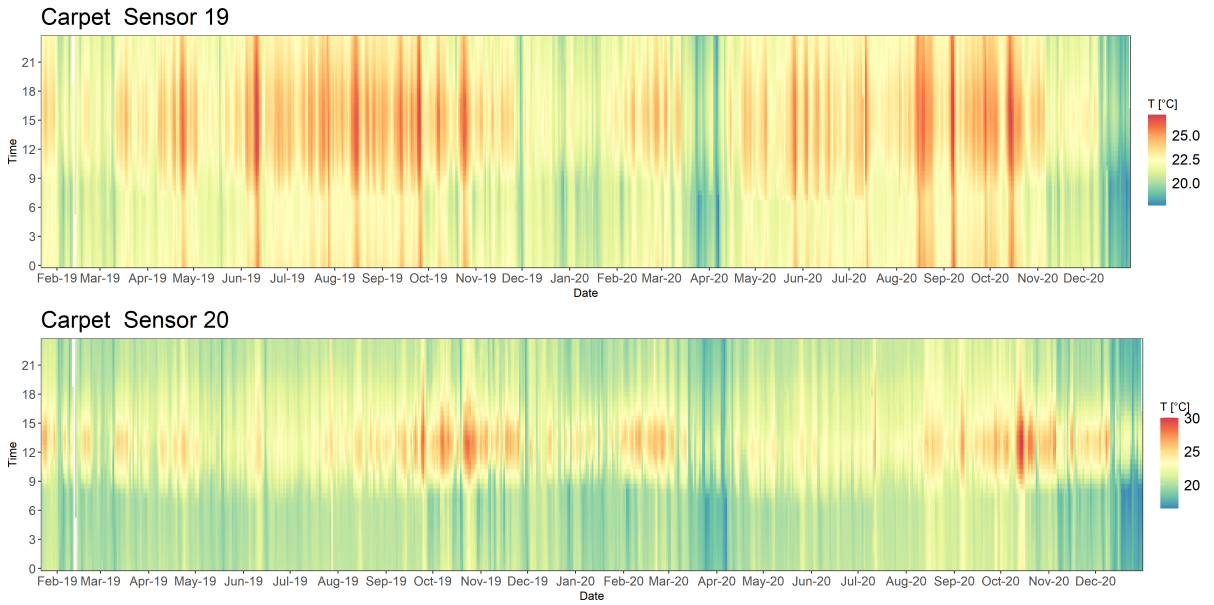


Figura 4.4: Prima di pre-processing: Carpetplot dei sensori di 19,20.

esempi sono mostrati nelle figure 4.3 e 4.4, di questi i sensori 19 e 20 appartengono alle zone esterne e i sensori 26 e 28 alle zone interne.

Per individuare con facilità i periodi di missing value è stato redatto un calendarplot riportante i giorni in cui sono presenti dei missing values, il medesimo è riportato nella figura 4.5. Il testo all'interno delle celle denota il giorno del mese, mentre il colore indica il numero dei sensori che possiedono dei missing values nel giorno indicato. Tali colori assumono tonalità più scure in base all'incremento del numero di sensori.

Il periodo della ricostruzione delle serie temporali è un trade-off tra il vantaggio di una quantità maggiore di dati e il costo computazionale della tecnica di imputazione dei missing values. Infatti, mentre nel mese di gennaio sarebbe risultato sconveniente ricostruire 49 profili temporali di 5 giorni per ottenere solo 6 giorni di dati in più per il training; nel mese di dicembre è stato vantaggioso sostituire i missing values di un sensore per un intervallo di 7 giorni poiché tale azione ha permesso di conquistare 20 giorni di dati in più. Dopo una serie di ragionamenti per effettuare l'analisi si è deciso di adottare il dataset che corrisponde al periodo che va dal 22/01/2019 al 31/12/2020.

Dopo l'eliminazione dei outliers, individuati con il metodo di z-score, e la sostituzione dei missing values, effettuata mediante l'interpolazione lineare degli outliers puntuali e il

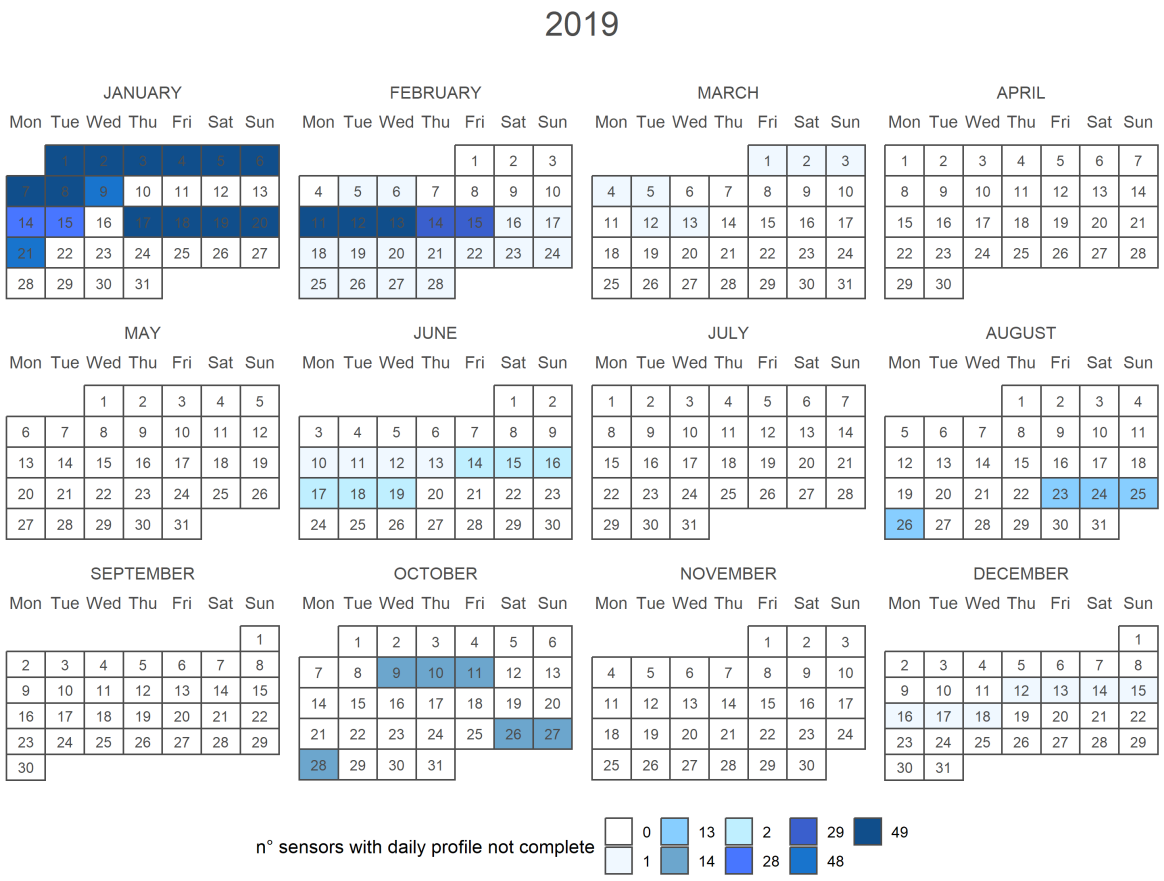


Figura 4.5: Calendar plot dei giorni presenti NA

K-Nearest Neighbors per intervalli dei missing values, si è ottenuto il dataset per svolgere le suddette analisi. Gli esempi dei risultati del Pre-processing possono essere visualizzati nella figura 4.6, contenente i carpetplot degli stessi sensori precedentemente mostrati, in modo da poterli paragonare tra di loro. A differenza, dai grafici della figura 4.7 è possibile vedere la ricostruzione dei valori evidenziati attraverso puntini di colore blu. Si osserva che i punti di sostituzione creati con il metodo KNN sono uniformi con i restanti valori del dataset. In aggiunta, la tecnica di sostituzione KNN è efficace per la ricostruzione sia di piccole sequenze costituite da decine di missing values, come mostrato nella figura 4.7.1, sia di sequenze composte da centinaia di NA, come mostrato nella figura 4.7.3.

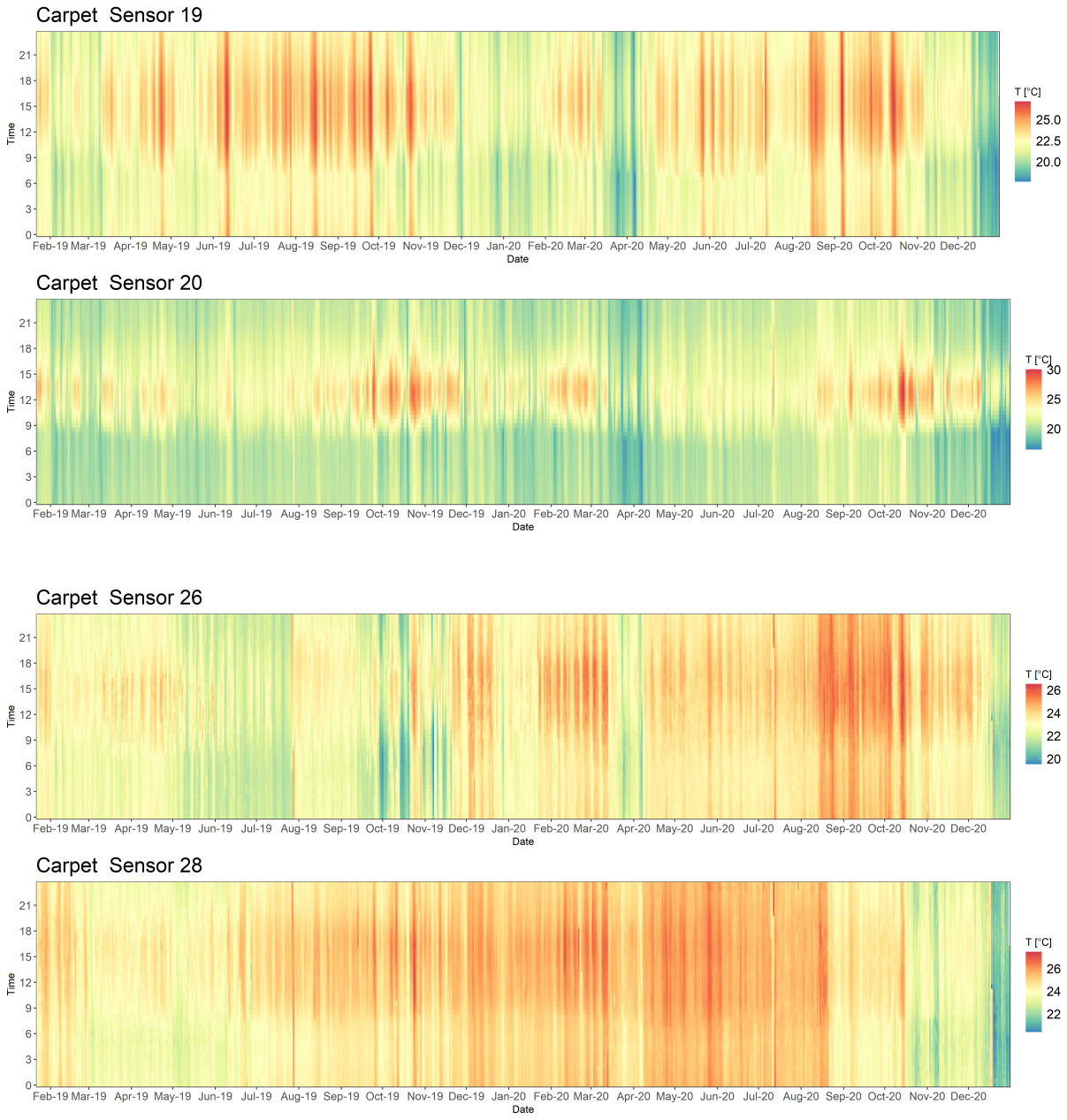


Figura 4.6: Dopo Pre-Processing: Carpetplot dei sensori 19,20,26,28.

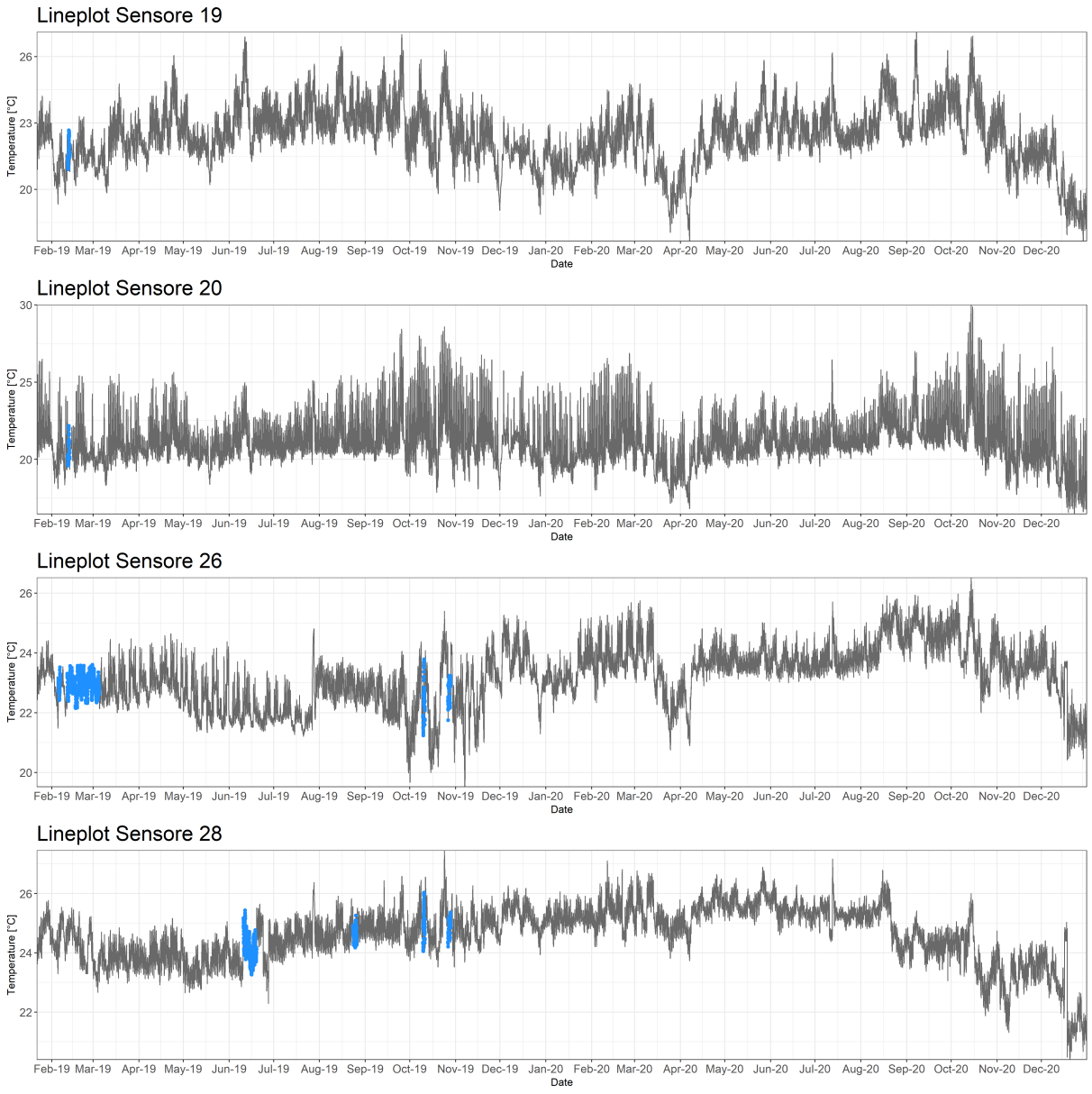


Figura 4.7: Lineplot dei sensori 19, 20, 26, 28.

4.3 Caratterizzazione Profili temporali giornalieri

Un primo fattore osservato è l'assenza di un evidente cambiamento della temperatura interna al variare del periodo stagionale. Tale fenomeno non si potrebbe verificare in Italia poiché la normativa in materia di edifici ad uso ufficio prevede un set-point attestato a

26°C in estate e a 20°C in inverno. La medesima condizione statica è riscontrata in tutti i giorni della settimana, quindi senza avere differenza tra i giorni festivi e i giorni feriali.

Come si può visionare tramite l'immagine 4.6 i range di temperatura rilevati dai sensori nelle diverse zone possono essere molto diversi tra di loro. Ad esempio il range del sensore 20 oscilla tra temperature che vanno dai 20°C ai 30°C, mentre quello del sensore 28 oscilla tra temperature che vanno dai 22 °C ai 26 °C. Sempre per i medesimi le ore più calde sono distribuiti in modo differente.

Infatti la variazione della temperatura nell'arco di una giornata può essere molto differente nelle varie zone, poiché alcune zone sono caratterizzate da una piccola variazione della temperatura dell'ambiente che oscilla intorno al set-point, mentre altre zone si distinguono per un'elevata alterazione della temperatura, la quale è forse meno influenzata dal sistema di riscaldamento e raffrescamento.

Per favorire la visione della variazione della temperatura nell'arco di una giornata è stata costruita una mappa sulla differenza media della temperatura di una giornata, la quale è presente nella figura 4.8. Il colore dei puntini si riferisce alla differenza media della temperatura in questo modo dalla legenda del grafico si può nota che in alcune zone la differenza supera i 3 °C, ciò potrebbe causare discomfort degli occupanti.

È stato ulteriormente osservato che le zone esposte ad Est hanno una differenza della temperatura minore rispetto a quelle esposte ad Ovest, nonostante i sistemi di riscaldamento e di raffrescamento, descritti nel capitolo precedente, gestiscono entrambe le zone. In particolare i sensori 20, 22 e 23 sono quelli con la differenza della temperatura maggiore e al contrario i sensori 13, 14, 36, 46 e 49 sono quelli con la variazione più bassa.

I sensori 50 e 60, nonostante risultano vicini a livello spaziale, hanno una differenza della temperatura abbastanza evidente. Infatti la differenza della temperatura misurata dal sensore 50 è inferiore rispetto al sensore 60, ciò probabilmente è dovuto dalla diversa posizione dei sensori, poiché il sensore 50 è stato installato a parete e appartiene al sistema BMS, mentre il sensore 60 è stato installato vicino alla postazione di lavoro, sotto il tavolo, da parte degli autori di [33] in un momento successivo.

4.3.1 Clustering

Allo scopo di capire quali sono i profili riscontrati all'interno dell'edificio e quali sono quelli dominanti, è stato effettuato un clustering dei profili temporali giornalieri completi.

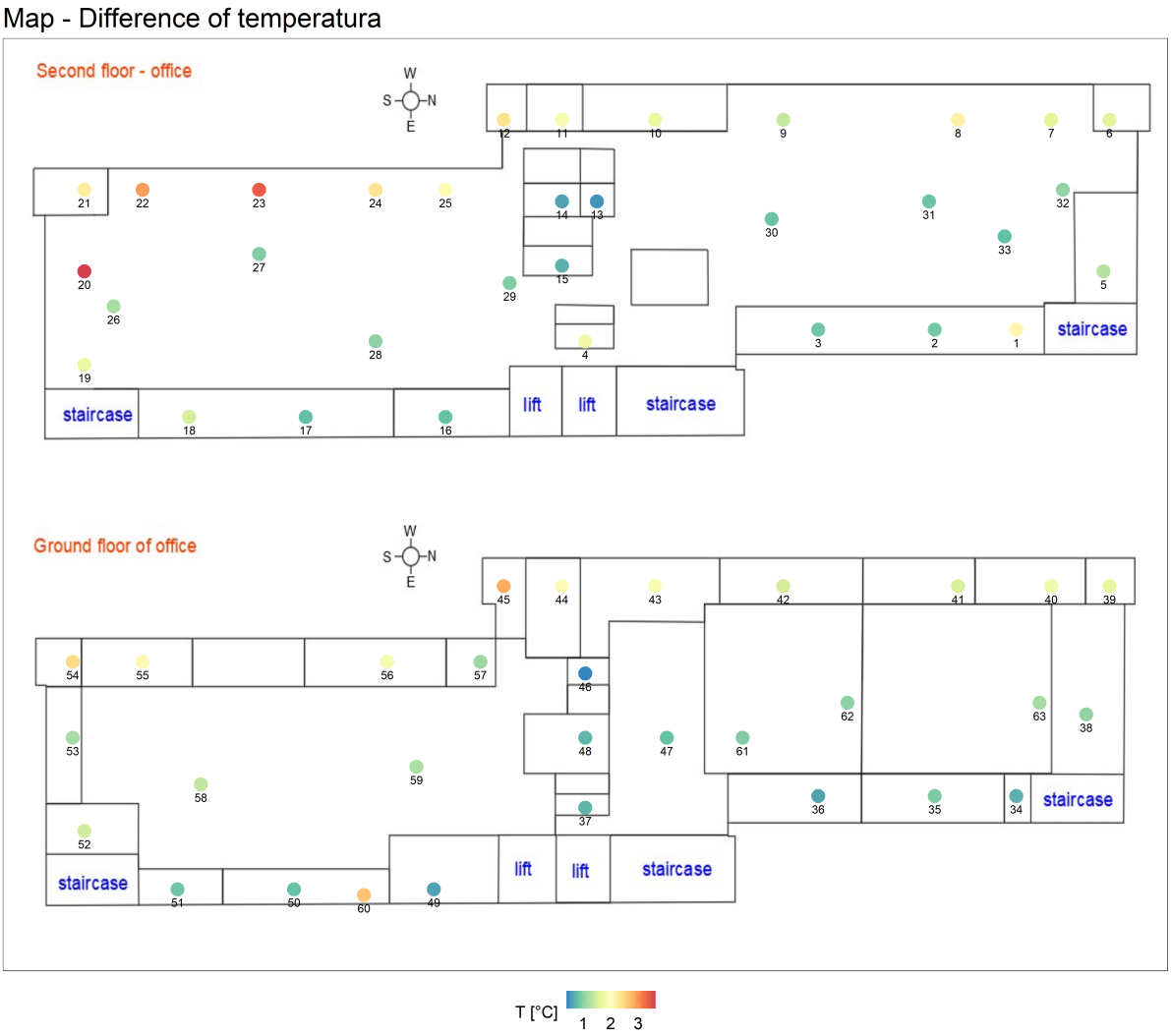


Figura 4.8: Mappa in funzione della media di differenza giornaliera di temperatura tra la minima e la massima

Applicando il sub-sequence clustering con k-means e la distanza Euclidea si ottengono i clusters plottati nella figura 4.9. Si trattano di clusters dissimili sia dal punto di vista della forma, sia dal punto di vista dell'intervallo dei valori della temperatura.

Quindi, dividendo i profili in 3 sotto-sequenze da 8 ore e stabilendo per ogni sotto-sequenza che k è uguale a 3, si ottengono 25 clusters in totale.

I profili temporali sono plottati con il color grigio, mentre i profili colorati di ciascun cluster sono la media complessiva dei precedenti. Questi ultimi non sempre sono

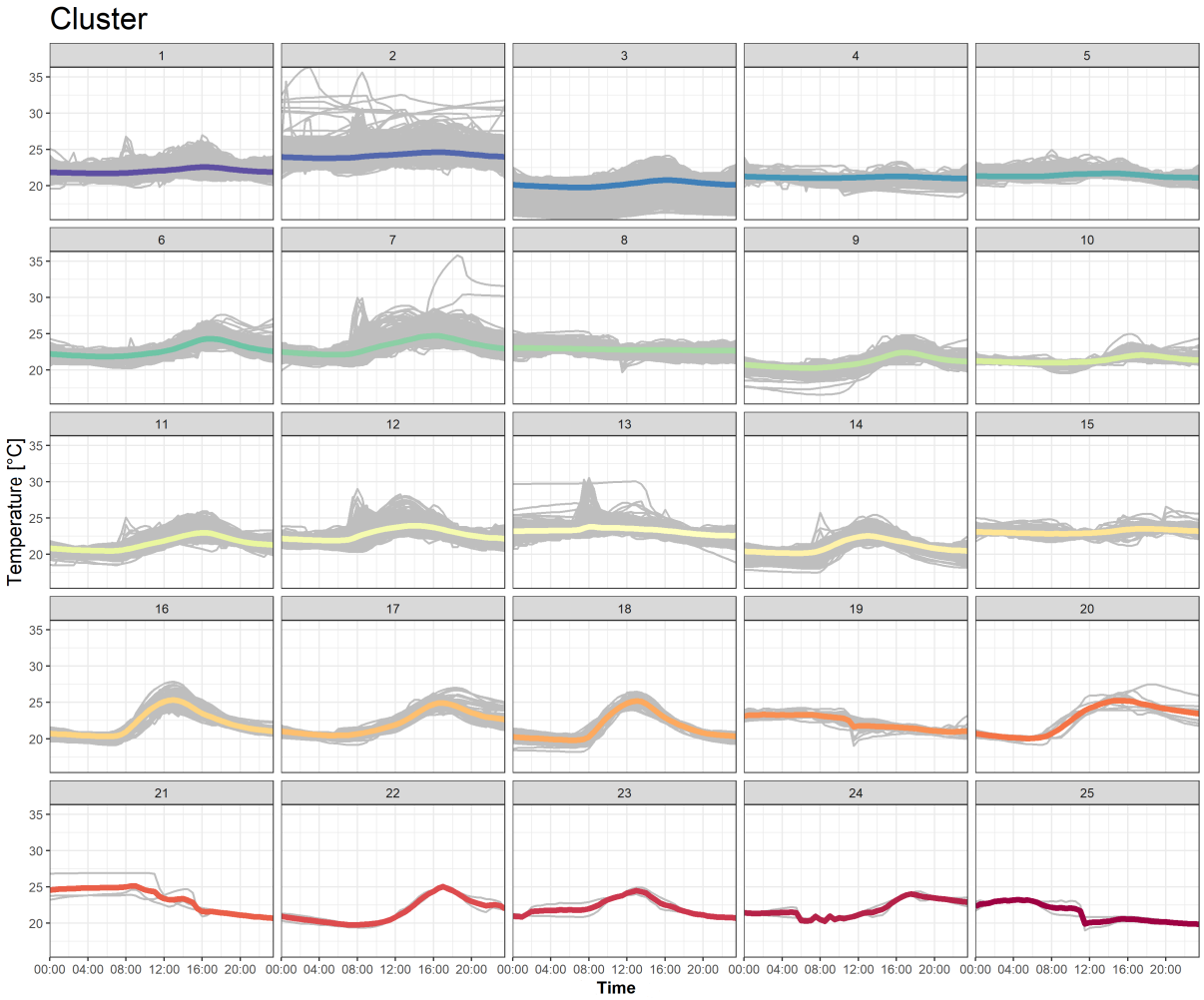


Figura 4.9: Cluster

necessariamente un profilo reale, ma descrivono in generale il pattern del cluster a cui appartengono.

I clusters sono ordinati in base alla numerosità dei profili, infatti mentre i primi clusters sono molto concentrati, gli ultimi contengono pochissimi profili. La percentuale di ogni cluster è stata riscontrata nella figura 4.10. Da essa è stato possibile osservare che i primi tre clusters hanno percentuali più alte se paragonate a quelle del resto dei clusters, infatti essi occupano circa i tre quarti del totale dei profili, non a caso già il primo ne contiene circa la metà.

Osservando la figura 4.9, si può riscontrare che i profili per la maggior parte sono

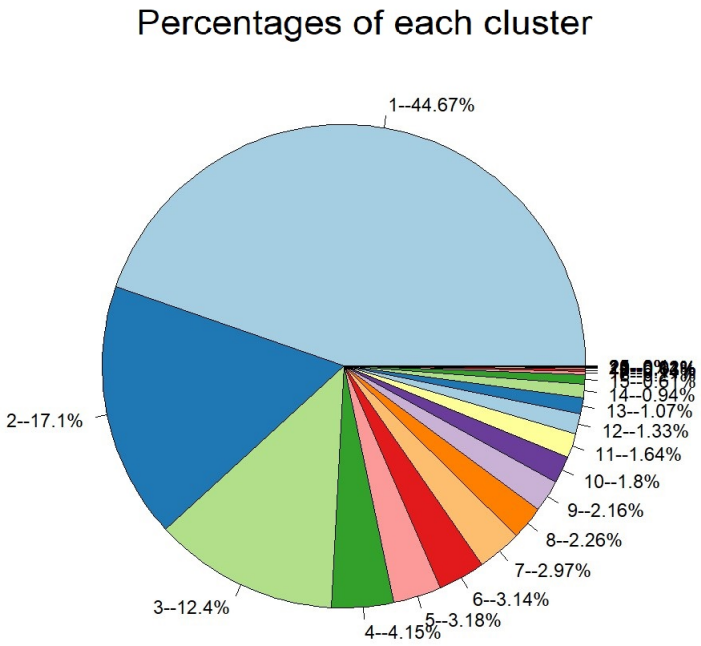


Figura 4.10: Diagramma a torta con i percentuali di profili in ciascun cluster

piatti, come nel caso dei centroidi dei primi cinque clusters. All'interno di questi cluster sono presenti dei profili che hanno una leggera curvatura, dipesa da una differenza di temperatura molto piccola. A differenza alcuni clusters, come il 16 e il 18, hanno registrato un aumento progressivo della temperatura nelle ore mattutine, la quale ha provocato una differenza di 4-5°C, mentre altri, come il 6 e il 9, hanno registrato un aumento progressivo della temperatura nelle ore pomeridiane. Infine alcuni profili non tipici, come gli ultimi cinque clusters, hanno contenuto pochissimi profili di conseguenza, non è stato possibile classificarli in ulteriori clusters per la loro forma.

I clusters, oltre che per la forma, si differenziano tra loro anche per valori di intervalli della temperatura. Ad esempio, i clusters 6 e 9 risultano simili per la forma, ma si differenziano tra loro per la temperatura, infatti il cluster 6 ha valori leggermente più alti. Allo stesso modo, presentano la stessa difformità anche i clusters 12 e 14 e, come loro tanti altri.

Riguardando il diagramma a torta, rappresentato nella figura 4.10, si può notare che i profili dei primi tre clusters, oltre a rappresentare circa il 74.2 % del totale, ciascuno di loro registrano una percentuale superiore al 10 %. Di conseguenza i restanti clusters non

arrivano neanche al 5 % e, perfino, gli ultimi undici clusters non raggiungono nemmeno l'1%.

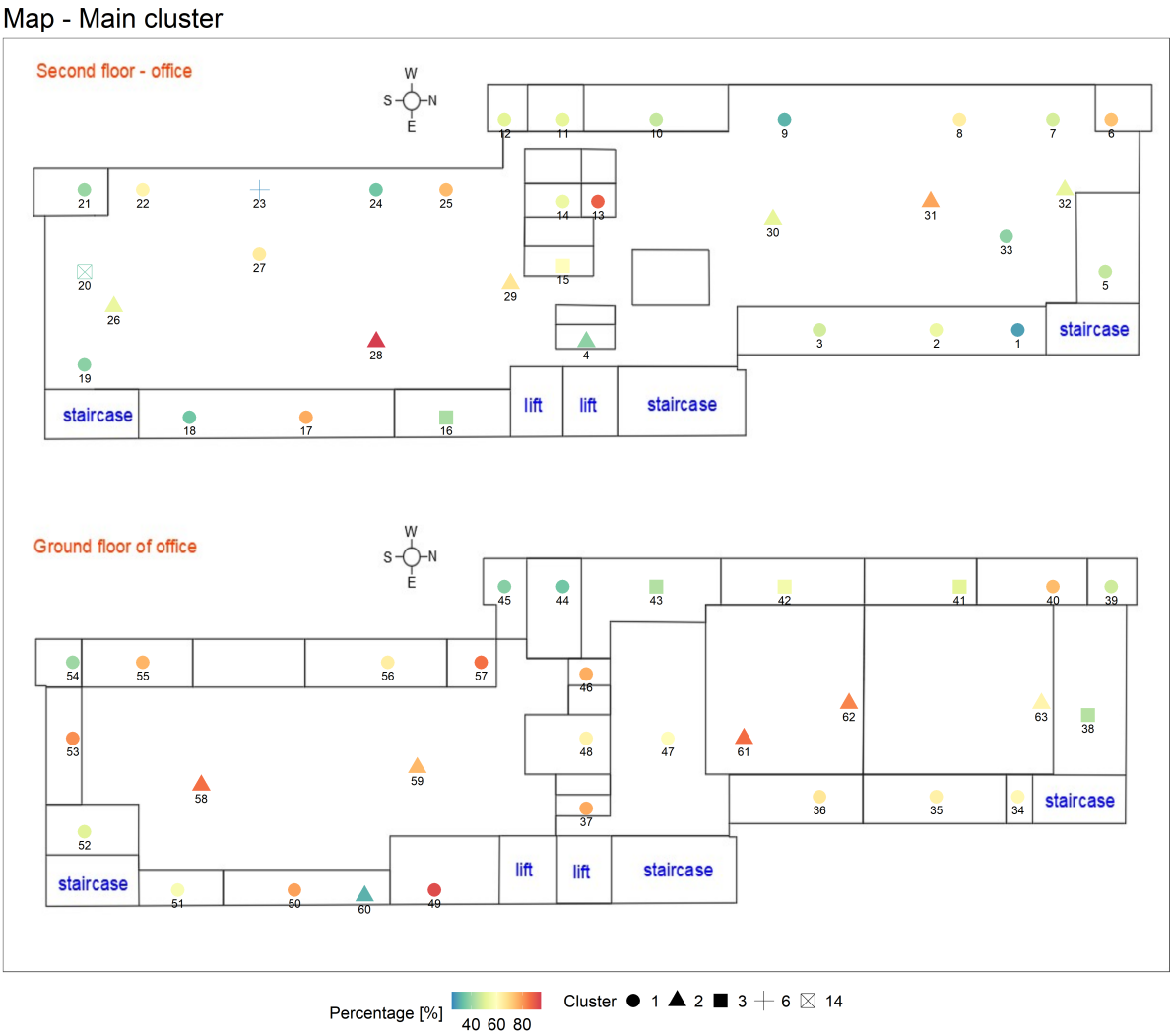


Figura 4.11: Mappa in relazione al Principale cluster

Attraverso la figura 4.11 è possibile constatare qual è il principale cluster a cui appartengono i profili della temperatura dei vari sensori. In questa la forma dei punti colorati rappresenta il cluster principale, il colore identifica la percentuale dei profili del cluster principale e il numero ne indica il sensore.

Innanzitutto è possibile affermare che sono presenti cinque clusters principali, rispettivamente l'1, il 2, il 3, il 6 e il 14, in particolare quest'ultimo è il cluster dominante

dei profili del sensore 20 con una bassissima percentuale, attestata al 35%; allo stesso modo, lo è il cluster 6 per il sensore 23 con una percentuale attestata al 25%. Questi due sensori sono stati argomento di discussione già precedentemente quando è stata presa in considerazione la media annuale della differenza della temperatura di un giorno, infatti questi sono risultati i due sensori in cui è stata constata una media di differenza superiore ai 3°C, la quale è stata osservata attraverso i centroidi dei clusters dominanti.

I sensori con bassa percentuale dei profili del cluster dominante registrano una variazione non stabile della temperatura e della forma, di conseguenza si ha una bassa frequenza dei profili degli altri clusters. In questo modo, è probabile che nell'analisi predittiva non sarà possibile registrare una buona metrica riguardo tali sensori.

Nella mappa è possibile vedere che i cluster dominanti sono principalmente l'1, il 2 e il 3, e che il primo risulta a essere il più presente, ciò è deducibile dalla percentuale dei clusters del diagramma a torta. A questo punto è interessante sottolineare che i sensori non esposti verso l'esterno hanno il cluster 2 come cluster dominante e che il cluster 3, nonostante è presente in pochi sensori, lo è in 3 sensori adiacenti.

Infine, si riscontra che i sensori con un'alta percentuale del cluster dominante sono il 13, il 28 e il 49, la quale è molto prossima o superiore al 90%.

Successivamente, l'analisi si è concentrata sui 3 principali clusters, in quanto attraverso questi è possibile descrivere la maggior parte dei casi e in quanto l'analisi attraverso il CART non riuscirebbe a prevedere il quarto cluster, data la sua percentuale molto bassa.

Nella figura 4.12 sono rappresentati i 3 cluster a seguito di un'eliminazione dei profili meno simili eseguita attraverso il cluster gerarchico con il single linkage, il quale ha portato a una rimozione di circa il 4% del totale di ciascun cluster. In linea con i precedenti grafici, i profili dei clusters sono colorati di grigio, mentre i centroidi del cluster assumono colori differenti. In questo caso i valori riportati dall'asse y sono diseguali tra i tre plot, quindi le forme dei centroidi sono pressoché identiche, ma i valori dei profili sono diversi. Infatti è possibile notare che i profili del cluster 1 variano tra i 20°C e i 26°C, quelli del cluster 2 oscillano tra i 22°C e i 28°C e infine quelli del cluster 3 fluttuano tra i 16 e i 24°C. Allacciandosi al discorso del cluster principale dei sensori, è ragionevole constatare che nelle zone interne le temperature siano leggermente superiori alle zone esposte all'esterno. Nel caso del cluster 3 è necessario evidenziare che nonostante siano presenti alcuni profili attestati a circa 16°C, questi sono sicuramente in quota minoritaria dato che le temperature del centroide sono di alcuni gradi in più.

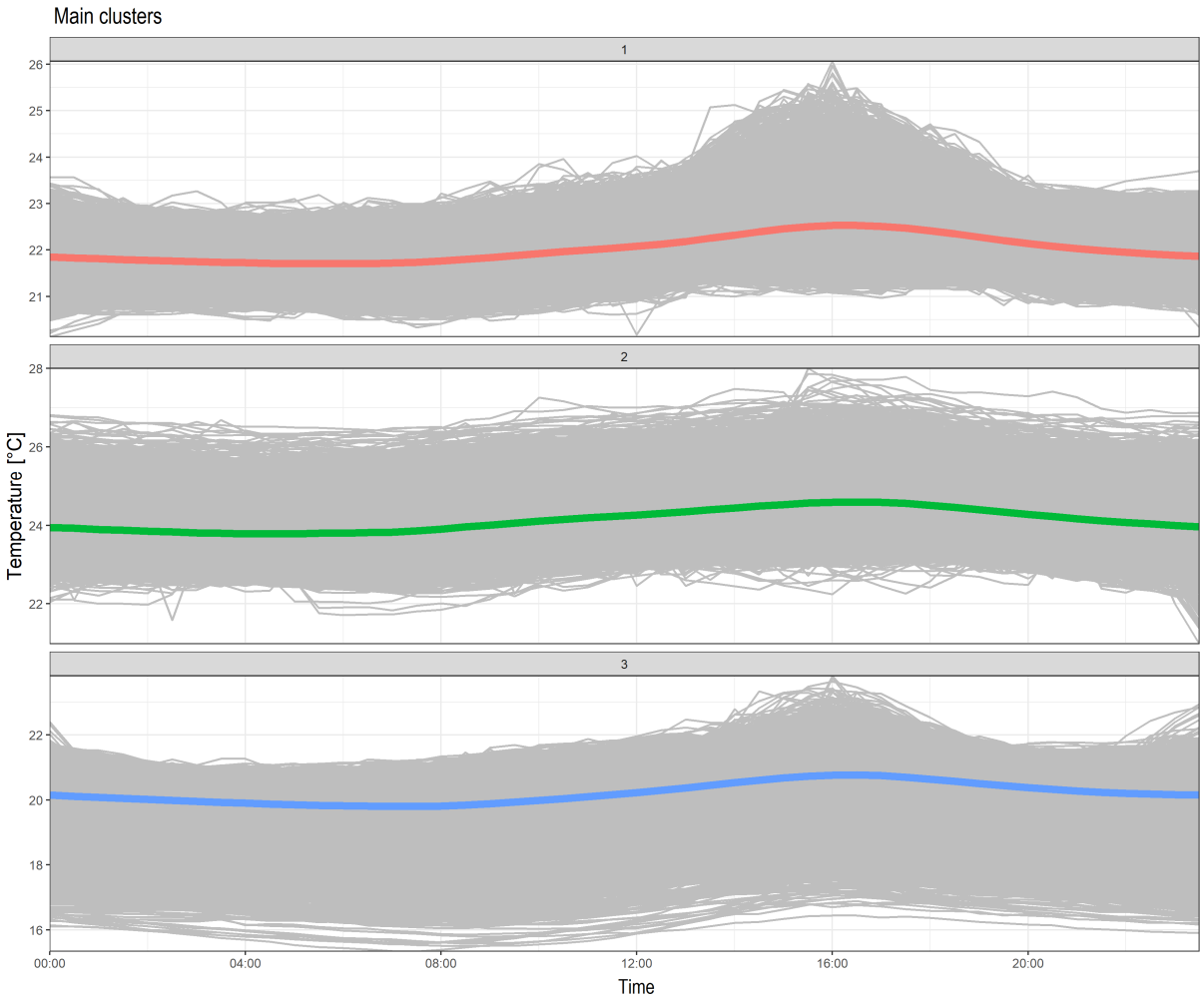


Figura 4.12: I tre principali clusters

Così, in conclusione è possibile sostenere che la differenza della temperatura all'interno dell'edificio è notevole.

4.3.2 Classification and Regression Tree

Come già anticipato, per motivi precedentemente enunciati, lo studio con il CART si è concentrato su i primi tre clusters. Le variabili indipendenti dell'albero si possono classificare in due distinte tipologie:

1. le variabili spaziali e temporali: il mese, il giorno della settimana, la posizione dei sensori e il piano;
2. le variabili esogene: la temperatura giornaliera, l'umidità relativa giornaliera e la radiazione solare giornaliera.

Inoltre, sono stati specificati il numero minimo di oggetti per ciascun nodo terminale, pari a 500, e il numero di cross-validation, uguale a 10.

I risultati del CART sono visibili attraverso la figura 4.13. Tra tutte le variabili

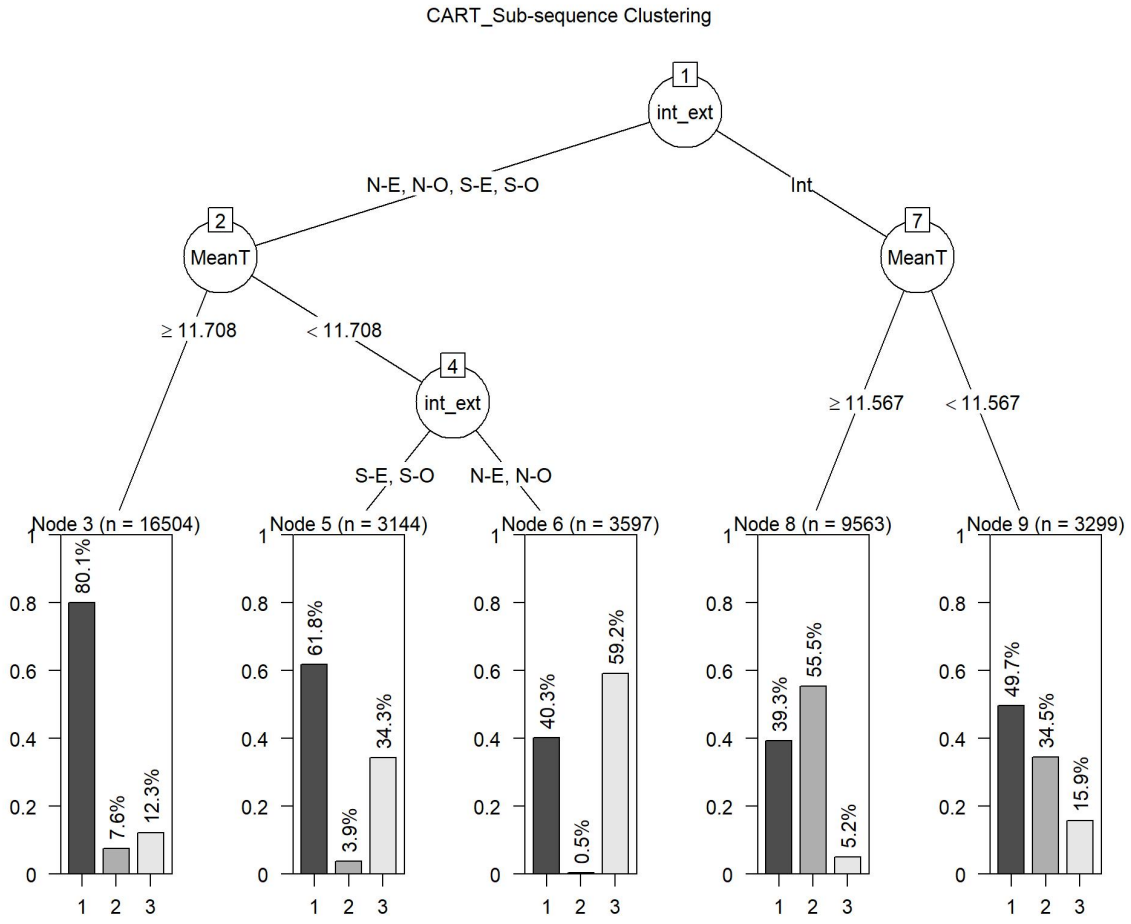


Figura 4.13: Albero di classificazione

indipendenti, solamente quella della temperatura media giornaliera e quella della posizione dei sensori risultano importanti per la classificazione dei profili temporali giornalieri.

Eposizione	Sensori
N-O (Nord-Ovest)	6, 7, 8, 9, 10, 11, 12, 39, 40, 41, 42, 43, 44, 45
N-E (Nord-Est)	1, 2, 3, 5, 34, 35, 36, 38
S-O (Sud-Ovest)	20, 21, 22, 23, 24, 25, 53, 54, 55, 56, 57
S-E (Sud-Est)	16, 17, 18, 19, 49, 50, 51, 52, 60
Interno	4, 13, 14, 15, 26, 27, 28, 29, 30, 31, 32, 33, 37, 46, 47, 48, 58, 59, 61, 62, 63

Tabella 4.1: Posizione dei sensori

La variabile che splitta l'albero in maniera decisiva è la posizione dei sensori, infatti tramite tale divisione si possono ottenere due rami: il primo costituito dai sensori posizionati nelle zone esterne e il secondo composto dai sensori posizionati nelle zone interne. Dopodiché, è la variabile della temperatura esterna a far suddividere ulteriormente entrambi i rami in altri sotto-rami, temperatura che si attesta per entrambi a 11.5°C. Per quanto riguarda i sensori delle zone interne è possibile notare che il cambiamento rilevato dai sensori non è così netto, infatti il cluster principale dei due nodi ha una percentuale molto prossima al secondo cluster. A differenza, i sensori adiacenti all'esterno hanno, per la maggior parte, dei profili appartenenti al cluster 1.

In generale, le percentuali del cluster dominante di ciascun nodo terminale non sono elevate; ciò è visibile attraverso la tabella 4.2, contenente le condizioni di decisione e le percentuali di ciascun cluster. Per quanto riguarda le metriche calcolate:

- Accuratezza: 67.13 %;
- Richiamo: 73.23 %;
- Precisione: 76.32 %.

Questi valori non risultano elevati, ma è necessario tenere in considerazione anche della numerosità degli oggetti. Infatti sono presenti più di 32 mila profili giornalieri, di conseguenza, più è consistente l'insieme degli oggetti, più la precisione e l'accuratezza si abbassano velocemente.

Condizione di decisione	Profili totali	Frequenza di occorrenza di ciascun cluster	Cluster Dominante
Se (Posizione != interno) AND (Media-T $\geq$ 11.708)	16504	80.1% - 7.6%-12.3%	1
Se (Media-T $<$ 11.708) AND (Posizione = S-E, S-O)	3144	61.8% - 3.9% - 34.4%	1
Se (Media-T $<$ 11.708) AND (Posizione = N-E, N-O)	3597	40.3% - 0.5% - 59.2%	3
Se (Posizione = interno) AND (Media-T $\geq$ 11.567)	9563	39.3% - 55.5% - 5.2%	2
Se (Posizione = interno) AND (Media-T $<$ 11.567)	3299	49.7% - 34.5% - 15.9%	1

Tabella 4.2: Condizioni di decisione per i pattern giornalieri di temperatura di generato dall'albero di decisione

4.4 Analisi Predittiva

Il dataset utilizzato dai modelli di previsione si divide in 2 parti, la prima parte contiene il dataset di training, implementato per allenare il modello in modo che apprenda le relazioni che intercorrono tra le variabili di input e di output, mentre la seconda parte contiene il dataset di testing, adoperato per verificare la qualità del modello allenato.

Gli input impiegati sono di 2 tipologie:

- i parametri esogeni, quali la temperatura dell'aria, l'umidità relativa e la radiazione solare;
- i dati storici della temperatura interna del sensore.

Nella figura 4.14 sono illustrate le distribuzioni della temperatura per i diversi mesi, sia dei dati del 2019 che dei dati del 2020. Le forme di distribuzione della temperatura dell'anno 2019 sono abbastanza differenti da quelle dell'anno 2020. È possibile notare che la forma della distribuzione della temperatura è simile nei mesi che vanno da marzo a settembre 2019 e lo stesso vale per gli ultimi tre mesi del medesimo anno. Inoltre, l'intervallo della distribuzione è pressoché lo stesso per tutti i mesi del 2019.

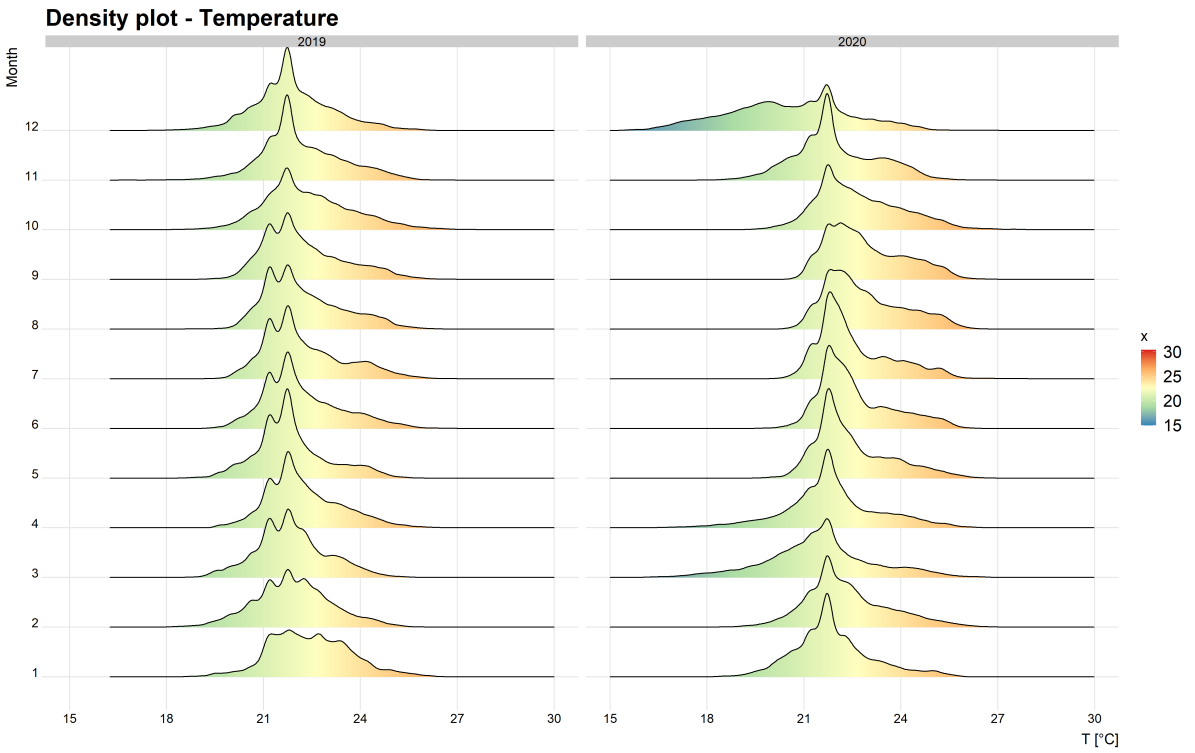


Figura 4.14: Ridgeplot della temperatura in relazione al mese

Diversamente le forme della densità della temperatura del 2020 sono differenti tra di loro così come lo è l'intervallo di distribuzione, il quale durante i mesi più caldi si sposta verso un range della temperatura maggiore.

La figura 4.15 mostra le distribuzioni della temperatura interna misurate da ciascun sensore, raccolte nel 2019 e nel 2020. È possibile osservare che le forme delle distribuzioni della temperatura interna sono molto diverse tra di loro. Infatti, alcune risultano essere monomodali, come nei casi dei sensori 25 e 53, alcune risultano essere bimodali, come nel caso del sensore 36, e altre risultano essere piatte, come nei casi dei sensori 1 e 60. Inoltre, gli intervalli dei valori delle distribuzioni sono diversi tra di loro, alcuni si concentrano intorno ai 22°C, altri variano tra i 21 e i 24°C, e altri ancora oscillano tra i 23 e i 26°C.

Tuttavia, non si riscontrano grandi variazioni nella forma e nell'intervallo dei valori delle distribuzioni della temperatura nei medesimi sensori nei due differenti anni.

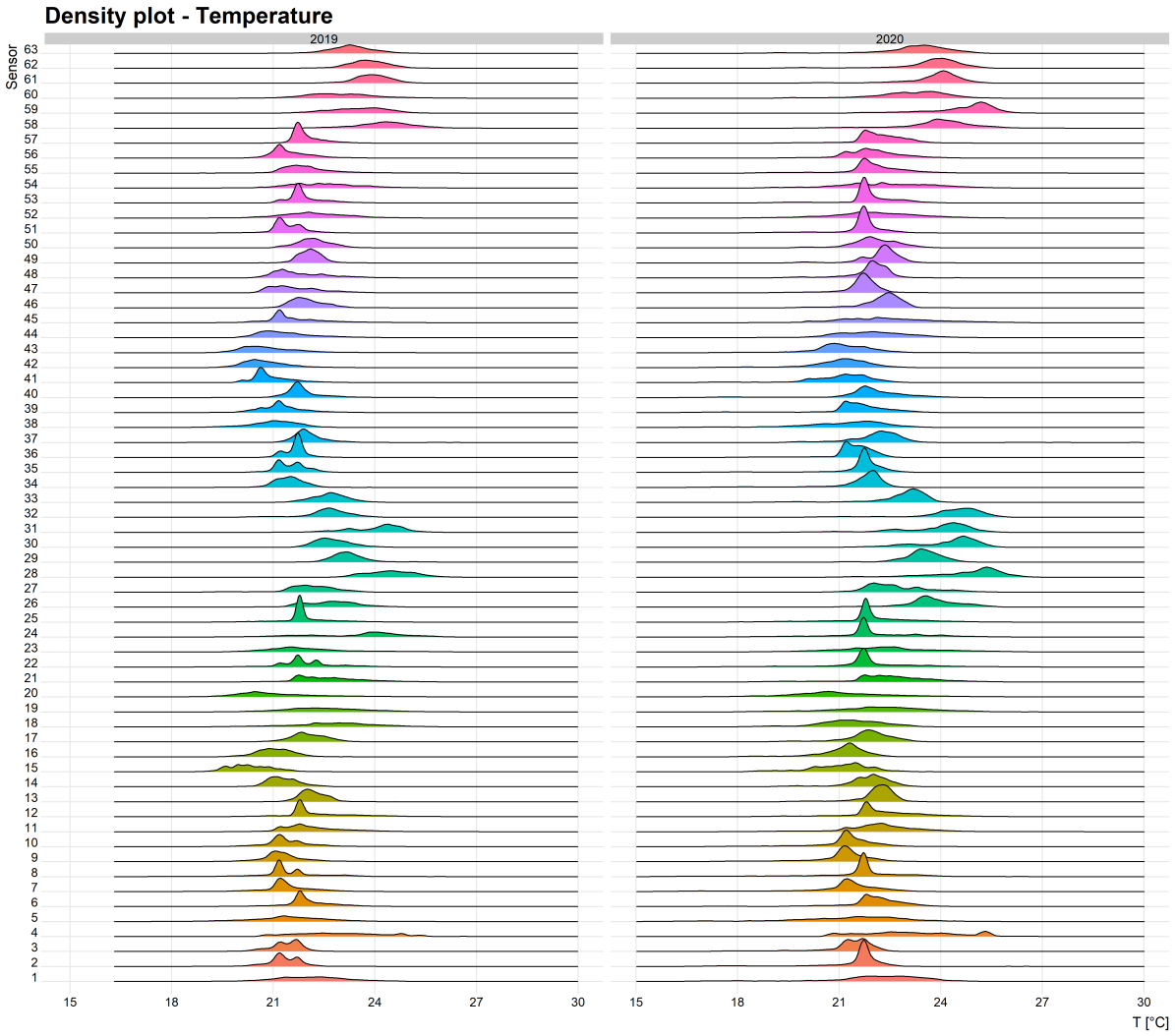


Figura 4.15: Ridgeplot della temperatura dei sensori

4.4.1 Graph Neural Network

Prima di effettuare l'analisi attraverso il GNN, è stato necessario introdurre le matrici di adiacenza e i grafi, poiché il lavoro di analisi si è concentrato principalmente sulla costruzione dei grafi, i quali sono stati implementati come input del modello.

Come già citato in precedenza sono stati utilizzati diversi approcci totalmente differenti tra loro per costruire la cosiddetta matrice di adiacenza:

- l'approccio spaziale, matrice costruita tramite le informazioni fisiche dei sensori e

dell'edificio.

- l'approccio basato sulla similarità, tramite il quale due nodi si collegano tra di loro per la similarità geometrica e per la correlazione delle serie storiche.

Le due matrici spaziali sono state create manualmente osservando la vicinanza dei sensori sulla mappa. I criteri di selezione sono diversi a secondo della zona di appartenenza, secondo una distinzione eseguita dagli autori del [33], quindi sono suddivisi in sensori appartenenti alla zona interna e in quelli appartenenti alla zona esterna. Le mappe fornite da quest'ultimi specificano la posizione precisa dei sensori installati nella zona interna, mentre i sensori installati nella zona esterna sono indicati in base a loro zone termiche dettagliatamente delimitate in cui sono presenti gli stessi sensori, quindi senza conoscerne la precisa collocazione.

La prima matrice di adiacenza, denominata 'sp1', è stata creata in base ai seguenti criteri:

- nel caso in cui un sensore è appartenente alla zona esterna vengono presi in esame i sensori molto prossimi al sensore considerato e il suo corrispondente verticale;
- nel caso in cui un sensore è appartenente alla zona interna vengono presi in considerazione i sensore più vicini, più precisamente le zone termiche più vicine.

Differentemente, la seconda matrice, denominata 'sp2', è stata creata con in base a una concezione della vicinanza fisica dei sensori più ampia, seguendo i seguenti criteri:

- nel caso in cui un sensore è appartenente alla zona esterna vengono presi in esame i sensori prossimi al sensore considerato, il suo corrispondente verticale e i sensori vicini al suo corrispondente verticale;
- nel caso in cui un sensore è appartenente alla zona interna vengono presi in esame i sensori prossimi al sensore considerato.

La matrice di adiacenza, costruita con l'approccio basato sulla similarità, in un primo momento viene creata con il metodo del KNN. Attraverso tale metodo per ogni elemento, in questo caso il sensore, vengono trovati i punti più vicini per la similarità delle serie storiche.

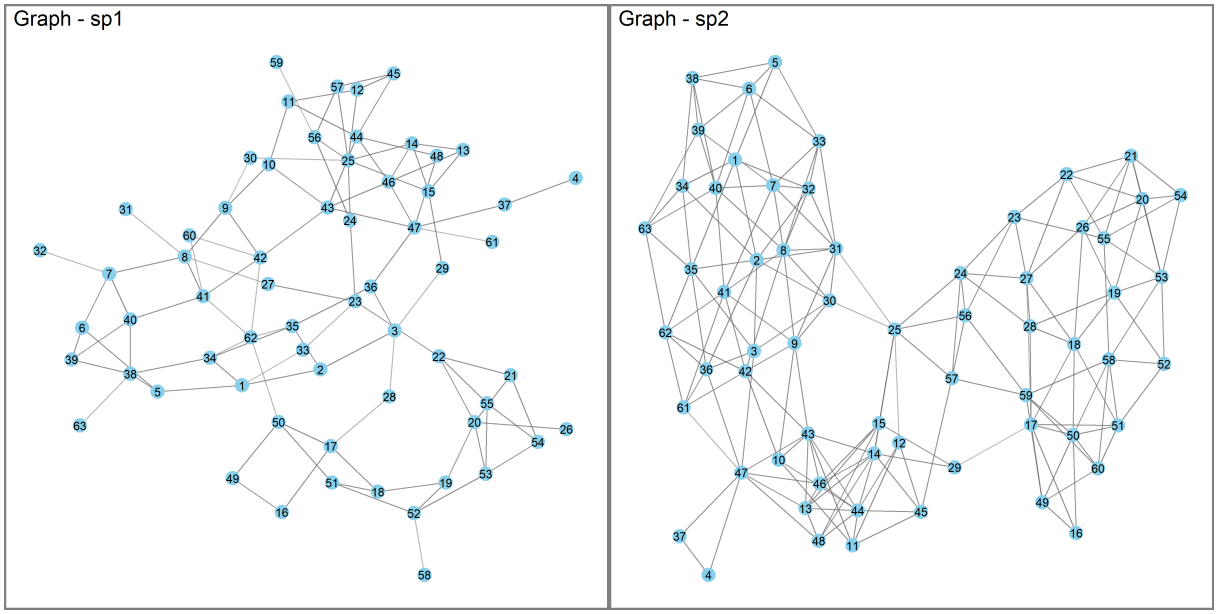


Figura 4.16: Grafi da matrici spaziali

Nell'individuazione del punto più vicino, all'aumento del valore assegnato a k si è notato che arrivando a 3 si viene a costituire un unico grafo, evitando quei casi che presentano più sotto-grafi, mostrati nella figura 4.17. Oltre a ciò sono stati creati dei grafi con il valore di k pari a 4 e a 5.

Dopo di che, oltre ai grafi ottenuti attraverso il KNN, sulla base di essi sono stati aggiunti ulteriori collegamenti trovati attraverso la correlazione di Pearson. In pratica, imposto un limite inferiore al coefficiente, se la correlazione tra questi due sensori risulta essere maggiore del limite stabilito si ammette l'esistenza del collegamento tra i due nodi. In questo caso, per non eccedere con i numeri dei collegamenti, sono stati scelti limiti pari a 0.7, a 0.75 e a 0.8. Questi grafi sono utili per svolgere l'analisi di sensibilità al fine di scegliere la configurazione migliore per fare un successivo paragone con i risultati ottenuti per mezzo della matrice spaziale e con i risultati di predizione del modello di LSTM.

I risultati del GNN dipendono dai collegamenti che sussistono tra i nodi, di conseguenza al mutare della matrice di adiacenza cambia il modello allenato, poiché nella prima variano le relazioni e nel secondo i risultati predetti.

Per tale motivo è stata eseguita una prima analisi di sensibilità, come mostrato nella figura 4.19. Tale analisi si è basata su un confronto delle metriche dei risultati di predizione

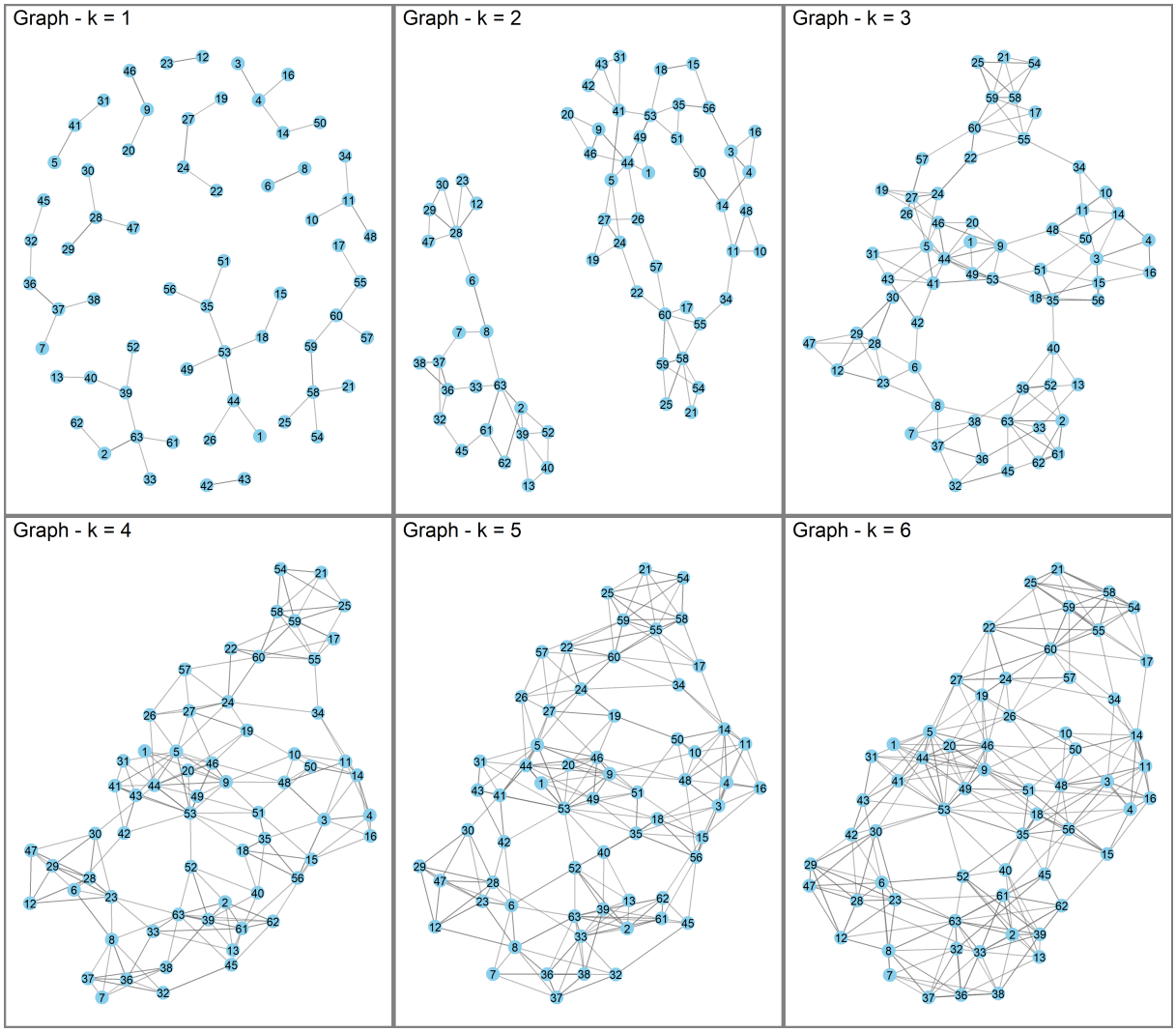


Figura 4.17: Grafi da matrici con KNN

a tre timestep in avanti nelle diverse configurazione della matrice di adiacenza. L'immagine si divide in due parti, a destra ci sono i boxplot delle matrici di similarità mentre a sinistra quelli delle matrici spaziali. In particolare, il grafico a sinistra sull'asse delle ascisse riporta il numero di punti più vicini e sul piano cartesiano contiene i boxplot raggruppati in base allo stesso valore di k.

In generale, si nota che non sussiste alcuna relazione tra i boxplot e il numero di k o il coefficiente di Pearson, poiché osservando solo i boxplot con il k pari a 4 o 5 si potrebbe affermare che i risultati di predizione peggiorano al crescere del limite del coefficiente,

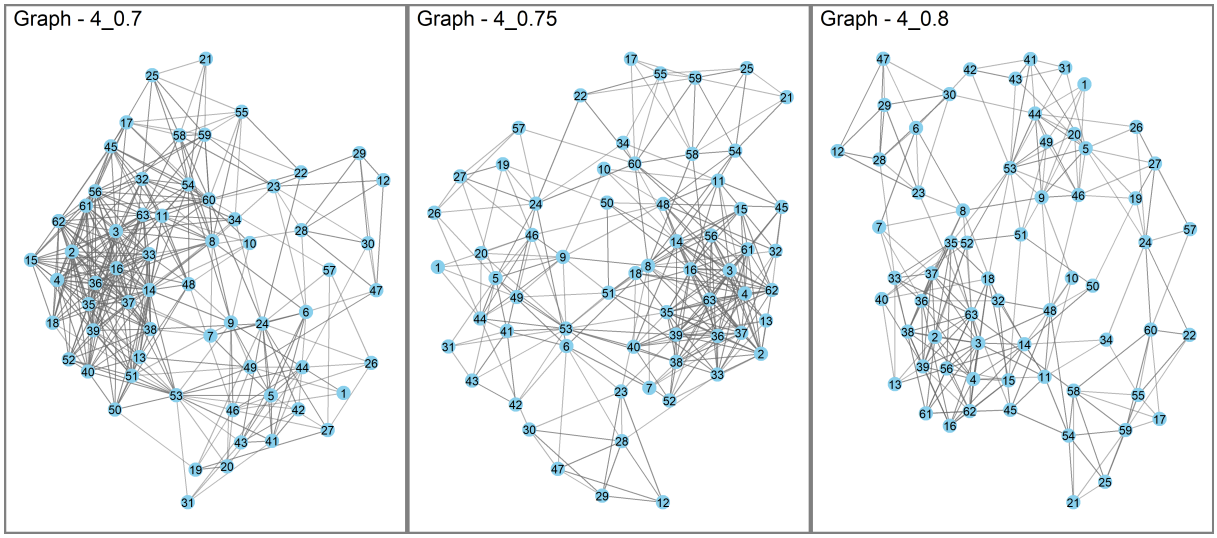


Figura 4.18: Grafi da matrici con KNN e correlazione di Pearson

però tale osservazione non risulta valida nel caso in cui il k è uguale a 3. Differentemente per quanto riguarda la parte spaziale, i boxplot della prima matrice potrebbero essere paragonati con gli altri boxplot delle matrici di similarità, mentre i boxplot della seconda matrice spaziale potrebbero essere paragonati con quelli delle matrici di similarità di migliore prestazione.

Successivamente, è stata compiuta un'analisi di sensibilità in relazione al numero dei collegamenti esistenti in un grafo, come nella figura 4.20, dove i boxplot sono messi in ordine crescente sull'asse x, i punti fanno riferimento all'asse y sulla destra e indica la mediana del numero di collegamenti di ciascun nodo. Attraverso il grafico è possibile confermare che le prestazioni dei modelli del GNN sono indipendenti dal numero di edge e dal numero di collegamenti del ciascun nodo del grafo, poiché dipendono dalla configurazione del grafo.

In seguito, lo studio si è concentrato sul confronto di due configurazioni dei grafi di migliore prestazione, una per ciascun tipologia di matrice, rispettivamente tra le matrici '4_0.7' e 'sp2'. La figura 4.21 mostra i violinplot delle metriche riguardo ai risultati delle due matrici, rispettivamente quella del MSE, quella del MAE e quella della MAPE. Inoltre, il grafico è costituito da due parti, rispettivamente i violinplot della matrice di similarità e quelli della matrice spaziale. In esso i valori sull'asse delle ascisse rappresentano il timestep di previsione in avanti, i punti all'estremità di ciascun violinplot rappresentano le massime e le minime dell'insieme delle metriche e il colore rappresenta il sensore.

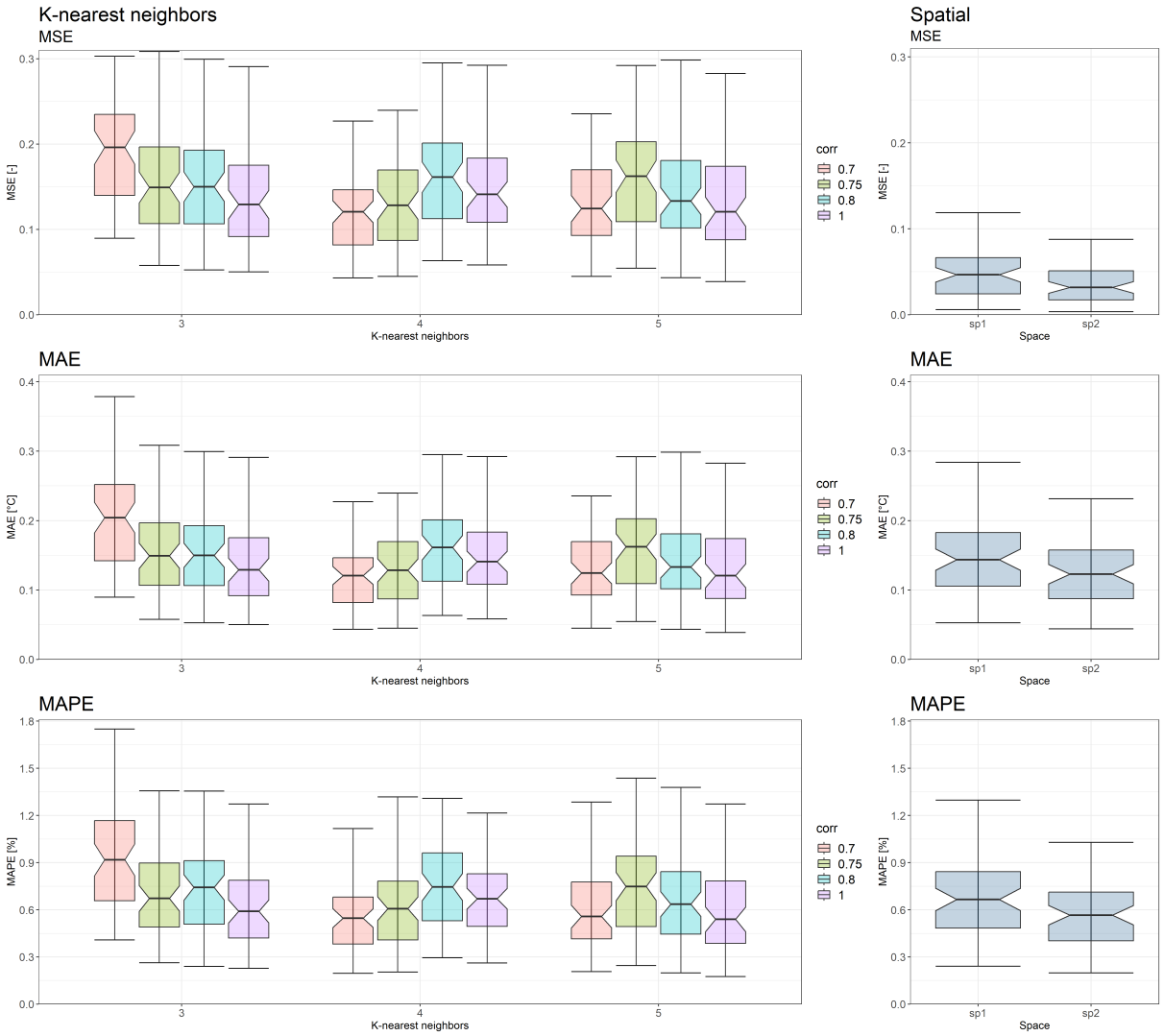


Figura 4.19: L'analisi di sensibilità attraverso con diverse configurazioni di matrici, spaziali e di similarità.

Come previsto, i valori delle metriche incrementano in correlazione al timestep di predizione. La forma del violinplot descrive la distribuzione dei valori, infatti se la forma è molto schiacciata il range dei valori è molto stretto e viceversa. Infatti, con l'aumentare dell'unità del timestep la forma dei violinplot si snellisce, in particolare, la sua parte superiore diventa più snella rispetto al restante e ciò va a indicare che solo in pochi sensori i valori di predizione sono peggiorati in maniera più evidente rispetto al complesso dei

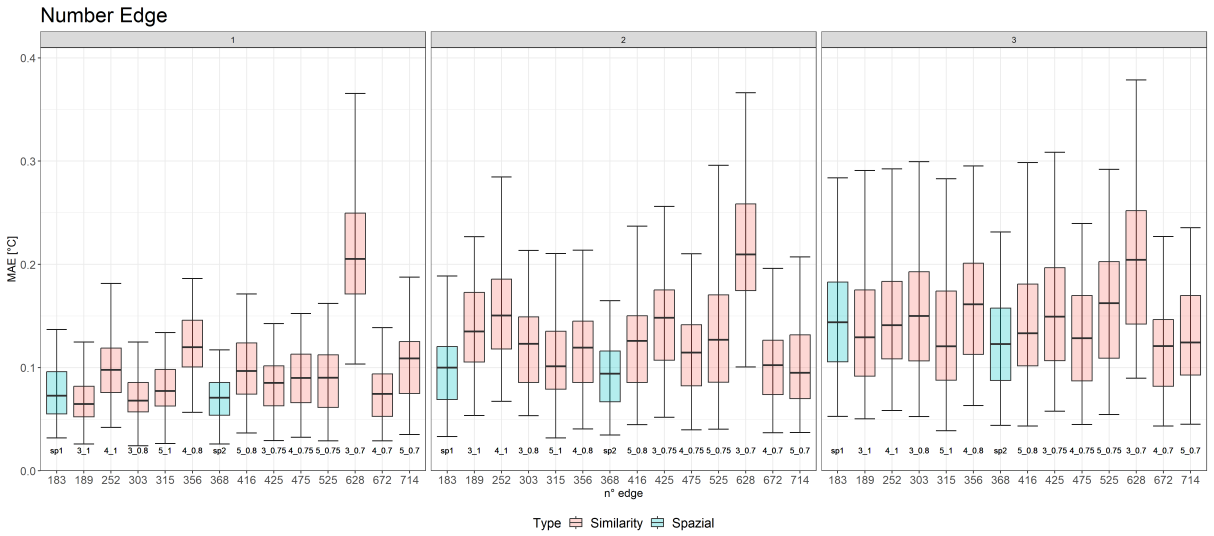


Figura 4.20: L'analisi di sensibilità attraverso il boxplot in relazione al numero di edge.

sensori. Il limite inferiore e la mediana del range dei valori sono aumentati leggermente, mentre il limite superiore è cresciuto di un ordine di grandezza.

Inoltre, è possibile constatare, in relazione alla tipologia delle metriche e al numero di timestep in avanti, che sono sempre gli stessi sensori ad avere una peggiore performance. Ad esempio, il sensore 23 risulta essere il peggiore sia in base alla configurazione per similarità che in base a quella spaziale. La stessa situazione non è riscontrata per i sensori con una migliore performance, di conseguenza è possibile ipotizzare che sono presenti molti sensori con ottime performance.

Per concludere, è possibile paragonare tra loro le metriche delle due configurazioni. Durante il confronto della difficoltà di creazione della matrice di adiacenza emerge che quella per similarità è molto più semplice e automatizzata.

Osservando le figure 4.22 e 4.23, nelle quali il colore di ciascun nodo rappresenta il MAE normalizzato, si può affermare che i sensori con cattive performance sono sempre gli stessi anche con il cambiare del timestep, indipendentemente dalla matrice di adiacenza, dal numero di collegamenti e dalla complessità del grafo. Infatti, una maggiore complessità del grafo non incide sull'ottimizzazione dei risultati, mentre la numerosità dei collegamenti di un nodo non condiziona la qualità dei risultati del nodo. Ad esempio, il sensore 23 è molto collegato ad altri nodi nella matrice di adiacenza di similarità e poco relazionato ad altri

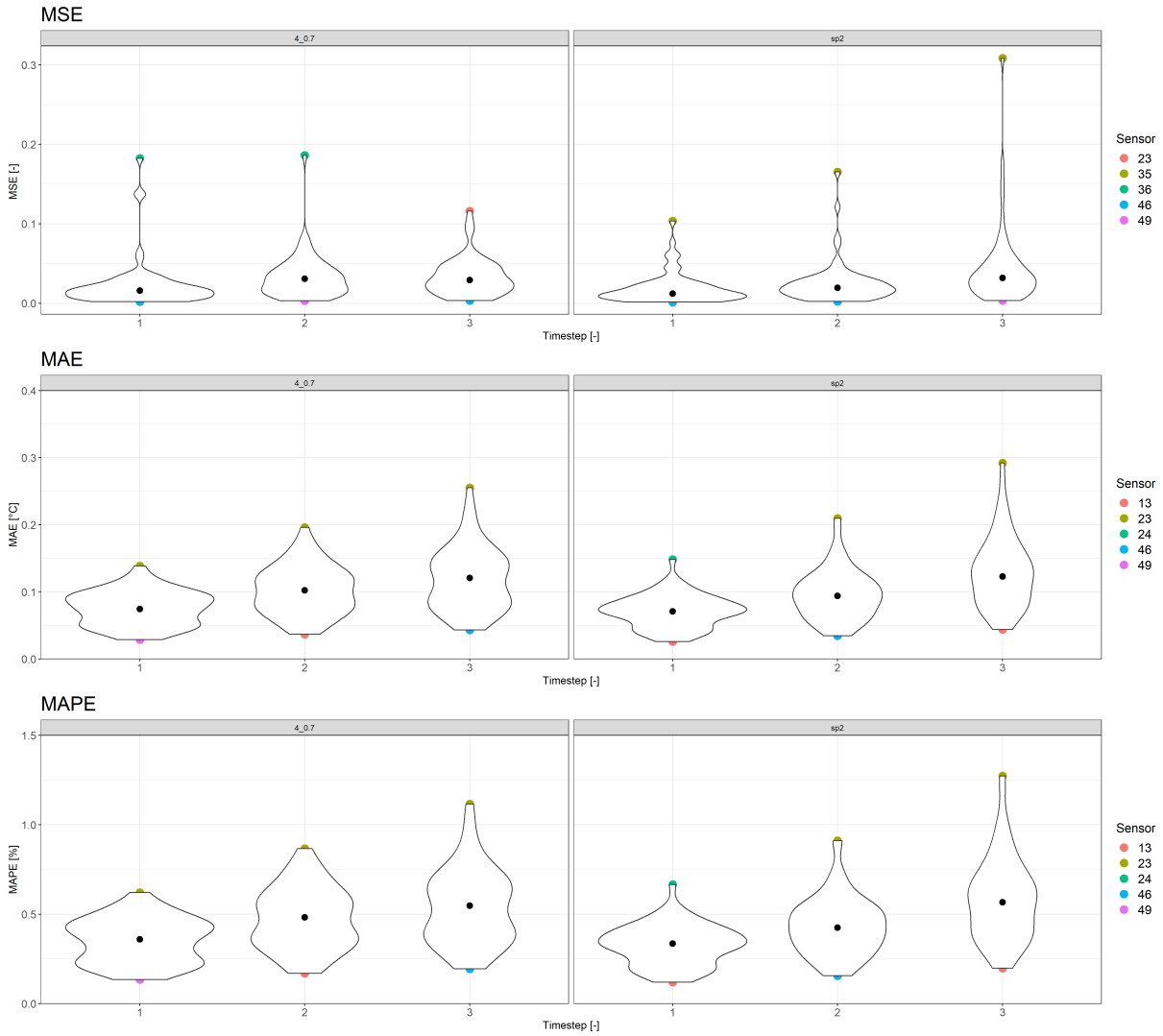


Figura 4.21: Violinplot delle metriche dei risultati di 2 configurazioni di matrice differenti

nodi nella matrice spaziale, nonostante risulta essere il sensore con peggiore performance in entrambi modelli di predizione.

La stessa affermazione è valida per i sensori con buone performance. Ad esempio, il sensore 13 ha molti edge rispetto ad altri sensori nella matrice spaziale e pochi collegamenti nella matrice di similarità, nonostante risulta essere il sensore con la migliore performance.

Il grafico 4.24 mostra i grafi della matrice di similarità in funzione del MAE calcolato mensilmente e con la stessa legenda. Il colore del nodo rappresenta il valore del MAE, di

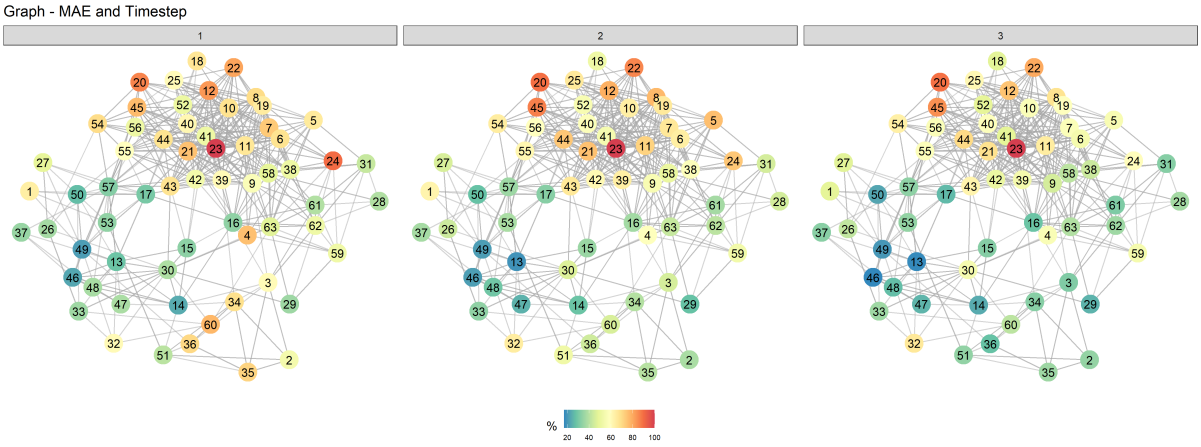


Figura 4.22: Grafi della matrice di similarità in relazione al MAE

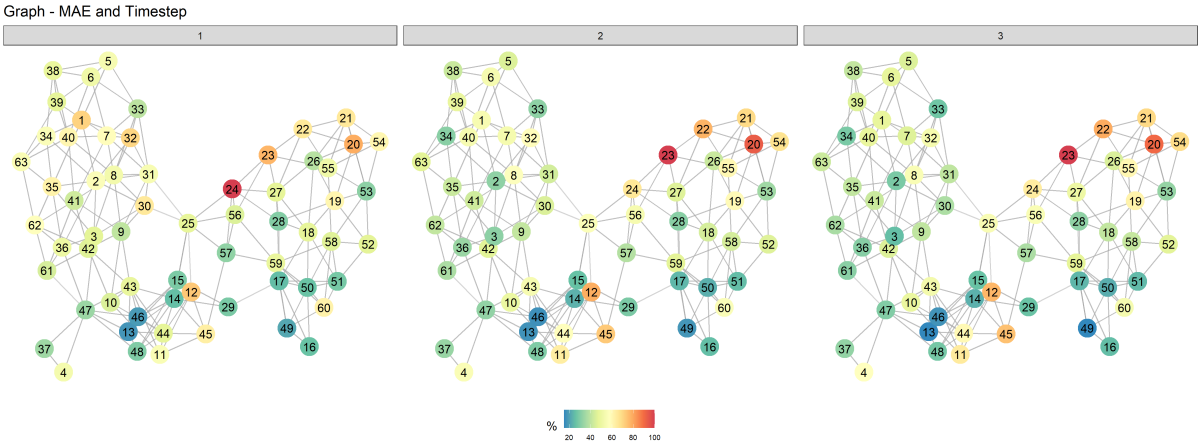


Figura 4.23: Grafi della matrice spaziale in relazione al MAE

conseguenza all'aumentare del valore il colore si associa a tonalità più calde. Si osserva che i valori di MAE variano in base al mese, nonostante rimangono costanti le performance dei sensori. I MAE dei sensori nei mesi che vanno da aprile a settembre, nel complesso, sono migliori degli altri mesi, che mostrano un discostamento maggiore dei valori di previsione rispetto ai valori reali.

In seguito, vengono illustrati i risultati di predizione dei sensori con la migliore e la peggiore performance rispetto ad entrambe le tipologie di matrice.

Il grafico è formato da più parti, una prima parte si concentra sui valori della temperatura registrati nell'anno 2019 che costituiscono il dataset di training, una seconda parte si

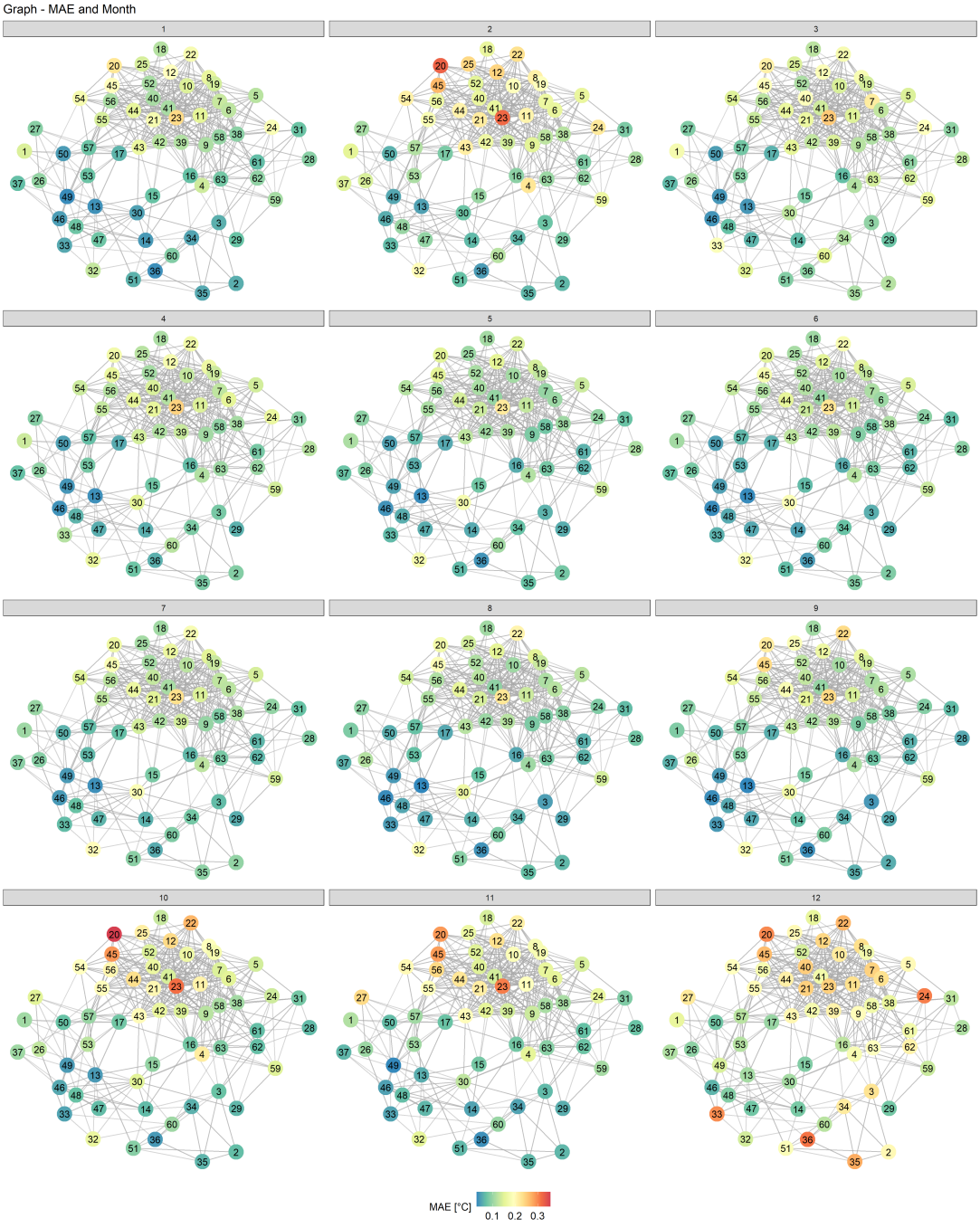


Figura 4.24: Grafi in relazione al MAE mensile della matrice di similarità

concentra su un'illustrazione dei valori predetti confrontati con i valori realmente misurati nell'anno 2020 e ultima parte si concentra su un focus dei risultati di febbraio e di dicembre. Questi due mesi sono stati scelti al fine di effettuare una comparazione, in quanto si è notato che: nel mese di dicembre la temperatura interna rilevata dai sensori ha registrato un calo della temperatura per tutti i sensori e quindi si è trattato di un mese in cui è risultato più difficile fare predizioni; e nel mese di febbraio i valori della temperatura sono risultati in linea con il dataset restante.

A questo punto sono stati presi in considerazione i risultati dei sensori a tre timestep in avanti. Nel caso della matrice spaziale sono stati ripresi i risultati dei sensori 13 e 23 in riferimento al MAE e al MAPE, mentre nel caso della matrice di similarità sono stati esaminati i risultati dei sensori 46 e 23, sempre in riferimento alle stesse metriche. I valori reali sono stati plottati con il color celeste, mentre i valori di previsione con il colore rosso.

Matrice spaziale

Sono stati rappresentati i risultati dei sensori 23 e 13, cioè i sensori che rappresentano rispettivamente la migliore performance e la peggiore performance.

Nel caso dei risultati del sensore 23 si notano le seguenti osservazioni:

- i valori misurati nel 2019 oscillano nell'intervallo tra i 21 °C e i 26 °C, facendo verificare una grande variazione della temperatura nell'arco di una giornata;

- le temperature interne sono minori nei mesi febbraio, marzo, novembre e dicembre, in particolare in essi la variazione è minore nell'intervallo giornaliero rispetto agli altri periodi.

- i valori realmente rilevati nel 2020 sono minori in confronto a quelli del 2019, così come lo sono le variazioni della temperatura.

- il modello non riesce a predire con accuratezza i picchi di temperatura per i valori superiori ai 25°C, mentre i valori di previsione di picco sono maggiori rispetto a quelli reali per i valori fino ai 25°C. In più in generale, i valori minimi predetti risultano essere inferiori a quelli reali.

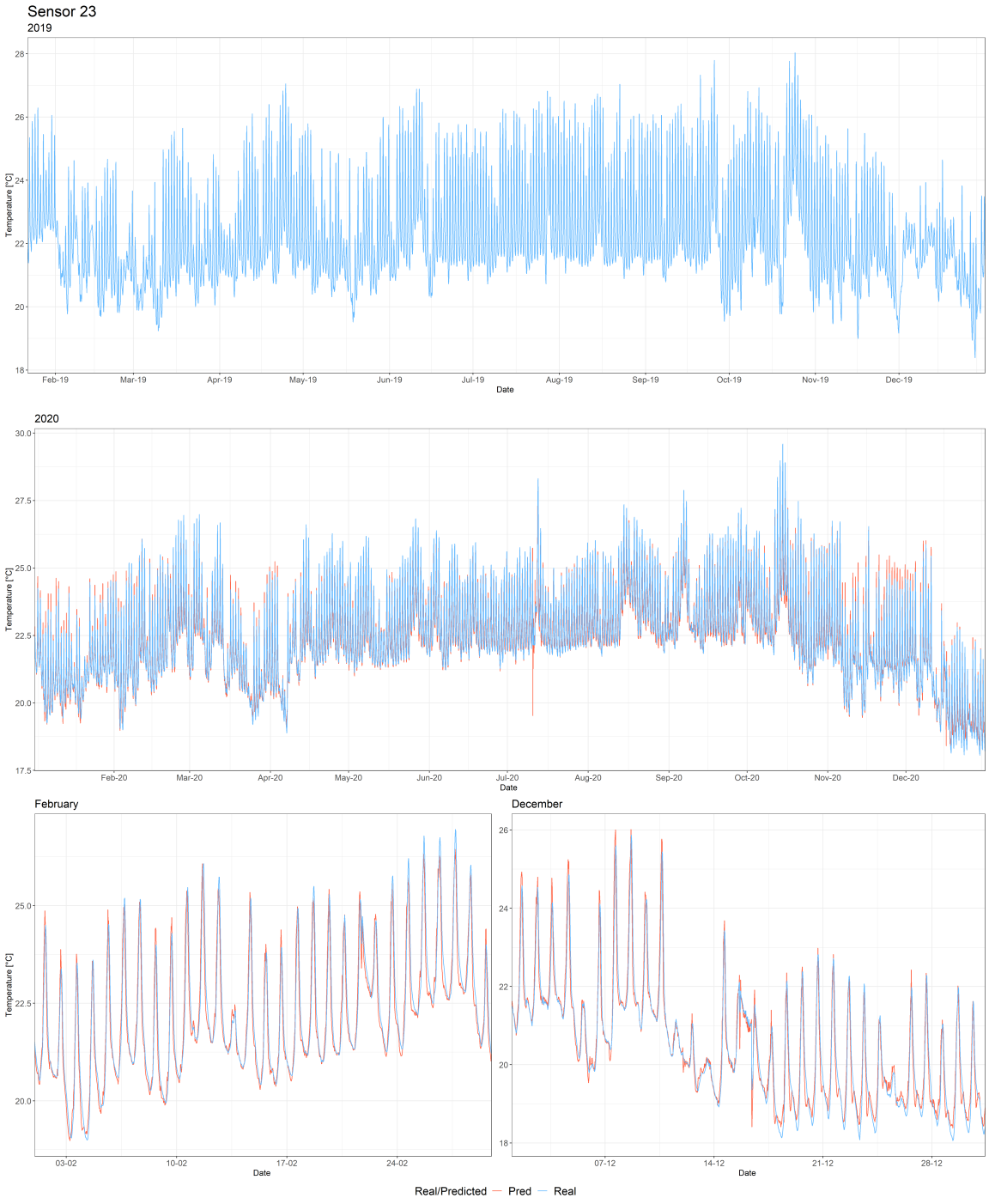


Figura 4.25: Plot del sensore 23.

- il mese di luglio ha una previsione insolita, in particolare in esso è avvenuto un abbassamento brusco della temperatura che non si è verificato nella realtà. Tale fenomeno è stato osservato anche nei altri sensori nello stesso periodo temporale.

- i risultati di febbraio hanno mostrato un andamento della temperatura predetta perfettamente in linea con i valori reali a eccezione dei picchi in cui la deviazione è stata di pochi decimi di gradi. I sensori volti a misurare non sono accurati a livello di laboratorio quindi è normale avere un errore di misura fino ai 0.5°C. Quindi, sono accettabili queste differenze tra i valori reali e quelli predetti.

- il mese di dicembre ha riscontrato un abbassamento della temperatura interna nella seconda metà del mese, infatti, mentre solitamente i valori della temperatura normale oscillano tra i 21 e i 25 °C, in questo caso si sono abbassati fino ai 18 °C, nonostante la precisione della previsione della temperatura non è diminuita.

Spostandosi sui risultati del sensore 13 di migliore performance si nota la seguente situazione.

In questo caso, al contrario del sensore 23, i valori misurati dal sensore 13 hanno un range di variazione della temperatura più stretto, infatti, in generale i valori fluttuano intorno ai 22°C, tranne per i pochi giorni in cui si registrano un massimo di 25°C.

Per cui, i valori predetti differiscono di meno di un decimo di grado dai valori reali sia nel mese di febbraio che nel mese di dicembre, nonostante hanno il medesimo abbassamento della temperatura. Nella tabella 4.3 sono presenti il riassunto delle metriche, i quali il MSE, il MAE e il MAPE, delle varie timestep di previsione. Sono presenti i valori delle metriche dei risultati della previsione. La tabella contiene il valore minio, la mediana, la media e il massimo del range dei valori delle metriche.

Matrice di similarità

Per quanto concerne i risultati di predizione del modello compiuto mediante la matrice di similarità, è possibile osservare che i valori del MAE e del MAPE del sensore 23 risultano essere i peggiori, come è accaduto nel caso della matrice spaziale. A differenza, il sensore 46 ha il migliore MAE e il migliore MAPE rispetto agli altri.

Complessivamente, nel caso del sensore 23 i valori predetti sono leggermente inferiori rispetto ai valori reali, con un discostamento massimo di 0.4 - 0.5 °C. Contrariamente

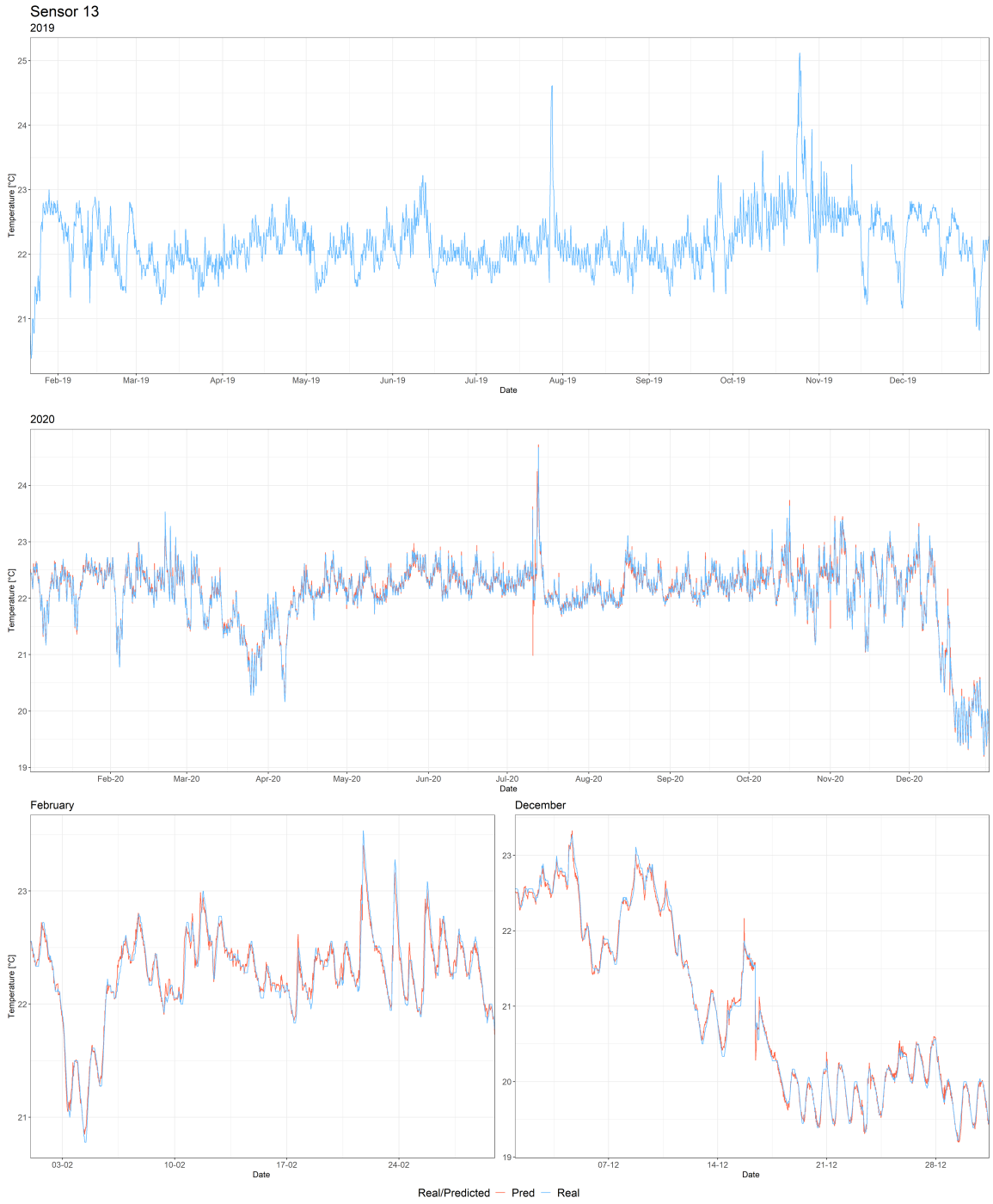


Figura 4.26: Plot del sensore 13.

MSE [-]				
Timestep	Min	Mediana	Media	Max
1	0.001	0.012	0.019	0.104
2	0.002	0.019	0.026	0.165
3	0.003	0.032	0.044	0.309
MAE [°C]				
Timestep	Min	Mediana	Media	Max
1	0.026	0.071	0.071	0.148
2	0.035	0.094	0.096	0.210
3	0.044	0.123	0.129	0.292
MAPE [%]				
Timestep	Min	Mediana	Media	Max
1	0.120	0.336	0.330	0.666
2	0.156	0.424	0.434	0.911
3	0.199	0.566	0.584	1.273

Tabella 4.3: Metriche del modello GNN - Spaziale

la differenza è minima quando sussiste una diminuzione della temperatura, come si è verificato nella seconda metà di dicembre.

Il sensore 46 ha la stessa caratteristica del sensore 13, cioè ha un intervallo di variazione molto piccolo se paragonato agli altri sensori, come ad esempio al sensore 23. Nel complesso, tale intervallo ha come estremi i 21 e i 23°C, nonostante solitamente a livello giornaliero ha variazioni minori. Questa variazione non è così periodica come si osserva nel caso degli altri sensori, per questo motivo i valori della previsione a differenza di quelli degli altri sensori non sembrano così accurati e l'imprecisione non si verifica solo nei picchi e nei minimi, di conseguenza generalmente i valori reali sono leggermente maggiori di quelli predetti. Nonostante ciò, il MAPE e il MAE hanno indicato che i risultati del sensore 46 sono i migliori per via del loro trascurabile discostamento. Nella tabella 4.4 sono presenti il riassunto delle metriche, i quali il MSE, il MAE e il MAPE, delle varie timestep di previsione. Sono presenti i valori delle metriche dei risultati della previsione. La tabella contiene il valore minio, la mediana, la media e il massimo del range dei valori delle metriche.

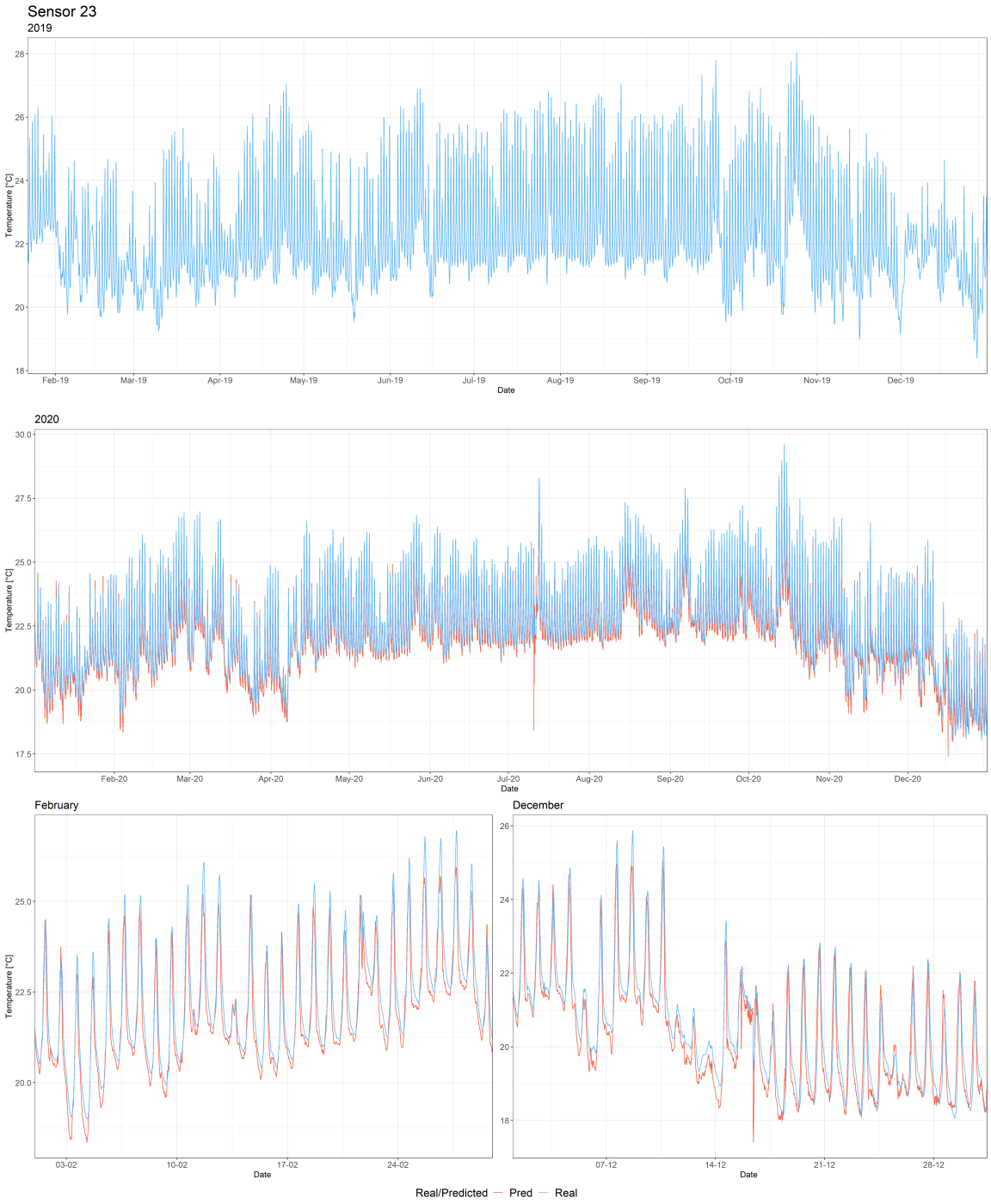


Figura 4.27: Plot del sensore 23.

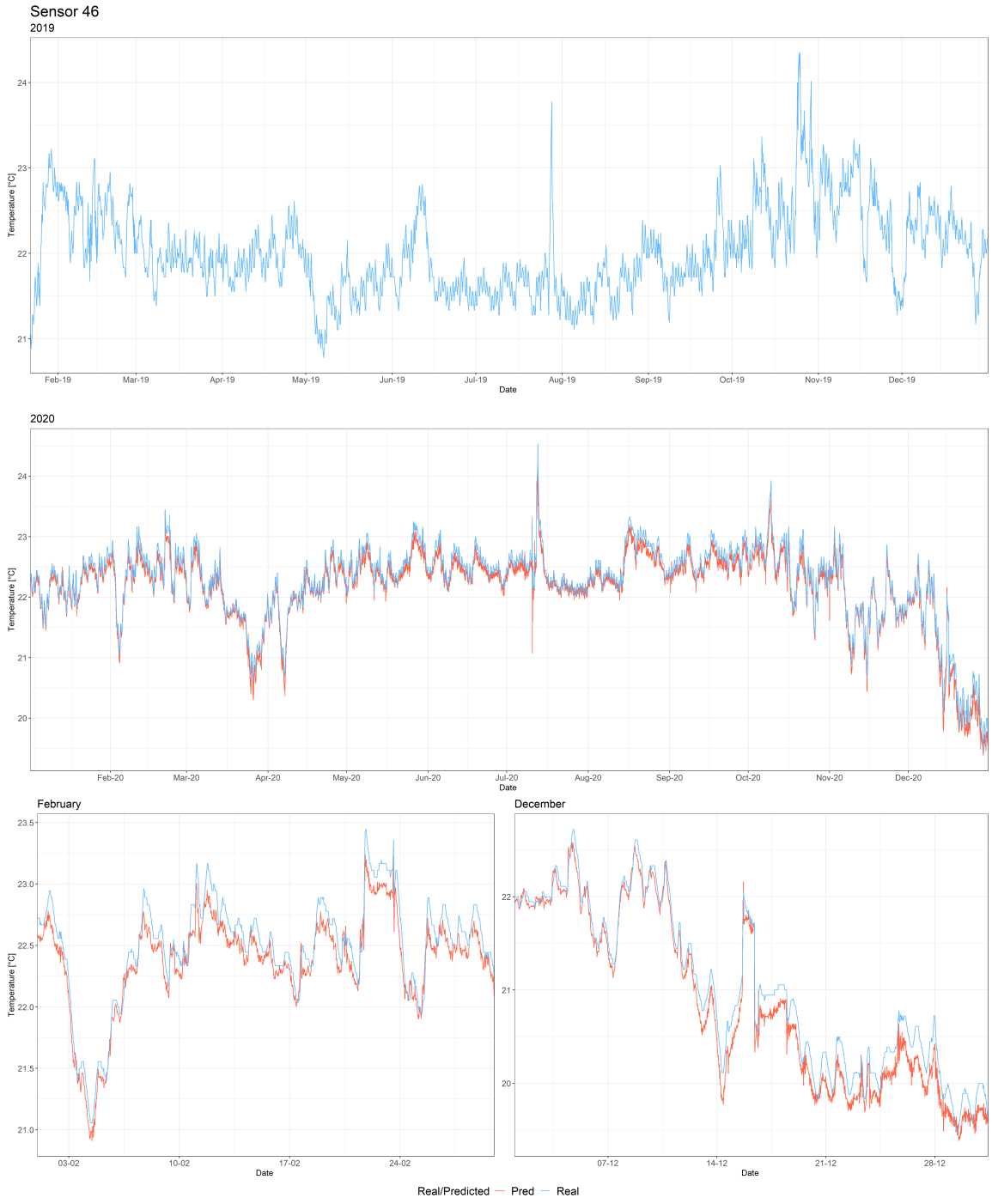


Figura 4.28: Plot del sensore 46.

MSE [-]				
Timestep	Min	Mediana	Media	Max
1	0.002	0.016	0.024	0.182
2	0.003	0.031	0.033	0.186
3	0.003	0.029	0.034	0.116
MAE [°C]				
Timestep	Min	Mediana	Media	Max
1	0.029	0.074	0.076	0.139
2	0.037	0.102	0.103	0.196
3	0.043	0.121	0.119	0.255
MAPE [%]				
Timestep	Min	Mediana	Media	Max
1	0.135	0.360	0.353	0.621
2	0.170	0.482	0.475	0.866
3	0.195	0.547	0.540	1.116

Tabella 4.4: Metriche del modello GNN - Similarità

4.4.2 Baseline - Long Short Term Memory

Per avere una buona comparazione tra i due modelli, al il modello di LSTM Neural Network sono stati dati gli stessi iper-parametri e la stessa impostazione. Di conseguenza si è tratto di un modello costituito da:

- un input layer, un output layer e due hidden layer, entrambi dei quali composti da 64 LSTM units;
- un RMSprop come optimezer;
- un lookback pari a 12 e tre timestep in avanti;
- un batch size uguale a 64;
- un learning rate corrispondente a 2 E-4.

L'utilizzo di questo algoritmo non richiede nessun grafo, infatti viene allenato un modello di predizione per ogni sensore costruendo in tutto 63 modelli diversi.

Nella figura 4.29 sono mostrati i boxplot delle metriche, rispettivamente di MSE, di MAE e di MAPE fino a tre timestep in avanti. In questo modo è possibile nota che i

limiti inferiori dell'intervallo dei valori al cambiare del timestep sono gli stessi, mentre le mediane e gli estremi superiori aumentano progressivamente, come ipotizzato, e nei limiti accettabili dei valori. Inoltre confrontando i boxplot delle MSE può notare che i valori del MSE del modello LSTM sono minori dei modelli di GNN, perché questa metrica è più sensibile agli errori in quanto sono al quadrato.

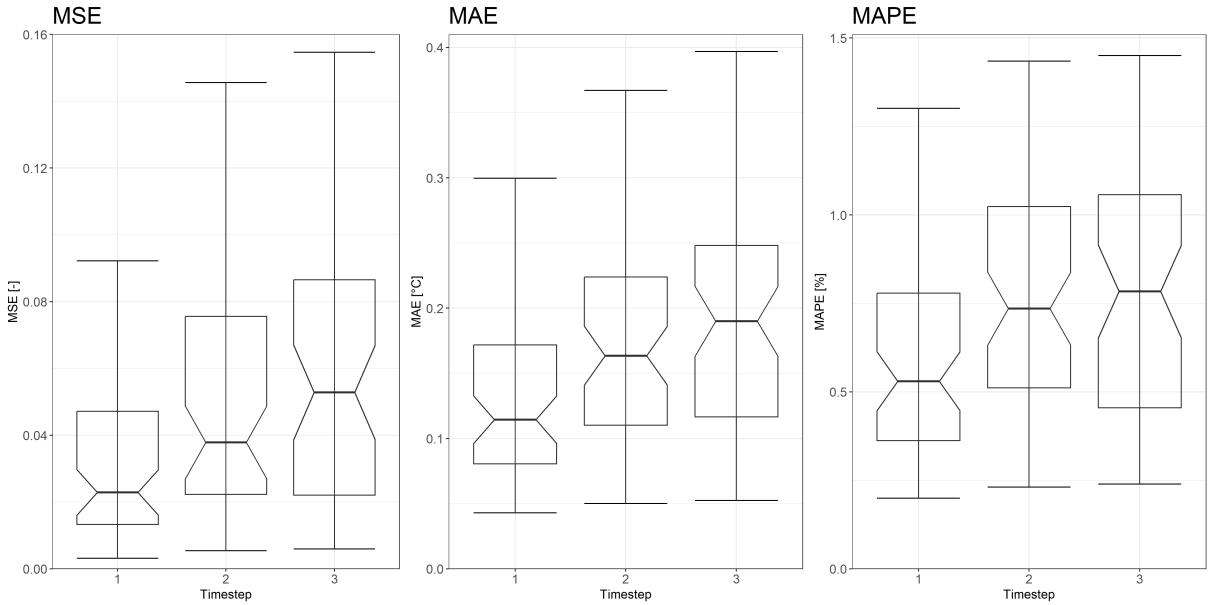


Figura 4.29: Boxplot delle metriche dei risultati dei modelli LSTM

Invece plottando le metriche di tutti modelli, come mostrato nella figura 4.30, attraverso i violinplot è possibile osservare che le forme dei violinplot delle metriche sono più slanciate e tendenti a valori superiori. Di conseguenza è possibile affermare che la prestazione dei modelli di LSTM è abbastanza buona ed è comparabile con il modello di GNN per alcuni sensori. Per tutte le metriche e per tutti i timestep i sensori con la migliore e la peggiore prestazione non risultano quasi mai essere gli stessi.

Guardando ai plot dei valori del sensore 45 nella figura 4.31, si evidenziano le seguenti considerazioni:

- la variazione della temperatura interna tra i picchi e i minimi è intorno ai 21 e i 25 °C, sia per l'anno 2019 che per il 2020, a eccezione dei mesi della stazione primaverile del 2020 per cui sia le variazioni giornalieri che i valori della temperatura diminuiscono, che di fatto è variata tra i 21 e i 23 °C.

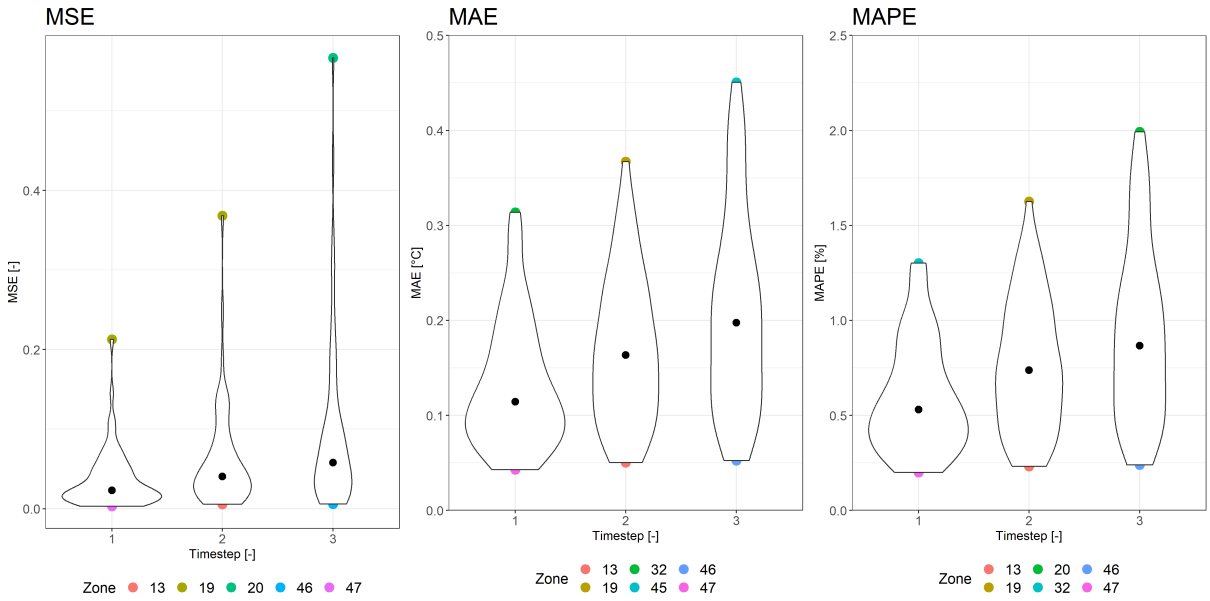


Figura 4.30: Violinplot delle metriche dei risultati dei modelli LSTM

- i valori di previsione di picchi sono minori dei valori reali, in caso di valori reali superiori a 25°C, lo discostamento può raggiungere 1°C. A differenza, nella seconda metà del dicembre i valori predetti sono perfettamente in linea con i valori reali.

Rispetto ai risultati del sensore 46 è possibile affermare che il modello di previsione è molto accurato e l'errore assoluto è molto piccolo, poiché è inferiore a 0.1°C ad eccezione, dei valori minori di 20°C, e in tal caso la temperatura di previsione è leggermente superiore alla temperatura reale, causando un errore di pochi decime di gradi. Nella tabella 4.5 sono presenti il riassunto delle metriche delle varie timestep di previsione. Sono presenti i valori delle metriche dei risultati della previsione. La tabella contiene le stesse grandezze calcolate per altri due modelli di previsione.

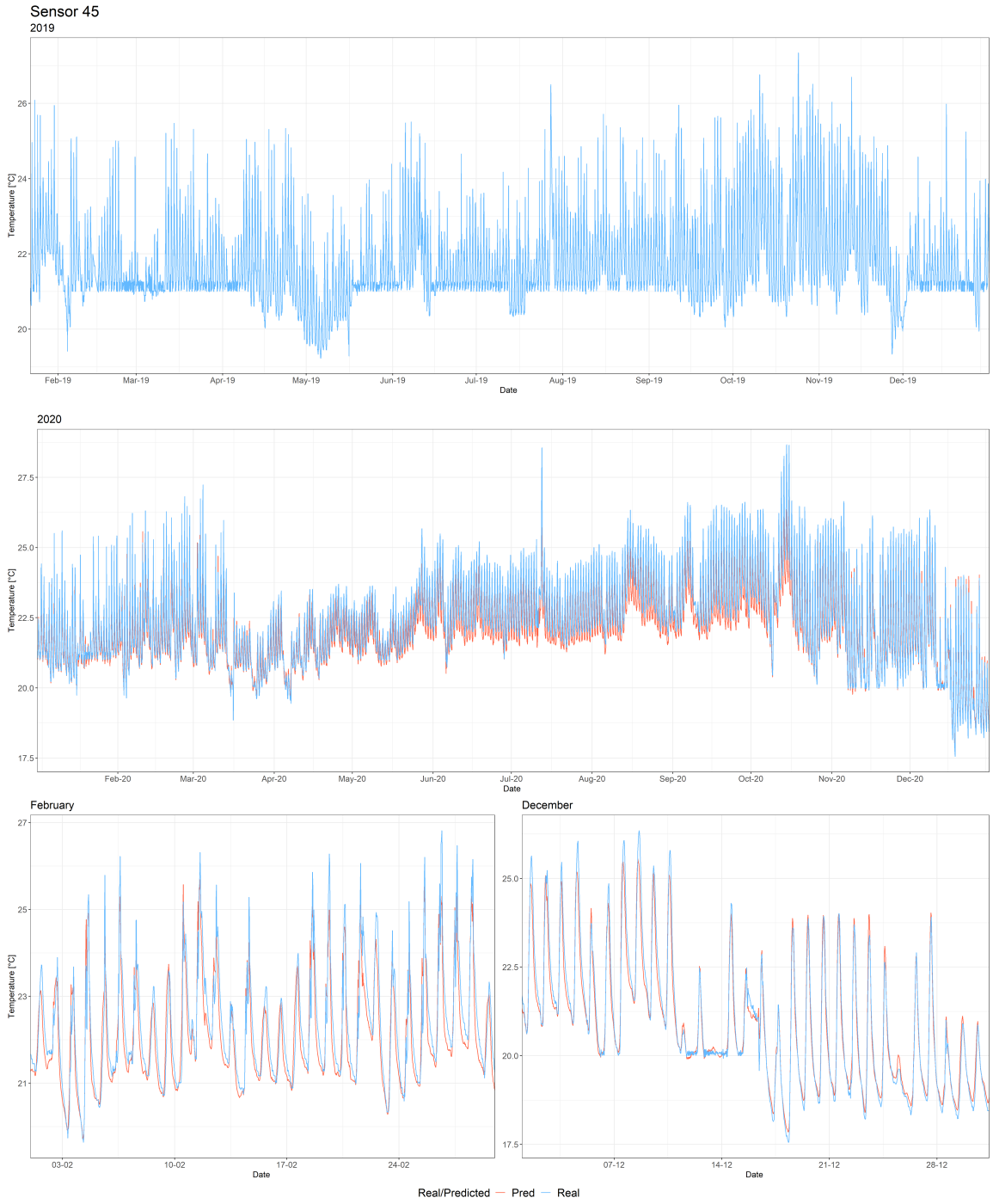


Figura 4.31: Plot del sensore 45.

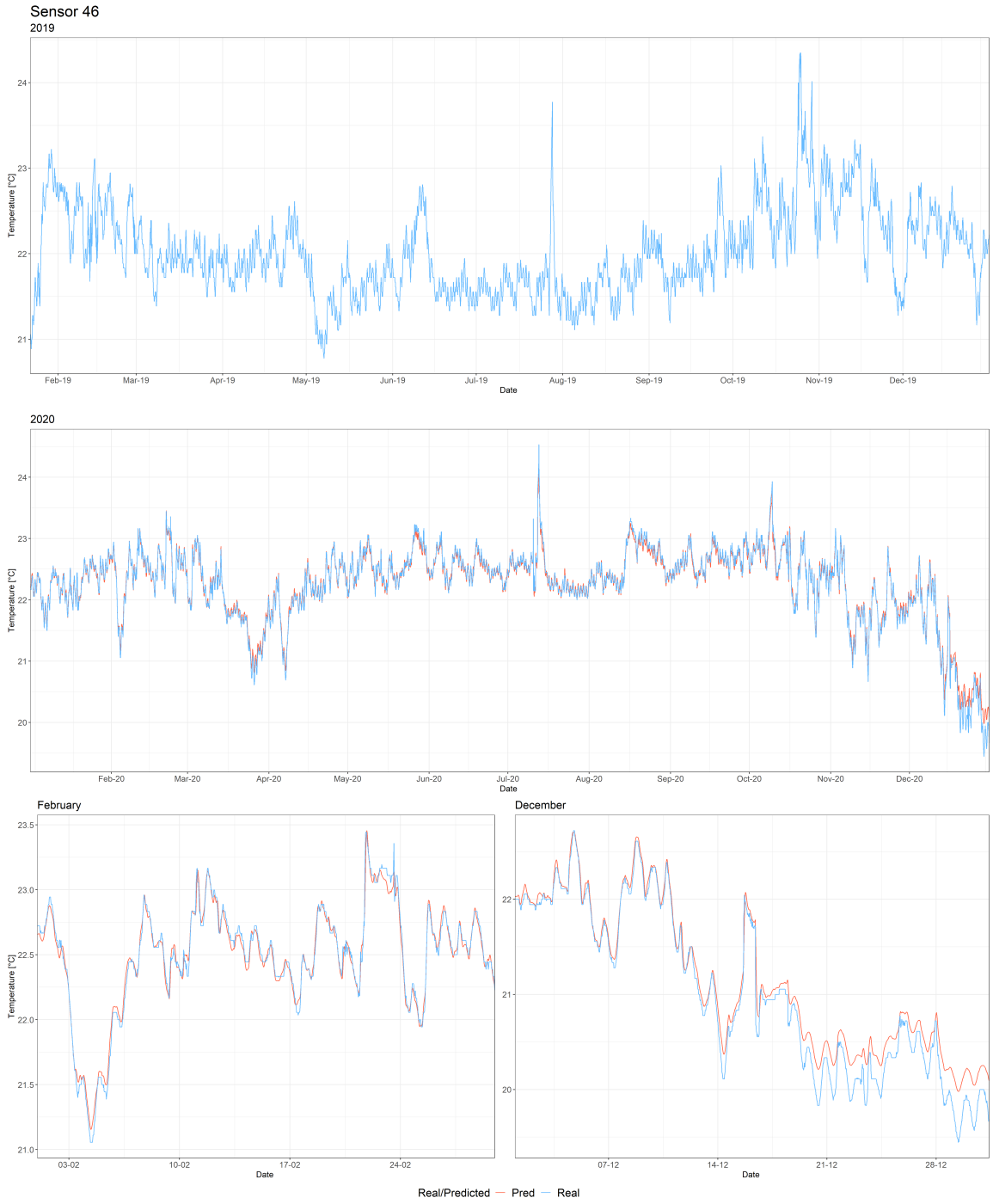


Figura 4.32: Plot del sensore 46.

MSE [-]				
Timestep	Min	Mediana	Media	Max
1	0.003	0.023	0.037	0.213
2	0.005	0.041	0.062	0.368
3	0.006	0.058	0.098	0.567
MAE [°C]				
Timestep	Min	Mediana	Media	Max
1	0.043	0.114	0.132	0.314
2	0.050	0.164	0.170	0.367
3	0.052	0.198	0.208	0.451
MAPE [%]				
Timestep	Min	Mediana	Media	Max
1	0.200	0.530	0.599	1.302
2	0.231	0.739	0.766	1.626
3	0.240	0.867	0.937	1.992

Tabella 4.5: Metriche del modello LSTM

4.4.3 Confronto tra modelli di previsione

Il confronto tra i tre modelli di predizione, come mostrato nella figura 4.33, è stato effettuato attraverso il MAE dei risultati di predizione a tre timestep in avanti, rispettivamente per il baseline (LSTM), il STGNN avendo come matrice di adiacenza con approccio di similarità e il STGNN con la matrice di adiacenza con approccio spaziale.

Nella mappa ogni punto rappresenta la posizione di ciascun sensore e il colore raffigura il valore del MAE in base a quanto stabilito attraverso la legenda nella parte inferiore del grafico, la quale risulta valida per tutte le mappe. In questo modo, all'aumentare del MAE il colore dei punti tende ad assumere tonalità più calde.

Nel complesso è possibile riscontrare che i sensori esposti a ovest hanno i MAE maggiori rispetto ai sensori esposti ad est, di conseguenza i punti ad ovest tendono a giallo-rosso, mentre i sensori esposti ad est tendono per la maggior parte al verde-blu. Questo fenomeno si verifica per tutti i modelli, quindi è indipendente dal modello di predizione e strettamente dipendente dai dati di ciascun sensore. Infatti, come già discusso prima, in riferimento alla figura 4.8, la media della differenza della temperatura tra la massima e la minima giornaliera è maggiore nei sensori esposti ad ovest, come ad esempio nel caso del sensore 23. Inoltre, come mostra la figura 4.11, la percentuale del cluster principale dei sensori è minore

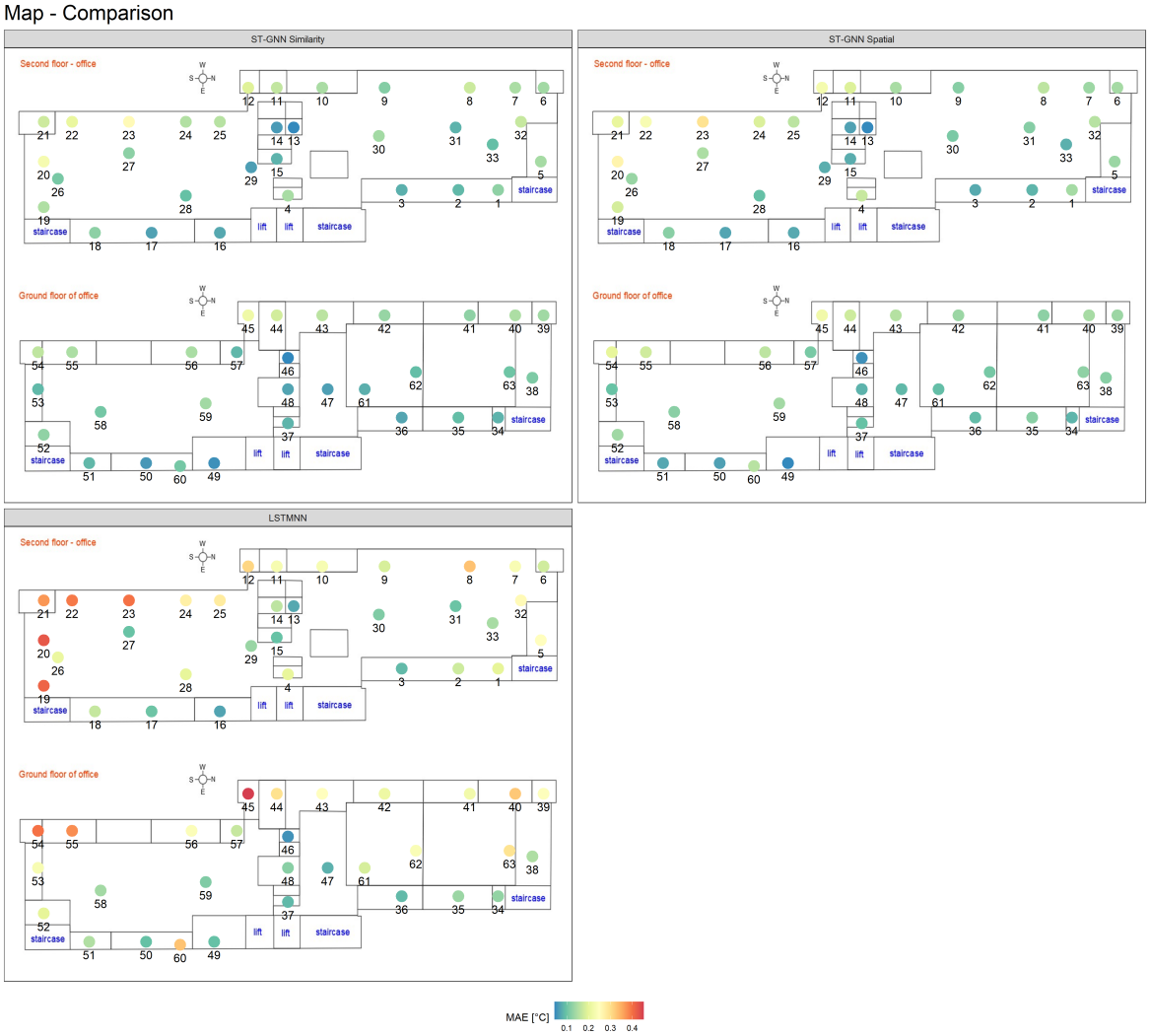


Figura 4.33: Mappa in relazione al MAE di tre modelli di predizione

rispetto a quelli esposti ad est. Tale percentuale minore nel primo cluster dominante dei profili indica che i profili giornalieri della temperatura sono meno ripetitivi. Di conseguenza, entrambi i motivi incidono sulla difficoltà di effettuare una predizione.

Per tutti e tre modelli, le prestazioni di predizione sono molto simili per la maggior parte dei sensori, soprattutto per i sensori che hanno le migliori e le peggiori performance, come i sensori 13, 46, 20 e 23.

Confrontando il MAE, quindi i colori dei sensori, è possibile notare che l'ordine di grandezza del MAE dei tre modelli è lo stesso, ma i colori dei sensori del modello LSTM

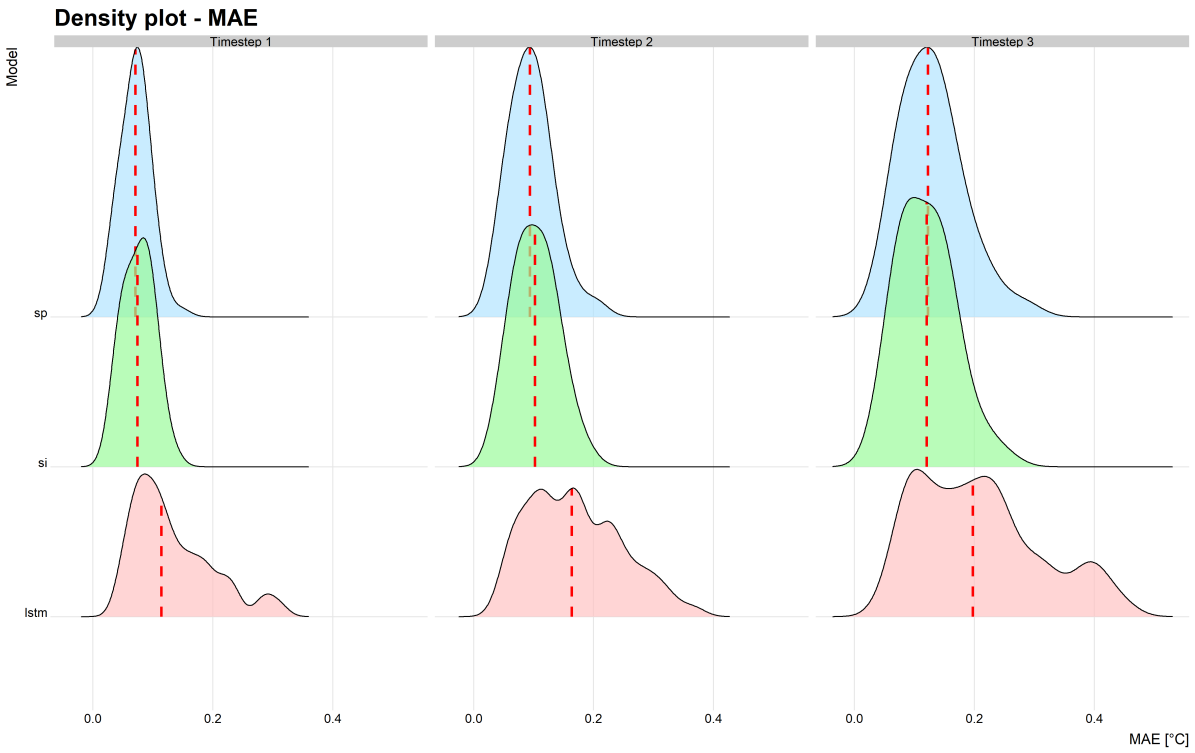


Figura 4.34: Ridgeline plot in relazione al MAE di tre modelli di predizione

sono più inclini al giallo-rosso, ciò è particolarmente evidente nei casi dei sensori 23, 20, 19, 45 e 55. Un'ulteriore osservazione effettuata sul medesimo modello ha constato che ci sono più sensori di colore rosso e meno sensori di colore blu scuro, come meglio mostrato nella figura 4.34.

Il grafico 4.34 riporta i ridgelineplot del Mean Absolut Error (MAE) dei tre modelli di previsione, ovvero il modello spaziale, il modello di somiglianza e il modello LSTM, a uno, due e tre unità di timestep in avanti. I ridgelineplot sono utili per studiare la distribuzione di una variabile numerica tra i diversi gruppi, in questo caso il MAE dei tre modelli e la mediana dei MAE.

Dall'analisi dei grafici si può notare che per tutti i timestep in avanti, la forma della distribuzione del MAE del modello spaziale e del modello di somiglianza è più slanciata e tendente a un valore più basso rispetto al modello LSTM. A differenza, la distribuzione del MAE del modello LSTM è più distribuita e la mediana è più spostata verso un valore superiore.

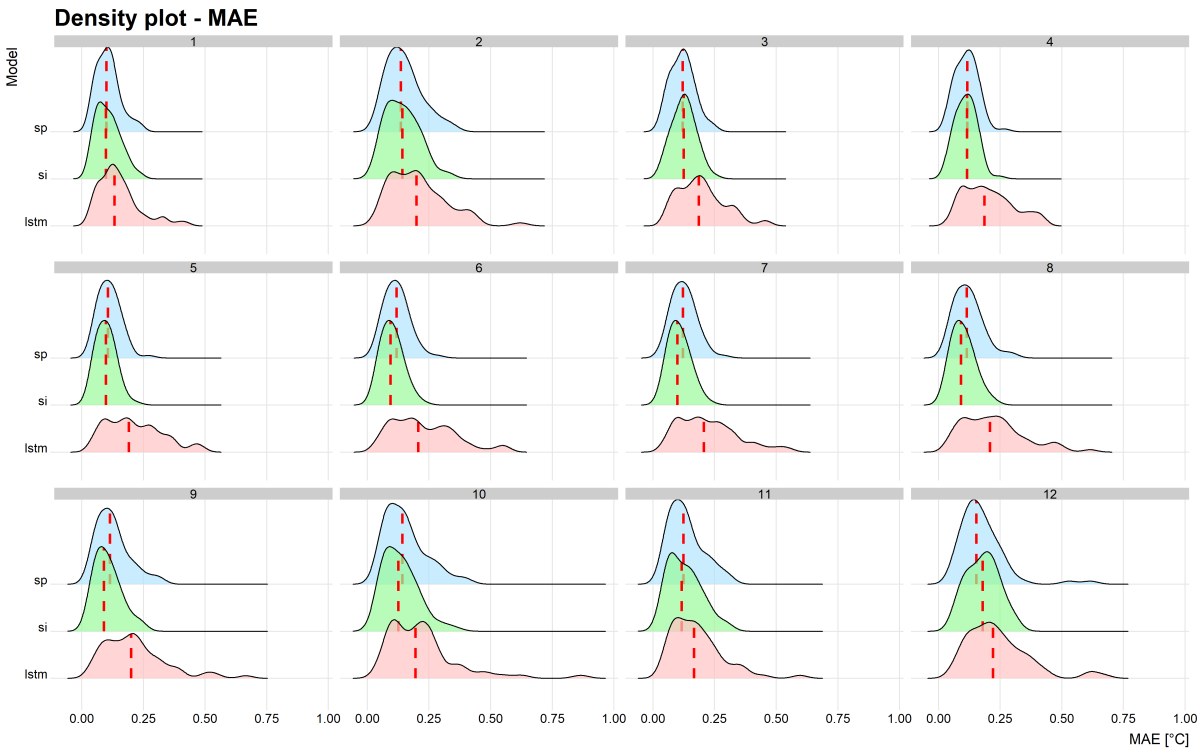


Figura 4.35: Ridgeline plot in relazione al MAE mensile di tre modelli di predizione

Lo stesso è visibile nella tabella 4.6, in cui è possibile leggere le metriche globali dei tre modelli a varie timestep di previsione. Nel caso dei due modelli di GNN i valori si discostano di circa 10%, mentre i valori del modello LSTM sono quasi il doppio dei valori dei due modelli GNN. Pertanto, la prestazione dell'algoritmo LSTM. è paragonabile con quella dei modelli GNN e i MAE, così come la sua mediana, sono leggermente maggiori.

Analizzando i grafici ridgelineplot relativi all'Mean Absolut Error (MAE) mensile a tre timestep in avanti dei tre modelli, presentati nella figura 4.35, si può osservare che le forme delle distribuzioni sono molto simili a quelle descritte in precedenza, così come i valori delle mediane. La differenza tra i modelli è minima, ma si nota che i valori delle mediane del modello GNN per similarità sono molto vicini a quelli del modello GNN spaziale nelle stagioni invernale e primaverile, e leggermente inferiori negli altri periodi.

Visto che è impossibile svolgere un confronto dettagliato tra tutti i sensori, sono stati selezionati i risultati di quattro sensori, rispettivamente il 13, il 46, il 23 e il 35, cioè i quattro sensori con la migliore o la peggiore prestazione in ciascun modello.

MSE [-]			
Timestep	GNN - Similarity	GNN - Spatial	LSTM
1	0.0237	0.0193	0.0364
2	0.0329	0.0258	0.0614
3	0.0336	0.044	0.0959
MAE [°C]			
Timestep	GNN - Similarity	GNN - Spatial	LSTM
1	0.076	0.071	0.132
2	0.103	0.096	0.170
3	0.119	0.129	0.208
MAPE [%]			
Timestep	GNN - Similarity	GNN - Spatial	LSTM
1	0.353	0.329	0.597
2	0.475	0.434	0.768
3	0.540	0.583	0.940

Tabella 4.6: Metriche globali dei tre modelli

Nelle figure 4.36, 4.38, 4.39, 4.37 sono presentati i valori di predizione dei tre modelli e i valori reali focalizzati in 2 giorni, rispettivamente il 12/07/2020 e il 29/12/2020. Si è deciso di scegliere un giorno nella stagione estiva, in quanto l'inverno e l'estate sono due stagioni a temperatura esterna estrema, di conseguenza è importante avere a disposizione una predizione accurata in entrambi i diversi momenti dell'anno. Calcolando il MAE dei risultati rilevati dal sensore 46 per tutti e tre i modelli, emerge che i valori per i due giorni selezionati sono più alti rispetto agli altri.

I valori reali e i valori predetti sono rappresentati con colori differenti. Osservando i grafici si riscontra che i valori di previsione del modello LSTM sono smorzati come se fossero descritti da un'equazione, mentre i modelli di GNN presentano frequenti oscillazioni come se fossero dei valori reali. Di conseguenza, i modelli di GNN riescono ad apprendere meglio rispetto al modello di LSTM, anche se è presente una minima variazione della temperatura.

Nell'insieme, i risultati di predizione del modello GNN spaziale e del modello GNN per similarità sono più accurati e prossimi ai valori reali, per tutti i sensori.

Per quanto riguarda i risultati del sensore 23, il quale costituisce il sensore con i peggiori risultati di predizione secondo le metriche del MAE e del MAPE per i modelli del GNN, i

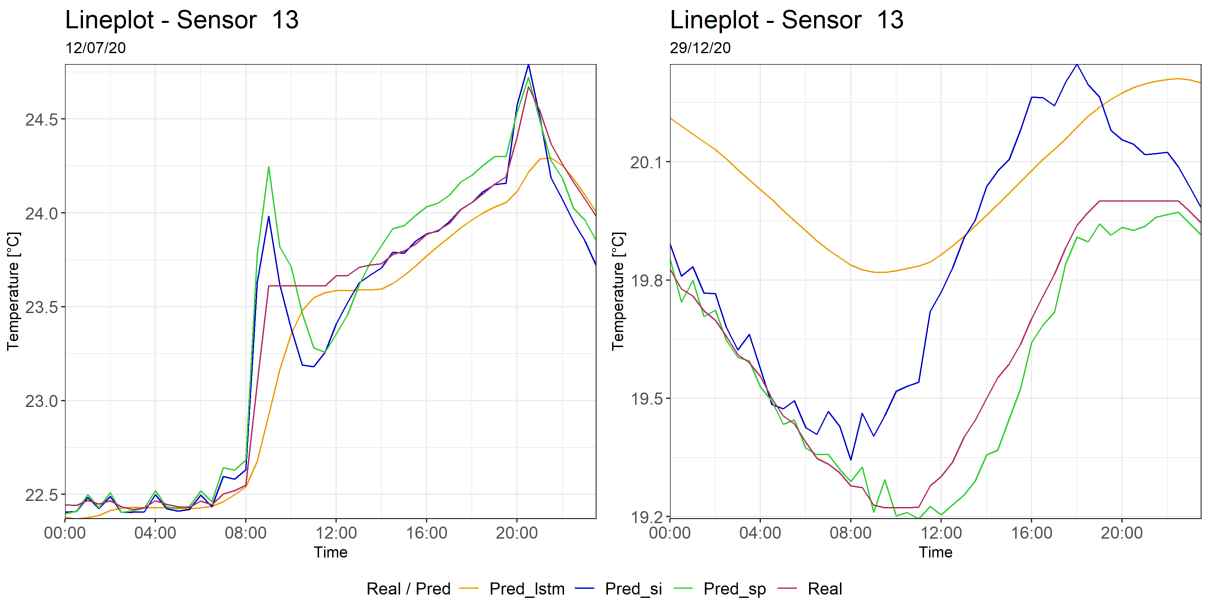


Figura 4.36: Lineplot dei risultati di predizione del sensore 13

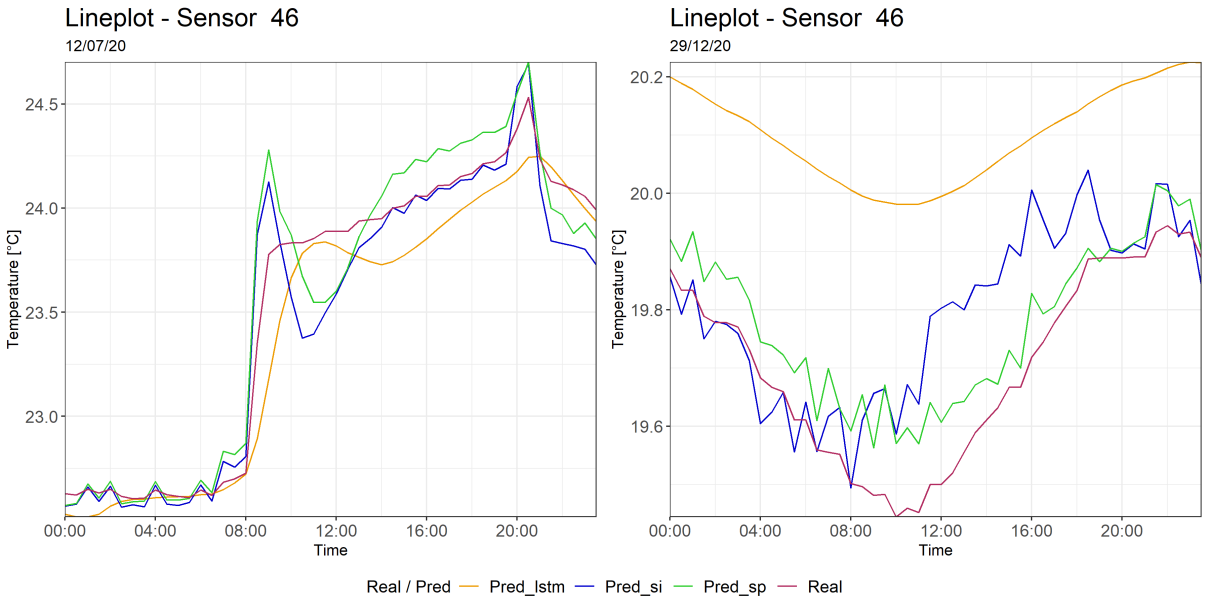


Figura 4.37: Lineplot dei risultati di predizione del sensore 46

valori di predizione dei diversi modelli non si discostano molto dai valori reali.

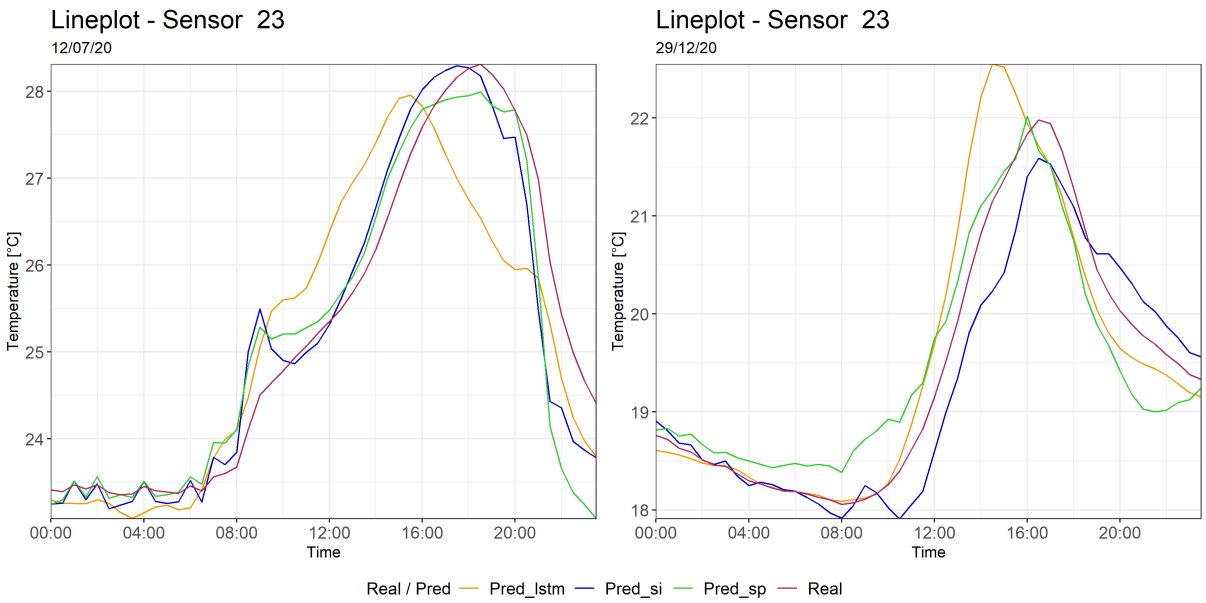


Figura 4.38: Lineplot dei risultati di predizione del sensore 23

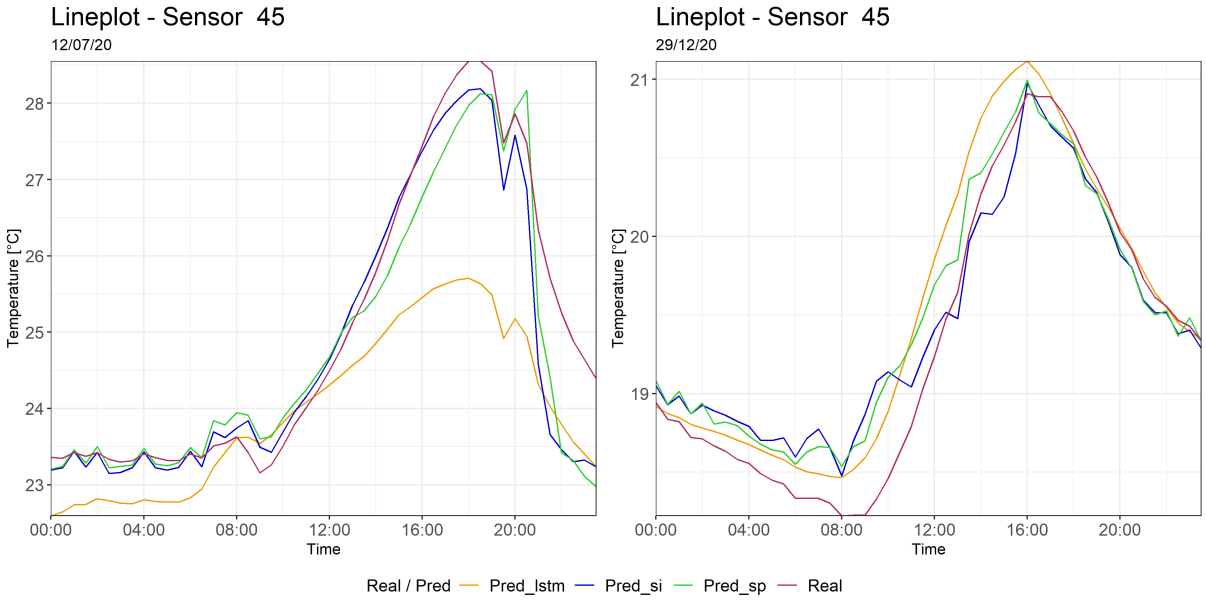


Figura 4.39: Lineplot dei risultati di predizione del sensore 45

Capitolo 5

Conclusione

Questo elaborato intende dimostrare l'importanza della caratteristica temporale e spaziale della temperatura interna ad un edificio del attraverso l'utilizzo del metodo STGNN. Nel caso studio analizzato il dataset è composto da valori della temperatura interna rilevati da 63 sensori disposti su due piani di un edificio ad uso ufficio.

Dalla valutazione del dataset sulla temperatura interna dell'aria sono emerse diverse conclusioni e punti di osservazione.

La prima osservazione derivata dall'analisi del dataset ha mostrato che la variazione giornaliera della temperatura dell'edificio può essere molto diversa a seconda della posizione all'interno dell'edificio, nonostante i sistemi di riscaldamento e di raffreddamento agiscono in modo simile in termini di velocità e di zone di appartenenza. Questa differenza è particolarmente evidente tra i sensori posti ad ovest e quelli posti ad est dell'edificio. Tale differenza di variazione giornaliera della temperatura interna ha permesso di identificare 25 cluster utilizzando il metodo di clustering delle sotto-sequenze con il k-means e con la distanza euclidea. Tra questi, tre cluster risultano particolarmente importanti poiché occupano circa il 75% dei profili giornalieri, mentre i restanti cluster hanno una percentuale inferiore al 5%. Questi tre cluster presentano profili piatti con una piccola curvatura nelle ore pomeridiane e si distinguono tra loro per il range dei valori.

L'albero di classificazione, effettuato tramite il metodo CART sui profili dei tre clusters dominanti, ha confermato l'importanza della caratteristica spaziale dei sensori di misura come precedentemente ipotizzato. Infatti, il root node dell'albero ha diviso i rami in base alla posizione dei sensori, distinguendo quelli della zona interna all'edificio da quelli della

zona esterna. Inoltre, la media giornaliera della temperatura esterna dell'aria, intorno ai 11.5°C, ha dimostrato la sua importanza per entrambi i rami. I profili della temperatura rilevati dai sensori interni all'edificio appartengono principalmente al cluster 1 quando la temperatura media esterna giornaliera supera i 11.5°C e al cluster 2 in caso contrario. A differenza, la maggior parte dei profili della temperatura rilevati dai sensori esterni all'edificio appartengono al cluster 1, mentre solo i profili dei sensori posizionati nella parte nord dell'edificio a condizione di una temperatura media esterna inferiore agli 11.7°C appartengono al cluster 3.

L'analisi predittiva realizzata attraverso il metodo STGNN è articolata in diversi punti. Il modello di previsione ha come input due tipologie di variabili, le variabili esogene (la temperatura dell'aria, l'umidità relativa e la radiazione solare) e la temperatura interna storica del sensore. I modelli di previsione vengono applicati ai dati di ben 63 sensori, ottenendo i risultati di previsione a uno, a due e a tre timestep in avanti per tutti i 63 sensori. Le prestazioni dei modelli di predizione sono state valutate attraverso le metriche per le variabili regressive, cioè il Mean Square Error (MSE), il Mean Absolut Error (MAE) e il Mean Average Percentage Error (MAPE). In seguito, tali metriche vengono applicate ai risultati riferiti ai singoli sensori per i diversi timestep di predizione. Una prima analisi di sensibilità viene svolta secondo diverse configurazioni della matrice di adiacenza e dei grafi. Infatti, attraverso i due approcci di creazione sono stati creati ben 12 configurazioni di grafi con l'approccio di similarità e 2 configurazioni con l'approccio spaziale. Dai risultati ottenuti, è stato evidenziato che non vi è una chiara differenza tra gli intervalli dei valori delle metriche calcolate nelle diverse configurazioni. Inoltre, è stato possibile osservare che le prestazioni dei risultati non sono influenzate dal numero dei punti vicini (k), dal threshold del coefficiente di Pearson, dal numero totale dei collegamenti all'interno di un grafo o dalla mediana del numero dei collegamenti di ogni nodo. Tuttavia, è stato notato che le prestazioni dipendono dalla configurazione del grafo, nonostante una configurazione più complessa del grafo non garantisce prestazioni migliori e viceversa.

Concentrandosi sui risultati di previsione ottenuti dalle due configurazioni dei grafi con le migliori prestazioni per ciascun approccio è emerso che le prestazioni dei sensori rimangono costanti a diverse unità di timestep di previsione indipendentemente dal mese, dal numero dei collegamenti dei sensori e dal timestep di previsione. Le prestazioni dei sensori sono influenzate principalmente dalla differenza giornaliera della temperatura e dalla ripetibilità dei profili.

Dai risultati delle metriche ottenute dai modelli di previsione basati sul GNN e sul LSTM si può affermare che i modelli GNN hanno prestazioni migliori rispetto a quelli basati sul LSTM, sia per quanto riguarda le metriche dei risultati dei singoli sensori che per quelle globali, a qualsiasi intervallo temporale di predizione e a qualsiasi timestep. Inoltre, al contrario del modello LSTM, i modelli GNN tengono in considerazione l'attributo spaziale dei sensori permettendo di allenare un solo modello per l'insieme dei dataset di tutti i sensori, riducendo i costi computazionali.

Al contrario, le metriche dei due modelli di GNN mostrano un'elevata somiglianza sia per quanto riguarda i risultati dei singoli sensori che quelli globali. Così, analizzando la distribuzione del MAE, è possibile notare che per alcuni mesi il modello con l'approccio di similarità produce risultati leggermente migliori. Di conseguenza, considerando anche la difficoltà di creare le matrici di adiacenza, il modello basato sull'approccio di similarità risulta più conveniente.

Bibliografia

[1] Tian Xing, Kailai Sun e Qianchuan Zhao. «MITP-Net: A Deep-Learning Framework for Short-Term Indoor Temperature Predictions in Multi-Zone Buildings». In: *Available at SSRN 4289282* ().

[2] Linh Nguyen, Guoqiang Hu e Costas J Spanos. «Spatio-temporal environmental monitoring for smart buildings». In: *2017 13th IEEE International Conference on Control & Automation (ICCA)*. IEEE. 2017, pp. 277–282.

[3] Brett Pollard, Fabian Held, Lina Engelen, Lauren Powell e Richard de Dear. «Data fusion in buildings: Synthesis of high-resolution IEQ and occupant tracking data». In: *Science of the Total Environment* 776 (2021), p. 146047.

[4] Joseph G Allen e John D Macomber. *Healthy buildings: How indoor spaces drive performance and productivity*. Harvard University Press, 2020.

[5] Ashani Wickramasinghe, Saman Muthukumarana, Dan Loewen e Matt Schaubroeck. «Temperature clusters in commercial buildings using k-means and time series clustering». In: *Energy Informatics* 5.1 (2022), pp. 1–14.

[6] Veerendra Sahu e Bhola Ram Gurjar. «Spatio-temporal variations of indoor air quality in a university library». In: *International Journal of Environmental Health Research* 31.5 (2021), pp. 475–490.

[7] Francisco Troncoso-Pastoriza, Miguel Martinez-Comesana, Ana Ogando-Martinez, Javier Lopez-Gomez, Pablo Eguia-Oller e Lara Febrero-Garrido. «IoT-based platform for automated IEQ spatio-temporal analysis in buildings using machine learning techniques». In: *Automation in Construction* 139 (2022), p. 104261.

[8] Slava Kisilevich, Florian Mansmann, Mirco Nanni e Salvatore Rinzivillo. *Spatio-temporal clustering*. Springer, 2010.

[9] Derya Birant e Alp Kut. «ST-DBSCAN: An algorithm for clustering spatial–temporal data». In: *Data & knowledge engineering* 60.1 (2007), pp. 208–221.

[10] KP Agrawal, Sanjay Garg, Shashikant Sharma e Pinkal Patel. «Development and validation of OPTICS based spatio-temporal clustering technique». In: *Information Sciences* 369 (2016), pp. 388–401.

[11] Hesam Izakian, Witold Pedrycz e Iqbal Jamal. «Clustering spatiotemporal data: An augmented fuzzy c-means». In: *IEEE transactions on fuzzy systems* 21.5 (2012), pp. 855–868.

[12] Marc Hüsch, Bruno U Schyska e Lueder von Bremen. «CorClustST—Correlation-based clustering of big spatio-temporal datasets». In: *Future Generation Computer Systems* 110 (2020), pp. 610–619.

[13] Yang Geng, Wenjie Ji, Yongxin Xie, Borong Lin e Weimin Zhuang. «A sub-sequence clustering method for identifying daily indoor environmental patterns from massive time-series data». In: *Automation in Construction* 139 (2022), p. 104303.

[14] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang e S Yu Philip. «A comprehensive survey on graph neural networks». In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.

[15] Hyeon-Ju Jeon, Min-Woo Choi e O-Joun Lee. «Day-Ahead Hourly Solar Irradiance Forecasting Based on Multi-Attributed Spatio-Temporal Graph Convolutional Network». In: *Sensors* 22.19 (2022), p. 7179.

[16] Seunghye J Wilson. «Data representation for time series data mining: time domain approaches». In: *Wiley Interdisciplinary Reviews: Computational Statistics* 9.1 (2017), e1392.

[17] Chunkai Zhang, Yingyang Chen, Ao Yin, Zhen Qin, Xing Zhang, Keli Zhang e Zoe L Jiang. «An improvement of PAA on trend-based approximation for time series». In: *Algorithms and Architectures for Parallel Processing: 18th International Conference, ICA3PP 2018, Guangzhou, China, November 15-17, 2018, Proceedings, Part II 18*. Springer. 2018, pp. 248–262.

[18] V Barnett e T Lewis. «Outliers in Statistical Data (Probability & Mathematical Statistics).(1994)». In: *V. Barnett and T. Lewis* (1994).

[19] Irfan Pratama, Adhistya Erna Permanasari, Igi Ardiyanto e Rini Indrayani. «A review of missing values handling methods on time-series data». In: *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*. 2016, pp. 1–6. DOI: 10.1109/ICITSI.2016.7858189.

[20] Novta Dany'el Irawan, Wijono Wijono e Onny Setyawati. «Perbaikan Missing value Menggunakan Pendekatan Korelasi Pada Metode K-Nearest Neighbor». In: *Jurnal Infotel* 9.3 (2017), pp. 305–311.

[21] Seyedjamal Zolhavarieh, Saeed Aghabozorgi e Ying Wah Teh. «A review of subsequence time series clustering». In: *The Scientific World Journal* 2014 (2014).

[22] Shraddha Shukla e S Naganna. «A review on K-means data clustering approach». In: *International Journal of Information & Computation Technology* 4.17 (2014), pp. 1847–1860.

[23] Haidi Rao, Xianzhang Shi, Ahoussou Kouassi Rodrigue, Juanjuan Feng, Yingchun Xia, Mohamed Elhoseny, Xiaohui Yuan e Lichuan Gu. «Feature selection based on artificial bee colony and gradient boosting decision tree». In: *Applied Soft Computing* 74 (2019), pp. 634–642.

[24] Monalisa Jena e Satchidananda Dehuri. «DecisionTree for Classification and Regression: A State-of-the Art Review». In: *Informatica* 44.4 (2020).

[25] *Decision Trees: What to Know and How to Construct Them*. URL: https://levelup.gitconnected.com/decision-trees-what-to-know-and-how-to-construct-them-818cf1b47ef3.

[26] Sepp Hochreiter e Jürgen Schmidhuber. «LSTM can solve hard long time lag problems». In: *Advances in neural information processing systems* 9 (1996).

[27] Greg Van Houdt, Carlos Mosquera e Gonzalo Nápoles. «A review on the long short-term memory model». In: *Artificial Intelligence Review* 53.8 (2020), pp. 5929–5955.

[28] *Comprehensive Guide to LSTM RNNs*. URL: https://www.turing.com/kb/comprehensive-guide-to-lstm-rnn.

[29] Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce e Alexander B Wiltschko. «A gentle introduction to graph neural networks». In: *Distill* 6.9 (2021), e33.

[30] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li e Maosong Sun. «Graph neural networks: A review of methods and applications». In: *AI open* 1 (2020), pp. 57–81.

[31] Bing Yu, Haoteng Yin e Zhanxing Zhu. «Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting». In: *arXiv preprint arXiv:1709.04875* (2017).

[32] Adam Auten, Matthew Tomei e Rakesh Kumar. «Hardware acceleration of graph neural networks». In: *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE. 2020, pp. 1–6.

[33] Na Luo, Zhe Wang, David Blum, Christopher Weyandt, Norman Bourassa, Mary Ann Piette e Tianzhen Hong. «A three-year dataset supporting research on building energy management and occupancy analytics». In: *Scientific Data* 9.1 (2022), pp. 1–15.