UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

Analisi di dati meteorologici con metodi di clustering per il posizionamento ottimo di sensori

Relatore: Prof. Andrea Zanella

Laureanda: Maria Teresa Pepaj

ANNO ACCADEMICO 2021 – 2022

Data di laurea 19/07/2022

Sommario

L'analisi e lo studio di eventi meteorologici è fondamentale per prevenire ed evitare disastri ambientali di grandi proporzioni. La creazione e l'utilizzo di dataset climatici di grandi dimensioni comporta alti costi di realizzazione, gestione ed elaborazione dei dati. Diventa quindi importante utilizzare tecniche adeguate che permettano una raccolta dati meno dispersiva e più informativa.

Questa tesi si occuperà di analizzare un set di dati meteorologici attraverso tecniche di *data mining* al fine di studiare il posizionamento ottimo di sensori meteorologici in un territorio. Sono state adottate tecniche di clustering al fine di suddividere le stazioni in sottogruppi e, allo stesso tempo, mantenere l'informazione originaria. Si sono verificati i risultati analizzando la varianza intra cluster, la Silhouette score e calcolando l'errore percentuale sulla ricostruzione tramite interpolazione spaziale delle stazioni. I risultati hanno mostrato che, per lo specifico caso considerato, 3 risulta il numero di cluster ottimale per la classificazione delle stazioni. Utilizzando i rappresentanti delle classi come sottogruppo di stazioni, l'errore risulta del 80% se si ricostruisce solo la variabile precipitazione, invece è del 21% se si considerano tutte le variabili raccolte dal dataset quali: temperatura media, precipitazione, umidità massima, radiazione solare globale e velocità del vento media. Le tecniche utilizzate non hanno identificato un sottogruppo in grado di ricostruire l'informazione originaria in modo soddisfacente, ma dal punto di vista della classificazione si sono formate delle classi abbastanza omogenee.

Indice

1 Introduzione **7**

2 Stato dell'arte **9**

2.1 Posizionamento di sensori 9

2.2 Analisi di dati meteorologici tramite tecniche di clustering 10

3 Dati e metodologia **13**

3.1 Introduzione al caso di studio 13

3.2 Analisi statistica generale 15

3.2.1 Analisi nel tempo 15

3.2.2 Correlazione tra variabili 16

3.2.3 Correlazione spaziale 18

3.3 Tecniche di clustering 19

3.3.1 Tecnica K-means 19

3.3.2 Tecnica clustering gerarchico 20

3.3.3 Clustering e serie temporali 21

3.3.4 Trovare il numero di clusters ottimali 22

3.4 Interpolazione Spaziale 24

4 Risultati sperimentali **26**

4.1 Tecniche scelte per l'analisi dei dati meteorologici 26

4.2 Caso 1: studio della variabile precipitazione su 99 stazioni 28

4.2.1 Verifica informazione del sottogruppo 30

4.3 Caso 2: studio di 5 variabili meteorologiche su 50 stazioni 32

4.3.1 Verifica informazione del sottogruppo 34

4.4 Confronto tra i due casi di studio 35

5 Conclusione **36**

Elenco delle figure

3.1 Mappa delle stazioni considerate. 14

3.2 Evoluzioni temporali delle variabili d'interesse (precipitazione, temperatura media, umidità massima, radiazione solare, velocità media del vento) per la stazione 195. 16

3.3 Stazioni considerate per la correlazione spaziale. 18

3.4 Metodi di ottimizzazione dei clusters. 23

3.5 Triangolazione di Delauney (linee sottili), poligoni di Thiessen (linee spesse). 25

4.1 Metodi di ottimizzazione dei clusters nel caso 1. 29

4.2 Mappa divisione in clusters nel caso 1. 30

4.3 Silhouette stazioni nel caso 1. 32

4.4 Metodi di ottimizzazione dei clusters nel caso 2. 33

4.5 Suddivisione stazioni nel caso 2. 33

4.6 Silhouette clusters nel caso 2. 35

Elenco delle tabelle

3.1 Esempio della struttura del dataset. 15

3.2 Descrizione del dataset. 15

3.3 Correlazioni tra diverse variabili per le stazioni 195, 230, 551. 17

3.4 Correlazione tra variabili dello stesso tipo per diverse stazioni. 19

4.1 Suddivisione stazioni in 3 classi per 99 stazioni nel caso 1. 29

4.2 Suddivisione stazioni in 5 classi per 99 stazioni nel caso 1. 30

4.3 Varianza intra cluster nel caso 1. 31

4.4 Suddivisione stazioni in 3 cluster per 50 stazioni nel caso 2 34

4.5 Varianza intra cluster nel caso 2. 34

Capitolo 1

Introduzione

I dati meteorologici sono la misurazione di parametri fisici atmosferici istantanei e possono essere utilizzati per lo studio di fenomeni meteorologici oppure fornire, se raccolti per decenni, lo stato climatico di un territorio.

L'analisi di dati meteorologici è rilevante per il monitoraggio ambientale, nell'ambito delle previsioni, ma soprattutto al fine di studiare le criticità ambientali di un determinato territorio: ad esempio, lo studio delle precipitazioni può essere fondamentale per evitare alluvioni in zone a rischio. Diventa dunque importante avere a disposizione una adeguata rete di stazioni che catturi le variabili climatiche al fine di avere una visione globale della situazione ambientale del territorio.

Si pone dunque il problema del posizionamento ottimo di sensori che consiste nel trovare la migliore disposizione di sensori in un certo spazio seguendo un determinato criterio di valutazione, come massimizzare l'informazione, oppure minimizzare il numero di sensori richiesti al fine di ridurre i costi della rete mantenendo un buon grado di accuratezza, oppure una combinazione delle due.

Lo scopo di questa tesi è lo studio del posizionamento ottimo di stazioni meteorologiche tramite tecniche di clustering per analisi e classificazione di dati, applicate su un dataset che descrive una serie di variabili meteorologiche. Quest'analisi fornisce una risposta alla questione in maniera sperimentale tramite l'analisi del caso di studio presentato.

La tesi sarà strutturata nella seguente maniera: nel secondo capitolo si propone una panoramica bibliografica su tecniche di clustering e posizionamento ottimo di sensori facendo un confronto con altre tecniche più avanzate di machine learning come SVM. Nel capitolo 3

introdurremo il caso preso in questione, relativo ai dati di una parte della regione Veneto presentando un'analisi statistica generale per studiare la struttura del dataset per poi continuare con l'analisi delle tecniche utilizzate. Nel quarto capitolo si parlerà in maniera dettagliata del problema dell'identificazione di un sottogruppo di stazioni prendendo in considerazione due casi: nel primo si prenderà in esame solo la variabile precipitazione nel secondo invece tutte le variabili raccolte. Infine, nel quinto capitolo si trarranno delle conclusioni sui risultati ottenuti e si discuteranno possibili sviluppi futuri.

Capitolo 2

Stato dell'arte

Lo studio di grandi dataset, specialmente di dati meteorologici, è un tema molto presente in letteratura. Nel seguito si descriveranno casi rilevanti allo scopo di questa tesi, riportando le principali metodologie e risultati fondamentali correlati al caso di studio.
Si riporteranno casi in letteratura riguardo il posizionamento ottimo di sensori e l'utilizzo di tecniche di clustering.

2.1 Posizionamento di sensori

In questo paragrafo si analizzeranno diverse tecniche di posizionamento ottimo presenti in letteratura basate sull'ottimizzazione dell'informazione.
Nell'articolo [1] si affronta il problema di posizionare n sensori per la misurazione di flusso e velocità lungo un canale fluviale in k possibili posizioni. In questo caso vengono calcolate le risposte dei sensori in una posizione di monitoraggio, in seguito si utilizza la matrice di Gram al fine di trovar quale delle $\binom{n}{k}$ combinazioni risulta essere quella più indipendente, cioè che contiene più informazione. In [2] si utilizzano tecniche di Supporting Vector Machine (SVM) al fine di studiare una rete di sensori idrometeorologici trovando quali siti sono veramente necessari per l'analisi dello stato idrologico del territorio. Sapendo a priori quali sensori avranno il ruolo di *supporting vector*, la tecnica SVM riesce a classificare i sensori della rete. Si dimostra poi che prendendo come sottoinsieme proprio il gruppo di support vectors l'informazione raccolta dalla rete rimane intatta con un minimo errore.
La tecnica utilizzata contrariamente alla tecnica di clustering, che mira a raggruppare sensori

con misurazioni simili, è un tipo di classificazione a priori, che richiede quindi di sapere quali saranno i *supporting vector*.

In [3] viene descritto il metodo Monte Carlo che assicura di massimizzare l'informazione della rete di sensori. Per applicare il metodo descritto bisogna conoscere le dimensioni del territorio su cui si vogliono posizionare i sensori e dividere il territorio in una griglia i cui nodi indicano possibili posizionamenti per i sensori. In seguito si fa uno studio sul variogramma, per stimare la dipendenza spaziale tra le osservazioni, e si utilizza il metodo per l'interpolazione spaziale Kriging per individuare quali sono le posizioni ottimali.

L'articolo segue con un esperimento per individuare il posizionamento di sensori per il monitoraggio di mercurio nella zona di Oak Ridge (Tennessee). Conoscendo il valore medio e la varianza della concentrazione di mercurio, si assume la distribuzione con una Gaussiana e viene utilizzato il metodo Monte Carlo, variando sia il numero N di sensori necessari sia il numero M di ripetizioni del metodo. Si conclude notando una relazione inversamente proporzionale tra M e la varianza minima.

Nell'articolo [4] si cerca di trovare il numero ottimale di sensori da mantenere al fine di non perdere informazione tramite tecniche di clustering. Le variabili considerate nello studio sono la temperatura, la luminosità e l'umidità all'interno di un palazzo. Il set iniziale prevede 31 sensori distribuiti uniformemente nell'area considerata, in seguito vengono applicate diverse tecniche di clustering trovando un sottogruppo ottimale di 6 sensori. La valutazione della bontà di tale sottogruppo viene fatto calcolando l'informazione persa dopo la ricostruzione della rete iniziale. In conclusione la perdita di informazione è molto esigua dunque il clustering è un metodo valido per trovare la posizione ottimale di sensori su un grande gruppo.

2.2 Analisi di dati meteorologici tramite tecniche di clustering

Come è stato detto precedentemente, studiare dati meteorologici è essenziale per determinare le condizioni di un territorio. Tuttavia, spesso, i set di dati meteorologici sono molto grandi e rendono necessarie nuove tecniche di analisi.

In [5] viene considerato un dataset di 6 variabili temporali appartenenti ad una singola sta-

zione meteorologica nella regione della Striscia di Gaza in un periodo di 9 anni.
I dati sono serie temporali. Come primo passo fondamentale viene fatta una pre-elaborazione sul dataset rendendo i dati omogenei ricostruendo quelli mancanti tramite l'interpolazione lineare. Una delle tecniche che viene presentata è proprio il clustering, in questo caso al fine di trovare delle caratteristiche prevalenti nel processo climatico. Viene utilizzato l'algoritmo di clustering K-means raggruppando i dati in 4 classi per descrivere il comportamento delle stagioni nel territorio considerato, vengono in seguito comparate altre tecniche di classificazione con metodi di predizione quali Least Median Squares Linear Regression e Neural Networks, che però non saranno descritti in questa sede in quanto al di fuori dello scopo di questo lavoro.
In [5] quindi, il clustering viene utilizzato per estrarre informazioni nel dominio del tempo. In [6], invece, viene utilizzata la tecnica Density Based Spatial Clustering of Applications with Noise (DBSCAN) al fine di suddividere la Turchia in regioni basate su delle caratteristiche climatiche. Il dataset preso in considerazione in questo caso comprende le variabili di temperatura minima e massima correlate alle coordinate nello spazio per 248 stazioni nel periodo dal 1930-1996.
Nel caso di grandi dataset come questo è stato fatto un ridimensionamento per evitare un tempo computazionale dell'algoritmo troppo elevato: sono state prese dunque solo le osservazioni del mese di luglio del 1996 per tutte le stazioni. Come risultato finale vengono identificate 4 regioni mettendo in evidenza le stazioni equivalenti al nucleo del cluster, trovando dunque delle zone con caratteristiche climatiche simili.
Nell'articolo [7] si prende in considerazione un dataset di 188 stazioni distribuite per tutta la Turchia, per un periodo di circa 31 anni dal 1967 al 1998. La variabile misurata è la precipitazione totale annua e si considera anche la variazione totale annua. Lo scopo è dividere la regione in zone con simili caratteristiche climatiche attraverso tecniche di clustering, confrontando due metodi diversi.
Si mettono a confronto la tecnica K-means e quella Fuzzy C-means (FCM), facendo un test dell'omogeneità della cluster e un "discordancy test" per studiare la struttura dei dati. Il metodo K-means è più rigido poiché non permette che una stazione venga inclusa in più clusters al contrario del FCM. Il numero di clusters ottenuti è 6 per FCM e 7 per l'algoritmo K-means ottenendo risultati sull'omogeneità migliori per il metodo FCM.

Si conclude quindi che il metodo Fuzzy C-means risulta essere più efficiente nel trovare delle zone di omogeneità per la variabile precipitazione.

In [8] viene studiato il territorio cinese, prendendo in considerazione dati giornalieri della variabile precipitazione nell'anno 2019 da 2472 siti meteorologici. Vengono messe a confronto 3 tecniche di clustering diverse: K-means, DBSCAN e clustering gerarchico per trovare delle regioni con simili distribuzioni di pioggia. Per trovare il numero di classi ottimali viene utilizzato il metodo di Elbow, che verrà descritto nel paragrafo 3.3.4, dando come risultato 8 classi. La classificazione viene fatta sia per i singoli mesi dell'anno considerato, sia per l'intero anno, mostrando le mappe del territorio diviso nei vari cluster durante il 2019. Si conclude notando che i metodi K-means e clustering gerarchico mostrano risultati molto simili, anche al variare dell'intervallo di tempo considerato, mentre DBSCAN si dimostra non appropriato rispetto ai dati considerati.

Capitolo 3

Dati e metodologia

In questo capitolo si introdurrà il caso di studio preso in considerazione, descrivendo il dataset utilizzato tramite un'analisi statistica generale mettendo a confronto grafici e correlazioni tra le variabili meteorologiche considerate. Infine si descriverà brevemente la metodologia che si è utilizzata per la risoluzione dei sotto problemi di identificazioni di un sottogruppo descritti nel capitolo seguente.

3.1 Introduzione al caso di studio

Per studiare il problema del posizionamento ottimo e le sotto domande conseguenti poste in questa tesi è stato costruito un dataset.

I dati sono stati raccolti dal sito dell' Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto (Arpav) nella sezione dati meteorologici ultimi 60 giorni [9]. È stata presa in considerazione un'area approssimativamente di dimensioni 10124 km^2 di forma approssimativamente quadrata, come mostrato in Fig. 3.1. Le coordinate spaziali (x, y, z) delle stazioni sono espresse in metri, e si è assicurati di includere diverse aree paesaggistiche (laguna, montagna, pianura).

Le misure ambientali riguardano precipitazioni, temperatura, umidità, radiazione solare e vento. Tuttavia, le stazioni possono monitorare solamente alcuni di questi parametri: in totale infatti si contano 99 stazioni per la precipitazione e un sottoinsieme di 50 stazioni con dati per tutte 5 le variabili.

La serie di dati temporali è stata raccolta con frequenza giornaliera, quindi il numero di

campioni sarà 60, dalla data 18/02/2022 al 18/04/2022. Come si può vedere nell'esempio in Tab. 3.1, ogni serie di dati è collegata al codice della stazione, alla sua posizione nello spazio tramite le coordinate (x, y, z) e alla data dell'acquisizione del campione. I dati sono stati precedentemente controllati in modo da recuperare eventuali osservazioni mancanti, pari allo 0.09% sul totale di dati, attraverso interpolazione lineare, in modo che tutte le stazioni prese in considerazione fossero omogenee e avessero ugual numero di osservazioni.

Per descrivere il dataset di studio è stata fatta un'analisi sui valori medi, deviazione standard, massimo, minimo e varianza tra stazioni per le 5 variabili considerate. Per la precipitazione si considereranno tutte le 99 stazioni, mentre per le altre 4 variabili ci si limita alle sole 50 stazioni che raccolgono l'insieme completo delle osservazioni. La deviazione standard non coinciderà con la radice quadra della varianza poiché il calcolo è stato fatto sulle n stazioni considerate per ognuno dei 60 giorni disponibili facendone in seguito la media. Dalla Tab. 3.2 si può notare come le variabili vento e precipitazione abbiano varianza minore mentre l'umidità mostra una varianza molto elevata. Dunque rispetto alla precipitazione in, generale, le altre variabili sono meno omogenee.

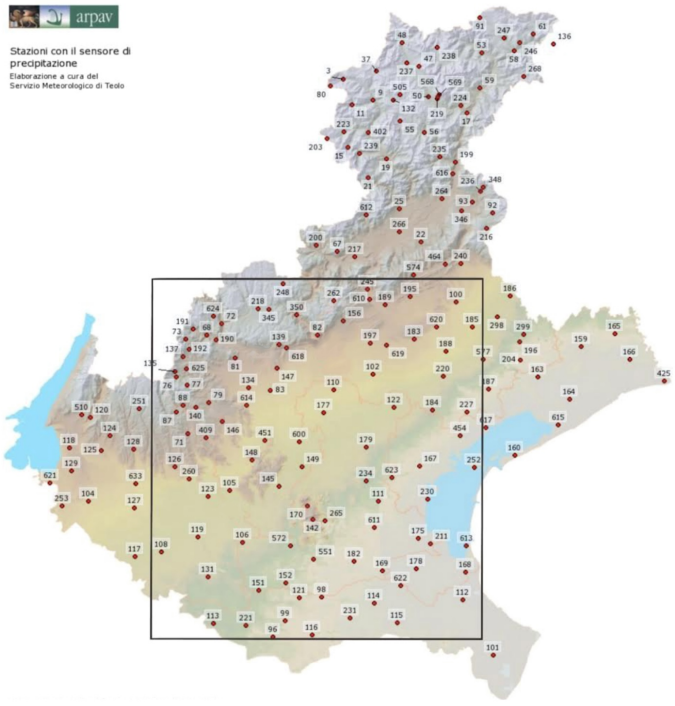


Figura 3.1: Mappa delle stazioni considerate.

stazione	data	coordinate	precipitazione	temperatura	umidità	radiazione	vento
68	18/02/22	x,y,z	...	...	...	...	...
...	...	...	...	...	...	...	...

Tabella 3.1: Esempio della struttura del dataset.

variabile	media	σ	σ^2	min	max
precipitazione (mm)	0.6	0.57	3.64	0	54.6
temperatura (C°)	8.31	2.14	4.67	-9.3	17.2
umidità (%)	88	8.4	95	29	100
radiazione globale (MJ/m^2)	15.75	1.61	3.7	1.3	27.6
vento (m/s)	1.5	0.66	0.53	0.1	8.8

Tabella 3.2: Descrizione del dataset.

3.2 Analisi statistica generale

Prima di utilizzare il dataset del caso di studio per risolvere i problemi principali, si farà un'analisi statistica e delle considerazioni sui dati a disposizione. Poiché il territorio considerato è molto vasto e comprende diversi tipi di paesaggi saranno considerate 3 stazioni: una in una zona di montagna, una in pianura e una in zona lagunare, scelte tra le stazioni che raccolgono tutte le variabili. Nel paragrafo 3.2.1 si farà un'analisi nel tempo delle variabili rappresentando i grafici dei singoli campioni e facendo un primo confronto. Nel paragrafo 3.2.2 si farà un'analisi della correlazione tra variabili per definire il grado di dipendenza tra di esse, specialmente con la precipitazione. Infine nel paragrafo 3.2.3 si studierà la correlazione spaziale tra stazioni vicine.

3.2.1 Analisi nel tempo

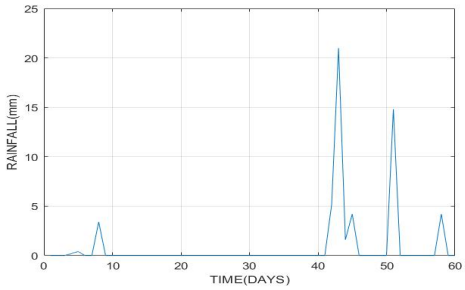
Per fare una prima analisi dei dati si confronteranno i grafici nel tempo delle serie di variabili, in cui le ascisse rappresentano la dimensione temporale.

Analizzare i grafici è un primo approccio utile anche per l'identificazione di outliers, ossia di valori anomali; in questo caso non si notano valori fuori norma.

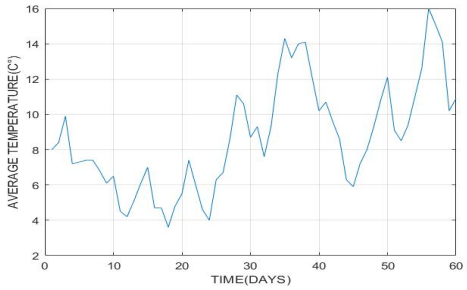
I dati sono relativi alla stazione 195, situata in zona di montagna, e osservando i grafici presentati in Fig. 3.2, sono già visibili le correlazioni tra le variabili.

Infatti, si nota come tra i giorni 42 e 52 in cui la pioggia ha i picchi maggiori, la radiazione solare è diminuita, mentre l'umidità in quei giorni oscilla intorno a valori vicino al 100 per tut-

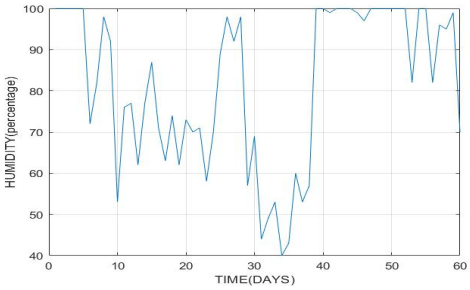
to il periodo di pioggia. Per quanto riguarda la temperatura questa diminuisce leggermente durante il periodo di pioggia, mentre per il vento, la correlazione non risulta evidente ad una prima ispezione visiva. Valuteremo i coefficienti di correlazione esatti nel prossimo paragrafo.



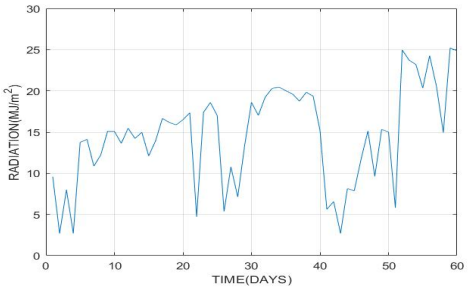
(a) Grafico precipitazione.



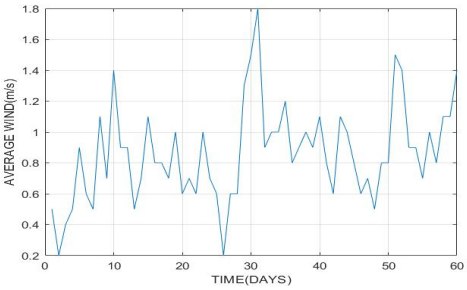
(b) Grafico temperatura media.



(c) Grafico umidità massima.



(d) Grafico radiazione solare.



(e) Grafico vento medio.

Figura 3.2: Evoluzioni temporali delle variabili d'interesse (precipitazione, temperatura media, umidità massima, radiazione solare, velocità media del vento) per la stazione 195.

3.2.2 Correlazione tra variabili

Al fine di descrivere al meglio il processo delle precipitazioni analizzeremo quali variabili sono maggiormente correlate ad essa. Per analizzare la correlazione tra le diverse variabili si deve considerare quale coefficiente possa adattarsi meglio ad un processo metereologico che non è lineare. Tra i coefficienti di correlazione si è deciso di impiegare il coefficiente

di Spearman in opposizione al coefficiente lineare di Pearson. Il coefficiente di Spearman ρ_s permette di calcolare la correlazione tra due variabili stimando quanto accuratamente è possibile descriverne la relazione con una funzione monotona. Il coefficiente può variare tra -1 e 1, maggiore è il modulo maggiore sarà la relazione tra le due; il segno rappresenta se la dipendenza è crescente (segno positivo) o decrescente (segno negativo). La formula per calcolare il coefficiente di Spearman è la seguente:

$$\rho_s = \frac{cov(R(x_i), R(x_j))}{\sigma_{R(x_i)}\sigma_{R(x_j)}},$$

dove $R(x_i)$ e $R(x_j)$ rappresentano le due serie, *cov* la covarianza e σ la deviazione standard. Come già accennato si vedranno le differenze tra 3 stazioni diverse che misurano tutte le variabili prese in considerazione, la stazione 195 in zona di montagna posizionata ad un'altitudine di 169 metri, la 551 in zona di pianura ad un'altitudine di 8 metri e la 230 in prossimità della laguna a 0 metri sopra il livello del mare. Dalla Tab. 3.3 che riporta i coefficienti di correlazioni tra precipitazione e altre variabili si nota immediatamente un'alta correlazione con la radiazione e l'umidità per tutte le stazioni, mentre per la temperatura la correlazione non è così marcata. Non si riscontrano grandi differenze tra le stazioni che esibiscono tutte la medesima struttura nei dati: una forte correlazione negativa tra precipitazioni e radiazione ρ_{p-r} e una forte correlazione positiva tra precipitazioni e umidità ρ_{p-u}, la correlazione precipitazione-temperatura ρ_{p-t} e quella precipitazione-vento ρ_{p-v} hanno invece valori prossimi a zero. Le correlazioni vento e temperatura non hanno un comportamento uniforme per tutte le stazioni. Infatti, prendendo in considerazione il vento si ha una correlazione che cambia di molto in base alla stazione senza seguire una determinata legge.

Quindi possiamo concludere che se i dati meteorologici venissero usati al fine di predire l'evento "precipitazione" si potrebbe alleggerire il set di dati eliminando quelle variabili che non possiedono una correlazione evidente, in questo caso vento e temperatura.

Coefficienti di correlazione Spearman				
stazione	ρ_{p-t}	ρ_{p-u}	ρ_{p-r}	ρ_{p-v}
195	-0.042	0.4041	-0.45	0.169
230	-0.006	0.4316	-0.374	-0.092
551	-0.046	0.399	-0.2105	0.03

Tabella 3.3: Correlazioni tra diverse variabili per le stazioni 195, 230, 551.

3.2.3 Correlazione spaziale

Nel capitolo seguente parleremo di clustering e interpolazione lineare considerando la posizione geografica delle stazioni. Dunque introduciamo indagini sulla correlazione spaziale al fine di far notare che le variabili meteorologiche prese in considerazione sono dipendenti dallo spazio, dimostrando che stazioni tra loro vicine sono fortemente correlate. Nel caso in questione si sono prese le stazioni 551, 245 e 182, con le stazioni 182 e 551 limitrofe e la 245 più lontana come si vede in Fig. 3.3, ed è stato calcolato il coefficiente di Spearman per tutte le coppie di variabili prendendo la stazione 182 come riferimento. Dalla Tab. 3.4 notiamo una forte correlazione tra i valori delle stazioni più vicine e valori leggermente minori per la stazione più lontana. Specialmente per quanto riguarda la variabile umidità il valore è molto minore, per il vento la correlazione ha addirittura segno opposto, mentre per le precipitazioni i due valori sono molto vicini anche se leggermente maggiore per la stazione più vicina. Il comportamento è coerente con quanto atteso e questo può indicare che il sistema considerato ha dati abbastanza simili tra loro: pertanto è prevedibile che il un numero di clusters risulti relativamente basso con cluster formati da stazioni vicine nello spazio.

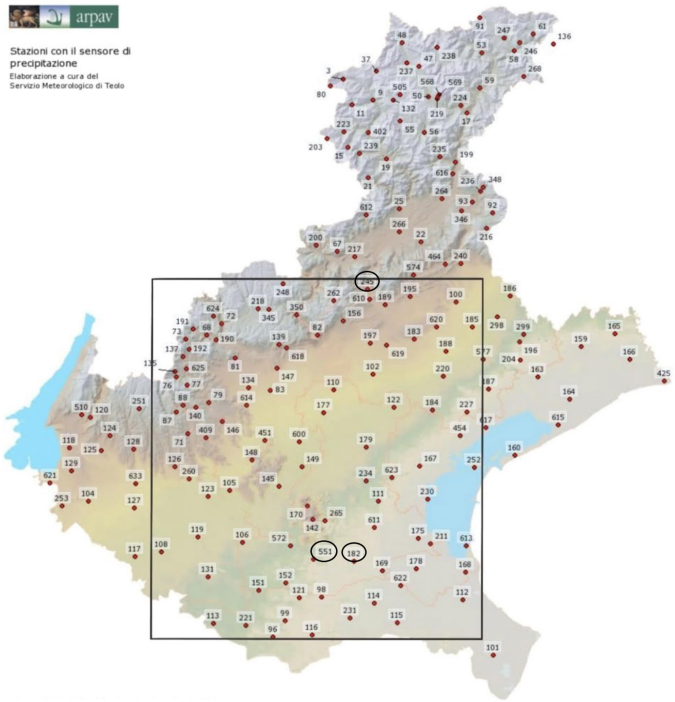


Figura 3.3: Stazioni considerate per la correlazione spaziale.

stazione 182					
stazione	ρ_p	ρ_t	ρ_u	ρ_r	ρ_v
551	0.63	0.98	0.98	0.97	0.02
245	0.6	0.92	0.57	0.79	-0.29

Tabella 3.4: Correlazione tra variabili dello stesso tipo per diverse stazioni.

3.3 Tecniche di clustering

Le tecniche di clustering vengono usate in vari ambiti per l'analisi di dati con diversi scopi tra cui: la classificazione non supervisionata per raggruppare dati simili, oppure per ridurre la dimensione di un dataset. Il concetto fondamentale del clustering è quello del raggruppamento in classi composte da oggetti il più simili tra loro avendo un fattore di indipendenza il più alto possibile con le altre classi.
Nei prossimi paragrafi sono presentati diversi algoritmi di clustering evidenziando le caratteristiche principali.

3.3.1 Tecnica K-means

Una tra le tecniche di clustering principali e più semplici è il metodo K-means. Conoscendo a priori il numero k di clusters l'obiettivo dell'algoritmo K-means è quello di riuscire a trovare la partizione ottimale delle osservazioni in quelle k classi.
Dopo aver deciso il numero k di clusters in cui si vuole dividere il set di dati, verranno scelti in maniera casuale i primi k elementi rappresentanti di ciascuna classe, che verranno chiamati centroidi. In seguito viene assegnata ogni osservazione ad una dell k classi utilizzando un determinato criterio. Solitamente si utilizza la formula della distanza euclidea. Prendendo in considerazione due vettori $X = (x_1, \ldots, x_n), Y = (y_1, \ldots, y_n)$ di lunghezza n, la distanza euclidea è calcolata con:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Ciascun elemento del dataset sarà assegnato dunque alla classe rappresentata dal centroide da cui ha minor distanza.
Successivamente si ricalcoleranno i centroidi, calcolando il baricentro per ogni classe, se questi cambiano si ritornerà al secondo passo e si riassegneranno i punti mancanti reiterando l'algoritmo fino al momento in cui i rappresentanti delle k classi rimangono inalterati, condizione

di temine dell'algoritmo.

Una variante molto simile al K-means è l'algoritmo K-medoids che si distingue semplicemente perché prende come centroide del cluster uno degli elementi del dataset, il più vicino al baricentro della classe, invece del baricentro stesso. Per un algoritmo K-means la complessità si dimostra essere:

$$T = O(nkl),$$

con n la dimensione del vettore, k il numero di classi e l le volte in cui l'algoritmo itera.

3.3.2 Tecnica clustering gerarchico

Il metodo di clustering gerarchico ha come principio quello di formare un albero di clusters, che può essere ottenuto con un approccio top down (per divisione) o bottom up (per agglomerazione).

Nell'algoritmo con approccio agglomerativo tutti i dati separati formeranno un singolo cluster e rappresenteranno le foglie, in seguito si raggrupperanno i clusters che sono più simili tra loro formando un nodo. Si procede con l'algoritmo finché non si completa l'albero cioè quando si arriva alla radice e tutte le osservazione sono racchiuse in un unico cluster.

Ci sono vari criteri per portare al completamento dell'albero basati su diverse formule per calcolare la distanza tra clusters. Il metodo più semplice (e più datato) è il *single linkage*, cioè si considera la distanza minima tra due membri delle classi prese in considerazione. Altri metodi sono il *complete linkage* che consiste nel calcolo della distanza massima tra due membri della classe, oppure si può calcolare la distanza dai centroidi delle classi considerate, infine il *metodo di Ward* che consiste nel calcolare quali classi hanno minima varianza se unite.

Come nel lavoro [10] per definire quindi il criterio di unificazione si può scrivere la formula generale, cioè *The Lance-Williams dissimilarity update*, formula che definisce, prese due osservazioni i e j unite in un nuovo cluster, la differenza tra il nuovo cluster e le altre classi k, e viene così definita:

$$d(i \cup j, k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)|,$$

nel quale i valori $\alpha_i, \alpha_j, \beta, \gamma$ rappresentano dei parametri che controllano il criterio che si vuole utilizzare per l'agglomerazione. Per il single linkage vale: $\alpha_i = 0.5, \alpha_j = 0.5, \beta = 0, \gamma = -0.5$.
Il metodo divisivo, molto meno utilizzato, si basa invece su un approccio top down cioè forma l'albero partendo da un unico cluster per poi andare a dividere, secondo i criteri di dissimilarità appena definiti, in più clusters fino al completamento dell'albero con le foglie cioè quando tutte le cluster coincidono con un unico elemento. Tra i criteri più usati per la divisione c'è quello di Ward.
La complessità dell'algoritmo è maggiore rispetto ad un algoritmo K-means. Infatti l'algoritmo di cluster gerarchico, sia implementato con approccio per agglomerazione che per divisione ha complessità:

$$T = O(n^3),$$

con n elementi di input.

3.3.3 Clustering e serie temporali

Per serie temporale si intende un vettore che rappresenta un numero di osservazioni temporali dal tempo t che rappresenta l'istante del primo campione, al tempo $t+n$ ultima osservazione considerata e si rappresenta come:

$$X = (x_t, x_{t+1}, x_{t+2} \ldots, x_{t+n}).$$

Si possono impiegare tecniche di clustering anche per analizzare dataset di serie nel tempo, che possono rappresentare la variazione di una grandezza in un certo lasso temporale.
L'articolo [11] divide la categorizzazione di serie temporali in base a dei criteri quali:

1. la loro similarità nel tempo, cioè se il loro grafico temporale ha un andamento simile,
2. hanno un forma simile cioè delle determinate caratteristiche in comune (hidden pattern),
3. mostrano tipi di variazioni simili nel tempo.

Tecniche di clustering possono essere utilizzate con i dati originali o con i dati elaborati secondo un determinato criterio, per esempio la normalizzazione. Per le serie temporali, in

particolare, è importante una fase di riduzione preliminare in modo da diminuire la dimensione dei vettori, importante per l'efficienza degli algoritmi di clustering. Alcuni metodi sono la *dinamic time warping*, come viene usato nell'articolo [11], oppure la *DFT* (Discrete Fourier Transformation).

Come anticipato in precedenza un criterio molto comune e valido anche per le serie temporali è la distanza euclidea; potrebbero essere utili anche altri criteri quali: il coefficiente di correlazione di Pearson sempre come viene citato nell'articolo [11] e il coefficiente di Spearman.

Scelto quindi il criterio per identificare la similarità tra le serie temporali si possono applicare le stesse tecniche descritte prima, come K-means o clustering gerarchico.

3.3.4 Trovare il numero di clusters ottimali

Le tecniche di clustering, al contrario di altri metodi di classificazione, non sono supervisionate ma hanno bisogno di fissare preventivamente il numero di gruppi in cui dividere i dati: nasce, quindi, il problema dell'ottimizzazione del numero di clusters. Per quanto riguarda il cluster gerarchico una soluzione può essere tagliare il dendogramma (una rappresentazione in forma di grafo-albero della suddivisione in cluster) rispetto ad un certo criterio di similarità come si fa vedere nella Fig. 3.4a tratta dal lavoro [12], oppure, un criterio valido in generale, si decide a priori il limite di errore che si è disposti ad accettare, trovando dunque la condizione di fine dell'algoritmo ottenendo le k classi richieste. L'errore può essere definito come SSE (Sum of Squares Error) dalla seguente formula:

$$SSE = \sum_{k} \sum_{i=1}^{n} ||x_i - c_k||^2,$$

con x_i si intende il dato i-esimo della serie di dati, con c_k si intende il centroide della k-esima classe e con $||.||$ si intende la norma del vettore. Un altro metodo basato sul calcolo del SSE è il metodo "Elbow" un semplice algoritmo che consiste nel plottare rispetto al numero di clusters l'errore SSE. Quando il valore avrà un decremento significativo formando un angolo, cioè quando la curva raggiunge il punto di "elbow" il valore del numero di cluster ad esso associato rappresenterà il numero di k ottimale per suddividere i dati.

Come si nota dalla Fig. 3.4b, tratta dal lavoro [13] che studia la suddivisione in cluster attraverso l'utilizzo di K-means e il metodo elbow, si nota che al cluster 4 si avrà il punto

elbow della curva, dunque 4 rappresenta il numero di clusters necessario.

Esiste anche una semplice regola empirica citata dall'articolo [14] che trova un numero indicativo di clusters in funzione del numero di dati considerati senza alcuna informazione aggiuntiva, la formula è la seguente:

$$k = \sqrt{\frac{N}{2}},$$

definendo N come il numero di dati utilizzati.

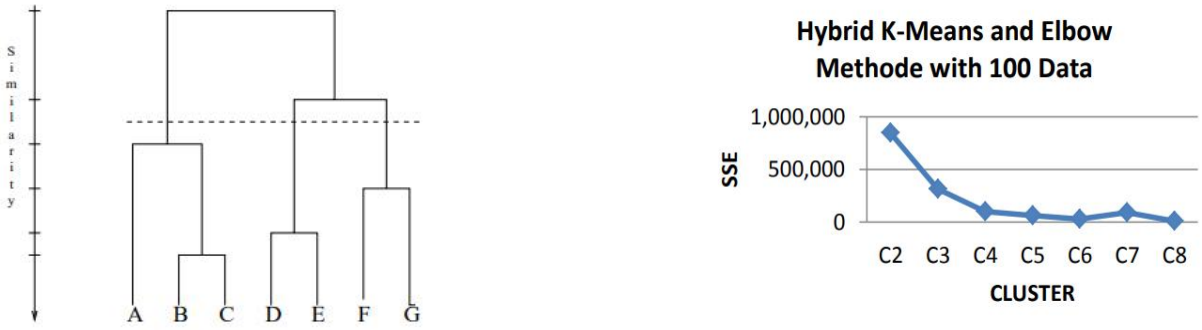
Matlab, in [15], inoltre implementa un metodo di valutazione di clustering basato sul criterio Calinski-Harabasz, basato sulla definizione stessa di suddivisione in clusters ottima, cioè si deve avere una grande varianza tra gruppi diversi mantenendo una varianza minima tra membri dello stesso cluster. Questa metrica è chiamata anche Variance Ratio Criterion (VRC), e si calcola nel seguente modo:

$$VRC = \frac{SS_b}{SSE} \times \frac{(N-k)}{(k-1)},$$

definendo N come il numero di dati, k il numero di clusters, SSE rappresenta la varianza tra clusters definita già prima come errore, e SS_b l'errore interno al cluster definite dalla seguente formula:

$$SS_b = \sum_{i=0}^{k} n_i ||m_i - m||,$$

dove k è il numero di cluster, n_i è il numero di osservazioni nel cluster i, m_i è il centroide del cluster i, m è la media complessiva dei dati del campione e $||m_i - m||$ è la norma tra i due vettori. Il numero di cluster ottimo sarà quello associato all'indice di Calinski-Harabasz più alto.



(a) Taglio del dendogramma.

(b) Metodo di Elbow.

Figura 3.4: Metodi di ottimizzazione dei clusters.

3.4 Interpolazione Spaziale

L'interpolazione spaziale viene usata spesso nella meteorologia: infatti non sempre sono presenti stazioni in tutti i luoghi di una regione e per predire quindi tutti i punti incogniti si utilizzano i valori delle stazioni conosciute.
Un metodo utilizzato per predire valori della precipitazione, nell'articolo [16], per l'interpolazione spaziale è quella delle Distanze Inverse Pesate (IDW) in cui vengono considerate le osservazioni delle stazioni conosciute come una base di vettori $V_1, V_2, V_3, ..., V_n$ e Y come il vettore da predire. La loro relazione si scrive nel seguente modo:

$$Y = \alpha_1 V_1 + \alpha_2 V_2 + \alpha_3 V_3 + ... + \alpha_n V_n,$$

con α_n i coefficienti della media pesata relativi alla distanza, perciò più vicino è la stazione n-esima al valore da predire più il suo coefficiente sarà elevato.
I coefficienti vengono calcolati nel seguente modo:

$$\alpha_i = \frac{\frac{1}{d(j,i)^\gamma}}{\sum_{i=1}^{k} \frac{1}{d(j,i)^\gamma}},$$

nella quale i rappresenta l'indice dell'i-esima stazione dei centroidi, j l'indice della stazione che si vuole ricostruire, $d(j,i)$ la distanza nello spazio tra le stazioni i e j e γ rappresenta un parametro di controllo generalmente posto a due.
Questo metodo può avere risultati limitati se la distribuzione delle stazioni non è uniforme: infatti, nel caso una zona sia densa di stazioni, i punti predetti in quella zona saranno molto più accurati rispetto ad altre in cui sono presenti meno stazioni.
Un altro metodo descritto nell'articolo [17], sempre per la misurazione di precipitazioni, è quello dell'interpolazione triangolare: cioè si divide il territorio in opportune zone triangolari, un metodo per fare ciò può essere l'algoritmo di Delauney.
Altrimenti si può fare una divisione in zone poligonali con il metodo di Thiessen (Fig. 3.5), per poi misurare i valori sconosciuti nelle singole zone con le misure delle stazioni conosciute che rientrano in quella zona.

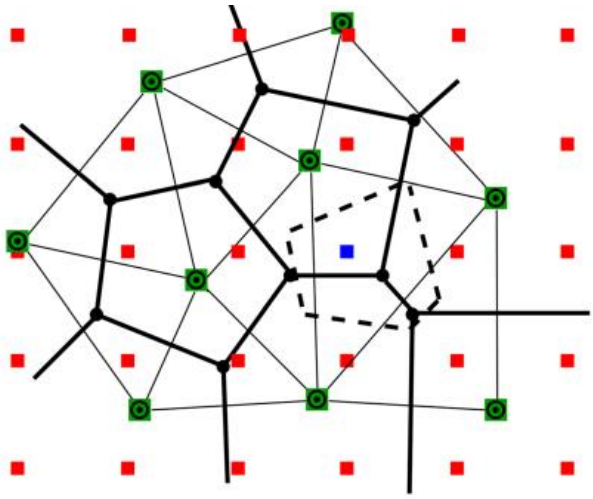


Figura 3.5: Triangolazione di Delauney (linee sottili), poligoni di Thiessen (linee spesse).

Capitolo 4

Risultati sperimentali

Nel seguente capitolo si riporteranno i dettagli sulle metodologie utilizzate nel caso di studio, riferendosi a quanto descritto nel capitolo precedente, e si illustrerà come verranno applicate al fine dello studio del posizionamento ottimo.
Si riporteranno i risultati al quesito del posizionamento ottimo nel caso di studio, trovando un sottogruppo tra le stazioni meteorologiche del territorio considerato, al fine di riuscire a descrivere la situazione ambientale con una rete di stazioni con un minor numero di nodi ma più informativa possibile.
Nel primo paragrafo dunque si specificherà quale tecniche sono state utilizzate, nel secondo si studierà la variabile precipitazione misurata da 99 stazioni nel territorio del Veneto, mentre nel terzo paragrafo si analizzerà un caso con 50 stazioni descritte da 5 variabili meteorologiche ,nel quarto paragrafo, infine, si farà un confronto dei risultati.

4.1 Tecniche scelte per l'analisi dei dati meteorologici

Come si è discusso prima tra gli algoritmi proposti si è deciso di utilizzare la tecnica del K-medoids, una variante del K-means. La scelta è dovuta al fatto che il K-means assegna come centroide un punto calcolato come la media dei punti della classe che non appartiene necessariamente al set di dati, invece K-medoids assegna semplicemente come rappresentante della classe l'osservazione più vicina al centroide del cluster.
L'algoritmo è stato implementato con la funzione *kmedoids*, presente nel "Statistical and Machine learning Tool Box" di Matlab, che richiede come input la matrice X rappresentante

i vettori di dati, k il numero di clusters in cui si vogliono raggruppare i dati e un input che rappresenti il criterio di confronto tra dati: nel seguente caso sarà utilizzato il coefficiente di Spearman con cui è stato fatto già un'analisi di correlazione nella Sez. 3.2.2, ma di default questo parametro della funzione è la distanza euclidea.

Come primo passo si è controllato che i dati fossero completi, ovvero che non mancassero osservazioni a qualche istante t.

Come è stato detto nei capitoli precedenti, le serie di dati spesso devono essere ridimensionate; in questo caso il dataset era già abbastanza esiguo, quindi si è deciso di usare i dati puri senza alcuna modifica, si è lavorato infatti con dati in uno span temporale di soli 60 giorni.

Per definire il numero di clusters si è utilizzato il metodo di "Elbow", descritto prima, e si è fatto un paragone con la regola empirica già menzionata al fine di controllare la precisione di un metodo non strettamente scientifico.

Il metodo K-means dà come output un vettore colonna $N \times 1$ contenente gli indici del cluster assegnato per ogni dato e un vettore contenente i centroidi per classe. Si assegnerà dunque ogni stazione alla propria classe e si visualizzerà il cluster sulla mappa.

Infine, per dimostrare se la scelta dei k centroidi è sufficiente per descrivere il territorio, entro certi limiti di errore, si userà l'interpolazione spaziale IDW controllando che la media pesata in base alla distanza riesca a definire i valori delle altre stazione con la seguente formula:

$$x_j = \sum_{i=0}^{k} a_i x_i,$$

con x_j la serie della stazione da voler ricostruire, k il numero di clusters, a_i i pesi relativi alla stazione rappresentante del i-esimo cluster rispetto alla stazione da ricostruire e x_i la serie della variabile del rappresentante del i-esimo cluster.

I coefficienti α_i vengono calcolati utilizzando la formula descritta al paragrafo 3.4, scegliendo due come parametro di controllo γ. Il coefficiente relativo alla stazione più vicina nello spazio avrà un valore più alto e la somma di tutti i coefficienti distanza dalla base di centroidi sarà 1.

Si definisce quindi il modello lineare che ha come variabili indipendenti le stazioni rappresentanti del cluster, che formeranno una base dello spazio delle stazioni nel territorio; la stazione i-esima che si vuole ricostruire come variabile dipendente e come parametri dell'equazione

lineare i coefficienti spaziali definiti in precedenza.
Tra i possibili coefficienti per la valutazione dell'errore relativo si è utilizzato l'errore relativo sul vettore ϵ calcolato nel seguente modo :

$$\epsilon = \frac{||X_o - X_p||}{||X_o||},$$

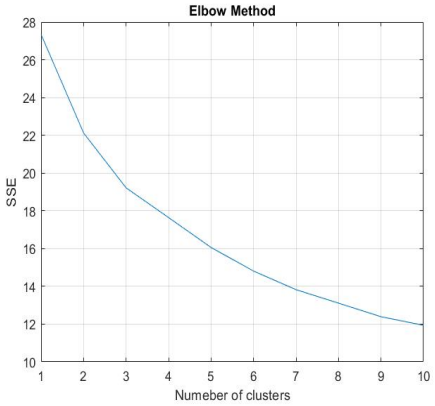
con X_o si intende il vettore osservato, X_p il vettore predetto tramite i centroidi e $||.||$ la norma del vettore.
Inoltre si è deciso di utilizzare la funzione *silhouette* di Matlab per valutare la classificazione delle singole stazioni in un cluster. Il metodo di Silhouette calcola un coefficiente per ogni stazione che varia da -1 a 1 secondo un certo criterio a scelta, (in questo caso si è usato la correlazione) se il coefficiente è vicino ad 1 significa che il dato è ben abbinato al suo cluster altrimenti, se è basso o negativo, significa che la classificazione non è accurata.
Per avere dei valori numerici sull'omogeneità di un cluster è stata fatta anche un'analisi della varianza all'interno di ciascuna classe.

4.2 Caso 1: studio della variabile precipitazione su 99 stazioni

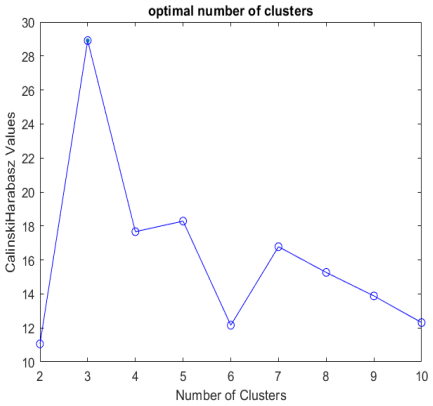
Per il primo caso di studio l'unica variabile presa in considerazione sarà la precipitazione e le stazioni incluse sono tutte quelle presenti nella zona evidenziata dalla Fig. 3.1. Per utilizzare il metodo di cluster, definito nel capitolo precedente, è stata creata una matrice in cui ogni riga corrisponde alla serie temporale della precipitazione di una stazione. Si ottiene quindi una matrice 99×60 cioè il numero di stazioni moltiplicata per i campioni rilevati.
Come si nota dalla Fig. 4.1a, il metodo di Elbow ha evidenziato come numero di clusters ottimale 3 classi. Si è adottato anche il metodo di Calinski-Harabasz, Fig. 4.1b, che ha confermato l'ipotesi del metodo grafico di Elbow.
Si sono in seguito suddivise le stazioni tramite K-medoids. In Tab. 4.1 è riportata la suddivisione in clusters ottenuta. Nella mappa in Fig. 4.2a si possono visualizzare le diverse zone che rappresentano i clusters. Come si può notare, i gruppi sono abbastanza separati tra loro nello spazio e si possono identificare delle aree precise anche se il cluster 1 e quello

2 si intersecano in alcune zone. Il numero di cluster risulta essere abbastanza basso: infatti, per quanto riguarda le precipitazioni, la correlazione spaziale era abbastanza simile anche tra stazioni lontane evidenziando una bassa varianza nel territorio.

Si evidenzia maggiormente il cluster 3, il più compatto, nella zona di montagna mentre il cluster 1 copre una zona più ampia la parte Sud-Est del Veneto, il cluster 2 si concentra nella parte Nord e centrale con qualche stazione nella zona del primo cluster. Vengono cerchiati invece i centroidi risultanti le stazioni 151 per il primo cluster, 102 per il cluster 2 e per il terzo la stazione 137. Notando che 3 è un numero abbastanza ridotto per 99 serie di dati, si è provato a dividere il dataset anche in 5 classi per verificare se anche dall'esperienza si deduce che 3 è il numero di clusters ottimale. Si riportano i risultati in Fig. 4.2b e Tab. 4.2, mentre il confronto tra i due casi è presentato nel paragrafo seguente.



(a) Metodo di Elbow.



(b) Metodo di Calinski-Harabas.

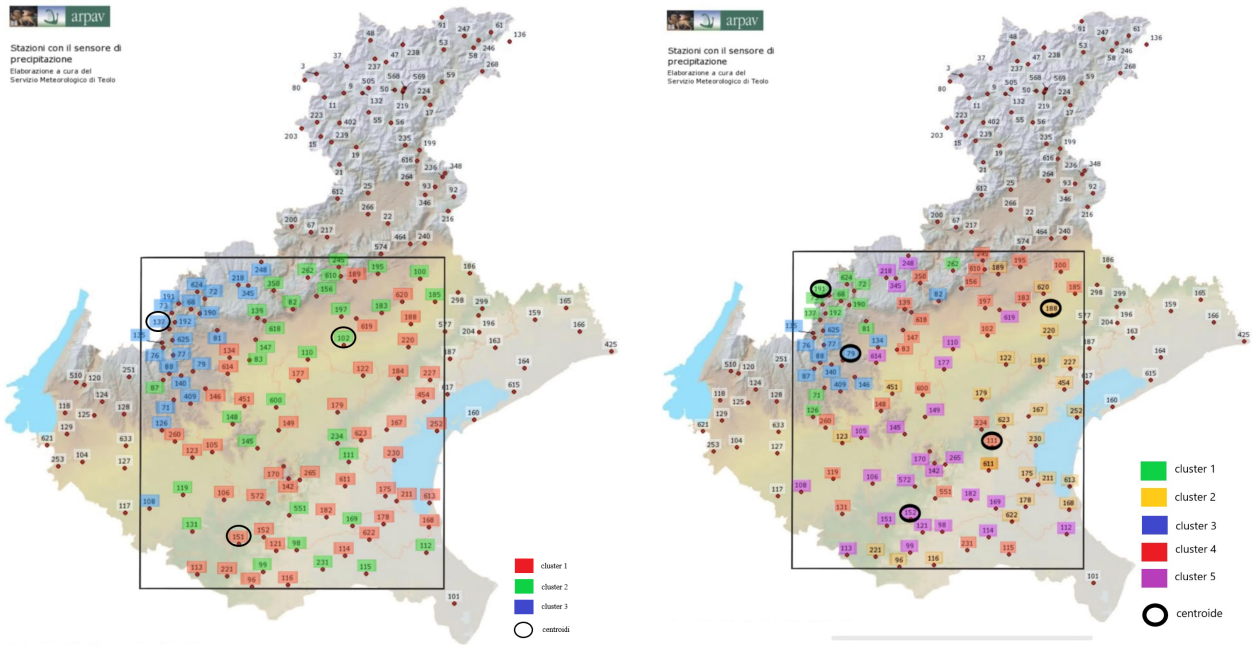
Figura 4.1: Metodi di ottimizzazione dei clusters nel caso 1.

cluster 1	cluster 2	cluster 3
96 105 106 113 114 116 121	82 83 87 98 99 100	68 71 72 73
122 123 142 146 149 151 152	102 110 111 112 115 119	76 77 79 81
167 168 170 175 177 178 179	131 134 139 145 147 148	88 108 126 135
182 184 188 189 211 220 221	156 169 183 185 195 197	137 140 190 191
227 230 252 265 451 454 572	231 234 245 260 262 350	192 218 248 345
611 613 614 619 620 622 623	551 600 610 618	409 624 625

Tabella 4.1: Suddivisione stazioni in 3 classi per 99 stazioni nel caso 1.

cluster 1	cluster 2	cluster 3	cluster 4	cluster5
68 71 72 73	96 116 122 123	76 77 79 82	83 100 102 111	98 99 105 106
81 126 137 190	167 168 175 178	87 88 134 135	115 119 131 139	108 110 112 113
191 192 262 624	179 184 188 189	140 146 409 625	147 148 156 183	114 121 142 145
	211 220 221 227		185 195 197 231	149 151 152 169
	230 252 451 454		234 245 260 350	170 177 182 218
	611 613 620 622		551 600 610 618	248 265 345 572
	623			614 619

Tabella 4.2: Suddivisione stazioni in 5 classi per 99 stazioni nel caso 1.



(a) Mappa dei 3 clusters per 99 stazioni.

(b) Mappa dei 5 clusters per 99 stazioni.

Figura 4.2: Mappa divisione in clusters nel caso 1.

4.2.1 Verifica informazione del sottogruppo

Il sottogruppo di stazioni è stato identificato prendendo i centroidi definiti dall'algoritmo di cluster. Essi sono una scelta ottimale per la definizione stessa dell'algoritmo: infatti sono i centroidi del cluster che garantiscono maggior somiglianza intra-cluster e indipendenza tra cluster diverse.

Lo scopo è controllare se, sapendo a priori solo i valori delle precipitazioni, si riescano a ricostruire le serie delle altre stazioni, stimandone la precisione. Le stazioni quindi che si sono utilizzate come base per l'interpolazione spaziale, nel caso della divisione in 3 cluster, sono la stazione 102, 137 e 151 rappresentanti delle 3 classi.

Per calcolare una predizione delle altre 96 stazioni dunque si è calcolato una matrice di coefficienti $\alpha_{i,102}, \alpha_{i,137}, \alpha_{i,151}$, che rappresentano i coefficienti di una media pesata in base alla distanza definiti precedentemente.

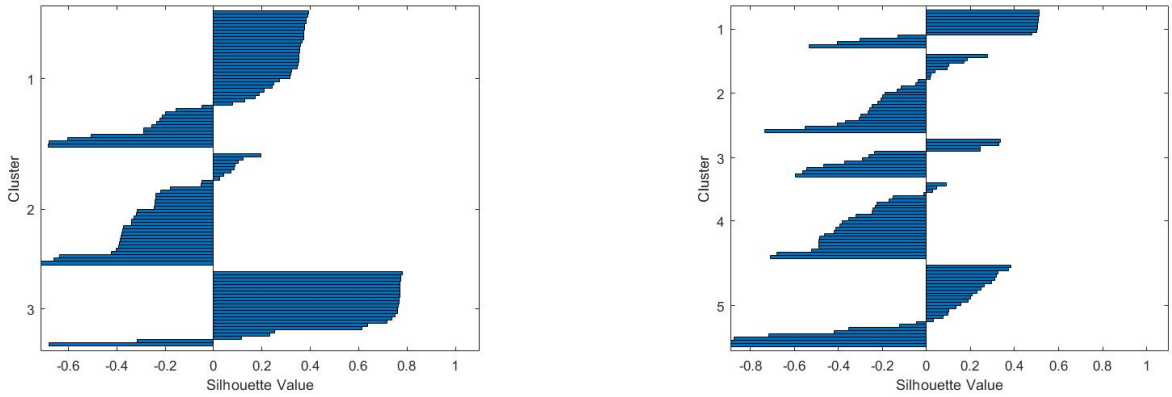
Per verificare l'accuratezza del modello descritto si è calcolato il coefficiente di errore relativo dei vettori, citato precedentemente, per ogni stazione e poi si è fatta la media, dando come risultato un errore medio del 80% , il modello quindi risulta essere poco accurato in generale. Inoltre alcune stazioni hanno un errore che arriva oltre il 200% mentre molte si mantengono comunque sotto il 50% di errore. Il metodo utilizzato non ha prodotto un modello accurato come potevamo aspettarci dallo studio della correlazione spaziale, che mostrava valori relativamente omogenei nello spazio per la variabile precipitazione, forse dovuto alle tecniche di interpolazione spaziale usate.

Per quanto riguarda i centroidi nel caso della divisione in 5 clusters, sono le stazioni 79, 111, 152, 188, 191. Anche in questo caso è stata fatta l'interpolazione IDW e si è ottenuto un errore medio del 82%, dunque l'aumento di classi non ha portato ad un miglioramento nella precisione della ricostruzione. Il metodo di Silhouette ha evidenziato per il caso delle 3 clusters in Fig. 4.3a valori alti per il cluster 3, che era infatti il più compatto, e valori negativi o comunque bassi per il cluster 1 e 2. Il cluster 3 si dimostra dunque essere abbastanza accurato mentre il cluster 1 e 2 non sono una buona configurazione. Per il secondo caso invece in Fig. 4.3b notiamo che eccetto il primo e il quinto cluster, gli altri hanno valori di Silhouette negativi rappresentando una divisione in cluster inadeguata. Dal confronto delle silhouette risulta dunque più efficace il raggruppamento in 3 classi.

Si è calcolata anche la varianza intra cluster al fine di avere un'ulteriore metrica di valutazione. Nella Tab. 4.3 si mettono a confronto i valori delle varianze interne alle classi: si nota che nel caso con 3 cluster i valori sono minori perciò risulta che i cluster sono più omogenei, confermando di nuovo che 3 è il numero ottimale in questo caso.

	cluster 1 σ^2	cluster 2 σ^2	cluster 3 σ^2	cluster 4 σ^2	cluster 5 σ^2
Caso 3 cluster	1.73	1.9	4.06	-	-
Caso 5 cluster	4.29	2.16	3.1	1.94	1.37

Tabella 4.3: Varianza intra cluster nel caso 1.



(a) Silhouette 3 clusters.

(b) Silhouette 5 clusters.

Figura 4.3: Silhouette stazioni nel caso 1.

4.3 Caso 2: studio di 5 variabili meteorologiche su 50 stazioni

Nel secondo caso di studio si sono considerate le 5 variabili definite nel paragrafo 3.1. Le stazioni per questo secondo caso sono un sottogruppo dell'insieme totale considerato precedentemente ovvero solo quelle che misurano tutte le variabili.

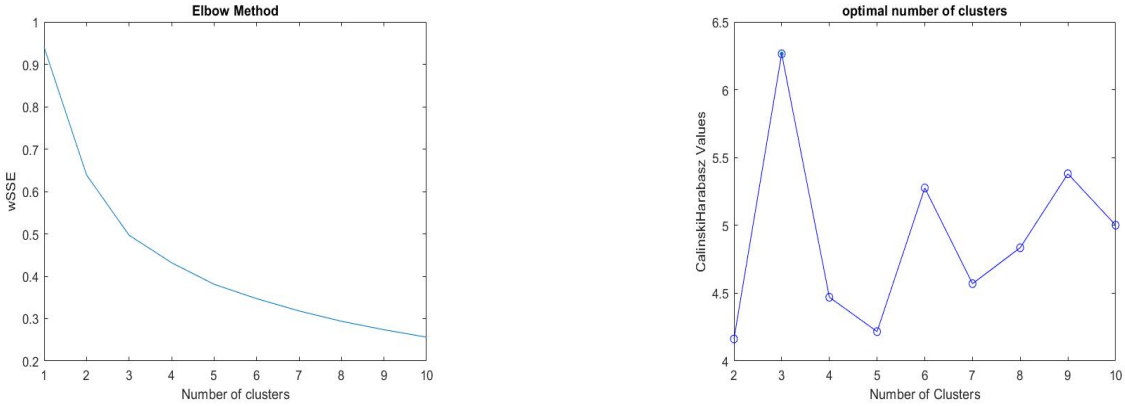
Il metodo cluster K-medoids, precedentemente definito, avrà come input una matrice con 50 righe, rappresentanti il numero di stazioni e 300 colonne rappresentanti le 5 serie temporali con 60 campioni ciascuna. Vengono riportati nella Fig.4.4 le immagini relative alla scelta del numero di classi ottimale. Anche per questo caso i metodi sono concordanti ed evidenziano 3 clusters, differentemente dai 5 clusters suggeriti dalla regola empirica.

La suddivisione delle stazioni in cluster è riportata nella Tab. 4.4.

Nella Fig. 4.5 si sono evidenziate i cluster con diverso colore. Si possono notare delle zone nettamente separate sempre perché si ottiene una rilevante correlazione spaziale tra stazioni vicine quando sono considerate tutte le variabili.

Le zone sono divise in una regione Sud centrale rappresentata dal cluster 1, una zona Nord-Est corrispondente al cluster 2 e una zona in montagna per il cluster 3.

Sono evidenziati i centroidi che sono in ordine per cluster rispettivamente la stazione 152, 102 e 218.



(a) Metodo di Elbow. (b) Metodo di Calinski-Harabas.

Figura 4.4: Metodi di ottimizzazione dei clusters nel caso 2.

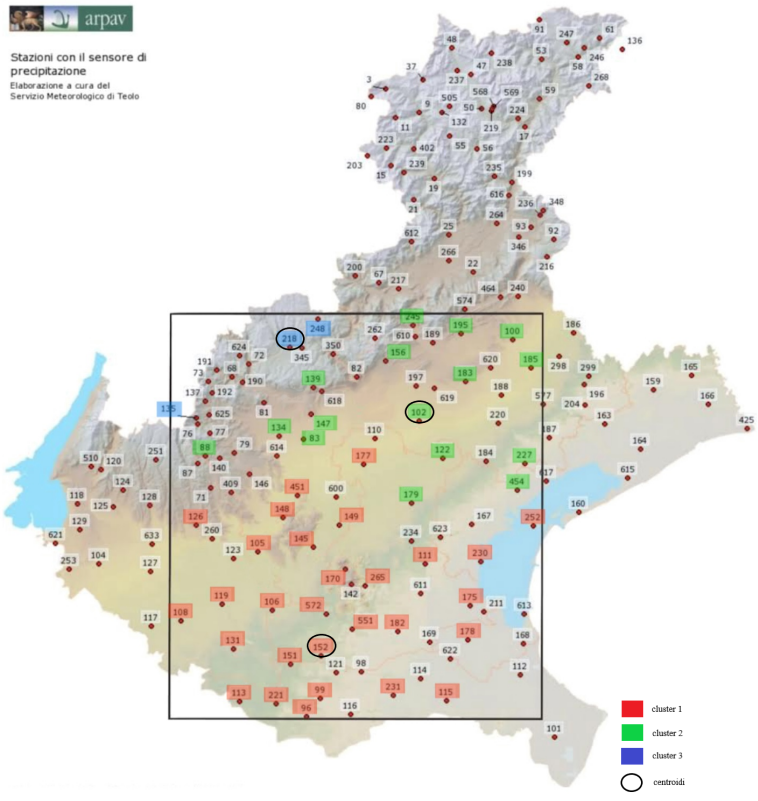


Figura 4.5: Suddivisione stazioni nel caso 2.

cluster 1	cluster 2	cluster 3
96 99 105 106 108 111 112	83 88 100 102	135 218 248
113 115 119 126 131 145 148	122 134 139 147	
149 151 152 170 175 177 178	156 179 183 185	
182 221 230 231 252 265 451	195 227 245 454	
551 572 600		

Tabella 4.4: Suddivisione stazioni in 3 cluster per 50 stazioni nel caso 2

4.3.1 Verifica informazione del sottogruppo

In questo caso le stazioni utilizzate sono stati i rappresentanti dei clusters, corrispondenti ai sensori 102, 152 e 218. I coefficienti sono stati calcolati nel modo precedentemente definito ma prendendo le coordinate dei nuovi centroidi.

Per il seguente caso di studio si sono ricostruiti i valori di tutte le variabili meteorologiche prese in considerazione e sempre con il coefficiente di errore tra vettori si è identificato l'errore sulla ricostruzione di tutte e 5 le variabili, dando come risultato un errore medio del 21%, relativamente buono.

L'analisi di clustering ha quindi in maniera soddisfacente trovato un sottogruppo ottimo che permette di stimare in maniera sufficientemente accurata anche i dati di altre stazioni.

Il metodo di Silhouette ha evidenziato in figura 4.6 valori per la maggior parte positivi per i cluster 1 e 3, mentre il cluster 2 ha una buona parte di valori negativi, rendendo la seconda classe meno accurata.

Per quanto riguarda la varianza riportata nella tabella 4.5, i valori del terzo cluster sono i più alti, eccetto che per la variabile vento, mentre le altre due classi sono abbastanza omogenee con una varianza abbastanza bassa. La varianza della variabile umidità è la più alta, infatti tra tutte le variabili era la più correlata allo spazio.

	cluster 1 σ^2	cluster 2 σ^2	cluster 3 σ^2
Precipitazione (mm)	0.88	1.33	2.64
Temperatura (C°)	0.5	1.57	8.4
Umidità (%)	65.8	114	133
Radiazione Globale (MJ/m^2)	1.63	2.53	3.84
Vento (m/s)	0.52	0.24	0.21

Tabella 4.5: Varianza intra cluster nel caso 2.

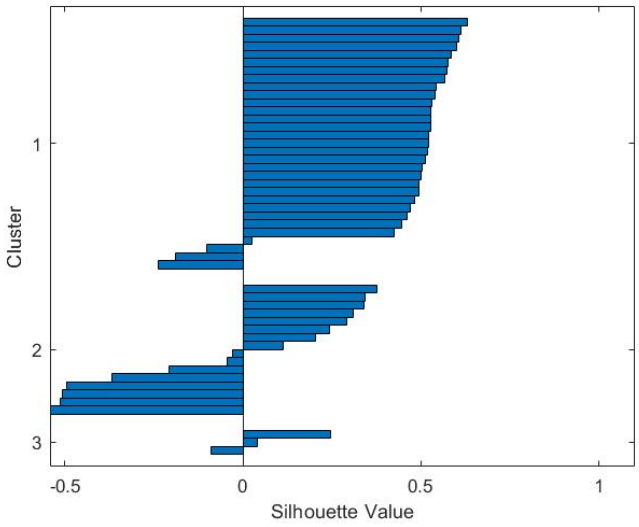


Figura 4.6: Silhouette clusters nel caso 2.

4.4 Confronto tra i due casi di studio

Si nota confrontando le zone dei clusters di entrambi i casi che i territori, coperti dai cluster, sono molto simili anche se nello studio sulle 50 stazioni si osservano delle zone molto più compatte e ben definite rispetto ai clusters nel primo caso: in cui il primo e secondo cluster si intersecano in alcune zone. Si era notato infatti, descrivendo i risultati sui coefficienti di correlazione spaziale, che mentre la precipitazione non aveva un'alta variabilità a seconda della posizione altre variabili, come l'umidità, presentano una variabilità maggiore.

Per quanto riguarda i centroidi, si osserva che la stazione 102 rimane rappresentante del cluster 2 in entrambi i casi; il cluster 1 è rappresentato da due stazioni limitrofe la 151 e la 152. Invece per il cluster 3 c'è una significativa differenza tra i due casi, perchè la diminuzione da 99 stazioni a 50 stazioni ha reso il cluster 3 molto esiguo nel secondo caso, in cui raggruppa solo 3 stazioni. Essendo comunque state mantenute le zone principali, il centroide rimane in una posizione molto simile.

Per quanto riguarda l'errore, i valori stimati considerando solo la variabile precipitazione risultano avere un errore medio molto maggiore rispetto al secondo caso di studio in cui sono state considerate più variabili; ciò può essere dovuto al fatto che la variabile precipitazione varia meno al variare dello spazio. Anche dal punto di vista della varianza, confrontando solo la variabile precipitazione, il caso con 50 stazioni risulta in una divisione più omogenea confermato anche dal metodo di Silhouette.

Capitolo 5

Conclusione

In questa tesi si sono analizzate tecniche di analisi dei dati al fine di studiare il posizionamento ottimo di sensori meteorologici. L'analisi statistica iniziale è stata fatta al fine di controllare la struttura del dataset per sottolineare le correlazioni tra variabili meteorologiche e precipitazioni oltre che evidenziare una correlazione spaziale. Tecniche di clustering sono state utilizzate per la classificazione non supervisionata dei dati. I casi presi in considerazione nello studio si riferiscono alle stazioni meteorologiche di una parte della regione Veneto e si è costruito il dataset con le singole stazioni con osservazioni che vanno dal 18/02/2022 al 18/04/22.

Lo studio del posizionamento ottimo è stato analizzato prima suddividendo le zone in cluster e in seguito, come verifica dell'informazione contenuta nella sottorete ottenuta, si è applicato il metodo di IDW per l'interpolazione spaziale per ricostruire i valori delle stazioni della rete originaria per poi calcolare l'errore medio sulla stima. Per quanto riguarda il primo caso di studio, che considera solo la precipitazione, il metodo applicato perde una buona parte dell'informazione originaria con un errore medio dell'80%, mentre per quanto riguarda il secondo caso le tecniche applicate si rivelano più accurate con un errore medio del 21%. Inoltre anche aumentando il numero di cluster a 5 i risultati non sembrano migliori anzi peggiorano leggermente, confermando che 3 risulta essere il numero di clusters ottimale per il primo caso di studio.

Le zone descritte in generale dai cluster si dimostrano molto simili ma decisamente più definite nel caso in cui si sono considerate tutte le variabili, come si nota dallo studio sulla varianza intra cluster.

Il metodo usato è uno tra i più semplici, infatti è stato utilizzato una tecnica di "hard" clustering, cioè ogni dato deve appartenere ad un singolo cluster mentre per dati meteorologici sarebbe stato forse meglio un FCM come si suggeriva in [7]. Il metodo di interpolazione utilizzato è vantaggioso perchè facile da implementare, ma poichè il sottogruppo trovato era molto esiguo i risultati ottenuti non si sono rivelati accurati.

Questo tipo di studi è importante al fine di trovare metodi per l'analisi di dataset meteorologici che diventano sempre più grandi.

Per ampliare questa ricerca in un possibile studio futuro, si potrebbe utilizzare un dataset con dati raccolti per un periodo maggiore a 60 giorni e con frequenza di raccolta maggiore di una volta al giorno. Inoltre, si potrebbero confrontare altri tipi di interpolazione spaziale e tecniche di clustering diverse.

Bibliografia

[1] S. Zubelzu, L. Rodríguez-Sinobas, D. Segovia-Cardozo, and A. Díez-Herrero, "Optimal locations for flow and velocity sensors along a river channel," *Hydrological Sciences Journal*, vol. 65, no. 5, pp. 800–812, 2020.

[2] W. H. Asquith, "The use of support vectors from support vector machines for hydrometeorologic monitoring network analyses," *Journal of Hydrology*, vol. 583, p. 124522, 2020.

[3] C. C. Castello, J. Fan, A. Davari, and R.-X. Chen, "Optimal sensor placement strategy for environmental monitoring using wireless sensor networks," in *2010 42nd Southeastern Symposium on System Theory (SSST)*, pp. 275–279, IEEE, 2010.

[4] D. Yoganathan, S. Kondepudi, B. Kalluri, and S. Manthapuri, "Optimal sensor placement strategy for office buildings using clustering algorithms," *Energy and Buildings*, vol. 158, pp. 1206–1225, 2018.

[5] S. N. Kohail and A. M. El-Halees, "Implementation of data mining techniques for meteorological data analysis," *Intl. Journal of Information and Communication Technology Research (JICT)*, vol. 1, no. 3, 2011.

[6] T. T. Bilgin and A. Y. Çamurcu, "A data mining application on air temperature database," in *International Conference on Advances in Information Systems*, pp. 68–76, Springer, 2004.

[7] F. Dikbas, M. Firat, A. C. Koc, and M. Gungor, "Classification of precipitation series using fuzzy cluster method," *International journal of climatology*, vol. 32, no. 10, pp. 1596–1603, 2012.

[8] H. Jin, X. Chen, P. Wu, C. Song, and W. Xia, "Evaluation of spatial-temporal distribution of precipitation in mainland china by statistic and clustering methods," *Atmospheric Research*, vol. 262, p. 105772, 2021.

[9] ARPAV, "Dati meteorologici ultimi 60 giorni," 18/04/2022. `www.arpa.veneto.it/bollettini/meteo60gg/Mappa_TEMP.htm`.

[10] F. Murtagh and P. Contreras, "Methods of hierarchical clustering," *arXiv preprint arXiv:1105.0121*, 2011.

[11] X. Zhang, J. Liu, Y. Du, and T. Lv, "A novel clustering method on time series data," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11891–11900, 2011.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[13] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration K-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, vol. 336, p. 012017, IOP Publishing, 2018.

[14] T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, 2012.

[15] Matlab, "Metodo Calinski-Harabasz," 30/05/2022. `https://it.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation.html`.

[16] F.-W. Chen and C.-W. Liu, "Estimation of the spatial rainfall distribution using Inverse Distance Weighting (IDW) in the middle of Taiwan," *Paddy and Water Environment*, vol. 10, no. 3, pp. 209–222, 2012.

[17] G. B. Lyra, T. P. Correia, J. F. de Oliveira-Júnior, and M. Zeri, "Evaluation of methods of spatial interpolation for monthly rainfall data over the state of Rio de Janeiro, Brazil," *Theoretical and Applied Climatology*, vol. 134, no. 3, pp. 955–965, 2018.