Applicare tecniche di clustering ai dati di qualità dell'aria urbana

Zanni Davide
Università degli studi di Modena e Reggio Emilia
Dipartimento di Ingegneria "Enzo Ferrari"
250960@studenti.unimore.it

***Abstract*—**

Questo report presenta una metodologia che sfrutta tecniche di machine learning per l'analisi delle concentrazioni di PM2.5 nell'atmosfera. L'analisi si basa su misurazioni raccolte in serie temporale da una rete di sensori ambientali distribuiti sul territorio italiano.

L'obiettivo è fornire un sistema di valutazione dell'inquinamento atmosferico che analizza automaticamente fattori ambientali e demografici.

Il lavoro si sviluppa in una pipeline strutturata che comprende: l'analisi preliminare, la preparazione e la pulizia dei dati, la visualizzazione esplorativa, il pre-processing e la selezione di modelli di clustering non supervisionati.

Nella fase iniziale del lavoro, sono state applicate tecniche di imputazione per la gestione dei valori mancanti, rimozione degli outlier e resampling temporale, al fine di ottenere un dataset con granularità oraria e giornaliera. Successivamente, la fase di visualizzazione dei dati ha permesso di esplorare le relazioni tra PM2.5 e variabili come posizione geografica e temperatura fornendo spunti utili per le analisi successive.. Nel pre-processing, le variabili numeriche sono state standardizzate e la dimensionalità dei dati è stata ridotta utilizzando la PCA, facilitando l'identificazione di schemi nei dati, e migliorando la comparabilità tra le feature per il clustering.

Per la fase di modellazione, sono stati valutati diversi algoritmi di clustering ,tra cui K-Means, Agglomerative Clustering e Spectral Clustering, ottimizzati tramite metriche come Silhouette Score ed Elbow Method. L'analisi dei risultati si è poi concentrata sugli aspetti geografici, valutando come i dispositivi e le città italiane si distribuiscono all'interno dei cluster individuati.

La fase conclusiva dell'analisi ha aggregato i dati per coordinate geografiche, identificando il cluster dominante per ciascuna area. Sono state elaborate statistiche descrittive sulla distribuzione dei cluster nelle città e realizzate visualizzazioni geografiche per evidenziare la diffusione territoriale. È stata inoltre condotta un'analisi della popolazione associata a ciascun cluster e valutata la correlazione tra variabili ambientali e demografiche, come popolazione, altitudine e concentrazione di PM2.5.

Particolare importanza è stata data alla valutazione del rischio ambientale rispetto alla soglia annuale di PM2.5 (25 µg/m³) e al confronto tra i cluster individuati e le reali condizioni climatiche osservate. I risultati dimostrano l'efficacia della metodologia proposta nell'individuare correlazioni complesse nei dati e nel fornire supporto decisionale per la gestione della qualità dell'aria, contribuendo concretamente alle strategie di controllo e riduzione dell'inquinamento atmosferico.

I. Introduzione

L'insieme delle particelle atmosferiche solide e liquide sospese in aria vengono definite materiale particolato (PM, dall'inglese *particulate matter*). Con PM2,5 si identificano quelle particelle il cui diametro è inferiore o uguale ai 2,5 micron. [1]

Queste particelle, per la loro dimensione ridotta, possono penetrare profondamente nel sistema respiratorio umano, raggiungendo gli alveoli polmonari e, in alcuni casi, entrando nel flusso sanguigno, rappresentano quindi uno degli inquinanti atmosferici più pericolosi per la salute umana.

Le fonti principali di PM2.5 includono il traffico urbano, il riscaldamento residenziale e le attività industriali.

A. Effetti sulla salute

L'esposizione al PM2.5 è stata associata a una serie di effetti negativi sulla salute, tra cui malattie cardiovascolari, problemi respiratori, asma e, in alcuni casi, cancro.

Studi epidemiologici hanno evidenziato un aumento della mortalità e dei ricoveri ospedalieri correlati all'inquinamento da particolato fine .

B. Normative e limiti

In Italia, la normativa vigente stabilisce un valore limite annuale di 25 μg/m³ per il PM2.5, in linea con le direttive europee.

Tuttavia, l'Organizzazione Mondiale della Sanità raccomanda un valore molto più restrittivo, pari a 5 μg/m³, per garantire una maggiore tutela della salute pubblica [2].

È importante sottolineare che, nonostante il rispetto dei limiti normativi, l'esposizione a livelli di PM2.5 seppur inferiori può comunque comportare rischi significativi per la salute.

C. Considerazioni per l'analisi dei dati

Nel contesto di questo progetto, l'attenzione è rivolta alle misurazioni di PM2.5 ottenute da una rete di sensori distribuiti sul territorio.

Per assicurare la qualità e la coerenza delle analisi, i dati vengono sottoposti a un processo di resampling orario.

Oltre al PM2.5, il dataset include variabili meteorologiche di contesto (velocità del vento, umidità relativa, temperatura, pressione atmosferica e precipitazioni), che possono essere integrate nell'analisi per valutare l'influenza dei fattori ambientali sulla concentrazione di particolato.

Le coordinate geografiche dei sensori rappresentano inoltre un elemento chiave per l'analisi spaziale e la caratterizzazione delle aree a maggiore rischio di inquinamento.

II. Definizione Problema

Il problema affrontato in questo progetto riguarda l'analisi e la comprensione delle dinamiche di concentrazione del particolato fine PM2.5 nell'atmosfera, utilizzando dati raccolti in serie temporale da una rete di sensori ambientali distribuiti sul territorio.

L'obiettivo è identificare pattern spaziali e temporali che influenzano la presenza di PM2.5, al fine di supportare strategie di monitoraggio e mitigazione dell'inquinamento atmosferico.

La complessità del problema deriva dalla natura multidimensionale dei dati: le misurazioni di PM2.5 sono influenzate da molteplici variabili e presentano una variabilità sia nello spazio che nel tempo.

La capacità di analizzare in modo automatizzato e affidabile le concentrazioni di PM2.5 consente di individuare aree e periodi a maggiore rischio, fornendo informazioni preziose per enti pubblici e cittadini.

Un'analisi approfondita delle relazioni tra PM2.5 e variabili ambientali può inoltre contribuire a una migliore comprensione dei meccanismi di formazione e dispersione del particolato, supportando interventi mirati di prevenzione e riduzione dell'esposizione.

III. Analisi Dati

A. Informazioni Dataset

Il dataset analizzato è costituito da oltre 7,3 milioni di osservazioni (7.323.392 istanze) e 15 colonne, raccolte da una rete di sensori ambientali distribuiti sul territorio italiano. Ogni istanza rappresenta una misurazione puntuale effettuata da un sensore in un determinato istante temporale e in una specifica posizione geografica.

Le principali feature del dataset sono:

- `created_at_utc_original`: timestamp della misurazione in formato UTC.
- `municipality`: nome del comune in cui è posizionato il sensore.
- `complete_address`: indirizzo completo del sensore.
- `device_id`: identificativo univoco del dispositivo di misura.
- `latitude` e `longitude`: coordinate geografiche del sensore.
- `pm1_sps30_ug_m3`, `pm2p5_sps30_ug_m3`, `pm4_sps30_ug_m3`, `pm10_sps30_ug_m3`: concentrazioni di particolato (PM1, PM2.5, PM4, PM10) espresse in microgrammi per metro cubo.
- `wind_speed_owm_m_s`: velocità del vento (m/s).
- `rh_percentage`: umidità relativa (
- `temperature_celsius`: temperatura dell'aria (°C).
- `pressure_pa`: pressione atmosferica (Pa).
- `rain_last_1h_mm`: precipitazioni nell'ultima ora (mm).

Le feature sono di tipo sia numerico (float64) sia categorico (object), con una prevalenza di misure ambientali continue.

B. Statistiche descrittive

L'analisi preliminare mostra che le variabili ambientali presentano una notevole variabilità. Ad esempio, la concentrazione media di PM2.5 è di circa 54,5 μg/m³, con una deviazione standard elevata (348,4 μg/m³), a testimonianza della presenza di valori estremi (outlier).

Le coordinate geografiche coprono un range coerente con il territorio italiano, mentre le variabili meteorologiche (temperatura, umidità, vento, pressione, pioggia) mostrano valori compatibili con le condizioni climatiche tipiche delle diverse regioni e stagioni.

C. Verifica di valori nulli

Un aspetto rilevante emerso dall'analisi esplorativa riguarda la presenza di valori mancanti. Complessivamente, sono stati rilevati 275.776 valori nulli distribuiti su diverse colonne.

In particolare, le variabili meteorologiche come `rh_percentage` (umidità relativa), `wind_speed_owm_m_s` (velocità del vento), `temperature_celsius`, `pressure_pa` e `rain_last_1h_mm` presentano il maggior numero di dati mancanti, mentre le misurazioni di particolato PM2.5 risultano complete.

La presenza di valori nulli richiede l'adozione di strategie di imputazione, a seconda della rilevanza delle variabili e dell'impatto sull'analisi successiva.

IV. Preparazione Dati

A. Rimozione delle Colonne Inutilizzate

Per ottimizzare l'analisi, sono state elminate alcune features considerate irrilevanti:

- `municipality` e `complete_address`: informazioni ridondanti rispetto alle coordinate geografiche.
- `pm1_sps30_ug_m3`, `pm4_sps30_ug_m3`, `pm10_sps30_ug_m3`: concentrazioni di particolato non rilevanti per l'analisi sul PM2.5.

Il dataset risultante conserva solo le informazioni essenziali relative all' identificazione temporale e spaziale, alla concentrazione PM2.5 e alle variabili meteorologiche

B. Imputazione Valori Mancanti

Per gestire i valori mancanti nelle variabili meteorologiche è stato utilizzato il metodo SimpleImputer con strategia *median*:

La scelta della mediana è motivata dalla sua minore sensibilità agli *outlier* rispetto alla media, garantendo così una stima più robusta e rappresentativa della distribuzione reale dei dati.

Al termine di questa operazione, il dataset risulta privo di valori nulli.

C. Resampling Temporale dei Dati

Per standardizzare la frequenza delle misurazioni e rendere il dataset omogeneo, è stato effettuato un resampling orario:

Questo processo consiste nel raggruppare le osservazioni per ciascun sensore e selezionare, per ogni ora, la misurazione più vicina all'ora esatta tramite il metodo "nearest".

- Conversione della colonna temporale in formato `datetime`;
- Impostazione della colonna temporale come indice;
- Raggruppamento per `device_id` e ricampionamento orario utilizzando il metodo `"nearest"`.

Il resampling ha comportato una significativa riduzione del numero di istanze:

- **Pre-resampling:** 7.323.392 record;
- **Post-resampling:** 2.335.357 record.

D. Rimozione degli Outlier

La qualità dei dati ambientali può essere compromessa dalla presenza di valori anomali, spesso dovuti a errori di misura, condizioni estreme o malfunzionamenti dei sensori.

Per identificare e rimuovere tali valori, è stato applicato un filtro basato sul criterio delle 3 deviazioni standard:

- Calcolo di media e deviazione standard per ogni variabile numerica;
- Rimozione delle osservazioni che si discostano dalla media di più di 3 deviazioni standard;
- Il filtro è stato applicato considerando tutte le variabili numeriche.

Al termine di questa fase, il dataset contiene **2.196.839** record, garantendo una maggiore robustezza statistica per le analisi successive.

Questa fase di preparazione dei dati ha prodotto un dataset più pulito, strutturato e statisticamente affidabile, pronto per le successive fasi di analisi esplorativa e modellazione.

V. Visualizzazione Dati

A. Analisi spaziale del particolato PM2.5

La mappa geografica delle concentrazioni medie di PM2.5 (Fig.1) evidenzia come i valori più elevati si concentrino nel Nord Italia, in particolare nelle aree urbane e industrializzate. In queste regioni si osserva anche una maggiore densità di punti di rilevamento, probabilmente legata a una più ampia rete di monitoraggio.

Procedendo verso Sud, le concentrazioni di PM2.5 tendono a diminuire progressivamente, riflettendo differenze sia nelle condizioni ambientali che nelle fonti di emissione.

B. Analisi del PM2.5 rispetto alla temperatura

La relazione tra PM2.5 e temperatura (Fig.2), illustrata tramite boxplot per intervalli di temperatura, mostra come le concentrazioni di particolato aumentino al diminuire della temperatura, soprattutto con valori elevati al di sotto degli 11°C. Questo andamento è riconducibile a fenomeni atmosferici come l'inversione termica, che nei mesi più freddi, ostacola

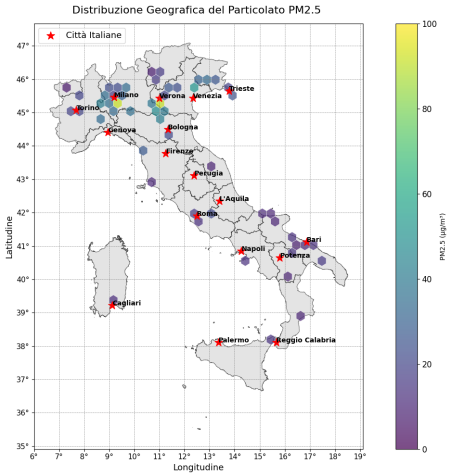


Fig. 1: Distribuzione Geografica del Particolato PM2.5.

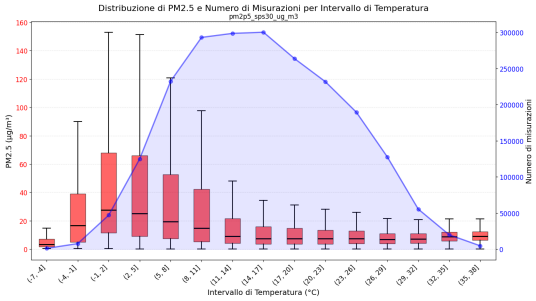


Fig. 2: Distribuzione di PM2.5 e Numero di Misurazioni per Intervallo di Temperatura.

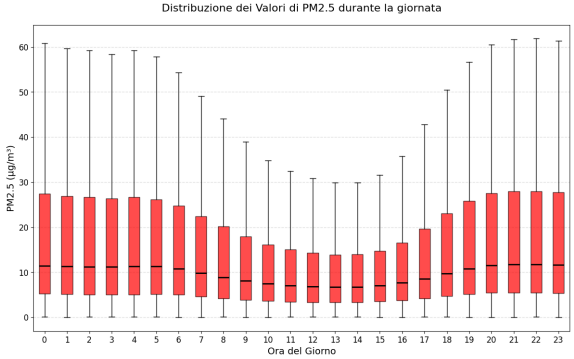


Fig. 3: Distribuzione dei Valori di PM2.5 durante la giornata.

la dispersione degli inquinanti e favorendone l'accumulo negli strati più bassi dell'atmosfera.

Al contrario, durante i periodi più caldi, la maggiore instabilità e turbolenza atmosferica contribuiscono a una più efficace dispersione del particolato, determinando concentrazioni di PM2.5 generalmente più basse.

C. Analisi temporale giornaliera del PM2.5

L'andamento orario dei valori di PM2.5 (Fig.3) evidenzia una marcata ciclicità giornaliera: le concentrazioni risultano più elevate durante le ore notturne e nelle prime ore del

mattino (indicativamente tra le 20 e le 7), mentre si riducono sensibilmente nelle ore centrali della giornata.

Questo comportamento è strettamente legato alle variazioni di temperatura e alle condizioni di stabilità atmosferica: durante la notte e al mattino, queste condizioni favoriscono l'accumulo degli inquinanti, mentre nelle ore più calde la maggiore ventilazione contribuisce a ridurre la concentrazione del particolato.

VI. Preprocessing

La fase di preprocessing ha l'obiettivo di preparare il dataset per le tecniche di clustering, garantendo uniformità tra le variabili, facilitando i confronti e semplificando la complessità dimensionale.

Questo processo ha incluso: analisi della matrice di correlazione, resampling temporale giornaliero, standardizzazione delle variabili numeriche e riduzione dimensionale tramite PCA.

A. Analisi della matrice di correlazione

La matrice di correlazione tra le variabili numeriche, visualizzata tramite heatmap, permette di individuare le relazioni lineari tra le diverse feature.

In particolare, si osservano correlazioni negative tra temperatura e PM2.5 (**-0.24**), e tra umidità e temperatura (**-0.45**), mentre la correlazione tra PM2.5 e le altre variabili meteorologiche risulta generalmente debole o moderata. Queste informazioni sono utili sia per la selezione delle feature che per l'interpretazione successiva dei risultati dei modelli di clustering.

B. Resampling temporale giornaliero

Per ottenere una serie temporale regolare e ridurre la variabilità dovuta a misurazioni multiple nello stesso giorno, i dati di ciascun sensore sono stati ricampionati su base giornaliera. Questo processo ha comportato una drastica riduzione del numero di istanze (circa **96.000**).

C. Standardizzazione delle variabili numeriche

Le variabili numeriche sono state standardizzate tramite lo *StandardScaler*, trasformando ciascuna feature affinché abbia media nulla e varianza unitaria. Questo processo è fondamentale per garantire che tutte le variabili contribuiscano allo stesso modo alle analisi clustering e per evitare che variabili con intervalli di valori più ampi dominino l'analisi e distorcano i risultati del clustering.

D. Riduzione dimensionale tramite PCA

Per facilitare la visualizzazione e migliorare l'efficacia dei modelli di clustering, è stata applicata la Principal Component Analysis (PCA). La PCA ha permesso di ridurre il numero di variabili mantenendo almeno l'85% della varianza totale, ottenendo così 5 componenti principali.

Questa trasformazione consente di sintetizzare l'informazione contenuta nelle variabili originali in un numero ridotto di dimensioni, semplificando l'analisi e limitando il rumore.

L'errore di ricostruzione si mantiene basso (circa **0.10**) e la varianza cumulativa spiegata raggiunge quasi il **90%**, confermando la validità della riduzione dimensionale ottenuta.

VII. Model Selection

La fase di selezione e valutazione dei modelli di clustering è fondamentale per identificare pattern nei dati e raggruppare i dispositivi in base a comportamenti simili. Questo processo considera contemporaneamente i livelli di particolato e le condizioni meteorologiche.

Per questa analisi di serie temporali multivariate, sono stati scelti tre algoritmi principali: K-Means, Agglomerative Clustering e Spectral Clustering.

A. Identificazione Numero Ottimale

Per identificare il numero ottimale di cluster sono state utilizzate due metriche principali:

- **Elbow Method**: consente di individuare il numero analizzando la curva del WCSS (Within-Cluster Sum of Squares). Dal grafico, il "gomito" suggerisce che il numero ideale di cluster si ha a 6, anche se il risultato non è del tutto chiaro e univoco.
- **Silhouette Score**: misura la coesione e la separazione tra i cluster. Per K-Means, il valore massimo di separazione tra i gruppi si ottiene con 6 cluster (Silhouette Score = 0.2375), mentre valori inferiori portano a una distinzione meno marcata tra i gruppi.

B. Descrizione e confronto dei modelli

- **K-Means**: Algoritmo rapido e scalabile, adatto a grandi dataset. Ha permesso di identificare 6 cluster principali, con una buona separazione tra i gruppi e una distribuzione equilibrata dei dispositivi tra i cluster.
- **Agglomerative Clustering**: Algoritmo gerarchico che costruisce una struttura ad albero (dendrogramma) dei cluster. L'analisi del dendrogramma mostra che il taglio ottimale si ottiene con 6 cluster, corrispondenti ai primi livelli di separazione significativa. Questo metodo è particolarmente utile per visualizzare la gerarchia e le relazioni tra i gruppi.
- **Spectral Clustering**: Algoritmo che sfrutta le proprietà spettrali della matrice di similarità tra i dati. Ha evidenziato una buona capacità di individuare cluster anche in presenza di strutture non lineari.

C. Analisi Dendogramma

Il dendrogramma generato dall'Agglomerative Clustering fornisce una rappresentazione visiva della gerarchia dei cluster. Tagliando l'albero a un livello di distanza appropriato, si ottengono cluster ben separati e interpretabili.

L'analisi visiva conferma che la scelta di 6 cluster rappresenta un buon compromesso tra granularità e coerenza interna dei gruppi.

D. Risultati quantitativi

- K-Means: 6 cluster, con una distribuzione dei dispositivi relativamente bilanciata.
- Agglomerative Clustering: 6 cluster, con alcuni gruppi più numerosi e altri più piccoli, riflettendo la struttura gerarchica dei dati.
- Spectral Clustering: 6 cluster, con una buona separazione tra i gruppi e la capacità di cogliere pattern complessi.

VIII. Rappresentazione Risultati di Clustering

La fase di rappresentazione e interpretazione dei risultati di clustering è fondamentale per comprendere la distribuzione spaziale e la composizione dei gruppi individuati dagli algoritmi.

Le visualizzazioni realizzate permettono di analizzare sia la localizzazione geografica dei cluster sia la numerosità dei dispositivi appartenenti a ciascun gruppo. Di seguito vengono illustrati i risultati ottenuti, prendendo come esempio il modello K-Means.

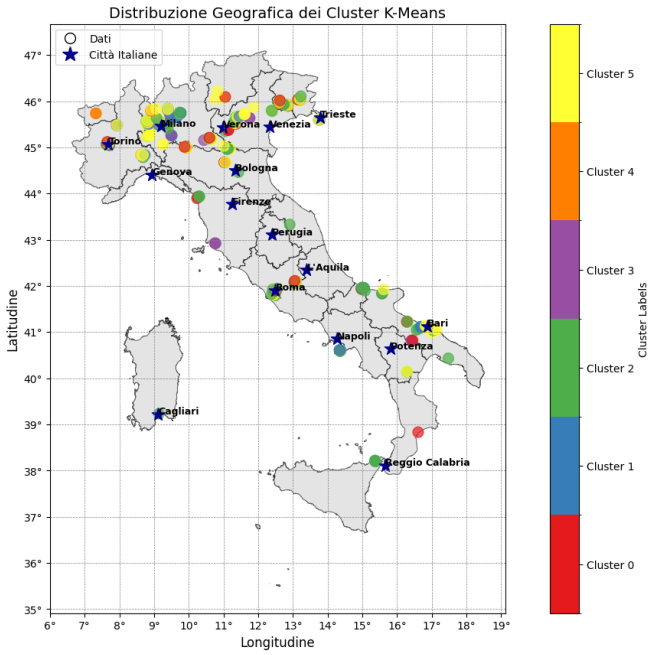


Fig. 4: Distribuzione geografica cluster K-Means

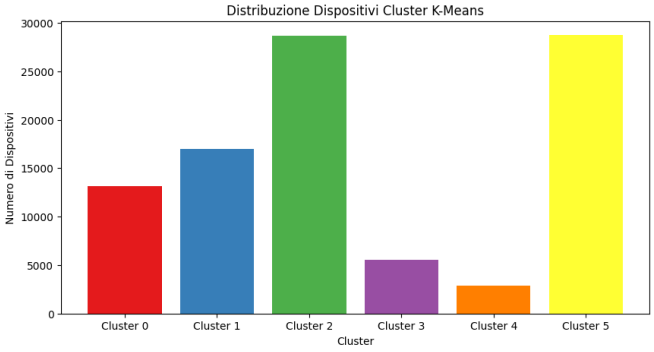


Fig. 5: Composizione cluster K-Means

A. Distribuzione geografica dei cluster

La mappa mostra la distribuzione dei cluster K-Means sul territorio italiano (Fig.4), associando a ciascuna posizione geografica il cluster più frequente tra le misurazioni del sensore. I punti colorati rappresentano i cluster assegnati, mentre le stelle indicano le principali città italiane.

Questa rappresentazione consente di evidenziare la presenza di pattern spaziali: alcuni cluster risultano prevalenti in specifiche aree geografiche, riflettendo differenze locali nelle condizioni ambientali e nei livelli di inquinamento da PM2.5.

B. Analisi composizione dei cluster

L'istogramma della distribuzione dei dispositivi per cluster fornisce una panoramica quantitativa della composizione dei gruppi individuati (Fig.5). Si nota una certa variabilità nella numerosità: alcuni cluster raccolgono un numero molto elevato di dispositivi, mentre altri risultano più ristretti.

L'analisi della composizione dei cluster è utile per identificare eventuali squilibri e per approfondire le peculiarità dei gruppi più rappresentati.

C. Metodologia di aggregazione

Per ogni sensore, sono state calcolate le coordinate medie e assegnato il cluster più frequente, in modo da rappresentare in modo sintetico la posizione e l'appartenenza di ciascun dispositivo.

Questo approccio consente di ottenere una mappa chiara e facilmente interpretabile, in cui ogni punto rappresenta la tipica e più frequente appartenenza a un cluster di una determinata area.

IX. Analisi Cluster

A. Selezione modello finale (KMeans)

Dopo aver confrontato diversi algoritmi di clustering, il modello K-Means è stato selezionato come soluzione finale per la sua capacità di individuare gruppi ben separati e facilmente interpretabili, sia dal punto di vista geografico che ambientale. Il numero ottimale di cluster è stato fissato a 6, sulla base delle metriche di valutazione.

B. Analisi caratteristiche dei cluster

L'analisi delle medie delle variabili principali per ciascun cluster ha permesso di caratterizzare i gruppi individuati (Fig.7):

- **Cluster 3** si distingue per i valori di PM2.5 molto elevati e per una grande variabilità interna, suggerendo la presenza di aree o periodi con condizioni di inquinamento particolarmente critiche.
- **Cluster 0** è caratterizzato da una velocità del vento superiore rispetto agli altri cluster, condizione che favorisce la dispersione del particolato e contribuisce a mantenere bassi i livelli di PM2.5.
- **Cluster 4** mostra valori elevati di umidità, pressione atmosferica e pioggia, indicando condizioni meteorologiche specifiche che possono influenzare la formazione e la permanenza del particolato.

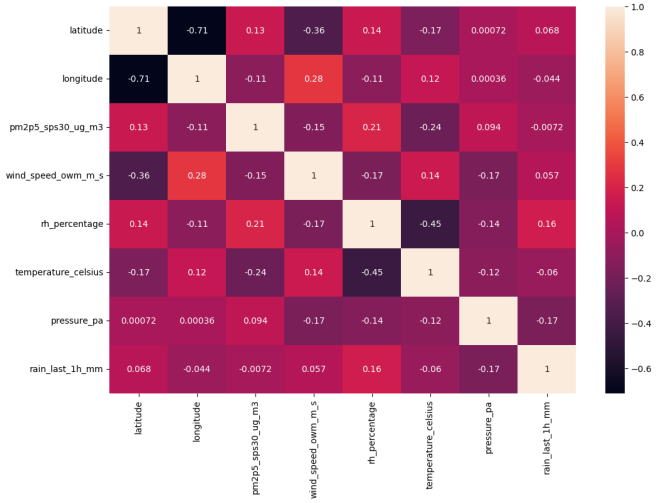


Fig. 6: Matrice di correlazione

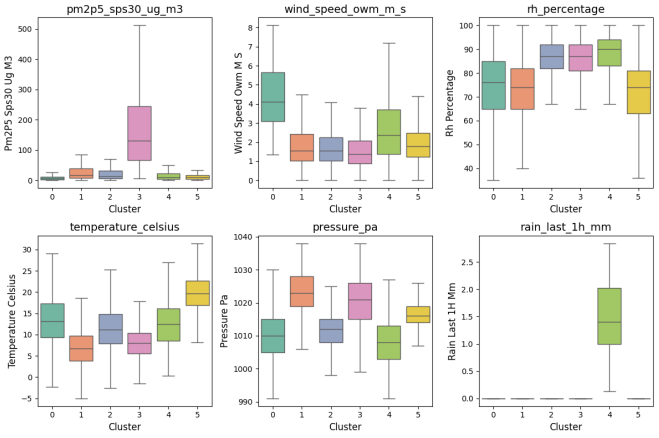


Fig. 7: Analisi medie delle variabili ambientali per cluster

- **Cluster 5** è associato a temperature più alte e umidità più bassa, tipiche di condizioni più calde e secche, che generalmente favoriscono la dispersione degli inquinanti.
- Gli altri cluster (**1 e 2**) presentano valori intermedi, con livelli di PM2.5 relativamente bassi e condizioni meteorologiche variabili.

L'analisi dei boxplot per ciascun cluster conferma queste osservazioni, evidenziando la variabilità interna e le differenze tra i gruppi per tutte le principali variabili ambientali.

Il confronto con la matrice di correlazione (Fig.6) rafforza alcune interpretazioni:

- La **velocità del vento** mostra una correlazione negativa con il PM2.5, confermando il ruolo del vento nella dispersione degli inquinanti.
- **Umidità relativa** e **temperatura** influenzano la concentrazione di PM2.5: condizioni di alta umidità e bassa temperatura favoriscono l'accumulo di particolato, come osservato nel Cluster 3.
- **Pioggia** e **pressione atmosferica** non mostrano correlazioni forti con il PM2.5, ma possono contribuire a definire condizioni meteorologiche particolari, come nel Cluster 4.

C. Assegnazione dei Sensori alle Città Italiane

Per una lettura territoriale dei risultati, i sensori sono stati associati alle principali città italiane sulla base della prossimità geografica, considerando un valore di soglia di distanza di circa 111km. Questo consente di analizzare la distribuzione dei cluster a livello urbano e di confrontare le condizioni ambientali tra le diverse città.

D. Analisi Territoriale e Demografica dei Cluster

TABLE I: Distribuzione delle città nei cluster

Città	Dominant Cluster	Ist. totali	% dominante
Perugia	2	455	37.31
L'Aquila	2	5428	35.90
Roma	2	8255	33.64
Torino	5	7375	32.71
Napoli	5	1645	32.38
Genova	5	6022	31.63
Milano	5	25986	31.38
Trieste	2	7772	31.24
Bologna	5	6276	31.08
Reggio Calabria	2	906	30.97
Firenze	5	3359	30.89
Potenza	5	4332	30.39
Verona	2	16671	30.21
Cagliari	5	552	29.98
Bari	2	9801	29.28
Venezia	2	12441	29.25

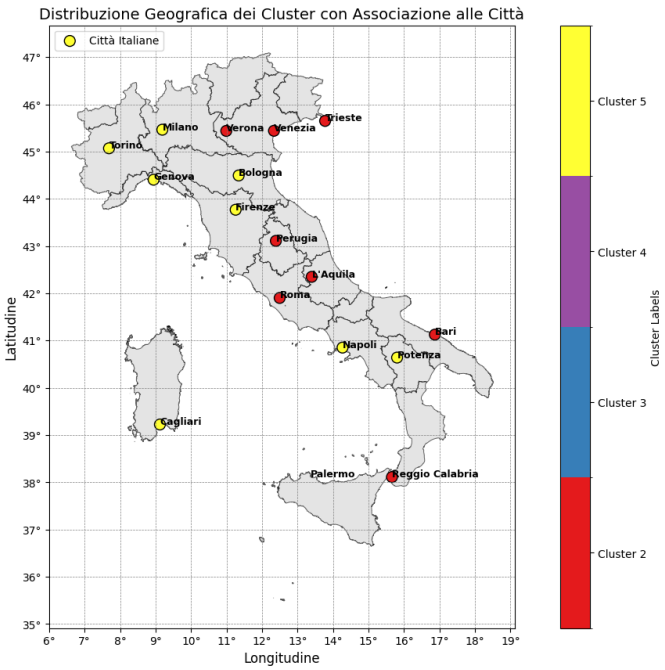


Fig. 8: Distribuzione geografica dei cluster con associazione alle città

L'assegnazione dei sensori alle città italiane (Fig.7) permette di approfondire l'analisi dei cluster anche dal punto di vista demografico e territoriale.

Per ciascuna città, è stata calcolata la distribuzione delle istanze tra i diversi cluster, individuando il cluster dominante e la sua incidenza percentuale sul totale delle osservazioni.

Questa analisi consente di confrontare le condizioni ambientali prevalenti tra le città e di identificare eventuali pattern ricorrenti a livello urbano.

Dai dati aggregati emerge che città come **Milano**, **Torino**, **Napoli**, **Firenze** e **Genova** sono prevalentemente associate al **Cluster 5**, caratterizzato da temperature più elevate e livelli di PM2.5 contenuti.

Al contrario, città come **Roma**, **Bari**, **Venezia**, **Trieste** e **Verona** risultano più frequentemente associate al **Cluster 2**, che presenta condizioni intermedie in termini di concentrazione di particolato e variabili meteorologiche.

Tuttavia, la percentuale di dominanza del cluster principale raramente supera il 35%, a conferma della forte variabilità interna delle condizioni ambientali anche all'interno della stessa città.

L'analisi della popolazione totale associata a ciascun cluster mostra che il **Cluster 5** raccoglie la quota maggiore di popolazione (oltre **4.7 milioni di abitanti**), seguito dal **Cluster 2** (circa **4.3 milioni**). Questo dato suggerisce che la maggior parte della popolazione delle principali città italiane vive in aree caratterizzate da condizioni ambientali relativamente favorevoli, con livelli di PM2.5 generalmente inferiori rispetto ad altri cluster.

Per ogni cluster è stato inoltre calcolato il valore medio di PM2.5, permettendo di associare a ciascun gruppo non solo una caratterizzazione meteorologica, ma anche un profilo di rischio ambientale. L'integrazione di queste informazioni consente di valutare l'esposizione della popolazione ai diversi livelli di inquinamento atmosferico.

E. Correlazioni tra Variabili

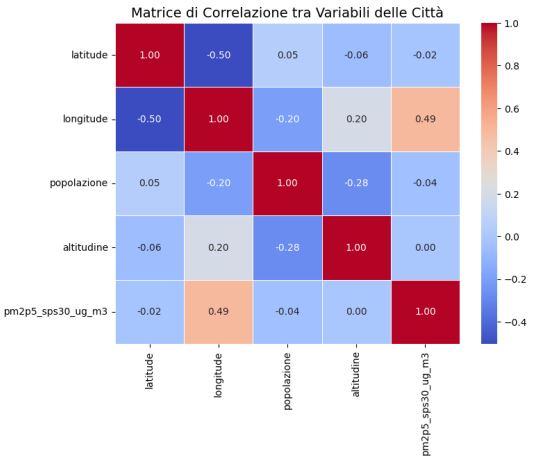


Fig. 9: Matrice di correlazione

L'analisi di correlazione tra le variabili (Fig.9) (latitudine, longitudine, popolazione, altitudine e PM2.5) evidenzia alcuni aspetti interessanti:

- **Longitudine** e **PM2.5** mostrano una correlazione (**0.49**), suggerendo che le città più a est tendono ad avere un inquinamento superiore.
- **Popolazione** e **PM2.5** non presentano una correlazione significativa (**-0.04**), indicando che i livelli di particolato non dipendono direttamente dalla densità abitativa, ma sono probabilmente influenzati da altri fattori, come le condizioni climatiche, la ventilazione e la presenza di fonti di emissione.
- **Altitudine** e **PM2.5** mostrano una correlazione trascurabile (**0.00**), confermando che l'altitudine non rappresenta un fattore determinante per la concentrazione di particolato nelle città analizzate.

F. Analisi Rischio Ambientale delle Città

L'analisi del rischio ambientale è stata condotta confrontando i valori medi di PM2.5 rilevati nei cluster urbani con il limite annuale stabilito dalla normativa europea, fissato a 25 μg/m³. Per ogni città e per ciascun cluster, è stato calcolato un indice di rischio come rapporto tra la concentrazione media di PM2.5 e il valore limite normativo. Un indice superiore a 1 indica il superamento della soglia di sicurezza, segnalando una potenziale criticità per la salute pubblica.

TABLE II: Indicatori ambientali per i cluster con confronto rispetto al limite normativo di PM2.5

Cluster	Popolazione	PM2.5	Rischio
2.0	4.332.823	28.53	1.14
5.0	4.774.812	15.96	0.64

TABLE III: Città associate a cluster con indice di rischio > 1

Città	Cluster
Roma	2.0
Venezia	2.0
Bari	2.0
Verona	2.0
Trieste	2.0
Reggio Calabria	2.0
Perugia	2.0
L'Aquila	2.0

Dall'analisi emerge che il **Cluster 2** presenta un indice di rischio superiore a 1 (**1.14**), mentre il **Cluster 5** si mantiene al di sotto della soglia (**0.64**).

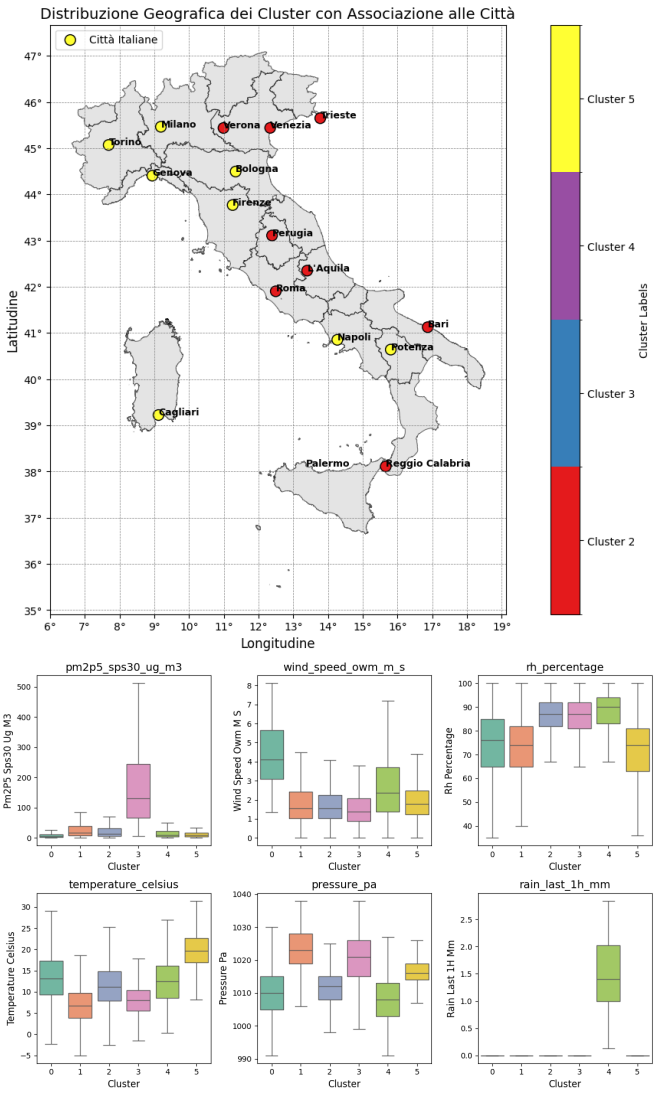
In particolare, città come **Roma**, **Venezia**, **Bari**, **Verona**, **Trieste**, **Reggio Calabria**, **Perugia** e **L'Aquila** risultano associate a cluster con un indice di rischio superiore al limite normativo, evidenziando la necessità di particolare attenzione e di strategie di mitigazione mirate in questi contesti urbani.

Questa valutazione quantitativa permette di identificare rapidamente le aree più esposte al rischio di superamento dei limiti di PM2.5, fornendo uno strumento utile per la pianificazione di interventi e per la comunicazione del rischio alla popolazione.

G. Analisi Completa dei Cluster e Confronto con le Condizioni Climatiche Reali

L'analisi dei cluster, basato sui boxplot delle variabili climatiche e sulla categorizzazione delle città, permette di

mettere in relazione i livelli di rischio ambientale con le condizioni meteorologiche.



a) Cluster 2: raggruppa città ad alta popolazione, spesso situate nel centro-sud e lungo la costa, caratterizzate da:

- **PM2.5**: valori medi prossimi o superiori al limite normativo.
- **Umidità relativa**: elevata, con mediane tra 80% e 90%;
- **Temperatura**: mediamente mite, intorno ai 10°C;
- **Vento**: moderato, circa 2 m/s;
- **Pressione atmosferica**: tendenzialmente bassa;
- **Pioggia**: quasi assente.

Tali condizioni climatiche, tipiche di molte città italiane costiere e del centro-sud, favoriscono la persistenza del particolato in atmosfera, soprattutto in presenza di umidità elevata e scarsa ventilazione.

b) Cluster 5: , invece, comprende città prevalentemente del nord e del centro, spesso a maggiore altitudine, con:

- **PM2.5**: valori medi inferiori al limite normativo;
- **Umidità relativa**: più bassa rispetto al Cluster 2 (mediana intorno al 70%);
- **Temperatura**: più elevata;
- **Vento**: moderato, simile al Cluster 2;
- **Pressione atmosferica**: più alta;
- **Pioggia**: trascurabile.

Queste città presentano condizioni più secche e ventilate, che favoriscono la dispersione degli inquinanti e contribuiscono a mantenere i livelli di PM2.5 sotto la soglia di rischio.

Il confronto con le condizioni climatiche reali conferma la coerenza dei cluster individuati:

le città del Cluster 2 sono effettivamente caratterizzate da alta umidità e temperature miti, mentre quelle del Cluster 5 godono di un clima più secco e ventilato.

Questo risultato sottolinea l'importanza di integrare dati ambientali e climatici nell'analisi del rischio, per una valutazione più accurata e per la definizione di strategie di prevenzione e mitigazione efficaci a livello urbano.

X. Conclusione

In conclusione, l'analisi condotta ha evidenziato come la combinazione tra dati ambientali, climatici e indicatori di rischio permetta di individuare con precisione le aree urbane maggiormente esposte a livelli critici di PM2.5.

In particolare, le città appartenenti al Cluster 2 richiedono interventi prioritari, poiché presentano condizioni climatiche che favoriscono l'accumulo di particolato e superano i limiti normativi di sicurezza. Al contrario, le città del Cluster 5 beneficiano di un clima più favorevole alla dispersione degli inquinanti, mantenendo i livelli di PM2.5 sotto la soglia di rischio.

Questi risultati sottolineano l'importanza di un approccio integrato nell'analisi del rischio ambientale, fondamentale per supportare decisioni informate e strategie di mitigazione efficaci a tutela della salute pubblica nelle aree.

XI. Fonti bibliografiche e sitografia

References

[1] Ancler. (2023) Pm 2.5: Cos'è, da dove deriva e quali sono i rischi. [Online]. Available: https://ancler.org/pm-25/

[2] ISPRA. (2025) Qualità dell'aria ambiente: particolato (pm2,5). Accesso: 23 aprile 2025. [Online]. Available: https://indicatoriambientali.isprambiente.it/it/qualita-dellaria/qualita-dellaria-ambiente-particolato-pm25

List of Figures

List of Tables