UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione



TESI DI LAUREA

Analisi del ruolo degli algoritmi e dei dati per la predizione di inquinanti

Analysis of the role of algorithms and data in predicting polluttants

Relatore

Prof. Domenico Ursino

Correlatore

Prof. Gianluca Bonifazi

Candidata

Alessandra D'Anna

ANNO ACCADEMICO 2025-2026

Ai miei nonni

Sommario

La presente tesi analizza l'impiego di modelli di previsione di serie temporali per lo studio dell'evoluzione delle principali grandezze associate alla qualità dell'aria. Il lavoro è stato sviluppato attraverso una fase di acquisizione, pulizia e organizzazione dei dati provenienti da sistemi di monitoraggio della qualità dell'aria, seguita da un'analisi esplorativa delle serie temporali considerate. Sono stati sperimentati differenti approcci di previsione, comprendenti modelli statistici, quali ARIMAX, SARIMA e ARIMA, modelli basati sulla decomposizione delle componenti temporali, quali Prophet e NeuralProphet, e modelli basati su reti neurali e approcci ibridi, tra cui LSTM, BiLSTM, Prophet–LSTM ed EMD-BiLSTM. Sono state analizzate tecniche basate su reti neurali grafiche e metodi di decomposizione avanzata, quali CEEMDAN, combinati con differenti modelli predittivi. Le prestazioni dei modelli sono state valutate mediante le metriche MAE, RMSE e MAPE e attraverso il confronto grafico tra valori osservati e previsti. Una parte dell'attività ha riguardato il passaggio a un dataset più esteso, comprendente diverse centraline distribuite sul territorio marchigiano. Le sperimentazioni hanno permesso di analizzare l'influenza della quantità di dati disponibili sulle capacità predittive dei modelli e di evidenziare differenze nel comportamento degli approcci considerati al variare della serie e della centralina analizzata.

Keyword: Inquinamento atmosferico, Prophet, NeuralProphet, Neural Network, Time series, LSTM, Machine Learning, Air quality forecasting

Indice

Introduzione **1**

1 La predizione degli inquinanti **4**
1.1 Problema dell'inquinamento 4
1.1.1 Le cause dell'inquinamento atmosferico 5
1.1.2 Conseguenze sulla salute 6
1.1.3 Importanza del monitoraggio 7
1.2 I principali inquinanti atmosferici 7
1.2.1 Inquinanti primari 7
1.2.2 Inquinanti secondari 9
1.3 Dal monitoraggio alla previsione 10
1.3.1 Dalla misura alla modellazione 10
1.3.2 Verso la previsione 10
1.3.3 Modello attuale 11
1.4 Importanza della previsione 12

2 Un primo dataset di riferimento **13**
2.1 Descrizione del dataset 13
2.2 Analisi esplorativa 15
2.2.1 Preprocessing e organizzazione 15
2.2.2 Analisi esplorativa 16
2.2.3 Analisi della copertura temporale 16
2.2.4 Analisi delle correlazioni tra variabili 17
2.3 Criticità del dataset 19

3 Applicazione di tecniche classiche al dataset di riferimento **22**
3.1 Definizione degli esperimenti 22
3.2 Modelli utilizzati 23
3.2.1 ARIMAX 23
3.2.2 SARIMA 25
3.2.3 Metriche di valutazione 26
3.3 Risultati sperimentali 27
3.3.1 Risultati ARIMAX 27
3.3.2 Risultati SARIMA 33
3.4 Discussione dei risultati 33

4 Analisi di tecniche innovative per la predizione degli inquinanti **38**
4.1 Introduzione a modelli predittivi avanzati . . . 38
4.2 Modelli basati sulla decomposizione delle componenti temporali . . . 39
4.2.1 Prophet . . . 39
4.2.2 NeuralProphet . . . 40
4.3 Approcci ibridi per la previsione delle serie temporali . . . 41
4.3.1 LSTM . . . 41
4.3.2 Prophet e LSTM . . . 42
4.4 Modelli basati su reti neurali grafiche . . . 43
4.4.1 GNN . . . 43
4.5 Tecniche di decomposizione avanzata delle serie temporali . . . 45
4.5.1 EEMD-ALSTM . . . 45
4.5.2 EMD-BiLSTM . . . 46
4.5.3 CEEMDAN . . . 47

5 Applicazione di tecniche innovative al dataset di riferimento **49**
5.1 Applicazione dei modelli di decomposizione . . . 49
5.1.1 Prophet . . . 50
5.1.2 NeuralProphet . . . 55
5.2 Applicazione dei modelli ibridi . . . 59
5.3 Applicazione dei modelli basati su reti neurali basate su grafi . . . 62
5.3.1 GNN . . . 63
5.3.2 GNN e RNN . . . 65
5.4 Applicazione delle tecniche di decomposizione . . . 67
5.4.1 BiLSTM ed EMD-BiLSTM . . . 67
5.4.2 CEEMDAN-BiLSTM . . . 68
5.5 Valutazione dei risultati . . . 69

6 Un secondo dataset di riferimento **72**
6.1 Descrizione generale del nuovo dataset . . . 72
6.2 Analisi preliminare dei dati . . . 74
6.3 Confronto con il dataset iniziale . . . 79

7 Applicazione di tecniche innovative al secondo dataset di riferimento **80**
7.1 Impostazione della sperimentazione . . . 80
7.2 Applicazione di ARIMA . . . 81
7.3 Applicazione di Prophet . . . 82
7.4 Applicazione di Prophet e LSTM . . . 86
7.5 Considerazioni finali . . . 87

8 Discussione in merito alle analisi svolte **89**
8.1 Sintesi delle sperimentazioni condotte . . . 89
8.2 Analisi delle metodologie più efficaci . . . 89
8.2.1 Integrazione tra modelli . . . 90
8.2.2 Ruolo della disponibilità dei dati nelle prestazioni . . . 90
8.3 Qualità dei dati e costo computazionale . . . 91

Conclusioni **93**

Bibliografia **95**

Sitografia **96**

Ringraziamenti **97**

Elenco delle figure

1.1 Confronto tra gli standard europei e i valori guida dell'OMS. 5
1.2 Conseguenze dell'inquinamento sulla salute 6
1.3 Ambiti di origine degli inquinanti principali 7

2.1 Pagina iniziale di AirQuino Web. 14
2.2 Dashboard sull'andamento della qualità dell'aria. 14
2.3 Informazioni più dettagliate sulle singole stazioni. 14
2.4 Pagina per l'acquisizione dei dati. 15
2.5 Ultime misurazioni relative a maggio 2026 da parte della stazione SMART137. 16
2.6 Percentuali di dati mancanti per ogni metrica e per ogni sensore. 17
2.7 Esempio di distribuzione temporale dei valori mancanti per la metrica NO2. . 18
2.8 Esempio di distribuzione temporale dei valori mancanti per la metrica CO. . 19
2.9 Esempio di distribuzione temporale dei valori mancanti per la metrica PM10. 20
2.10 Matrice di correlazione per la stazione SMART137. 21
2.11 Matrice di correlazione per la stazione SMART58. 21

3.1 Risultati sperimentali per il modello ARIMAX. 28
3.2 Predizione della CO_2 per la stazione SMART58 usando 0_3. 28
3.3 Intervallo di confidenza sulla predizione della CO_2 per la stazione SMART58. 29
3.4 Predizione della CO_2 per la stazione SMART58 sull'intero mese di gennaio 2026. 29
3.5 Predizione del PM10 per la stazione SMART58. 29
3.6 Intervallo di confidenza sulla predizione del PM10 per la stazione SMART58. 30
3.7 Predizione del PM10 per la stazione SMART58 sull'intero mese di ottobre 2025. 30
3.8 Predizione del PM10 per la stazione SMART137. 30
3.9 Intervallo di confidenza sulla predizione del PM10 per la stazione SMART137. 31
3.10 Predizione del PM10 per la stazione SMART137 sull'intero mese di dicembre 2025. 31
3.11 Metriche ottenute dalle diverse configurazioni ARIMAX per la previsione della PM10. 32
3.12 Metriche ottenute dalle diverse configurazioni ARIMAX per la previsione della CO_2. 32
3.13 Confronto tra le diverse configurazioni ARIMAX per la previsione della CO_2. 34
3.14 Confronto tra le diverse configurazioni ARIMAX per la previsione del PM10. 35
3.15 Risultati sperimentali per il modello SARIMA. 36

3.16 Andamento mensile completo per il mese di gennaio 2026 per la stazione SMART58 utilizzando il modello SARIMA. 36
3.17 Previsione sugli ultimi 5 giorni del mese di gennaio 2026 per la stazione SMART58 utilizzando il modello SARIMA. 36
3.18 Intervallo di confidenza per la previsione del mese di gennaio 2026 per la stazione SMART58 utilizzando il modello SARIMA. 37

5.1 Risultati ottenuti da Prophet nella configurazione mensile. 51
5.2 Previsione di Prophet sugli ultimi cinque giorni del mese di test nella configurazione mensile. 51
5.3 Andamento completo del mese di test previsto da Prophet nella configurazione mensile. 52
5.4 Risultati ottenuti da Prophet utilizzando l'intero periodo storico disponibile per l'addestramento. 52
5.5 Previsione degli ultimi cinque giorni del periodo di test ottenuta con Prophet addestrato sull'intera serie temporale disponibile. 52
5.6 Confronto delle prestazioni ottenute dalle configurazioni analizzate mediante grid search per Prophet. 54
5.7 Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART58-PM10 ottenuta dalla migliore configurazione Prophet individuata tramite grid search. 54
5.8 Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART58-CO_2 ottenuta dalla migliore configurazione Prophet individuata tramite grid search. 55
5.9 Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART137-PM10 ottenuta dalla migliore configurazione Prophet individuata tramite grid search. 55
5.10 Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART137-CO_2 ottenuta dalla migliore configurazione Prophet individuata tramite grid search. 55
5.11 Confronto delle prestazioni ottenute da NeuralProphet con e senza interpolazione dei valori mancanti. 57
5.12 Dettaglio della previsione di NeuralProphet sull'intervallo di test. 57
5.13 Dettaglio della previsione di NeuralProphet con il relativo intervallo di confidenza. 57
5.14 Andamento completo del mese di test previsto da NeuralProphet. 58
5.15 Dettaglio della previsione di NeuralProphet sull'intervallo di test usando l'interpolazione. 58
5.16 Dettaglio della previsione di NeuralProphet con il relativo intervallo di confidenza usando l'interpolazione. 58
5.17 Andamento completo del mese di test previsto da NeuralProphet usando l'interpolazione. 59
5.18 Valori ottenuti nelle metriche di valutazione considerate per le configurazioni di NeuralProphet testate. 60
5.19 Previsione degli ultimi cinque giorni del periodo di test per SMART58-PM10 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search. 60
5.20 Previsione degli ultimi cinque giorni del periodo di test per SMART58-CO_2 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search. 60

5.21 Previsione degli ultimi cinque giorni del periodo di test per SMART137-PM10 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search. . . . 61
5.22 Previsione degli ultimi cinque giorni del periodo di test per SMART137-CO_2 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search. . . . 61
5.23 Risultati ottenuti dall'approccio ibrido Prophet-LSTM. . . . 62
5.24 Dettaglio della previsione degli ultimi cinque giorni del periodo di test ottenuta dall'approccio Prophet-LSTM per la combinazione SMART58-PM10. . . . 62
5.25 Dettaglio della previsione degli ultimi cinque giorni del periodo di test con relativo intervallo di confidenza ottenuta dall'approccio Prophet-LSTM per la combinazione SMART58-PM10. . . . 62
5.26 Andamento complessivo del mese di test previsto dall'approccio Prophet-LSTM per la combinazione SMART58-PM10. . . . 63
5.27 Struttura del grafo generato. . . . 64
5.28 Risultati ottenuti dalla GCN considerando la sola componente spaziale. . . . 64
5.29 Dettaglio della previsione ottenuta dalla GCN sull'intervallo di test. . . . 64
5.30 Dettaglio della previsione ottenuta dalla GCN con relativo intervallo di confidenza. . . . 65
5.31 Andamento completo del periodo di test previsto dalla GCN. . . . 65
5.32 Risultati ottenuti dall'approccio combinato GCN-GRU. . . . 66
5.33 Dettaglio della previsione ottenuta dall'approccio GCN-GRU sull'intervallo di test. . . . 66
5.34 Dettaglio della previsione ottenuta dall'approccio GCN-GRU con relativo intervallo di confidenza. . . . 66
5.35 Andamento completo del periodo di test previsto dall'approccio GCN-GRU. . . 67
5.36 Risultati ottenuti dall'approccio EMD-BiLSTM nelle diverse combinazioni centralina-variabile. . . . 68
5.37 Dettaglio della previsione ottenuta dall'approccio EMD-BiLSTM sull'intervallo di test. . . . 68
5.38 Dettaglio della previsione ottenuta dall'approccio EMD-BiLSTM con relativo intervallo di confidenza. . . . 68
5.39 Andamento completo del periodo di test previsto dall'approccio EMD-BiLSTM. 69
5.40 Risultati ottenuti dall'approccio CEEMDAN-BiLSTM. . . . 69
5.41 Previsione ottenuta dall'approccio CEEMDAN-BiLSTM sull'intervallo di test. 70

6.1 Home page del sito dell'ARPAM. . . . 72
6.2 Sezione dei grafici del sito dell'ARPAM. . . . 73
6.3 Giorni e metriche comuni. . . . 74
6.4 Dati mancanti nelle metriche comuni. . . . 75
6.5 Heatmap per la centralina di Ancona. . . . 76
6.6 Heatmap per la centralina di Chiaravalle. . . . 77
6.7 Heatmap per la centralina di Fano. . . . 77
6.8 Heatmap per la centralina di Falconara. . . . 78
6.9 Heatmap per la centralina di San Benedetto. . . . 78

7.1 Grafico delle previsioni settimanali di CO del sensore ad Ancona mediante metodo multistep. . . . 82
7.2 Grafico delle previsioni settimanali di PM10 del sensore ad Ancona mediante metodo multistep. . . . 82

7.3 Grafico delle previsioni settimanali di CO del sensore ad Ancona mediante metodo rolling one-step. 82
7.4 Grafico delle previsioni settimanali di PM10 del sensore ad Ancona mediante metodo rolling one step. 83
7.5 Previsioni di PM10 per la centralina di Ancona ottenute con la configurazione base di Prophet. 84
7.6 Previsioni di PM10 per la centralina di Chiaravalle ottenute con la configurazione base di Prophet. 84
7.7 Risultati della grid search di Prophet. 85
7.8 Previsioni di PM10 per la centralina di Ancona ottenute con Prophet. 85
7.9 Previsioni di PM10 per la centralina di Chiaravalle ottenute con Prophet. . . . 85
7.10 Previsioni di PM10 per la centralina di Fano ottenute con Prophet. 85
7.11 Previsioni di PM10 per la centralina di Falconara ottenute con Prophet. 86
7.12 Previsioni di PM10 per la centralina di San Benedetto ottenute con Prophet. . 86
7.13 Previsioni di PM10 per la centralina di Ancona ottenute con Prophet e LSTM. 87
7.14 Previsioni di PM10 per la centralina di Chiaravalle ottenute con Prophet e LSTM. 87
7.15 Previsioni di PM10 per la centralina di Fano ottenute con Prophet e LSTM. . . 87
7.16 Previsioni di PM10 per la centralina di Falconara ottenute con Prophet e LSTM. 88
7.17 Previsioni di PM10 per la centralina di San Benedetto ottenute con Prophet e LSTM. 88

8.1 Previsione della metrica CO_2 per la centralina SMART58 mediante NeuralProphet. 90
8.2 Previsione della metrica PM10 per la centralina SMART58 mediante NeuralProphet. 91
8.3 Previsione della metrica CO_2 per la centralina SMART137 mediante NeuralProphet. 91
8.4 Previsione della metrica PM10 per la centralina SMART137 mediante NeuralProphet. 91

Elenco delle tabelle

Introduzione

La qualità dell'aria rappresenta un aspetto rilevante per la tutela della salute e per la gestione del territorio. La concentrazione degli inquinanti atmosferici è il risultato dell'interazione di numerosi fattori e può variare nel tempo in funzione delle condizioni ambientali, meteorologiche e delle diverse sorgenti emissive. Per questo motivo, il monitoraggio delle concentrazioni degli inquinanti attraverso reti di centraline costituisce uno strumento fondamentale per descrivere lo stato della qualità dell'aria e individuare eventuali situazioni critiche. Accanto alla semplice misurazione dei valori osservati, assume particolare interesse la possibilità di prevederne l'evoluzione futura, fornendo informazioni che possano essere utilizzate per anticipare condizioni potenzialmente problematiche e supportare le attività di monitoraggio e pianificazione. Nel contesto italiano, tale attività viene svolta anche attraverso sistemi previsionali sviluppati dalle Agenzie Regionali per la Protezione dell'Ambiente, che integrano informazioni provenienti dalle reti di monitoraggio con ulteriori dati ambientali e meteorologici.

La previsione della qualità dell'aria costituisce, tuttavia, un problema complesso. Le concentrazioni degli inquinanti presentano, infatti, dinamiche temporali differenti e possono essere caratterizzate dalla presenza contemporanea di trend, stagionalità, variazioni rapide e fenomeni non facilmente riconducibili alla sola evoluzione storica della variabile. A ciò si aggiungono le caratteristiche dei dati provenienti dalle centraline, che possono presentare valori mancanti, periodi privi di osservazioni, anomalie e differenti livelli di completezza. La qualità e la quantità delle informazioni disponibili possono influenzare direttamente la possibilità di apprendere correttamente le dinamiche della serie e, di conseguenza, le prestazioni dei modelli previsivi.

In questo scenario, l'utilizzo di modelli statistici e di tecniche di apprendimento automatico consente di affrontare la previsione degli inquinanti secondo prospettive differenti. Tra le soluzioni proposte in letteratura rientrano metodi basati sulla decomposizione delle serie temporali, modelli spazio–temporali e approcci ibridi che combinano più tecniche predittive. Poiché le prestazioni possono dipendere significativamente dalle caratteristiche delle serie considerate, il confronto tra tali soluzioni deve tenere conto non soltanto dell'accuratezza delle previsioni, ma anche della disponibilità dei dati, della strategia adottata e del costo computazionale. Su queste considerazioni si inserisce l'attività sperimentale svolta nel presente lavoro, nella quale le diverse metodologie vengono analizzate in relazione a questi aspetti e alle caratteristiche dei dataset considerati.

Nel corso della tesi vengono inizialmente analizzati i dati raccolti da cinque stazioni AirQuino distribuite nel territorio urbano di Pesaro. Dopo una fase di acquisizione, pulizia e organizzazione dei dati, viene effettuata un'analisi esplorativa volta a valutarne la qualità, la

copertura temporale e la presenza di valori mancanti e anomalie. Su questo primo dataset vengono sperimentate differenti metodologie di previsione, partendo dai modelli ARIMAX e SARIMA e proseguendo con approcci più avanzati. Le diverse configurazioni vengono valutate attraverso metriche quantitative e mediante il confronto grafico tra valori osservati e previsti, così da analizzarne sia le prestazioni numeriche sia la capacità di seguire l'effettivo andamento delle serie.

L'analisi viene successivamente estesa a un secondo dataset di riferimento, introdotto per verificare l'influenza della maggiore disponibilità e qualità dei dati sulle prestazioni dei modelli. Il nuovo insieme di dati comprende diverse centraline distribuite sul territorio marchigiano e permette di disporre di serie temporali più estese e strutturate. L'analisi viene condotta inizialmente attraverso ARIMA, utilizzato come modello di riferimento, e successivamente attraverso Prophet e un modello ibrido Prophet–LSTM, con l'obiettivo di valutare il comportamento di tali metodologie in una diversa configurazione sperimentale.

L'attività sperimentale consente di analizzare diversi aspetti del problema, considerando non soltanto la scelta del modello, ma anche la qualità e la quantità dei dati disponibili, la strategia di previsione e il costo computazionale associato alle diverse metodologie. Particolare attenzione viene posta ai modelli che hanno mostrato risultati più interessanti, nonché al rapporto tra accuratezza delle previsioni e risorse necessarie per ottenere tali risultati. L'obiettivo di tale attività non è individuare una metodologia universalmente migliore, ma mettere in evidenza le caratteristiche e i limiti dei diversi approcci e valutare come questi possano variare al cambiare del dataset e dello scenario di previsione.

Questa tesi è strutturata come di seguito specificato:

- *Capitolo 1: la predizione degli inquinanti.* Viene introdotto il problema dell'inquinamento atmosferico, descrivendo i principali inquinanti considerati, le modalità di monitoraggio e il ruolo dei sistemi previsionali nella valutazione della qualità dell'aria.

- *Capitolo 2: un primo dataset di riferimento.* Viene descritto il dataset ottenuto dalle stazioni AirQuino installate nell'area urbana di Pesaro. Sono illustrate le operazioni di preprocessing e l'analisi esplorativa svolta per valutarne qualità, completezza e copertura temporale.

- *Capitolo 3: applicazione di tecniche classiche al dataset di riferimento.* Vengono presentati gli esperimenti condotti mediante ARIMAX e SARIMA, utilizzati come primi approcci previsivi e come riferimento per le successive sperimentazioni.

- *Capitolo 4: analisi di tecniche innovative per la predizione degli inquinanti.* Vengono introdotte le metodologie avanzate considerate nel lavoro, tra cui Prophet, NeuralProphet, LSTM, modelli ibridi, reti neurali grafiche e tecniche di decomposizione delle serie temporali.

- *Capitolo 5: applicazione delle tecniche innovative.* Le metodologie introdotte nel capitolo precedente vengono applicate al primo dataset e confrontate sulla base delle prestazioni ottenute.

- *Capitolo 6: un secondo dataset di riferimento.* Viene introdotto il secondo insieme di dati, proveniente dalle centraline distribuite sul territorio marchigiano, e tale insieme viene analizzato attraverso un confronto con il dataset utilizzato nelle precedenti sperimentazioni.

- *Capitolo 7: applicazione di tecniche innovative al secondo dataset di riferimento.* Vengono applicati ARIMA, Prophet e Prophet–LSTM al secondo dataset, introducendo la strategia rolling one-step e confrontandone il comportamento con l'approccio multistep.

- *Capitolo 8: discussione in merito alle analisi svolte.* Vengono messi in relazione i risultati delle diverse sperimentazioni, analizzando il ruolo della qualità e della quantità dei dati, delle metodologie utilizzate e del costo computazionale.

CAPITOLO 1

La predizione degli inquinanti

L'inquinamento atmosferico è un fenomeno complesso, influenzato da molteplici sorgenti emissive e da numerosi fattori ambientali che ne determinano la diffusione, la trasformazione chimica e l'evoluzione temporale. La comprensione di tali dinamiche richiede non solo l'analisi delle concentrazioni misurate attraverso le reti di monitoraggio, ma anche lo sviluppo di metodologie in grado di interpretare e modellare il comportamento degli inquinanti.

La previsione delle concentrazioni atmosferiche rappresenta infatti un problema multidisciplinare che coinvolge diverse fonti informative, tra cui dati di monitoraggio, condizioni meteorologiche, caratteristiche delle emissioni e modelli matematici o statistici. Negli ultimi anni, alla modellistica atmosferica tradizionale si sono affiancati approcci basati sull'apprendimento automatico, capaci di sfruttare la crescente disponibilità di dati ambientali per individuare relazioni complesse e pattern temporali utili alla previsione.

Nel presente capitolo vengono introdotti il contesto dell'inquinamento atmosferico, i principali inquinanti considerati e l'evoluzione dei sistemi utilizzati per il monitoraggio e la previsione della qualità dell'aria, descrivendo il passaggio dai modelli fisico-chimici tradizionali agli approcci basati su tecniche di machine learning.

1.1 Problema dell'inquinamento

Con il termine *inquinamento atmosferico* si definisce una particolare forma di inquinamento che comprende l'insieme di tutti gli agenti fisici, chimici e biologici che modificano le caratteristiche naturali dell'atmosfera terrestre, causando un effetto dannoso su esseri viventi e ambiente; infatti, esso si inserisce tra le maggiori cause del riscaldamento globale.

Secondo la European Environment Agency (EEA) l'inquinamento atmosferico rappresenta il rischio maggiore per la salute in Europa. Negli ultimi due decenni la qualità dell'aria è migliorata significativamente grazie all'attuale legislazione UE. Nonostante ciò, un gran numero di persone continua a essere esposto a livelli di inquinamento atmosferico che superano i limiti di sicurezza. La situazione è più grave se le concentrazioni riportate vengono valutate rispetto ai limiti delle linee guida dell'Organizzazione Mondiale della Sanità (OMS) sulla qualità dell'aria (AQG).

La Figura 1.1, tratta dal sito dell'EEA, mostra la percentuale di stazioni di monitoraggio che, nel 2024, hanno registrato concentrazioni annuali inferiori agli attuali limiti dell'Unione europea, ai limiti previsti per il 2030 e ai valori guida dell'OMS. Nella figura, l'Organizzazione Mondiale della Sanità è indicata mediante la sigla inglese WHO, acronimo di *World Health Organization*.

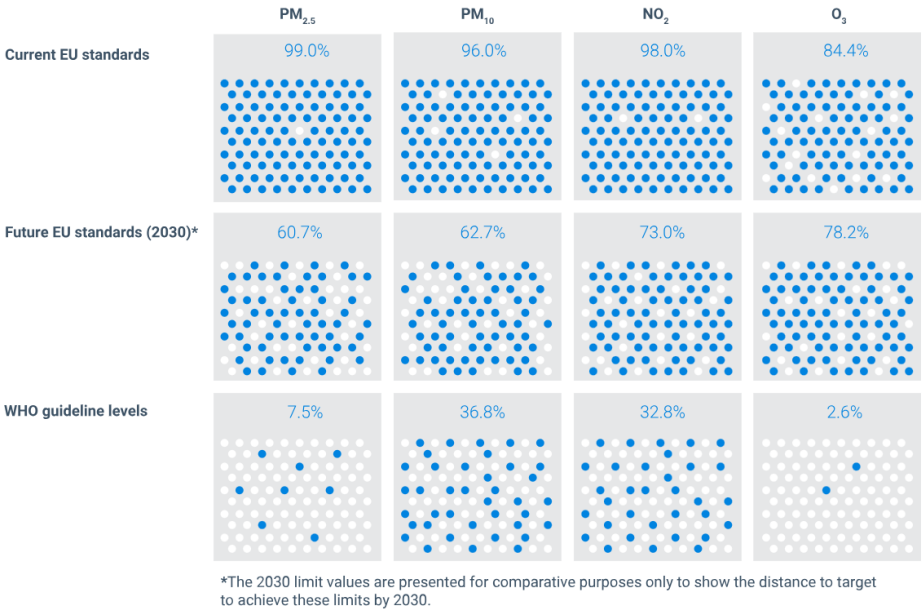


Figura 1.1: Confronto tra gli standard europei e i valori guida dell'OMS.

1.1.1 Le cause dell'inquinamento atmosferico

Secondo il *Programma delle Nazioni Unite per l'Ambiente* (UNEP) le cause concorrenti all'alimentazione dell'inquinamento atmosferico possono essere di origine antropica, come:

- produzione di energia;
- trasporti;
- attività industriali;
- agricoltura;
- incenerimento e gestione dei rifiuti;
- consumo domestico di energia per cucinare e riscaldare gli edifici.

Le cause, però, possono anche essere di orgine naturale, derivanti da fenomeni ambientali come:

- eruzioni vulcaniche;
- incendi boschivi;
- tempeste di sabbia.

Le cause di natura umana elencate rivestono un ruolo fondamentale sia per l'economia mondiale sia per la vita quotidiana. Tuttavia, il loro sviluppo si è a lungo basato sull'impiego di combustibili fossili e su sistemi produttivi poco attenti alle conseguenze ambientali. Sebbene oggi la transizione verso modelli più sostenibili rappresenti uno degli obiettivi principali di molti governi e settori industriali, una parte significativa dei problemi ambientali deriva da processi di industrializzazione, urbanizzazione e crescita economica avvenuti nei decenni precedenti.

In questo contesto, nel 2016 è entrato in vigore l'Accordo di Parigi, un trattato internazionale giuridicamente vincolante in materia di cambiamenti climatici. Sebbene l'accordo riguardi principalmente la riduzione delle emissioni di gas serra, molte delle misure previste, come la riduzione dell'uso dei combustibili fossili e la promozione delle fonti energetiche rinnovabili, possono contribuire anche al miglioramento della qualità dell'aria.

Quasi tutti i Paesi partecipano all'Accordo di Parigi e sono tenuti a presentare periodicamente i propri impegni climatici, denominati *Nationally Determined Contributions* (NDC).

L'Accordo riconosce il principio delle responsabilità comuni ma differenziate. Tutti i Paesi sono chiamati a contribuire al contrasto ai cambiamenti climatici, ma quelli sviluppati, disponendo di maggiori risorse economiche e avendo una maggiore responsabilità storica nella produzione delle emissioni, sono invitati a guidare la transizione e a sostenere finanziariamente e tecnologicamente i Paesi in via di sviluppo. Questi ultimi, infatti, possono incontrare maggiori difficoltà nel sostenere i costi necessari per l'adozione di politiche ambientali e per la trasformazione dei propri sistemi energetici e produttivi.

1.1.2 Conseguenze sulla salute

Le ripercussioni che l'inquinamento atmosferico ha sulla salute sono svariate; basti pensare che secondo le stime dell'EEA se si fossero seguite le normative dell'OMS nel 2022 si sarebbero potuti evitare:

- 239.000 decessi attribuibili al PM2.5;
- 70.000 decessi attribuibili all'ozono;
- 48.000 decessi attribuibili al biossido di azoto.

Alcuni dei sintomi derivanti da un eccessiva presenza di inquinanti nell'aria sono riportati in Figura 1.2

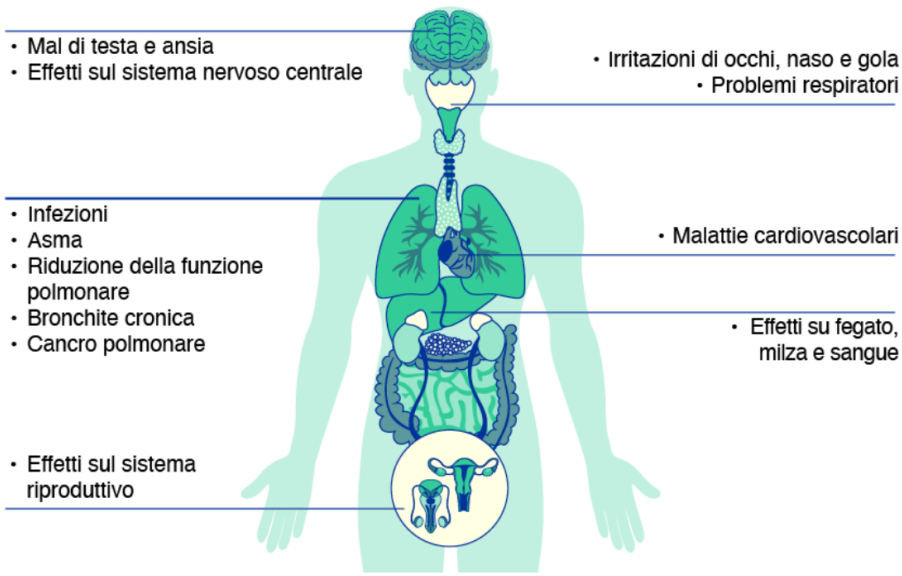


Figura 1.2: Conseguenze dell'inquinamento sulla salute

1.1.3 Importanza del monitoraggio

Data la varietà delle sorgenti e dei fattori che influenzano la concentrazione degli inquinanti, il monitoraggio rappresenta uno strumento essenziale per descrivere in modo oggettivo lo stato della qualità dell'aria. Le misurazioni consentono di verificare il rispetto dei valori limite, di individuare situazioni critiche e di analizzare l'andamento delle concentrazioni nel corso del tempo. In questo modo è possibile ottenere una conoscenza più accurata dell'esposizione della popolazione e valutare gli effetti delle politiche di riduzione delle emissioni.

1.2 I principali inquinanti atmosferici

Gli inquinanti atmosferici non sono tutti uguali, ma si distinguono in due categorie che si basano sulle modalità di emissione. Queste sono:

- *Inquinanti primari*: vengono emessi dalla sorgente senza subire trasformazioni in atmosfera.
- *Inquinanti secondari*: si formano in atmosfera a partire da sostanze già presenti, in seguito a reazioni chimico-fisiche innescate dalla luce solare o dall'interazione tra diverse molecole.

In Figura 1.3 è possibile osservare i principali ambiti da cui si originano alcuni degli inquinanti che verranno analizzati

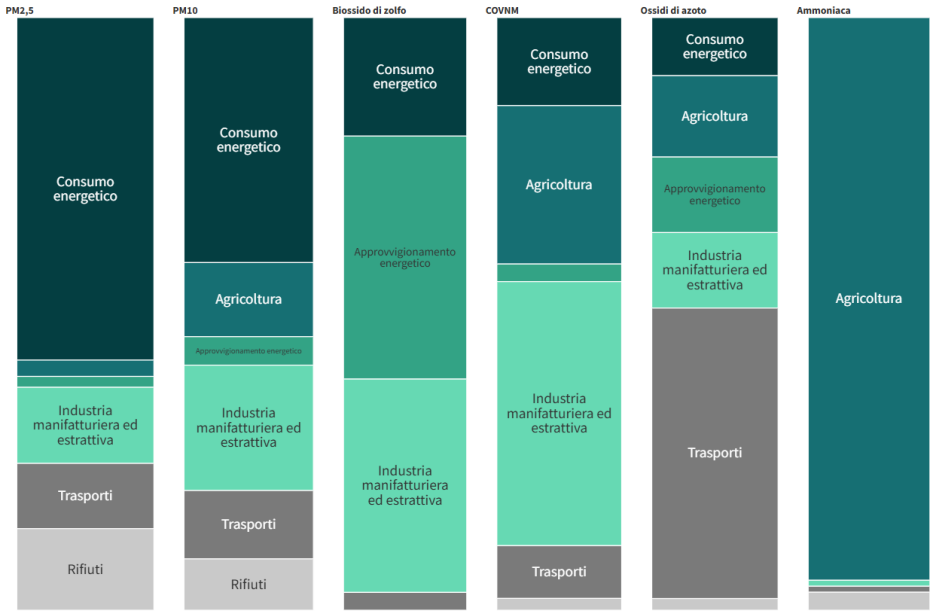


Figura 1.3: Ambiti di origine degli inquinanti principali

1.2.1 Inquinanti primari

Monossido di carbonio - CO

Il monossido di carbonio è un gas incolore e inodore, risultato di processi di combustione che, se inalato e presente in concentrazione sufficiente, può causare gravi problemi di salute.

Il CO presente nell'aria proviene dal fumo di tabacco o da fonti di combustione non dotate di idonea aspirazione, come caldaie, scaldabagni, caminetti e stufe.

Il monossido di carbonio inalato si lega con l'emoglobina formando una carbossiemoglobina (COHb). Le conseguenze sono un'impossibilità per l'ossigeno di legarsi all'emoglobina stessa, in quanto satura; ciò porta ad una mancata ossigenazione del sangue. In concentrazioni maggiori si possono verificare cefalea, confusione, disorientamento, visione alterata e nausea.

Trattandosi di un inquinante legato ad attività antropiche, alcuni fattori che ne aumentano la concentrazione possono essere: la mancanza di ricambio d'aria, un uso improprio di fuochi, gas di scarico delle macchine, fumo e incendi.

Biossido di zolfo - SO_2

Il biossido di zolfo è un gas incolore, non infiammabile e caratterizzato da elevata solubilità in acqua. Le principali emissioni di SO_2 derivano dalla combustione di combustibili contenenti zolfo, in particolare carbone e petrolio, utilizzati nei processi industriali, nella produzione di energia e nel riscaldamento domestico. Negli ultimi anni le emissioni sono diminuite significativamente grazie al miglioramento della qualità dei combustibili e all'introduzione di sistemi di controllo delle emissioni.

L'esposizione al biossido di zolfo può provocare danni all'apparato respiratorio, favorendo irritazioni delle vie aeree, bronchiti e un peggioramento delle patologie respiratorie preesistenti. Dal punto di vista ambientale, il biossido di zolfo contribuisce alla formazione delle piogge acide, con conseguenze sulla vegetazione, sugli ecosistemi acquatici e sui materiali.

Le concentrazioni di biossido di zolfo possono aumentare in presenza di elevate emissioni da attività industriali e processi di combustione, soprattutto in condizioni meteorologiche che limitano la dispersione degli inquinanti, come assenza di vento e inversioni termiche.

Ossidi di azoto - NOx

Gli ossidi di azoto sono un gruppo di gas inquinanti composti da azoto e ossigeno, generati in tutti i processi di combustione che usano l'aria come comburente. Tra gli ossidi più significativi per l'inquinamento si annoverano il monossido di azoto (NO) e il biossido di azoto (NO_2), unico ad avere rilevanza tossicologica. Il biossido di azoto (NO_2) è un gas di colore rosso-bruno, caratterizzato da un odore forte e pungente e da un elevato potere irritante. Si forma principalmente per ossidazione del monossido di azoto (NO), estremamente reattivo.

Gli ossidi di azoto sono da ritenersi fra gli inquinanti atmosferici maggiormente pericolosi perchè, oltre ad essere tossici, svolgono un ruolo determinante nella formazione dello "smog fotochimico": sono infatti importanti precursori dell'ozono troposferico in estate e del PM10 in inverno. La principale fonte degli ossidi è il traffico veicolare e l'entità delle emissioni varia in funzione del combustibile, dello stato del motore e delle modalità di utilizzo dello stesso. Altre fonti sono gli impianti di riscaldamento, le centrali per la produzione di energia, ma anche fonti naturali, come vulcani, incendi e fenomeni temporaleschi.

Si può dedurre che la concentrazione degli ossidi di azoto (NOx), e in particolare di NO_2, aumenta quando crescono le emissioni prodotte dai processi di combustione e quando le condizioni atmosferiche ne limitano la dispersione.

Composti organici volatili - COV

I composti organici volatili sono sostanze chimiche a base di carbonio che passano facilmente allo stato gassoso nell'aria anche a temperatura ambiente; questa caratteristica li

rende facilmente dispersibili nell'aria contribuendo all'inquinamento atmosferico. Le fonti che danno origine a queste sostanze includono prodotti per la pulizia, vernici e processi industriali.

I COV possono essere causa di una vasta gamma di effetti che vanno dall'irritazione di occhi, naso e gola fino a lesioni al sistema nervoso centrale, problemi al fegato e ai reni. Sono anche precursori dell'ozono troposferico, reagendo con la luce solare e con gli ossidi di azoto

Particolato primario - PM10 e PM2.5

Per particolato intendiamo un'insieme di piccole particelle, solide o liquide, emesse direttamente in atmosfera da fonti specifiche, come la combustione, l'erosione del suolo, l'usura dei freni e dell'asfalto. Le particelle che costituiscono il particolato possono avere dimensioni diverse e, proprio sulla base di ciò, si differenziano in PM10 e PM2.5.

Il particolato è l'inquinante atmosferico che provoca maggiori danni alla salute umana in Europa. Alcune di queste particelle sono così piccole da poter penetrare in profondità nei polmoni e raggiungere gli alveoli (PM2.5) mentre quelle più grossolane (PM10) si depositano nelle vie aeree superiori. Alcuni dei danni includono malattie respiratorie (come asma, bronchite cronica o riduzione della funzione polmonare), problemi cardiovascolari (aumento del rischio di infarto, ictus e alterazioni della pressione sanguigna) e danni sistemici (stress ossidativo, infiammazioni diffuse e maggior rischio di tumori). Oltre ai danni provocati alla salute si aggiungono anche danni alla vegetazione, dati dalla deposizione delle particelle sulle foglie, al clima, alla visibilità e, perfino, al patrimonio.

Fattori che ne possono aumentare la concentrazione sono: il riscaldamento, il traffico veicolare o i processi produttivi e di combustione delle industrie.

1.2.2 Inquinanti secondari

Ozono - O_3

L'ozono è un gas instabile e fortemente ossidante che agisce come inquinante secondario nocivo negli strati bassi dell'atmosfera, mentre nella stratosfera forma uno scudo protettivo contro i raggi ultravioletti. Tale inquinante ha origine dalla reazione tra precursori antropici e naturali con irraggiamento solare.

I precursori di origine antropica sono gli ossidi di azoto e sostanze organiche volatili emesse da attività umane, mentre quelle naturali sono composti organici volatili biogenici prodotti dalle piante. L'ozono ha origine principalmente in contesti fortemente antropizzati, con elevata emissione di precursori, soprattutto durante il periodo estivo, caratterizzato da alto irraggiamento solare, elevate temperature e condizioni meteorologiche stagnanti.

Un'esposizione cronica all'ozono comporta cambiamenti significativi nelle vie aeree al livello dei bronchioli limitandone la funzionalità e favorendo lo sviluppo di malattie come l'asma. La reversibilità di questo tipo di lesioni è un punto che deve essere ancora chiarito. L'assenza di studi dedicati non ci consente di legare l'esposizione prolungata all'ozono con effetti cronici sulla salute

Particolato secondario

Il particolato secondario è un tipo di inquinante atmosferico costituito da piccole particelle solide o liquide che non vengono emesse direttamente, ma si formano in aria attraverso reazioni chimiche e fisiche tra gas inquinanti detti precursori. I precursori principali sono gli ossidi di azoto (NOx), il biossido di zolfo (SO_2), l'ammoniaca (NH_3) e i composti organici volatili (COV).

Tra le fonti che determinano la formazione del particolato secondario troviamo quelle elencate per il particolato primario a cui si affianca il settore dell'agricoltura e dell'allevamento a causa del rilascio di pesticidi e fertilizzanti.

Acidi secondari

Gli acidi secondari sono composti che si formano in atmosfera attraverso reazioni chimiche che coinvolgono gli inquinanti primari. I principali sono l'acido solforico (H_2SO_4), derivante dall'ossidazione del biossido di zolfo (SO_2), e l'acido nitrico (HNO_3), prodotto dall'ossidazione degli ossidi di azoto (NOx).

La formazione di questi composti è favorita dalla presenza di radiazione solare, ossidanti atmosferici, come ozono e perossido di idrogeno, e dalla presenza di acqua nelle nubi o nella nebbia. Gli acidi possono successivamente contribuire alla formazione di particolato secondario e alla deposizione acida, fenomeno responsabile dell'acidificazione dei suoli e delle acque superficiali.

Dal punto di vista sanitario, gli acidi secondari contribuiscono alla formazione di particolato fine secondario che può aumentare il rischio di danni respiratori e cardiovascolari.

1.3 Dal monitoraggio alla previsione

1.3.1 Dalla misura alla modellazione

La valutazione della qualità dell'aria si è sviluppata parallelamente alla crescente consapevolezza degli effetti dell'inquinamento atmosferico sulla salute umana e sull'ambiente.

In Italia, le prime misure sistematiche della qualità dell'aria risalgono agli anni Cinquanta e Sessanta e riguardavano principalmente il biossido di zolfo (SO_2), un inquinante fortemente associato ai processi di combustione industriale e alla produzione energetica.

In questa prima fase gli strumenti di valutazione erano costituiti principalmente da stazioni di misura puntuali, localizzate soprattutto nei grandi centri urbani e nelle aree caratterizzate da una maggiore presenza industriale. L'obiettivo era quello di quantificare la presenza degli inquinanti tradizionali prodotti direttamente dalle sorgenti emissive.

1.3.2 Verso la previsione

Durante gli anni Ottanta furono introdotti criteri e metodi standardizzati per la misura dei principali inquinanti atmosferici, permettendo di rendere più confrontabili i dati raccolti dalle diverse stazioni di monitoraggio. In questa fase l'obiettivo principale rimaneva la descrizione dello stato della qualità dell'aria attraverso misure effettuate direttamente sul territorio, ma iniziò a emergere la necessità di interpretare i dati considerando anche i processi che regolano la dispersione e la trasformazione degli inquinanti.

La sola misura puntuale, infatti, presenta alcuni limiti: una centralina fornisce informazioni accurate relative alla zona in cui è installata, ma non permette di descrivere completamente la distribuzione degli inquinanti su aree più estese. Per superare questo limite, negli anni successivi si è sviluppata l'integrazione tra dati osservativi, informazioni sulle emissioni, condizioni meteorologiche e modelli matematici in grado di simulare il comportamento degli inquinanti in atmosfera.

I primi modelli utilizzati erano principalmente modelli deterministici basati su equazioni fisiche che descrivevano il trasporto e la dispersione degli inquinanti. Questi modelli utilizzavano informazioni relative alle sorgenti emissive, alla velocità e direzione del vento, alla turbolenza atmosferica e alle condizioni meteorologiche per stimare la distribuzione degli inquinanti nello spazio e nel tempo. Negli anni Novanta questa esigenza divenne

ancora più evidente con l'affermarsi dello studio degli inquinanti secondari, come ozono (O_3), biossido di azoto (NO_2) e particolato fine (PM10 e PM2.5), la cui concentrazione non dipende esclusivamente dalle emissioni dirette, ma anche da complessi processi chimico-fisici che avvengono in atmosfera. Questo portò allo sviluppo di modelli atmosferici sempre più complessi, capaci di integrare processi meteorologici, chimici ed emissivi.

Nel contesto italiano, un esempio importante è rappresentato dal sistema modellistico MINNI (Modello Integrato Nazionale a supporto della Negoziazione Internazionale sui temi dell'inquinamento atmosferico), sviluppato da ENEA per supportare la valutazione della qualità dell'aria e le politiche ambientali.

1.3.3 Modello attuale

I sistemi previsionali attualmente utilizzati rappresentano l'evoluzione dei modelli atmosferici sviluppati negli anni precedenti. Oggi la previsione della qualità dell'aria viene ottenuta attraverso sistemi integrati che combinano modelli numerici, dati osservativi provenienti dalle reti di monitoraggio e tecniche di assimilazione dei dati, con l'obiettivo di migliorare l'accuratezza delle stime prodotte.

Parallelamente, la crescente disponibilità di serie storiche ambientali e l'aumento della capacità computazionale hanno favorito lo sviluppo di approcci basati sull'Intelligenza Artificiale e sul machine learning. Questi modelli, a differenza dei modelli fisico-chimici tradizionali, non descrivono direttamente i processi atmosferici, ma apprendono dai dati le relazioni esistenti tra concentrazioni degli inquinanti e variabili ambientali.

Negli ultimi anni tali approcci sono stati sempre più studiati e applicati sia come strumenti complementari ai modelli tradizionali, ad esempio per correggerne gli errori sistematici o migliorarne la risoluzione spaziale e temporale, sia come modelli autonomi basati esclusivamente sull'apprendimento dai dati storici.

Tra i modelli maggiormente utilizzati in questo ambito rientrano algoritmi di machine learning, come Random Forest e Gradient Boosting, e modelli di deep learning, tra cui reti neurali ricorrenti come le Long Short-Term Memory (LSTM), particolarmente adatte alla modellazione di serie temporali ambientali.

Nel contesto italiano, la valutazione della qualità dell'aria è affidata alle reti di monitoraggio gestite dalle Agenzie Regionali per la Protezione dell'Ambiente (ARPA), attraverso catene modellistiche che integrano modelli chimico-trasportistici, dati meteorologici, inventari emissivi e osservazioni delle centraline. A livello nazionale sono, inoltre, disponibili sistemi modellistici, come MINNI e kAIROS

Le reti sono costituite principalmente da centraline fisse installate in punti rappresentativi del territorio, utilizzate per il monitoraggio continuo delle concentrazioni dei principali inquinanti atmosferici. A queste si affiancano campagne di monitoraggio tramite stazioni mobili, utilizzate principalmente in situazioni specifiche, come eventi accidentali o studi temporanei finalizzati alla caratterizzazione di particolari sorgenti emissive.

I dati raccolti dalle reti regionali vengono utilizzati sia per la valutazione dello stato della qualità dell'aria sia per la verifica del rispetto dei limiti normativi stabiliti dalla legislazione europea. Le informazioni elaborate vengono successivamente trasmesse agli enti competenti e comunicate alla Commissione Europea nell'ambito delle procedure previste dalla normativa comunitaria. Parallelamente, i sistemi previsionali regionali permettono di stimare l'evoluzione delle concentrazioni degli inquinanti nei giorni successivi, generalmente fino a un orizzonte temporale di 5 giorni, fornendo informazioni utili alla gestione delle eventuali situazioni critiche.

1.4 Importanza della previsione

Come precedentemente descritto, i sistemi previsionali della qualità dell'aria permettono di stimare con alcuni giorni di anticipo l'evoluzione delle concentrazioni degli inquinanti atmosferici. In Italia, le ARPA elaborano quotidianamente tali previsioni al fine di fornire informazioni utili sia alle amministrazioni locali sia alla popolazione.

Il principale valore della previsione risiede nella possibilità di individuare anticipatamente situazioni potenzialmente critiche e consentire l'adozione di interventi preventivi. Una previsione corretta permette, infatti, alle autorità competenti di attuare misure prima che si verifichino episodi di elevato inquinamento, riducendo l'esposizione della popolazione e limitando gli effetti negativi sulla salute pubblica.

Le informazioni prodotte dai sistemi previsionali costituiscono, quindi, uno strumento di supporto alle decisioni. Le strategie di contrasto all'inquinamento atmosferico sono definite all'interno dei piani regionali di risanamento della qualità dell'aria e coinvolgono diversi settori responsabili delle emissioni, poiché sono richiesti interventi coordinati su più sorgenti emissive.

In ambito urbano, particolare rilevanza assumono le emissioni derivanti dal settore dei trasporti e dal riscaldamento degli edifici, soprattutto durante i mesi invernali, quando condizioni meteorologiche favorevoli all'accumulo degli inquinanti determinano solitamente episodi di superamento dei limiti normativi. Per tale motivo, le amministrazioni possono adottare differenti tipologie di intervento, distinguibili principalmente in misure strutturali e misure emergenziali.

Le misure strutturali sono interventi di medio e lungo periodo finalizzati a migliorare stabilmente la qualità dell'aria, come la riduzione delle emissioni industriali, la promozione della mobilità sostenibile, il miglioramento dell'efficienza energetica degli edifici e la regolamentazione degli impianti di riscaldamento. Le misure emergenziali, invece, vengono applicate in presenza di condizioni critiche, o previste come tali, e comprendono azioni temporanee quali limitazioni alla circolazione dei veicoli, provvedimento delle targhe alterne, restrizioni sull'utilizzo degli impianti termici o altre limitazioni alle attività maggiormente emissive.

Un ulteriore esempio di intervento è rappresentato dalle cosiddette "domeniche ecologiche", che non hanno come obiettivo principale una riduzione significativa e immediata delle concentrazioni degli inquinanti, ma rappresentano soprattutto uno strumento di sensibilizzazione della popolazione riguardo alle cause e alle conseguenze dell'inquinamento atmosferico, promuovendo comportamenti individuali maggiormente sostenibili.

CAPITOLO 2

Un primo dataset di riferimento

Il presente capitolo avrà come focus il dataset utilizzato come riferimento per lo sviluppo dei modelli predittivi, che saranno analizzati nel corso della presente tesi. Verranno inizialmente illustrate la provenienza e le caratteristiche principali dei dati disponibili, per poi descrivere le attività di Exploratory Data Analysis (EDA) svolte al fine di valutarne la qualità e individuare eventuali criticità.

2.1 Descrizione del dataset

Il dataset analizzato proviene da 5 stazioni di monitoraggio AirQuino distribuite nel territorio urbano della città di Pesaro. Tali dati rappresentano la base informativa utilizzata nell'ambito della piattaforma FUTURA, il cui obiettivo è l'analisi dei dati ambientali e lo sviluppo di strumenti previsionali finalizzati alla valutazione dell'evoluzione della qualità dell'aria.

AirQuino è una piattaforma tecnologica sviluppata dal Consiglio Nazionale delle Ricerche (CNR) per il monitoraggio della qualità dell'aria in ambito urbano. Il sistema è composto da stazioni di rilevamento, fisse o mobili, in configurazione indoor o outdoor, in grado di acquisire informazioni di carattere metereologico o relative ad agenti inquinanti. I dati raccolti vengono trasmessi, tramite moduli di connessione integrati, a un server cloud e successivamente resi disponibili attraverso la piattaforma web AirQuino, la cui pagina iniziale è riportata in Figura 2.1.

Attraverso l'area amministrativa della piattaforma è stato possibile accedere al progetto relativo alla città di Pesaro, consultare le informazioni associate alle stazioni coinvolte e visualizzare una dashboard riassuntiva dell'andamento della qualità dell'aria, riportata in Figura 2.2. La piattaforma consente, inoltre, di accedere ad informazioni dettagliate relative alle singole stazioni di monitoraggio, come mostrato in Figura 2.3.

Per la costruzione del dataset utilizzato nelle successive fasi di analisi, sono stati esportati i dati storici relativi alle cinque stazioni considerate. L'estrazione è stata effettuata mediante l'apposita sezione della piattaforma, che permette di selezionare il progetto, la stazione e l'intervallo temporale di interesse, come riportato in Figura 2.4.

Durante la fase di acquisizione dei dati è emersa una limitazione della piattaforma relativa all'intervallo temporale massimo selezionabile per ciascuna esportazione, fissato a 31 giorni. Per questo motivo, i dati disponibili sono stati acquisiti separatamente per ciascun mese, a partire da luglio 2025, primo mese per il quale sono risultate disponibili informazioni.

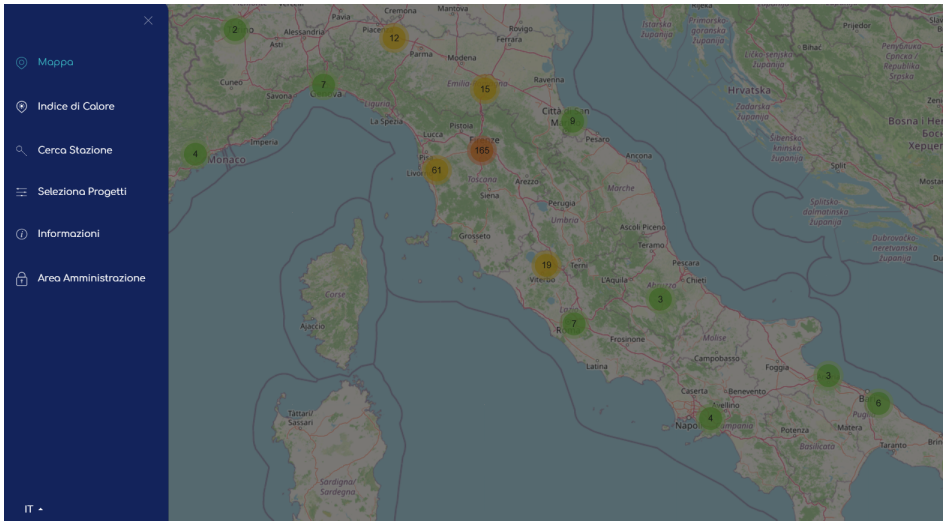


Figura 2.1: Pagina iniziale di AirQuino Web.

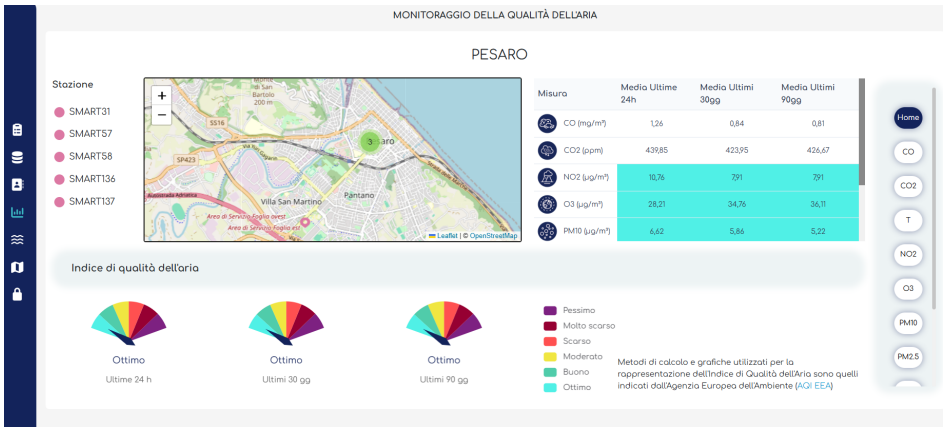


Figura 2.2: Dashboard sull'andamento della qualità dell'aria.

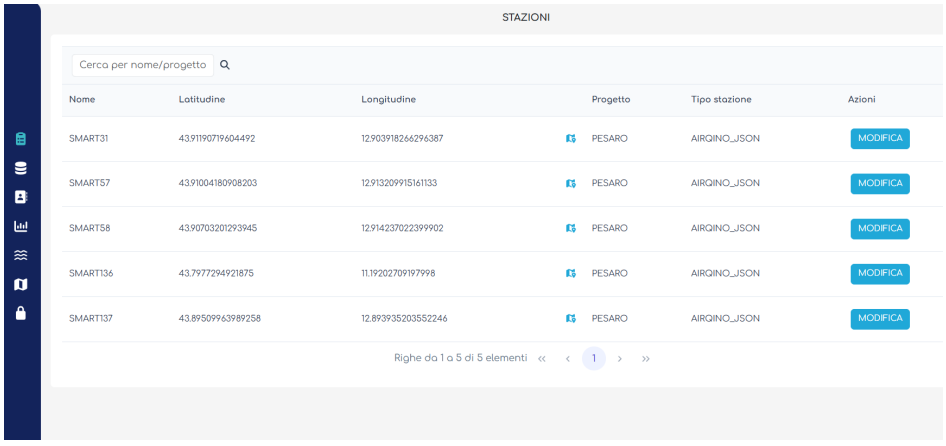


Figura 2.3: Informazioni più dettagliate sulle singole stazioni.

Inizialmente il periodo temporale considerato si estendeva da luglio 2025 a gennaio 2026, ma successivamente è stato esteso fino a giugno 2026, per un totale di 12 mesi di osservazioni.

I dati raccolti presentano informazioni relative a inquinanti atmosferici, parametri ambientali e del dispositivo. In particolare, essi sono i seguenti:

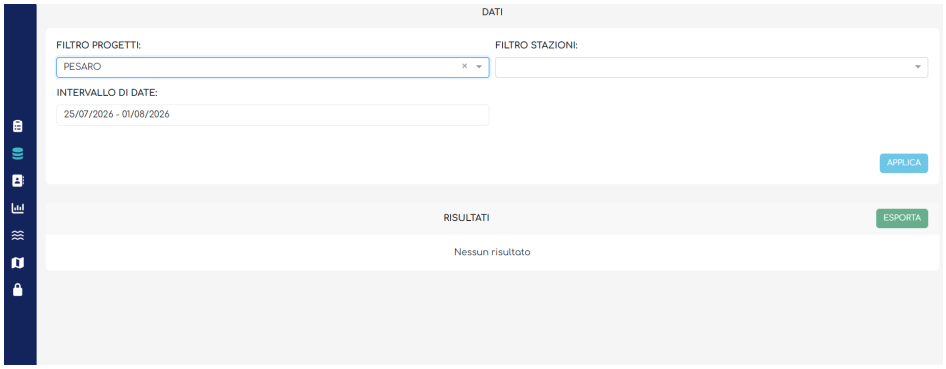


Figura 2.4: Pagina per l'acquisizione dei dati.

- *Data Acquired*: data e ora di acquisizione della misura;
- *CO2*: anidride carbonica (ppm);
- *T*: temperatura dell'aria (°C);
- *RH*: umidità relativa (%);
- *T. int.*: temperatura interna del sensore (°C);
- *O3*: ozono troposferico ($\mu g/m^3$);
- *NO2*: biossido di azoto ($\mu g/m^3$);
- *CO*: monossido di carbonio (mg/m^3);
- *PM2.5*: particolato fine con diametro inferiore a 2.5 μm ($\mu g/m^3$);
- *PM10*: particolato fine con diametro inferiore a 10 μm ($\mu g/m^3$).

2.2 Analisi esplorativa

I file CSV ottenuti dalla piattaforma AirQuino non presentavano una struttura direttamente adatta alle successive analisi. È stato, quindi, necessario effettuare una fase preliminare di preprocessing finalizzata alla pulizia, all'organizzazione e all'uniformazione dei dati raccolti.

2.2.1 Preprocessing e organizzazione

I file relativi ai dati raccolti durante ciascun mese sono stati generati mediante l'aggregazione di più blocchi di informazioni (chunk) esportati dalla piattaforma. Questa modalità di generazione ha introdotto alcune problematiche legate alla presenza di intestazioni ripetute all'interno dello stesso file e alla possibile variazione dell'ordine con cui venivano riportate le metriche.

Per questo motivo è stata effettuata una fase preliminare di pulizia dei dati, finalizzata all'individuazione e alla rimozione degli header duplicati, al riordinamento delle colonne e all'eliminazione delle osservazioni duplicate. Tali operazioni hanno permesso di ottenere una struttura uniforme e coerente tra i diversi file.

Successivamente, i dati relativi ai diversi mesi sono stati organizzati sulla base della stazione di appartenenza. Al termine di questa fase sono stati ottenuti cinque dataset distinti, uno per ciascuna centralina: SMART31, SMART57, SMART58, SMART136 e SMART137.

Un'ulteriore caratteristica riscontrata nella struttura dei dati mensili riguarda la presenza delle prime acquisizioni temporali appartenenti al mese successivo. In particolare, i file relativi a un determinato mese possono includere una o due misurazioni aggiuntive corrispondenti alle prime ore del mese seguente. Ad esempio, in Figura 2.5 è riportato il file relativo alla stazione SMART137 per il mese di maggio 2026, nel quale è possibile osservare come le ultime due righe corrispondano alle misurazioni effettuate alle ore 00:00 e alle ore 01:00 del giorno successivo. Tali valori non risultano successivamente presenti nel file del mese di giugno.

Questa caratteristica deve essere considerata durante le analisi basate sulla suddivisione mensile dei dati, poiché può portare alla generazione di grafici relativi a mesi per i quali è disponibile soltanto una singola osservazione. Per evitare tale problematica nella fase di visualizzazione degli andamenti mensili delle metriche, sono stati esclusi dalla rappresentazione grafica i mesi caratterizzati da meno di tre acquisizioni disponibili.

```
737  2026-05-31T17:00, 406.33334,30.191666,44.633335,24.61703,13.224815,0.6875467,1.0,4.4166665
738  2026-05-31T18:00, 406.58334,29.4,44.983334,24.615086,12.779911,0.6812758,1.75,4.3333335
739  2026-05-31T19:00, 408.75,28.325,46.566666,24.11558,12.764448,0.68851995,3.4166667,6.416666
740  2026-05-31T20:00, 416.0,27.075,50.433334,23.060217,13.901771,0.7067892,2.6666667,7.0
741  2026-05-31T21:00, 444.66666,25.341667,57.00833,13.965857,24.09852,0.7864558,4.0,9.0
742  2026-05-31T22:00, 421.5,25.983335,50.508335,22.015366,15.079185,0.6935808,2.25,5.8333335
743  2026-05-31T23:00, 416.91666,25.733335,49.466667,23.307579,13.026213,0.67358327,2.0,4.25
744  2026-06-01T00:00, 427.25,24.15,53.44167,22.776873,13.467379,0.6963367,2.1666667,4.75
745  2026-06-01T01:00, 423.66666,23.733334,51.408333,23.132858,12.640459,0.66834915,1.75,4.5
746
```

Figura 2.5: Ultime misurazioni relative a maggio 2026 da parte della stazione SMART137.

2.2.2 Analisi esplorativa

Dopo aver creato i 5 file con i dati puliti si è passati alla fase di Exploratory Data Analysis vera e propria, partendo da uno studio dei valori mancanti nel dataset. La presenza di valori mancanti rappresenta un aspetto rilevante nell'analisi di dati provenienti da reti di monitoraggio ambientale, poiché eventuali interruzioni nelle acquisizioni possono compromettere sia la continuità delle serie storiche sia la successiva fase di modellazione predittiva.

L'analisi è stata effettuata considerando separatamente ciascuna stazione di monitoraggio e ciascuna variabile acquisita. Per ogni combinazione sensore-metrica sono stati calcolati il numero di osservazioni mancanti e la relativa percentuale rispetto al totale dei campioni disponibili. Il risultato è una tabella contenente le percentuali di dati mancanti per ogni metrica e per ogni sensore, parzialmente riportata in Figura 2.6.

Una prima criticità emersa durante la fase di acquisizione dei dati riguarda i mesi di luglio, agosto e settembre 2025. Sebbene luglio 2025 rappresenti il primo mese per il quale risultano disponibili acquisizioni, la copertura temporale delle osservazioni risulta estremamente limitata. In particolare, le misure disponibili riguardano soltanto alcune ore di alcuni giorni del mese, tanto che alcuni file relativi alle stazioni considerate contengono appena sette osservazioni complessive. Nessuno dei sensori disponibili presenta misurazioni per il mese di agosto 2025 e, nel caso dello SMART 136, nemmeno per settembre 2025.

2.2.3 Analisi della copertura temporale

Oltre alla valutazione quantitativa dei valori mancanti, è stata effettuata un'analisi della distribuzione temporale delle osservazioni disponibili, con l'obiettivo di individuare eventuali periodi caratterizzati da assenze prolungate di dati.

Per ogni combinazione mese-metrica è stato, quindi, generato un grafico rappresentativo dell'andamento giornaliero della variabile considerata. La presenza di intervalli temporali

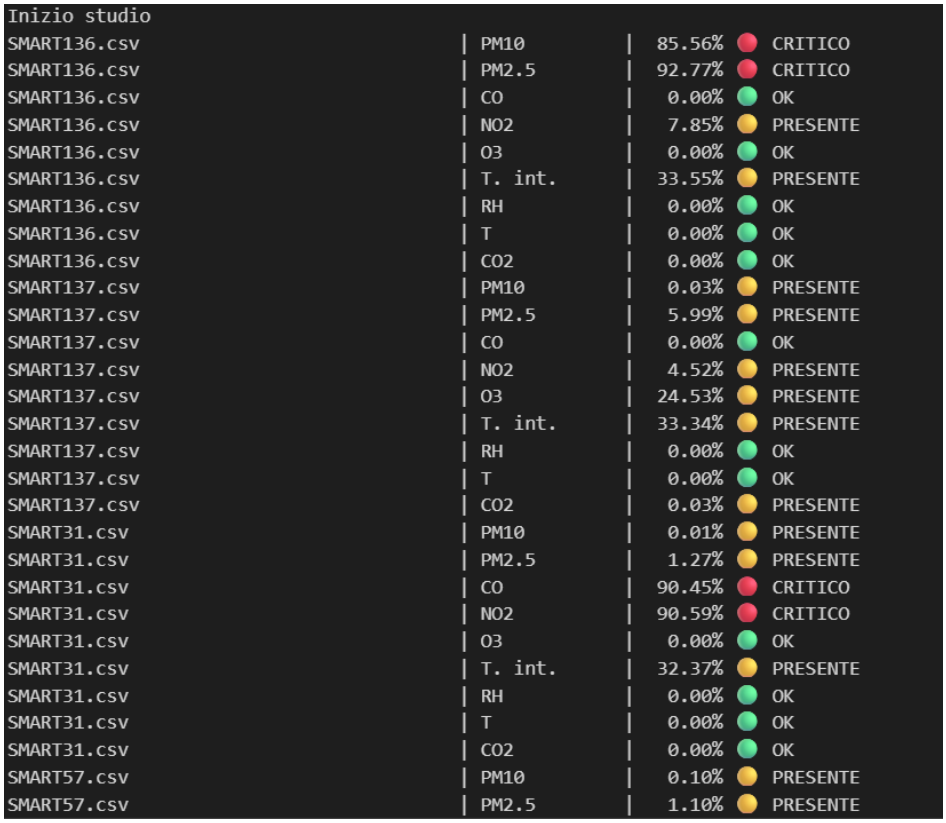


Figura 2.6: Percentuali di dati mancanti per ogni metrica e per ogni sensore.

privi di valori risulta immediatamente individuabile dall'assenza di punti relativi alle giornate corrispondenti.

Considerato l'elevato numero di rappresentazioni generate, di seguito vengono riportati solo alcuni esempi relativi alla valutazione di differenti metriche valutate nello stesso mese (dicembre 2025); in particolare, in Figura 2.7 sono riportati i valori per la NO2, in Figura 2.8 sono riportati i valori per la CO e in Figura 2.9 sono riportati i valori per la PM10.

Gli esempi riportati sono stati selezionati in quanto rappresentativi di una tendenza ricorrente osservata durante l'analisi: per diverse metriche, la maggior parte delle stazioni presenta una disponibilità temporale limitata, caratterizzata dalla presenza di numerosi intervalli privi di acquisizioni. Solo le stazioni SMART58 e SMART137 mostrano una maggiore continuità nelle osservazioni, con un numero significativamente superiore di dati disponibili rispetto alle altre centraline considerate.

Tale comportamento è risultato evidente anche dall'analisi complessiva dei valori mancanti e della copertura temporale delle serie storiche.

2.2.4 Analisi delle correlazioni tra variabili

L'ultima fase dell'analisi esplorativa ha riguardato lo studio delle correlazioni presenti tra le diverse variabili acquisite dalle stazioni di monitoraggio. Tale analisi è stata effettuata attraverso il coefficiente di correlazione di Pearson, con l'obiettivo di individuare eventuali relazioni lineari tra gli inquinanti monitorati e i parametri ambientali disponibili.

Poiché il dataset presenta una disponibilità non uniforme delle osservazioni, la correlazione non è stata calcolata considerando l'intera serie temporale disponibile, ma soltanto le cosiddette "ore complete", ovvero gli istanti temporali per i quali risultavano disponibili valori validi per tutte le variabili analizzate. Questa scelta permette di evitare che la presenza di

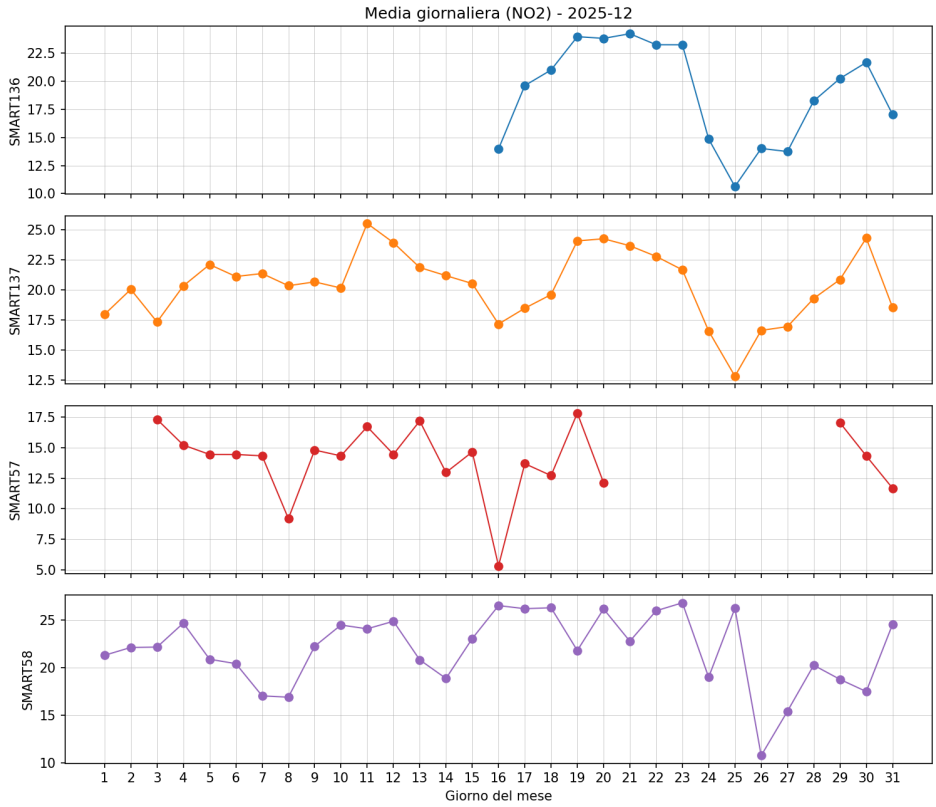


Figura 2.7: Esempio di distribuzione temporale dei valori mancanti per la metrica NO2.

valori mancanti in alcune metriche possa influenzare il calcolo della correlazione e garantisce che il confronto tra variabili venga effettuato sulla medesima base temporale.

L'analisi è stata ripetuta separatamente per ciascuna stazione di monitoraggio, ottenendo una matrice di correlazione per ogni sensore. Contestualmente alla matrice è stato riportato anche il numero di ore complete utilizzate per il calcolo, così da valutare l'affidabilità statistica dei risultati ottenuti. Il numero di ore complete per ciascun sensore è:

- *SMART31*: 635 ore complete.
- *SMART57*: 1.332 ore complete.
- *SMART58*: 1.999 ore complete.
- *SMART136*: 228 ore complete.
- *SMART137*: 2.653 ore complete.

Sono state prese in considerazione le matrici risultanti dalle stazione con il maggior numero di ore complete, ovvero SMART137 e SMART58. I risultati sono riportati, rispettivamente, in Figura 2.10 e in Figura 2.11

Dall'analisi delle matrici di correlazione considerate emergono alcune relazioni significative tra le variabili; in particolare, alcune risultano coerenti con quanto riportato in letteratura. La relazione più evidente riguarda la correlazione positiva tra le concentrazioni di PM2.5 e PM10, con un coefficiente di Pearson pari a 0.93 per entrambe le stazioni, comportamento atteso considerando la natura comune delle due variabili, entrambe rappresentative della presenza di particolato atmosferico.

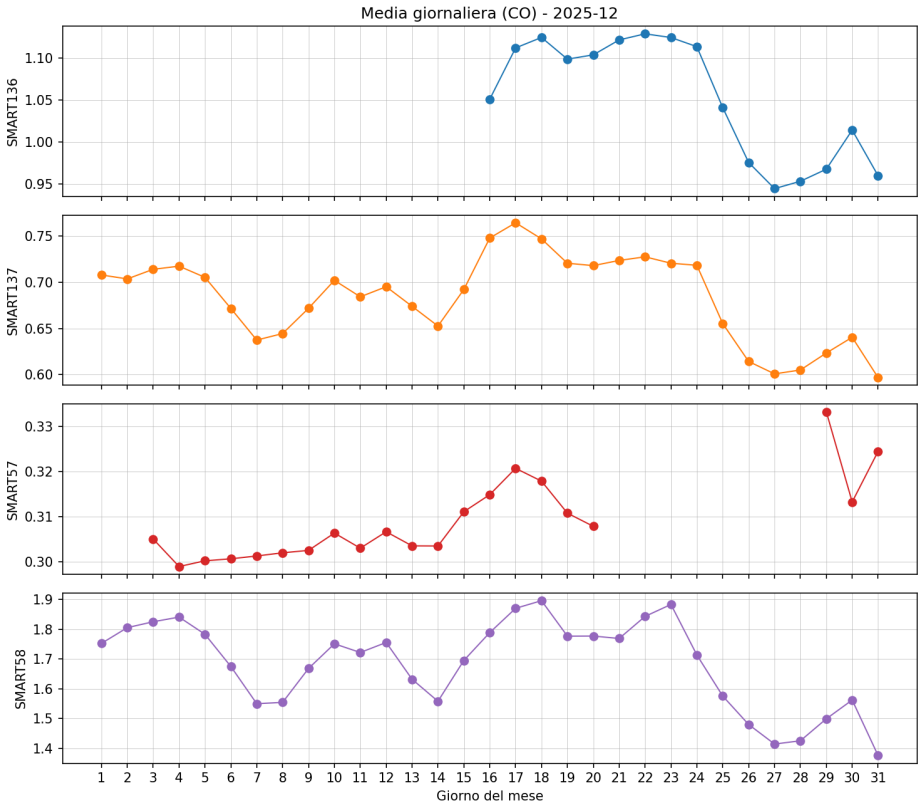


Figura 2.8: Esempio di distribuzione temporale dei valori mancanti per la metrica CO.

Un'altra relazione significativa è quella osservata tra NO_2 e O_3, caratterizzata da una forte correlazione negativa (-0.97 per SMART137 e -0.85 per SMART58). Tale comportamento è coerente con le dinamiche chimiche che regolano la presenza di questi inquinanti in atmosfera, in particolare, con il fatto che il biossido di azoto sia un precursore dell'ozono troposferico.

Infine, risulta evidente la correlazione tra temperatura esterna e temperatura interna del sensore, con valori prossimi alla correlazione perfetta (0.99 per SMART137 e 1.00 per SMART58). Tale relazione è attribuibile al fatto che la temperatura interna del dispositivo risente delle condizioni ambientali esterne, seguendone, quindi, l'andamento temporale.

2.3 Criticità del dataset

L'analisi esplorativa effettuata ha evidenziato alcune criticità, legate principalmente alla disponibilità temporale delle osservazioni. La problematica più rilevante riguarda la presenza di numerosi intervalli privi di acquisizioni, distribuiti in modo non uniforme tra le diverse stazioni e tra le differenti metriche monitorate.

La presenza di tali lacune ha rappresentato un elemento critico, soprattutto nella fase iniziale di sviluppo dei modelli predittivi. Infatti, i modelli di previsione delle serie temporali richiedono generalmente sequenze di dati sufficientemente complete e regolari per poter apprendere correttamente le dinamiche temporali della variabile analizzata. La presenza di ampi periodi mancanti ha, quindi, reso complessa la costruzione delle finestre temporali necessarie all'addestramento, riducendo il numero di campioni effettivamente utilizzabili e compromettendo potenzialmente la capacità del modello di identificare pattern temporali significativi.

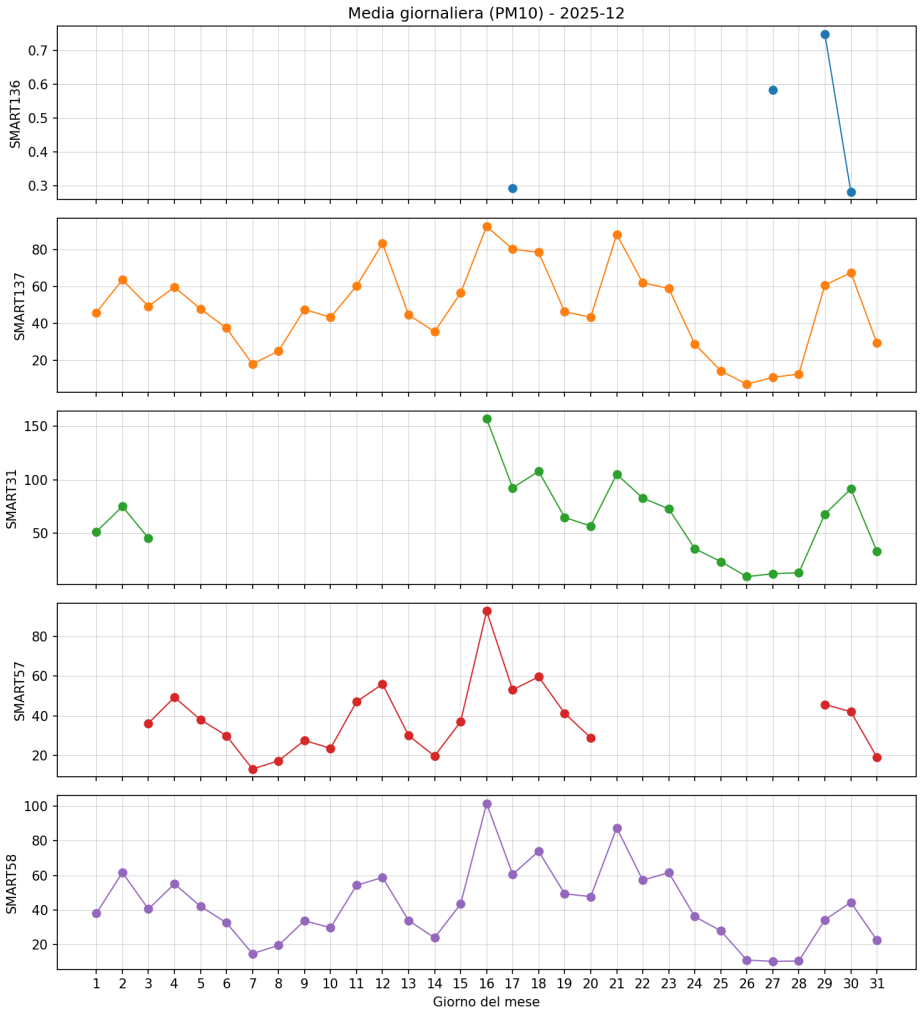


Figura 2.9: Esempio di distribuzione temporale dei valori mancanti per la metrica PM10.

Inoltre, la distribuzione non uniforme delle osservazioni ha introdotto una differenza nella qualità informativa delle diverse stazioni. Alcuni sensori presentavano, infatti, una copertura temporale insufficiente per supportare un'analisi predittiva affidabile, mentre altre centraline, caratterizzate da una maggiore continuità delle acquisizioni, hanno consentito di ottenere serie storiche più complete e rappresentative.

Per tale motivo, le analisi successive sono state indirizzate principalmente verso le stazioni con la maggiore disponibilità di dati, in particolare SMART58 e SMART137, individuate attraverso le precedenti analisi di copertura temporale e completezza delle osservazioni.

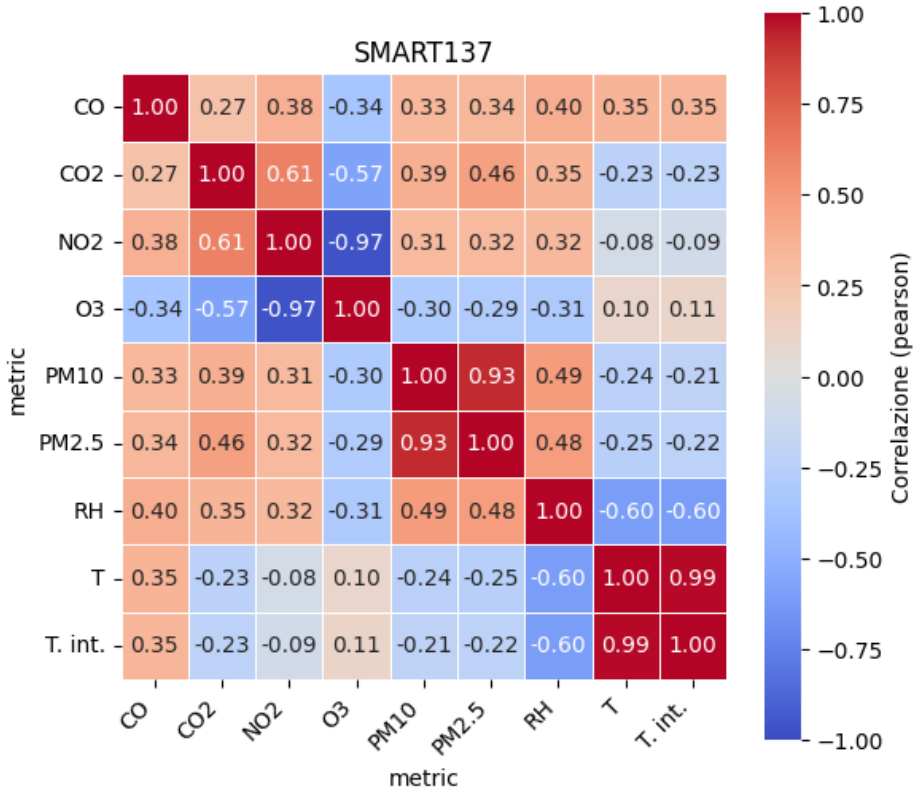


Figura 2.10: Matrice di correlazione per la stazione SMART137.

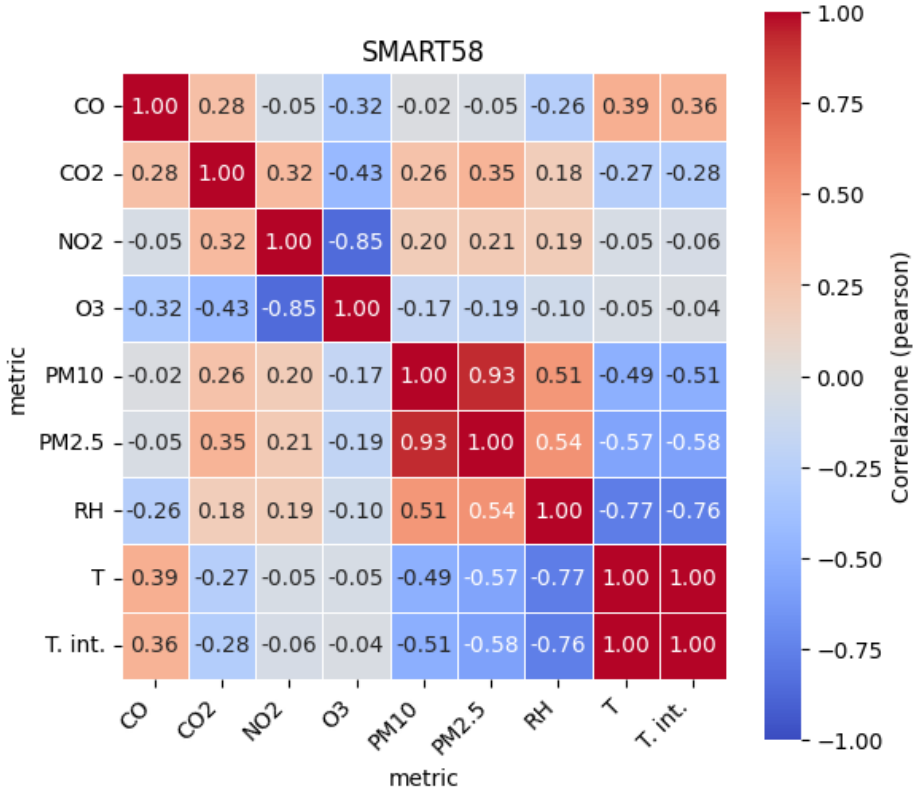


Figura 2.11: Matrice di correlazione per la stazione SMART58.

CAPITOLO 3

Applicazione di tecniche classiche al dataset di riferimento

Nel presente capitolo vengono descritti i primi esperimenti di previsione effettuati sul dataset descritto nel capitolo precedente mediante l'applicazione di tecniche classiche per l'analisi e la modellazione delle serie temporali.

In particolare, vengono analizzati due modelli appartenenti alla famiglia dei modelli autoregressivi integrati a media mobile: ARIMAX e SARIMA. Il primo permette di integrare, oltre alla dinamica storica della variabile analizzata, anche informazioni provenienti da variabili esogene, mentre il secondo sfrutta le caratteristiche temporali della serie, includendo possibili componenti stagionali.

Verranno descritte la configurazione sperimentale adottata per l'applicazione dei modelli al dataset di riferimento, la preparazione delle serie temporali, la suddivisione dei dati per l'addestramento e la valutazione. Successivamente verrà presentato un resoconto dei modelli utilizzati, illustrandone le caratteristiche principali e il ruolo delle loro diverse componenti nella previsione delle serie temporali, nonchè le metriche utilizzate per il confronto delle prestazioni ottenute.

Infine, verranno presentati e discussi i risultati sperimentali, evidenziando le capacità e i limiti dell'applicazione di tali modelli al problema della previsione delle concentrazioni degli inquinanti atmosferici. Questi esperimenti costituiscono una baseline di riferimento con cui confrontare approcci successivi.

3.1 Definizione degli esperimenti

Per l'implementazione degli esperimenti sono state utilizzate diverse librerie Python finalizzate alla gestione dei dati, alla loro analisi e all'applicazione dei modelli predittivi. In particolare, pandas e NumPy sono state impiegate per la manipolazione delle serie temporali e le operazioni numeriche, Matplotlib per la visualizzazione dei risultati, mentre Statsmodels e Scikit-learn hanno fornito gli strumenti necessari, rispettivamente, per la costruzione dei modelli e per la valutazione delle prestazioni ottenute.

Prima dell'applicazione dei modelli previsivi, è stato necessario effettuare una fase di organizzazione dei dati, con l'obiettivo di ottenere serie temporali strutturate secondo un formato coerente con i requisiti dei modelli utilizzati. A partire dai file CSV puliti, sono state considerate inizialmente le stazioni di monitoraggio con la maggiore disponibilità di osservazioni, ovvero SMART137 e SMART58. Per ciascuna centralina sono stati caricati i dati disponibili relativi alle diverse misurazioni ambientali.

La colonna relativa alla data di acquisizione è stata convertita in formato datetime e utilizzata come indice della serie temporale. Dallo studio fatto sui dati sappiamo che non tutte le acquisizioni presentano la medesima disponibilità di variabili; pertanto, è stato necessario verificarne la presenza. Nel caso di assenza di una determinata variabile, il valore

corrispondente è stato mantenuto come dato mancante, così da preservare la struttura temporale originale senza introdurre valori artificialmente stimati.

I modelli ARIMAX e SARIMA richiedono una sequenza temporale regolare, nella quale ogni osservazione possa essere associata ad un intervallo temporale definito. Sebbene i dati acquisiti dalle centraline siano, teoricamente, caratterizzati da una frequenza oraria, nella pratica le acquisizioni non risultano sempre distribuite in modo uniforme nel tempo, con possibili intervalli privi di misurazioni tra due osservazioni consecutive. Per questo motivo, le serie temporali sono state riportate su una griglia temporale oraria, inserendo gli istanti temporali mancanti e associando a tali posizioni valori nulli in assenza di una misura disponibile.

Tuttavia, la presenza di valori nulli all'interno delle serie temporali rappresenta una criticità rilevante nell'applicazione di modelli autoregressivi. Questo perchè tali modelli basano la previsione sull'analisi delle dipendenze tra osservazioni passate e presenti. Nel caso di periodi caratterizzati da acquisizioni incomplete, il modello non è in grado di distinguere tra una reale variazione della concentrazione dell'inquinante e una semplice assenza di misura. La presenza di intervalli privi di osservazioni può, quindi, compromettere la rappresentazione della dinamica temporale della serie e influenzare negativamente la capacità predittiva del modello.

A causa di ciò, prima dell'applicazione degli algoritmi previsivi, è stato necessario individuare i mesi caratterizzati dall'assenza di valori nulli. Per ogni mese completo è stata effettuata una suddivisione temporale interna, utilizzando le prime settimane come periodo di addestramento e gli ultimi cinque giorni come periodo di valutazione.

3.2 Modelli utilizzati

Nei primi esperimenti di predizione sono stati usati i modelli ARIMAX e SARIMA, che sono modelli classici di serie temporali e nascono come estensioni del modello ARIMA.

3.2.1 ARIMAX

Il modello ARIMAX (*AutoRegressive Integrated Moving Average with eXogenous variables*) rappresenta un'estensione del modello ARIMA tradizionale, introdotta per includere all'interno della previsione anche informazioni provenienti da variabili esterne alla serie temporale analizzata.

L'obiettivo degli esperimenti condotti è stato quello di valutare se l'introduzione di una variabile esogena fosse in grado di fornire informazioni utili al miglioramento della previsione, rispetto ad un modello basato esclusivamente sull'andamento storico della variabile target.

Infatti mentre un modello ARIMA costruisce la previsione basandosi esclusivamente sull'evoluzione passata della variabile di interesse, ARIMAX estende tale approccio introducendo informazioni aggiuntive provenienti da una o più variabili esogene. Questa caratteristica è di particolare interesse nell'ambito dello studio della qualità dell'aria, poiché le concentrazioni degli inquinanti dipendono non solo dalla loro evoluzione nel tempo, ma anche dalle interazioni con altre grandezze ambientali, quali le concentrazioni di altri inquinanti e i parametri meteorologici.

Il modello ARIMAX combina quattro componenti principali, ovvero:

- una componente autoregressiva (AR), che descrive la relazione tra il valore corrente della serie e le sue osservazioni passate;

- una componente integrata (I), introdotta attraverso il processo di differenziazione della serie, necessaria per ottenere una condizione di stazionarietà;
- una componente a media mobile (MA), che considera l'influenza degli errori commessi dal modello nelle osservazioni precedenti;
- una componente esogena (X), che incorpora l'effetto di variabili esterne sulla serie temporale.

Dal punto di vista matematico, il modello ARIMAX può essere espresso come segue:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{m} \beta_k X_{k,t} + \varepsilon_t \tag{3.2.1}$$

dove:

- Y_t rappresenta il valore della variabile target al tempo t;
- c rappresenta il termine costante del modello;
- ϕ_i sono i coefficienti associati alla componente autoregressiva;
- Y_{t-i} rappresenta i valori precedenti della serie temporale;
- θ_j sono i coefficienti associati alla componente di media mobile;
- ε_{t-j} rappresenta gli errori residui ai tempi precedenti;
- $X_{k,t}$ indica la k-esima variabile esogena disponibile al tempo t;
- β_k sono i coefficienti associati alle variabili esogene;
- ε_t rappresenta il termine di errore del modello.

Nel caso specifico del presente lavoro, la variabile Y_t coincide con la concentrazione dell'inquinante che si vuole prevedere, mentre la variabile $X_{k,t}$ rappresenta una grandezza ambientale aggiuntiva fornita al modello.

Il modello ARIMAX è, inoltre, caratterizzato da tre parametri (p, d, q), che definiscono, rispettivamente, l'ordine della componente autoregressiva, il grado di differenziazione applicato alla serie e l'ordine della componente a media mobile. Più specificatamente:

- p indica l'ordine della componente autoregressiva e determina il numero di valori passati della serie temporale utilizzati per stimare il valore corrente della variabile analizzata;
- d controlla il numero di differenziazioni applicate alla serie temporale, al fine di rimuovere eventuali trend e ottenere una serie maggiormente stazionaria;
- q definisce il numero di errori precedenti considerati nella componente di media mobile.

La selezione dei parametri ottimali rappresenta un passaggio fondamentale nella costruzione del modello poiché una configurazione non adeguata può portare a una descrizione non corretta della dinamica temporale della serie. Per evitare una selezione manuale dei parametri, si è fatto uso della procedura automatica *auto-ARIMA*, implementata mediante la libreria *pmdarima*. Tale procedura consente di esplorare automaticamente differenti combinazioni dei parametri che caratterizzano la famiglia dei modelli ARIMA. La configurazione

finale viene selezionata sulla base del criterio informativo di Akaike (AIC), scegliendo il modello caratterizzato dal valore più basso di tale indice.

I primi esperimenti sono stati condotti utilizzando i dati disponibili fino al mese di gennaio 2026. Per ciascuna coppia di variabile target e variabile esogena sono stati individuati i mesi con osservazioni complete per entrambe le variabili coinvolte. Tali mesi sono stati successivamente utilizzati per la fase di valutazione predittiva.

3.2.2 SARIMA

Il modello SARIMA (*Seasonal AutoRegressive Integrated Moving Average*) rappresenta un'estensione del modello ARIMA, sviluppata per descrivere serie temporali caratterizzate dalla presenza di componenti stagionali, ovvero comportamenti periodici che si ripetono a intervalli regolari nel tempo.

Nel contesto della qualità dell'aria, questa aggiunta risulta particolarmente rilevante poiché alcune concentrazioni di inquinanti possono presentare variazioni periodiche legate, ad esempio, a cicli giornalieri o settimanali associati alle attività antropiche e alle condizioni ambientali.

Il modello SARIMA è identificato attraverso la notazione:

$$SARIMA(p,d,q)(P,D,Q,s) \tag{3.2.2}$$

dove il primo gruppo di parametri (p,d,q) descrive la componente non stagionale del modello, mentre il secondo gruppo (P,D,Q) rappresenta la componente stagionale. Il parametro s indica, invece, la lunghezza del periodo associato alla stagionalità.

Dal punto di vista matematico, il modello SARIMA può essere espresso mediante l'operatore ritardo B come:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)\varepsilon_t \tag{3.2.3}$$

dove:

- $\phi(B)$ rappresenta il polinomio autoregressivo della componente non stagionale;
- $\Phi(B^s)$ rappresenta il polinomio autoregressivo della componente stagionale;
- $(1-B)^d$ indica l'operazione di differenziazione non stagionale;
- $(1-B^s)^D$ rappresenta la differenziazione stagionale;
- $\theta(B)$ rappresenta il polinomio della componente di media mobile non stagionale;
- $\Theta(B^s)$ rappresenta il polinomio della componente di media mobile stagionale;
- ε_t indica il termine di errore del modello.

Analogamente agli esperimenti ARIMAX, anche per il modello SARIMA la configurazione dei parametri non è stata definita manualmente, ma determinata mediante auto-ARIMA. In questo caso la ricerca ha riguardato sia i parametri della componente non stagionale sia quelli relativi alla componente stagionale.

Considerando la frequenza oraria delle serie temporali analizzate, è stato impostato un periodo stagionale pari a 24 osservazioni, corrispondente al ciclo giornaliero. La procedura ha, quindi, individuato, per ciascuna serie temporale considerata, la configurazione $SARIMA(p,d,q)(P,D,Q,24)$ caratterizzata dal miglior criterio statistico tra quelli analizzati.

La procedura sperimentale è stata mantenuta coerente con quella adottata per ARIMAX: per ciascuna serie temporale è stata utilizzata la medesima suddivisione tra dati di addestramento e periodo di valutazione, corrispondente agli ultimi cinque giorni del mese considerato. La previsione generata dal modello è stata confrontata con i valori reali degli ultimi cinque giorni del mese considerato.

3.2.3 Metriche di valutazione

La valutazione delle prestazioni predittive dei modelli è stata effettuata confrontando le previsioni generate sul periodo di test con i valori effettivamente osservati della serie temporale. A tale scopo sono state utilizzate tre metriche di errore: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) e Mean Absolute Percentage Error (MAPE). La scelta di utilizzare più metriche permette di ottenere una valutazione più accurata delle prestazioni dei modelli.

Le metriche sono state calcolate separatamente per ciascun esperimento, ovvero per ogni combinazione costituita dalla stazione di monitoraggio, dal mese considerato e dalla configurazione del modello. Ciò ha permesso di analizzare la variabilità delle prestazioni al variare del periodo temporale e della configurazione utilizzata.

Mean Absolute Error

Il Mean Absolute Error (MAE) rappresenta la media del valore assoluto della differenza tra il valore osservato e quello previsto dal modello.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.2.4}$$

dove y_i rappresenta il valore reale, $\hat{y}_i$ il valore previsto dal modello e n il numero di osservazioni considerate.

Questa metrica fornisce una misura diretta dell'errore medio commesso dal modello nella stessa unità di misura della variabile analizzata. Valori inferiori di MAE indicano una maggiore vicinanza tra valori previsti e valori osservati.

Root Mean Squared Error

Il Root Mean Squared Error (RMSE) misura la radice dell'errore quadratico medio tra valori osservati e valori predetti.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.2.5}$$

Rispetto al MAE, la presenza del termine quadratico attribuisce un peso maggiore agli errori più elevati. Questa caratteristica rende l'RMSE adatto ad evidenziare situazioni in cui il modello produce previsioni caratterizzate da deviazioni significative rispetto ai valori reali.

Anche in questo caso, valori più bassi indicano una migliore capacità predittiva del modello.

Mean Absolute Percentage Error

Il Mean Absolute Percentage Error (MAPE) esprime l'errore medio in termini percentuali rispetto ai valori osservati.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{3.2.6}$$

Questa metrica permette di valutare l'errore relativo commesso dal modello, indipendentemente dalla scala della variabile analizzata, rendendo più semplice il confronto tra serie temporali caratterizzate da valori di concentrazione differenti.

3.3 Risultati sperimentali

In questa sezione vengono presentati i risultati ottenuti nei diversi esperimenti condotti sui modelli di previsione visti. Verrà seguito l'ordine con cui sono stati definiti gli esperimenti, riportando, per ciascuna configurazione, le prestazioni quantitative ottenute e alcune rappresentazioni grafiche significative delle previsioni generate.

Le tabelle riepilogative riportano i valori delle metriche di errore calcolate per ogni esperimento. Per facilitare il confronto tra configurazioni caratterizzate da metriche differenti, i risultati sono stati ordinati secondo il valore del MAPE. Nei casi in cui l'analisi riguarda una singola metrica di valutazione, l'ordinamento è stato, invece, effettuato rispetto al MAE, in quanto produce lo stesso ordinamento delle configurazioni considerate.

3.3.1 Risultati ARIMAX

Gli esperimenti ARIMAX sono stati inizialmente condotti considerando configurazioni costituite da una variabile target e da una singola variabile esogena. La variabile riportata a sinistra della freccia rappresenta la grandezza oggetto della previsione, mentre quella riportata a destra corrisponde all'informazione aggiuntiva fornita al modello.

Le configurazioni considerate sono state:

- $CO \rightarrow$ Temp;
- $CO_2 \rightarrow O_3$;
- $CO_2 \rightarrow$ PM2.5;
- $NO_2 \rightarrow O_3$;
- $O_3 \rightarrow NO_2$;
- PM10 $\rightarrow$ PM2.5;
- PM10 $\rightarrow$ Temp;
- PM2.5 $\rightarrow$ Temp;
- PM10 $\rightarrow$ RH;
- PM2.5 $\rightarrow$ RH.

Per ciascuna configurazione ARIMAX analizzata sono stati calcolati i valori delle metriche MAE, RMSE e MAPE, insieme ai parametri del modello individuati dalla procedura *auto-ARIMA*. Queste informazioni hanno permesso di confrontare quantitativamente le diverse combinazioni di variabili target e variabili esogene considerate.

In Figura 3.1 è riportata la tabella riepilogativa dei risultati ottenuti.

Per ciascun esperimento sono state generate tre rappresentazioni: una relativa all'intero mese valutato, una seconda riportante gli ultimi cinque giorni del mese, ovvero il periodo di

	sensor	month	target	exog	order	MAE	RMSE	MAPE_%
4	SMART58	2026-01	CO2	O3	(4, 1, 2)	7.393404	8.930193	1.732260
3	SMART58	2025-10	CO2	O3	(2, 1, 1)	9.099258	12.106494	2.053471
2	SMART58	2026-01	CO	T	(2, 1, 1)	0.109718	0.138821	6.553505
11	SMART137	2025-12	CO	T	(3, 1, 3)	0.041979	0.050663	6.984023
0	SMART58	2025-10	CO	T	(2, 1, 2)	0.156082	0.196114	8.905140
1	SMART58	2025-12	CO	T	(1, 0, 3)	0.237743	0.277863	17.460118
8	SMART58	2025-10	PM10	RH	(0, 1, 1)	5.014599	7.604380	33.110838
5	SMART58	2025-10	PM10	T	(0, 1, 1)	6.217543	9.945890	37.463203
6	SMART58	2025-12	PM10	T	(1, 1, 0)	18.555996	28.195338	64.178388
9	SMART58	2025-12	PM10	RH	(1, 1, 0)	20.031867	27.602268	97.586902
7	SMART58	2026-01	PM10	T	(1, 1, 1)	14.019039	16.901281	122.750473
10	SMART58	2026-01	PM10	RH	(1, 1, 1)	13.278872	15.223344	135.355789
13	SMART137	2025-12	PM10	RH	(3, 0, 3)	25.846524	34.955273	194.920175
12	SMART137	2025-12	PM10	T	(3, 0, 3)	39.193519	45.444819	468.212805

Figura 3.1: Risultati sperimentali per il modello ARIMAX.

test, e una terza incentrata sempre sugli ultimi cinque giorni, ma contenente l'intervallo di confidenza della predizione.

Nelle Figure 3.2, 3.3 e 3.4 vengono riportati esempi relativi a configurazioni caratterizzate da buone prestazioni predittive. Si osservi come il modello riesca a seguire l'andamento generale della serie, mantenendo una distanza contenuta rispetto ai valori osservati.

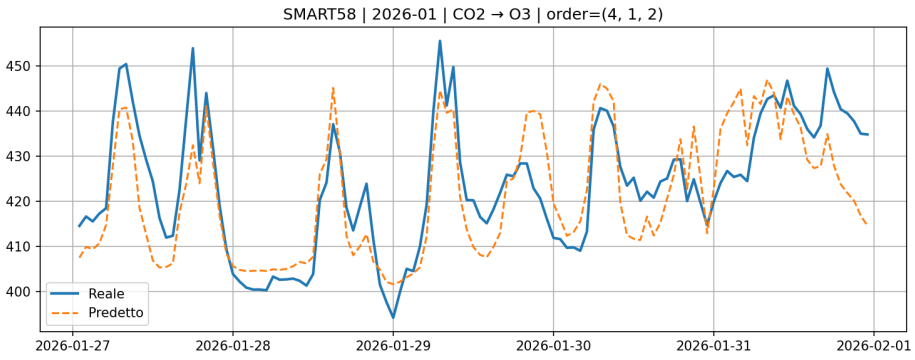


Figura 3.2: Predizione della CO_2 per la stazione SMART58 usando 0_3.

Al contrario, nelle Figure 3.5, 3.6 e 3.7 vengono riportati esempi relativi a configurazioni caratterizzate da prestazioni peggiori. In questi casi il modello riesce a descrivere l'andamento generale della serie, ma presenta difficoltà nel riprodurre alcune variazioni locali e i picchi improvvisi, determinando un aumento dell'errore di previsione rispetto ai casi precedentemente analizzati.

Infine, nelle Figure 3.8, 3.9 e 3.10 si mostra una configurazione caratterizzata da prestazioni più critiche. In tali situazioni la previsione tende ad allontanarsi maggiormente dai valori osservati, evidenziando un'incapacità del modello di adattarsi alla dinamica della serie temporale.

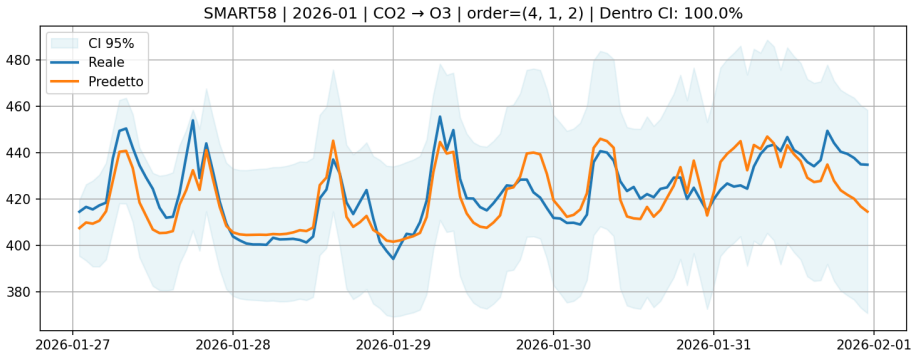


Figura 3.3: Intervallo di confidenza sulla predizione della CO_2 per la stazione SMART58.

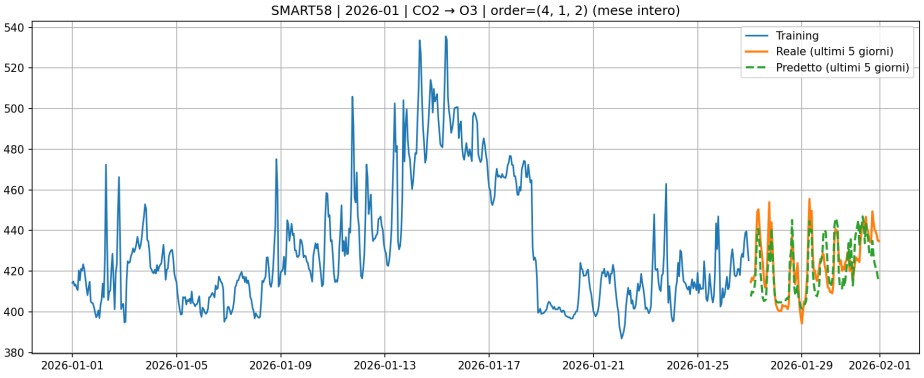


Figura 3.4: Predizione della CO_2 per la stazione SMART58 sull'intero mese di gennaio 2026.

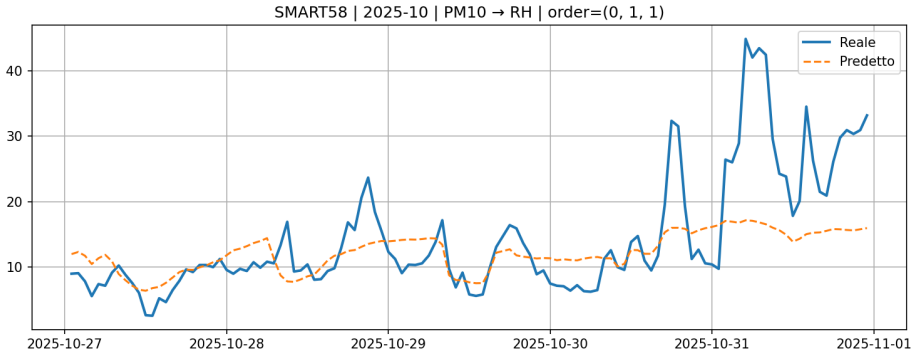


Figura 3.5: Predizione del PM10 per la stazione SMART58.

Valutazione di configurazioni con variabili esogene multiple

L'analisi delle configurazioni iniziali con ARIMAX ha permesso di individuare le variabili sulle quali effettuare ulteriori approfondimenti, mediante l'introduzione di più variabili esogene contemporaneamente. La scelta delle variabili target da analizzare non è stata effettuata esclusivamente sulla base delle metriche di errore ottenute, ma considerando anche la disponibilità delle osservazioni nel dataset.

In particolare, sono state selezionate la CO_2 e il PM10, in quanto caratterizzate da una maggiore presenza di osservazioni complete rispetto ad altre variabili analizzate e, quindi, più adatti a ulteriori valutazioni sperimentali con configurazioni più complesse.

Le configurazioni analizzate sono state:

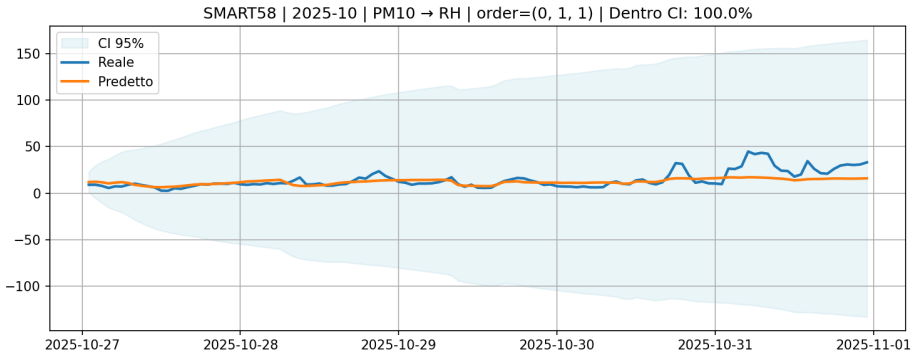


Figura 3.6: Intervallo di confidenza sulla predizione del PM10 per la stazione SMART58.

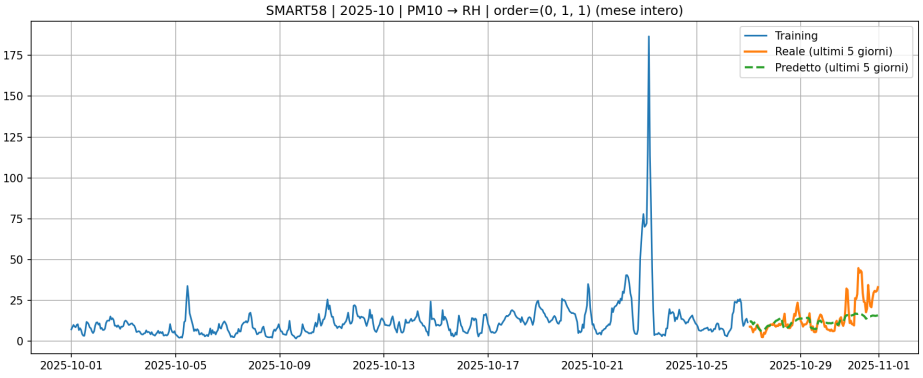


Figura 3.7: Predizione del PM10 per la stazione SMART58 sull'intero mese di ottobre 2025.

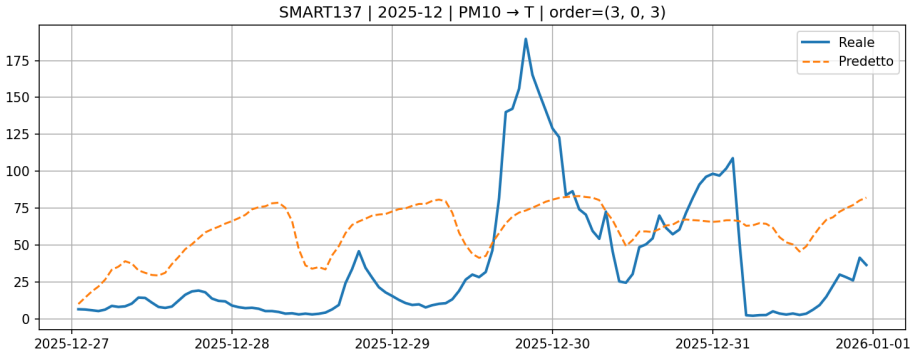


Figura 3.8: Predizione del PM10 per la stazione SMART137.

- $CO_2 \rightarrow$ [O3, RH]
- $CO_2 \rightarrow$ [T, O3, RH]
- $CO_2 \rightarrow$ [RH, PM10]
- $CO_2 \rightarrow$ [T, PM10]
- $CO_2 \rightarrow$ [NO2, PM10]
- $CO_2 \rightarrow$ [T, O3, PM10]
- PM10 $\rightarrow$ [T, RH]

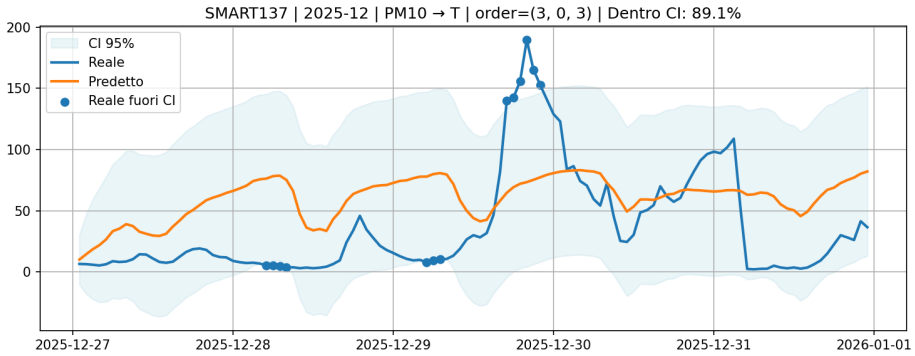


Figura 3.9: Intervallo di confidenza sulla predizione del PM10 per la stazione SMART137.

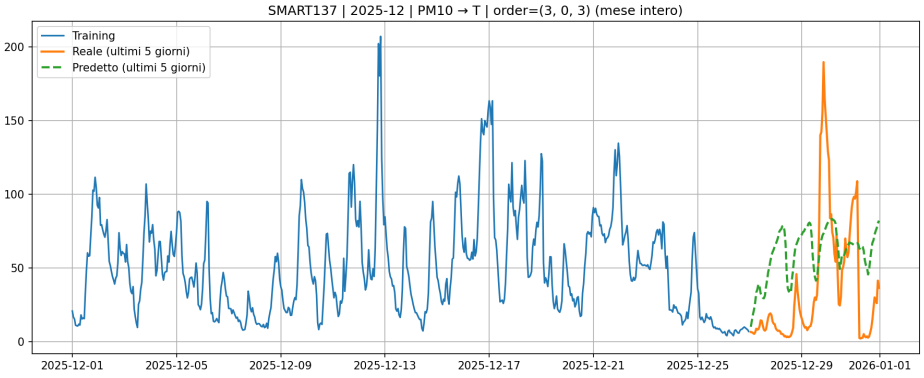


Figura 3.10: Predizione del PM10 per la stazione SMART137 sull'intero mese di dicembre 2025.

- PM10 $\rightarrow$ [T, CO_2]
- PM10 $\rightarrow$ [CO_2, RH]
- PM10 $\rightarrow$ [T, CO_2, RH]

I risultati quantitativi ottenuti sono riportati in Figura 3.12 per la CO_2 e in Figura 3.11 per il PM10. Le tabelle riportano, per ciascun esperimento, il sensore considerato, il mese di valutazione, la configurazione delle variabili esogene utilizzate, la configurazione dei parametri caratteristici e le metriche di errore ottenute.

Oltre alla valutazione quantitativa, è stata effettuata anche un'analisi grafica comparativa. Per ogni esperimento sono stati generati grafici relativi al periodo di test, confrontando la serie osservata con le previsioni ottenute dal modello ARIMAX. Successivamente, per facilitare il confronto tra configurazioni differenti, le previsioni caratterizzate dalla stessa stazione di monitoraggio, dalla stessa variabile target e dallo stesso mese di valutazione sono state riportate all'interno della medesima rappresentazione grafica.

Le valutazioni sono state effettuate esclusivamente sui mesi caratterizzati dalla disponibilità completa delle variabili coinvolte nell'esperimento. In particolare:

- per la CO_2:
  - SMART58: ottobre 2025 e gennaio 2026;
  - SMART137: dicembre 2025.
- per il PM10:

	sensor	month	target	exog	order	MAE	RMSE	MAPE_%
12	SMART58	2025-10	PM10	RH+O3	(0, 1, 1)	4.326777	6.795666	26.016249
8	SMART58	2025-10	PM10	CO2+RH	(0, 1, 1)	4.267140	6.557387	29.231386
6	SMART58	2025-10	PM10	T+CO2	(0, 1, 1)	5.038904	7.992139	32.590086
0	SMART58	2025-10	PM10	RH	(0, 1, 1)	5.014599	7.604380	33.110838
10	SMART58	2025-10	PM10	T+CO2+RH	(0, 1, 1)	4.254062	5.547905	37.701802
3	SMART58	2025-10	PM10	T+RH	(0, 1, 1)	4.762776	6.300793	39.960885
13	SMART58	2025-12	PM10	RH+O3	(1, 1, 0)	16.265641	22.715729	86.911832
14	SMART58	2026-01	PM10	RH+O3	(1, 1, 1)	11.832032	14.450634	90.870210
1	SMART58	2025-12	PM10	RH	(1, 1, 0)	20.031867	27.602268	97.586902
7	SMART58	2026-01	PM10	T+CO2	(1, 1, 1)	13.459556	16.650212	102.905546
11	SMART58	2026-01	PM10	T+CO2+RH	(1, 1, 1)	13.004820	15.128321	113.977399
9	SMART58	2026-01	PM10	CO2+RH	(1, 1, 1)	13.218675	15.226596	120.959670
2	SMART58	2026-01	PM10	RH	(1, 1, 1)	13.278872	15.223344	135.355789
5	SMART58	2026-01	PM10	T+RH	(1, 1, 1)	13.276695	15.216223	135.425682
15	SMART137	2025-12	PM10	RH	(3, 0, 3)	25.846524	34.955273	194.920175
4	SMART58	2025-12	PM10	T+RH	(1, 1, 0)	26.922304	32.857232	198.068775
18	SMART137	2025-12	PM10	CO2+RH	(3, 0, 3)	23.751128	31.425604	223.020150
16	SMART137	2025-12	PM10	T+RH	(3, 0, 3)	30.582591	37.871950	297.688900
19	SMART137	2025-12	PM10	T+CO2+RH	(3, 0, 3)	29.482884	36.422731	324.785718
17	SMART137	2025-12	PM10	T+CO2	(3, 0, 3)	34.196413	39.434163	382.308368

Figura 3.11: Metriche ottenute dalle diverse configurazioni ARIMAX per la previsione della PM10.

	sensor	month	target	exog	order	MAE	RMSE	MAPE_%
3	SMART58	2026-01	CO2	O3+RH	(4, 1, 2)	7.216215	8.720699	1.691884
5	SMART58	2026-01	CO2	T+O3+RH	(4, 1, 2)	7.254394	8.760179	1.710277
1	SMART58	2026-01	CO2	O3	(4, 1, 2)	7.393404	8.930193	1.732260
4	SMART58	2025-10	CO2	T+O3+RH	(2, 1, 1)	7.931829	10.730686	1.792157
2	SMART58	2025-10	CO2	O3+RH	(2, 1, 1)	8.190343	11.031730	1.850926
0	SMART58	2025-10	CO2	O3	(2, 1, 1)	9.099258	12.106494	2.053471
10	SMART58	2025-10	CO2	T+O3+PM10	(2, 1, 1)	9.548136	12.542626	2.159219
12	SMART137	2025-12	CO2	RH+PM10	(2, 1, 2)	10.525116	15.488035	2.399260
7	SMART58	2026-01	CO2	RH+PM10	(4, 1, 2)	11.090291	13.762251	2.613074
9	SMART58	2026-01	CO2	T+PM10	(4, 1, 2)	11.189776	13.689120	2.642209
11	SMART58	2026-01	CO2	T+O3+PM10	(4, 1, 2)	12.313928	14.265980	2.884024
6	SMART58	2025-10	CO2	RH+PM10	(2, 1, 1)	13.431557	17.203886	3.033703
8	SMART58	2025-10	CO2	T+PM10	(2, 1, 1)	15.557769	20.689857	3.494893
14	SMART137	2025-12	CO2	NO2+PM10	(2, 1, 2)	17.249709	23.116303	3.907616
13	SMART137	2025-12	CO2	T+PM10	(2, 1, 2)	17.925521	21.175637	4.198236

Figura 3.12: Metriche ottenute dalle diverse configurazioni ARIMAX per la previsione della CO_2.

- SMART58: ottobre 2025, dicembre 2025 e gennaio 2026;
- SMART137: dicembre 2025.

Considerato l'elevato numero di configurazioni analizzate, non vengono riportate tutte

le rappresentazioni grafiche generate. Sono state selezionate le configurazioni che hanno ottenuto i migliori risultati, in termini di metriche di errore, per ciascuna variabile target considerata.

In Figura 3.13 viene riportato il confronto tra le configurazioni ARIMAX relative alla previsione della CO_2 per il mese di gennaio 2026, mentre in Figura 3.14 viene mostrato il caso relativo al PM10 per il mese di gennaio 2026. In entrambe le rappresentazioni la prima curva prevista corrisponde alla configurazione baseline, ottenuta mediante l'utilizzo di una singola variabile esogena, mentre le successive rappresentano l'effetto dell'introduzione di ulteriori variabili ambientali.

3.3.2 Risultati SARIMA

Poiché il modello SARIMA è progettato per modellare serie temporali caratterizzate dalla presenza di andamenti periodici ricorrenti, gli esperimenti condotti hanno avuto l'obiettivo di valutare la capacità del modello di sfruttare la struttura temporale e le eventuali regolarità stagionali presenti nelle serie analizzate. In particolare, sono state considerate le serie temporali relative alle seguenti variabili:

- NO_2;
- O_3;
- PM10;
- CO_2.

La selezione di tali variabili è stata effettuata considerando sia la disponibilità delle osservazioni nel dataset sia il loro interesse nell'ambito del monitoraggio della qualità dell'aria.

Le previsioni generate sono state confrontate con i valori osservati negli ultimi cinque giorni del mese selezionato, utilizzando le metriche MAE, RMSE e MAPE definite nella configurazione sperimentale.

I risultati quantitativi ottenuti dagli esperimenti SARIMA sono riportati in Figura 3.15. Per ciascun esperimento vengono indicati il sensore considerato, il mese di valutazione, la variabile prevista, la configurazione del modello individuata dalla procedura automatica e le relative metriche di errore.

Anche in questo caso sono stati generati tre grafici, uno riguardante il mese completo, uno con focus sul periodo di test e uno che esplicita l'intervallo di confidenza della previsione. È stato selezionato un caso rappresentativo corrispondente all'esperimento caratterizzato dal valore di MAPE minore tra quelli analizzati.

Nelle Figure 3.16, 3.17 e 3.18 sono riportati, rispettivamente, l'andamento mensile completo, il dettaglio del periodo di previsione e l'intervallo di confidenza relativo all'esperimento selezionato. Nonostante il caso selezionato rappresenti il miglior risultato ottenuto dal modello SARIMA, si osserva come la previsione non riesca a riprodurre adeguatamente l'evoluzione temporale della serie osservata. Il modello tende a fornire una stima caratterizzata da una limitata variabilità, risultando poco efficace nel seguire l'andamento complessivo della concentrazione nel periodo di valutazione.

3.4 Discussione dei risultati

Gli esperimenti condotti mediante i modelli ARIMAX e SARIMA hanno permesso di valutare la capacità dei modelli classici di serie temporali nel descrivere l'evoluzione delle concentrazioni degli inquinanti atmosferici presenti nel dataset di riferimento.

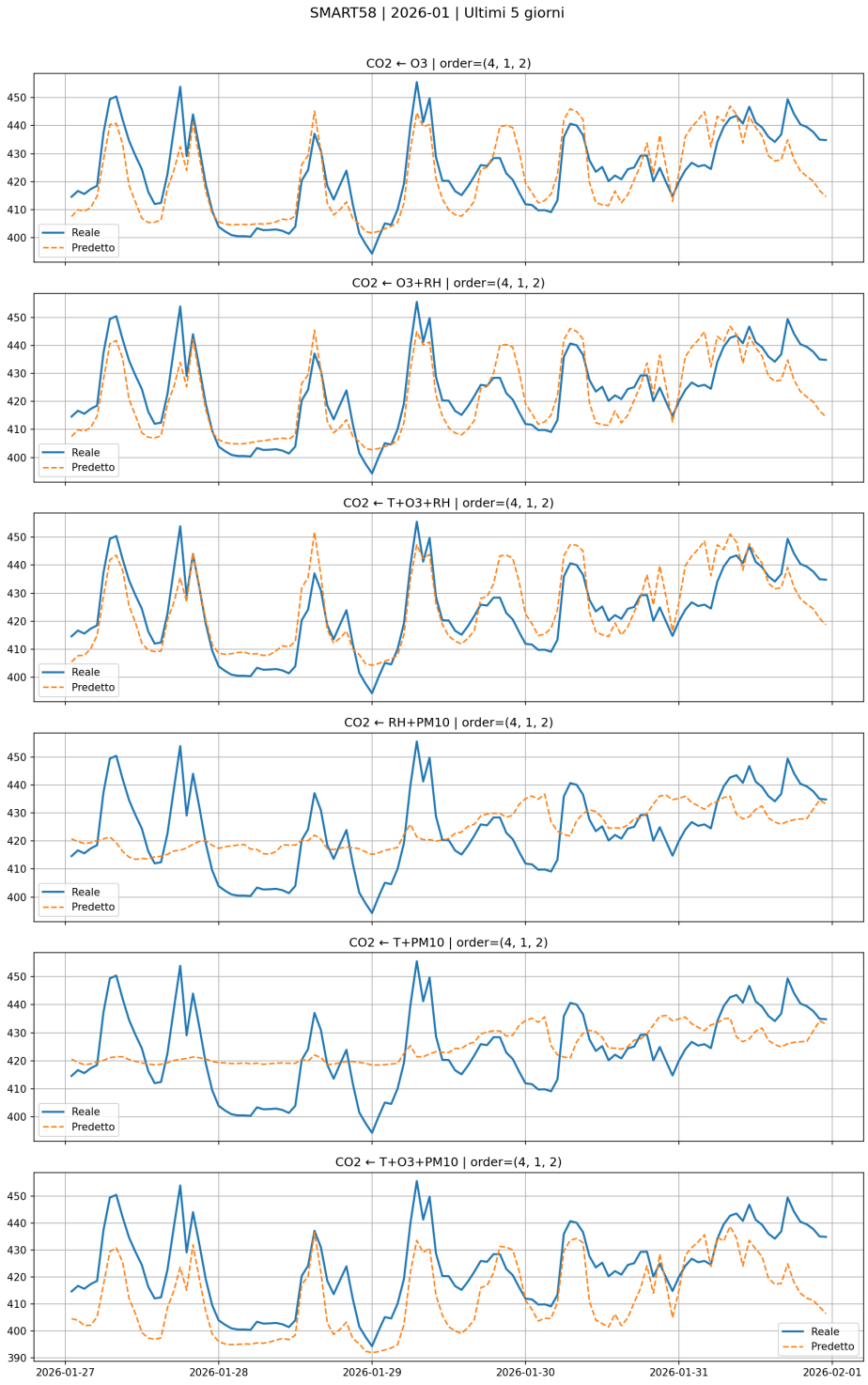


Figura 3.13: Confronto tra le diverse configurazioni ARIMAX per la previsione della CO_2.

Nel complesso, il confronto tra i due approcci evidenzia come l'introduzione di variabili esogene possa fornire un contributo rilevante rispetto ad un modello basato esclusivamente sulla dinamica storica della serie. Il modello SARIMA, infatti, sfruttando soltanto le informazioni contenute nella variabile target, ha mostrato un'incapacità nel riprodurre l'andamento degli inquinanti. Invece, il modello ARIMAX ha evidenziato, in alcuni casi, un miglioramento delle prestazioni grazie alla possibilità di integrare informazioni provenienti da altre

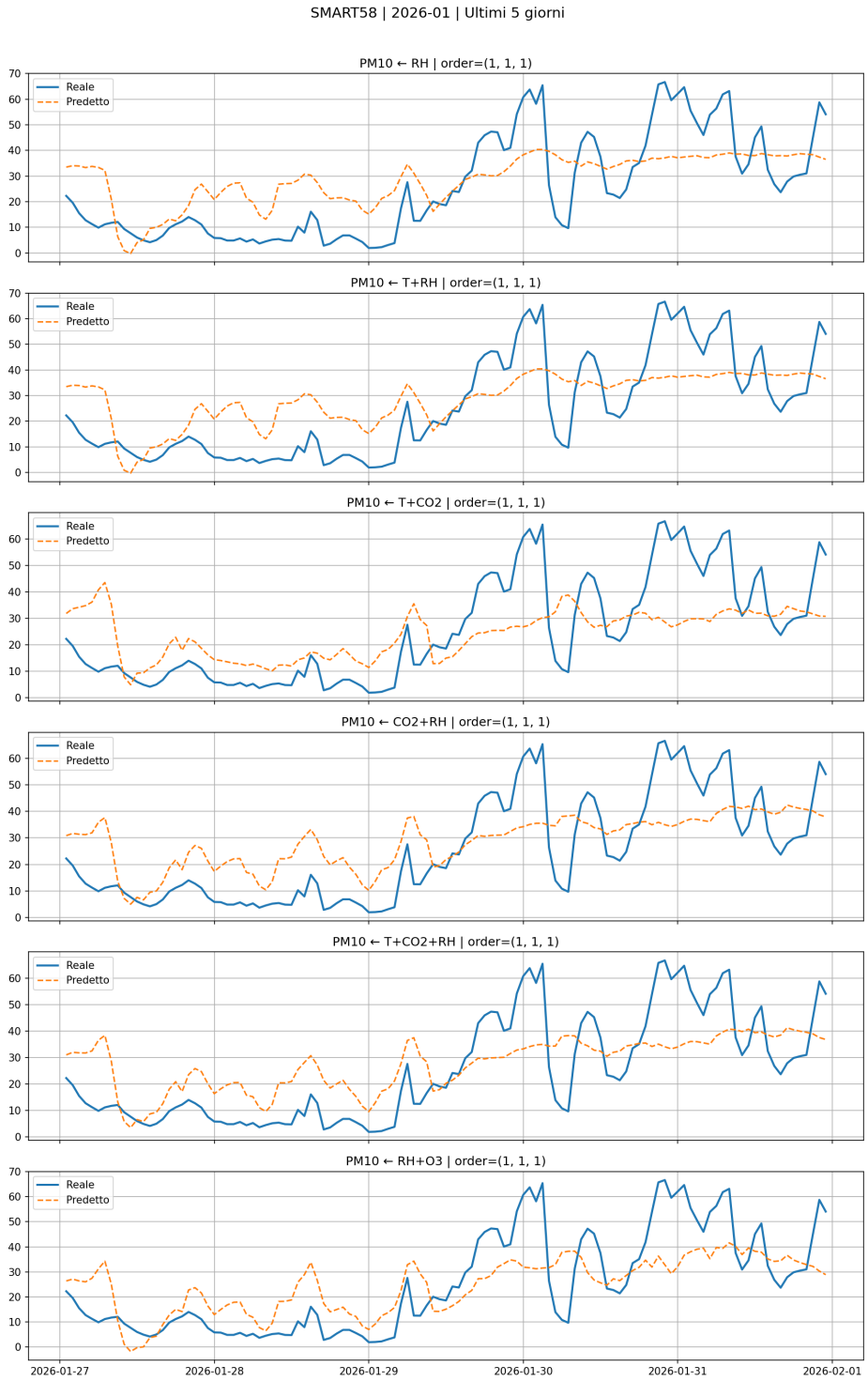


Figura 3.14: Confronto tra le diverse configurazioni ARIMAX per la previsione del PM10.

grandezze ambientali disponibili nella centralina di monitoraggio.

In particolare, i risultati ottenuti dal modello ARIMAX mostrano prestazioni più soddisfacenti per alcune variabili, come la CO_2 e il CO.

Questo comportamento può essere ricondotto alla natura stessa degli inquinanti considerati. Per alcune specie chimiche, infatti, le concentrazioni misurate possono essere legate alla presenza di altri composti atmosferici o a condizioni ambientali; basti pensare al fatto che

	sensor	month	target	order	seasonal_order	MAE	RMSE	MAPE_%
7	SMART58	2026-01	CO2	(1, 1, 2)	(2, 0, 0, 24)	12.536119	15.054600	2.996825
10	SMART137	2025-12	CO2	(1, 1, 0)	(1, 0, 1, 24)	23.684743	32.743298	5.301750
6	SMART58	2025-10	CO2	(2, 1, 2)	(1, 0, 2, 24)	31.482127	40.515598	7.069506
8	SMART137	2025-12	NO2	(2, 0, 2)	(1, 0, 1, 24)	4.839195	5.479497	28.738746
2	SMART58	2026-01	O3	(2, 1, 1)	(0, 0, 2, 24)	14.810474	17.014446	34.376857
3	SMART58	2025-10	PM10	(0, 1, 1)	(0, 0, 0, 24)	6.149552	9.602933	41.768821
1	SMART58	2025-12	O3	(2, 1, 1)	(2, 0, 0, 24)	21.769138	28.182816	90.179095
4	SMART58	2025-12	PM10	(1, 1, 0)	(2, 0, 0, 24)	24.349710	34.810369	92.635287
0	SMART58	2025-10	O3	(1, 1, 2)	(1, 0, 1, 24)	22.475286	31.018625	111.190655
9	SMART137	2025-12	PM10	(1, 0, 0)	(2, 0, 1, 24)	27.524332	40.121315	226.268808
5	SMART58	2026-01	PM10	(1, 1, 1)	(0, 0, 0, 24)	19.681485	21.842017	248.838872

Figura 3.15: Risultati sperimentali per il modello SARIMA.

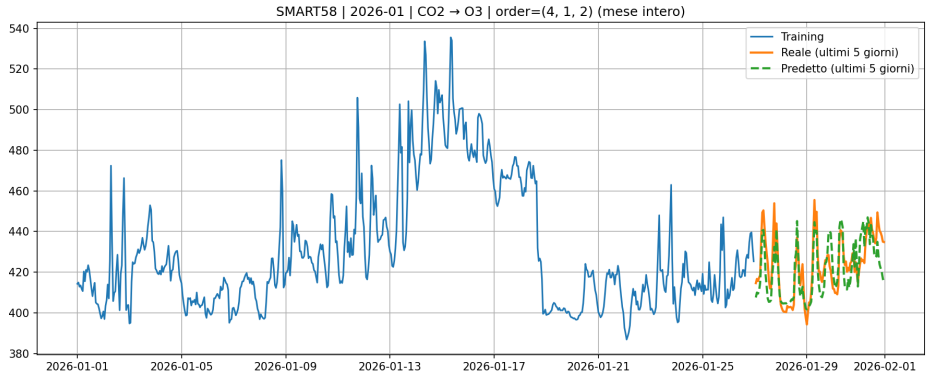


Figura 3.16: Andamento mensile completo per il mese di gennaio 2026 per la stazione SMART58 utilizzando il modello SARIMA.

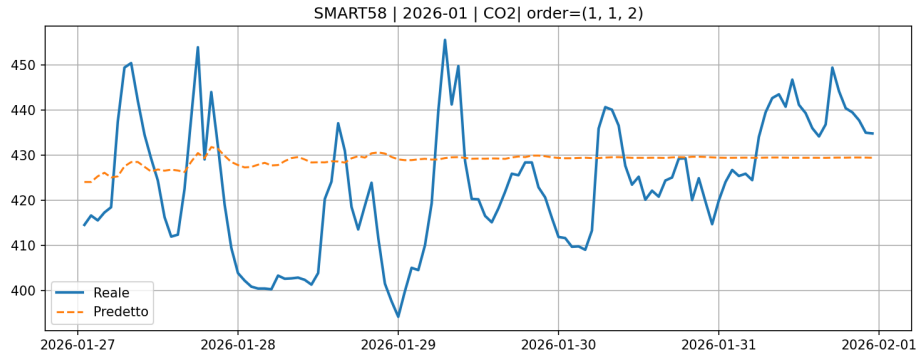


Figura 3.17: Previsione sugli ultimi 5 giorni del mese di gennaio 2026 per la stazione SMART58 utilizzando il modello SARIMA.

il CO è un precursore dell'ozono troposferico (O_3) e che la CO_2 può essere influenzata da fattori meteorologici, come la temperatura e l'umidità relativa.

D'altra parte, per gli inquinanti appartenenti alla frazione del particolato atmosferico, in particolare PM10 e PM2.5, i risultati ottenuti sono meno soddisfacenti. Nonostante il particolato possa essere sensibile ad influenze esterne, la sua concentrazione dipende fortemente anche da sorgenti emissive dirette, come il traffico veicolare e altre attività antropiche.

Per questo motivo, l'utilizzo di sole variabili ambientali, misurate dalla centralina, non è

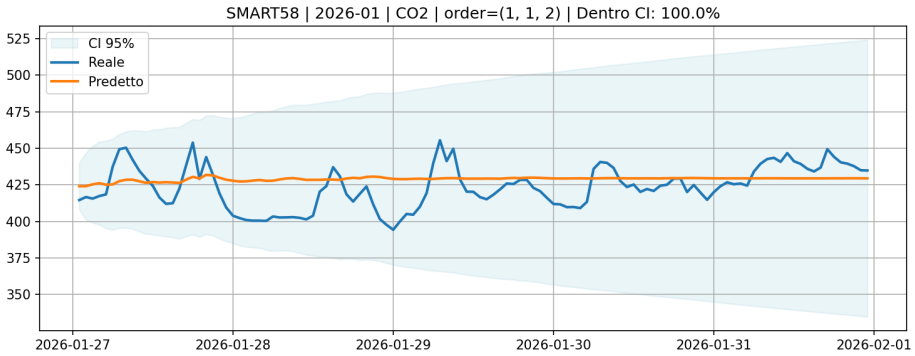


Figura 3.18: Intervallo di confidenza per la previsione del mese di gennaio 2026 per la stazione SMART58 utilizzando il modello SARIMA.

risultato sufficiente per descrivere in maniera accurata tutte le variazioni osservate nella serie temporale. In particolare, eventi caratterizzati da aumenti improvvisi delle concentrazioni risultano difficilmente prevedibili poiché possono essere associati a fenomeni locali o a variazioni delle sorgenti emissive.

Considerando la natura del particolato atmosferico, è stata valutata anche la possibilità che una modellazione in grado di catturare maggiormente le componenti stagionali potesse fornire informazioni aggiuntive. Tuttavia, anche il modello SARIMA non ha mostrato risultati sufficientemente accurati. Ciò suggerisce che la sola presenza di una componente stagionale non sia sufficiente a descrivere la complessità del fenomeno analizzato.

Una possibile causa delle prestazioni ottenute è costituita dalla disponibilità limitata di informazioni del dataset. Il periodo di osservazione disponibile non consente, infatti, di analizzare ciclicità stagionali su intervalli temporali più lunghi, come, ad esempio, confronti tra differenti stagioni o anni consecutivi. Una maggiore disponibilità di dati storici avrebbe potuto permettere una caratterizzazione più completa degli effetti stagionali e delle variazioni associate alle diverse condizioni ambientali.

I risultati ottenuti evidenziano come, sebbene i modelli classici siano in grado di sfruttare alcune relazioni temporali e ambientali presenti nei dati, la previsione della qualità dell'aria rimanga un problema complesso, influenzato da molteplici fattori difficilmente rappresentabili attraverso modelli statistici tradizionali. Queste considerazioni motivano l'esplorazione di approcci maggiormente flessibili basati su tecniche di apprendimento automatico.

CAPITOLO 4

Analisi di tecniche innovative per la predizione degli inquinanti

Nel seguente capitolo veranno esplorate le tecniche avanzate per la predizione degli inquinanti atmosferici utilizzate nel lavoro sperimentale. Verranno approfondite le idee alla base di tali metodologie, come queste si appplicano al caso in questione e l'uso che ne è stato fatto in lettaratura, oltre alla descrizione dei pregi e dei difetti di ciascuno di questi.

4.1 Introduzione a modelli predittivi avanzati

Dopo aver analizzato i risultati ottenuti mediante i modelli autoregressivi a media mobile, sono emerse alcune limitazioni nella previsione delle serie temporali relative agli inquinanti atmosferici. In particolare, le configurazioni analizzate hanno mostrato difficoltà nel riprodurre variazioni locali e fenomeni caratterizzati da dinamiche non lineari, evidenziando la necessità di valutare metodologie alternative più adatte a descrivere la complessità dei dati ambientali.

Per questo motivo sono stati analizzati approcci predittivi più recenti, inizialmente concentrando l'attenzione sui modelli Prophet e NeuralProphet, sviluppati per la previsione di serie temporali caratterizzate da trend, stagionalità e possibili componenti non lineari. Successivamente, attraverso l'analisi della letteratura scientifica relativa alla previsione degli inquinanti atmosferici, sono stati individuati diversi filoni metodologici che hanno mostrato risultati promettenti in questo ambito.

Tenendo conto di Prophet e NeuralProphet, sono stati considerati quattro principali approcci:

- Modelli basati sulla decomposizione delle componenti temporali, che mirano a separare la serie originale nelle sue principali componenti, come trend e stagionalità, rendendo più semplice la modellazione delle diverse dinamiche presenti.
- Modelli ibridi per la previsione delle serie temporali, che combinano modelli statistici e tecniche di apprendimento automatico o deep learning al fine di sfruttare contemporaneamente caratteristiche lineari e non lineari della serie.
- Modelli basati su reti neurali grafiche, che introducono la componente spaziale del problema attraverso la rappresentazione del problema come un grafo, permettendo di modellare le relazioni tra differenti aree geografiche.

- Tecniche di decomposizione avanzata delle serie temporali che consentono di scomporre segnali complessi e non stazionari in componenti più semplici successivamente utilizzabili per la previsione.

4.2 Modelli basati sulla decomposizione delle componenti temporali

Una possibile strategia per affrontare la previsione di serie temporali complesse consiste nell'effettuare la previsione attraverso una decomposizione preliminare della serie nelle sue componenti temporali principali, anzichè modellare direttamente il valore futuro della variabile analizzata.

Generalmente, una serie temporale può essere rappresentata come la combinazione di:

- Una componente di trend, associata all'evoluzione generale del fenomeno nel lungo periodo.
- Una componente stagionale, relativa alla presenza di variazioni periodiche ricorrenti.
- Una componente residua, legata a variazioni non spiegate dalle componenti precedenti o a fenomeni casuali.

Questa rappresentazione risulta particolarmente interessante nel contesto della previsione degli inquinanti atmosferici, poiché le concentrazioni misurate dalle centraline possono essere influenzate contemporaneamente da fenomeni caratterizzati da differenti scale temporali. Ad esempio, possono essere presenti variazioni giornaliere legate al ciclo delle attività antropiche, variazioni stagionali associate alle condizioni climatiche e cambiamenti improvvisi dovuti a eventi ambientali specifici.

I modelli basati sulla decomposizione delle componenti sfruttano questa struttura per separare i diversi contributi presenti nella serie, consentendo una modellazione più interpretabile e, in alcuni casi, più efficace rispetto agli approcci che considerano il segnale nella sua forma complessiva.

Tra questi approcci rientrano Prophet e NeuralProphet.

4.2.1 Prophet

Prophet è una libreria software open-source sviluppata da Meta Open Source per la previsione di serie temporali. A differenza dei modelli autoregressivi tradizionali, come ARIMA e SARIMA, Prophet non cerca di modellare direttamente la dipendenza tra osservazioni successive, ma assume che la serie temporale osservata sia composta da differenti componenti interpretabili. La struttura del modello permette, quindi, di analizzare separatamente i principali fattori che influenzano l'andamento della serie.

La formulazione generale del modello è:

$$y(t) = g(t) + s(t) + h(t) + e(t) \tag{4.2.1}$$

dove:

- $g(t)$: rappresenta la componente di trend, ovvero l'evoluzione generale della serie nel tempo;
- $s(t)$: rappresenta la componente stagionale, associata alla presenza di variazioni periodiche, come cicli giornalieri, settimanali o annuali;

- $h(t)$: rappresenta gli effetti dovuti a festività o eventi specifici che possono modificare temporaneamente il comportamento della serie;
- $e(t)$: rappresenta la componente residua, associata alla variabilità non spiegata dal modello.

La presenza di una struttura decompositiva rappresenta uno dei principali vantaggi di Prophet, poiché permette di ottenere previsioni facilmente interpretabili e di analizzare separatamente il contributo delle diverse componenti della serie temporale.

Un ulteriore vantaggio del modello riguarda la capacità di gestire automaticamente alcune caratteristiche tipiche delle serie reali, come dati mancanti, cambiamenti nel trend e presenza di più componenti stagionali. Inoltre, la configurazione del modello richiede un numero limitato di parametri definiti manualmente, rendendo Prophet facilmente applicabile anche a serie temporali caratterizzate da strutture complesse.

L'addestramento del modello originale viene effettuato attraverso una procedura basata sull'algoritmo L-BFGS implementato tramite il framework probabilistico Stan. Tale procedura permette di stimare i parametri delle diverse componenti del modello e ottenere la previsione finale della serie temporale.

Nonostante questi vantaggi, Prophet presenta alcune limitazioni. La struttura decompositiva, pur garantendo una maggiore interpretabilità, può risultare meno efficace nella modellazione di variazioni locali improvvise e fenomeni caratterizzati da relazioni fortemente non lineari.

Per superare alcune di queste limitazioni è stato successivamente introdotto NeuralProphet.

4.2.2 NeuralProphet

NeuralProphet nasce come evoluzione del modello Prophet con l'obiettivo di superare alcuni suoi limiti mantenendo, allo stesso tempo, le caratteristiche positive che hanno reso Prophet uno dei modelli più interessanti nella previsione di serie temporali. Il modello combina infatti approcci tipici dell'analisi delle serie temporali tradizionale con tecniche basate su reti neurali, sfruttando il framework PyTorch come ambiente di sviluppo.

L'idea alla base di NeuralProphet deriva dalla necessità di colmare il divario esistente tra due categorie di modelli: da un lato i modelli statistici classici, come ARIMA e Prophet, caratterizzati da una maggiore interpretabilità ma limitati nella capacità di rappresentare relazioni complesse e non lineari; dall'altro i modelli di deep learning, generalmente più flessibili e capaci di apprendere pattern complessi, ma spesso difficili da interpretare.

NeuralProphet propone un approccio ibrido, mantenendo la struttura decompositiva di Prophet e integrandola con componenti neurali in grado di modellare dinamiche temporali più complesse.

La caratteristica principale di NeuralPophet è la sua struttura modulare. Il modello è composto da diverse componenti indipendenti, ciascuna responsabile di modellare un particolare comportamento della serie temporale. La previsione finale viene ottenuta combinando le diverse componenti:

$$\hat{y}(t) = T(t) + S(t) + E(t) + F(t) + A(t) + L(t) \tag{4.2.2}$$

dove:

- T(t) rappresenta la componente di trend e descrive l'evoluzione generale della serie;
- S(t) rappresenta gli effetti stagionali ovvero le variazioni periodiche all'interno della serie temporale;

- E(t) rappresenta gli effetti dovuti a eventi e festività;
- A(t) rapprsenta la componente autoregressiva che permette al modello di utilizzare i valori precedenti della serie;
- L(t) rappresenta l'effetto delle variabili esogene ritardate (lagged regressors) che permette di includere variabili esterne alla serie target.

Questa struttura permette di considerare il contesto locale della previsione, ovvero l'influenza esercitata dalle osservazioni più recenti sull'evoluzione futura della variabile analizzata.

Un'altra differenza importante rispetto a Prophet riguarda il metodo di addestramento, NeuralProphet sostituisce l'approccio di Prophet con un addestramento basato su *mini-batch Stochastic Gradient Descent* (SGD) utilizzando PyTorch. Tale scelta permette:

- Una maggiore scalabilitò su dataset di grandi dimensioni.
- L'integrazione diretta di componenti basate su reti neurali.
- Una maggiore flessibilità nell'estensione del modello.

Dal punto di vista applicativo, NeuralProphet risulta particolarmente interessante per problemi caratterizzati da serie temporali ad alta frequenza e da fenomeni influenzati contemporaneamente da componenti periodiche e variazioni locali. Questo aspetto rende il modello adatto anche alla previsione della qualità dell'aria.

4.3 Approcci ibridi per la previsione delle serie temporali

I modelli statistici tradizionali e i modelli basati su reti neurali presentano caratteristiche differenti che possono risultare complementari nell'analisi delle serie temporali.

Per questo motivo, negli ultimi anni sono stati sviluppati numerosi approcci ibridi, con l'obiettivo di combinare modelli caratterizzati da proprietà differenti al fine di ottenere una previsione più completa della dinamica osservata. L'idea alla base di questi metodi consiste nell'affidare a ciascun modello la componente della serie temporale maggiormente adatta ad essere descritta.

Tra le architetture neurali maggiormente utilizzate per questo tipo di applicazioni troviamo le Long Short-Term Memory (LSTM) e, nel corso della sezione, verrà introdotta una soluzione basata sull'uso combinato di Prophet e LSTM.

4.3.1 LSTM

Il modello LSTM (*Long Short-Term Memory*) è una particolare architettura di *Recurrent Neural Network* (RNN) progettata per l'analisi di dati sequenziali, come testi, segnali e serie temporali. La sua introduzione deriva dalla necessità di superare uno dei principali limiti delle RNN tradizionali, ovvero la difficoltà nel mantenere informazioni relative a eventi lontani nel passato.

Nelle RNN classiche, infatti, l'informazione proveniente dalle osservazioni precedenti viene mantenuta attraverso uno stato nascosto, aggiornato ad ogni passo temporale. Durante la fase di addestramento, l'errore viene propagato all'indietro attraverso la sequenza temporale mediante l'algoritmo di *Backpropagation Through Time* (BPTT). Tuttavia, durante questo processo il gradiente può progressivamente diminuire fino ad assumere valori prossimi allo zero, causando il problema noto come *vanishing gradient*.

Quando ciò avviene, gli aggiornamenti dei pesi della rete diventano estremamente ridotti e il modello perde la capacità di apprendere relazioni tra osservazioni temporalmente distanti. Di conseguenza, le RNN tradizionali risultano poco efficaci nella modellazione di dipendenze temporali lunghe.

Per risolvere questo problema, la LSTM introduce una struttura interna denominata *memory cell*, progettata per separare il processo di memorizzazione delle informazioni da quello di utilizzo delle stesse. La cella di memoria permette di conservare informazioni per intervalli temporali più lunghi attraverso un percorso dedicato chiamato *Constant Error Carousel* (CEC), che consente al gradiente di propagarsi attraverso la sequenza riducendo il problema della scomparsa del gradiente.

La gestione delle informazioni contenute nella memoria viene effettuata attraverso specifici meccanismi di controllo chiamati *gate*. Questi permettono alla rete di decidere quali informazioni mantenere, modificare o utilizzare durante l'elaborazione della sequenza.

In particolare, la struttura LSTM comprende tre principali componenti:

- *Forget gate*: controlla quali informazioni presenti nella memoria devono essere eliminate, permettendo alla rete di rimuovere informazioni non più rilevanti;
- *Input gate*: determina quali nuove informazioni provenienti dall'input corrente devono essere memorizzate nello stato interno;
- *Output gate*: controlla quali informazioni presenti nella memoria devono essere utilizzate per generare l'output della rete.

Dal punto di vista dell'apprendimento, il gradiente può quindi fluire all'interno della memoria attraverso un percorso controllato, mentre le modifiche provenienti da percorsi non desiderati vengono limitate attraverso i gate. Questo permette alla rete di apprendere dipendenze temporali lunghe mantenendo una complessità computazionale comparabile a quella delle reti ricorrenti tradizionali.

Grazie a queste caratteristiche, le LSTM sono particolarmente utilizzate nella previsione di serie temporali caratterizzate da dinamiche complesse, rumore e variazioni non lineari. Nel contesto della previsione degli inquinanti atmosferici, esse risultano interessanti poiché consentono di modellare le dipendenze temporali presenti nelle concentrazioni misurate, considerando l'influenza delle osservazioni precedenti sull'evoluzione futura del fenomeno.

4.3.2 Prophet e LSTM

Una delle tecniche ibride più diffuse vede la combinazione degli approcci Prophet, per modellare le componenti strutturate della serie temporale, e di una rete neurale LSTM, utilizzata per apprendere le variazioni più complesse e difficilmente rappresentabili attraverso un modello statistico. Come già detto Prophet considera la serie come una combinazione di diverse componenti, permettendo di ottenere una previsone interpretabile. Tuttavia, presenta qualche limitazione quando la serie contiene variazioni locali irregolari o pattern complessi, come nel caso della previsione dell'andamento di particolato fine.

Per affrontare il problema viene introdotta una rete LSTM per la sua capacità di apprendimento non lineare. Esistono diversi modi con cui legare questi due modelli, quello di maggiore interesse per lo studio proposto si basa sulla correzione dei residui. A differenza degli approcci nei quali Prophet e LSTM producono due previsioni indipendenti, successivamente combinate, in questo caso i due modelli vengono utilizzati in maniera sequenziale. I passaggi seguiti sono:

1. Prophet effettua una prima previsione della serie temporale;

2. vengono calcolati gli errori prodotti da Prophet;
3. LSTM viene addestrata usando questi residui;
4. la previsione finale corregge quella di Prophet attraverso il contributo della rete neurale.

Tale approccio riduce il compito della rete neurale, perchè non deve imparare nuovamente trend e stagionalità già modellati, ma solo le componenti difficili e non spiegate. Per una previsione di inquinanti atmosferici tale approccio è particolarmente adatto perchè consente di separare i fenomeni periodici da anomalie e variazioni rapide legate alla meteorologia o alle emissioni.

4.4 Modelli basati su reti neurali grafiche

I modelli descritti nelle sezioni precedenti considerano principalmente la componente temporale del fenomeno analizzato, cercando di individuare pattern presenti nella sequenza storica della variabile considerata. Tuttavia, in numerosi problemi reali, l'evoluzione di una grandezza non dipende esclusivamente dal suo andamento passato, ma anche dalle relazioni esistenti tra differenti punti di osservazione.

Questo aspetto assume particolare rilevanza nel caso della previsione della qualità dell'aria, dove la concentrazione degli inquinanti, registrata in una determinata posizione, può essere influenzata dalle condizioni presenti in aree vicine. Per affrontare questo limite sono state introdotte le Graph Neural Network (GNN), una famiglia di modelli di deep learning progettata per lavorare su dati rappresentabili attraverso grafi. In questo tipo di rappresentazione, gli elementi del problema vengono descritti attraverso nodi e archi.

Nel contesto della qualità dell'aria, ogni stazione di monitoraggio può essere rappresentata come un nodo del grafo, mentre le connessioni tra le stazioni possono essere definite sulla base della distanza geografica, delle condizioni meteorologiche condivise o di altre relazioni individuate tra i punti di misura. Attraverso il processo di aggregazione delle informazioni provenienti dai nodi vicini, le GNN sono in grado di apprendere contemporaneamente le dipendenze spaziali e quelle temporali, fornendo una rappresentazione più completa del fenomeno analizzato.

4.4.1 GNN

Le reti neurali a grafo (GNN) sono una famiglia di modelli di deep learning progettata per lavorare su dati rappresentabili mediante grafi. Esse permettono di modellare direttamente relazioni spaziali non euclidee, rappresentando il problema attraverso un grafo.

Le GNN ricevono come input una struttura grafica composta da nodi, archi e caratteristiche associate agli elementi del grafo. Durante il processo di apprendimento queste informazioni vengono trasformate in rappresentazioni numeriche (embedding) che permettono al modello di apprendere le relazioni presenti nella struttura.

Per addestrare le reti a grafo si creano sottografi che preservano le proprietà essenziali del grafo principale di grandi dimensioni. Le GNN utilizzano un meccanismo di trasmissione dei messaggi per aggregare le informazioni dai nodi vicini, consentendo loro di catturare le complesse relazioni nei grafi.

ARIMAX, SARIMA, Prophet e anche LSTM trattano principalmente la serie temporale come una sequenza dove il valore futuro dipende dai valori precedenti. Tuttavia, nel caso degli inquinanti atmosferici, questa rappresentazione è limitata perché la concentrazione misurata in una stazione non dipende esclusivamente dalla sua storia temporale, ma anche da ciò che avviene nelle aree circostanti.

Quindi il problema diventa prevedere una variabile che dipende contemporaneamente dalla dimensione temporale e dalla relazione spaziale tra diversi punti di osservazione. Questo è il motivo per cui le GNN risultano interessanti.

Il concetto fondamentale da introdurre è, dunque, quello del grafo. Una GNN lavora infatti su una struttura:

$$G = (V, E) \tag{4.4.1}$$

dove:

- V rappresenta l'insieme dei nodi.
- E rappresenta le connessioni tra i nodi.

Nel caso specifico del dataset di interesse per l'elaborato, i nodi vanno a rappresentare le centraline di monitoraggio, ed ogni nodo contiene delle caratteristiche come la concentrazione degl inquinanti, la temperatura esterna ed interna, umidità e coordinate geografiche

Il funzionamento base delle GNN si basa sul concetto di *message passing*. In pratica ogni nodo possiede informazioni proprie, riceve informazioni dai nodi collegati e combina queste informazioni per poi aggiorna il proprio stato.

Le GNN permettono, quindi, di modellare la dipendenza spaziale tra differenti punti di osservazione. Tuttavia, quando vengono applicate a fenomeni dinamici come la qualità dell'aria, è necessario integrare anche la componente temporale associata all'evoluzione delle serie misurate. Questo ha portato allo sviluppo di architetture ibride capaci di combinare la rappresentazione spaziale dei grafi con modelli per l'analisi temporale.

ST-GNN

Le *Spatial-Temporal Graph Neural Network* (ST-GNN) sono una particolare categoria di reti neurali grafiche progettata per analizzare fenomeni caratterizzati contemporaneamente da una componente spaziale e da una componente temporale.

A differenza delle GNN tradizionali, che considerano principalmente le relazioni tra gli elementi di un grafo in un determinato istante, le ST-GNN sono progettate per modellare l'evoluzione nel tempo delle caratteristiche associate ai nodi del grafo. Nel caso della previsione degli inquinanti atmosferici, in cui il grafo rappresenta la rete di monitoraggio, ciascun nodo non contiene una singola informazione, ma una sequenza temporale relativa alla variabile analizzata.

L'obiettivo di una ST-GNN è quello di apprendere contemporaneamente due tipologie di dipendenze:

- Le dipendenze spaziali: determinate dalle relazioni tra nodi collegati nel grafo.
- Le dipendenze temporali: legate all'evoluzione della variabile osservata nei diversi istanti temporali.

Durante il processo di elaborazione, la componente grafica permette al modello di aggregare le informazioni provenienti dai nodi vicini, identificando eventuali correlazioni spaziali tra differenti punti di osservazione. Successivamente, una componente temporale analizza l'evoluzione delle informazioni raccolte, permettendo di individuare pattern e dipendenze presenti nella sequenza temporale.

In questo modo, una ST-GNN non considera più ogni serie temporale come indipendente dalle altre, ma interpreta il fenomeno come un sistema interconnesso nel quale l'evoluzione futura di un nodo può essere influenzata sia dal proprio andamento passato sia dalle informazioni provenienti dai nodi spazialmente correlati.

GNN e RNN

Un approccio alternativo consiste nel combinare una GNN con una rete neurale ricorrente (RNN), separando maggiormente la modellazione della componente spaziale da quella temporale.

In questo caso, la GNN viene utilizzata per estrarre le caratteristiche spaziali presenti nella struttura a grafo, aggregando le informazioni provenienti dai nodi vicini. Le informazioni ottenute vengono successivamente fornite a una rete ricorrente, come LSTM o GRU, incaricata di modellare l'evoluzione temporale della variabile.

Questo approccio permette di sfruttare separatamente i punti di forza delle due architetture.

Nel presente lavoro è stata considerata la possibilità di utilizzare questo tipo di approccio sebbene le ST-GNN rappresentino una soluzione più generale per la modellazione congiunta delle due dimensioni.

Tale scelta è stata determinata principalmente dalla maggiore disponibilità di implementazioni e lavori di riferimento relativi ad architetture GNN combinate con reti ricorrenti, che hanno permesso una più semplice analisi della metodologia e una maggiore disponibilità di esempi applicativi.

Inoltre, trattandosi di una tecnica introdotta recentemente nell'ambito dell'analisi svolta, l'utilizzo di un'architettura maggiormente documentata ha permesso di valutare più agevolmente la possibilità di applicazione al dataset considerato.

4.5 Tecniche di decomposizione avanzata delle serie temporali

Un'ulteriore strategia utilizzata nella previsione delle serie temporali consiste nella decomposizione del segnale originale in un insieme di componenti caratterizzate da differenti proprietà temporali. L'idea alla base di questi approcci è che una serie osservata possa essere composta da molteplici oscillazioni sovrapposte, ciascuna associata a differenti scale temporali e livelli di complessità.

A differenza dei modelli basati sulla decomposizione esplicita di componenti come trend e stagionalità, come nel caso di Prophet, le tecniche di decomposizione avanzata non assumono a priori una struttura specifica della serie, ma cercano di estrarre automaticamente le componenti che costituiscono il segnale osservato.

La decomposizione permette di separare il segnale in componenti più semplici, che possono essere successivamente modellate attraverso algoritmi predittivi differenti. In particolare, molte delle tecniche proposte in letteratura combinano metodi di decomposizione del segnale con reti neurali, sfruttando la capacità di queste ultime di apprendere relazioni non lineari presenti nelle singole componenti estratte.

Tra gli approcci maggiormente utilizzati rientrano le tecniche basate su Empirical Mode Decomposition (EMD) e le sue evoluzioni, come Ensemble Empirical Mode Decomposition (EEMD) e Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). Questi metodi vengono spesso integrati con reti neurali ricorrenti, come LSTM e BiLSTM, dando origine ad architetture ibride in grado di combinare la capacità di analizzare le diverse componenti del segnale con quella di apprendere dipendenze temporali complesse.

4.5.1 EEMD-ALSTM

EEMD-ALSTM è un modello avanzato di previsione per serie storiche che unisce tre tecniche: la scomposizione dei dati con rumore assistito (EEMD), una rete neurale a memoria a lungo-breve termine (LSTM) e un meccanismo di attenzione (Attention). Viene usato per

analizzare dati complessi e non lineari (come l'inquinamento atmosferico o l'energia eolica) con alta precisione.

Uno dei limiti più importanti di alcuni dei modelli visti, come l'LSTM, è il fatto che, in casi di inquinanti paerticolarmente irregolari, come il particolato, fanno grande difficoltà a seguirne l'andamento, poiché la rete deve apprendere contemporaneamente tutte le componenti presenti nel segnale originale, rendendo più complessa l'estrazione delle informazioni utili.

Per affrontare questo problema, viene applicata una fase preliminare di decomposizione del segnale attraverso la tecnica Ensemble Empirical Mode Decomposition (EEMD). La EEMD deriva dalla tecnica EMD (Empirical Mode Decomposition), un metodo sviluppato per analizzare segnali caratterizzati da comportamento non lineare e non stazionario.

L'idea alla base della decomposizione è trasformare una serie temporale complessa in un insieme di componenti più semplici, chiamate Intrinsic Mode Functions (IMF). Ogni componente IMF rappresenta una diversa scala temporale presente nel segnale originale.

Il vantaggio principale è che la rete neurale non deve più modellare direttamente un segnale altamente complesso, ma può apprendere separatamente componenti caratterizzate da dinamiche più regolari. La tecnica EMD presenta, però, un problema chiamato *mode mixing*, ovvero il fenomeno di sovrapposizione di oscillazioni appartenenti a scale temporali differenti all'interno della stessa componente.

La EEMD cerca di ridurre questo problema introducendo una procedura basata sull'aggiunta di rumore bianco gaussiano al segnale originale. Oltre alla decomposizione, come abbiamo già detto, il modello proposto integra un meccanismo di attenzione (Attention Mechanism). L'obiettivo dell'attention è permettere alla rete di identificare quali parti della sequenza temporale contengano informazioni maggiormente rilevanti per la previsione, assegnando pesi diversi agli elementi in sequenza.

Il modello finale proposto dagli autori prende il nome di EEMD-ALSTM.

4.5.2 EMD-BiLSTM

L'EMD-BiLSTM è un modello ibrido che combina una tecnica di decomposizione delle serie temporali con una rete neurale ricorrente bidirezionale.

In particolare, il modello integra la *Empirical Mode Decomposition* (EMD) con una rete *Bidirectional Long Short-Term Memory* (BiLSTM), impiegata per apprendere le dipendenze temporali presenti nelle componenti estratte.

L'idea alla base di questo approccio consiste nel considerare la serie temporale della metrica considerata come un segnale composto da diverse oscillazioni sovrapposte, caratterizzate da differenti frequenze e scale temporali.

Formalmente, una serie temporale può essere rappresentata come:

$$x(t) = \sum_{i=1}^{n} IMF_i(t) + r(t) \tag{4.5.1}$$

dove $IMF_i(t)$ rappresenta la i-esima componente oscillatoria estratta dalla decomposizione, mentre $r(t)$ rappresenta il residuo finale della serie.

L'obiettivo della decomposizione è quindi trasformare una serie temporale complessa in un insieme di sottosegnali caratterizzati da una maggiore semplicità. In questo modo la rete neurale può modellare separatamente le diverse componenti ottenute.

La procedura EMD è basata su un processo iterativo denominato *sifting process*. Il procedimento può essere riassunto nei seguenti passaggi:

1. vengono individuati i massimi e i minimi locali del segnale originale;

2. attraverso un'interpolazione spline vengono costruiti l'inviluppo superiore e l'inviluppo inferiore del segnale;
3. viene calcolato il valore medio dei due inviluppi;
4. la componente media viene sottratta dal segnale originale;
5. il procedimento viene ripetuto fino all'ottenimento di una componente oscillatoria che soddisfa le condizioni necessarie per essere considerata una IMF.

La prima componente estratta rappresenta generalmente le oscillazioni caratterizzate dalla frequenza più elevata. Successivamente, tale componente viene rimossa dal segnale e il procedimento viene applicato nuovamente al residuo ottenuto, permettendo di estrarre componenti associate a frequenze progressivamente inferiori.

Dopo la decomposizione, ciascuna componente viene utilizzata come ingresso della rete BiLSTM, che sfrutta le informazioni temporali contenute nelle sequenze per produrre le previsioni relative alle singole componenti. La previsione finale viene successivamente ricostruita combinando i contributi delle diverse IMF e del residuo.

4.5.3 CEEMDAN

La *Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (CEEMDAN) è una tecnica avanzata di decomposizione delle serie temporali sviluppata per analizzare segnali caratterizzati da comportamento non lineare e non stazionario. La decomposizione ottenuta può essere rappresentata come la precedente, in quanto entrambi i metodi fanno uso dell'EMD.

CEEMDAN deriva, infatti, dall'evoluzione di due tecniche precedenti: la *Empirical Mode Decomposition* (EMD) e la *Ensemble Empirical Mode Decomposition* (EEMD). L'obiettivo principale di queste evoluzioni è quello di migliorare la stabilità della decomposizione e ridurre alcuni problemi presenti nel metodo originale.

Sappiamo che, nella pratica, la EMD presenta alcune limitazioni. Un primo problema è rappresentato dal fenomeno del *mode mixing*, ovvero la presenza all'interno della stessa IMF di oscillazioni caratterizzate da frequenze differenti. Questo può rendere difficile l'interpretazione delle componenti estratte, poiché fenomeni distinti possono risultare combinati all'interno dello stesso sottosegnale.

Il problema risulta particolarmente evidente in presenza di: rumore, variazioni improvvise del segnale e transizioni rapide tra differenti condizioni del fenomeno analizzato.

Un ulteriore limite riguarda la possibilità che la somma delle componenti ottenute non ricostruisca perfettamente il segnale originale, introducendo una perdita di informazione durante il processo di decomposizione.

Per ridurre il problema del *mode mixing* è stata introdotta la tecnica EEMD (*Ensemble Empirical Mode Decomposition*). L'idea alla base consiste nell'aggiungere al segnale originale differenti realizzazioni di rumore bianco e applicare successivamente la decomposizione EMD a ciascuna versione del segnale.

Tuttavia, anche la EEMD presenta alcuni limiti, tra cui:

- Una ricostruzione del segnale non sempre perfetta.
- La possibile presenza di residui del rumore introdotto.
- Una dipendenza dei risultati dal numero di decomposizioni effettuate e dai parametri scelti.

La CEEMDAN introduce un ulteriore miglioramento rispetto alla EEMD attraverso un'iniezione del rumore adattiva durante il processo di decomposizione. Questo permette di ottenere una decomposizione più stabile e una migliore separazione delle componenti del segnale.

I principali vantaggi introdotti dalla CEEMDAN sono: una maggiore separazione delle componenti oscillanti, una maggiore stabilità rispetto alla decomposizione EMD originale e una migliore capacità di ricostruzione del segnale iniziale.

CAPITOLO 5

Applicazione di tecniche innovative al dataset di riferimento

Nel capitolo precedente sono stati introdotti e descritti gli approcci innovativi considerati per la previsione degli inquinanti atmosferici, evidenziandone il funzionamento, le caratteristiche principali e le motivazioni alla base del loro utilizzo.

In questo capitolo vengono presentati gli esperimenti effettuati applicando tali tecniche al dataset di riferimento. Per ciascun modello analizzato vengono descritte la configurazione adottata, le modalità di applicazione al problema considerato e i risultati ottenuti in termini di capacità predittiva.

L'obiettivo di questa fase sperimentale è valutare l'efficacia dei diversi approcci nel contesto specifico della previsione degli inquinanti atmosferici, analizzando come le differenti caratteristiche dei modelli possano influenzare le prestazioni ottenute.

I risultati vengono confrontati attraverso le metriche di valutazione definite, al fine di evidenziare vantaggi e limitazioni dei diversi metodi considerati. Infine, viene effettuata un'analisi complessiva dei risultati ottenuti, con l'obiettivo di individuare gli approcci maggiormente promettenti per il problema analizzato e comprendere i fattori che ne influenzano le prestazioni.

Gli esperimenti descritti nel presente capitolo sono stati condotti considerando differenti intervalli temporali in funzione della tecnica analizzata. Per i modelli Prophet e NeuralProphet è stato utilizzato il periodo compreso tra luglio 2025 e marzo 2026, mentre per i successivi approcci, basati su modelli ibridi, reti neurali grafiche e tecniche di decomposizione, è stato possibile utilizzare un intervallo temporale più esteso, comprendente i dati fino a giugno 2026.

5.1 Applicazione dei modelli di decomposizione

Il primo esperimento svolto ha riguardato l'applicazione di Prophet ai dati per valutare quanto questo metodo possa essere in grado di generare previsioni coerenti con i valori reali. Si è poi proseguito valutando NeuralProphet e, quindi, analizzando come l'aggiunta di una rete neurale avrebbe influito sui risultati.

I modelli proposti, così come gli altri descritti nel presente capitolo, sono stati applicati alle serie temporali relative alle centraline SMART137 e SMART58 considerando come variabili target CO_2 e PM10. A differenza degli studi svolti sui modelli ARIMAX e SARIMA, non è presente una libreria per la selezione automatica dei valori migliori dei parametri per la modellazione di Prophet e NeuralProphet e, pertanto, gli studi si sono sviluppati in diverse fasi, culminate in una ricerca mediante gridsearch della migliore configurazione dei parametri.

5.1.1 Prophet

L'applicazione del modello è stata effettuata considerando diverse configurazioni sperimentali, con l'obiettivo di valutare sia il comportamento del modello, quando addestrato su intervalli temporali limitati, sia la capacità di sfruttare l'intera storia disponibile della serie temporale. Sono state considerate due differenti modalità di addestramento, ovvero:

- La prima modalità, definita analisi mensile, in cui ogni mese valido viene considerato indipendentemente dagli altri. Il modello viene addestrato utilizzando le osservazioni disponibili nella prima parte del mese, mentre gli ultimi cinque giorni vengono esclusi e utilizzati come periodo di test.
- La seconda modalità dove, invece, il modello viene addestrato utilizzando l'intera storia temporale disponibile per la serie considerata. Anche in questo caso gli ultimi cinque giorni dell'ultimo mese completo vengono mantenuti come insieme di valutazione.

Per poter essere utilizzata da Prophet, ogni serie temporale viene convertita nel formato richiesto dalla libreria, costituito da una colonna temporale *ds* e dalla variabile da prevedere *y*.

Il modello è addestrato esclusivamente sulla porzione di serie temporale definita come training set e, successivamente, utilizzato per generare la previsione relativa all'intervallo temporale di test.

Prima dell'addestramento è stata effettuata una fase di controllo della qualità dei dati, finalizzata ad individuare gli intervalli temporali utilizzabili per la modellazione. In particolare, per ogni combinazione centralina-variabile, è stata analizzata la percentuale di valori mancanti presenti nei singoli mesi. Sono stati mantenuti soltanto i mesi caratterizzati da una quantità limitata di dati mancanti, fissando una soglia massima pari all'1% delle osservazioni.

Tale scelta deriva dalla capacità di Prophet di gestire piccole irregolarità nella serie temporale, mentre la presenza di intervalli estesi, privi di osservazioni, potrebbe compromettere la stima delle componenti del modello. Non avendo inizialmente una conoscenza approfondita del comportamento del modello, si è scelto di partire da una sua configurazione standard e volutamente semplice, così da ottenere una prima baseline di riferimento, riportata nel Listato 5.1.

```
model = Prophet(
    daily_seasonality=True,
    weekly_seasonality=True,
    yearly_seasonality=False
)
```

Listing 5.1: Configurazione iniziale del modello Prophet.

È stata disabilitata la stagionalità annuale perché, come già detto, la limitata estensione temporale del dataset non ne permette una stima affidabile.

Prophet mensile

Viene sfruttata la prima modalità di addestramento e vengono testate tutte le combinazioni centralina-variabile di interesse. Per ogni combinazione:

1. viene estratta la serie;
2. vengono trovati i mesi validi;
3. per ogni mese:

4. (a) viene creato train/test;
   (b) viene addestrato Prophet;
   (c) viene effettuata la previsione;
   (d) vengono calcolate le metriche.

La previsione viene confrontata con gli ultimi 5 giorni del mese. I risultati ottenuti per le diverse combinazioni di centralina e variabile sono sintetizzati nella Figura 5.1.

	sensor	metric	month	n_test_points	mae	rmse	mape
0	SMART137	PM10	2025-11	120	35.413288	40.342986	395.513453
1	SMART137	PM10	2025-12	120	66.748780	77.099256	512.352495
2	SMART137	PM10	2026-01	120	16.958117	21.288122	130.660964
3	SMART137	PM10	2026-03	119	7.422280	9.336752	138.067818
4	SMART137	CO2	2025-11	120	66.301836	69.164446	15.873793
5	SMART137	CO2	2025-12	120	31.966924	37.053173	7.331314
6	SMART137	CO2	2026-01	120	26.074185	28.177838	6.246243
7	SMART137	CO2	2026-03	119	9.724298	12.679668	2.332613
8	SMART58	PM10	2025-10	120	8.607932	11.131636	73.136811
9	SMART58	PM10	2025-11	120	16.242371	19.181168	126.850281
10	SMART58	PM10	2025-12	120	47.041005	52.278022	367.995235
11	SMART58	PM10	2026-01	120	23.703370	27.999566	128.658284
12	SMART58	PM10	2026-03	119	4.144278	5.919264	50.961874
13	SMART58	CO2	2025-10	120	20.920058	29.702641	4.676874
14	SMART58	CO2	2026-01	120	26.481675	28.950688	6.311057
15	SMART58	CO2	2026-03	119	10.523076	12.758634	2.515535

Figura 5.1: Risultati ottenuti da Prophet nella configurazione mensile.

Per la configurazione che ha ottenuto le migliori prestazioni sono stati analizzati anche i grafici relativi al periodo di previsione. Nelle Figure 5.2 e 5.3 vengono riportati, rispettivamente, il dettaglio dell'intervallo previsto e l'andamento completo del mese.

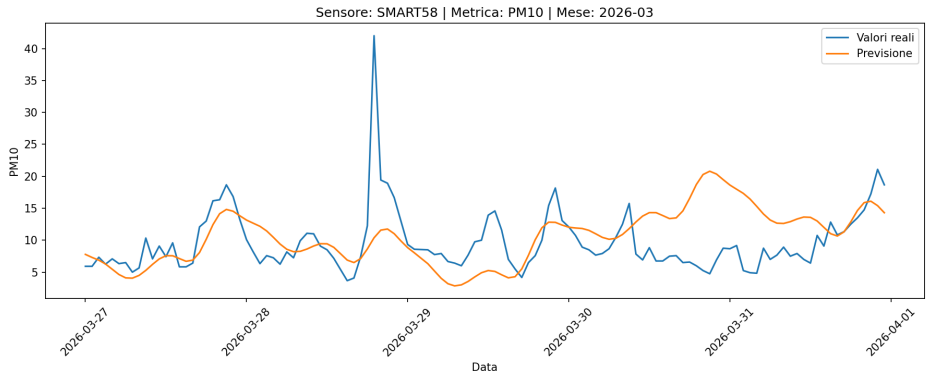


Figura 5.2: Previsione di Prophet sugli ultimi cinque giorni del mese di test nella configurazione mensile.

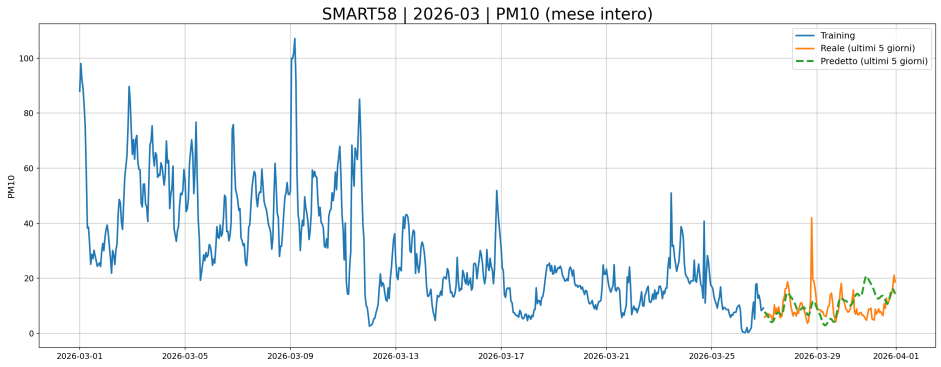


Figura 5.3: Andamento completo del mese di test previsto da Prophet nella configurazione mensile.

Configurazione sull'intero periodo disponibile

Nella seconda configurazione, il modello viene addestrato utilizzando l'intero periodo temporale disponibile per la serie considerata, anziché analizzare separatamente ciascun mese. Gli ultimi cinque giorni dell'ultimo mese completo vengono esclusi dalla fase di addestramento e utilizzati come insieme di test.

In questo modo Prophet può sfruttare l'intera storia disponibile, non più i singoli mesi, per apprendere le caratteristiche della serie e per generare la previsione relativa al periodo di test. Le prestazioni ottenute nelle diverse combinazioni centralina-variabile sono riportate nella Figura 5.4.

	sensor	metric	last_month	n_test_points	mae	rmse	mape
0	SMART137	PM10	2026-03	119	7.701711	9.405008	151.872000
1	SMART137	CO2	2026-03	119	12.696437	16.183367	3.003554
2	SMART58	PM10	2026-03	119	15.026626	15.950138	195.463884
3	SMART58	CO2	2026-03	119	12.930322	15.380701	3.068542

Figura 5.4: Risultati ottenuti da Prophet utilizzando l'intero periodo storico disponibile per l'addestramento.

A completamento della valutazione quantitativa, per la configurazione con le migliori prestazioni in Figura 5.5 viene mostrato il grafico rappresentante la predizione degli ultimi cinque giorni del mese di marzo.

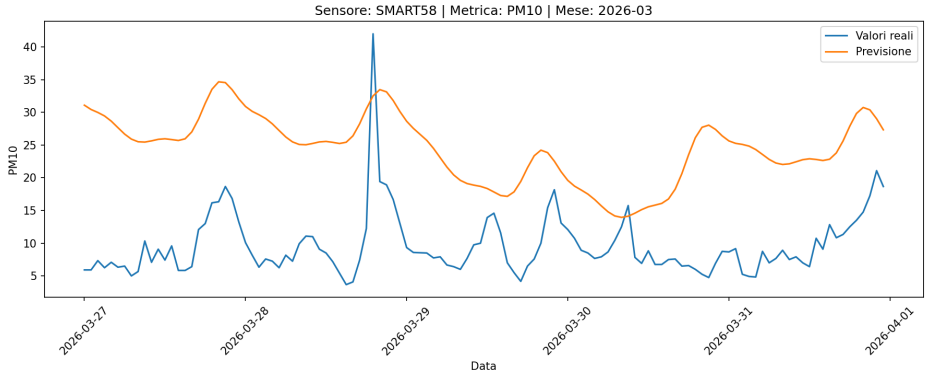


Figura 5.5: Previsione degli ultimi cinque giorni del periodo di test ottenuta con Prophet addestrato sull'intera serie temporale disponibile.

Prophet grid search

Dopo una prima valutazione effettuata utilizzando una configurazione standard di Prophet, è stata introdotta una fase di ottimizzazione dei parametri mediante grid search.

La ricerca ha riguardato alcuni parametri in grado di influenzare il comportamento del modello; nello specifico:

- *changepoint prior scale*, che controlla la flessibilità del modello nell'adattamento ai cambiamenti del trend.
- *seasonality prior scale*, che regola l'importanza delle componenti stagionali.
- *seasonality mode*, che permette di scegliere tra una modellazione additiva o moltiplicativa della stagionalità.

Per ogni combinazione di parametri è stato addestrato un modello Prophet. La configurazione caratterizzata dal valore minimo della metrica RMSE è stata selezionata come migliore configurazione per ciascuna serie analizzata.

Prima di procedere con la ricerca degli iperparametri, è utile considerare la configurazione di default di Prophet, ovvero quella ottenuta quando i parametri analizzati nella grid search non vengono specificati esplicitamente. Tale configurazione è riportata nel Listato 5.2.

```
model = Prophet(
    growth="linear",
    changepoint_prior_scale=0.05,
    n_changepoints=25,
    changepoint_range=0.8,
    yearly_seasonality=False,
    weekly_seasonality=True,
    daily_seasonality=True,
    seasonality_mode="additive",
    seasonality_prior_scale=10.0,
    holidays_prior_scale=10.0
)
```

Listing 5.2: Configurazione iniziale del modello Prophet utilizzata negli esperimenti.

Partendo da ciò abbiamo effettuato una grid search sulle combinazioni riportate nel Listato 5.3.

```
prophet_param_grid = [
    {
        "changepoint_prior_scale": 0.001,
        "seasonality_prior_scale": 0.1,
        "seasonality_mode": "additive"
    },
    {
        "changepoint_prior_scale": 0.01,
        "seasonality_prior_scale": 1.0,
        "seasonality_mode": "additive"
    },
    {
        "changepoint_prior_scale": 0.05,
        "seasonality_prior_scale": 1.0,
        "seasonality_mode": "additive"
    },
    {
        "changepoint_prior_scale": 0.1,
```

```
        "seasonality_prior_scale": 10.0,
        "seasonality_mode": "additive"
    },
    {
        "changepoint_prior_scale": 0.5,
        "seasonality_prior_scale": 10.0,
        "seasonality_mode": "additive"
    },
    {
        "changepoint_prior_scale": 0.01,
        "seasonality_prior_scale": 1.0,
        "seasonality_mode": "multiplicative"
    },
    {
        "changepoint_prior_scale": 0.05,
        "seasonality_prior_scale": 10.0,
        "seasonality_mode": "multiplicative"
    }
]
```

Listing 5.3: Configurazioni dei parametri analizzate mediante grid search per Prophet.

La Figura 5.6 sintetizza i risultati ottenuti dalle diverse configurazioni analizzate durante la grid search, permettendo di confrontare le prestazioni delle combinazioni di iperparametri considerate.

	changepoint_prior_scale	seasonality_prior_scale	seasonality_mode	n_series	mae_mean	rmse_mean	mape_mean	mae_std
0	0.05	1.0	additive	4	20.292411	23.651002	208.733441	7.344351
1	0.05	10.0	additive	4	20.297018	23.651960	207.786695	7.172342
2	0.01	1.0	additive	4	20.864377	24.651439	220.567126	8.619041
3	0.05	10.0	multiplicative	4	22.302244	25.690987	227.911187	9.611281
4	0.10	10.0	additive	4	23.170433	26.688638	227.984687	7.547534

Figura 5.6: Confronto delle prestazioni ottenute dalle configurazioni analizzate mediante grid search per Prophet.

La configurazione risultata migliore per ciascuna combinazione centralina-variabile è stata successivamente utilizzata per generare i grafici relativi al periodo di previsione, riportati nelle Figure 5.7, 5.8, 5.9 e 5.10.

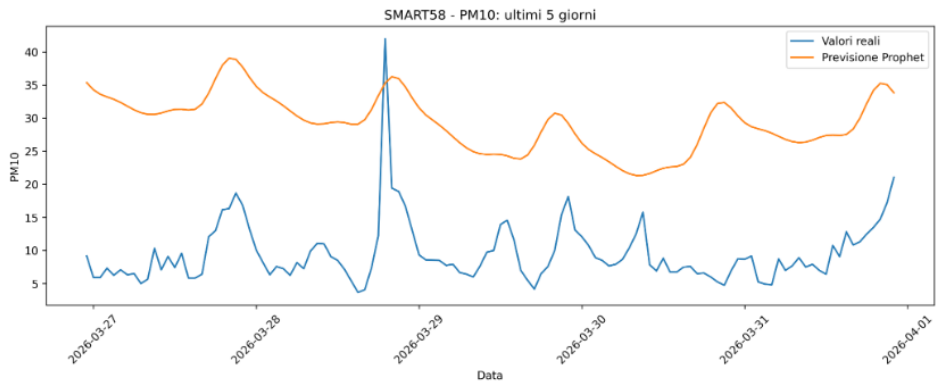


Figura 5.7: Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART58-PM10 ottenuta dalla migliore configurazione Prophet individuata tramite grid search.

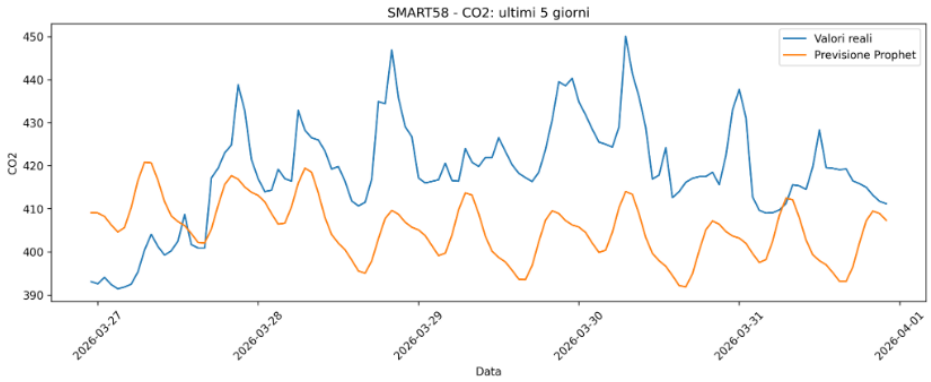


Figura 5.8: Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART58-CO_2 ottenuta dalla migliore configurazione Prophet individuata tramite grid search.

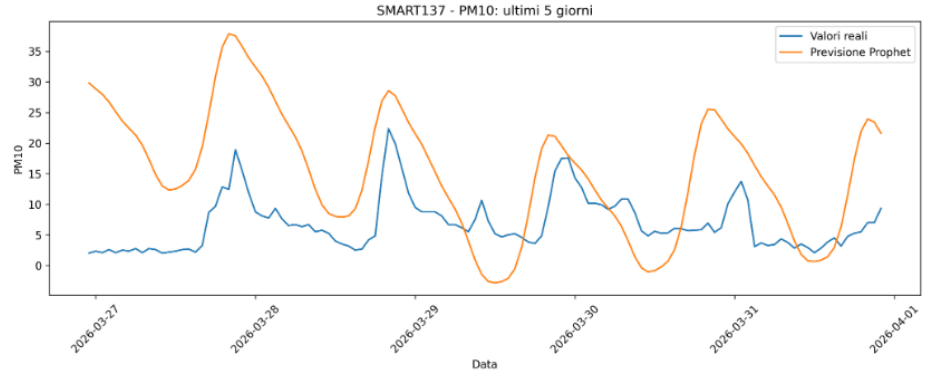


Figura 5.9: Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART137-PM10 ottenuta dalla migliore configurazione Prophet individuata tramite grid search.

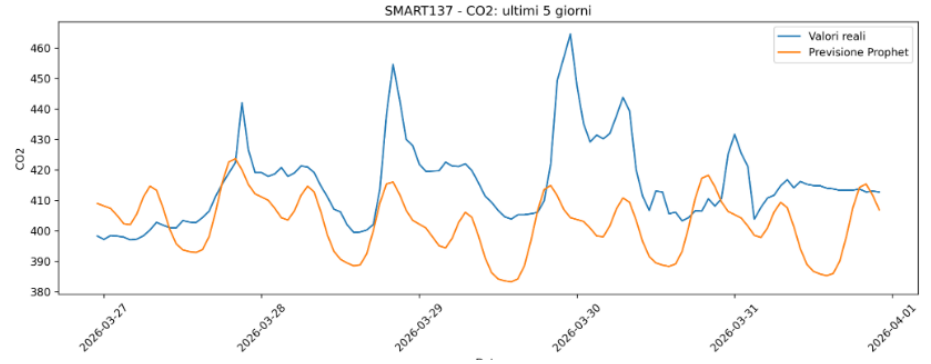


Figura 5.10: Previsione degli ultimi cinque giorni del periodo di test per la combinazione SMART137-CO_2 ottenuta dalla migliore configurazione Prophet individuata tramite grid search.

5.1.2 NeuralProphet

Si è, dunque, passati all'applicazione di NeuralProphet mantenendo attive le componenti di trend e stagionalità già utilizzate per Prophet e introducendo una componente autoregressiva attraverso il parametro `n_lags`. I dati sono stati convertiti nel formato richiesto dalla libreria, costituito dalle colonne temporali *ds* e dal valore osservato y.

Per NeuralProphet è stata utilizzata direttamente l'intera storia temporale disponibile per ciascuna combinazione centralina-variabile, ovvero la seconda modalià di addestramento, poichè la rete neurale richiede un numero maggiore di dati per apprendere correttamente il

comportamento della serie temporale.

La configurazione iniziale adottata è caratterizzata dai seguenti parametri principali:

- *n_lags* = 48: indica il numero di osservazioni precedenti utilizzate dalla componente autoregressiva. Considerando la frequenza oraria dei dati, il modello utilizza, quindi, le ultime 48 ore della serie temporale come informazione aggiuntiva per generare la previsione.
- *learning rate*=0.1: rappresenta il passo utilizzato durante l'ottimizzazione dei parametri della rete neurale.
- *epochs*=150: indica il numero di iterazioni effettuate durante il processo di addestramento.
- *trend_reg*=1.0: controlla la regolarizzazione applicata alla componente di trend, limitando eventuali variazioni eccessivamente rapide.

I valori inizialmente assegnati agli iperparametri sono stati scelti come configurazione di riferimento per una prima valutazione del modello, prima di procedere con la ricerca mediante grid search. Essa viene riportata nel Listato 5.4:

```
model = NeuralProphet(
    growth="linear",
    yearly_seasonality=False,
    weekly_seasonality=True,
    daily_seasonality=True,
    seasonality_mode="additive",
    n_lags=48,
    learning_rate=0.1,
    epochs=150,
    trend_reg=1.0,
    drop_missing=True
)
```

Listing 5.4: Configurazione iniziale del modello NeuralProphet utilizzata negli esperimenti.

La presenza della componente autoregressiva introduce un vincolo aggiuntivo rispetto al modello Prophet. Per costruire gli ingressi della rete neurale, infatti, NeuralProphet necessita di sequenze temporali consecutive, caratterizzate dalla presenza dei valori precedenti. Di conseguenza, la gestione dei valori mancanti assume un ruolo particolarmente importante, poiché intervalli privi di osservazioni possono impedire la costruzione delle finestre temporali necessarie all'addestramento.

Per la gestione dei valori mancanti sono state usate due strategie, caratterizzate dall'uso, o dal mancato uso, dell'interpolazione. Il confronto tra i risultati ottenuti con le due strategie è riportato nel Listato 5.11, mentre una descrizione più approfondita delle stesse sarà presentata nei prossimi paragrafi.

NeuralProphet senza interpolazione

Nel primo esperimento sono stati utilizzati esclusivamente i dati originali disponibili, senza applicare alcuna procedura di ricostruzione dei valori mancanti.

In questa configurazione sono stati mantenuti soltanto i periodi temporali caratterizzati da una percentuale di dati mancanti pari a zero, così da garantire la continuità della sequenza richiesta dalla componente autoregressiva del modello.

Tra le configurazioni analizzate, la combinazione SMART137-PM10 ha ottenuto le migliori prestazioni. Nelle Figure 5.12, 5.13 e 5.14 vengono riportati i grafici relativi il dettaglio

	modello	sensor	metric	last_month	train_size	test_size	mae	rmse	mape	coverage_90
0	NeuralProphet senza forward fill	SMART137	PM10	2025-12	624	120	36.451979	51.079624	194.071933	28.333333
1	NeuralProphet senza forward fill	SMART137	CO2	2025-12	624	120	23.325809	30.098771	5.264517	38.333333
2	NeuralProphet senza forward fill	SMART58	PM10	2026-01	1368	120	27.131266	31.307728	374.967789	50.000000
3	NeuralProphet senza forward fill	SMART58	CO2	2026-01	624	120	16.204710	20.395667	3.843896	23.333333
4	NeuralProphet con forward fill	SMART137	PM10	2026-01	2088	120	36.109804	41.147418	647.639346	38.333333
5	NeuralProphet con forward fill	SMART137	CO2	2026-01	2088	120	14.471459	18.494474	3.492714	80.833333
6	NeuralProphet con forward fill	SMART58	PM10	2026-01	2832	120	34.998251	39.209868	445.719101	45.000000
7	NeuralProphet con forward fill	SMART58	CO2	2026-03	624	120	19.898570	24.678487	4.817072	17.500000

	modello	mae_medio	rmse_medio	mape_medio	coverage_90_media
0	NeuralProphet con forward fill	26.369521	30.882562	275.417058	45.416667
1	NeuralProphet senza forward fill	25.778441	33.220448	144.537034	35.000000

Figura 5.11: Confronto delle prestazioni ottenute da NeuralProphet con e senza interpolazione dei valori mancanti.

dell'intervallo previsto, i grafici relativi allo stesso periodo ma completo di intervallo di confidenza e l'andamento completo del mese.

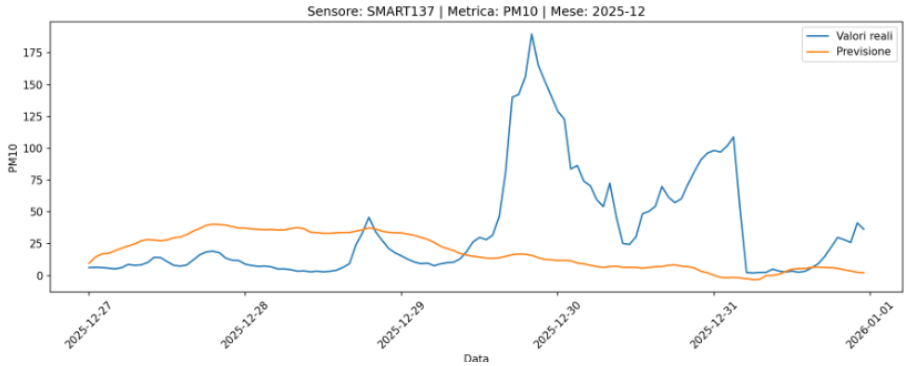


Figura 5.12: Dettaglio della previsione di NeuralProphet sull'intervallo di test.

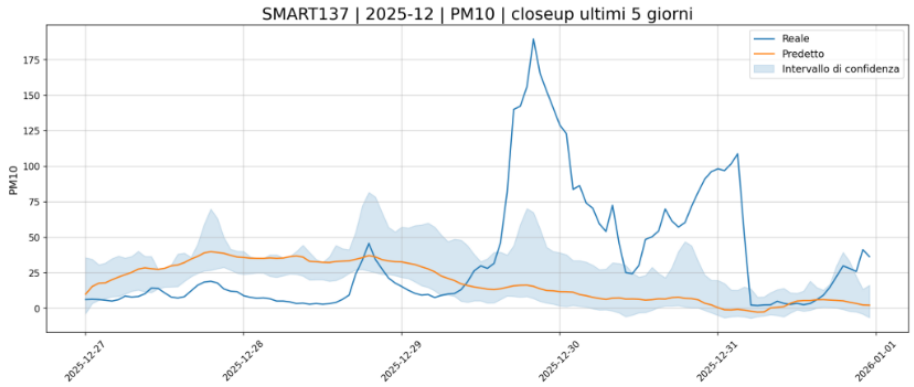


Figura 5.13: Dettaglio della previsione di NeuralProphet con il relativo intervallo di confidenza.

NeuralProphet con interpolazione

Successivamente è stata valutata una configurazione nella quale è stata introdotta una procedura di riempimento dei valori mancanti, limitata ai soli intervalli brevi.

Nello specifico, sono stati individuati i periodi caratterizzati dalla presenza di lacune temporali non superiori a tre ore consecutive. Tali valori sono stati ricostruiti mediante forward fill, escludendo gli intervalli caratterizzati da assenze di dati più estese.

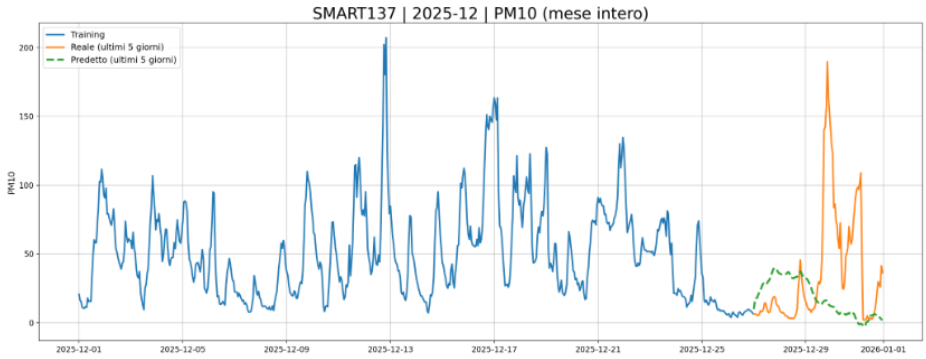


Figura 5.14: Andamento completo del mese di test previsto da NeuralProphet.

Tale scelta deriva dalla scarsa disponibilità di dati conformi alle limitazioni imposte precedentemente. Anche in questa configurazione, la combinazione SMART137-PM10 ha fornito i risultati più favorevoli. I relativi grafici di previsione sono riportati nelle Figure 5.15, 5.16 e 5.17.

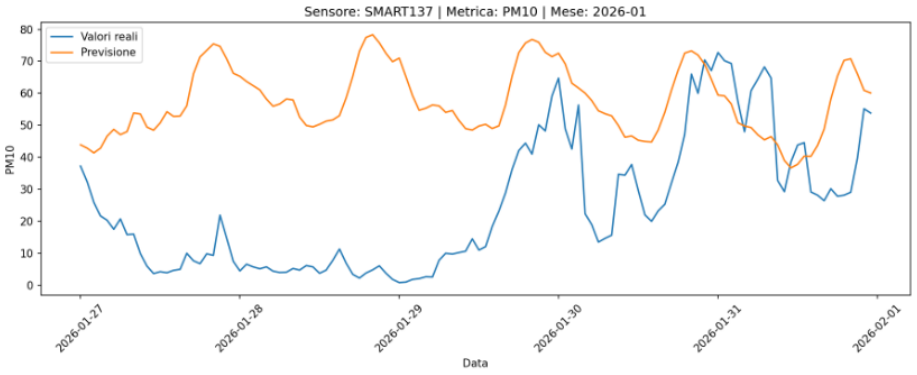


Figura 5.15: Dettaglio della previsione di NeuralProphet sull'intervallo di test usando l'interpolazione.

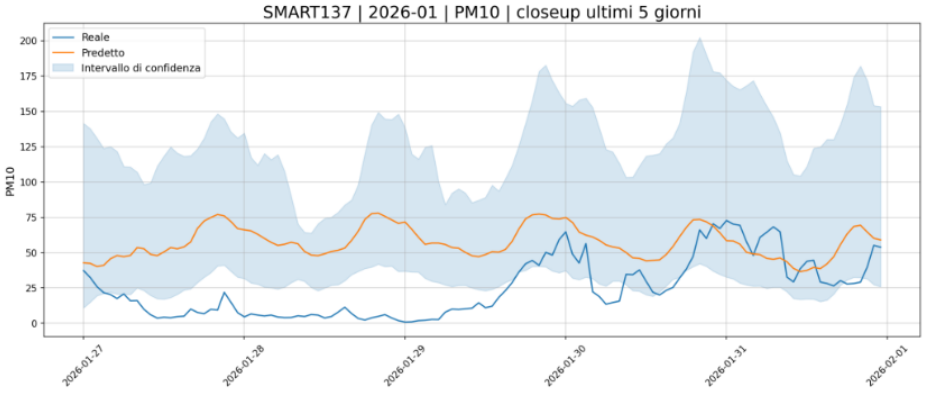


Figura 5.16: Dettaglio della previsione di NeuralProphet con il relativo intervallo di confidenza usando l'interpolazione.

NeuralProphet grid search

Dopo una prima valutazione del modello mediante una configurazione iniziale, è stata effettuata una ricerca sistematica degli iperparametri attraverso una procedura di grid search.

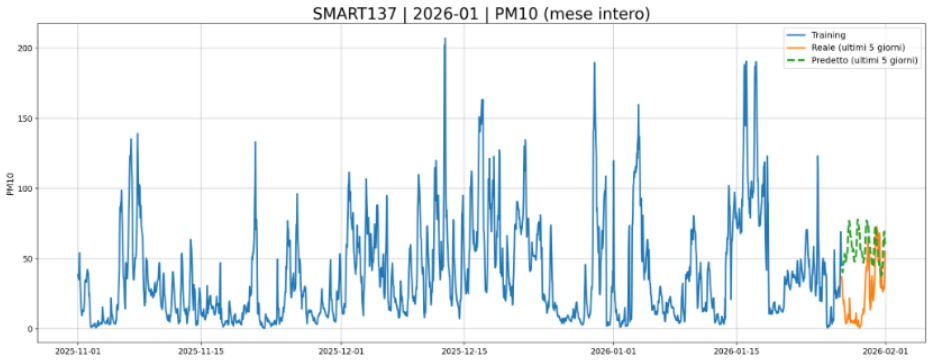


Figura 5.17: Andamento completo del mese di test previsto da NeuralProphet usando l'interpolazione.

La ricerca ha riguardato i parametri maggiormente influenti sul comportamento della componente autoregressiva e sul processo di addestramento della rete neurale. Più nello specifico, nel Listato 5.5 vengono mostrate le configurazioni prese in considerazione.

```
param_grid = [
    {"n_lags": 24, "learning_rate": 0.01,  "epochs": 100, "trend_reg":
       1.0},
    {"n_lags": 24, "learning_rate": 0.01,  "epochs": 100, "trend_reg":
       0},
    {"n_lags": 24, "learning_rate": 0.01,  "epochs": 100, "trend_reg":
       0.5},
    {"n_lags": 24, "learning_rate": 0.005, "epochs": 150, "trend_reg":
       1.0},
    {"n_lags": 24, "learning_rate": 0.1,   "epochs": 150, "trend_reg":
       1.0},
    {"n_lags": 24, "learning_rate": 0.05,  "epochs": 150, "trend_reg":
       1.0},
    {"n_lags": 48, "learning_rate": 0.01,  "epochs": 100, "trend_reg":
       1.0},
    {"n_lags": 48, "learning_rate": 0.005, "epochs": 150, "trend_reg":
       1.0},
    {"n_lags": 48, "learning_rate": 0.1,   "epochs": 150, "trend_reg":
       1.0},
    {"n_lags": 48, "learning_rate": 0.05,  "epochs": 150, "trend_reg":
       1.0}
]
```

Listing 5.5: Configurazioni analizzate mediante grid search per NeuralProphet.

I risultati ottenuti mediante la ricerca effettuata con la tecnica di grid search, riportati in Figura 5.18 hanno permesso di individuare, tra le configurazioni analizzate, quelle caratterizzate dalle migliori prestazioni. Nelle Figura 5.19, 5.20, 5.21 e 5.22 vengono riportati i risultati delle configurazioni migliori individuate rispettivamente in termini di MAE, RMSE e MAPE.

5.2 Applicazione dei modelli ibridi

Dopo aver analizzato le prestazioni ottenute mediante Prophet e NeuralProphet, è stato implementato un ulteriore approccio ibrido basato sulla combinazione tra Prophet e una rete neurale LSTM.

	n_lags	learning_rate	epochs	trend_reg	mean_rank	mean_relative_rmse	max_relative_rmse	mean_rmse	mean_mae	mean_mape	n_series
0	24	0.100	150	1.0	2.25	1.023151	1.083694	19.953067	16.709633	128.318766	4
1	24	0.050	150	1.0	2.75	1.149775	1.341498	22.513684	19.059737	175.292069	4
2	24	0.005	150	1.0	3.00	1.382702	2.390098	28.734898	25.246550	255.273570	4
3	48	0.050	150	1.0	3.50	1.226670	1.539443	24.258687	20.508067	192.366594	4
4	48	0.100	150	1.0	4.75	1.200383	1.278086	23.510749	19.531990	160.431369	4
5	48	0.005	150	1.0	5.50	1.599189	2.128802	31.286669	27.503933	238.734263	4
6	24	0.010	100	1.0	6.75	1.846830	2.551621	36.116250	32.511587	264.182157	4
7	48	0.010	100	1.0	7.50	1.942877	2.634239	37.739019	33.687305	276.992769	4

Figura 5.18: Valori ottenuti nelle metriche di valutazione considerate per le configurazioni di NeuralProphet testate.

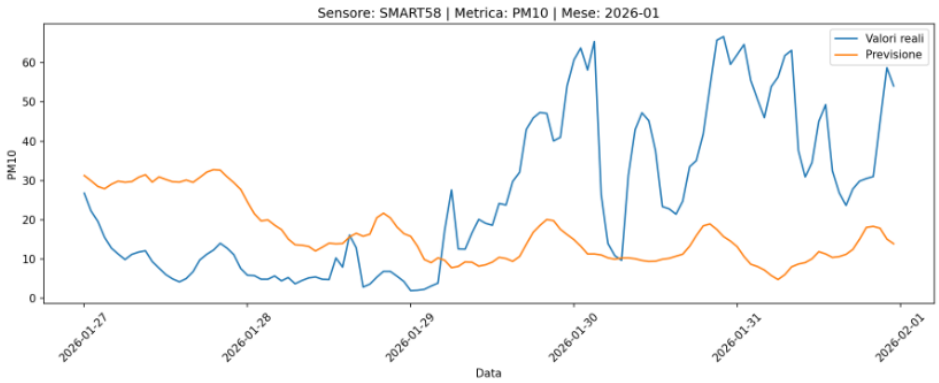


Figura 5.19: Previsione degli ultimi cinque giorni del periodo di test per SMART58-PM10 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search.

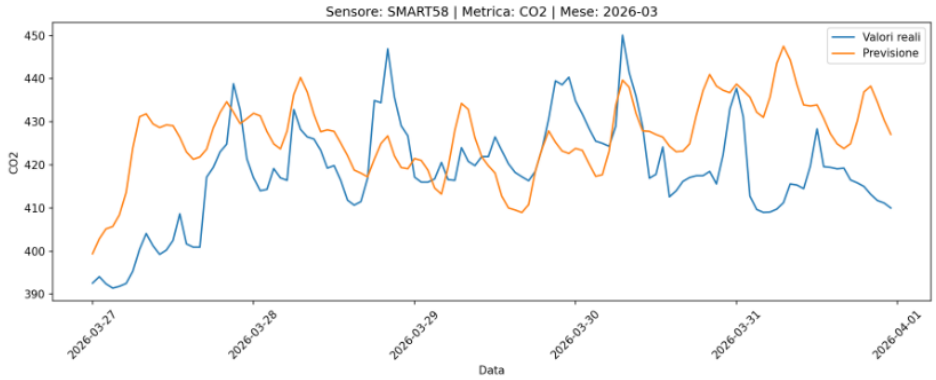


Figura 5.20: Previsione degli ultimi cinque giorni del periodo di test per SMART58-CO_2 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search.

L'idea alla base di questo modello consiste nello sfruttare le caratteristiche complementari dei due approcci: Prophet viene utilizzato per modellare la struttura generale della serie temporale, attraverso l'analisi delle componenti di trend e stagionalità, mentre la rete LSTM viene impiegata per apprendere le dinamiche temporali più complesse non catturate dal modello iniziale.

La strategia adottata non prevede l'utilizzo diretto della LSTM per la previsione della variabile di interesse, ma introduce un approccio basato sulla previsione dei residui. La previsione finale viene ottenuta combinando la stima prodotta da Prophet con una correzione appresa dalla rete neurale.

La prima fase dell'approccio ibrido consiste nell'addestramento del modello Prophet sulla serie temporale relativa al periodo di training. La configurazione utilizzata mantiene le impostazioni definite negli esperimenti precedenti, considerando una componente stagionale

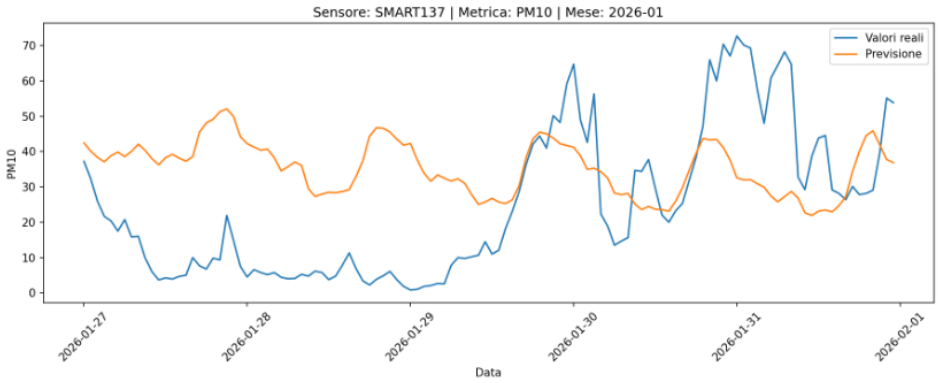


Figura 5.21: Previsione degli ultimi cinque giorni del periodo di test per SMART137-PM10 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search.

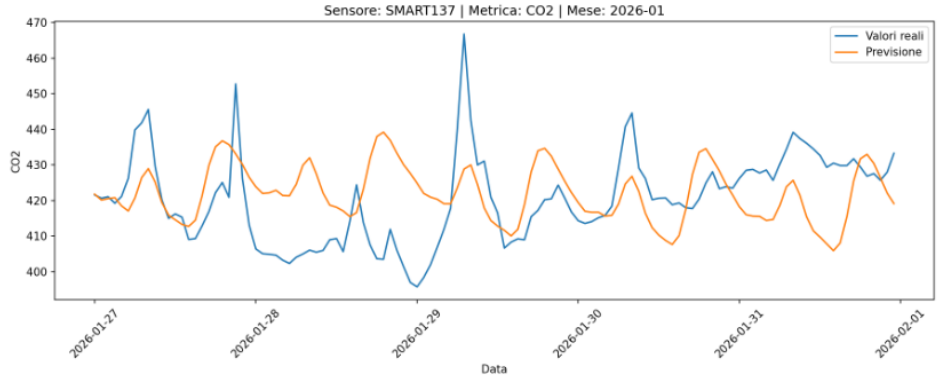


Figura 5.22: Previsione degli ultimi cinque giorni del periodo di test per SMART137-CO_2 ottenuta dalla migliore configurazione NeuralProphet individuata tramite grid search.

giornaliera e settimanale, mentre la stagionalità annuale viene esclusa a causa della limitata disponibilità temporale del dataset.

Una volta ottenuta la previsione di Prophet sul periodo di addestramento, sono stati calcolati i residui della serie temporale secondo la relazione:

$$r(t) = y(t) - \hat{y}_{Prophet}(t) \tag{5.2.1}$$

dove $y(t)$ rappresenta il valore osservato e $\hat{y}_{Prophet}(t)$ la previsione ottenuta dal modello Prophet.

I residui così ottenuti rappresentano la componente della serie temporale non spiegata dal modello decompositivo e vengono utilizzati come ingresso per la rete LSTM. Per la costruzione degli esempi di addestramento, la sequenza dei residui viene trasformata mediante una finestra temporale mobile. Sono state considerate le ultime 48 osservazioni disponibili per prevedere direttamente l'intero orizzonte futuro relativo al periodo di test. Al termine dell'addestramento della rete LSTM, la previsione finale viene ottenuta sommando la componente stimata da Prophet con la correzione prevista dalla rete neurale:

$$\hat{y}(t) = \hat{y}_{Prophet}(t) + \hat{r}_{LSTM}(t) \tag{5.2.2}$$

dove $\hat{y}_{Prophet}(t)$ rappresenta la previsione ottenuta mediante Prophet, mentre $\hat{r}_{LSTM}(t)$ rappresenta la previsione del residuo effettuata dalla rete LSTM.

I risultati ottenuti dalle diverse combinazioni di centralina e variabile sono sintetizzati nella Figura 5.23.

	model	sensor	metric	last_month	train_size	test_size	input_window	forecast_horizon	prophet_mae	prophet_rmse	prophet_mape	hybrid_mae	hybrid_rmse	hybrid_mape
0	Prophet_LSTM_residual	SMART137	PM10	2026-06	2808	120	48	120	2.907298	5.388735	41.179747	2.555988	4.934717	34.664543
1	Prophet_LSTM_residual	SMART137	CO2	2026-06	2808	120	48	120	10.988583	13.723719	2.524113	10.665337	13.442100	2.467952
2	Prophet_LSTM_residual	SMART58	PM10	2026-06	2808	120	48	120	2.006154	2.490004	24.970829	2.093274	2.565573	26.288281
3	Prophet_LSTM_residual	SMART58	CO2	2026-06	2808	120	48	120	15.071598	17.915364	3.598846	15.562082	18.720914	3.716401

Figura 5.23: Risultati ottenuti dall'approccio ibrido Prophet-LSTM.

La configurazione che ha ottenuto le migliori prestazioni è stata successivamente analizzata attraverso i grafici relativi al periodo di previsione. In particolare, nelle Figure 5.24, 5.25 e 5.26 vengono riportati il dettaglio dell'intervallo previsto, il relativo intervallo di confidenza e l'andamento complessivo del periodo di test.

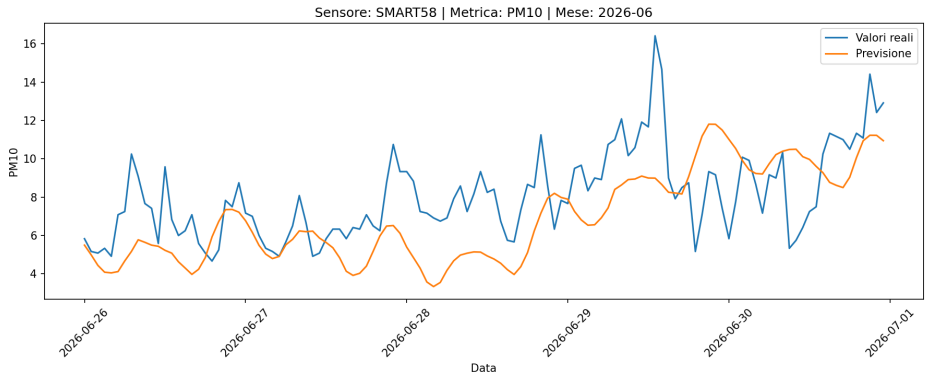


Figura 5.24: Dettaglio della previsione degli ultimi cinque giorni del periodo di test ottenuta dall'approccio Prophet-LSTM per la combinazione SMART58-PM10.

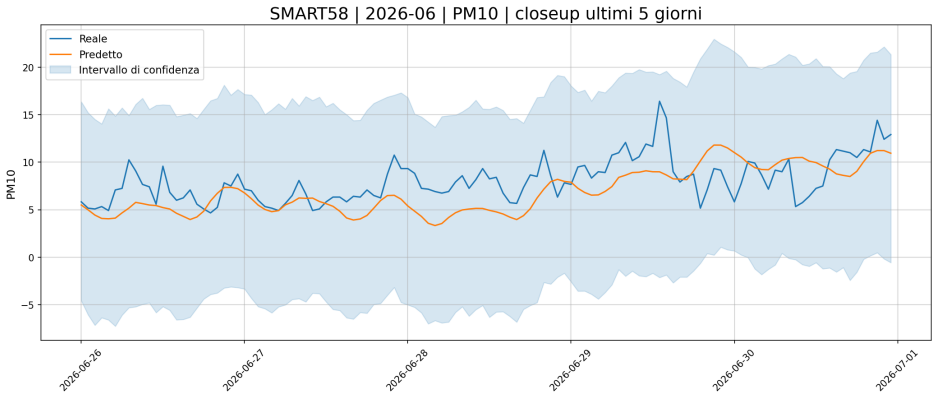


Figura 5.25: Dettaglio della previsione degli ultimi cinque giorni del periodo di test con relativo intervallo di confidenza ottenuta dall'approccio Prophet-LSTM per la combinazione SMART58-PM10.

5.3 Applicazione dei modelli basati su reti neurali basate su grafi

L'applicazione delle reti neurali basate su grafi è stata sviluppata progressivamente introducendo, inizialmente, un modello caratterizzato esclusivamente dalla componente spaziale, per poi integrare una componente temporale in grado di modellare l'evoluzione della serie.

Questo approccio incrementale ha permesso, dapprima, di valutare l'effettiva utilità delle informazioni spaziali, derivanti dalla presenza di più centraline di monitoraggio, e

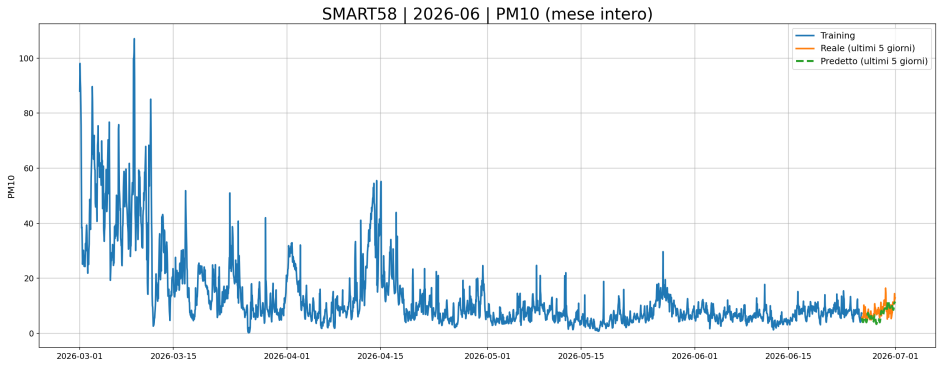


Figura 5.26: Andamento complessivo del mese di test previsto dall'approccio Prophet-LSTM per la combinazione SMART58-PM10.

successivamente di analizzare il contributo delle dinamiche temporali all'interno del modello predittivo.

Un aspetto che distingue l'applicazione delle GNN dagli approcci precedentemente analizzati riguarda il numero di centraline considerate. Nei modelli precedenti l'analisi era stata concentrata sulle centraline SMART58 e SMART137, mentre la struttura a grafo richiede di considerare l'insieme dei nodi, e quindi di tutte le stazioni, che compongono il grafo.

5.3.1 GNN

La prima fase dell'applicazione delle reti neurali basate su grafi ha riguardato la definizione della struttura del grafo utilizzato dal modello. Nel contesto analizzato, le centraline di monitoraggio sono state rappresentate mediante nodi, mentre le connessioni tra esse sono state definite sulla base della distanza geografica. La distanza tra le centraline è stata calcolata utilizzando le informazioni relative alla latitudine e alla longitudine associate a ciascun sensore.

Per evitare la generazione di connessioni tra punti eccessivamente distanti, è stata introdotta una soglia massima (*threshold*) pari a 3 km: due centraline vengono, cioè, collegate soltanto quando la loro distanza geografica risulta inferiore a tale valore.

In questo modo il modello considera esclusivamente le relazioni spaziali tra stazioni potenzialmente influenzate dagli stessi fenomeni locali, ottenendo una rappresentazione del problema mediante una GNN spaziale, che prende il nome di GCN (Graph Convolutional Network).

La GCN viene usata in questo caso poichè, grazie al processo di convoluzione grafica, ciascun nodo aggiorna la propria rappresentazione combinando le informazioni associate al nodo stesso con quelle provenienti dai nodi adiacenti. Di conseguenza, la previsione relativa ad una specifica centralina non dipende esclusivamente dalla propria storia temporale, ma anche dall'evoluzione delle stazioni connesse nel grafo.

Dall'analisi della struttura del grafo generato, riportata in Figura 5.27, è stato possibile individuare una criticità relativa alla centralina SMART136, identificata come nodo isolato e, pertanto, non collegato alle altre centraline presenti.

Una successiva analisi delle coordinate geografiche associate alla centralina, e una ricerca sul sito di AirQuino, hanno evidenziato un'anomalia nella configurazione della posizione, in quanto il sensore risultava localizzato in un'area differente rispetto alle altre stazioni del dataset, più precisamento nell'area urbana di Firenze.

Per tale motivo, la centralina SMART136 non è stata considerata nella costruzione finale del grafo, evitando di introdurre una relazione spaziale non coerente con il problema analizzato.

```
Grado uscente: [3, 2, 3, 0, 2]
Grado entrante: [3, 2, 3, 0, 2]
Grado totale: [6, 4, 6, 0, 4]
Numero sensori isolati: 1
Sensori isolati:
SMART136

Archi del grafo:
SMART31 -> SMART57 | distanza = 0.793 km
SMART31 -> SMART58 | distanza = 1.000 km
SMART31 -> SMART137 | distanza = 2.049 km
SMART57 -> SMART31 | distanza = 0.793 km
SMART57 -> SMART58 | distanza = 0.356 km
SMART58 -> SMART31 | distanza = 1.000 km
SMART58 -> SMART57 | distanza = 0.356 km
SMART58 -> SMART137 | distanza = 2.097 km
SMART137 -> SMART31 | distanza = 2.049 km
SMART137 -> SMART58 | distanza = 2.097 km
```

Figura 5.27: Struttura del grafo generato.

Per una prima valutazione della tecnica, si è scelto di considerare una sola variabile target, individuata nella concentrazione di PM10. La rete viene utilizzata per prevedere l'evoluzione futura della variabile considerando contemporaneamente:

- la storia temporale della centralina analizzata;
- le informazioni provenienti dalle centraline collegate nel grafo.

Le prestazioni ottenute dalla GCN per le diverse centraline sono sintetizzate nella Figura 5.28.

	model	sensor	metric	last_month	train_size	test_size	input_window	forecast_horizon	MAE	RMSE	MAPE	error_std_empirical
0	GNN_GCN	SMART31	PM10	2026-06	1344	120	24	120	2.738734	4.465054	27.694803	3.202745
1	GNN_GCN	SMART57	PM10	2026-06	1344	120	24	120	1.712539	2.185519	25.196537	2.327718
2	GNN_GCN	SMART58	PM10	2026-06	1344	120	24	120	2.600253	3.344019	28.047007	2.913201
3	GNN_GCN	SMART137	PM10	2026-06	1344	120	24	120	2.650819	4.942365	30.492672	2.378300

Figura 5.28: Risultati ottenuti dalla GCN considerando la sola componente spaziale.

La combinazione centralina-variabile caratterizzata dalle migliori prestazioni è stata successivamente analizzata attraverso la rappresentazione grafica delle previsioni. I relativi risultati sono riportati nelle Figure 5.29, 5.30 e 5.31.

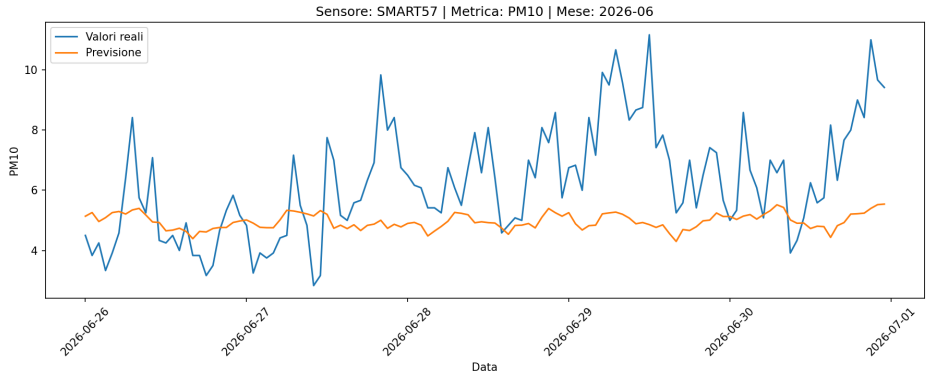


Figura 5.29: Dettaglio della previsione ottenuta dalla GCN sull'intervallo di test.

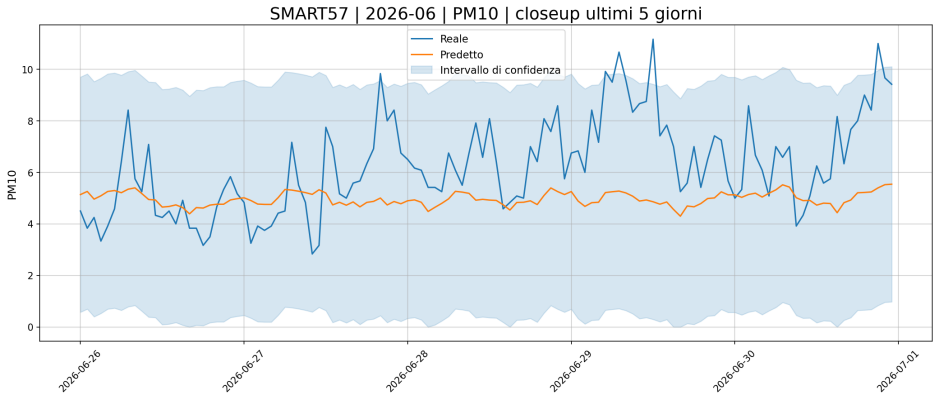


Figura 5.30: Dettaglio della previsione ottenuta dalla GCN con relativo intervallo di confidenza.

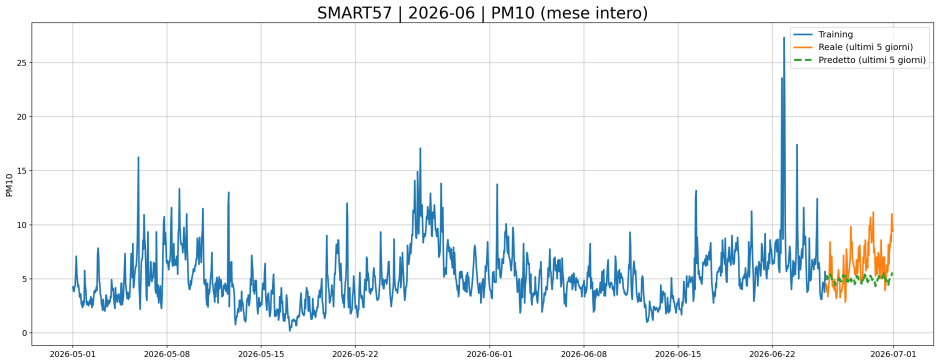


Figura 5.31: Andamento completo del periodo di test previsto dalla GCN.

5.3.2 GNN e RNN

I risultati ottenuti mediante la sola GCN hanno mostrato il limite di un approccio in grado di modellare esclusivamente le relazioni spaziali. Per migliorare la capacità predittiva del modello è stata introdotta una componente ricorrente, con l'obiettivo di affiancare alla modellazione delle relazioni spaziali l'analisi dell'evoluzione temporale delle variabili considerate.

Tuttavia, le RNN tradizionali possono presentare problemi durante l'addestramento quando la sequenza temporale diventa estesa, principalmente a causa dei fenomeni di *vanishing gradient* ed *exploding gradient*. Questi problemi possono limitare la capacità della rete di apprendere informazioni relative a intervalli temporali lontani.

Per questo motivo è stata utilizzata una Gated Recurrent Unit (GRU), una variante delle reti ricorrenti progettata per migliorare la gestione delle dipendenze temporali attraverso l'introduzione di due meccanismi di controllo, ovvero:

- *update gate*: decide quanta informazione passata mantenere;
- *reset gate*: decide quanta informazione passata ignorare quando viene elaborato un nuovo ingresso.

Ciò è preferibile anche rispetto ad una LSTM, poichè ne semplifica la struttura e diminuisce la complessità computazionale. L'applicazione di questa architettura ha permesso di valutare il contributo dell'integrazione tra informazione spaziale e temporale nella previsione della concentrazione di PM10. Nel modello sviluppato, la GCN viene, quindi, utilizzata per estrarre le caratteristiche spaziali dal grafo, mentre la GRU riceve tali informazioni e ne analizza l'evoluzione temporale.

A differenza del primo esperimento, nel quale era stata considerata esclusivamente la concentrazione di PM10, in questa configurazione è stato analizzato anche il CO_2. Il modello è stato applicato mantenendo la struttura del grafo precedentemente definita e considerando, per ciascuna variabile, le centraline appartenenti al grafo.

I risultati ottenuti dalla combinazione GCN-GRU sono riportati nella Figura 5.32.

	sensor	metric	last_month	test_size	input_window	forecast_horizon	MAE	RMSE	MAPE
0	SMART31	PM10	2026-06	120	48	120	2.371364	4.012190	27.040917
1	SMART57	PM10	2026-06	120	48	120	1.690286	2.164277	25.707749
2	SMART58	PM10	2026-06	120	48	120	1.943694	2.592394	22.435400
3	SMART137	PM10	2026-06	120	48	120	2.668041	4.940849	31.815857
4	SMART31	CO2	2026-06	120	48	120	18.258963	20.269964	4.250599
5	SMART57	CO2	2026-06	120	48	120	9.888468	11.859608	2.394905
6	SMART58	CO2	2026-06	120	48	120	14.720062	16.585570	3.421967
7	SMART137	CO2	2026-06	120	48	120	12.744538	17.732782	2.872222

Figura 5.32: Risultati ottenuti dall'approccio combinato GCN-GRU.

Per valutare graficamente il comportamento del modello è stata analizzata la configurazione che ha ottenuto le migliori prestazioni tra le combinazioni di centralina e variabile considerate. I relativi grafici di previsione sono riportati nelle Figure 5.33, 5.34 e 5.35.

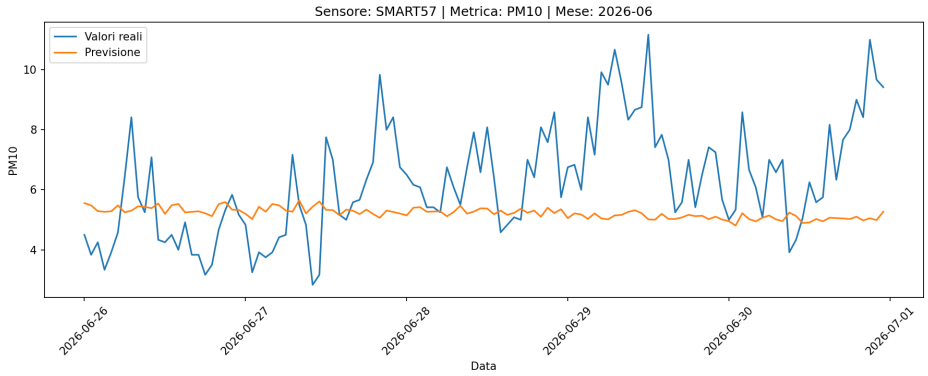


Figura 5.33: Dettaglio della previsione ottenuta dall'approccio GCN-GRU sull'intervallo di test.

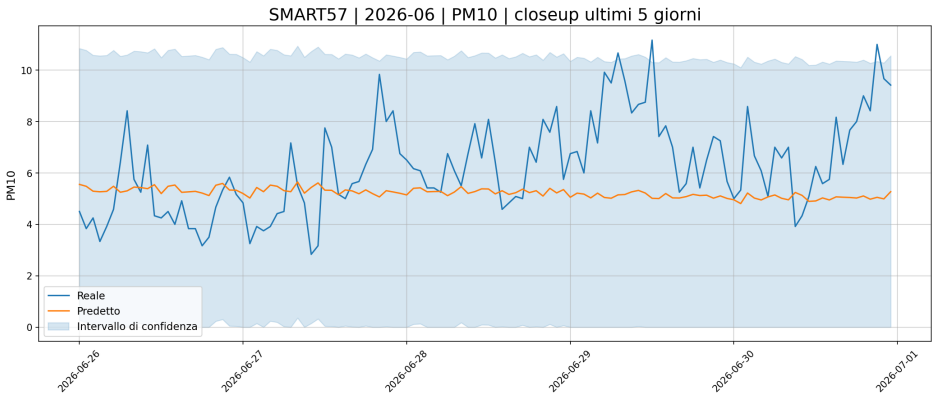


Figura 5.34: Dettaglio della previsione ottenuta dall'approccio GCN-GRU con relativo intervallo di confidenza.

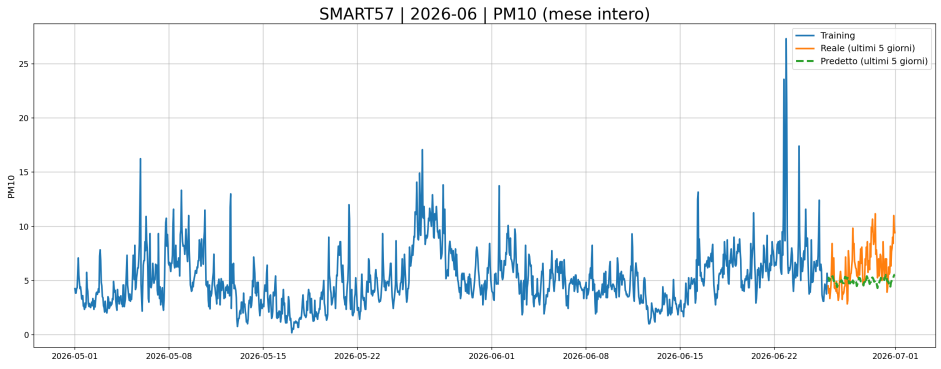


Figura 5.35: Andamento completo del periodo di test previsto dall'approccio GCN-GRU.

5.4 Applicazione delle tecniche di decomposizione

Anche l'applicazione delle tecniche di decomposizione è stata sviluppata introducendo progressivamente diversi livelli di complessità del modello. Inizialmente è stata valutata la capacità predittiva di una rete neurale ricorrente applicata direttamente alla serie temporale, utilizzando una BiLSTM come modello di riferimento.

In seguito, per verificare se una diversa rappresentazione del segnale potesse migliorare le prestazioni ottenute, è stata introdotta la tecnica CEEMDAN, utilizzata per scomporre la serie originale in componenti caratterizzate da differenti frequenze e complessità. Le componenti ottenute vengono, poi, utilizzate come ingresso per modelli predittivi dedicati.

5.4.1 BiLSTM ed EMD-BiLSTM

Per l'applicazione delle tecniche di decomposizione è stata scelta una rete Bidirectional Long Short-Term Memory (BiLSTM) come modello predittivo di riferimento. La scelta è stata effettuata considerando anche gli approcci presenti nella letteratura scientifica relativa alla previsione degli inquinanti atmosferici mediante tecniche di decomposizione avanzata.

Infatti, numerosi lavori basati su tecniche come EMD, EEMD e CEEMDAN utilizzano reti ricorrenti, e in particolare architetture BiLSTM, per modellare le componenti ottenute dalla decomposizione.

Come primo esperimento è stata valutata una configurazione costituita dalla sola BiLSTM e, come già descritto in precedenti approcci, la previsione è stata formulata come un problema multistep diretto. In particolare, il modello utilizza una finestra temporale di 48 osservazioni orarie per prevedere simultaneamente i successivi 120 valori della serie, corrispondenti a cinque giorni.

La rete BiLSTM utilizzata è costituita da due livelli ricorrenti, in particolare:

- il primo livello contiene 64 unità e restituisce una rappresentazione della sequenza;
- il secondo livello è costituito da 32 unità e produce una rappresentazione più compatta delle informazioni temporali.

La rappresentazione ottenuta viene utilizzata dai livelli densi per generare direttamente l'intero orizzonte di previsione.

La configurazione basata sulla sola BiLSTM costituisce il punto di riferimento per il successivo esperimento EMD-BiLSTM. In quest'ultimo caso, la serie temporale originale viene preliminarmente scomposta mediante EMD nelle relative Intrinsic Mode Functions (IMF) e nella componente residua. Le singole componenti ottenute vengono modellate separatamente

mediante BiLSTM e le previsioni generate vengono successivamente ricombinate per ottenere la previsione finale della serie originale.

I risultati ottenuti dall'approccio EMD-BiLSTM nelle diverse combinazioni di centralina e variabile sono riportati nella Figura 5.36.

	model	sensor	metric	last_month	test_size	input_window	forecast_horizon	MAE	RMSE	MAPE
0	EMD_BiLSTM	SMART137	PM10	2026-06	120	72	120	2.641305	4.589500	39.303338
1	EMD_BiLSTM	SMART58	PM10	2026-06	120	72	120	2.046251	2.775140	23.273249
2	EMD_BiLSTM	SMART137	CO2	2026-06	120	72	120	11.493186	14.606682	2.681553
3	EMD_BiLSTM	SMART58	CO2	2026-06	120	72	120	11.379427	14.775610	2.716622

Figura 5.36: Risultati ottenuti dall'approccio EMD-BiLSTM nelle diverse combinazioni centralina-variabile.

La combinazione centralina-variabile che ha ottenuto le migliori prestazioni è stata successivamente analizzata attraverso i grafici relativi al periodo di previsione, riportati nelle Figure 5.37, 5.38 e 5.39.

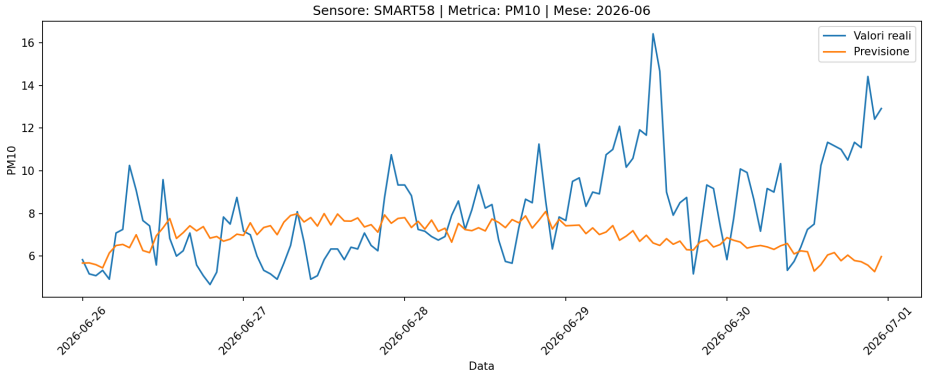


Figura 5.37: Dettaglio della previsione ottenuta dall'approccio EMD-BiLSTM sull'intervallo di test.

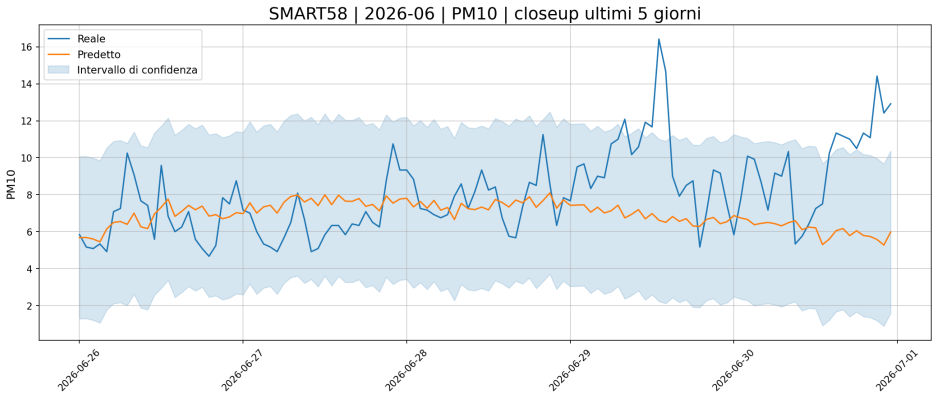


Figura 5.38: Dettaglio della previsione ottenuta dall'approccio EMD-BiLSTM con relativo intervallo di confidenza.

5.4.2 CEEMDAN-BiLSTM

A seguito degli esperimenti descritti, è stata introdotta la tecnica CEEMDAN (*Complete Ensemble Empirical Mode Decomposition with Adaptive Noise*) con l'obiettivo di trasformare la serie temporale originale in un insieme di componenti più semplici da modellare.

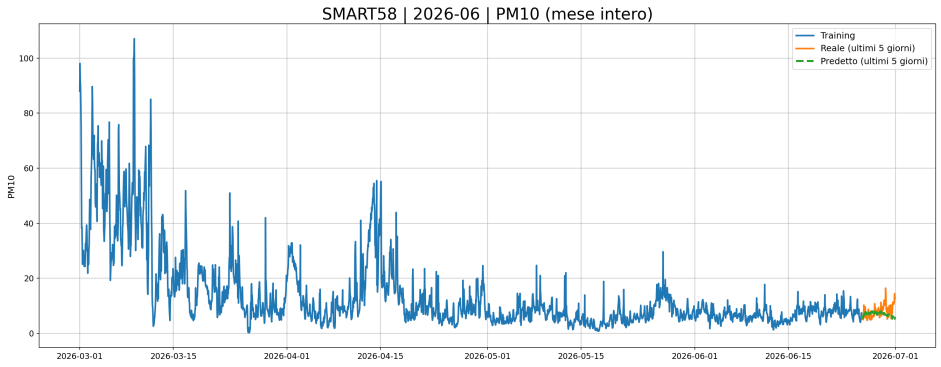


Figura 5.39: Andamento completo del periodo di test previsto dall'approccio EMD-BiLSTM.

La decomposizione CEEMDAN permette di rappresentare la serie originale come una combinazione di diverse Intrinsic Mode Functions (IMF) e di una componente residua che descrive generalmente la componente di fondo della serie, cioè ciò che rimane dopo aver rimosso le componenti oscillatorie individuate dalle IMF.

Dopo la decomposizione, ciascuna componente ottenuta viene utilizzata separatamente per l'addestramento di una rete BiLSTM. L'obiettivo è permettere alla rete neurale di apprendere singolarmente le diverse dinamiche presenti nel segnale, evitando di dover modellare direttamente una serie caratterizzata dalla presenza contemporanea di trend, oscillazioni e rumore.

Le previsioni ottenute dalle singole componenti vengono, infine, ricombinate per ottenere la previsione finale della serie temporale originale.

I risultati ottenuti mediante la decomposizione CEEMDAN e la successiva previsione con BiLSTM sono sintetizzati nella Figura 5.40.

	sensor	metric	model	n_components	n_test_points	mae	rmse	r2
0	SMART137	PM10	CEEMDAN_BiLSTM	6	120	7.430886	9.292714	-3.702481
1	SMART58	PM10	CEEMDAN_BiLSTM	6	120	4.923181	6.808948	-1.020656
2	SMART137	CO2	CEEMDAN_BiLSTM	6	120	12.399186	15.289212	-0.297613
3	SMART58	CO2	CEEMDAN_BiLSTM	6	120	10.717938	13.801579	-0.377772

Figura 5.40: Risultati ottenuti dall'approccio CEEMDAN-BiLSTM.

La configurazione che ha prodotto le migliori prestazioni è stata ulteriormente analizzata attraverso la rappresentazione grafica della previsione. Il risultato relativo al periodo di test è riportato nella Figura 5.41.

5.5 Valutazione dei risultati

Dopo aver completato la fase di studio e applicazione dei vari metodi per la predizione degli inquinanti atmosferici, è stato possibile analizzare concretamente quanto ottenuto.

È importante evidenziare che non per tutte le metodologie sviluppate sono stati riportati tutti i grafici relativi al periodo di previsione. Infatti, per alcuni approcci considerati come fase preliminare dello studio, quali Prophet e la GNN, non sono stati generati tutti i dettagli grafici comprendenti il confronto sugli ultimi cinque giorni, l'intervallo di confidenza e l'andamento completo del mese.

Tale scelta è stata effettuata considerando i risultati quantitativi ottenuti nelle prime valutazioni. Poiché tali modelli non hanno mostrato prestazioni sufficientemente promettenti, si è ritenuto non necessario dedicare ulteriori risorse computazionali alla generazione completa

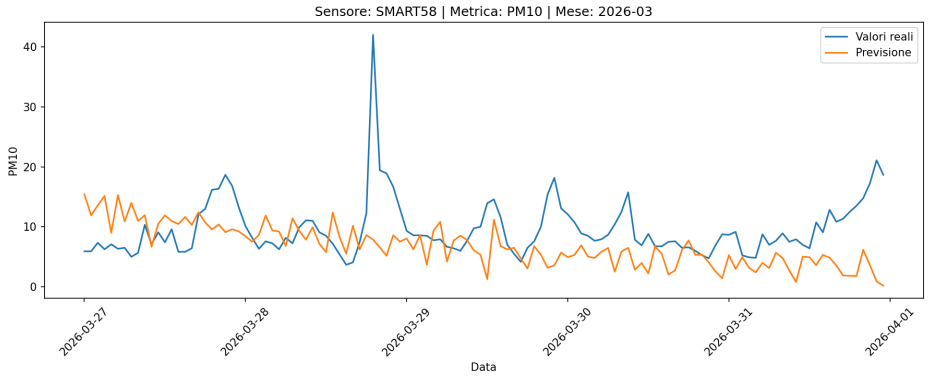


Figura 5.41: Previsione ottenuta dall'approccio CEEMDAN-BiLSTM sull'intervallo di test.

delle rappresentazioni grafiche, concentrando, invece, l'analisi sui modelli che risultano, a livello teorico, maggiormente interessanti.

I risultati ottenuti hanno evidenziato, però, come gli approcci più complessi non abbiano necessariamente garantito un miglioramento rispetto ai modelli di riferimento. In particolare, alcune delle metodologie più innovative hanno mostrato difficoltà nel modellare correttamente l'andamento delle serie temporali considerate.

Per quanto riguarda Prophet, il modello ha mostrato una limitata capacità di seguire le variazioni della serie temporale osservata. Le previsioni ottenute risultano poco sensibili alle variazioni improvvise della concentrazione degli inquinanti e non riescono a riprodurre adeguatamente i picchi presenti nei dati reali.

Un comportamento analogo è stato osservato anche per NeuralProphet, nonostante l'introduzione della componente autoregressiva basata su rete neurale. In questo caso sono stati osservati risultati incoraggianti per la variabile CO_2, nella quale l'andamento previsto risulta più coerente con quello osservato. Tuttavia, tale miglioramento non si è riscontrato nella previsione del PM10, per il quale il modello continua a mostrare difficoltà nell'individuazione dei picchi di concentrazione.

Anche l'approccio ibrido Prophet-LSTM non ha mostrato il miglioramento atteso. Ciò suggerisce che le limitazioni non siano dovute esclusivamente alla capacità del modello di apprendere le relazioni temporali, ma siano fortemente legate alla qualità e alla struttura dei dati disponibili.

Uno degli aspetti maggiormente rilevanti, emersi durante la sperimentazione, riguarda, infatti, la qualità del dataset utilizzato.

Le serie temporali analizzate presentano numerosi intervalli privi di osservazioni, caratteristica che ha influenzato significativamente la fase di addestramento dei modelli. La presenza di valori mancanti ha comportato l'esclusione di dati per svariati mesi, riducendo la quantità effettiva di dati disponibili per l'apprendimento delle dinamiche della serie.

Questa limitazione risulta particolarmente critica nel caso di modelli basati su reti neurali, che richiedono generalmente una quantità elevata di esempi per apprendere pattern complessi e generalizzare correttamente su nuovi periodi temporali.

Inoltre, la necessità di utilizzare una previsione multistep diretta ha rappresentato un ulteriore elemento di difficoltà. In tale configurazione il modello deve stimare simultaneamente un intero intervallo futuro di cinque giorni utilizzando esclusivamente le informazioni disponibili nella finestra temporale precedente. Questo approccio, sebbene permetta di evitare l'accumulo degli errori tipico delle strategie ricorsive, rende più complesso il compito predittivo soprattutto quando la serie presenta elevata variabilità e fenomeni improvvisi.

Una possibile alternativa consiste nell'utilizzare una strategia di previsione iterativa, nella

quale il modello genera una previsione alla volta e utilizza il valore stimato come ingresso per il passo successivo. Tale approccio, noto come previsione step-by-step, verrà analizzato nei capitoli successivi con l'obiettivo di valutare se possa risultare più adatto alla natura delle serie temporali considerate.

Alla luce delle problematiche evidenziate, si è scelto di estendere successivamente l'analisi utilizzando un nuovo dataset fornito da ARPAM. Il nuovo insieme di dati presenta una maggiore estensione temporale e una struttura più completa rispetto al dataset inizialmente utilizzato.

CAPITOLO 6

Un secondo dataset di riferimento

In questo capitolo verrà presentato un secondo dataset, utilizzato per successive fasi di analisi e sperimentazione. Inizialmente verranno descritte le principali caratteristiche dei dati e delle centraline considerate. Successivamente, verranno effettuate alcune analisi preliminari, volte a valutare la completezza, la distribuzione temporale e le relazioni tra le variabili disponibili. Infine, il nuovo dataset sarà confrontato con quello utilizzato nelle precedenti fasi del lavoro, evidenziandone le differenze e le motivazioni che hanno portato al suo utilizzo.

6.1 Descrizione generale del nuovo dataset

Il dataset alternativo proposto è stato ottenuto a partire dai dati messi a disposizione dall'ARPAM, l'Agenzia Regionale per la Protezione Ambientale delle Marche, attraverso il portale dedicato al monitoraggio della qualità dell'aria, la cui pagina principale è riportata in Figura 6.1.

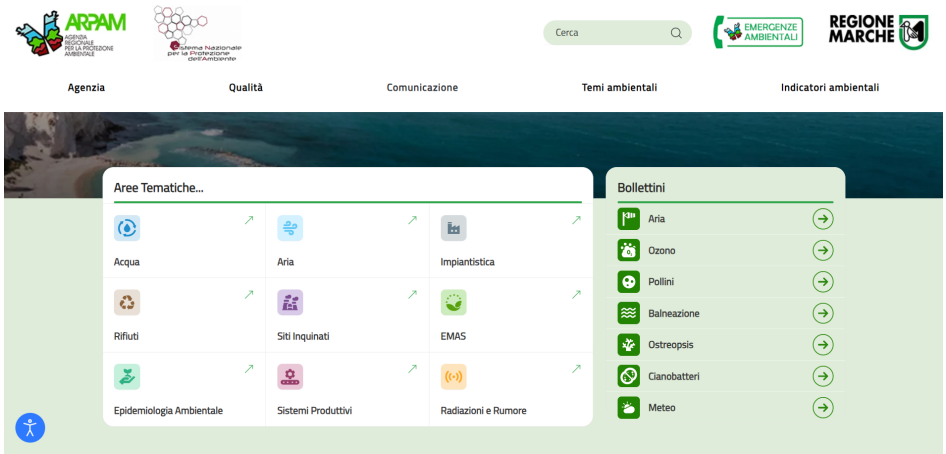


Figura 6.1: Home page del sito dell'ARPAM.

Attraverso la sezione dedicata al monitoraggio della qualità dell'aria, da ricercare nel blocco "Bollettini", è possibile visualizzare la distribuzione geografica delle centraline di rilevamento, consultare i dati relativi agli inquinanti monitorati e analizzarne l'andamento nel tempo mediante apposite rappresentazioni grafiche. Tale sezione è riportata in Figura 6.2.



Figura 6.2: Sezione dei grafici del sito dell'ARPAM.

Per le nostre finalità, è stato necessario accedere alla sezione dedicata ai grafici, che consente di visualizzare l'andamento temporale delle concentrazioni degli inquinanti monitorati dalle diverse centraline. Da qui è stato possibile esportare i dati in formato CSV, riguardanti le centraline di Ancona, Chiaravalle, Falconara, Fano e San Benedetto.

Per quanto riguarda le variabili monitorate, le diverse centraline non presentano esattamente lo stesso insieme di metriche. Sono, tuttavia, presenti alcune grandezze comuni, che permettono di effettuare un confronto omogeneo tra le diverse stazioni. Le metriche comuni individuate sono:

- Benzene [$\mu g/m^3$].
- NO_2 [$\mu g/m^3$].
- NOx [$\mu g/m^3$].
- PM10 [$\mu g/m^3$].
- PM2.5 [$\mu g/m^3$].
- SO_2 [$\mu g/m^3$].

Per alcune delle centraline considerate è, inoltre, disponibile la concentrazione di CO [mg/m^3]. Tale variabile è presente per le centraline di Ancona, Chiaravalle e Fano e viene, pertanto, considerata nelle loro analisi.

La presenza di un insieme comune di variabili costituisce un elemento importante per le successive analisi, poiché permette di confrontare il comportamento delle stesse grandezze misurate in diversi punti del territorio regionale.

Un'ulteriore caratteristica distintiva del nuovo dataset riguarda la frequenza temporale delle osservazioni. A differenza del dataset precedente, costituito da dati acquisiti con frequenza oraria, si è scelto di esportare le informazioni su base giornaliera.

Questa differenza deve essere tenuta in considerazione nell'interpretazione dei risultati, poichè la rappresentazione giornaliera non consente di osservare le variazioni che avvengono all'interno della singola giornata e tende a ridurre l'effetto delle oscillazioni di breve periodo. Di conseguenza, le serie temporali risultano generalmente più regolari, proprio perchè variazioni rapide e picchi di durata limitata possono non essere direttamente rappresentati.

Dal punto di vista temporale, tutte le centraline considerate presentano dati a partire dal gennaio 2022 e arrivano fino al 12 maggio 2026. L'estensione temporale del nuovo dataset

risulta, quindi, significativamente maggiore rispetto a quella del dataset precedentemente utilizzato, permettendo di disporre di una storia temporale più lunga per l'analisi degli inquinanti.

6.2 Analisi preliminare dei dati

Prima di procedere con l'utilizzo del dataset nelle successive fasi sperimentali, è stata effettuata un'analisi preliminare finalizzata a verificarne la struttura, la completezza e le principali caratteristiche statistiche. L'analisi ha riguardato la distribuzione temporale delle osservazioni, la presenza di valori mancanti e le relazioni tra le variabili disponibili.

Questa fase risulta particolarmente importante nel contesto della previsione di serie temporali, poiché la presenza di valori mancanti, una diversa distribuzione temporale delle osservazioni o relazioni significative tra le variabili possono influenzare le successive fasi di elaborazione e modellazione.

L'analisi è stata realizzata mediante Python, utilizzando alcune delle principali librerie per la manipolazione, l'analisi e la visualizzazione dei dati. In particolare, sono state utilizzate:

- *pandas*, per leggere a manipolare il dataset;
- *numpy*, per effettuare operazioni matematiche e statistiche;
- *matplotlib*, per la creazione di grafici;
- *seaborn*, per la creazione di grafici più complessi e per la visualizzazione dei dati, come le heatmap;
- *os e glob*, per la gestione dei file e delle directory.

Una volta caricati i dataset relativi alle cinque centraline considerate, per ciascuna di esse sono state analizzate le colonne disponibili, così da individuare le metriche comuni, precedentemente elencate.

Successivamente, è stata analizzata la distribuzione temporale delle osservazioni. Per ogni centralina sono state individuate la prima e l'ultima data disponibile, così da verificare l'estensione temporale delle serie.

I risultati di questa prima analisi, insieme all'elenco delle metriche comuni alle diverse centraline, sono riportati in Figura 6.3.

```
dataset_ancona: 2022-01-01 00:00:00 --> 2026-12-05 00:00:00
dataset_chiaravalle: 2022-01-01 00:00:00 --> 2026-12-05 00:00:00
dataset_falconara: 2022-01-01 00:00:00 --> 2026-12-05 00:00:00
dataset_fano: 2022-01-01 00:00:00 --> 2026-12-05 00:00:00
dataset_san_benedetto: 2022-01-01 00:00:00 --> 2026-12-05 00:00:00
Metriche comuni a tutti i sensori:
['Benzene [ug/m3]', 'NO2 [ug/m3]', 'NOx [ug/m3]', 'PM10 [ug/m3]', 'PM2.5 [ug/m3]']
```

Figura 6.3: Giorni e metriche comuni.

Successivamente è stata analizzata la presenza di valori mancanti in tali metriche. Per ciascuna combinazione centralina-metrica è stato calcolato sia il numero complessivo di osservazioni mancanti sia la relativa percentuale rispetto al numero totale di osservazioni disponibili.

L'analisi è stata effettuata, dapprima considerando le metriche comuni a tutte le centraline, e, necessariamente, includendo anche la variabile CO per le sole centraline nelle quali essa risulta disponibile.

	Sensor	Metric	Total_missing	Total_values	Missing_pct
0	dataset_ancona	Benzene [ug/m3]	156	1611	9.683426
1	dataset_ancona	NO2 [ug/m3]	40	1611	2.482930
2	dataset_ancona	NOx [ug/m3]	43	1611	2.669150
3	dataset_ancona	PM10 [ug/m3]	96	1611	5.959032
4	dataset_ancona	PM2.5 [ug/m3]	110	1611	6.828057
5	dataset_ancona	SO2 [ug/m3]	89	1611	5.524519
6	dataset_ancona	CO [mg/m3]	72	1611	4.469274
7	dataset_chiaravalle	Benzene [ug/m3]	217	1611	13.469894
8	dataset_chiaravalle	NO2 [ug/m3]	131	1611	8.131595
9	dataset_chiaravalle	NOx [ug/m3]	118	1611	7.324643
10	dataset_chiaravalle	PM10 [ug/m3]	98	1611	6.083178
11	dataset_chiaravalle	PM2.5 [ug/m3]	155	1611	9.621353
12	dataset_chiaravalle	SO2 [ug/m3]	76	1611	4.717567
13	dataset_chiaravalle	CO [mg/m3]	84	1611	5.214153
14	dataset_falconara	Benzene [ug/m3]	57	1611	3.538175
15	dataset_falconara	NO2 [ug/m3]	31	1611	1.924271
16	dataset_falconara	NOx [ug/m3]	23	1611	1.427685
17	dataset_falconara	PM10 [ug/m3]	90	1611	5.586592
18	dataset_falconara	PM2.5 [ug/m3]	96	1611	5.959032
19	dataset_falconara	SO2 [ug/m3]	38	1611	2.358783
20	dataset_fano	Benzene [ug/m3]	55	1611	3.414029
21	dataset_fano	NO2 [ug/m3]	38	1611	2.358783
22	dataset_fano	NOx [ug/m3]	37	1611	2.296710
23	dataset_fano	PM10 [ug/m3]	84	1611	5.214153
24	dataset_fano	PM2.5 [ug/m3]	73	1611	4.531347
25	dataset_fano	SO2 [ug/m3]	95	1611	5.896958
26	dataset_fano	CO [mg/m3]	68	1611	4.220981
27	dataset_san_benedetto	Benzene [ug/m3]	84	1611	5.214153
28	dataset_san_benedetto	NO2 [ug/m3]	33	1611	2.048417
29	dataset_san_benedetto	NOx [ug/m3]	29	1611	1.800124
30	dataset_san_benedetto	PM10 [ug/m3]	81	1611	5.027933
31	dataset_san_benedetto	PM2.5 [ug/m3]	264	1611	16.387337
32	dataset_san_benedetto	CO [mg/m3]	69	1611	4.283054

Figura 6.4: Dati mancanti nelle metriche comuni.

I risultati relativi alla presenza di dati mancanti sono riportati in Figura 6.4.

Tale analisi, però, non ha costituito il principale criterio utilizzato per la selezione delle variabili da impiegare nelle successive fasi di previsione. La scelta è stata effettuata principalmente sulla base della possibilità di mantenere un confronto diretto con le precedenti sperimentazioni condotte sul primo dataset. Sono state, quindi, selezionate le metriche già considerate nella fase precedente e presenti anche nel nuovo insieme di dati, in particolare PM10, NO_2 e, dove disponibile, CO.

L'utilizzo delle stesse variabili permette di confrontare i risultati ottenuti sui due dataset, pur tenendo conto delle differenze esistenti. Il confronto non è, dunque, finalizzato a stabilire

una corrispondenza diretta tra i singoli valori osservati, ma a valutare il comportamento dei modelli e la loro capacità di riprodurre l'andamento delle serie temporali nei due differenti contesti.

Un'ulteriore analisi è stata effettuata per valutare le relazioni esistenti tra le diverse variabili misurate dalle singole centraline. A questo scopo è stato utilizzato il coefficiente di correlazione di Pearson, che consente di quantificare la presenza e la direzione di una relazione lineare tra due variabili. L'analisi è stata effettuata separatamente per ciascuna centralina. Per ogni sensore sono state considerate le metriche per le quali erano disponibili osservazioni; infatti un aspetto importante riguarda la gestione dei valori mancanti.

Prima del calcolo della correlazione sono state mantenute esclusivamente le osservazioni complete, ovvero le righe per le quali tutte le metriche considerate presentavano contemporaneamente un valore disponibile. In questo modo, il calcolo della correlazione viene effettuato utilizzando lo stesso insieme di osservazioni per tutte le variabili considerate all'interno della singola centralina.

La matrice di correlazione ottenuta per ciascun sensore è stata successivamente rappresentata mediante una heatmap. In queste rappresentazioni, le righe e le colonne corrispondono alle diverse metriche, mentre ciascuna cella riporta il coefficiente di correlazione di Pearson relativo alla coppia di variabili considerata, dove valori prossimi a 1 o -1 indicano una forte relazione lineare, rispettivamente positiva e negativa, mentre valori prossimi a 0 indicano l'assenza di una relazione significativa tra le due variabili.

Le matrici di correlazione ottenute per le cinque centraline sono riportate nelle Figure 6.5, 6.6, 6.7, 6.8 e 6.9.

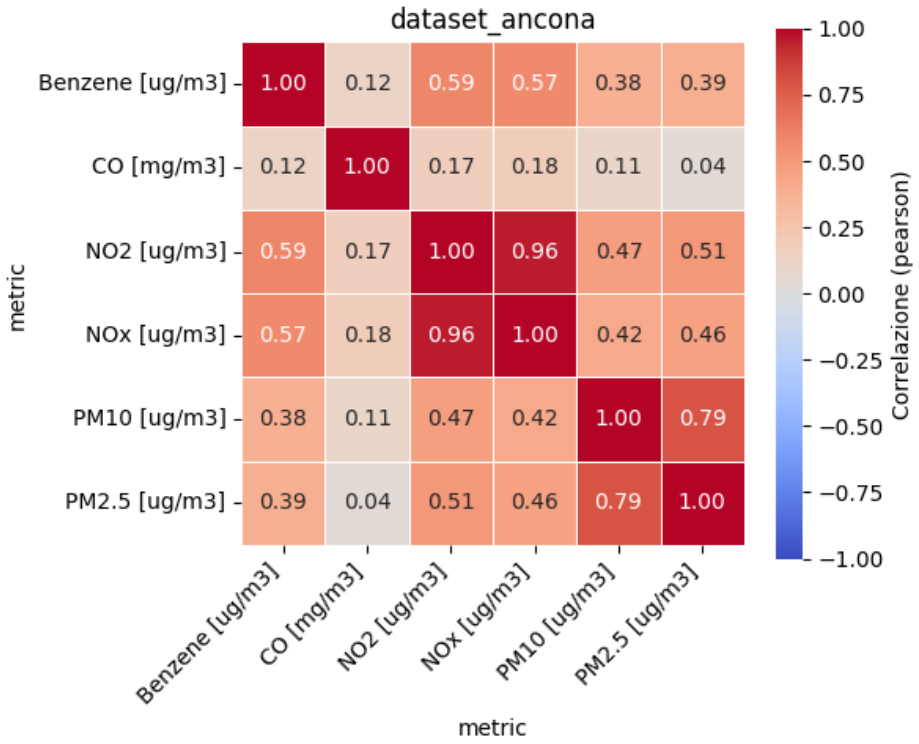


Figura 6.5: Heatmap per la centralina di Ancona.

L'osservazione delle matrici di correlazione permette di individuare alcune relazioni ricorrenti tra le diverse centraline. La relazione maggiormente evidente riguarda NO_2 e NOx, che presentano coefficienti di correlazione elevati in tutti i sensori analizzati, compresi tra

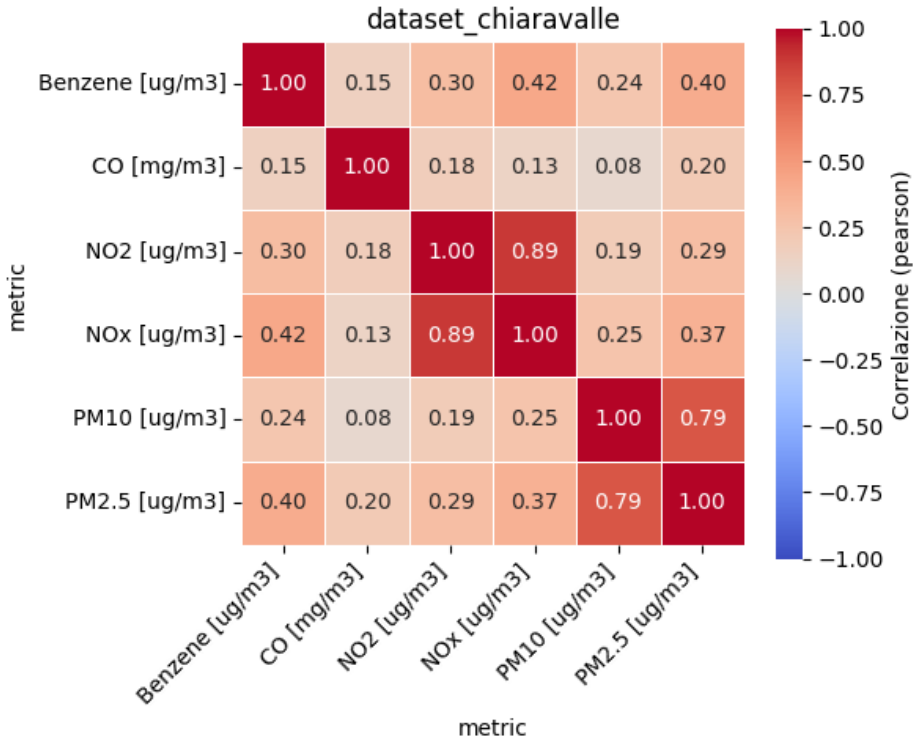


Figura 6.6: Heatmap per la centralina di Chiaravalle.

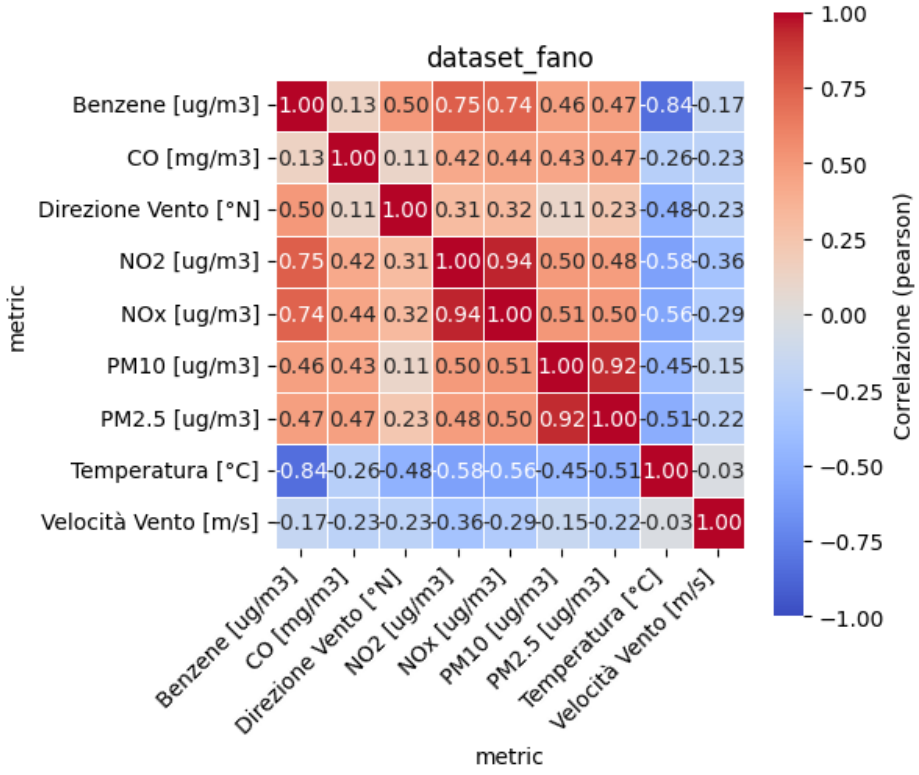


Figura 6.7: Heatmap per la centralina di Fano.

0.89 e 0.96. Il comportamento risulta particolarmente stabile e indica una forte associazione lineare tra le due variabili nel periodo considerato.

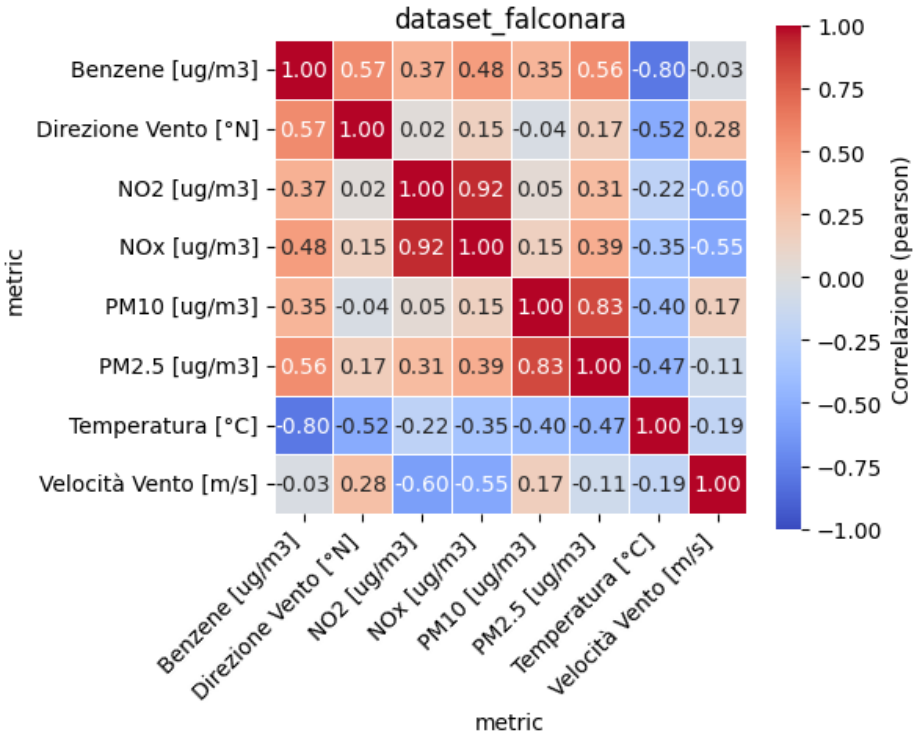


Figura 6.8: Heatmap per la centralina di Falconara.

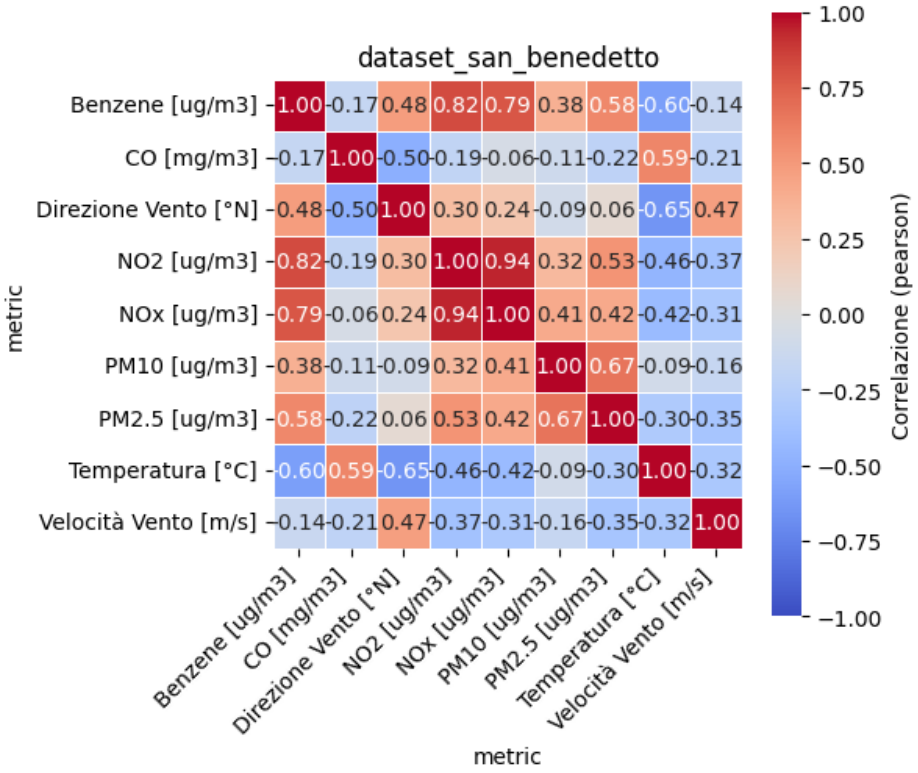


Figura 6.9: Heatmap per la centralina di San Benedetto.

Una seconda relazione rilevante è quella tra PM10 e PM2.5, caratterizzata da coefficienti positivi e relativamente elevati in tutte le centraline, compresi tra 0.67 e 0.92. La correlazione

è particolarmente marcata nella centralina di Fano, dove raggiunge un valore pari a 0.92, mentre assume il valore più contenuto, pur rimanendo positivo, a San Benedetto, con 0.67.

Sono presenti, anche, alcune relazioni tra il Benzene e gli ossidi di azoto. La correlazione con NO_2 e NOx risulta più elevata nelle centraline di Fano e San Benedetto, mentre assume valori più moderati nelle altre stazioni. Le associazioni tra queste variabili sono presenti in più centraline, ma la loro intensità varia in funzione del contesto considerato.

Le variabili meteorologiche mostrano un comportamento meno uniforme. La temperatura presenta diverse correlazioni negative con gli inquinanti, particolarmente evidenti per il Benzene nelle centraline di Falconara e Fano. La velocità del vento presenta, in alcune stazioni, correlazioni negative con NO_2 e NOx. Le relazioni cambiano tra le diverse centraline e non permettono di individuare un comportamento univoco per l'intero dataset.

6.3 Confronto con il dataset iniziale

Il dataset ARPAM presenta alcune differenze sostanziali rispetto al dataset di AirQuino utilizzato nella prima fase sperimentale. Una prima differenza riguarda la distribuzione spaziale delle centraline; infatti, mentre nel dataset precedente i sensori erano collocati esclusivamente nell'area urbana della città di Pesaro, il nuovo dataset comprende centraline distribuite nella totalità del territorio marchigiano. Questa caratteristica permette di disporre di osservazioni provenienti da contesti territoriali maggiormente diversificati.

Una seconda differenza riguarda l'estensione temporale dei dati. Il nuovo dataset contiene osservazioni relative a un periodo significativamente più lungo rispetto al precedente, permettendo di disporre complessivamente di un numero maggiore di osservazioni. Sebbene i dati siano organizzati con frequenza giornaliera e non oraria come nel dataset precedente, la maggiore estensione temporale consente, comunque, di ottenere una quantità di dati superiore e di analizzare il comportamento degli inquinanti su un arco temporale più ampio.

Un'ulteriore differenza riguarda le variabili disponibili. A differenza del dataset precedente, nel quale le metriche considerate erano maggiormente uniformi tra le centraline, il nuovo dataset presenta una maggiore eterogeneità delle variabili monitorate. Accanto agli inquinanti atmosferici, presenti in modo più ricorrente tra le diverse stazioni, alcune centraline includono anche variabili di natura meteorologica, come direzione e velocità del vento e temperatura. Di conseguenza, l'insieme delle metriche disponibili varia in funzione della centralina considerata, con alcune stazioni caratterizzate da un numero maggiore di variabili rispetto ad altre.

Infine, l'analisi dei valori mancanti ha evidenziato una maggiore completezza delle serie relative alle variabili considerate nell'analisi. Sebbene siano presenti valori mancanti anche nel nuovo dataset e alcune metriche presentino una quantità di dati assenti più elevata, tali variabili non sono state considerate nelle successive fasi dello studio. Per le grandezze di interesse, invece, la minore presenza di intervalli privi di osservazioni permette di disporre di una quantità maggiore di dati effettivamente utilizzabili.

CAPITOLO 7

Applicazione di tecniche innovative al secondo dataset di riferimento

In questo capitolo vengono applicate alcune delle metodologie viste nei capitoli precedenti con l'obiettivo di valutare l'effetto della maggiore disponibilità e qualità dei dati sulle prestazioni dei modelli di previsione.

La sperimentazione introduce, inoltre, una diversa strategia di previsione, basata sull'approccio rolling one-step, e ne confronta il comportamento con quello della strategia multistep utilizzata nelle precedenti analisi.

L'analisi parte da un modello statistico semplice, ARIMA, utilizzato come riferimento, per passare successivamente a Prophet e, infine, al modello ibrido Prophet-LSTM.

7.1 Impostazione della sperimentazione

Nel capitolo precedente è stato introdotto un secondo dataset di riferimento; il ricorso ad esso è motivato dalla necessità di verificare il comportamento dei modelli in un contesto caratterizzato da una diversa disponibilità e organizzazione dei dati.

Oltre a tale cambio, è stata modificata anche la strategia di previsione. Negli esperimenti precedenti era stata utilizzata una modalità multistep, nella quale il modello produceva direttamente l'intero intervallo di previsione a partire dai dati disponibili al termine dell'addestramento. Viene, ora, adottata una strategia di tipo rolling one-step, nella quale il modello produce una previsione

Prima dell'applicazione delle strategie di previsione, i dati sono stati suddivisi temporalmente in un insieme di training e uno di test. Il training comprende tutte le osservazioni precedenti al 1 gennaio 2026, mentre il test comprende il periodo compreso tra il 1 gennaio 2026 e la data dell'ultima misurazione disponibile, ovvero il 12 maggio 2026. La suddivisione viene effettuata inizialmente sui dati giornalieri, mantenendo separati i due intervalli temporali, ma, per la generazione dei grafici relativi, i dati vengono accorpati settimanalmente. Questa operazione viene effettuata per ottenere una visualizzazione più comprensibile di predizioni fatte su un lasso di tempo molto ampio, di quasi 5 mesi.

La previsione viene effettuata mediante una strategia di tipo rolling one-step. Ad ogni itera zione il modello viene addestrato utilizzando tutta la storia disponibile fino a quel momento e viene prodotta una previsione relativa al giorno successivo. Una volta effettuata la previsione, il valore reale del giorno appena osservato viene aggiunto alla storia utilizzata dal modello. La procedura viene ripetuta fino al completamento dell'intero periodo di test. In questo modo, a differenza dell'approccio multistep adottato nelle precedenti sperimentazioni, il modello non deve prevedere in un'unica fase un intervallo temporale esteso. Ogni previsione

viene effettuata utilizzando anche l'ultima osservazione reale disponibile, permettendo al modello di aggiornarsi progressivamente durante il periodo di test.

Il dataset utilizzato comprende le centraline di Ancona, Fano, Chiaravalle, Falconara e San Benedetto e le metriche selezionate nelle analisi precedenti, quindi: NO_2, PM10 e, dove presente, CO. La procedura viene applicata separatamente a ciascuna combinazione centralina–metrica.

7.2 Applicazione di ARIMA

Poiché uno degli obiettivi della nuova sperimentazione è valutare l'effetto della strategia *rolling one-step*, è stato inizialmente scelto un modello semplice da utilizzare come baseline. A questo scopo è stato adottato il modello ARIMA (*AutoRegressive Integrated Moving Average*), ampiamente utilizzato nell'analisi e nella previsione delle serie temporali.

Si è partiti da una configurazione base del modello ARIMA riportata di seguito:

$$ARIMA(1, 0, 0) \tag{7.2.1}$$

La configurazione adottata corrisponde, di fatto, a un modello autoregressivo di primo ordine. Il parametro $p = 1$ indica che la previsione utilizza la dipendenza dall'osservazione precedente della serie, mentre $d = 0$ indica che non viene applicata alcuna differenziazione. Il valore $q = 0$ indica, invece, l'assenza di una componente a media mobile basata sugli errori precedenti.

La scelta di una configurazione così semplice come primo esperimento sul nuovo dataste è stata fatta con l'obiettivo di verificare quanto una metodologia basata esclusivamente sulla dipendenza temporale della variabile considerata fosse in grado di descriverne l'andamento. Non sono state introdotte variabili esogene né componenti stagionali; in questo modo il comportamento del modello viene valutato utilizzando esclusivamente la storia della singola serie.

Siccome ARIMA viene applicato alla serie settimanale, i dati giornalieri sono stati aggre gati mediante il calcolo della media dei valori disponibili all'interno di ciascuna settimana. La settimana è stata definita utilizzando la frequenza W-SUN, con termine la domenica. Il procedimento di aggregazione è riportato nel Listato 7.1.

```
def weekly_aggregation(df, weekly_rule="W-SUN"):
    return df.resample(weekly_rule).mean()

train_weekly = weekly_aggregation(train_daily)
test_weekly = weekly_aggregation(test_daily)
```

Listing 7.1: Funzione per l'aggregazione settimanale dei dati giornalieri.

Per ciascuna combinazione centralina–metrica sono stati generati i relativi grafici di previsione. Considerato l'elevato numero di grafici prodotte, nel seguito vengono riportati alcuni casi rappresentativi, selezionati per la centralina di Ancona e per le metriche CO e PM10. Per valutare l'effetto della nuova strategia di previsione sono stati generati anche i risultati ottenuti mediante l'approccio multistep precedentemente adottato.

Le Figure 7.1 e 7.2 mostrano i risultati ottenuti con l'approccio multistep, rispettivamente per il CO e il PM10 della centralina di Ancona. Le Figure 7.3 e 7.4 riportano gli stessi esperimenti utilizzando la strategia rolling one-step.

Il confronto visivo evidenzia una maggiore aderenza delle previsioni ai valori osservati nel caso della strategia rolling one-step.

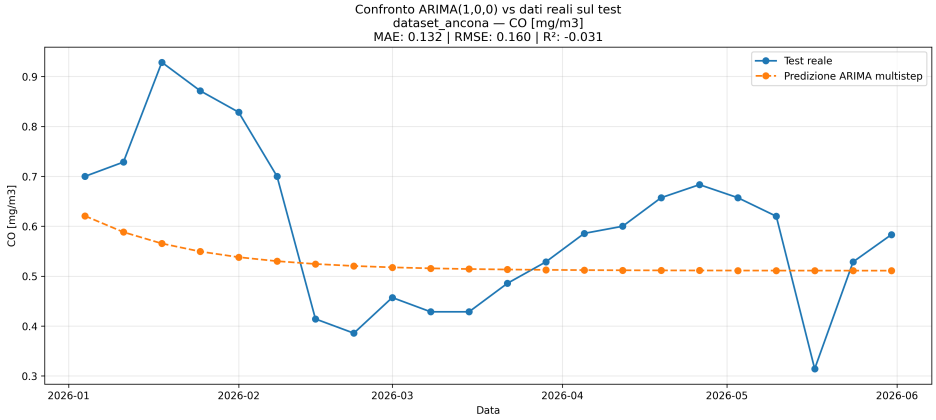


Figura 7.1: Grafico delle previsioni settimanali di CO del sensore ad Ancona mediante metodo multistep.

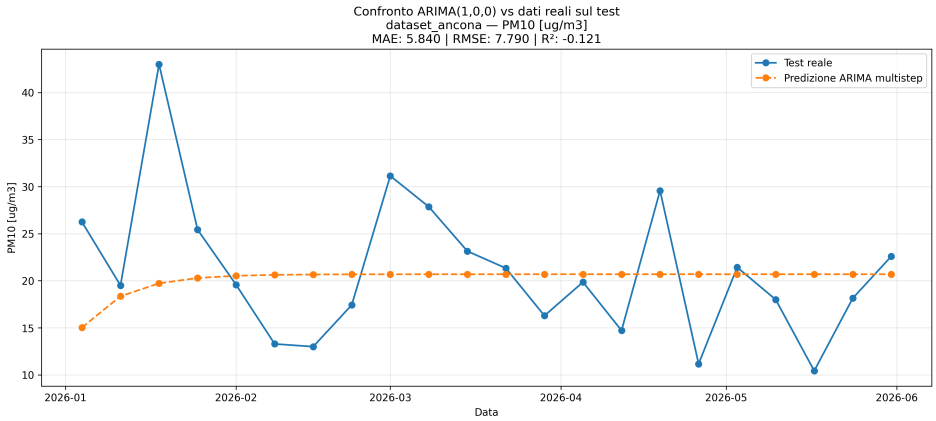


Figura 7.2: Grafico delle previsioni settimanali di PM10 del sensore ad Ancona mediante metodo multistep.

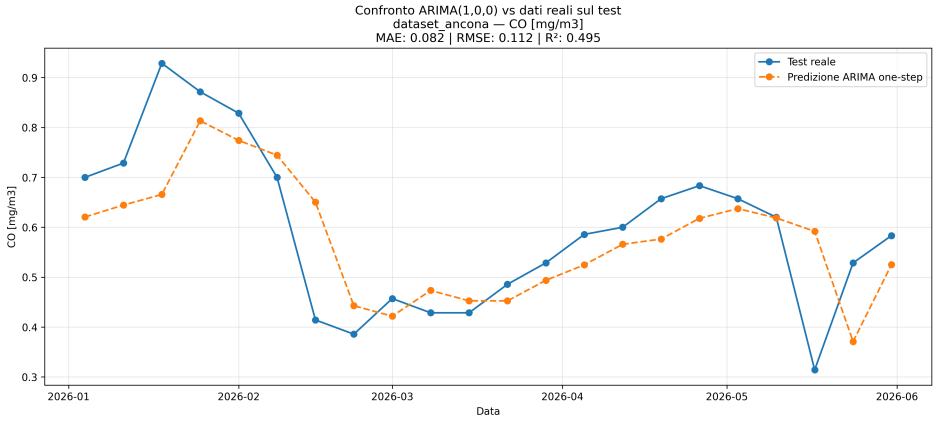


Figura 7.3: Grafico delle previsioni settimanali di CO del sensore ad Ancona mediante metodo rolling one-step.

7.3 Applicazione di Prophet

Visti i risultati ottenuti con il modello ARIMA, ci si è chiesti se queste differenze fossero rilevabili anche mediante l'applicazione di modelli più complessi, come Prophet. Specifi-

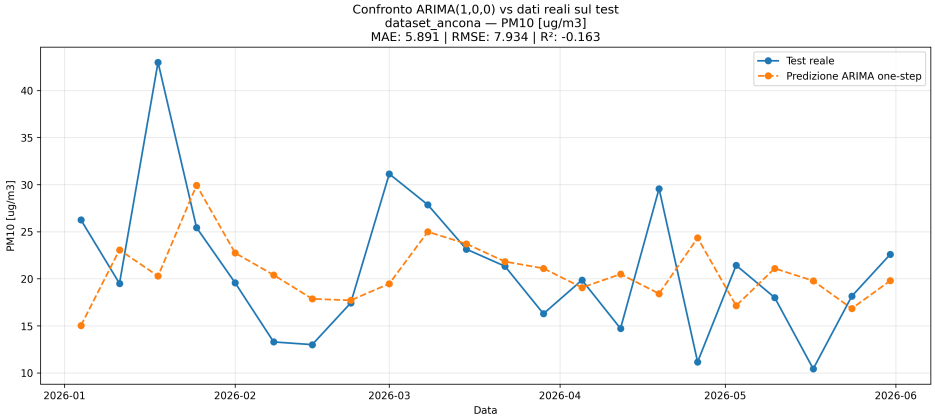


Figura 7.4: Grafico delle previsioni settimanali di PM10 del sensore ad Ancona mediante metodo rolling one step.

chiamo che l'analisi non è stata estesa a NeuralProphet, considerati gli eleveati tempi di elaborazione richiesti da quest'ultim algoritmo.

A differenza di ARIMA, Prophet viene addestrato sulla serie giornaliera. Le previsioni vengono successivamente aggregate su base settimanale, in modo da rendere più comprensibile il confronto tra i valori osservati e quelli previsti nell'intero periodo di test.

La configurazione iniziale che abbiamo assegnato a Prophet è quella che ha permesso di ottenere i migliori risultati nei capitoli precedenti, riportata nel Listato 7.2.

```
model = Prophet(
    growth="linear",
    changepoint_prior_scale=0.05,
    n_changepoints=25,
    changepoint_range=0.8,
    yearly_seasonality=False,
    weekly_seasonality=True,
    daily_seasonality=True,
    seasonality_mode="additive",
    seasonality_prior_scale=10.0,
    holidays_prior_scale=10.0
)
```

Listing 7.2: Configurazione iniziale del modello Prophet.

Tale configurazione è stata utilizzata come riferimento per valutare l'effetto della successi va ottimizzazione dei parametri. Riportiamo due grafici risultanti da questa configurazione, in modo da poterli, poi, confrontare con le sucessive condigurazioni. In particolare, riportia mo i grafici ottenuti dalla centralina di Ancona e Chiaravalle per il PM10, rispettivamente raffigurati nelle Figure 7.5 e 7.6.

Successivamente è stata effettuata una ricerca a griglia con l'obiettivo di individuare una singola configurazione di Prophet in grado di fornire complessivamente le migliori prestazioni sull'insieme delle centraline e delle metriche considerate.

Sono state valutate le combinazioni riportate nel Listato 7.3.

```
prophet_param_grid = [
    {"changepoint_prior_scale": 0.001, "seasonality_prior_scale": 0.1},
    {"changepoint_prior_scale": 0.01,  "seasonality_prior_scale": 1.0},
    {"changepoint_prior_scale": 0.05,  "seasonality_prior_scale": 1.0},
```

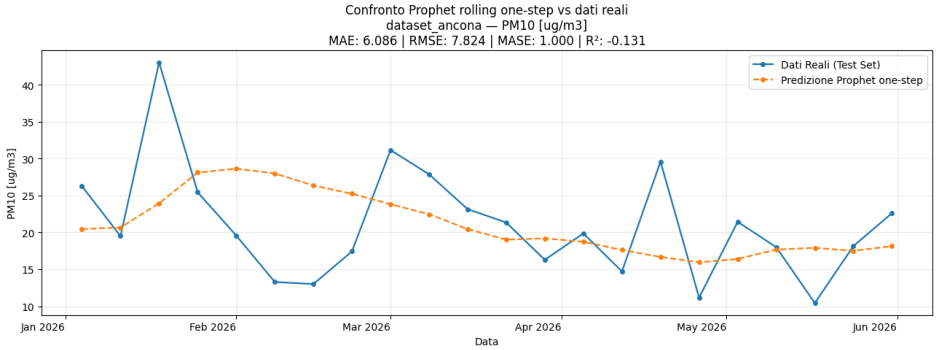


Figura 7.5: Previsioni di PM10 per la centralina di Ancona ottenute con la configurazione base di Prophet.

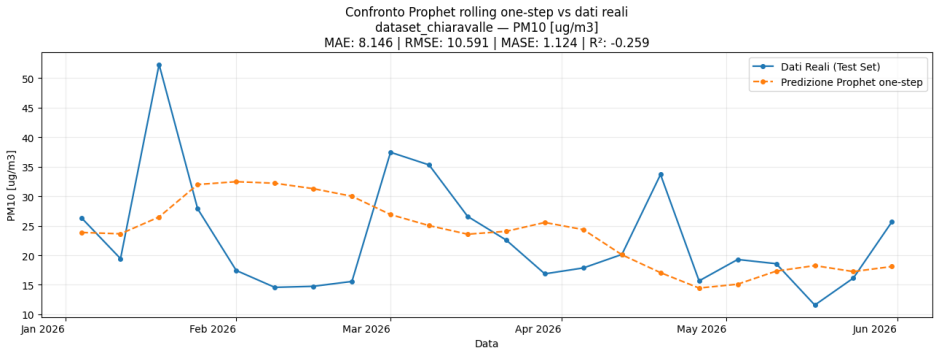


Figura 7.6: Previsioni di PM10 per la centralina di Chiaravalle ottenute con la configurazione base di Prophet.

```
    {"changepoint_prior_scale": 0.1,   "seasonality_prior_scale": 10.0},
    {"changepoint_prior_scale": 0.5,   "seasonality_prior_scale": 10.0},
    {"changepoint_prior_scale": 0.5,   "seasonality_prior_scale": 10.0},
    {"changepoint_prior_scale": 0.01,  "seasonality_prior_scale": 1.0},
    {"changepoint_prior_scale": 0.05,  "seasonality_prior_scale": 10.0},
]
```

Listing 7.3: Configurazione per la grid search del modello Prophet.

Per ciascuna configurazione, il modello è stato addestrato sulle serie di training e valutato sul periodo di test per tutte le combinazioni centralina–metrica considerate. Per ogni combinazione è stato calcolato il valore di RMSE; i valori ottenuti sono stati successivamente aggregati mediante la loro media, ottenendo un unico valore di Mean-RMSE per ciascuna configurazione. La configurazione associata al valore più basso di Mean-RMSE è stata selezionata come configurazione generale di Prophet e utilizzata nelle successive sperimentazioni.

Il risultato è riportato in Figura 7.7.

Anche per Prophet sono stati prodotti grafici relativi a ciascuna combinazione centralina-metrica. Nelle Figure 7.8, 7.9, 7.10, 7.11 e 7.12 vengono riportati cinque esempi relativi alla metrica PM10, uno per ciascuna delle centraline considerate, ovvero: Ancona, Chiaravalle, Fano, Falconara e San Benedetto.

changepoint_prior_scale	0.05
seasonality_prior_scale	10.0
changepoint_range	0.8
seasonality_mode	multiplicative
yearly_seasonality	True
growth	linear
Mean_RMSE	0.138158
Median_RMSE	0.138158
Mean_MAE	0.124688
Number_of_series	1

Name: 4, dtype: object

Figura 7.7: Risultati della grid search di Prophet.

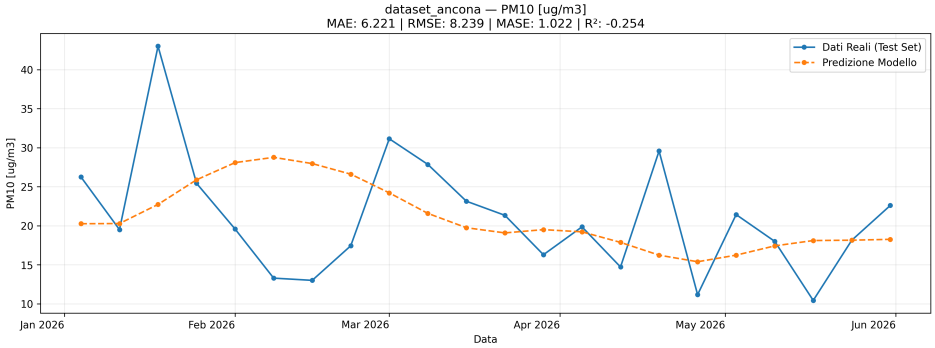


Figura 7.8: Previsioni di PM10 per la centralina di Ancona ottenute con Prophet.

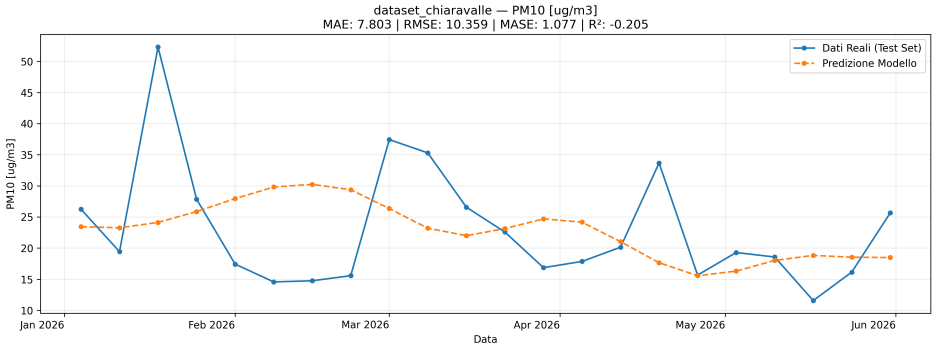


Figura 7.9: Previsioni di PM10 per la centralina di Chiaravalle ottenute con Prophet.

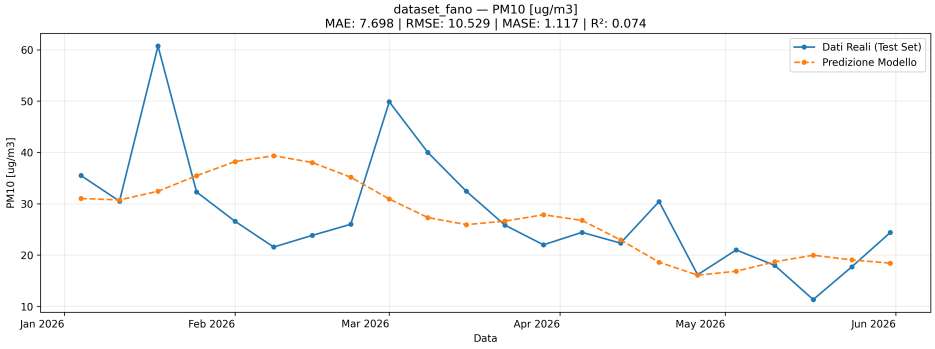


Figura 7.10: Previsioni di PM10 per la centralina di Fano ottenute con Prophet.

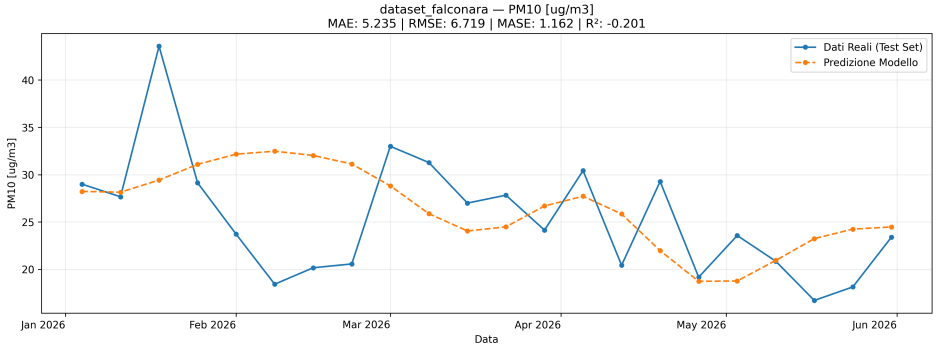


Figura 7.11: Previsioni di PM10 per la centralina di Falconara ottenute con Prophet.

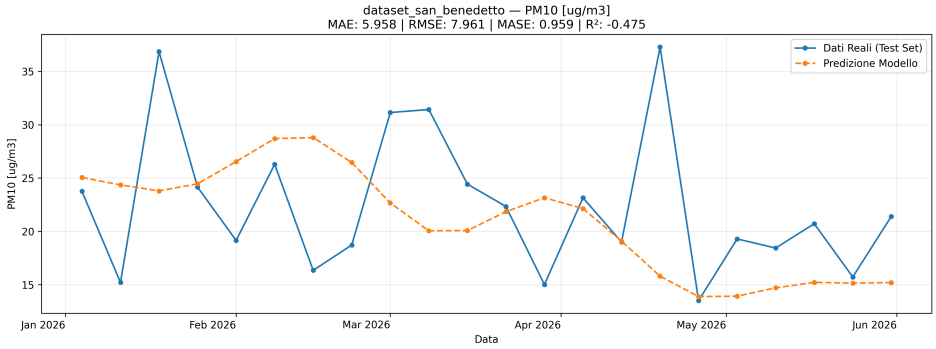


Figura 7.12: Previsioni di PM10 per la centralina di San Benedetto ottenute con Prophet.

7.4 Applicazione di Prophet e LSTM

Nonostante l'utilizzo della configurazione selezionata mediante la grid search, le previsioni ottenute con Prophet mostrano ancora alcune difficoltà nel seguire le variazioni più marcate della serie, in particolare in corrispondenza dei picchi.

Per questo motivo è stato ripreso il modello ibrido Prophet–LSTM, già considerato nelle precedenti sperimentazioni e risultato particolarmente interessante per il rapporto tra qualità delle previsioni e costo computazionale. La configurazione di partenza di Prophet è quella che ha ottenuto i valori migliori durante la ricerca mediante griglia, fatta in precedenza.

Una volta addestrato Prophet sui dati di training, vengono calcolati i residui della serie come differenza tra il valore osservato e quello stimato dal modello. I residui rappresentano la componente della serie non spiegata dal modello Prophet e vengono utilizzati come variabile di ingresso per la rete LSTM.

Nel presente esperimento la LSTM presenta la stessa configurazione usata negli studi precedenti e riportata nel Listato 7.4.

```
INPUT_WINDOW = 30
LEARNING_RATE = 0.001
EPOCHS = 100
BATCH_SIZE = 16
```

Listing 7.4: Configurazione della LSTM utilizzata nel modello ibrido.

La finestra temporale utilizzata dalla rete è, quindi, costituita da 30 osservazioni consecutive. La previsione del residuo viene effettuata da una rete composta da due livelli LSTM, rispettivamente di 64 e 32 unità, seguiti da livelli di dropout e da uno strato denso utilizzato per produrre la previsione scalare del residuo.

Anche per il modello ibrido sono stati generati i grafici relativi alle diverse combinazioni centralina-metrica. Per quanto riguarda i grafici, a causa dell'elevato numero di risultati prodotti, nelle Figure 7.13, 7.14, 7.15, 7.16 e 7.17 vengono riportati cinque esempi relativi alla metrica PM10, uno per ciascuna centralina.

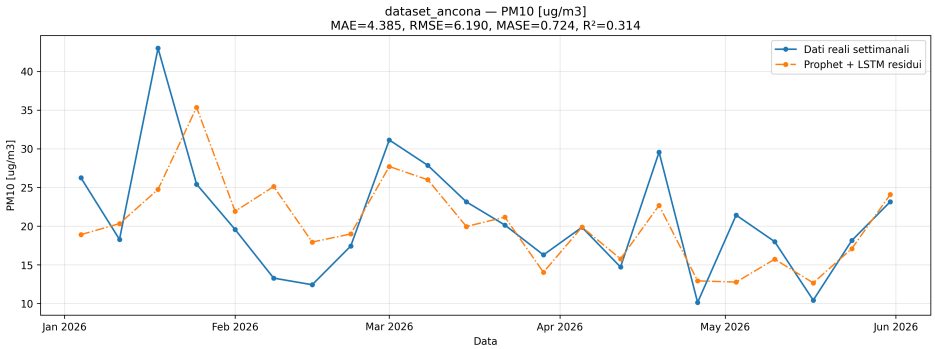


Figura 7.13: Previsioni di PM10 per la centralina di Ancona ottenute con Prophet e LSTM.

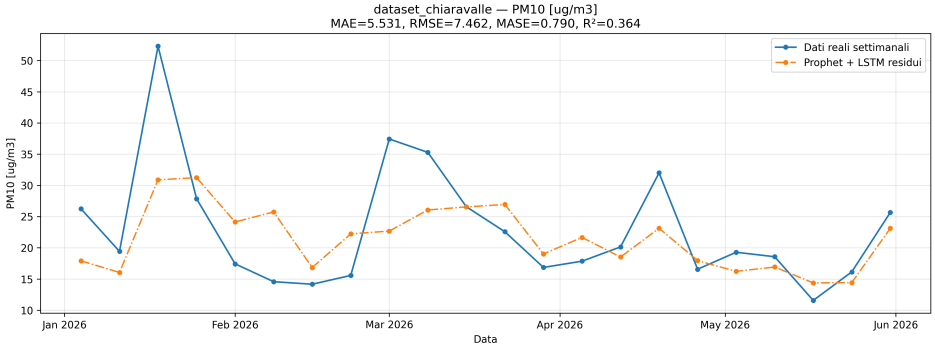


Figura 7.14: Previsioni di PM10 per la centralina di Chiaravalle ottenute con Prophet e LSTM.

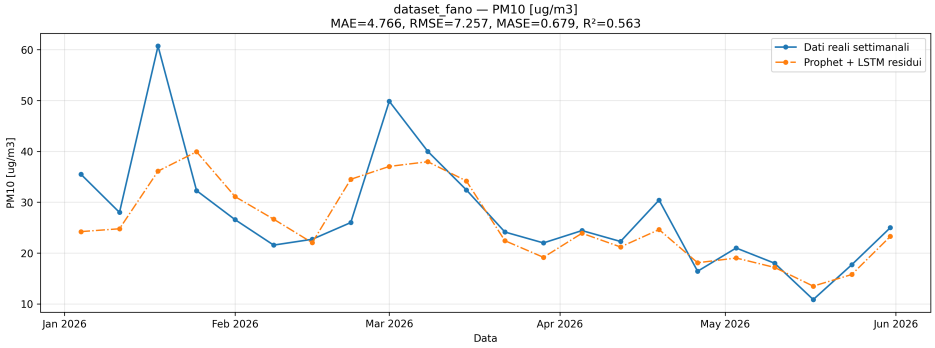


Figura 7.15: Previsioni di PM10 per la centralina di Fano ottenute con Prophet e LSTM.

7.5 Considerazioni finali

La sperimentazione condotta sul secondo dataset ha permesso di valutare sia l'effetto della diversa qualità dei dati sia quello della strategia di previsione rolling one-step. Il confronto iniziale effettuato con ARIMA ha mostrato una maggiore aderenza delle previsioni ai valori osservati rispetto all'approccio multistep. L'aggiornamento progressivo della storia

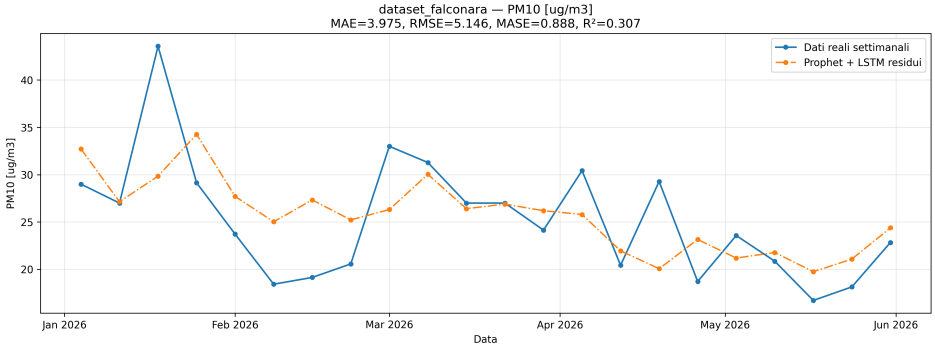


Figura 7.16: Previsioni di PM10 per la centralina di Falconara ottenute con Prophet e LSTM.

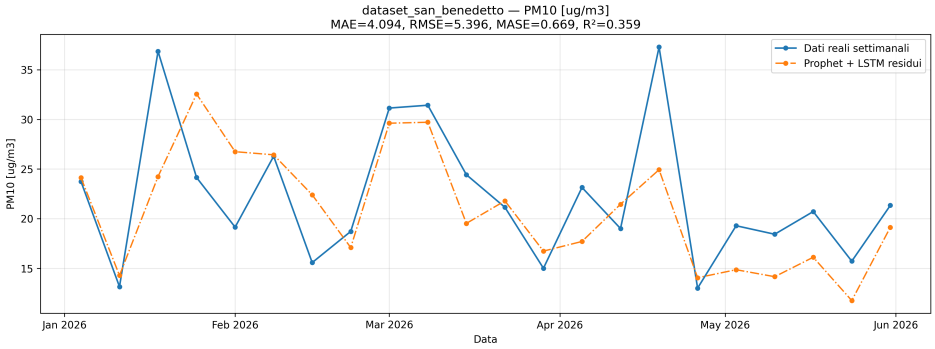


Figura 7.17: Previsioni di PM10 per la centralina di San Benedetto ottenute con Prophet e LSTM.

disponibile permette al modello di utilizzare le osservazioni reali più recenti durante il periodo di previsione.

Nel caso di Prophet si è visto, però, come tali cambiamenti non influiscano in maniera marcata sui risultati ottenuti. Difatti le previsioni rimangono, come le precedenti, più ancorate alla riproduzione dell'andamento generale con grandi difficoltà nella rappresentazione delle variazioni più rapide e dei picchi.

Il modello Prophet con LSTM è stato introdotto con l'obiettivo di affrontare proprio questa limitazione. L'utilizzo della LSTM sui residui permette di aggiungere alla previsione di Prophet una componente in grado di modellare le dinamiche temporali non catturate dal modello principale. I risultati ottenuti dimostrano quanto l'integrazione tra una componente statistica e una componente neurale possa migliorare la capacità predittiva di Prophet, con curve di previsione sorprendentemente accurate, anche per variabili problematiche come il PM10, motivo per il quale si è scelto di riportarlo.

Nel complesso, il nuovo dataset e la strategia rolling one-step costituiscono un'evoluzione rispetto alla configurazione utilizzata nelle precedenti sperimentazioni. Il confronto tra ARIMA, Prophet e Prophet-LSTM permette di analizzare progressivamente l'effetto dell'aumento della complessità del modello, mantenendo costante il periodo di valutazione e la struttura dei dati utilizzati.

CAPITOLO 8

Discussione in merito alle analisi svolte

In questo capitolo vengono analizzati nel loro complesso i risultati ottenuti nel corso della sperimentazione, con particolare attenzione al ruolo della qualità dei dati, alla scelta delle metodologie predittive e alle risorse necessarie per la loro applicazione.

L'obiettivo è mettere in evidenza gli aspetti emersi dall'analisi e individuare gli elementi che possono orientare successive sperimentazioni.

8.1 Sintesi delle sperimentazioni condotte

Riprendiamo brevemente il percorso che è stato seguito nel corso del presente elaborato. L'obiettivo ultimo è quello della valutazione dell'inquinamento atmosferico, basandoci sui dati estrapolati da centraline preposte alla raccolta di dati sugli inquinanti.

La prima fase ha riguardato l'uso del dataset fornito da AirQuino, costituito dalle osservazioni raccolte da centraline situate nell'area urbana di Pesaro e caratterizzato da una frequenza di campionamento oraria. Su questi dati sono state sperimentate diverse metodologie predittive, con particolare attenzione a Prophet e NeuralProphet. Le prestazioni ottenute non sono risultate soddisfacenti e hanno portato ad ampliare la sperimentazione verso ulteriori approcci di previsione.

L'analisi dei risultati ha evidenziato, però, che le difficoltà non erano riconducibili esclusivamente alla scelta del modello. Le caratteristiche del dataset, in particolare la disponibilità e la qualità delle osservazioni, rappresentavano un ulteriore elemento di criticità per l'applicazione delle tecniche considerate. Si è, dunque, scelto di concentrare la sperimentazione di modelli su una nuova fonte di dati, ricorrendo al dataset fornito da ARPAM (Agenzia Regionale per la Proptezione Ambientale delle Marche) e relativo a diverse centraline distribuite sul territorio marchigiano.

Dopo le analisi preliminari e la selezione delle variabili di interesse, sono state riprese alcune delle metodologie che avevano mostrato gli aspetti più interessanti nelle sperimentazioni precedenti. Tra queste, particolare attenzione è stata dedicata al modello ibrido Prophet–LSTM.

8.2 Analisi delle metodologie più efficaci

Tra le diverse metodologie sperimentate, due approcci hanno mostrato particolare interesse: il modello ibrido Prophet–LSTM e NeuralProphet. I risultati ottenuti con questi

approcci hanno permesso di evidenziare aspetti rilevanti legati, rispettivamente, alla strategia di previsione e alla quantità di dati disponibili per l'addestramento.

8.2.1 Integrazione tra modelli

Il modello Prophet–LSTM ha fornito risultati particolarmente soddisfacenti nel secondo dataset. In questo approccio Prophet viene utilizzato per descrivere la componente principale della serie temporale, mentre la differenza tra i valori osservati e le relative previsioni viene utilizzata per ottenere i residui. Questi ultimi vengono successivamente elaborati dalla LSTM, con l'obiettivo di stimare la componente non catturata da Prophet. La previsione finale deriva dalla combinazione delle due componenti.

L'efficacia del modello è risultata evidente nell'utilizzo della strategia rolling one-step, nella quale il modello viene progressivamente aggiornato con le nuove osservazioni reali disponibili. Questa modalità consente di utilizzare le informazioni più recenti della serie ad ogni previsione e ha portato, nel caso considerato, a una maggiore aderenza tra valori osservati e previsti.

8.2.2 Ruolo della disponibilità dei dati nelle prestazioni

Un comportamento differente è stato osservato per NeuralProphet. Le prime sperimentazioni, effettuate sul dataset iniziale e con una quantità più limitata di osservazioni, avevano prodotto risultati deludenti.

L'aumento delle osservazioni disponibili ha determinato un miglioramento nella capacità predittiva del modello pur non consentendo di replicare completamente l'andamento reale delle serie considerate. Le previsioni risultano maggiormente aderenti alle variazioni osservate rispetto alle precedenti sperimentazioni. A titolo rappresentativo, le Figure 8.1, 8.2, 8.3 e 8.4 riportano i risultati ottenuti per le metriche CO_2 e PM10 nelle centraline Smart58 e Smart137.

In questo caso la sperimentazione è stata condotta mediante un approccio multistep, nel quale il modello produce direttamente le previsioni relative all'intervallo temporale futuro considerato.

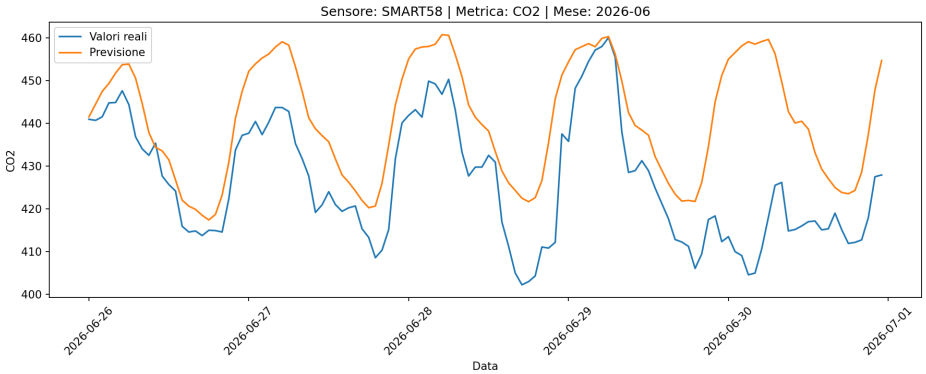


Figura 8.1: Previsione della metrica CO_2 per la centralina SMART58 mediante NeuralProphet.

Le due metodologie risultano interessanti per motivazioni differenti, che devono essere considerate anche alla luce delle risorse computazionali necessarie per la loro applicazione.

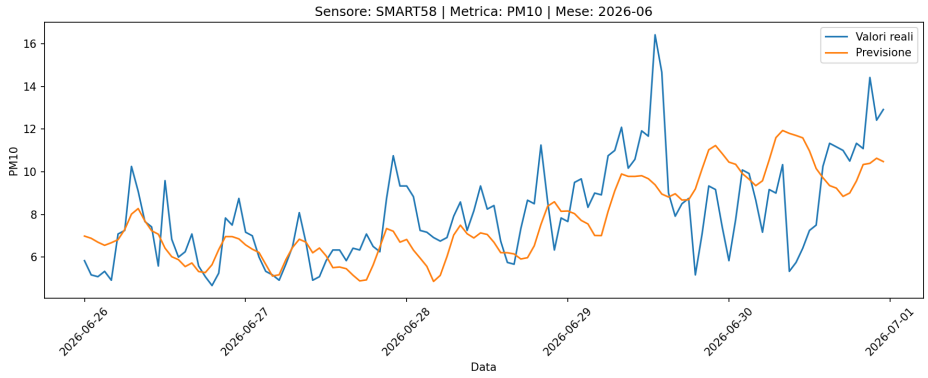


Figura 8.2: Previsione della metrica PM10 per la centralina SMART58 mediante NeuralProphet.

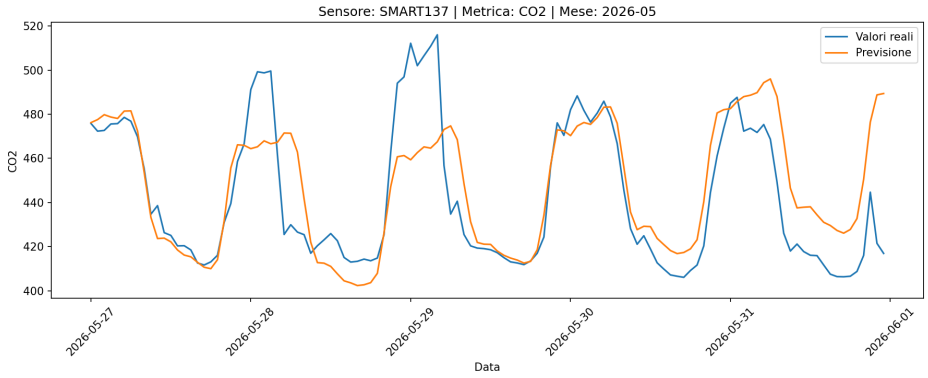


Figura 8.3: Previsione della metrica CO_2 per la centralina SMART137 mediante NeuralProphet.

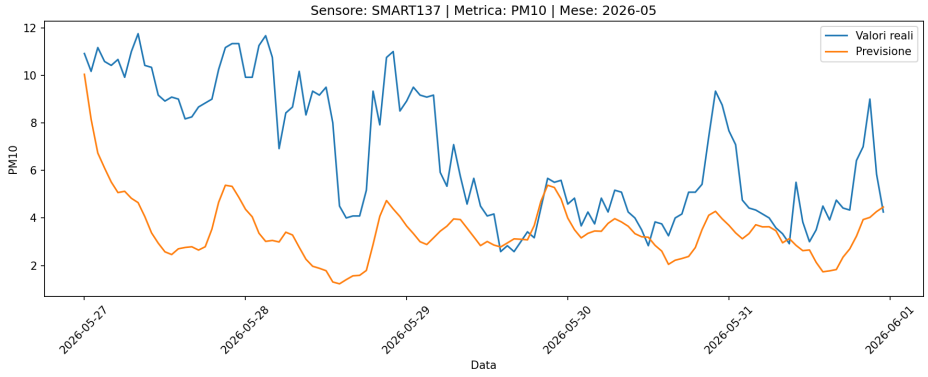


Figura 8.4: Previsione della metrica PM10 per la centralina SMART137 mediante NeuralProphet.

8.3 Qualità dei dati e costo computazionale

Uno degli aspetti più interessanti emersi durante le sperimentazioni riguarda l'influenza della quantità di dati disponibili sulle prestazioni dei modelli. Questo aspetto è risultato particolarmente evidente nel caso di NeuralProphet.

È noto che, soprattutto nei casi di modelli caratterizzati da una componente neurale, la disponibilità di una storia temporale sufficientemente estesa rappresenti un elemento rilevante.

Un numero maggiore di osservazioni permette al modello di disporre di una più ampia rappresentazione dell'andamento della serie e apprendere con maggiore efficacia le dinami-

che ricorrenti delle metriche considerate. La quantità dei dati non costituisce, però, l'unico elemento da considerare: la presenza di valori mancanti e di anomalie può ridurre il numero di informazioni effettivamente utilizzabile dal modello e influire sulla qualità delle previsioni.

La quantità di dati richiesta deve essere valutata anche in relazione al costo computazionale necessario per l'addestramento. Nel caso di NeuralProphet, l'aumento delle osservazioni utilizzate ha comportato un incremento dei tempi di elaborazione, rendendo l'utilizzo del modello progressivamente più oneroso all'aumentare della dimensione del dataset. Il miglioramento delle previsioni ottenuto deve essere considerato insieme alle risorse necessarie per ottenerlo.

Tale aspetto ha assunto particolare rilevanza nel confronto con il modello Prophet–LSTM. Quest'ultimo ha mostrato, nel corso delle sperimentazioni, un rapporto particolarmente interessante tra qualità delle previsioni e costo computazionale. La combinazione delle due componenti consente, infatti, di sfruttare la capacità di Prophet di modellare le componenti strutturate della serie, come trend e stagionalità, e quella della LSTM di modellare la componente residua, mantenendo tempi di elaborazione più contenuti.

La scelta di approfondire modelli come Prophet e NeuralProphet è legata alla loro capacità di gestire piccoli buchi nei dati nonchè di rappresentare differenti componenti delle serie temporali, quali il trend e le componenti stagionali. Questo aspetto risulta particolarmente interessante nel contesto della qualità dell'aria, il cui andamento è influenzato sia dalle condizioni ambientali sia dalle attività antropiche, come le variazioni del traffico veicolare e dei cicli produttivi.

Conclusioni

Il lavoro svolto ha avuto come obiettivo lo studio e la valutazione di differenti metodologie per la previsione dell'andamento degli inquinanti atmosferici, analizzandone le prestazioni in relazione alle caratteristiche dei dati disponibili. In una prima fase è stato analizzato un dataset costituito dalle osservazioni raccolte da cinque centraline AirQuino distribuite nell'area urbana di Pesaro. I dati, caratterizzati da una frequenza di campionamento oraria, sono stati sottoposti a operazioni di pulizia e organizzazione, attraverso le quali sono emerse alcune criticità legate principalmente alla limitata estensione temporale delle serie e alla presenza di valori mancanti e anomalie. Su questo primo dataset sono state sperimentate differenti metodologie di previsione, partendo dai modelli statistici ARIMAX e SARIMA e proseguendo con approcci basati sulla decomposizione delle serie temporali, reti neurali e modelli ibridi. In particolare, sono stati analizzati Prophet, NeuralProphet, LSTM, EMD-BiLSTM, reti neurali a grafo e il modello ibrido Prophet–LSTM. Le diverse metodologie sono state confrontate attraverso metriche quantitative e mediante l'analisi grafica delle previsioni, evidenziando prestazioni differenti in funzione della serie considerata e della metodologia adottata. Tra gli approcci sperimentati, il modello Prophet–LSTM ha mostrato i risultati più interessanti sul primo dataset, combinando la capacità di Prophet di rappresentare le principali componenti della serie con quella delle LSTM di modellare la componente residua non catturata dalla previsione iniziale.

Le criticità riscontrate nel primo dataset hanno evidenziato, tuttavia, l'importanza della disponibilità di una quantità maggiore di osservazioni per l'applicazione delle metodologie considerate. Per questo motivo l'analisi è stata successivamente estesa a un secondo dataset, fornito dall'ARPAM, comprendente diverse centraline distribuite sul territorio marchigiano e caratterizzato da un periodo di osservazione più esteso, pari a circa quattro anni, con frequenza di campionamento giornaliera. Il passaggio al nuovo dataset ha permesso di disporre di serie temporali più complete e ha consentito di verificare nuovamente il comportamento delle metodologie che avevano mostrato risultati più interessanti nelle precedenti sperimentazioni. In questa seconda fase è stato inizialmente utilizzato ARIMA come modello di riferimento, per poi applicare Prophet e il modello ibrido Prophet–LSTM. È stata, inoltre, introdotta una strategia di previsione rolling one-step, nella quale il modello viene progressivamente aggiornato mediante le nuove osservazioni disponibili, confrontandone il comportamento con quello della strategia multistep. Le sperimentazioni condotte sul secondo dataset hanno mostrato una maggiore aderenza tra valori osservati e previsti, in particolare nel caso del modello Prophet–LSTM.

Nel complesso, i risultati ottenuti hanno evidenziato come la qualità e la quantità dei dati disponibili costituiscano elementi determinanti per l'applicazione dei modelli previsivi

e come la scelta della metodologia debba essere valutata congiuntamente alla strategia di previsione e alle risorse computazionali necessarie. Un ulteriore elemento emerso durante il lavoro riguarda proprio il rapporto tra accuratezza delle previsioni e costo computazionale. Le sperimentazioni con NeuralProphet hanno mostrato un miglioramento delle prestazioni all'aumentare della quantità di dati utilizzata per l'addestramento, ma anche un incremento dei tempi necessari per l'elaborazione. Il modello Prophet–LSTM ha evidenziato un compromesso più favorevole tra qualità delle previsioni e risorse computazionali, risultando, pertanto, particolarmente interessante nel contesto delle sperimentazioni effettuate.

Il lavoro svolto può essere considerato come un punto di partenza per ulteriori sperimentazioni, volte ad approfondire il ruolo dei dati, delle metodologie previsive e delle strategie di previsione nel contesto della qualità dell'aria. Un primo possibile sviluppo riguarda la disponibilità dei dati. L'impiego di sensori caratterizzati da una maggiore accuratezza e l'estensione delle campagne di raccolta permetterebbero di ottenere serie temporali più complete e rappresentative.

Parallelamente, potrebbe essere approfondito l'impiego di tecniche specifiche per la gestione dei valori mancanti e delle anomalie, con l'obiettivo di aumentare la quantità di informazioni effettivamente utilizzabile durante l'addestramento.

Un ulteriore sviluppo potrebbe riguardare l'integrazione di dati ausiliari all'interno dei modelli predittivi. Le concentrazioni degli inquinanti possono essere influenzate non soltanto dall'evoluzione storica della singola variabile, ma anche dalle interazioni con altri inquinanti e da fattori esterni. Potrebbe, quindi, essere valutata l'integrazione di informazioni meteorologiche, dati satellitari e caratteristiche geografiche dell'area analizzata, verificando il contributo di tali informazioni al miglioramento delle previsioni.

Un ulteriore ambito di approfondimento riguarda la componente spaziale del problema. Sebbene nel corso dello studio sia stato valutato l'utilizzo congiunto delle informazioni provenienti da più centraline, tale approccio non ha prodotto i risultati attesi nel caso analizzato. Potrebbe, però, essere interessante proseguire la ricerca attraverso metodologie in grado di rappresentare esplicitamente le relazioni spaziali tra le diverse stazioni, ad esempio mediante modelli basati su grafi o approcci ibridi che combinino la componente temporale con quella spaziale.

Infine, potrebbe essere approfondito il confronto tra strategie rolling one-step e multistep, valutandone l'efficacia in funzione dell'orizzonte di previsione e della frequenza temporale dei dati. Un'analisi condotta su serie orarie, giornaliere e settimanali permetterebbe di verificare in quali condizioni sia più utile ottenere una previsione aggiornata del valore successivo e in quali, invece, sia preferibile disporre della previsione di un intervallo temporale più esteso.

Nel complesso, le sperimentazioni hanno evidenziato come la previsione della qualità dell'aria richieda un equilibrio tra disponibilità e qualità dei dati, accuratezza delle previsioni, scelta della metodologia, strategia di previsione e risorse computazionali. Gli sviluppi proposti potranno consentire di approfondire tali aspetti su serie temporali più estese e attraverso un insieme più ampio di informazioni, contribuendo a una valutazione più sistematica delle metodologie applicabili alla previsione degli inquinanti atmosferici.

Bibliografia

AFRICAN DEVELOPMENT BANK (2023), «Financing Climate Resilience and a Just Energy Transition in Africa», Rap. tecn., African Development Bank.

ASURAZL (2025), «EMD_BiLSTM», https://github.com/asurazl/EMD_BiLSTM.

BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. e LJUNG, G. M. (2015), *Time Series Analysis: Forecasting and Control*, Wiley.

HOCHREITER, S. e SCHMIDHUBER, J. (1997), «Long Short-Term Memory», *Neural Computation*, vol. 9 (8), p. 1735–1780.

MINISTERO DELLA SALUTE (2015), «Monossido di carbonio (CO)», Rap. tecn., Ministero della Salute, Roma.

RYBARCZYK, Y. e ZALAKEVICIUTE, R. (2021), «Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review», *Applied Sciences*.

SCHRAUFNAGEL, D. E. e OTHERS (2019), «Air Pollution and Noncommunicable Diseases: A Review by the Forum of International Respiratory Societies' Environmental Committee, Part 1», *Chest*, vol. 155 (2), p. 409–416.

SHUOWANG-AI (2025), «PM2.5-GNN», https://github.com/shuowang-ai/PM2.5-GNN.

TAYLOR, S. J. e LETHAM, B. (2017), «Forecasting at Scale», *PeerJ Preprints*, vol. 5, p. e3190v2.

TRIEBE, O., HEWAMALAGE, H., PILYUGINA, P., LAPTEV, N., BERGMEIR, C. e RAJAGOPAL, R. (2021), «NeuralProphet: Explainable Forecasting at Scale», *arXiv preprint arXiv:2111.15397*.

WORLD HEALTH ORGANIZATION (2021), «WHO Global Air Quality Guidelines: Particulate Matter, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide», Rap. tecn., World Health Organization, Geneva.

ZHANG e OTHERS (2025), «Assessing the Policy Gaps for Achieving China's Carbon Neutrality Goals», .

ZHANG, Y., BOCQUET, M., MALLET, V., SEIGNEUR, C. e BAKLANOV, A. (2012), «Real-time Air Quality Forecasting, Part I: History, Techniques, and Current Status», *Atmospheric Environment*, vol. 60, p. 632–655.

Sitografia

- AirQuino, sistema per il monitoraggio della qualità dell'aria – https://www.airquino.it
- ARPA Piemonte, Ossidi di Azoto – NOx – https://www.arpa.piemonte.it
- ARPA Umbria, Agenzia Regionale per la Protezione Ambientale dell'Umbria – https://www.arpa.umbria.it
- ARPAM, Agenzia Regionale per la Protezione Ambientale delle Marche – https://www.arpa.marche.it
- GitHub, repository "EMD_BiLSTM" – https://github.com/asurazl/EMD_BiLSTM
- GitHub, repository "PM2.5-GNN" – https://github.com/shuowang-ai/PM2.5-GNN
- NeuralProphet, NeuralProphet documentation – https://neuralprophet.com
- Prophet, Forecasting at scale – https://facebook.github.io/prophet/
- TeleAmbiente, "Qualità dell'aria, come avviene il monitoraggio?" – YouTube – https://www.youtube.com

Ringraziamenti

Vorrei dedicare questo spazio per ringraziare tutte le persone che mi hanno supportato durante il percorso di realizzazione di questa tesi.

Innazitutto alla mia famiglia. Ai miei genitori, Armando e Amedea, per avermi dato tutto il supporto e l'affetto di cui avevo bisogno, e per avermi sempre incoraggiata nelle mie scelte. A mio fratello Lorenzo e mia cognata Grazia, per essermi stati vicini e per avermi sempre sostenuta; siete riusciti a fare in modo che la distanza non pesasse così tanto. A mia cugina Laura perchè è sempre stata un punto di riferimento con un instacabile voglia di divertirsi. A mio zio Daniele e a Martina, alla dolcezza che mi riservano ogni volta che ci vediamo. A mio nonno Antonio, alle pizzette, ai gelati, all'anguria e alla frittata di spaghetti, al fatto che quando ti guardo ritorno sempre bambina. A nonna Margherita, nonna Ilde, nonno Aldo, zia Livia e zia Sidonia.

A Laika, Nina, Maddy e Betty.

Ai miei compagni di studio Laura, Giansimone, Walter e Luca; al fatto che ci siamo stretti la mano 6 anni fa e non ce la siamo più lasciata; grazie per tutto quello che abbiamo condiviso, per le risate, le chiacchierate e i momenti di sconforto. A Gaia C., Gaia P., Sofia, Giulia, Tiziano e Daniele; grazie perchè siete una famiglia meravigliosa e perchè mi sono resa conto quanto sia poco scontato avere delle persone con cui riuscire a mantenere questo genere di rapporto. A Elsi, all'impegno, alla delicatezza e alla poca voglia di vivere che la contraddistinguono. Sei forse l'unica persona che darebbe realmente un rene ad un amico, ed è una cosa meravigliosa. A Vitto, che ha assunto il ruolo di diario segreto che, anche se non vuole, emana una luce fortissima in grado di mettere a propio agio tutti quelli che ne sono colpiti, piuttosto conveniente per una dannata psicologa. A Nene; condividere questo traguardo con te è la cosa più bella che potesse accadere. Tempo fa di dicesti che ti sentivi sull'orlo di un precipizio e che ero arrivata a trattenerti dal cadere; ti devo confessare, però, che ci stiamo trattenendo a vicenda da anni, in un equilibrio perfetto, per il quale non esiste una senza l'altra. A Mirco, mi sembra superfluo aggiungere altro. A Mati, perchè quelle due volte l'anno che ci vediamo restano sempre le mie preferite e non vedo l'ora che arrivino. A tutti gli altri amici che mi hanno supportato e incoraggiato durante questo percorso.

Infine, desidero esprimere la mia gratitudine al mio relatore, il Prof. Domenico Ursino, per la sua guida, la pazienza e i preziosi consigli durante tutto il processo di ricerca, assieme al correlatore, il Prof. Gianluca Bonifazi.