



UNIVERSITÀ
DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E
DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

MONITORAGGIO DELL'ESPOSIZIONE AMBIENTALE NELLA SCLEROSI
MULTIPLA: SVILUPPO DI UNA PIATTAFORMA NELL'AMBITO DEL
PROGETTO BRAINTEASER

Candidato: Alessandro Puppi

Relatore: Prof. Pietro Bosoni

Correlatrice: Prof.ssa Arianna Dagliati

A.A. 2025/2026

SOMMARIO

1	INTRODUZIONE	1
1.1	Descrizione e obiettivi della tesi	1
1.2	Il progetto BRAINTEASER	3
1.3	Struttura della tesi	4
2	SCLEROSI MULTIPLA E INQUINAMENTO	5
2.1	Sclerosi Multipla	5
2.1.1	Forme e sintomi	6
2.1.2	Expanded Disability Status Scale	8
2.1.3	Cure e mortalità	9
2.2	Inquinamento atmosferico	9
2.2.1	Gli inquinanti e le loro linee guida	11
2.2.2	Air Quality Index	14
2.2.3	Inquinamento atmosferico e sclerosi multipla	15
3	INFRASTRUTTURA DATI E MODELLISTICA	17
3.1	Database AirQuality	17
3.1.1	Schema dei dati relativi agli inquinanti	18
3.1.2	Popolamento delle tabelle relative agli inquinanti	21
3.2	Dati aggiuntivi	24
3.2.1	Dati geografici	24
3.2.2	Dati ambientali	27
3.3	Dati Pazienti	35
3.4	Stima spazio-temporale degli inquinanti	37
3.4.1	L'Inverse Distance Weighting	38
3.4.2	Land Use Regression	45
3.5	I metodi ibridi	54
3.6	Valutazione comparativa dei modelli	55
3.6.1	Confronto nel Data Imputation	57
3.6.2	Confronto nella stima in un luogo privo di stazione	62
3.7	Applicazione dei metodi	67
4	PROGETTAZIONE E SVILUPPO DI DASHBOARD TRAMITE GRAFANA PER LA VISUALIZZAZIONE DEI DATI	70
4.1	Dashboard Home	72
4.2	Dashboard di analisi per area geografica	76
4.3	Dashboard per l'analisi della qualità dell'aria nel CAP selezionato	80
4.4	Dashboard Paziente	82

4.5 Dashboard per l'analisi della qualità dell'aria a livello provinciale e regionale	90
5 CONCLUSIONI E PROSPETTIVE FUTURE	94
APPENDICE A- Confronto dei modelli per l'imputazione dei dati in zone con bassa densità di stazioni	97
A.1 Confronto modelli per il SO ₂	97
A.2 Confronto modelli per il PM ₁₀	97
A.3 Confronto modelli per l'O ₃	97
A.4 Confronto modelli per il NO ₂	98
A. 5 Confronto modelli per il CO	98
A.6 Confronto modelli per il C ₆ H ₆	99
A.7 Confronto modelli per il PM _{2.5}	99
APPENDICE B- Confronto dei modelli per l'imputazione dei dati in zone con alta densità di stazioni	100
B.1 Confronto modelli per il SO ₂	100
B.2 Confronto modelli per il PM ₁₀	100
B.3 Confronto modelli per l'O ₃	101
B.4 Confronto modelli per il NO ₂	101
B.5 Confronto modelli per il CO	101
B.5 Confronto modelli per il C ₆ H ₆	102
B.7 Confronto modelli per il PM _{2.5}	102
APPENDICE C- Confronto dei modelli in comuni privi di stazioni all'interno di una zona geografica con bassa densità di stazioni	103
C.1 Confronto modelli per il SO ₂	103
C.2 Confronto modelli per il PM ₁₀	103
C.3 Confronto modelli per l'O ₃	104
C.4 Confronto modelli per il NO ₂	104
C.5 Confronto modelli per il CO	104
C.6 Confronto modelli per il C ₆ H ₆	105
C.7 Confronto modelli per il PM _{2.5}	105
APPENDICE D- Confronto dei modelli in comuni privi di stazioni all'interno di una zona geografica con alta densità di stazioni	107
D.1 Confronto modelli per il SO ₂	107
D.2 Confronto modelli per il PM ₁₀	107
D.3 Confronto modelli per l'O ₃	108

D.4 Confronto modelli per il NO ₂	108
D.5 Confronto modelli per il CO	109
D.6 Confronto modelli per il C ₆ H ₆	109
D.7 Confronto modelli per il PM _{2.5}	110
BIBLIOGRAFIA	111

INDICE DELLE FIGURE

Figura 2.1: Confronto tra un nervo sano e un nervo affetto da sclerosi multipla [7]	5
Figura 2.2: Punteggi della scala EDSS (0-10) e corrispondenti livelli di disabilità nella sclerosi multipla [11].	8
Figura 2.3: Mappa globale delle concentrazioni di PM _{2.5} nel 2025. La legenda in basso indica il livello di inquinamento, dalle tonalità chiare/verdi (aria pulita) a quelle scure/viola (aria sporca) [14]	10
Figura 2.4: Gli Interim Target (IT) e le Air Quality Guidelines (AQG) definiti dalla World Health Organization per i principali inquinanti atmosferici [16]	11
Figura 3.1: Diagramma ER del database AirQuality	20
Figura 3.2: Procedura di upload nella modalità file tramite API [35].	22
Figura 3.3: Andamento temporale della concentrazione media giornaliera dell'O ₃ dall'8 settembre 2021 al 7 marzo 2026	27
Figura 3.4: Schermata di scaricamento del database da E-OBSpre1950 [33].	29
Figura 3.5: Classi CLC presenti nella griglia 8x8 estratta per un punto geografico casuale.	33
Figura 3.6: strade rilevabili nella città di Pavia da OSM	34
Figura 3.7: strade rilevabili nel comune di Ghiffa da OSM.	35
Figura 3.8: Boxplot degli errori assoluti e relativi valori di RMSE per le diverse combinazioni di parametri PM _{2.5} . In verde è rappresentata la Penisola Iberica, in arancione l'Italia.	44
Figura 3.9: Boxplot degli errori assoluti commessi da ciascun metodo di Data Imputation per il CO nelle zone ricche di stazioni, suddivisi per contesto: tutte le stazioni (blu), urban (verde), suburban (rosso) e rural (arancione).	59
Figura 3.10: Boxplot degli errori assoluti commessi da ciascun metodo di Data Imputation per il PM ₁₀ nelle zone ricche di stazioni, suddivisi per contesto: tutte le stazioni (blu), urban (verde), suburban (rosso) e rural (arancione).	60
Figura 3.11: Boxplot degli errori assoluti commessi da ciascun metodo di stima per il CO nelle zone ricche di stazioni, suddivisi per contesto: tutte le stazioni (blu), urban (verde), suburban (rosso) e rural (arancione).	65
Figura 4.1: Variabili, time picker e menù presenti nella parte superiore della dashboard Home	72
Figura 4.2: Rappresentazione geografica delle stazioni di monitoraggio presenti su AirQuality mediante il pannello Geomap. In verde le stazioni in pianura, in rosso quelle in collina e in blu quelle in montagna.	74
Figura 4.3: Finestra informativa contenente le informazioni geografiche associate alla stazione di monitoraggio.	74
Figura 4.4: Suddivisione del numero di stazioni per area_type, altitude_zone e station_type. In basso, range temporale coperto da AirQuality.	75
Figura 4.5: Andamento temporale della concentrazione media giornaliera degli inquinanti misurati dalle stazioni di monitoraggio che rispettano i vincoli impostati dall'utente attraverso le variabili di selezione. In questo caso, tutte le variabili sono impostate su All. La legenda relativa agli inquinanti rappresentati, insieme ai rispettivi valori medio, massimo e minimo, è riportata in Figura 4.6.	75

Figura 4.6: Legenda degli inquinanti, con indicazione del relativo colore e dei valori medio, minimo e massimo.	76
Figura 4.7: Andamento temporale delle concentrazioni medie giornaliere di PM _{2.5} negli ultimi 5 anni, suddiviso per macroregione (Nord, Centro e Sud Italia). La legenda è riportata nella Figura 4.8.	77
Figura 4.8: Legenda della Figura 4.7. Codifica dei colori per le aree geografiche (Nord, Centro, Sud) e per le soglie normative (IT, AQG), integrata con i valori medi e massimi per macroregione	78
Figura 4.9: Pannello contenente le fasce di classificazione associate da ciascuna macroregione considerando i valori di PM _{2.5} rilevati negli ultimi 5 anni	79
Figura 4.10: Andamento temporale delle concentrazioni medie giornaliere di O ₃ negli ultimi 5 anni, suddiviso per fascia altimetrica (pianura in viola, collina in azzurro e montagna in blu).	80
Figura 4.11: Banner contenente l'indicazione dell'inquinante caratterizzato dalla maggiore frequenza di superamento della soglia Good presso la stazione di Piazza Minerva a Pavia negli ultimi cinque anni. È inoltre riportata la percentuale di misurazioni in cui tale soglia risulta superata.	81
Figura 4.12: Andamento temporale delle concentrazioni medie orarie di SO ₂ negli ultimi 5 anni misurate presso la stazione di Via Folperti a Pavia.	82
Figura 4.13: Informazioni relative alla qualità dell'aria e alla ricaduta nel comune di residenza del paziente n. 37735, con i relativi dati comunali.	84
Figura 4.14: Dashboard integrata per il monitoraggio del singolo paziente. L'interfaccia visualizza, dall'alto verso il basso: l'andamento dell'AQI nel comune di residenza nei 7 giorni antecedenti e successivi alla data selezionata; l'evoluzione dell'EDSS prima e dopo la data di riferimento; il quadro clinico aggiornato all'ultima visita effettuata.	85
Figura 4.15: Andamento stagionale dell'AQI in Galliate nei due anni precedenti il 10-10-2020	87
Figura 4.16: Andamento temporale della disabilità del paziente n. 37092, misurata attraverso l'EDSS_software (in verde) e l'EDSS_clinical (in giallo).	88
Figura 4.17: Andamento temporale della concentrazione media giornaliera di PM _{2.5} a Vercelli dal 15/5/2010 al 15/5/2012.	88
Figura 4.18: A sinistra, l'AQI stimato nei sette giorni antecedenti e successivi alla date_relapse selezionata. A destra è riportata la suddivisione dei livelli di AQI osservati nei due anni precedenti rispetto alla data selezionata.	89
Figura 4.19: Mappa raffigurante i valori di AQI associati ai diversi CAP italiani.	91
Figura 4.20: Andamento stagionale e mensile dell'AQI misurato a livello provinciale (a sinistra) e regionale (a destra) nei due anni precedenti alla date_relapse selezionata.	92
Figura 4.21: Andamento temporale della concentrazione media giornaliera di PM _{2.5} stimata nel comune di Pavia, nella relativa provincia e a livello regionale (Lombardia).	92
Figura 4.22: Valori minimi, massimi e medi di PM _{2.5} stimati nel comune di Pavia, nella provincia di Pavia e in Regione Lombardia.	93

INDICE DELLE TABELLE

Tabella 3.1: Sottoinsieme della tabella pollutants limitato agli inquinanti con misurazioni in measurements	21
Tabella 3.2: Conteggio annuale delle misurazioni e la relativa percentuale di misure valide presenti in AirQuality (2009-2026).	23
Tabella 3.3: Esempio della tabella view_cap_italy per il comune di Abano Terme. Le coordinate geografiche (LAT, LON) sono espresse in gradi decimali, l'altitudine in metri e il totale indica il numero di abitanti.	26
Tabella 3.4: Macrocategorie, sottocategorie e numero di classi del dataset CLC.	31
Tabella 3.5: Limiti massimi di concentrazione e relativi codici per gli inquinanti considerati (valori espressi in mg/m ³ per il CO e in µg/m ³ per gli altri).	42
Tabella 3.6: Valori ottimali di R (espressi in km) e p, con relativi valori di RMS, per ciascuna combinazione di macroarea (Penisola Iberica = IB, Italia = IT) e inquinante.	43
Tabella 3.7: I regressori utilizzati per il modello LUR. Per ciascun regressore sono riportati il codice identificativo assegnato all'interno del modello e il corrispondente codice della fonte originale	48
Tabella 3.8: Le restrizioni applicate sul segno dei coefficienti: valore positivo (>0) o negativo (<0)	51
Tabella 3.9: Regressori selezionati, intercetta e relativi coefficienti per ciascun inquinante.	53
Tabella 3.10: Metodi di Data Imputation ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e complessivo) in aree ad alta densità di stazioni.	58
Tabella 3.11: Metodi di Data Imputation ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e complessivo) in aree a bassa densità di stazioni.	61
Tabella 3.12: Metodi di stima ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e globale) in aree ricche di stazioni	63
Tabella 3.13: Metodi di stima ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e globale) in aree carenti di stazioni.	66
Tabella 4.1: Metodo selezionato e relativo codice associato per ciascun inquinante	84
Tabella A.1: Confronto nel data imputing in termini di RMSE dei modelli per il SO ₂	97
Tabella A.2: Confronto nel data imputing in termini di RMSE dei modelli per il PM ₁₀	97
Tabella A.3: Confronto nel data imputing in termini di RMSE dei modelli per l'O ₃	97
Tabella A.4: Confronto nel data imputing in termini di RMSE dei modelli per il NO ₂	98
Tabella A.5: Confronto nel data imputing in termini di RMSE dei modelli per il CO	98
Tabella A.6: Confronto nel data imputing in termini di RMSE dei modelli per il C ₆ H ₆	99

Tabella A.7: Confronto nel data imputing in termini di RMSE dei modelli per il PM _{2.5}	99
Tabella B.1: Confronto nel data imputing in termini di RMSE dei modelli per il SO ₂	100
Tabella B.2: Confronto nel data imputing in termini di RMSE dei modelli per il PM ₁₀	100
Tabella B.3: Confronto nel data imputing in termini di RMSE dei modelli per l'O ₃	101
Tabella B.4: Confronto nel data imputing in termini di RMSE dei modelli per il NO ₂	101
Tabella B.5: Confronto nel data imputing in termini di RMSE dei modelli per il CO	101
Tabella B.6: Confronto nel data imputing in termini di RMSE dei modelli per il C ₆ H ₆	102
Tabella B.7: Confronto nel data imputing in termini di RMSE dei modelli per il PM _{2.5}	102
Tabella C.1: Confronto modelli in termini di RMSE per il SO ₂	103
Tabella C.2: Confronto modelli in termini di RMSE per il PM ₁₀	103
Tabella C.3: Confronto modelli in termini di RMSE per l'O ₃	104
Tabella C.4: Confronto modelli in termini di RMSE per il NO ₂	104
Tabella C.5: Confronto modelli in termini di RMSE per il CO	104
Tabella C.6: Confronto modelli in termini di RMSE per il C ₆ H ₆	105
Tabella C.7: Confronto modelli in termini di RMSE per il PM _{2.5}	105
Tabella D.1: Confronto modelli in termini di RMSE per il SO ₂	107
Tabella D.2: Confronto modelli in termini di RMSE per il PM ₁₀	107
Tabella D.3: Confronto modelli in termini di RMSE per l'O ₃	108
Tabella D.4: Confronto modelli in termini di RMSE per il NO ₂	108
Tabella D.5: Confronto modelli in termini di RMSE per il CO	109
Tabella D.6: Confronto modelli in termini di RMSE per il C ₆ H ₆	109
Tabella D.7: Confronto modelli in termini di RMSE per il PM _{2.5}	110

INDICE DELLE EQUAZIONI

(2.1)	_____	14
(3.1)	_____	39
(3.2)	_____	39
(3.3)	_____	43
(3.4)	_____	52
(3.5)	_____	52

1 INTRODUZIONE

1.1 Descrizione e obiettivi della tesi

Le malattie neurodegenerative rappresentano una delle principali cause di morte e disabilità a livello globale. Secondo l'ultimo report dell'Organizzazione Mondiale della Sanità (World Health Organization, WHO) sullo stato globale della neurologia pubblicato nell'ottobre 2023, nel 2021 circa 3.4 miliardi di persone hanno sofferto di disturbi legati al sistema nervoso [1].

La WHO ha classificato 37 condizioni patologiche che possono causare danni neurologici di varia natura; tra queste, oltre all'ictus e alla malattia di Alzheimer che sono le più diffuse, rientra la sclerosi multipla (SM), alla quale sarà dedicato un approfondimento specifico nei capitoli successivi.

Negli ultimi anni, l'attenzione si è rivolta anche al possibile ruolo dei fattori ambientali nell'insorgenza e nella progressione delle malattie neurodegenerative. Tra questi, l'inquinamento atmosferico costituisce un elemento di particolare interesse, considerata la variabilità spazio-temporale dei livelli di esposizione e la loro potenziale influenza sul decorso clinico di tali malattie. In questo contesto si inserisce il presente lavoro di tesi, il cui obiettivo è fornire al medico uno strumento di supporto che consenta di monitorare i livelli di inquinamento atmosferico ai quali il paziente è esposto, con la possibilità di correlarli con i periodi di insorgenza delle ricadute e con il relativo quadro sintomatologico.

Pertanto, nel presente lavoro sono state sviluppate varie metodologie per stimare le concentrazioni dei diversi inquinanti atmosferici anche in località prive di una stazione di monitoraggio, utilizzando le informazioni disponibili presso le stazioni presenti sul territorio limitrofo e integrandole con variabili ambientali e meteorologiche.

A tal fine, il lavoro è partito dalla prima versione del database AirQuality, sviluppato nella prima fase del progetto BRAINTEASER (BRinging Artificial INTElligence home for a better cAre of amyotrophic lateral Sclerosis and multiple sclERosis), ampliato con i dati relativi alla qualità dell'aria provenienti

da 1694 stazioni di monitoraggio distribuite sul territorio spagnolo, portoghese e italiano. Successivamente, a partire dai valori rilevati presso le stazioni e integrando informazioni provenienti da misure meteorologiche e relative all'uso e alla copertura del suolo, sono stati sviluppati due differenti approcci di stima degli inquinanti: Land Use Regression (LUR) e Inverse Distance Weighting (IDW). I due approcci sono stati successivamente confrontati in differenti condizioni geografiche e di disponibilità delle stazioni, al fine di individuare le configurazioni maggiormente adatte alla stima delle concentrazioni in punti privi di monitoraggio.

Le stime ottenute sono state quindi utilizzate per caratterizzare l'esposizione agli inquinanti atmosferici nei luoghi di interesse per il paziente. In particolare, sono state considerate le coordinate associate al codice di avviamento postale (CAP) di residenza dei pazienti.

Infine, per rendere le informazioni ottenute facilmente consultabili e interpretabili dal personale medico, è stata sviluppata un'interfaccia grafica mediante Grafana. La dashboard integra le informazioni relative alla qualità dell'aria con i dati clinici disponibili, consentendo di visualizzare l'esposizione agli inquinanti in relazione al periodo temporale di interesse e, in particolare, agli episodi di ricaduta della SM.

L'obiettivo finale non è stabilire un rapporto causale tra l'esposizione agli inquinanti atmosferici e l'insorgenza delle ricadute, ma fornire al medico uno strumento di supporto che permetta di individuare eventuali condizioni ambientali particolarmente critiche per lo specifico paziente. Tali informazioni possono quindi essere utilizzate nell'ambito della valutazione clinica e, ove ritenuto opportuno, per fornire indicazioni volte a limitare l'esposizione a livelli di inquinamento particolarmente elevati.

1.2 Il progetto BRAINTEASER

Il progetto europeo BRAINTEASER nasce nel febbraio 2021 attraverso la collaborazione di 11 partner appartenenti a università, centri di ricerca e strutture ospedaliere provenienti da Spagna, Italia, Portogallo e Serbia [2].

Il coordinamento del progetto è affidato all'Università Politecnica di Madrid (UPM), mentre in Italia sono coinvolte l'Università degli Studi di Padova, l'Università degli Studi di Torino e la Fondazione Istituto Neurologico Nazionale Casimiro Mondino di Pavia, con il supporto del Laboratorio di Informatica Biomedica “Mario Stefanelli” dell’Università degli Studi di Pavia.

Questo progetto si basa sull'utilizzo dell'intelligenza artificiale e sull'elaborazione di grandi quantità di dati per analizzare la complessa interazione tra fattori ambientali, caratteristiche individuali, stili di vita e informazioni cliniche relative a due patologie neurodegenerative quali la SLA e la SM. La Fondazione Mondino e l’Università di Pavia, in particolare, hanno ricoperto un ruolo centrale nell'integrazione di dati ambientali all'interno di modelli basati sull'intelligenza artificiale, con l'obiettivo di studiare i fattori associati al rischio di sviluppo e alla progressione delle due patologie neurodegenerative.

Un altro obiettivo del progetto è ottimizzare la gestione del paziente tramite lo sviluppo di strumenti digitali e applicazioni in grado di supportare il medico nel processo decisionale, favorendo una diagnosi tempestiva e accurata e una gestione terapeutica maggiormente personalizzata.

Negli ultimi anni l'intelligenza artificiale ha assunto un ruolo sempre più rilevante in ambito sanitario, grazie alla capacità di analizzare grandi volumi di dati e individuare pattern difficilmente riconoscibili attraverso metodologie tradizionali. In questo contesto, BRAINTEASER rappresenta un esempio dell'evoluzione della medicina verso un approccio sempre più integrato con l'intelligenza artificiale, nel quale dati clinici, ambientali e personali vengono combinati per sviluppare strumenti innovativi a supporto della ricerca e della pratica clinica.

La collaborazione tra competenze mediche, scientifiche e tecnologiche ha permesso quindi di affrontare patologie neurodegenerative complesse attraverso

un approccio multidisciplinare, con l'obiettivo finale di migliorare la comprensione della malattia e le prospettive di cura dei pazienti.

1.3 Struttura della tesi

Il lavoro di tesi è stato suddiviso in due sezioni principali. La prima riguarda la componente di backend, nella quale saranno descritte l'aggiornamento del database di AirQuality, le sorgenti di dati utilizzati, le procedure di acquisizione e trattamento dei dati, nonché gli approcci metodologici adottati per la realizzazione dei modelli di stima degli inquinanti.

La seconda sezione riguarda invece la componente di frontend, nella quale verrà illustrata l'implementazione di un'interfaccia grafica realizzata mediante Grafana, con l'obiettivo di rendere le informazioni contenute nel database facilmente consultabili e interpretabili da parte del medico.

2 SCLEROSI MULTIPLA E INQUINAMENTO

2.1 Sclerosi Multipla

Nel 2023 si stimavano circa 2.9 milioni di persone affette da SM in tutto il mondo, mentre in Italia, secondo l'ultimo barometro pubblicato dalla Associazione Italiana Sclerosi Multipla (AISM) nel maggio 2026 in occasione della giornata mondiale della SM, i casi stimati sono circa 150 mila [3], [4].

La SM è considerabile una malattia autoimmune in quanto si basa sulla degradazione della guaina mielinica da parte del sistema immunitario, in particolare dall'infiltrato linfocitario perivascolare e dai macrofagi [5], [6].

Le guaine mieliniche sono un rivestimento di natura lipidica che avvolge quasi totalmente gli assoni del neurone, garantendo una trasmissione dello stimolo a velocità estremamente elevate. Grazie alla presenza di interruzioni tra un segmento di mielina e l'altro, detti nodi di Ranvier, l'impulso salta da un punto ad un altro rendendo il passaggio estremamente più rapido rispetto a quanto avverrebbe in assenza della guaina mielinica. L'attacco del sistema immunitario e le conseguenti lesioni mieliniche compromettono fortemente la conduzione saltatoria, determinando una significativa perdita della velocità di conduzione dello stimolo elettrico, causando deficit motori, cognitivi e di linguaggio. Tali lesioni, denominate placche, si sviluppano vicino a piccole vene, provocando inizialmente edema e successivamente danno all'assone. Le differenze tra un nervo sano e un nervo affetto da SM sono riportate nella Figura 2.1.

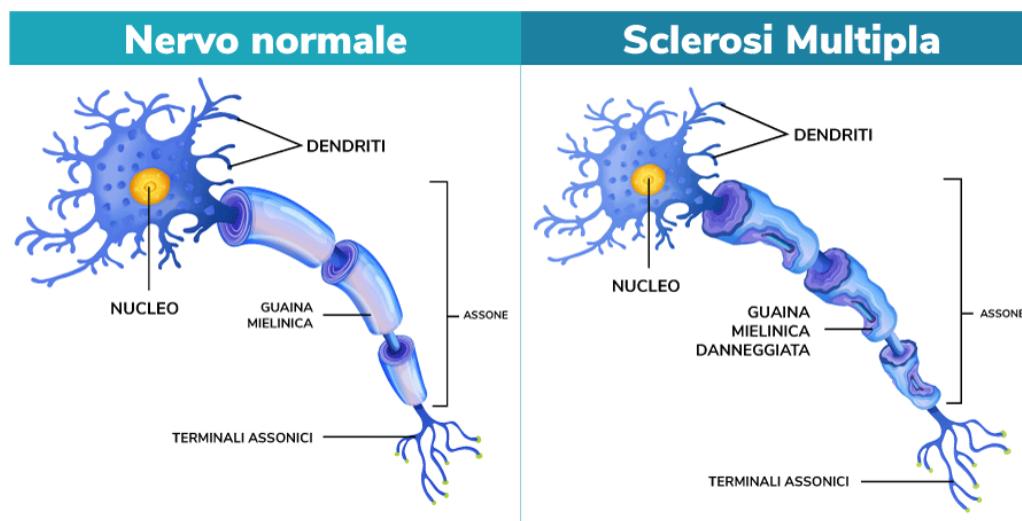


Figura 2.1: Confronto tra un nervo sano e un nervo affetto da sclerosi multipla [7]

Le cause dell'insorgenza della SM non sono ancora completamente note. Sono state formulate diverse ipotesi sui fattori eziologici della malattia; tra queste assumono particolare rilevanza i fattori genetici, i quali sono supportati anche da dati che mostrano una maggiore incidenza nel sesso femminile rispetto a quello maschile. In Italia, l'ultimo report pubblicato nel 2025 riporta un rapporto donne/uomini pari a 5:1 [4].

Sono stati identificati una decina di geni associati ad un significativo aumento della probabilità di insorgenza della patologia relativi alla via dell'ubiquitina-proteasoma, il sistema cellulare responsabile della degradazione delle proteine. Le alterazioni genetiche in questo punto portano a degradare sostanze errate, contribuendo allo sviluppo di patologie autoimmuni. Analoghe alterazioni, infatti, sono riscontrabili anche nel diabete di tipo 1 [6]. Nonostante ciò, l'origine genetica di tale patologia non è del tutto certa.

Per questo motivo, numerosi studi hanno approfondito anche il ruolo dei fattori ambientali. Studi recenti hanno mostrato una correlazione positiva molto forte tra l'incidenza della SM e l'area geografica di appartenenza. Questo, oltre a supportare le teorie genetiche, suggerisce che i fattori ambientali svolgono un ruolo importante nell'insorgenza della patologia [6],[7],[8].

Tra i fattori ambientali maggiormente studiati rientra l'inquinamento atmosferico. È ormai noto che un'esposizione prolungata ad aria inquinata sia dannosa per l'organismo; nel caso della SM, molti studi in letteratura sottolineano che tale esposizione può aumentare il rischio generale di sviluppare la malattia e favorisce la comparsa di ricadute più frequenti e più gravi [9].

2.1.1 Forme e sintomi

La SM è caratterizzata da un decorso clinico molto variabile. La frequenza delle ricadute, intesa come la comparsa di sintomi nuovi o l'acutizzarsi di sintomi preesistenti, consente di classificare la SM in diverse forme cliniche, caratterizzate da una differente gravità. Sono state classificate sette forme principali di SM [5]:

- **Sindrome clinicamente isolata (SIC):** consiste in un deterioramento isolato della guaina mielinica. È il primo evento della malattia;

- Recidivante-remittente (RR): è la forma più comune e la più tipica nelle fasi iniziali. È caratterizzata dalla comparsa di sintomi nuovi o dall'aggravamento di quelli preesistenti, della durata massima di due giorni;
- Progressiva primaria (PP): meno comune, caratterizzata da un progressivo peggioramento del quadro clinico fin dall'esordio;
- Secondariamente progressiva (SP): è l'evoluzione della forma RR, con un progressivo peggioramento che compare generalmente dopo 10-15 anni;
- Progressiva recidivante (PR): caratterizzata da un peggioramento continuo della disabilità, associato a ricadute acute sovrapposte;
- Fulminante: forma rare, caratterizzata da ricadute multiple e una rapida progressione della disabilità;
- Benigna: ricadute rare e di lieve entità, con progressione limitata della disabilità.

Ognuna di queste forme si manifesta attraverso sintomi estremamente variabili, che possono essere di lieve o di elevata gravità. In generale troviamo peggioramenti che coinvolgono:

- Disturbi visivi: dolore nei movimenti, perdita della vista, visione doppia;
- Disturbi vestibolari: vertigini, perdita di equilibrio.
- Disturbi motori: debolezza, affaticamento, tremolio
- Disturbi Cognitivi: memoria, funzioni esecutive e concentrazione.
- Disturbi Sensoriali: sensibilità, parestesie, disestesie.

Generalmente tali sintomi tendono a regredire spontaneamente nel giro di qualche giorno o settimana; tuttavia, con il progredire della malattia possono comparire fenomeni di riacutizzazione (esacerbazioni), caratterizzati da una maggiore frequenza e gravità degli episodi, con conseguente peggioramento del quadro clinico.

2.1.2 Expanded Disability Status Scale

L'accumulo progressivo dei deficit neurologici può determinare diversi gradi di disabilità. Per monitorarne l'evoluzione nel tempo e per valutare l'eventuale progressione della malattia viene comunemente utilizzata la Expanded Disability Status Scale (EDSS).

L'EDSS è una scala di valutazione che assume numeri da 0 a 10, nella quale il punteggio 0 indica la totale assenza di sintomi, mentre il punteggio 10 indica la morte a causa della SM.

I valori intermedi sono assegnati con incrementi di 0.5 punti e rappresentano livelli progressivamente crescenti di disabilità (Figura 2.2) [10].



Figura 2.2: Punteggi della scala EDSS (0-10) e corrispondenti livelli di disabilità nella sclerosi multipla[11].

Il valore della scala viene ottenuto valutando otto sistemi funzionali (FS):

1. Funzione visiva;
2. Funzioni del tronco encefalico;
3. Sistema piramidale;
4. Sistema cerebellare;
5. Sistema sensitivo;
6. Funzioni intestinali e vescicali;
7. Funzioni cerebrali;
8. Altre funzioni neurologiche.

Fino a un livello di EDSS pari a 4, il punteggio è determinato unicamente dal grado di compromissione di questi otto FS. Dal valore di 4.5, la capacità di deambulazione subentra nel calcolo per la definizione del punteggio finale. Nelle

varie visite presso il medico competente, dunque, il paziente viene valutato periodicamente attraverso questa scala, monitorando il progresso della malattia.

2.1.3 Cure e mortalità

Attualmente non esistono cure risolutive per la SM. Inoltre, data la sporadicità dei sintomi risulta complessa una diagnosi precoce. La diagnosi si basa principalmente sulla risonanza magnetica, integrata con l'analisi del sangue e del liquido cerebrospinale. Una diagnosi tempestiva è considerata fondamentale per rallentare la progressione della malattia. Infatti, se iniziati nelle fasi iniziali, i trattamenti modificanti la malattia (Disease-Modifying Therapies, DMT) risultano significativamente più efficaci [3].

Queste terapie agiscono modulando il sistema immunitario e consentono di ridurre la frequenza delle ricadute e rallentare l'accumulo della disabilità. Nel breve termine l'obiettivo delle DMT è ridurre l'attività infiammatoria e il numero delle nuove lesioni; nel lungo termine mirano invece a rallentare la progressione della malattia e limitare la comparsa della forma secondariamente progressiva. L'introduzione dei DMT ha determinato una significativa riduzione della mortalità, portandola a valori sempre più vicini a quelli della popolazione generale [7].

Di conseguenza, la SM è oggi una patologia generalmente più controllabile rispetto al passato e, nella maggior parte dei casi, meno aggressiva di quanto osservato nelle decadi precedenti.

2.2 Inquinamento atmosferico

L'inquinamento atmosferico rappresenta una dei possibili fattori di rischio per lo sviluppo e la progressione della SM [9].

Secondo il World Air Quality Report 2025 pubblicato da IQAir soltanto il 14% delle quasi 9000 città analizzate, distribuite in 143 paesi, rispetta i limiti di concentrazione del particolato fine (PM_{2.5}) raccomandati dalla WHO [12].

Osservando la mappa (Figura 2.3), emerge come l'Asia e l'Oceania siano i continenti più inquinati, dato confermato anche dalla classifica delle città con i livelli di PM_{2.5} più elevati [13]. In Europa, invece, le concentrazioni più elevate

le troviamo principalmente nell'area orientale, con la Bosnia-Erzegovina che guida la classifica. L'Italia occupa l'undicesima posizione, a causa dei picchi registrati nella Pianura Padana e nei principali centri urbani.

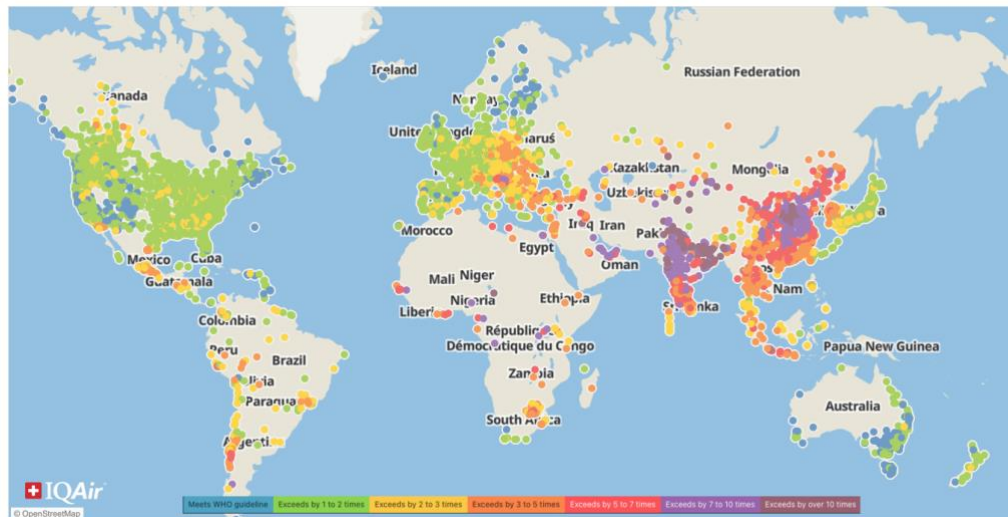


Figura 2.3: Mappa globale delle concentrazioni di PM2.5 nel 2025. La legenda in basso indica il livello di inquinamento, dalle tonalità chiare/verdi (aria pulita) a quelle scure/viola (aria sporca) [14]

Alla luce di questi dati, numerosi studi hanno analizzato gli effetti dell'esposizione cronica agli inquinanti atmosferici sulla salute umana.

L'University Medical Center di Amburgo, in Germania, ha condotto una importante review valutando l'impatto dell'esposizione ai principali inquinanti atmosferici considerando anche gli effetti del cambiamento climatico, responsabile dell'aumento delle temperature medie e la crescente frequenza di eventi meteorologici estremi [15]. Gli studi inclusi nella review hanno interessato le popolazioni europee e nordamericane, aree nelle quali ondate di calore, incendi boschivi ed eventi climatici estremi risultano sempre più frequenti. I risultati mostrano che l'esposizione prolungata a concentrazioni elevate di ozono (O_3) e al particolato inalabile (PM_{10}), associata a temperature elevate, è correlata a un incremento della mortalità per qualsiasi tipo di causa.

Inoltre, l'esposizione cronica agli inquinanti atmosferici, combinata con il calore estremo, è stata associata a un aumento del rischio di sviluppare diverse patologie croniche, tra cui malattie cardiovascolari, respiratorie, neurologiche e disturbi della salute mentale. In particolare, il $PM_{2.5}$, il PM_{10} , l' O_3 e le temperature elevate rappresentano i fattori maggiormente coinvolti in tali associazioni.

2.2.1 Gli inquinanti e le loro linee guida

La WHO periodicamente aggiorna le proprie linee guida sulla qualità dell'aria, definendo i livelli massimi compatibili con la tutela della salute pubblica [16].

Poiché le condizioni ambientali differiscono notevolmente tra le diverse aree geografiche, soprattutto tra Paesi industrializzati e Paesi in via di sviluppo, la WHO ha introdotto due differenti livelli di riferimento.

Il primo riferimento è rappresentato dagli Air Quality Guidelines (AQG) levels, che identificano i valori ottimali da raggiungere per garantire la massima protezione della salute. Il secondo riferimento comprende gli Interim Targets (IT), strutturati in un massimo di quattro livelli. Gli IT sono concepiti come step intermedi per i Paesi che presentano livelli di inquinamento particolarmente elevati, consentendo di pianificare una riduzione graduale delle concentrazioni fino al raggiungimento dell'AQG. La Figura 2.4 mostra ogni IT e di ogni AQG dei principali inquinanti.

Pollutant	Averaging time	Interim target				AQG level
		1	2	3	4	
PM_{2.5}, µg/m³	Annual	35	25	15	10	5
	24-hour ^a	75	50	37.5	25	15
PM₁₀, µg/m³	Annual	70	50	30	20	15
	24-hour ^a	150	100	75	50	45
O₃, µg/m³	Peak season ^b	100	70	–	–	60
	8-hour ^a	160	120	–	–	100
NO₂, µg/m³	Annual	40	30	20	–	10
	24-hour ^a	120	50	–	–	25
SO₂, µg/m³	24-hour ^a	125	50	–	–	40
CO, mg/m³	24-hour ^a	7	–	–	–	4

^a 99th percentile (i.e. 3–4 exceedance days per year).

^b Average of daily maximum 8-hour mean O₃ concentration in the six consecutive months with the highest six-month running-average O₃ concentration.

Figura 2.4: Gli Interim Target (IT) e le Air Quality Guidelines (AQG) definiti dalla World Health Organization per i principali inquinanti atmosferici [16]

Le soglie previste dalla WHO sono generalmente espresse come concentrazioni medie annuali o giornaliere. L'O₃ costituisce un'eccezione, poiché la sua concentrazione varia significativamente in funzione dell'irraggiamento solare e

della stagione; per questo motivo vengono utilizzati indicatori specifici, quali la massima concentrazione media su otto ore (maximum daily 8-hour mean) e la concentrazione media durante la stagione di massima esposizione (peak season) [15],[17].

Il PM_{2.5} comprende particelle solide e liquide sospese nell'aria con diametro aerodinamico inferiore o uguale a 2,5 µm. Grazie alle dimensioni estremamente ridotte, tali particelle riescono a raggiungere gli alveoli polmonari e, in parte, il circolo sanguigno. Pertanto, l'esposizione a concentrazioni elevate di PM_{2.5} può contribuire all'insorgenza e al peggioramento di patologie respiratorie e cardiovascolari. Le principali fonti emissive sono rappresentate dai processi di combustione, quali traffico veicolare, impianti industriali, riscaldamento domestico e incendi [18].

Il PM₁₀ comprende invece particelle con diametro aerodinamico inferiore o uguale a 10 micrometri. Rispetto al PM_{2.5}, presenta dimensioni maggiori, ma può comunque penetrare nelle vie respiratorie e, in presenza di determinate condizioni atmosferiche, essere trasportato per lunghe distanze. L'esposizione al PM₁₀ è stata associata a un aumento della mortalità generale e del rischio di patologie cardiovascolari e respiratorie. Le sue concentrazioni possono aumentare sia a seguito di fenomeni naturali, quali l'erosione delle rocce a causa del vento e le eruzioni vulcaniche, sia a causa di attività antropiche, principalmente legate a combustioni di vario genere [19].

Un altro importante inquinante atmosferico è l'O₃. È uno dei principali elementi che formano lo smog e si forma nell'atmosfera attraverso reazioni fotochimiche che coinvolgono principalmente ossidi di azoto e composti organici. Pertanto, presenta generalmente picchi nei mesi più caldi e nelle ore con maggiore irraggiamento solare. La capacità di diffondersi a grande distanza dai precursori originari consente a questo agente di contribuire all'inquinamento atmosferico anche in aree remote rispetto alle fonti di emissione principali. L'esposizione a concentrazioni elevate di O₃ può causare effetti negativi sulla salute simili a quelli riportati per il PM₁₀, con anche un possibile coinvolgimento nello sviluppo anche di malattie della pelle [20].

Il biossido di azoto (NO_2) è un gas caratterizzato dall'odore forte e da proprietà irritanti, sia per i polmoni che per gli occhi. Le principali fonti di emissione sono rappresentate dai veicoli, dai processi di riscaldamento e dalle attività industriali. L'esposizione prolungata è correlata soprattutto a patologie respiratorie, in particolare all'asma e al peggioramento della funzionalità polmonare, ma anche a problematiche in sistemi diversi da quello respiratorio, come irritazione degli occhi, della gola e del naso [21].

Un altro gas inquinante, in questo caso inodore, è l'anidride solforosa (SO_2), prodotta principalmente dalla combustione di combustibili contenenti zolfo, in particolare nell'ambito della produzione di energia elettrica e di alcuni processi industriali. L'esposizione a tale inquinante può causare effetti avversi nell'apparato respiratorio ed è pertanto correlata a patologie respiratorie [22].

Il monossido di carbonio (CO) è il gas più abbondante in atmosfera e deriva da una combustione incompleta di composti organici. Le principali sorgenti emissive comprendono il traffico veicolare, gli incendi boschivi e gli impianti di combustione. Il CO presenta un'elevata affinità per l'emoglobina, riducendo la capacità del sangue di trasportare ossigeno ai tessuti. Pertanto, l'esposizione a concentrazioni elevate può determinare sintomi quali cefalea, affaticamento e, nei casi più gravi, ipossia [23].

Infine, il benzene (C_6H_6) è un idrocarburo aromatico presente principalmente nelle emissioni dei veicoli e derivante da alcuni processi industriali. È classificato come cancerogeno certo per l'uomo ed è associato, in particolare, allo sviluppo di neoplasie del sistema ematopoietico, tra cui alcune forme di leucemia. Per tale motivo la WHO non identifica una soglia di sicurezza, mentre la normativa europea stabilisce un limite di concentrazione media annuale di C_6H_6 pari a $5 \mu\text{g}/\text{m}^3$ [24], [25].

2.2.2 Air Quality Index

L'andamento temporale della concentrazione degli inquinanti atmosferici, confrontato con i limiti raccomandati dalla WHO, rappresenta il metodo più accurato per valutare il livello di esposizione di un individuo all'inquinamento atmosferico. Tuttavia, essendo l'interpretazione diretta delle concentrazioni poco intuitiva, sono stati sviluppati diversi indici di qualità dell'aria il cui obiettivo è sintetizzare le concentrazioni dei principali inquinanti in un unico valore di immediata comprensione [26].

Tali indici sono comunemente utilizzati nelle mappe della qualità dell'aria e nelle applicazioni di monitoraggio ambientale [27], [28]. Tra gli indici maggiormente utilizzati figurano l'European Index (EAQI), sviluppato dall'Agenzia Europea dell'Ambiente (European Environment Agency, EEA) e l'Air Quality Index (AQI) proposto dalla United States Environmental Protection Agency (EPA) [29], [30].

L'EAQI classifica la qualità dell'aria in sei categorie qualitative (Good, Fair, Moderate, Poor, Very Poor ed Extremely Poor), determinate sulla base delle concentrazioni rilevate per i principali inquinanti atmosferici e dei corrispondenti intervalli di concentrazione definiti per ciascuna categoria. L'indice complessivo corrisponde alla categoria peggiore tra quelle associate ai singoli inquinanti misurati. Tale approccio consente una comunicazione immediata dello stato della qualità dell'aria, ma presenta una limitazione: tutte le concentrazioni appartenenti alla medesima categoria vengono considerate equivalenti, senza distinguere situazioni caratterizzate da differenti livelli di esposizione.

L'AQI dell'EPA adotta invece un approccio quantitativo. Per il CO, il NO₂, la SO₂, il PM₁₀, il PM_{2.5} e l'O₃, la concentrazione misurata viene convertita, mediante una funzione di interpolazione lineare basata su specifici breakpoint, in un indice numerico compreso tra 0 e 500. La conversione avviene mediante l'Equazione (2.1)

$$AQI = \frac{I_{HI} - I_{LO}}{BP_{HI} - BP_{LO}} (C - BP_{LO}) + I_{LO} \quad (2.1)$$

dove C rappresenta la concentrazione misurata dell'inquinante, BP_{LO} e BP_{HI} sono rispettivamente il breakpoint inferiore e superiore dell'intervallo in cui ricade la concentrazione, mentre I_{LO} e I_{HI} rappresentano i corrispondenti valori dell'AQI associati ai due breakpoint. I breakpoint sono definiti e aggiornati periodicamente dall'EPA e sono specifici per ciascun inquinante.

Successivamente, il valore ottenuto viene associato a una delle sei classi qualitative (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy e Hazardous). Analogamente all'EAQI, anche in questo caso l'indice finale corrisponde al valore peggiore tra quelli calcolati per i singoli inquinanti. Entrambi gli indici sono inoltre accompagnati da raccomandazioni rivolte alla popolazione, formulate in funzione del livello di qualità dell'aria rilevato. In particolare, l'EPA associa inoltre a ciascun intervallo di AQI specifici messaggi per la salute, differenziati per la popolazione generale e per particolari gruppi sensibili (ad esempio soggetti affetti da asma o patologie cardiovascolari), fornendo raccomandazioni sulle attività all'aperto, sui possibili effetti sulla salute e sulle precauzioni da adottare in funzione del livello di inquinamento.

Rispetto all'approccio a classi discrete dell'EAQI, l'utilizzo da parte dell'AQI di un indice numerico continuo consente di discriminare differenti livelli di esposizione anche all'interno della stessa categoria qualitativa. Per tale ragione, nel presente lavoro di tesi si è scelto di adottare l'indice proposto dall'EPA.

2.2.3 Inquinamento atmosferico e sclerosi multipla

L'esposizione cronica agli inquinanti atmosferici determina numerosi effetti biologici potenzialmente coinvolti nello sviluppo della SM. Tra i principali meccanismi studiati vi è l'aumento dello stress ossidativo, dovuto alla produzione di specie reattive dell'ossigeno (reactive oxygen species, ROS), che possono superare le capacità antiossidanti dell'organismo [31].

L'eccesso di ROS favorisce l'attivazione di processi infiammatori sistemici e il danno cellulare, contribuendo ad alterare l'integrità della barriera emato-encefalica. Tale alterazione potrebbe facilitare il passaggio di cellule immunitarie autoreattive, in particolare linfociti T, all'interno del sistema

nervoso centrale (SNC), dove possono partecipare alla risposta autoimmune diretta contro la mielina.

Oltre alla via ematica, gli inquinanti possono penetrare nel SNC anche per via trans-olfattiva, attraverso il sistema olfattivo.

Dal punto di vista clinico e radiologico, l'esposizione a breve e lungo termine al PM_{2.5}, al PM₁₀, al NO₂ e all'O₃ è stata associata non solo a un maggior rischio di insorgenza della malattia, ma anche a un incremento significativo del tasso di ricadute e alla comparsa di lesioni attive alla risonanza magnetica.

Inoltre, l'inquinamento atmosferico sembra interferire negativamente con la vascolarizzazione cerebrale, favorendo la formazione di lesioni e "buchi neri" cronici per carenza di apporto ematico e ossigenazione. Infine, l'ingresso di inquinanti nel SNC è in grado di indurre modificazioni epigenetiche ed attivare la microglia, la quale gioca un ruolo primario nella neurodegenerazione e nella progressione della disabilità valutata tramite l'EDSS.

Sebbene le evidenze attualmente disponibili non consentano ancora di dimostrare un rapporto causale diretto tra inquinamento atmosferico e SM, i risultati della letteratura scientifica supportano l'ipotesi che l'inquinamento rappresenti un importante fattore ambientale in grado di contribuire allo sviluppo e alla progressione della malattia [31].

3 INFRASTRUTTURA DATI E MODELLISTICA

Il presente capitolo illustra la tipologia di dati impiegati nel progetto di tesi, l'architettura aggiornata del database AirQuality e i modelli utilizzati per la stima delle concentrazioni dei principali inquinanti atmosferici. In prima istanza, verranno esaminati i dati ambientali e clinici forniti nell'ambito del progetto BRAINTEASER. Successivamente, l'analisi si focalizzerà su ulteriori sorgenti di dati, utilizzate come regressori all'interno delle metodologie di stima spazio-temporale adottate in questo lavoro di ricerca.

3.1 Database AirQuality

La prima versione del database AirQuality, sviluppata nel contesto del progetto BRAINTEASER, ospitava i dati relativi ai livelli di inquinamento atmosferico per il periodo compreso tra il 2013 e il 2025. Le informazioni derivavano da due fonti principali, entrambe open source e gestite nell'ambito di programmi dell'Unione Europea. La prima fonte era costituita dal portale della European Environment Agency (EEA), impiegato per reperire le concentrazioni dei diversi inquinanti misurate dalle stazioni di monitoraggio dislocate sul territorio europeo, unitamente alle relative coordinate geografiche [32].

A seguito di un aggiornamento del backend del servizio di download del portale EEA avvenuto nell'ottobre 2025, durante questo lavoro di tesi si è resa necessaria una revisione delle procedure di scaricamento e di integrazione dei dati. Tale modifica ha richiesto un aggiornamento dell'architettura del modello Entità-Relazione (ER), finalizzato all'inserimento di nuove tabelle contenenti dettagli aggiuntivi sulle stazioni e sulle misurazioni. Il database è stato successivamente esteso includendo i dati rilevati fino al 30 aprile 2026. Questo aggiornamento ha compreso l'integrazione di nuove stazioni di monitoraggio e delle rispettive misurazioni per Italia, Spagna e Portogallo, oltre all'introduzione di ulteriori sorgenti dati, approfondite nel Capitolo 3.2.

La seconda fonte informativa era costituita dal portale Copernicus Climate Change Service [33]. Anche in questo caso, la pipeline di pre-elaborazione è

stata aggiornata per consentire l'integrazione e l'estensione delle serie storiche fino al 30 aprile 2026

3.1.1 Schema dei dati relativi agli inquinanti

Il nucleo del database è rappresentato dalla tabella `measurements`, la quale contiene le misurazioni degli inquinanti effettuate dalle diverse stazioni di monitoraggio.

Ogni riga corrisponde a una singola misura, caratterizzata da attributi (variabili) presenti nelle colonne. Alcune di queste variabili costituiscono chiavi di collegamento con le altre tabelle del database, consentendo quindi di associare a ciascuna misurazione informazioni aggiuntive relative, ad esempio, alla stazione, al punto di campionamento, al Paese e all'inquinante selezionato.

Le tabelle poste in relazione con la tabella `measurements` sono le seguenti:

- `Validity_flags`: accessibile tramite la variabile `validity`, contiene le diverse tipologie di validità associate a ciascuna misurazione. Un valore di -99 e -1 indica una misura considerata non valida, altrimenti un valore pari a 1, 2, 3, 4 indica una misura considerata valida;
- `Verification_status`: accessibile tramite la variabile `verification` e contiene le informazioni relative allo stato di verifica della misurazione. Un valore pari a 3 indica una misura non verificata, un valore pari a 2 indica una misura sottoposta almeno a una verifica preliminare e infine un valore pari a 1 indica una misura verificata.
- `Sampling_points`: indica il punto di campionamento e, di conseguenza, il sensore che ha effettuato la misura;
- `Stations`: accessibile tramite lo `station_code` presente nella tabella `sampling_points`, identifica la stazione di monitoraggio che effettua la misura e permette di reperire le relative informazioni, compresa la posizione geografica;
- `Countries`: accessibile anch'essa tramite la tabella `sampling_points`, contiene la codifica delle nazioni presenti nel database, ovvero Italia (IT), Spagna (ES) e Portogallo (PT);

- Pollutants: contiene i codici(pollutant_code) utilizzati per identificare i diversi inquinanti presenti nel database, riportati in Tabella 3.1

È presente, inoltre, la tabella postcode_places. Quest'ultima non risulta collegata direttamente alla tabella measurements e contiene informazioni relative alla geolocalizzazione delle località appartenenti ai tre Paesi analizzati, ovvero Italia, Spagna e Portogallo.

Ogni riga identifica uno specifico punto geografico attraverso la variabile geonameid. Vengono classificati, oltre ai comuni e ai principali centri abitati, anche località di dimensioni ridotte e una più ampia varietà di punti di interesse geografici appartenenti. Sono infatti presenti strutture quali hotel, stazioni e altri elementi geografici, come i fiumi.

La struttura e le relazioni tra le varie tabelle del database sono rappresentate nel diagramma Entità-Relazione (ER) riportato in Figura 3.1

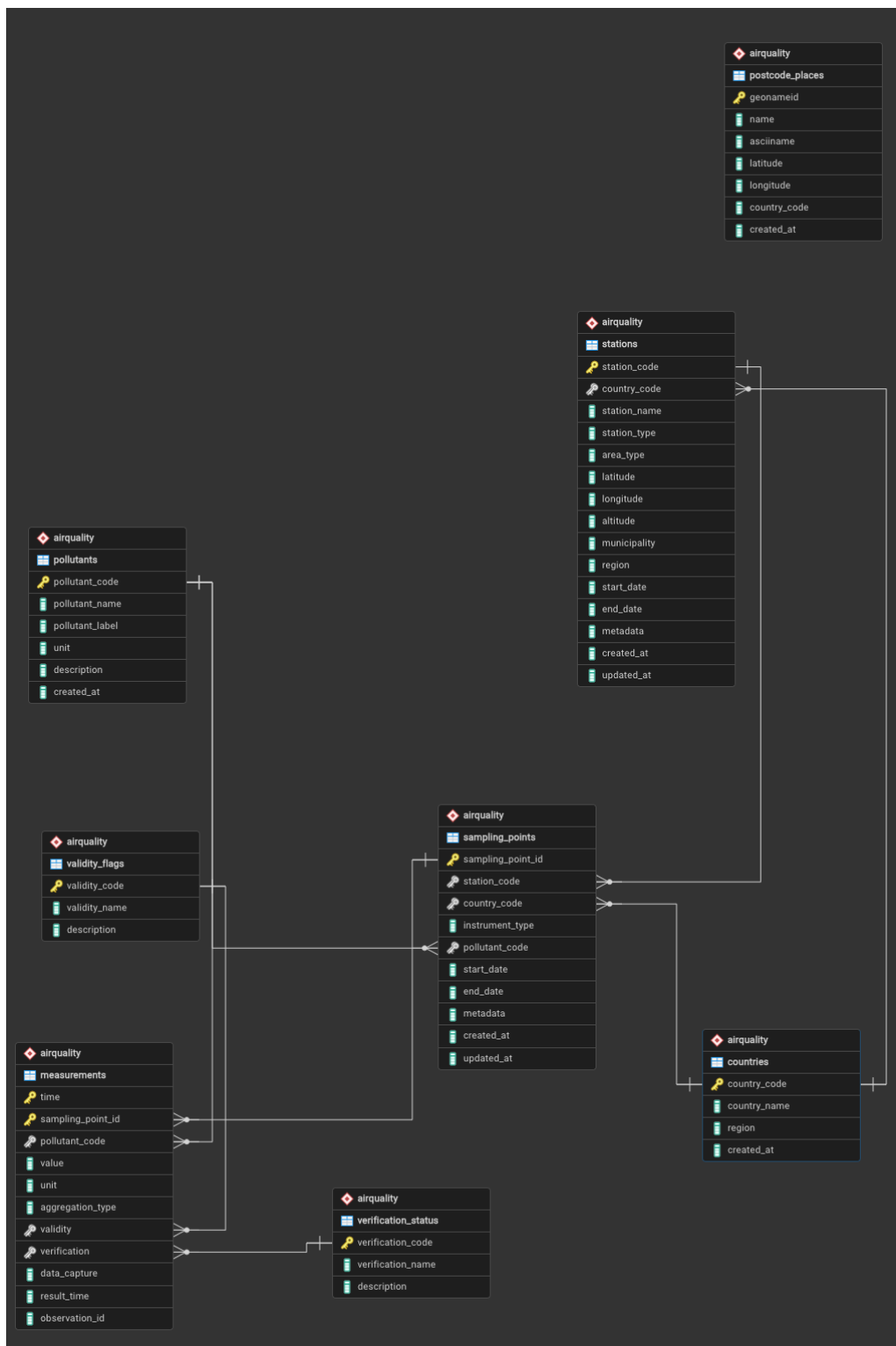


Figura 3.1: Diagramma ER del database AirQuality

Tabella 3.1: Sottinsieme della tabella pollutants limitato agli inquinanti con misurazioni in measurements

POLLUTANT_CODE	POLLUTANT_NAME	POLLUTANT_LABEL
1	SO ₂	Sulphur dioxide (air)
5	PM ₁₀	Particulate matter < 10 µm (aerosol)
7	O ₃	Ozone (air)
8	NO ₂	Nitrogen dioxide (air)
10	CO	Carbon monoxide (air)
20	C ₆ H ₆	Benzene (air)
6001	PM _{2.5}	Particulate matter < 2.5 µm (aerosol)

3.1.2 Popolamento delle tabelle relative agli inquinanti

La procedura di popolamento delle tabelle stations, sampling_points e measurements prevede innanzitutto il reperimento dei nuovi dati da inserire nel database.

Accedendo al portale dell'EEA è possibile selezionare i tipi di dati desiderati mediante appositi filtri relativi all'inquinante, al Paese e al periodo storico di interesse, per procedere successivamente al loro download [32], [34].

Per evitare di appesantire la fase di upload, i dati sono stati scaricati separatamente per ciascun Paese e suddivisi nelle due categorie temporali E1a e E2a.

La categoria E1a comprende i dati di misurazione verificati dal 2009 in poi; la categoria E2a comprende invece i dati di misurazione non verificati a partire dall'anno 2025.

Per quanto concerne le stazioni di monitoraggio, i relativi dati sono stati reperiti attraverso la sezione Metadata, dalla quale sono state scaricate le stazioni italiane, spagnole e portoghesi assenti nella prima versione della tabella stations. Entrambe le procedure di caricamento sono progettate per riconoscere eventuali dati già presenti, evitando così la presenza di duplicati. La fase di upload avviene tramite un'API dedicata, progettata all'interno del Laboratorio di Informatica Biomedica "Mario Stefanelli", ed è gestita mediante una procedura guidata [35].

Una volta scaricato il file di input nel formato list of URL, sono disponibili tre possibilità di caricamento dei dati nel database: sync, batch e file. Le modalità sync e batch sono adatte per caricamenti di dimensioni ridotte in quanto richiedono l'inserimento di singoli URL. Al contrario, per dataset di grandi dimensioni come quello in esame, viene impiegata la modalità file, la quale permette di caricare un file contenente un insieme di URL.

Nella Figura 3.2 è riportata l'interfaccia della modalità file: in questa schermata è sufficiente caricare nell'apposito campo (denominato File) il documento di testo precedentemente scaricato e contenente l'insieme degli URL, avviando così la procedura automatizzata di caricamento nel database.

ETL - Files Asynchronous ETL from uploaded Parquet files

POST /api/v1/etl/async/file Upload Url File

Upload a text file containing URLs for batch processing with SAFE concurrency control.
File format: One URL per line (text/plain).
Process:
1. Parse URLs from file
2. Divide into batches of 50 URLs
3. Execute max 3 batches concurrently (safe limit)
4. Return master job ID for tracking
Example: `bash curl -X POST http://localhost:8000/api/v1/etl/async/file \ -F "file=@urls.txt"`
Returns: Master job ID and processing information

Parameters

Cancel Reset

Name	Description
upsert boolean (query)	false

Request body required

multipart/form-data

file * required
string(binary)

Scogli file Nessun file selezionato

Figura 3.2: Procedura di upload nella modalità file tramite API [35].

Per evitare rallentamenti durante il caricamento e un'eccessiva occupazione dello spazio disponibile nel database, i file sono stati suddivisi in blocchi da circa 150-200 URL. Al termine di ciascuna operazione, è possibile verificare lo spazio disponibile sul sistema tramite codici da terminali, monitorando l'eventuale saturazione dello spazio associato al database AirQuality. Per ottimizzare ulteriormente lo spazio occupato, è inoltre possibile svolgere la compressione dei file inseriti, selezionando l'apposita opzione nell'interfaccia dell'API.

A seguito delle operazioni di aggiornamento e popolamento descritte, il database AirQuality comprende complessivamente circa 462 milioni di misurazioni provenienti da 1694 stazioni di monitoraggio. Circa l'89% delle misurazioni

risulta contemporaneamente verificato (verification <3) e valido (validity>=1). La distribuzione delle misurazioni in funzione dell'anno e dell'inquinante è riportata nella Tabella 3.2. Come si può notare, negli ultimi anni si registra un volume nettamente inferiore di dati validi, in quanto tali misurazioni non sono ancora state sottoposte al processo di verifica e convalida da parte dell'EEA.

Tabella 3.2: Conteggio annuale delle misurazioni e la relativa percentuale di misure valide presenti in AirQuality (2009-2026), più basse negli ultimi due anni.

ANNO	SO ₂	PM ₁₀	O ₃	NO ₂	CO	C ₆ H ₆	PM _{2.5}	VALIDE
2009	0	0	2637271	0	0	0	0	100
2010	0	0	2991439	0	0	0	0	100
2011	0	0	3167827	0	0	0	34004	100
2012	0	0	3220607	0	0	0	33997	99.99
2013	4184204	2885381	6101496	8493284	3834229	1220101	1207170	94.19
2014	3305193	3229213	5923791	8510506	3366057	1154728	1391348	95.68
2015	4661254	3315299	6042958	9101323	3475196	2255410	1404905	95.00
2016	4806238	3512341	6358504	9419947	3566222	2308166	1501344	95.05
2017	5028137	3772080	6578772	9872675	3663192	2519780	1604155	92.38
2018	5167845	3849210	6992235	10348959	3746835	2675007	1666757	92.86
2019	5285952	3973940	7075503	10369122	3752758	2748840	1815791	92.34
2020	5376326	4365894	7211335	10560670	3737375	2923553	1979305	91.96
2021	5546644	4464271	7339692	10820319	4035987	2969769	2455345	92.47
2022	5329718	5072460	7295030	10832860	3854173	2969385	2875496	92.47
2023	5232282	5182947	7247857	10759536	3773812	2834340	3022662	93.26
2024	5218919	5408007	7302813	10754671	3760252	2913808	3261985	93.83
2025	4438247	5413884	6970358	10117475	3425386	2351053	3257393	50.44

ANNO	SO ₂	PM ₁₀	O ₃	NO ₂	CO	C ₆ H ₆	PM _{2.5}	VALIDE
2026	1164030	926576	1102893	1678209	572473	402458	531480	5.61

3.2 Dati aggiuntivi

I dati riportati nella presente sezione costituiscono informazioni aggiuntive rispetto a quelle contenute nel database AirQuality e sono stati raccolti separatamente, al fine di integrare le informazioni necessarie per le successive analisi.

3.2.1 Dati geografici

Per poter visualizzare le distribuzioni degli inquinanti in relazione ai comuni di residenza di un paziente è stato necessario predisporre un elenco completo dei codici di avviamento postale (CAP) definiti all'interno del database come postcode, associando a ciascuno di essi i relativi valori di latitudine, longitudine e altitudine. Reperire un file completo e aggiornato da una sorgente open source affidabile risultato particolarmente complesso.

Un'ulteriore difficoltà è emersa per il fatto che i comuni possono cambiare nel tempo. In particolare, alcuni comuni possono essere soppressi o accorpati con altri, con la conseguente possibilità di una variazione dei CAP associati. Tali cambiamenti rendono quindi necessario fare riferimento, per quanto possibile, alla versione più aggiornata dell'elenco ufficiali dei comuni e dei relativi codici postali, al fine di garantire la coerenza e l'aggiornamento dei dati utilizzati.

Per le analisi all'interno di questo lavoro di tesi ci si è concentrati sulla costruzione di una tabella contenente i CAP italiani. In particolare, è stata usata una fonte secondaria che consente di rilevare, per l'intero insieme dei CAP, le coordinate di latitudine e longitudine aggiornate al 2026 [36]. Sebbene non si configuri come una banca dati primaria, tale risorsa è stata ritenuta idonea e considerata nell'analisi poiché, come attestato sulla piattaforma stessa, le informazioni derivano direttamente dai dati ufficiali forniti dall'Istituto Nazionale di Statistica (ISTAT), garantendo in tal modo un elevato livello di affidabilità e rigore scientifico

Oltre ai valori di latitudine e longitudine, sono stati associati a ciascun comune altri due parametri: l'altitudine e il numero di residenti.

Il numero di abitanti è stato considerato in quanto può rappresentare un potenziale fattore associato a un aumento dei livelli di inquinamento atmosferico. Numerosi studi lo utilizzano infatti come variabili esplicative con coefficiente positivo [37], [38]. In generale, a una maggiore popolazione possono corrispondere una maggiore presenza di attività antropiche e una maggiore rilascio di fonti emissive, tra cui il traffico veicolare, il riscaldamento degli edifici e le attività produttive. Di conseguenza, ci si può attendere un incremento delle concentrazioni degli inquinanti atmosferici. Il numero di abitanti per ogni comune italiano è stato reperito attraverso il sito dell'ISTAT [39].

Il file ottenuto dal sito riporta i residenti suddivisi per sesso e per età. Tramite uno script sviluppato in Python, è stato quindi possibile sommare le diverse categorie e ottenere il numero complessivo di residenti per ciascun comune.

Un altro parametro considerato è l'altitudine, fattore frequentemente preso in esame nei modelli presenti in letteratura. La revisione di Rueda-Torres, Warthon-Ascarza e Pacsi-Valdivia analizza nel dettaglio il legame tra altitudine e inquinamento atmosferico [40]. All'aumentare dell'altitudine si osserva una progressiva riduzione della pressione atmosferica, a cui si associa una fisiologica rarefazione degli insediamenti umani e, di conseguenza, una contrazione delle attività antropiche direttamente responsabili delle emissioni inquinanti.

In base alle relazioni descritte dalla legge dei gas ideali, una riduzione della pressione comporta una diminuzione della densità dell'aria e, secondo questo approccio, ci si può quindi aspettare di trovare una concentrazione minore dell'inquinante. Il rapporto tra altitudine e inquinamento è tuttavia più complesso. La concentrazione degli inquinanti è infatti influenzata da numerosi fattori meteorologici e atmosferici, quali temperatura, vento, umidità e condizioni di inversione termica, oltre che dalle caratteristiche del territorio e dalla distribuzione delle emissioni. Rimane comunque una variabile rilevante nei modelli di qualità dell'aria, anche in relazione all'esposizione della popolazione. Ad alta quota, la minore pressione atmosferica determina infatti una minore

densità dell'aria e modifica le condizioni di esposizione e respirazione. Pertanto, a parità di concentrazione dell'inquinante, gli effetti sull'organismo possono differire rispetto a quelli osservabili a quote inferiori. Questo aspetto evidenzia l'importanza dell'altitudine nella valutazione dell'esposizione all'inquinamento atmosferico, pur senza prevedere, nel presente lavoro, alcuna correzione dei valori in funzione della quota.

Per questo motivo, si è proceduto al reperimento dei dati relativi all'altitudine dei vari comuni italiani tramite l'apposita sezione sul sito dell'ISTAT [41]. I valori scaricati sono riferiti al dicembre del 2021. Nel frattempo, alcuni comuni italiani hanno subito alcune modifiche: alcuni comuni sono stati soppressi, mentre altri sono stati accorpati. Di conseguenza, il dataset risultava parzialmente ridotto e non perfettamente coincidente con l'elenco ufficiale e aggiornato di tutti i comuni, per i quali erano già disponibili i dati relativi alla popolazione [42], [43]. Si è proceduto a inserire manualmente i valori di altitudine relativi ai comuni di nuova istituzione, utilizzando le informazioni disponibili nel sito di Tuttitalia [44].

La medesima operazione è stata svolta anche per i valori di latitudine e longitudine. Al termine di questo processo è stato ottenuto un dataset completo, nel quale ogni comune italiano è associato a tutte le informazioni disponibili e necessarie al progetto, proveniente dalle diverse fonti consultate, tra cui i dati ISTAT.

Nella Tabella 3.3 sono schematizzate le varie variabili a disposizione per caratterizzare ogni comune.

Tabella 3.3: Esempio della tabella view_cap_italy per il comune di Abano Terme. Le coordinate geografiche (LAT, LON) sono espresse in gradi decimali, l'altitudine in metri e il totale indica il numero di abitanti.

CODICE_ISTAT	28001
COMUNE	Abano Terme
CAP	35031
SIGLA_PROVINCIA	PD
DENOMINAZIONE_PROVINCIA	Padova
CODICE_REGIONE	5
DENOMINAZIONE_REGIONE	Veneto
LAT	45,3593736

LON	11,7886941
ALTITUDINE	14
TOTALE	20479

È stata inoltre creata una seconda tabella, strutturalmente analoga alla precedente, contenente soltanto i capoluoghi delle province italiane.

Entrambe le tabelle, contenenti le informazioni relative a tutti i comuni e capoluoghi, sono state inserite nel database PostgreSQL tramite una semplice query CREATE su PgAdmin4. Successivamente, per facilitare l'accesso e l'utilizzo nelle successive attività di analisi, sono state create due View, denominate view_cap_italy e view_capoluoghi_italy.

3.2.2 Dati ambientali

Il meteo incide significativamente sui valori dei diversi inquinanti. Un esempio è l'O₃, la cui concentrazione presenta picchi durante i mesi estivi, in quanto la sua formazione è fortemente influenzata dalla radiazione solare. Come mostrato nella Figura 3.3, è infatti possibile osservare una maggiore concentrazione dell'O₃ nei periodi aventi una maggiore radiazione solare.



Figura 3.3: Andamento temporale della concentrazione media giornaliera dell'O₃ dall'8 settembre 2021 al 7 marzo 2026

Anche altri fenomeni meteorologici possono influenzare la concentrazione degli inquinanti. La presenza di vento, ad esempio, favorisce la dispersione e il

trasporto delle sostanze inquinanti, mentre le precipitazioni possono contribuire alla loro rimozione dall'atmosfera attraverso il fenomeno della wet deposition. Questo processo consiste nella rimozione degli inquinanti atmosferici mediante le precipitazioni, che possono inglobare o trascinare le sostanze presenti nell'atmosfera e depositarle successivamente sulla superficie terrestre.

Di conseguenza, la disponibilità di informazioni meteorologiche relative al periodo e alla posizione considerati rappresenta un elemento importante per poter ottenere una stima il più possibile verosimile con le condizioni ambientali effettivamente presenti.

Tra le fonti dei dati meteorologici analizzate rientrano i dataset ERA5 e ERA5-land, disponibili attraverso il servizio europeo Copernicus Climate Change Service e il dataset E-OBSpre1950, basato sui dati del progetto European Climate Assessment & Dataset [45], [46],[33].

ERA5 presenta una risoluzione temporale oraria e una risoluzione spaziale di $0,25^\circ \times 0,25^\circ$, mentre ERA5-land raggiunge una risoluzione spaziale di $0,1^\circ \times 0,1^\circ$ ed è focalizzato sulle superfici terrestri. Entrambi i dataset consentono inoltre di accedere a un numero molto elevato di variabili meteorologiche.

E-OBSpre1950, invece, fornisce dati meteorologici giornalieri a partire dal 1950 e permette di selezionare una risoluzione spaziale pari a $0,1^\circ$ o $0,25^\circ$.

La scelta della fonte dati è stata effettuata considerando, oltre alla qualità e alla varietà delle informazioni disponibili, anche le esigenze di gestione ed elaborazione dei dati all'interno del progetto. In particolare, il periodo di interesse, compreso tra il 2009 e il 2026, rende necessario considerare un numero considerevole di osservazioni. Una risoluzione temporale oraria, pur offrendo un elevato livello di dettaglio, comporterebbe infatti un volume di dati significativo, soprattutto considerando l'intero periodo di analisi. La risoluzione temporale giornaliera offerta da E-OBSpre1950 è risultata pertanto maggiormente adeguata alle esigenze del presente lavoro, consentendo una gestione più agevole dei dati nelle successive fasi di elaborazione.

Il dataset fornisce la possibilità di reperire otto variabili meteorologiche, che sono:

- Mean Temperature (TG): temperatura media giornaliera, misurata in gradi Celsius;
- Minimum Temperature (TN): temperatura minima giornaliera, espressa in gradi Celsius;
- Maximum Temperature (TX): temperatura massima giornaliera, espressa in gradi Celsius;
- Precipitation Sum (RR): quantità complessiva di precipitazioni giornaliera, espressa in millimetri;
- Sea Level Pressure (PP): pressione atmosferica media giornaliera ridotta al livello del mare, espressa in hPa;
- Relative Humidity (HU): umidità relativa media giornaliera, espressa in percentuale;
- Wind Speed (FG): velocità media giornaliera del vento, espressa in m/s. Non viene indicato la direzione del vento;
- Global radiation (QQ): quantità di radiazione solare globale ricevuta dalla superficie terrestre, espressa in W/m².

I dati sono stati reperiti per tutti i giorni compresi tra il 1° gennaio del 2009 e il 30 aprile 2026, ultima data disponibile al momento dello scaricamento dei dati. Nella Figura 3.4 è riportata la schermata di scaricamento con le due versioni di risoluzione spaziale disponibili.

Version 33.0e	Ensemble mean
0.1 deg. regular grid	TG TN TX RR PP HU FG QQ
0.25 deg. regular grid	TG TN TX RR PP HU FG QQ

Figura 3.4: Schermata di scaricamento del database da E-OBSpre1950 [33].

Il dataset è fornito in formato NetCDF e presenta una griglia regolare; nella configurazione scelta, con risoluzione 0.1°, la distanza tra i punti della griglia è di circa 10km, variabile in funzione della latitudine. Per ciascun punto della griglia è disponibile la serie storica delle variabili meteorologiche considerate.

Per associare i dati meteorologici a una specifica stazione di monitoraggio o a un determinato CAP, è stato sviluppato uno script in Python che, ricevute le relative coordinate geografiche, individua il punto della griglia più vicino. Per fare ciò, è stata utilizzata la libreria xarray, attraverso il metodo nearest.

In alcune aree, in particolare nelle zone costiere e insulari, il punto della griglia più vicino poteva tuttavia essere in mare e dunque non contenere dati meteorologici. È stato dunque necessario implementare una procedura aggiuntiva che ricerca, entro un raggio massimo di 0.4° , il punto più vicino contenente valori validi. Nel caso in cui non venga individuato alcun punto disponibile entro tale distanza, il dato non viene associato alla relativa stazione o al CAP.

Alla fine di questa procedura sono stati ottenuti tre dataset contenenti tutte le informazioni meteorologiche associate, rispettivamente, alle stazioni di monitoraggio, ai comuni italiani e ai capoluoghi di provincia italiani, includendo nel primo caso anche le stazioni presenti in Spagna e Portogallo.

Altre informazioni di particolare interesse per il modello LUR riguardano l'uso e la copertura del suolo. Tali variabili sono presenti in molti studi, tra cui quello di De Hoogh, poiché le concentrazioni dei vari inquinanti possono dipendere significativamente dalle caratteristiche del territorio circostante [37].

La presenza di determinate attività o caratteristiche territoriali può infatti influenzare localmente il valore degli inquinanti. Per esempio, la presenza di aree industriali come residenziali, infrastrutture o altre attività antropiche può determinare livelli di inquinamento differenti rispetto ad aree caratterizzate da vegetazione o corpi idrici. Per ottenere una stima più accurata è quindi necessario descrivere la composizione dell'area circostante il punto geografico considerato attraverso opportune informazioni relative all'uso e alla copertura del suolo.

A questo scopo è stato utilizzato il dataset CORINE Land Cover (CLC) sviluppato nell'ambito del progetto europeo COoRdination of INformation on the Environment (CORINE) per fornire informazioni omogenee sulla copertura e sull'uso del suolo del territorio europeo [47]. Nel presente lavoro è stata utilizzata la versione del 2018.

La classificazione CLC organizza il territorio europeo in 5 macrocategorie, suddivise complessivamente in 15 sottocategorie, che comprendono a loro volta 44 classi di copertura del suolo [48].

La struttura gerarchica della classificazione è riportata nella Tabella 3.4.

Tabella 3.4: Macrocategorie, sottocategorie e numero di classi del dataset CLC.

MACROCATEGORIA	SOTTOCATEGORIE	NUMERO DI CLASSI
Territori modellati artificialmente	1.1 Zone urbanizzate 1.2 Zone industriali, commerciali e reti di comunicazione 1.3 Zone estrattive, cantieri, discariche e terreni artefatti 1.4 Zone verdi artificiali non agricole	11
Territori agricoli	2.1 Seminativi 2.2 Colture permanenti 2.3 Prati stabili 2.4 Zone agricole eterogenee	11
Territori boscati e ambienti seminaturali	3.1 Boschi 3.2 Ambienti arbustivi ed erbacei 3.3 Aree aperte con vegetazione diradata o assente	13
Zone umide	4.1 Zone umide interne 4.2 Zone umide marittime	5
Corpi idrici	5.1 Acque continentali 5.2 Acque marine	4

Il dataset CLC presenta una griglia con risoluzione spaziale pari a 100 metri. A differenza dei dati meteorologici, la griglia è definita in coordinate metriche,

anziché in gradi decimali. È pertanto necessario trasformare le coordinate geografiche del luogo di interesse nel sistema di riferimento EPSG:3035, specificamente progettato per rappresentare il territorio europeo in coordinate metriche e consentire analisi spaziali coerenti sull'intera area. La trasformazione è stata effettuata mediante Python, utilizzando una funzione di conversione del sistema di riferimento che, a partire dalle coordinate geografiche espresse nel sistema EPSG:4326, restituisce le corrispondenti coordinate piane nel sistema EPSG:3035.

Una volta fatto ciò, è possibile definire un'area di interesse attorno al punto geografico considerato e calcolare la percentuale di superficie occupata da ciascuna delle 44 classi di uso del suolo. Per ogni luogo, pertanto, ciascuna classe assume un valore compreso tra 0 e 1, rappresentativo della proporzione di territorio occupata dalla specifica classe di suolo all'interno della zona considerata.

La Figura 3.5 mostra il funzionamento di questa procedura. Una volta individuato il punto geografico di interesse, viene considerata una porzione della griglia circostante composta da 8x8 celle di seguito. Ogni cella è associata ad una specifica classe, identificata attraverso un codice e rappresentata mediante un colore differente. Attraverso uno script in Python, per ogni CAP e stazione di monitoraggio è stata quindi calcolata la percentuale di superficie occupata da ciascuna classe rispetto all'area complessivamente considerata.

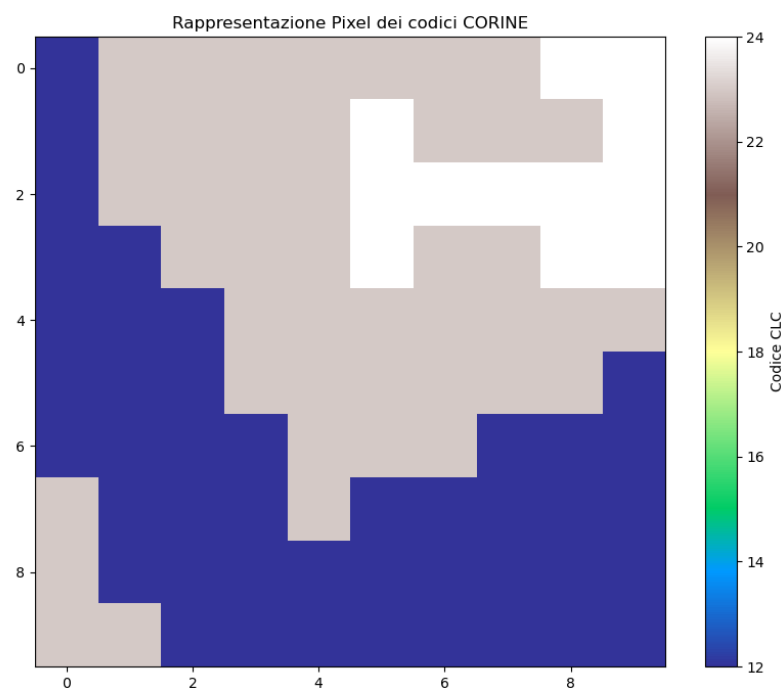


Figura 3.5: Classi CLC presenti nella griglia 8x8 estratta per un punto geografico casuale.

Le ultime informazioni necessarie per l'applicazione del modello LUR riguardano la presenza e la distribuzione della rete stradale nell'intorno del punto di interesse.

Anche CLC contiene informazioni relative alle infrastrutture stradali, tuttavia non risultano sufficientemente dettagliate per gli obiettivi del presente lavoro. Infatti, CLC permette di ricavare informazioni sulla copertura del territorio, ma non consente di ottenere la lunghezza complessiva delle strade presenti nell'area di interesse, eventualmente suddivisa in base alla loro tipologia.

Questa informazione risulta invece utile per il modello LUR e viene utilizzata, ad esempio, nello studio di De Hoogh, nel quale la presenza della rete stradale costituisce uno dei possibili predittori delle concentrazioni di inquinanti atmosferici [37].

A seguito della ricerca delle fonti disponibili, si è quindi scelto di utilizzare OpenStreetMap (OSM), un grande database geografico collaborativo contenente informazioni relative a numerosi elementi del territorio, tra cui la rete stradale.

La disponibilità di dati dettagliati sulla tipologia delle strade permette di calcolare la lunghezza complessiva delle strade presenti all'interno dell'area e di distinguere tra differenti categorie di strade.

Anche in questo caso, per poter effettuare correttamente in calcoli spaziali, le coordinate sono state convertite nel sistema di riferimento EPSG:3035, così da poter ottenere direttamente le distanze espresse in metri.

I vari tipi di strade disponibili nel dataset utilizzato sono riportati nella bibliografia del dataset [49].

A titolo di esempio, vengono confrontati due casi: la città di Pavia (Figura 3.6), rappresentativa di un contesto urbano, e il comune di Ghiffa (Figura 3.7) situato sul Lago Maggiore e caratterizzato da un territorio prevalentemente montuoso e dalla presenza del Sacro Monte della Santissima Trinità. Nel secondo caso è possibile osservare come il dataset permetta di rilevare anche strade di dimensioni ridotte, come i sentieri presenti nel parco, evidenziando l'elevato livello di dettaglio offerto dal dataset.

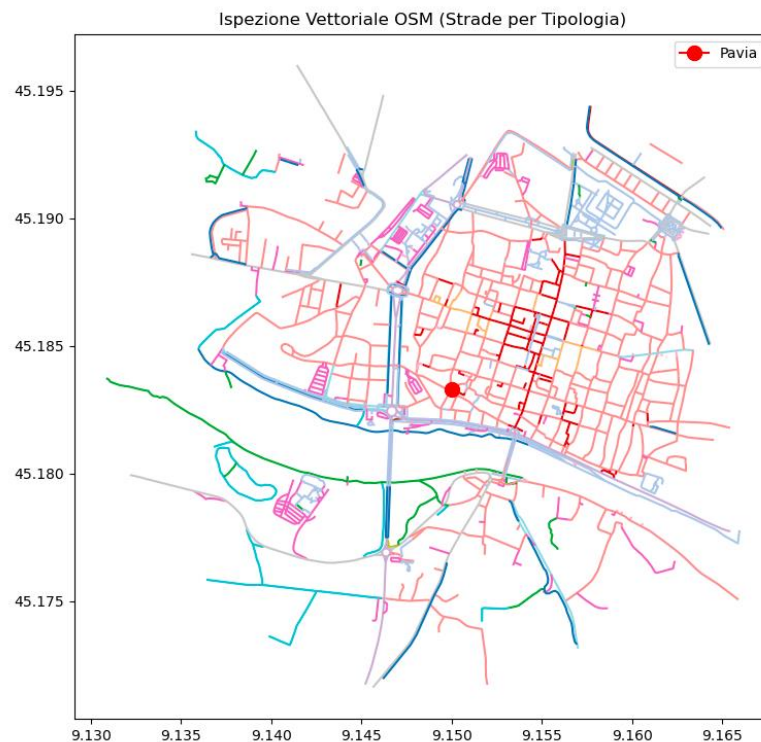


Figura 3.6: strade rilevabili nella città di Pavia da OSM

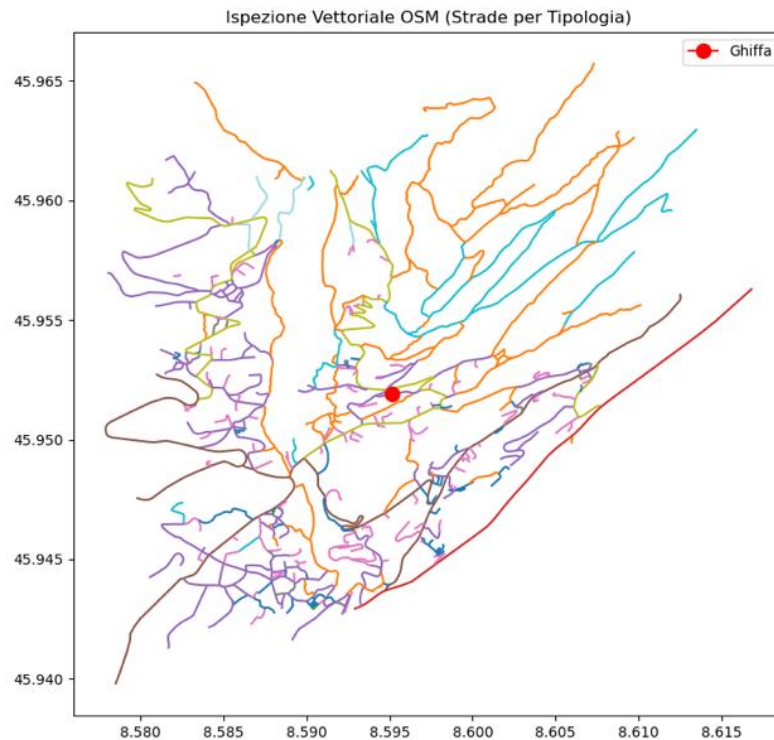


Figura 3.7: strade rilevabili nel comune di Ghiffa da OSM.

3.3 Dati Pazienti

Come discusso negli scorsi capitoli, il presente progetto di tesi si inserisce all'interno del progetto BRAINTEASER. Grazie a ciò è stato possibile accedere a dati clinici relativi a pazienti affetti da SM, forniti dalla Fondazione Istituto Neurologico Nazionale Casimiro Mondino di Pavia.

I dati sono presenti in un file Excel suddiviso in tre fogli distinti, in una versione ridotta e opportunamente anonimizzata.

Il primo foglio è denominato Anagrafica, caricato come View su PgAdmin4 dal nome `view_patients`, e contiene le informazioni relative a 1103 pazienti. Per ogni paziente, identificato tramite un opportuno codice, sono disponibili il comune di residenza, il relativo CAP e la provincia di appartenenza. Il codice del paziente è fondamentale per accedere alle informazioni presenti negli altri due fogli, relativi alle visite cliniche e alle eventuali ricadute.

Il secondo foglio, denominato Visite, contiene le informazioni relative alle visite effettuate dai pazienti. Ogni visita è codificata e comprende diversi parametri misurati durante la visita. In particolare, sono disponibili i punteggi relativi agli

8 FS utilizzati per il calcolo dell'EDSS. Per ciascuna visita sono inoltre presenti due differenti valori di EDSS: l'EDSS calcolato in modo automatico, applicando i criteri previsti dalla scala, e il valore di EDSS come valutato dal clinico, sulla base della valutazione clinica del paziente.

Sono inoltre disponibili informazioni relative all'evoluzione dello stato clinico rispetto alla visita precedente. Oltre alla possibilità di monitorare l'eventuale progressione della disabilità attraverso il valore di EDSS rilevato durante ciascuna visita, il database riporta anche l'eventuale insorgenza di neoplasie, infezioni gravi, herpes zoster o altre condizioni patologiche. Viene inoltre indicata l'eventuale presenza, al momento della visita, di trattamenti farmacologici o terapeutici specifici.

Il terzo e ultimo foglio contiene esclusivamente le informazioni relative alle ricadute sviluppate dal paziente nel corso della malattia. Questi dati rivestono particolare interesse ai fini del presente progetto. Poiché i fattori responsabili dell'insorgenza di una ricaduta in uno specifico momento non sono completamente noti, risulta interessante osservare l'eventuale esposizione del paziente a fattori di rischio ambientale nel periodo corrispondente alla manifestazione dell'evento.

Nel presente lavoro, l'attenzione è rivolta in particolare alla valutazione dell'esposizione all'inquinamento atmosferico nel periodo temporale associato alla ricaduta, verificando se il paziente sia stato esposto a concentrazioni al di sopra delle AQG raccomandate dalla WHO.

Per ciascuna ricaduta, sono disponibili le seguenti informazioni:

- Data di inizio e durata dell'evento;
- FS interessato;
- Livello di recupero dopo le relapse, classificato come completo o parziale;
- Impatto sulle attività della vita quotidiana (ADL);
- Gravità dell'evento;
- Tipo di trattamento somministrato, distinguendo tra trattamento cortisonico, ambulatoriale o domiciliare.

Non tutte le informazioni risultano presenti per ogni ricaduta. Tuttavia, se presenti, tali variabili consentono di fornire una descrizione completa e dettagliata dell'evento clinico e del relativo quadro evolutivo del paziente.

3.4 Stima spazio-temporale degli inquinanti

Per ottenere una misura diretta dell'esposizione sarebbe possibile, specie per studi condotti su coorti di pazienti di dimensioni ridotte, utilizzare campionatori personali. In letteratura viene descritto, ad esempio, l'uso di campionatori passivi per misurare la concentrazioni degli inquinanti negli ambienti frequentati dai pazienti, anche attraverso sensori indossabili [50].

Sebbene questa soluzione permetta di ottenere una misura molto verosimile all'effettiva esposizione individuale, essa presenta evidenti difficoltà nel caso in cui si voglia estendere l'analisi a un numero elevato di soggetti e a una ampia varietà di località geografiche. Nel presente lavoro, inoltre, non sono disponibili per tutti i luoghi geografici considerati misure dirette provenienti dalle stazioni di monitoraggio.

Le stazioni di monitoraggio che misurano tali inquinanti sono in numero limitato e distribuite in modo non uniforme sul territorio, creando di conseguenza aree caratterizzate da una densità di stazioni estremamente ridotta. Inoltre, una stessa stazione può essere temporaneamente inattiva: per motivi tecnici, in un determinato istante potrebbe non essere in grado di rilevare l'inquinante in questione, riducendo ulteriormente il numero di stazioni attive nell'area presa in considerazione.

Questa situazione comporta una disuniformità sia temporale sia spaziale dei dati, rendendo impossibile associare direttamente il valore registrato da una singola stazione ad un'intera città, poiché tale operazione porterebbe ad una stima non rappresentativa della reale concentrazione dell'inquinante nell'aria.

Per stimare il valore di un determinato inquinante in uno specifico punto geografico è stato dunque necessario sviluppare modelli di stima spazio-temporale che, ricevendo in ingresso le coordinate geografiche di una qualsiasi città, siano in grado di restituire una stima verosimile della concentrazione di inquinante relativo a quel periodo e a quella specifica zona. A seguito di ricerche

bibliografiche, tra le diverse metodologie impiegate per il monitoraggio e la stima della qualità dell'aria sono stati sviluppati l>IDW e il LUR, poiché rappresentano le metodologie maggiormente diffuse e utilizzate in letteratura [51],[37], [52], [53].

Per consentire lo sviluppo di queste due metodologie, è stata necessaria la costruzione di un dataset di dati aggiuntivi citati nel Capitolo 3.2 .

Infatti, per l>IDW è stato necessario reperire le coordinate geografiche dei comuni. Per il LUR, invece, sono stati necessari diversi dati ambientali, tra cui parametri meteorologici giornalieri, le informazioni relative all'uso del suolo e la distribuzione della rete stradale. Sebbene il LUR possa prevedere l'utilizzo di moltissime altre tipologie di variabili, la difficoltà nel reperire tali informazioni in modo affidabile, omogeneo, gratuito e soprattutto con una copertura completa per tutto il Paese considerato, ha portato a basare il modello su questo insieme di variabili, che rappresenta comunque un sottoinsieme significativo e adeguato agli obiettivi del progetto.

Per quanto riguarda il contesto geografico, sebbene il database AirQuality comprenda dati relativi a Italia, Spagna e Portogallo, la costruzione del dataset dei dati aggiuntivi è stata approfondita principalmente per il territorio italiano. Tale scelta è motivata dal focus del presente lavoro sull'applicazione dei metodi di stima spazio-temporale a una coorte di pazienti esclusivamente italiana e dalle maggiori difficoltà riscontrate nel reperimento di dati ambientali e geografici completi e omogenei per il contesto iberico.

3.4.1 L'Inverse Distance Weighting

L>IDW è uno dei metodi più diffusi per effettuare una stima spaziale. È utilizzato in numerosi studi relativi alla qualità dell'aria e anche in contesti affini [51], [54], [55].

Il principio alla base dell'algoritmo è intuitivo: dato un insieme di punti geografici nei quali è noto il valore delle variabili in studio, il valore relativo a un punto nel quale la variabile non è misurata viene stimato attribuendo un peso maggiore alle misure provenienti dai punti più vicini e un peso progressivamente minore a quelle provenienti dai punti più lontani.

Nel contesto del presente lavoro, i valori sono rappresentati dalle concentrazioni degli inquinanti misurate dalle stazioni di monitoraggio. Tali stazioni possono tuttavia trovarsi a distanze molto diverse dal punto per il quale si vuole ottenere la stima. Non è quindi possibile assumere che la concentrazione misurata da una stazione rappresenti allo stesso modo qualsiasi punto nell'area circostante. L'IDW permette di tenere conto di questa differenza, attribuendo maggiore peso alle misure provenienti dalle stazioni più vicine. L'Equazione (3.1) riporta la formula base dell'IDW, dove $u(x)$ rappresenta il valore stimato della variabile nel punto x , mentre u_i è il valore osservato presso la i -esima stazione di monitoraggio.

$$u(x) = \frac{\sum_{i=1}^N \frac{1}{d_i^p} u_i}{\sum_{i=1}^N \frac{1}{d_i^p}} \quad (3.1)$$

Nell'equazione (3.1) figurano due parametri. Il primo parametro è la distanza d tra la stazione di monitoraggio e il luogo per il quale si vuole ottenere la stima, calcolata sfruttando le coordinate geografiche sia delle stazioni sia dei diversi postcode italiani, spagnoli e portoghesi. In una prima fase si era ipotizzato di utilizzare direttamente la distanza euclidea bidimensionale. Successivamente si è scelto di utilizzare la distanza di Harvesine (Equazione (3.2)), maggiormente appropriata per il calcolo delle distanze sulla superficie terrestre in quanto tiene conto della curvatura della Terra.

$$d = 2r \arcsin \left(\sqrt{\sin^2((lat_2 - lat_1)/2) + \cos(lat_1) \cos(lat_2) \sin^2((lon_2 - lon_1)/2)} \right) \quad (3.2)$$

È stata inoltre valutata la possibilità di considerare, nella misura della distanza, anche la differenza di altitudine tra i punti considerati. Tuttavia, in assenza di metodologia sufficientemente consolidata in letteratura, tale approccio non è stato adottato nel presente lavoro.

Il secondo parametro è l'esponente p , che determina quanto rapidamente diminuisce il peso associato a una stazione all'aumentare della distanza. Un

valore maggiore di p attribuisce infatti una maggiore importanza alle misure più vicine, facendo diminuire più rapidamente il contributo delle stazioni lontane.

Per la scelta dei parametri sono stati valutati diversi possibili approcci.

Una prima possibilità consiste nel definire un numero minimo di stazioni da utilizzare per ciascuna stima. Tale approccio garantisce la disponibilità di un numero sufficiente di osservazioni, ma presenta alcune criticità nelle aree caratterizzate da una bassa densità di stazioni di monitoraggio. In questi casi, per raggiungere il numero minimo stabilito, sarebbe infatti necessario includere stazioni situate a distanze potenzialmente elevate e appartenenti a contesti geografici differenti.

Un approccio alternativo consiste nel definire un raggio massimo (R) entro il quale ricercare le stazioni di monitoraggio da utilizzare nella stima. Questo approccio consente di limitare l'analisi alle stazioni sufficientemente prossime al punto considerato, evitando di includere osservazioni provenienti da aree geograficamente troppo distanti.

La seconda strategia è stata ritenuta maggiormente adeguata agli obiettivi del presente lavoro. L'IDW attribuisce infatti un peso maggiore alle stazioni più vicine, indipendentemente dalla distanza assoluta alla quale esse si trovano.

Di conseguenza, la stazione più vicina riceverebbe comunque il peso maggiore anche qualora si trovasse a una distanza considerevole. Ad esempio, in un'area urbana la stazione più vicina potrebbe trovarsi a pochi chilometri, mentre in un'area montana potrebbe trovarsi a decine o centinaia di chilometri. In entrambi i casi la stazione più vicina avrebbe il peso maggiore, nonostante la diversa affidabilità associata alla distanza.

Per questo motivo, nel presente lavoro è stato definito un valore di R specifico per ogni inquinante, utilizzato per individuare le stazioni di monitoraggio da considerare nella stima. Sebbene R non compaia direttamente nella formula dell'IDW, esso consente di definire a priori quali stazioni possono essere incluse nel calcolo, considerando esclusivamente quelle situate all'interno dell'area stabilita da R .

Se all'interno di R non è presente alcuna stazione, non viene restituita una stima.

Se è presente una sola stazione, la stima coincide con il valore misurato da quest'ultima, mentre in presenza di più stazioni viene applicata la formulazione classica dell'IDW. Osservando in letteratura emerge che non esiste un valore universalmente valido per la scelta del valore di R e di p .

Per p vengono frequentemente utilizzati valori pari a 2 o a 3, mentre R può variare significativamente in base al contesto, in relazione con le caratteristiche dell'area di studio, dalla distribuzione delle stazioni e dagli obiettivi del lavoro. In studi relativi ad aree relativamente limitate sono stati utilizzati, ad esempio, valori di R nell'ordine di alcune decine di chilometri, mentre in contesti più estesi possono essere fissati valori maggiori [37], [53].

Nel presente lavoro, anziché fissare arbitrariamente tali parametri, si è deciso di determinarli attraverso una procedura basata sulla Leave-One-Out-Cross-Validation (LOOCV).

La procedura consiste nel rimuovere temporaneamente la misura di una specifica stazione di monitoraggio e nel tentare di ricostruirne il valore utilizzando le misure disponibili nelle stazioni circostanti. La stima viene effettuata mediante IDW, ripetendo il procedimento per ogni combinazione di R e p considerata. Il valore stimato viene quindi confrontato con il valore effettivamente misurato dalla stazione esclusa.

Sono state testate le combinazioni ottenute utilizzando come possibili valori di p 2 e 3, mentre per R sono stati considerati 15,25,50,75 e 100 km. Tali valori sono stati individuati sulla base degli approcci proposti in letteratura e successivamente adattati alle caratteristiche del presente lavoro, in particolare alla distribuzione spaziale delle stazioni di monitoraggio e all'estensione dell'area geografica considerata [53],[56].

Per limitare il costo computazionale e contemporaneamente ottenere un campione rappresentativo dell'intero periodo temporale, la procedura LOOCV è stata applicata a un sottoinsieme opportunamente campionato dei dati. Nel dettaglio, per garantire una distribuzione equilibrata sia dal punto di vista spaziale sia temporale nell'intero periodo compreso tra il 2009 e il 2026, sono state selezionate le prime sette misure giornaliere disponibili di ciascun inquinante nei mesi dispari di ogni anno.

Tali dati sono stati prelevati dal database AirQuality, considerando esclusivamente i valori giornalieri classificati come validi e verificati.

Nel caso in cui non fosse disponibile la misura giornaliera associata alla variabile `aggregation_type='day'`, l'algoritmo verificava la presenza, per la stessa stazione e nello stesso giorno, di almeno il 75 % delle misure orarie previste, soglia indicata in [57]. In caso positivo, tali misure venivano utilizzate per calcolare direttamente la media giornaliera dell'inquinante.

Sia per le misure giornaliere sia per quelle orarie sono state considerate esclusivamente misure valide ($\text{validity} \geq 1$) e verificate ($\text{verification} < 3$). Sono inoltre stati esclusi i valori negativi, presenti anche per misure ritenute valide e verificate. Infine, sulla base delle analisi dei dati disponibili, sono stati definiti dei valori massimi per ciascun inquinante. Tali soglie sono state fissate identificando concentrazioni palesemente non plausibili rispetto agli ordini di grandezza osservati nei dati. L'applicazione di tali limiti, presenti nella Tabella 3.5, ha consentito di escludere valori manifestamente errati, pur in presenza di indicatori di validità e verifica positivi.

Tabella 3.5: Limiti massimi di concentrazione e relativi codici per gli inquinanti considerati (valori espressi in mg/m^3 per il CO e in $\mu\text{g}/\text{m}^3$ per gli altri).

CODICE INQUINANTE	INQUINANTE	LIMITE MASSIMO
1,7	SO ₂ , O ₃	400
5	PM ₁₀	500
10	CO	50
20	C ₆ H ₆	200
6001	PM _{2.5}	450
8	NO ₂	600

In questo modo, è stato possibile ottenere un campione distribuito in maniera uniforme sull'intero periodo temporale considerato, corrispondente a circa l'11% dei dati complessivamente disponibili.

Il campione utilizzato è stato suddiviso in due macroaree geografiche: Italia e Penisola Iberica, quest'ultima comprendente Spagna e Portogallo, uniti sia a causa della minore disponibilità di stazioni di monitoraggio rispetto all'Italia, sia per la loro contiguità geografica. Tale scelta è rilevante in quanto, avendo individuato separatamente la combinazione ottimale di R e p per le diverse aree, una località situata in prossimità del confine può avere, all'interno del proprio valore di R, stazioni appartenenti ad entrambi i Paesi. Per ciascuna combinazione di parametri è stato calcolato il Root Mean Square Error (RMSE) tra il valore misurato dalla stazione (y_i) e quello stimato mediante IDW ($\hat{y}_i$). Il funzionale è riportato nell'equazione (3.3)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.3)$$

La combinazione avente il valore minimo di RMSE è stata selezionata come configurazione ottimale per lo specifico inquinante e per la relativa zona geografica. Una volta completata la procedura, sono stati ottenuti, per ogni inquinante e per ciascuna delle due aree geografiche, i valori di R e p ottimali. I valori delle configurazioni ottimali ottenuti sono riassunti nella Tabella 3.6, mentre la Figura 3.8 illustra il boxplot relativo al caso esempio del PM_{2.5}.

Tabella 3.6: Valori ottimali di R (espressi in km) e p, con relativi valori di RMS, per ciascuna combinazione di macroarea (Penisola Iberica = IB, Italia = IT) e inquinante.

INQUINANTE	REGIONE	CONFIGURAZIONE MIGLIORE	RMSE
SO ₂	IB	R=15; p=2	4.1682
SO ₂	IT	R=100; p=2	3.1029
PM ₁₀	IB	R=15; p=2	8.1370
PM ₁₀	IT	R=15; p=2	9.4213
O ₃	IB	R=15; p=2	10.8107
O ₃	IT	R=15; p=2	12.4899

INQUINANTE	REGIONE	CONFIGURAZIONE MIGLIORE	RMSE
NO ₂	IB	R=15; p=2	9.1785
NO ₂	IT	R=100; p=2	12.7965
CO	IB	R=75; p=2	0.2082
CO	IT	R=100; p=2	0.3301
C ₆ H ₆	IB	R=15; p=2	1.0193
C ₆ H ₆	IT	R=50; p=2	0.9381
PM _{2.5}	IB	R=25; p=2	4.7266
PM _{2.5}	IT	R=15; p=2	7.2433

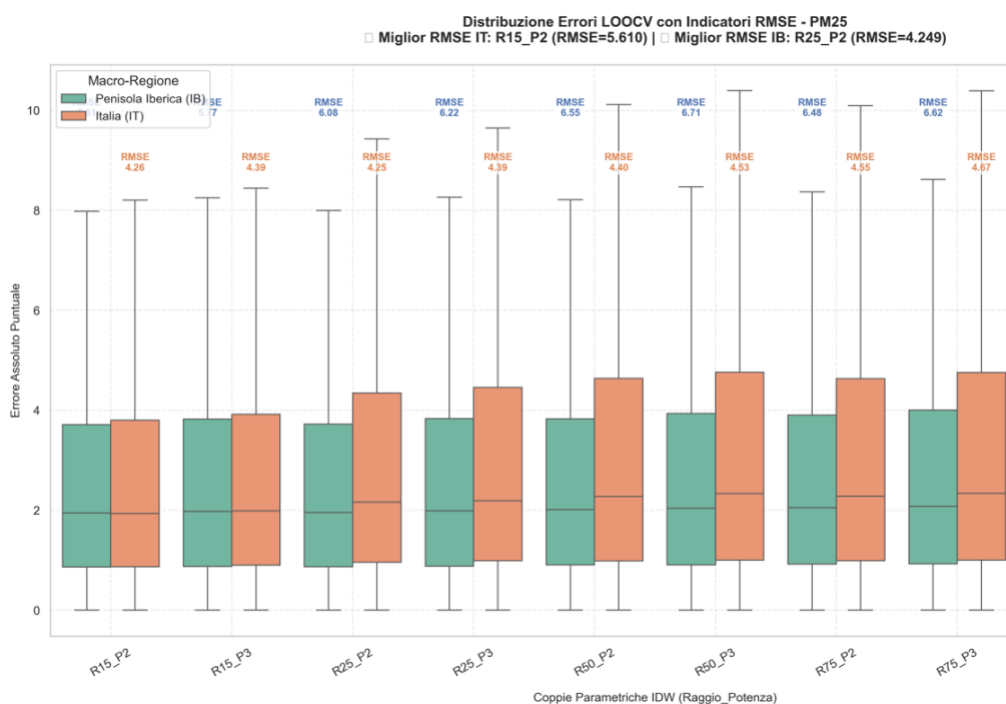


Figura 3.8: Boxplot degli errori assoluti e relativi valori di RMSE per le diverse combinazioni di parametri PM_{2.5}. In verde è rappresentata la Penisola Iberica, in arancione l'Italia.

Il boxplot sintetizza la distribuzione degli errori assoluti ottenuti per ciascuna combinazione di parametri: si nota come la Penisola Iberica mantenga una variabilità e un RMSE tendenzialmente inferiori rispetto all'Italia.

Una volta individuata la coppia ottimale di R e p per ciascun inquinante e area geografica, è stato implementato in Python l'algoritmo IDW definitivo.

L'approccio basato sul solo R presenta tuttavia una criticità; nei giorni in cui nessuna stazione attiva è presente all'interno di R, il metodo non è in grado di produrre una stima e pertanto restituisce un valore nullo.

Per ovviare a questa eventualità, è stata introdotta una procedura aggiuntiva. In particolare, il giorno viene escluso dalla stima solamente nel caso in cui, per l'inquinante considerato, non sia disponibile alcuna stazione nell'interno dataset. Se invece sono presenti stazioni attive ma nessuna di esse ricade all'interno del valore di R trovato per il relativo inquinante, viene utilizzata la stazione più vicina al punto da stimare, a condizione che questa si trovi entro una distanza massima di 200 km.

La scelta di tale valore deriva dall'osservazione del dataset: tendenzialmente, non è necessario estendere R al di sopra di quello ottimale. Nei casi in cui ciò si rende necessario, la stazione più vicina tende, comunque, a trovarsi a una distanza leggermente maggiore. È stato pertanto introdotto il limite di 200 km come soglia massima oltre la quale la stima viene considerata non sufficientemente affidabile.

L'algoritmo ottenuto, denominato IDW_vicino, permette quindi di ridurre il numero di stime nulle, mantenendo al contempo il vincolo di distanza massima ritenuta accettabile tra il punto da stimare e la stazione di monitoraggio utilizzata.

3.4.2 Land Use Regression

Il secondo metodo utilizzato per la stima spazio-temporale della concentrazione di inquinanti è la LUR, una delle metodologie, insieme alla IDW, maggiormente usate negli studi sulla qualità dell'aria.

Questo approccio statistico permette di stimare la concentrazione degli inquinanti atmosferici tramite l'utilizzo di diverse variabili predittive, definite regressori.

Nel presente lavoro sono state considerate variabili relative alle condizioni meteorologiche, alla copertura e all'uso del suolo, alla presenza di strade, all'altitudine e al numero di abitanti.

Uno dei principali punti di forza del LUR è la possibilità di ottenere una stima senza dipendere dalla presenza e dal funzionamento delle stazioni di

monitoraggio. Pertanto, la metodologia LUR permette di ottenere una stima anche in località nelle quali non sono disponibili misure dirette. L'unica eccezione è rappresentata da alcune variabile meteorologiche, che possono risultare occasionalmente assenti per specifici luoghi. Gli altri regressori utilizzati nel modello, essendo prevalentemente variabili geografiche statiche, possono invece essere associati alle località indipendentemente dalla presenza di stazioni di monitoraggio e risultano pertanto generalmente disponibili sull'intera area di studio.

Questa caratteristica rende tale metodo complementare all'IDW; l'assenza di una stima ottenuta mediante una delle due metodologie non implica necessariamente l'impossibilità di ottenere una stima mediante l'altra tipologia. L'utilizzo congiunto permette dunque di ridurre la probabilità di ottenere valori mancanti della concentrazione di un inquinante in un determinato luogo.

In merito alla costruzione del modello LUR adottato nel presente lavoro sono stati presi come riferimento diversi studi presenti in letteratura, sia per la scelta dei regressori sia per le metodologie utilizzate per individuare quelli più adatti, tra i disponibili, alla stima delle concentrazioni degli inquinanti [37], [58], [59]. Osservando inoltre il numero di stazioni e l'estensione delle aree geografiche coinvolte nei diversi studi, è stato preso particolare riferimento al lavoro di De Hoogh [37].

Nello sviluppo di un modello LUR esistono diverse modalità per individuare il migliore sottoinsieme di regressori in grado di massimizzare la qualità della stima. Tra le principali si possono citare la backward selection, che parte da un modello contenente tutti i regressori disponibili e procede eliminando man mano quelli meno informativi, la forward selection, che parte da un modello privo di regressori e li aggiunge progressivamente, e la stepwise regression, che combina i due approcci valutando sia l'aggiunta sia l'eventuale rimozione dei regressori durante il processo di selezione.

Nel presente lavoro è stata adottata una procedura di stepwise regression, in quanto rappresenta uno degli approcci maggiormente utilizzati per la selezione dei regressori nei modelli LUR e risulta inoltre coerente con l'approccio adottato

nello studio di riferimento di de Hoogh per lo sviluppo di modelli LUR su scala europea [37], [59].

Prima di procedere alla selezione dei regressori, è stato tuttavia necessario definire opportunamente alcuni dei regressori derivanti dal dataset CLC e da OSM.

Come evidenziato negli studi esaminati, per i regressori relativi alla copertura del suolo è necessario definire una area di interesse attorno al punto geografico preso in considerazione [37], [59]. Il valore associato a ciascuna classe di copertura del suolo derivata da CLC dipende infatti dall'area entro la quale viene analizzata la sua presenza.

Nel presente lavoro sono stati utilizzati due differenti raggi: 1 km e 20 km. Il primo permette di descrivere le caratteristiche presenti nell'area immediatamente circostante al punto geografico considerato, mentre il secondo permette di rappresentare un contesto territoriale più ampio, includendo dunque una porzione maggiore di area.

Considerando le 44 classi di copertura del suolo previste da CLC e i due raggi considerati, si sarebbero ottenuti in totale 88 potenziali regressori di questo tipo. Un numero così elevato di variabili avrebbe aumentato la complessità del modello e il rischio di overfitting. Per ridurre il numero di regressori mantenendo comunque una descrizione sufficientemente completa del territorio, è stata quindi sfruttata la struttura gerarchica di CLC, utilizzando le 15 sottoclassi, ottenendo un numero di variabile più contenuto ma ancora rappresentativo delle principali tipologie di coperture del suolo [60].

Per quanto riguarda la presenza e la tipologia delle strade, è stato seguito l'approccio presentato da De Hoogh, dal quale sono state riprese sia le principali categorie stradali considerate rilevanti ai fini della stima dell'inquinamento atmosferico, sia i raggi di ricerca utilizzati per determinare la presenza nell'intorno del punto geografico considerato [37]. In particolare, le main roads comprendono autostrade(motorway), superstrade(trunk), strade principali (primary e secondary) e i relativi svincoli e raccordi (motorway, trunk, primary, secondary link).

A questi regressori sono stati aggiunti l'altitudine e il numero di abitanti, ottenendo complessivamente 46 possibili regressori statici, ovvero variabili il cui valore rimane invariato nel tempo. A questi si sono aggiunti i regressori definiti dinamici, ovvero le variabili meteorologiche, il cui valore varia giorno per giorno. Tutti i possibili regressori sono presenti nella Tabella 3.7.

Tabella 3.7: I regressori utilizzati per il modello LUR. Per ciascun regressore sono riportati il codice identificativo assegnato all'interno del modello e il corrispondente codice della fonte originale

CODICE ASSEGNATO	CODICE DELLA FONTE	DESCRIZIONE
land_1	alrd50	Lunghezza totale in km della rete stradale entro 50 m
land_2	alrd100	Lunghezza totale in km della rete stradale entro 100 m
land_3	alrd300	Lunghezza totale in km della rete stradale entro 300 m
land_4	alrd1000	Lunghezza totale in km della rete stradale entro 1 km
land_5	mjrd100	Lunghezza totale in km delle sole strade principali entro 100 m
land_6	mjrd200	Lunghezza totale in km delle sole strade principali entro 200 m
land_7	urb_fabr1000	Percentuale di superficie urbana e tessuto impermeabilizzato entro 1 km
land_8	ind_comm_transp1000	Percentuale di aree industriali, commerciali e nodi di trasporto entro 1 km
land_9	mine_dump_constr1000	Percentuale di cantieri, miniere e siti di smaltimento rifiuti entro 1 km
land_10	green_area1000	Percentuale di aree verdi urbane e ricreative entro 1 km
land_11	arable_land1000	Percentuale di seminativi e terreni a rotazione culturale entro 1 km

CODICE ASSEGNATO	CODICE DELLA FONTE	DESCRIZIONE
land_12	perma_crops1000	Percentuale di colture permanenti (vigneti, frutteti, oliveti) entro 1 km
land_13	past1000	Percentuale di pascoli e praterie permanenti entro 1 km
land_14	agri_area1000	Percentuale di zone agricole eterogenee e complesse entro 1 km
land_15	forest1000	Percentuale di superfici boscate e foreste entro 1 km
land_16	scrubs_vegetation1000	Percentuale di vegetazione arbustiva e di transizione entro 1 km
land_17	open_space1000	Percentuale di spazi aperti con vegetazione scarsa o assente entro 1 km
land_18	in_wet_zone1000	Percentuale di zone umide interne (paludi, torbiere) entro 1 km
land_19	maritime_wetland1000	Percentuale di zone umide marittime (saline, lagune) entro 1 km
land_20	in_waters1000	Percentuale di acque interne (fiumi, laghi, corsi d'acqua) entro 1 km
land_21	maritime_waters1000	Percentuale di mari, oceani e specchi d'acqua costieri entro 1 km
land_22 – land_36	(Corine 20 km)	Identici ai codici land_7–land_21, ma calcolati su un raggio esteso di 20 km
weather_1	pp	Pressione media giornaliera al livello del mare in hPa.
weather_2	tn	Temperatura minima giornaliera in gradi Celsius
weather_3	tx	Temperatura massima giornaliera in gradi Celsius.

CODICE ASSEGNATO	CODICE DELLA FONTE	DESCRIZIONE
weather_4	tg	Temperatura media giornaliera in gradi Celsius.
weather_5	fg	Velocità media giornaliera del vento in m/s.
weather_6	hu	Umidità relativa media giornaliera, espressa in percentuale.
weather_7	qq	Radiazione globale media giornaliera in W/m ²
weather_8	rr	Precipitazione totale giornaliera in mm
reg_altitudine	altitudine	Altitudine sul livello del mare (m s.l.m.)
reg_totale	Popolazione	Numero dei residenti nel comune

Per individuare, per ciascun inquinante, il sottoinsieme ottimale di regressori per il modello LUR, è stata effettuata una procedura di selezione utilizzando lo stesso dataset descritto nel paragrafo 3.4.1.

Da un modello inizialmente privo di variabili, è stata valutata, a ogni iterazione, l'aggiunta di un nuovo regressore. La scelta è stata effettuata considerando diversi indicatori statici, tra cui il Variance Inflation Factor (VIF), il p-value e il coefficiente di determinazione R^2 .

Il VIF permette di valutare il livello di multicollinearità tra i regressori. Nel presente lavoro è stato imposto un valore massimo di 4, così da evitare l'inserimento di regressori eccessivamente correlati tra di loro.

Il p-value è stato utilizzato per valutare la significatività statistica del regressore inserito, considerando come soglia significativa un valore pari a 0.05.

Il valore di R^2 è stato usato per valutare il contributo della nuova variabile alla capacità esplicativa del modello. In particolare, un nuovo regressore è stato incluso solamente qualora il suo inserimento determinasse un incremento dell' R^2 di almeno 0,005 rispetto al set di regressori precedenti.

Per alcuni regressori sono state inoltre applicate delle restrizioni sul segno dei coefficienti. Tale scelta deriva dal fatto che, sulla base delle conoscenze disponibili e degli studi presenti in letteratura, alcune variabili sono associate ad un aumento della concentrazione delle sostanze inquinanti, mentre altre sono associati ad un suo calo. Per i primi è stato quindi imposto un valore del coefficiente positivo, mentre per i secondi negativo. I regressori per i quali sono stati inseriti tali vincoli sono stati presi sempre dalla revisione di De Hoogh e sono riportati nella Tabella 3.8.

Tabella 3.8: Le restrizioni applicate sul segno dei coefficienti: valore positivo (>0) o negativo (<0)

CODICE	DESCRIZIONE	VALORE COEFFICIENTE
alrd50	Lunghezza totale in km della rete stradale entro 50 m	Positivo.
alrd100	Lunghezza totale in km della rete stradale entro 100 m	positivo
alrd300	Lunghezza totale in km della rete stradale entro 300 m	positivo
alrd1000	Lunghezza totale in km della rete stradale entro 1 km	positivo
mjrd100	Lunghezza totale in km delle sole strade principali entro 100 m	positivo
mjrd200	Lunghezza totale in km delle sole strade principali entro 200 m	positivo
urb_fabr1000	Percentuale di superficie urbana e tessuto impermeabilizzato entro 1 km.	Positivo
ind_comm_transp1000	Percentuale di aree industriali, commerciali e nodi di trasporto entro 1 km	positivo
urb_fabr20000	Percentuale di superficie urbana e tessuto impermeabilizzato entro 20 km	positivo
green_area1000	Percentuale di aree verdi urbane e ricreative entro 1 km	negativo

CODICE	DESCRIZIONE	VALORE COEFFICIENTE
ind_comm_transp20000	Percentuale di aree industriali, commerciali e nodi di trasporto entro 20 km	Positivo
green_area20000	Percentuale di aree verdi urbane e ricreative entro 20 km	Negativo
Reg_altitudine	altitudine	Positivo

Una volta definiti il set di regressori e i relativi vincoli, è stato quindi possibile procedere alla stima dei coefficienti del modello. I coefficienti dei regressori selezionati per ciascun inquinante sono stati calcolati esclusivamente nel contesto italiano, tramite la funzione `lscipy.optimize.lsqr_linear`. Tale funzione permette di individuare i coefficienti che minimizzano il funzionale dei minimi quadrati rispettando contemporaneamente i vincoli precedentemente definiti. Nell'equazione (3.4) è riportato il funzionale da minimizzare.

$$\min_{\beta} \frac{1}{2} \|X\beta - y\|_2^2 \quad (3.4)$$

dove X rappresenta la matrice dei regressori, β il vettore dei coefficienti usati nel modello, y il vettore dei valori osservati e $\|X\beta - y\|_2^2$ la norma euclidea al quadrato.

Trovati i parametri, è stato quindi possibile applicare il modello LUR mediante la classica formula della regressione lineare (3.5)

$$y = a + \sum_{h=1}^k b_h x_h \quad (3.5)$$

dove a rappresenta l'intercetta del modello, b_h il coefficiente associato al regressore h -esimo, x_h il valore assunto dal regressore h -esimo e k il numero complessivo di regressori inclusi nel modello.

Nella Tabella 3.9 sono riportati i regressori ottimali e i relativi coefficienti e ottenuti per ogni inquinante.

Tabella 3.9: Regressori selezionati, intercetta e relativi coefficienti per ciascun inquinante.

INQUINANTE	INTERCETTA (a)	REGRESSORI SELEZIONATI	COEFFICIENTI (b _h)
SO ₂	2.948	land_1, land_2, land_3, land_5, land_6, land_7, land_22	3.192, -6.514, - 212.706, 1.503, - 165.130, 16.201, - 19.039
PM ₁₀	36.299	land_5, land_8, land_23, land_10, land_25, weather_1, weather_2	139.399, -2.307, 11.827, -0.724, - 0.016, -0.299, - 0.034
O ₃	16.130	land_5, land_8, land_25, weather_1, weather_2, weather_3, weather_4, weather_5, weather_6, weather_7, weather_8	1.032, 6.194, - 4.855e-27, 0.314, 0.140, 22.803, 39.518, 18.456, 47.129, 23.398, - 0.194
NO ₂	18.986	land_8, land_23, land_10, weather_3, weather_5, weather_8, reg_altitudine	-2.450, 6.520, - 0.800, -10.958, - 20.044, 0.148, - 12.590
CO	0.725	land_2, land_5, land_8, land_23, land_10, weather_2	-0.816, 3.202, - 0.028, -0.140, - 0.017, -0.001
C ₆ H ₆	1.857	land_5, land_8, land_10, land_25, weather_1, weather_2	5.106, -0.136, - 0.055, -0.0004, - 0.011, -0.002
PM _{2.5}	32.341	land_5, land_8, land_23, land_10, land_25, weather_1, weather_2, reg_totale	100.006, -1.843, 5.326, -0.918, - 0.014, -0.237, - 0.045, -12.590

I risultati del modello mostrano una generale coerenza con i meccanismi fisici che determinano la presenza degli inquinanti atmosferici. In particolare, la presenza di coefficienti fortemente positivi associati al land_5 (major roads nel raggio di 100 m) per il PM_{2.5} e il PM₁₀ evidenzia il contributo delle caratteristiche antropiche alla presenza del PM_{2.5} e del PM₁₀, mentre la marcata dipendenza dell'O₃ dalle variabili meteorologiche è coerente con la natura dei processi che ne regolano la sua formazione.

3.5 I metodi ibridi

Il metodo LUR presenta anche alcuni limiti. In particolare, la relazione tra concentrazione degli inquinanti e regressori ambientali non è necessariamente sufficiente a descrivere tutte le condizioni locali che possono determinare variazioni delle concentrazioni.

I regressori utilizzati nel presente studio, in particolare quelli meteorologici, costituiscono infatti una rappresentazione necessariamente semplificata delle condizioni atmosferiche reali. Variabili quali temperatura, precipitazione, umidità, vento e radiazione solare sono certamente rilevanti, ma non sono da sole sufficienti a descrivere completamente l'atmosfera. La concentrazione di un inquinante può infatti dipendere dall'interazione simultanea di numerosi fenomeni, tra cui condizioni meteorologiche locali e presenza di sorgenti emissive. La costruzione di un modello in grado di rappresentare tutte queste interazioni risulta pertanto estremamente complessa, soprattutto considerando il numero limitato di regressori disponibili e la loro risoluzione spaziale e temporale.

Per cercare di ridurre l'errore commesso, è stata quindi sviluppata una metodologia ibrida, basata sull'utilizzo congiunto del LUR e dell'interpolazione dei residui mediante IDW.

Il residuo viene calcolato come differenza tra il valore misurato dalla stazione di monitoraggio e quello stimato dal modello LUR. I residui vengono quindi interpolati spazialmente mediante IDW e utilizzati quindi per correggere la stima finale del LUR. In questo modo, se in una determinata zona il modello tende sistematicamente a sottostimare o sovrastimare le concentrazioni misurate,

l'interpolazione dei residui permette di tenere conto di tale andamento locale e di correggere dunque la stima ottenuta.

Nel presente lavoro sono state sviluppate due varianti del metodo:

- LUR_ibrido, che applica IDW sui residui secondo la configurazione classica;
- LUR_vicino, che qualora all'interno del valore di R individuato per lo specifico inquinante non siano presenti stazioni, utilizza, se disponibile entro una distanza massima di 200 km, la stazione più vicina.

In entrambi i casi, qualora per l'area geografica considerata e per lo specifico istante non siano disponibili stazioni di monitoraggio utilizzabili, la stima coincide con quella ottenuta mediante il LUR classico, senza quindi nessuna correzione. Sono pertanto disponibili tre metodi di stima che verranno confrontati nel capitolo 3.6.

3.6 Valutazione comparativa dei modelli

Una volta sviluppati i modelli LUR e IDW e le relative varianti, è stato necessario condurre una valutazione comparativa delle diverse metodologie sviluppate, confrontandole anche con alcuni approcci di base comunemente utilizzati per problemi analoghi.

Il confronto è stato condotto utilizzando i dati validi e verificati disponibili per ogni inquinante, concentrandosi esclusivamente nel contesto italiano. Considerata l'elevata numerosità del database AirQuality, il confronto è stato effettuato su un sottoinsieme di dati, selezionando tutte le misure disponibili relative all'anno 2024, che risulta essere uno degli anni con il maggior numero di misure.

Sono state considerate le 842 stazioni di monitoraggio italiane presenti nel database e, per ciascuna di esse, è stata valutata la capacità dei diversi modelli di stimare la concentrazione dell'inquinante effettivamente misurato dalla stazione nel medesimo istante. Per valutare la bontà del comportamento dei modelli in differenti condizioni geografiche, le stazioni sono state suddivise in categoria sulla base della variabile `area_type` presente nella tabella `stations`.

La categoria Urban identifica le aree urban caratterizzate da una edificazione continua, mentre la categoria Suburban comprende le aree urban ampiamente edificate, ma caratterizzate da una densità edilizia inferiore rispetto alle aree urban [61], [62].

La categoria rural, invece, comprende tutte le aree che non rientrano nelle precedenti classificazioni[63]. Quest'ultima è ulteriormente suddivisa, in funzione della distanza dalla zona urban o suburban più vicina, in tre sottocategorie: nearcity, per le aree rural situate a meno di 10 km da una zona urban o suburban; regional, per le aree comprese tra 10 e 50 km; e remote, per quelle situate a più di 50 km [64], [65], [66].

Considerando il numero ridotto di stazioni appartenenti alle sottoclassi di rural, queste ultime sono state incorporate nella categoria rural. In questo modo è stato quindi possibile distinguere tre principali contesti: urban, suburban e rural.

È stato inoltre tenuto in considerazione la disponibilità di stazioni di monitoraggio attive nell'intorno del punto di interesse. Per ciascun inquinante, il numero di stazioni presenti è stato valutato utilizzando il valore di R individuato durante la fase di ottimizzazione del metodo IDW. Sono state quindi definite due condizioni: una zona ricca di stazioni (zona_ricca), nella quale sono presenti almeno due stazioni attive all'interno di R, e una zona carente di stazioni (zona_povera), nella quale è presente al massimo una stazione.

Il confronto è stato quindi articolato secondo due differenti scenari.

Il primo valuta l'utilizzo dei metodi per il data imputation, ossia la ricostruzione di una misura mancante in una determinata stazione di monitoraggio.

Il secondo invece riguarda la stima della concentrazione dell'inquinante in un punto geografico privo di una stazione di monitoraggio, che rappresenta lo scenario applicativo principale per il quale sono stati sviluppati tali modelli.

Per valutare le performance dei diversi metodi è stato utilizzato l'RMSE, una delle metriche più comunemente utilizzate per valutare l'errore commesso dalla stima rispetto al valore osservato. Nel presente lavoro, per ciascuna combinazione di inquinante, metodo e contesto geografico, il valore di RMSE è stato ottenuto tramite l'Equazione (3.3), confrontando le concentrazioni stimate dal modello con quelle effettivamente misurate dalle stazioni di monitoraggio.

Un valore di RMSE più basso indica quindi una maggiore accuratezza del modello, poiché corrisponde a una minore differenza media tra i valori stimati e i valori osservati. Pertanto, per ciascun inquinante e per ciascuna condizione analizzata, è stato considerato come metodo migliore quello caratterizzato dal valore di RMSE più basso.

3.6.1 Confronto nel Data Imputation

In questo primo scenario è stata valutata la capacità dei diversi metodi di ricostruire una misura mancante presso una stazione di monitoraggio.

Sono stati confrontati i seguenti modelli:

- LUR_puro, modello LUR classico senza correzione mediante residui;
- LUR_ibrido, combina la stima LUR con l'interpolazione IDW dei residui;
- LUR_vicino, analogo al precedente, ma che, qualora non siano presenti stazioni all'interno di R, utilizza la stazione più vicina disponibile entro una distanza massima di 200 km;
- IDW, interpolazione classica che sfrutta la distanza;
- IDW_vicino, che utilizza la stazione più vicina entro 200 km qualora non siano presenti stazioni all'interno di R.

La disponibilità dei diversi metodi dipende tuttavia dalla densità delle stazioni presenti nell'area considerata. Infatti, nelle zone carenti di stazioni, l'IDW non è applicabile, mentre l'IDW_vicino consente di ottenere comunque una stima qualora sia disponibile almeno una stazione entro la distanza massima di 200 km.

Analogamente, quando non sono presenti stazioni utilizzabili per la correzione dei residui entro R, il modello LUR_ibrido coincide con il LUR_puro.

Tali modelli sono stati confrontati con alcune metodologie di base comunemente utilizzate per il data imputation [67].

Il primo metodo è l'interpolazione lineare (linear), adatta alla ricostruzione di brevi interruzioni, nella quale il valore mancante viene ottenuto interpolando linearmente tra il valore immediatamente precedente e quello successivo.

Il secondo è lo spline, che, come l'interpolazione lineare, utilizza i valori circostanti per ricostruire il dato, ma sfruttando una funzione curva anziché una relazione lineare.

Il terzo metodo è il Last Observation Carried Forward (LOCF), che assegna al dato mancante l'ultimo valore disponibile. Nel presente lavoro, visto l'impossibilità del ripetersi lo stesso valore per giorni consecutivi, è stato imposto un limite massimo di tre giorni consecutivi.

Infine, è stato utilizzato l'Exponentially Weighted Moving Average (EWMA), che calcola una media pesata dei valori circostanti attribuendo un peso maggiore alle osservazioni temporalmente più vicine.

Questi metodi sono stati confrontati utilizzando le misurazioni relative al 2024 e considerando esclusivamente le stazioni italiane. Nella Tabella 3.10 sono riportati i risultati ottenuti per ciascun inquinante in ciascun contesto.

Tabella 3.10: Metodi di Data Imputation ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e complessivo) in aree ad alta densità di stazioni.

INQUINANTE	GLOBALE (RMSE)	URBAN (RMSE)	SUBURBAN (RMSE)	RURAL (RMSE)
NO ₂	Linear / Spline (4.6045)	Linear / Spline (5.2098)	Linear / Spline (4.2613)	Linear / Spline (2.3896)
C ₆ H ₆	Linear / Spline (0.4019)	Linear / Spline (0.4077)	Linear / Spline (0.4299)	Linear / Spline (0.1241)
CO	Linear / Spline (0.0838)	Linear / Spline (0.0924)	Linear / Spline (0.0675)	Linear / Spline (0.0540)
O ₃	Linear / Spline (8.5467)	Linear / Spline (8.7598)	Linear / Spline (8.4399)	Linear / Spline (8.2833)
SO ₂	Linear / Spline (0.9978)	Linear / Spline (0.7926)	Linear / Spline (1.2252)	Linear / Spline (0.9725)
PM ₁₀	IDW (7.3559)	IDW (6.9214)	IDW (7.8040)	IDW (8.3573)

INQUINANTE	GLOBALE (RMSE)	URBAN (RMSE)	SUBURBAN (RMSE)	RURAL (RMSE)
PM _{2.5}	IDW (4.7114)	IDW (4.6620)	IDW (4.9670)	IDW (4.0342)

In generale, nello scenario di data imputation, i metodi Linear e Spline risultano essere quelli aventi le migliori prestazioni per praticamente tutti gli inquinanti e in tutti i contesti considerati. Questo risultato è particolarmente evidente, ad esempio, per il CO, per il quale i modelli sviluppati mostrano valori di RMSE quasi doppi rispetto ai metodi di interpolazione temporale (Figura 3.9).

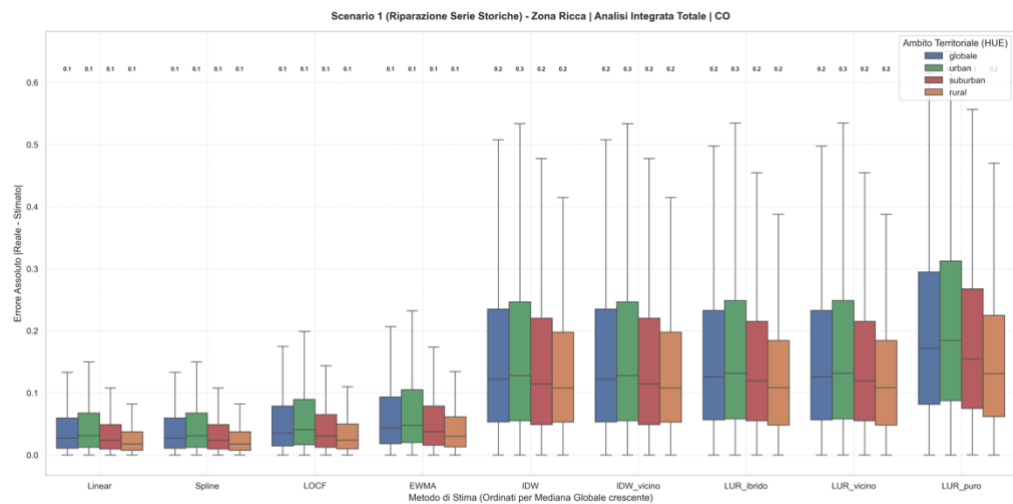


Figura 3.9: Boxplot degli errori assoluti commessi da ciascun metodo di Data Imputation per il CO nelle zone ricche di stazioni, suddivisi per contesto: tutte le stazioni (blu), urban (verde), suburban (rosso) e rural (arancione).

Tale comportamento è riconducibile al fatto che, quando si deve ricostruire una breve interruzione temporale di una stazione, le misure immediatamente precedenti e successive al dato mancante contengono informazioni nettamente più correlate al valore da ricostruire rispetto alle misure contemporaneamente rilevate da stazioni geograficamente distanti. Allo stesso modo, i regressori meteorologici e territoriali, pur fornendo informazioni utili sulle condizioni ambientali del luogo, non riescono a descrivere con sufficiente precisione le variazioni della concentrazione dell'inquinante.

Unica eccezione è rappresentata dal PM_{2.5} e dal PM₁₀ (Figura 3.10), per i quali il metodo IDW ottiene risultati competitivi, soprattutto nelle zone caratterizzate da un'elevata presenza di stazioni attive.

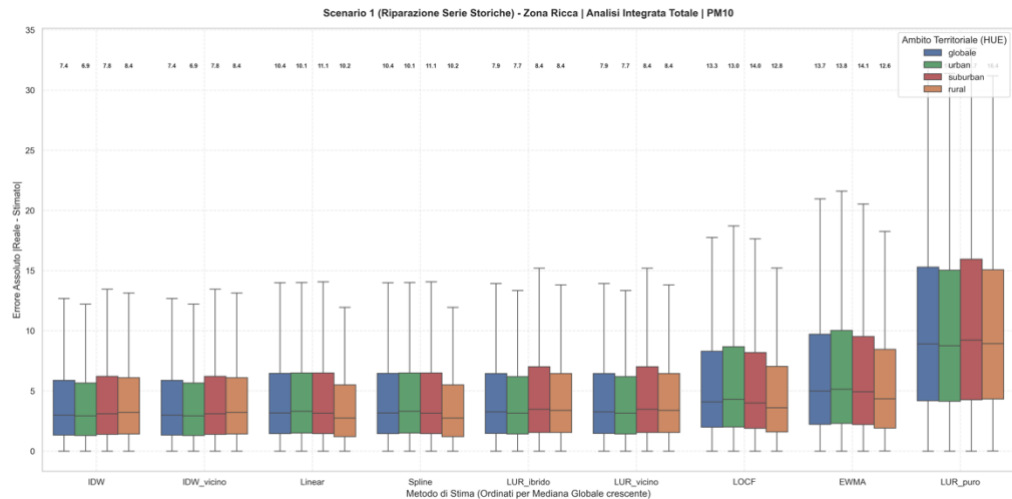


Figura 3.10: Boxplot degli errori assoluti commessi da ciascun metodo di Data Imputation per il PM₁₀ nelle zone ricche di stazioni, suddivisi per contesto: tutte le stazioni (blu), urban (verde), suburban (rosso) e rural (arancione).

Questo comportamento può essere ricondotto alla maggiore omogeneità spaziale di queste due sostanze rispetto agli altri inquinanti maggiormente influenzati da sorgenti locali. Pertanto, quando sono disponibili diverse stazioni nell'intorno del punto preso in esame, l>IDW può sfruttare efficacemente l'informazione spaziale disponibile.

Analizzando le tre categorie territoriali, si osservano inoltre differenze nelle prestazioni in funzione del contesto. In generale, le stazioni collocate in ambiente rural presentano valori di RMSE inferiori rispetto a quelle urban, con differenze particolarmente evidenti per inquinanti fortemente associati al traffico e a attività antropiche, come il NO₂, il CO, il PM₁₀ e il PM_{2.5}.

Nel contesto suburban si osservano valori intermedi, mentre le stazioni urban presentano valori maggiori. Tale comportamento può essere ricondotto sia alla maggiore complessità del contesto urban, sia alla maggiore presenza di picchi di concentrazione difficilmente stimabili.

Il risultato può inoltre essere influenzato dalla diversa distribuzione delle stazioni nelle tre aree. L'O₃ presenta invece un comportamento più uniforme tra i tre contesti territoriali, coerentemente con la sua maggiore componente di

fondo e con i processi atmosferici che ne regolano la sua formazione e il suo trasporto.

In generale, i modelli sviluppati nel presente lavoro mostrano prestazioni peggiori rispetto ai metodi di interpolazione temporale per lo specifico problema del data imputation. Tra questi, il LUR_puro risulta generalmente il meno performante. Tale risultato non rappresenta una bocciatura di tale modello, poiché quest'ultimo è stato sviluppato principalmente per la stima spaziale della concentrazione in punti privi di stazioni di monitoraggio, e non dunque per la ricostruzione di brevi lacune temporali.

Anche LOCF e EWMA presentano generalmente prestazioni leggermente inferiori rispetto al Linear e Spline, in quanto non sfruttano in maniera analoga l'informazione successiva al giorno con la misurazione mancante.

Nella zona povera di stazioni si osserva un ulteriore peggioramento dei metodi che dipendono dalla presenza di stazioni nel vicinato.

Anche in questo caso, Linear e Spline risultano generalmente i migliori in tutti i contesti. Per il PM_{2.5} e il PM₁₀, tuttavia, l>IDW mantiene prestazioni competitive, confermando il ruolo della componente spaziale per queste tipologie di inquinanti. I risultati ottenuti nella zona povera sono presenti nella Tabella 3.11.

Tabella 3.11: Metodi di Data Imputation ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e complessivo) in aree a bassa densità di stazioni.

INQUINANTE	GLOBALE (RMSE)	URBAN (RMSE)	SUBURBAN (RMSE)	RURAL (RMSE)
NO ₂	Linear / Spline (1.0990)	N/D	N/D	Linear / Spline (1.0990)
C ₆ H ₆	Linear / Spline (0.3253)	Linear / Spline (0.3593)	Linear / Spline (0.1644)	Linear / Spline (0.2305)
CO	Linear / Spline (0.1020)	Linear / Spline (0.0636)	Linear / Spline (0.1045)	Linear / Spline (0.1323)
O ₃	Linear / Spline (8.4393)	Linear / Spline (8.5549)	Linear / Spline (8.7156)	Linear / Spline (8.0305)

INQUINANTE	GLOBALE (RMSE)	URBAN (RMSE)	SUBURBAN (RMSE)	RURAL (RMSE)
SO ₂	Linear / Spline (0.3543)	Linear / Spline (0.3543)	N/D	N/D
PM ₁₀	Linear / Spline (9.3619)	IDW_vicino (9.2245)	IDW_vicino (8.5794)	Linear / Spline (8.8943)
PM _{2.5}	Linear / Spline (5.0615)	Linear / Spline (5.2922)	Linear / Spline (4.6929)	Linear / Spline (4.7308)

3.6.2 Confronto nella stima in un luogo privo di stazione

In questo secondo scenario è stata valutata la capacità dei metodi di stimare la concentrazione di un inquinante in un punto geografico nel quale non è disponibile una stazione di monitoraggio. Per simulare tale condizione, il valore misurato dalla stazione è stato temporaneamente escluso dal processo di stima e successivamente ricostruito utilizzando le informazioni disponibili presso le stazioni limitrofe e, nel caso dei modelli LUR, sfruttando i regressori ambientali e meteorologici.

Anche in questo caso le prestazioni sono state valutate distinguendo tra le categorie territoriali e tra zone ricche e carenti di stazioni.

I modelli sviluppati sono stati confronti con altre metodologie di riferimento basate direttamente sulle stazioni limitrofe. In particolare, sono state considerate la media e la mediana dei valori misurati nello stesso istante dalle 3 o 5 stazioni più vicine (Media_3_Vicini, Media_5_Vicini, Mediana_3_Vicini, Mediana_5_Vicini) e il metodo Stazione_Più_Vicina, che utilizza direttamente la misura della stazione spazialmente più vicina. Nel caso in cui il numero di stazioni presenti entro il valore di R, specifico per ogni inquinante, fosse inferiore a quello richiesto dal metodo, la stima è stata effettuata utilizzando tutte le stazioni disponibili.

Se invece non sono presenti stazioni di monitoraggio nell'area considerata, i metodi basati sulla media e sulla mediana restituiscono valori nulli per tutti gli inquinanti. Non si è scelto di estendere R poiché ciò avrebbe comportato

l'utilizzo di valori misurati in contesti geografici molto differenti rispetto al luogo oggetto di stima.

Nella Tabella 3.12 vengono riportati i modelli vincitori per ogni inquinante nella zona ricca di stazioni.

Tabella 3.12: Metodi di stima ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e globale) in aree ricche di stazioni

INQUINANTE	GLOBALE (RMSE)	URBAN (RMSE)	SUBURBAN (RMSE)	RURAL (RMSE)
NO ₂	LUR_ibrido (7.8117)	LUR_ibrido (8.5141)	LUR_ibrido (7.4435)	LUR_ibrido (5.5389)
C ₆ H ₆	LUR_ibrido (0.6993)	Media_5_Vicini (0.7372)	LUR_ibrido (0.6050)	LUR_ibrido (0.3562)
CO	Media_5_Vicini (0.2333)	Media_5_Vicini (0.2493)	Media_5_Vicini (0.2075)	LUR_ibrido (0.1666)
PM ₁₀	Media_5_Vicini (7.2150)	Media_5_Vicini (6.7233)	Media_5_Vicini (7.6000)	Mediana_3_Vicini (8.1970)
PM _{2.5}	IDW (4.7114)	IDW (4.6620)	Media_3_Vicini (4.6960)	IDW (4.0342)
O ₃	IDW (12.0060)	Mediana_5_Vicini (11.1929)	IDW (10.5946)	IDW (14.3778)
SO ₂	Media_5_Vicini (2.3228)	IDW (1.9790)	Media_5_Vicini (2.7661)	Media_5_Vicini (2.1280)

A differenza dello scenario precedente, nel quale Linear e Spline risultavano generalmente dominanti, in questo caso i risultati sono maggiormente diversificati e dipendono fortemente dalle caratteristiche dell'inquinante, nonché anche dalla disponibilità di stazioni nell'intorno.

È possibile individuare, in generale, tre differenti comportamenti dei vari inquinanti. Una prima categoria comprende gli inquinanti avente una variabilità prevalentemente locale, come il C₆H₆ e il NO₂, per i quali la presenza di strutture

e sorgenti emissive localizzate può determinare differenze significative anche su distanze ridotte. Una seconda categoria comprende inquinanti aventi una maggiore omogeneità spaziale, come la SO_2 , l' O_3 , il $\text{PM}_{2.5}$ e il PM_{10} . Il CO presenta invece un comportamento intermedio tra le due classi.

Per gli inquinanti maggiormente dipendenti dalle condizioni locali, i modelli LUR_vicino e LUR_ibrido risultano particolarmente competitivi, poiché possono utilizzare regressori territoriali in grado di descrivere la presenza di sorgenti emissive, come strade e presenza di determinate infrastrutture.

Nel caso del NO_2 , il modello LUR_ibrido risulta performante in tutti i contesti considerati, mostrando valori di RMSE inferiori rispetto agli altri metodi. Ad esempio, l'IDW presenta un RMSE pari a $9.4072 \mu\text{g}/\text{m}^3$, risultando dunque meno efficace rispetto al modello LUR_ibrido.

Per il C_6H_6 , nel contesto urban caratterizzato da una maggiore presenza di stazioni, la media delle cinque stazioni più vicine risulta il metodo migliore, mentre i due metodi LUR presentano prestazioni solo leggermente inferiori. Questo evidenzia come, quando la rete di monitoraggio è sufficientemente densa, l'utilizzo delle misure disponibili possa fornire una stima molto efficace anche per inquinanti caratterizzati da una marcata variabilità locale.

Per gli inquinanti caratterizzati da una maggiore omogeneità, risultano invece efficaci i metodi basati direttamente sulle misure delle stazioni, come media, mediana e i metodi IDW. In questi casi, infatti, una stazione distante può comunque fornire informazioni utili per stimare la concentrazione nel punto geografico in studio.

Per il CO si osserva un comportamento intermedio: nelle zone urban e suburban, caratterizzate dalla presenza di più stazioni nel raggio di pochi chilometri, i metodi basati direttamente sulle misure delle stazioni risultano maggiormente competitivi. Nei contesti rural, invece, dove le stazioni sono presenti entro il valore di R ma risultano meno fitte, acquistano maggiore importanza i metodi LUR.

A titolo di esempio, la distribuzione degli errori ottenuti dai diversi metodi per il CO nella zona ricca di stazioni è riportata nella Figura 3.11. Il boxplot, ordinato dal metodo con l'RMSE più basso a quello con l'RMSE più alto, calcolato

considerando complessivamente le tre aree geografiche, evidenzia le differenze di prestazione tra i diversi approcci. Come si può osservare, i modelli IDW e IDW_vicino mostrano prestazioni particolarmente buone, confermando l'importanza della disponibilità di stazioni di monitoraggio in contesti caratterizzati da una maggiore densità della rete.

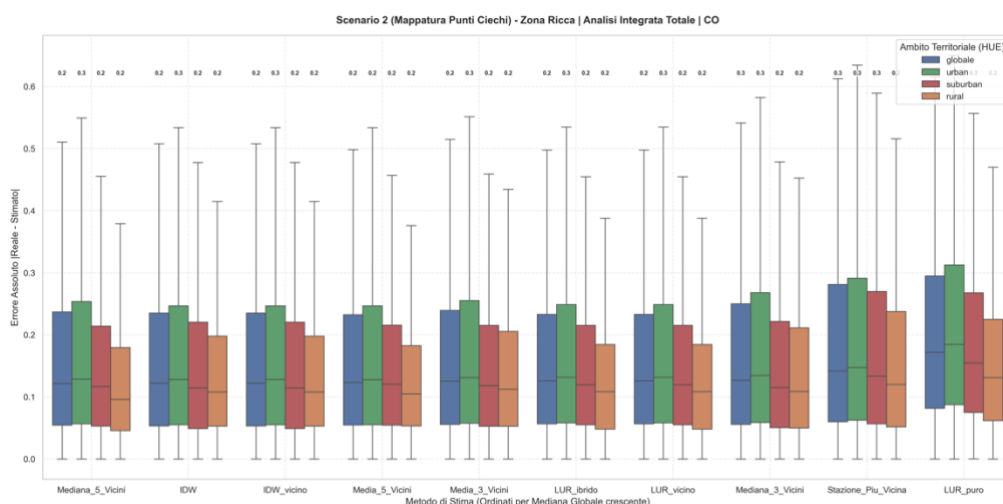


Figura 3.11: Boxplot degli errori assoluti commessi da ciascun metodo di stima per il CO nelle zone ricche di stazioni, suddivisi per contesto: tutte le stazioni (blu), urban (verde), suburban (rosso) e rural (arancione).

Nel complesso, nelle zone ricche di stazioni i risultati mostrano che i modelli sviluppati nel presente lavoro sono competitivi rispetto ai metodi di riferimento. In particolare, le varianti LUR che integrano la correzione mediante i residui risultano per diversi inquinanti i più performanti. Il LUR_puro invece presenta generalmente prestazioni inferiori, così come il metodo Stazione_Piu_Vicina. Tale risultato è coerente con la natura stessa di quest'ultimo modello, il quale non considera né la distanza dalle stazioni né le caratteristiche ambientali del luogo da stimare, ma si limita ad assumere come stima il valore misurato dalla stazione più vicina. Per alcuni inquinanti, tuttavia, la presenza di una rete di monitoraggio particolarmente densa consente al metodo di ottenere comunque risultati soddisfacenti.

Per quanto riguarda la zona carente di stazione, i risultati ottenuti sono i seguenti (Tabella 3.13):

Tabella 3.13: Metodi di stima ottimali per ciascun inquinante e contesto geografico (urban, suburban, rural e globale) in aree carenti di stazioni.

INQUINANTE	GLOBALE (RMSE)	URBAN (RMSE)	SUBURBAN (RMSE)	RURAL (RMSE)
NO ₂	LUR_vicino (4.0791)	N/D	N/D	LUR_vicino (4.0791)
C ₆ H ₆	LUR_ibrido (0.7787)	Media_3_Vicini (0.7857)	LUR_ibrido (0.5429)	Media_5_Vicini (0.4687)
CO	LUR_vicino (0.2915)	Media_3_Vicini (0.2080)	LUR_vicino (0.2662)	LUR_vicino (0.3402)
PM ₁₀	Media_5_Vicini (8.6565)	Media_5_Vicini (7.8208)	Media_3_Vicini (7.4533)	Media_5_Vicini (11.6019)
PM _{2.5}	Media_5_Vicini (6.0806)	Media_5_Vicini (5.8321)	Media_3_Vicini (5.2676)	LUR_vicino (7.4937)
O ₃	Media_5_Vicini (15.7752)	Media_5_Vicini (12.7962)	Media_5_Vicini (14.1839)	LUR_vicino (18.4431)
SO ₂	Mediana_5_Vicini (1.2994)	Mediana_5_Vicini (1.2994)	N/D	N/D

Come atteso, si osserva un peggioramento delle prestazioni dei metodi che dipendono direttamente dalla presenza di stazioni di monitoraggio. In numerosi casi non sono infatti disponibili stazioni per applicare alcuni dei metodi sviluppati, motivo per cui nella tabella compare N/D. inoltre, quando il numero di stazioni è comunque molto ridotto, i valori di RMSE devono essere valutati con cautela, poiché possono essere influenzati dall'utilizzo di un numero limitato di osservazioni.

In questo contesto, i modelli LUR assumono maggiore rilevanza, in quanto la loro applicazione non richiede una elevata densità di stazioni di monitoraggio. Per gli inquinanti caratterizzati da una forte dipendenza locale, le varianti LUR_ibrido e LUR_vicino risultano molto competitive, poiché possono

sfruttare le informazioni presenti nei regressori ambientali. Un comportamento analogo può essere osservato anche per il CO.

Per gli inquinanti dotati di una maggiore omogeneità spaziale, media e mediana continuano a fornire stime affidabili, poiché anche stazioni relativamente distanti possono continuare a fornire informazioni utili per la stima della concentrazione di questi inquinanti. Anche in questo caso il LUR_ibrido risulta tuttavia competitivo, mentre l'IDW mostra un peggioramento più marcato rispetto alla situazione caratterizzata da una maggiore presenza di stazioni.

Il peggioramento dell'IDW tra zona ricca e zona povera è particolarmente evidente nel presente lavoro anche in relazione alla configurazione adottata. Infatti, come descritto nell'opportuno capitolo, quando non sono disponibili stazioni all'interno di R, il metodo si affida totalmente alla stazione più vicina, la cui distanza può essere considerevole.

Al contrario, i modelli LUR_vicino e LUR_ibrido mantengono prestazioni più stabili nei due contesti, poiché non dipendono fortemente dalla presenza di stazioni nella città considerata. Questo risultato evidenzia uno dei principali vantaggi di tali modelli: quando la rete di monitoraggio è poco densa, l'utilizzo di regressori ambientali e meteorologici permette di ottenere una stima anche in condizioni nelle quali le informazioni spaziali sono estremamente limitate.

3.7 Applicazione dei metodi

Una volta discussi i diversi metodi e individuati quelli con le migliori performance nei vari contesti, è stato necessario definire sia i punti geografici sui quali effettuare le stime sia i modelli da utilizzare nell'applicazione finale.

Come definito nei precedenti capitoli, non è stato possibile effettuare le stime su tutti i punti geografici presenti nella tabella `postcode_places`. Infatti, il reperimento dei regressori meteorologici e ambientali per ciascuno di questi luoghi avrebbe infatti comportato un carico computazionale e un consumo di memoria non indifferente. È stato dunque necessario compiere una selezione dei punti di interesse, concentrandosi esclusivamente sui CAP corrispondenti alla residenza dei pazienti in cura presso l'Istituto Neurologico Nazionale Casimiro

Mondino di Pavia e, per ottenere una rappresentazione geografica più ampia del territorio italiano, sui CAP associati ai capoluoghi di provincia.

Questo ha permesso di ottenere quindi una stima verosimile della qualità dell'aria nel comune di residenza dei pazienti e, tramite i CAP associati ai capoluoghi di provincia, di estendere l'analisi al territorio nazionale, fornendo un quadro più ampio sulla distribuzione della qualità dell'aria nel contesto geografico considerato.

È stato inoltre necessario selezionare i metodi da implementare nella stima finale. I risultati del confronto hanno evidenziato prestazioni generalmente inferiori per alcuni metodi introdotti, in particolare per il LUR_puro e per la Stazione_Più_Vicina, che pertanto sono stati esclusi. Per ragioni computazionali e di memoria, si è scelto di introdurre solamente due metodi, uno per ciascuna delle due principali famiglie considerate: IDW_vicino, per la famiglia dei metodi IDW e LUR_vicino per quella dei metodi LUR.

La scelta di queste due versioni non è stata determinata solamente dalle prestazioni ottenute nei confronti, ma anche dalla loro capacità di ridurre il numero di stime nulle. Infatti, entrambi i metodi prevedono, qualora non siano presenti stazioni di monitoraggio all'interno di R individuato per ciascun inquinante, il ricorso alla stazione più vicina entro una distanza massima di 200 km, quando ovviamente disponibile.

Questa caratteristica, unita al fatto che i due metodi risultano essere indipendenti tra di loro, permette di ridurre il rischio di ottenere una stima nulla per una determinata località e specifica finestra temporale.

È importante sottolineare che la selezione dei punti geografici e dei metodi è stata effettuata esclusivamente in funzione dei limiti computazionali e di memoria del presente lavoro. Le metodologie sviluppate sono infatti potenzialmente applicabili anche al contesto spagnolo e portoghese, a condizione di disporre di un numero adeguato di stazioni di monitoraggio avente misure valide e verificate e dei regressori meteorologici e ambientali necessari per poter applicare la stima tramite modello LUR.

È stato quindi sviluppato uno script in Python che, ricevuto in ingresso le coordinate geografiche dei luoghi selezionati, restituisce per ciascuno di essi la

stima della concentrazione dei diversi inquinanti mediante i metodi IDW_vicino e LUR_vicino, per l'intero range temporale compreso tra il 2009 e il 2026. I risultati ottenuti sono stati successivamente caricati in PostgreSQL tramite PgAdmin4 e organizzati all'interno del database in apposite viste. In particolare, sono state create le seguenti viste, le quali si aggiungono alla view_patients precedentemente descritta:

- view_cap_estimation, contenente le stime delle concentrazioni degli inquinanti per i CAP considerati;
- view_cap_italy: contenente l'elenco di tutti i CAP d'Italia e le relative informazioni geografiche;
- view_capoluoghi_italy, contenente i CAP associati ai capoluoghi di provincia e le informazioni relative;
- view_estimation_code, contenente il codice identificativo del metodo utilizzato, con valore 1 per l>IDW_vicino e valore 2 per il LUR_vicino.

Le viste così ottenute costituiscono quindi la base dati utilizzata nelle successive fasi di visualizzazione dei risultati.

4 PROGETTAZIONE E SVILUPPO DI DASHBOARD TRAMITE GRAFANA PER LA VISUALIZZAZIONE DEI DATI

Per rendere i dati raccolti all'interno di AirQuality immediatamente accessibili, è stato sviluppato uno strumento di visualizzazione che fosse il più possibile intuitivo e di facile fruizione. L'obiettivo principale è stato quello di realizzare un'interfaccia in grado di fornire al personale clinico sia una panoramica sintetica della qualità dell'aria, sia la possibilità di approfondire nel dettaglio specifici inquinanti, intervalli temporali e aree geografiche di interesse.

Le diverse dashboard sono state realizzate tramite Grafana, una piattaforma software nata nel 2014 e specializzata nella visualizzazione e nell'analisi interattiva dei dati [68].

Nel presente lavoro, Grafana è stato collegata al database AirQuality, ospitato su PostgreSQL e gestito tramite PgAdmin4, dal quale vengono recuperati i dati necessari alla costruzione delle diverse rappresentazioni grafiche.

Questa integrazione ha permesso di realizzare molteplici dashboard, ciascuna progettata per uno specifico obiettivo di analisi e basata sulle tabelle e sulle viste precedentemente descritte nel Capitolo 3

Grafana permette di strutturare tali dashboard inserendo diversi pannelli, ciascuno dei quali può essere configurato per rappresentare i dati secondo modalità differenti, in particolare:

- Time series, utilizzato per rappresentare l'andamento temporale delle concentrazioni;
- Bar chart, utilizzato per rappresentazioni mediante grafici a barre;
- Stat, utilizzato per mostrare singoli valori o indicatori sintetici;
- Pie chart, utilizzato per rappresentare la distribuzione delle stazioni e valori tra le differenti categorie;
- Status history, utilizzato per mostrare l'evoluzione temporale di specifici valori;
- Geomap, utilizzato per la rappresentazione geografica di stazioni e dei vari CAP;

- XY Chart, utilizzato per rappresentare la relazione tra due variabili (per sostituire il pannello Time Series in alcune dashboard, in particolare nei casi in cui la gestione dell'intervallo temporale è stata affidata alla variabile `date_relapse`).

Ogni pannello interroga direttamente il database mediante una query SQL, attraverso la quale vengono selezionati i dati necessari alla specifica rappresentazione. I risultati delle query possono successivamente essere elaborati attraverso le funzionalità messe a disposizione da Grafana. Tra queste assumono particolare importanza le *transformations*, che consentono di modificare e riorganizzare i dati prima della loro visualizzazione sul pannello, e gli *overrides*, i quali permettono di definire specifiche modalità di rappresentazione per specifici valori o condizioni.

Oltre ai pannelli, Grafana mette a disposizione alcuni strumenti globali che consentono all'utente di interagire dinamicamente con la dashboard. Tra questi rientra il *time picker*, attraverso il quale è possibile selezionare l'intervallo di tempo di interesse, specificando manualmente la data iniziale e finale oppure scegliendo tra alcuni intervalli predefiniti.

Un ulteriore strumento è rappresentato dalle *variables*. Tramite apposite query SQL è stato possibile definire variabili i cui valori possono essere selezionati dall'utente. Questa funzionalità permette di modificare direttamente il sottoinsieme di dati visualizzato nei diversi pannelli, rendendo pertanto le dashboard adattabili alle specifiche esigenze dell'utente (ad esempio modificando in modo dinamico l'inquinante, l'area geografica, la tipologia di stazione e l'intervallo di tempo analizzato).

Una limitazione riscontrata durante lo sviluppo riguarda l'organizzazione del *time picker* e delle variabili, che risultano fissate automaticamente nella parte superiore della dashboard e non possono essere spostati arbitrariamente all'interno dell'interfaccia. Si è riscontrato, inoltre, che il *time picker* non si aggiorna automaticamente in risposta alla selezione di una data specifica all'interno della rispettiva variabile, evidenziando una mancanza di sincronizzazione tra i due componenti.

Le funzionalità descritte costituiscono la base utilizzata per lo sviluppo delle dashboard del presente lavoro. La progettazione delle singole schermate è stata quindi orientata non soltanto alla corretta rappresentazione dei dati, ma anche alla loro immediatezza e facilità di interpretazione, con particolare attenzione alle informazioni potenzialmente utili al personale clinico nella valutazione del contesto ambientale a cui il paziente è esposto.

4.1 Dashboard Home

L'interfaccia Home presenta una panoramica generale del database AirQuality, attraverso statistiche descrittive relative alle concentrazioni degli inquinanti e alle stazioni di monitoraggio presenti nel database, senza considerare le stime ottenute mediante i modelli LUR e IDW precedentemente sviluppati.

Nella parte superiore della dashboard sono presenti le variabili che consentono di filtrare i dati visualizzabili, insieme al selettore temporale (time picker) e al menù che permette di spostarsi tra le diverse dashboard (Figura 4.1).

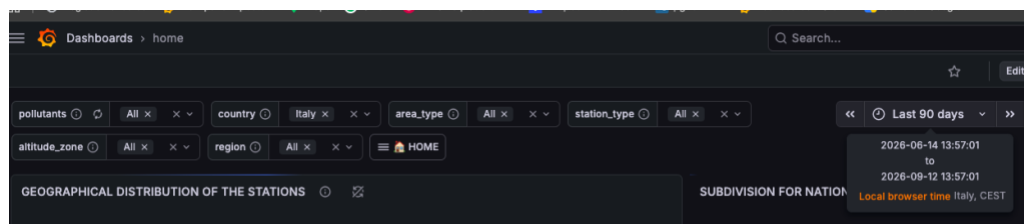


Figura 4.1: Variabili, time picker e menù presenti nella parte superiore della dashboard Home

Tutte le variabili permettono di selezionare più valori contemporaneamente e includono l'opzione "All" attraverso la quale è possibile includere tutti i valori disponibili.

Le variabili utilizzate come filtri sono:

- Pollutants, che permette di selezionare uno o più inquinanti di interesse
- Country, che consente di selezionare Spagna, Portogallo, Italia;
- Station_type, che permette di distinguere tra stazioni di tipo industrial, situate in prossimità di aree industriali, background, rappresentative dell'esposizione media della popolazione, e traffic, situate in prossimità di una strada.

- Area_type, che permette di distinguere tra stazioni situate in area urban, suburban e rural;
- Altitude_zone: che suddividono le stazioni in pianura(altitudine minore di 300 m), collina(altitudine tra 300 m e 650 m) e montagna(altitudine maggiore di 650 m) [69];
- Region, che consente di selezionare le diverse sotto-regioni geografiche. Per l'Italia è possibile scegliere tra Nord, Centro e Sud, mentre per Spagna e Portogallo la scelta include Penisola Iberica, Isole Baleari, Canarie e Azzorre&Madeira.

L'utilizzo di uno o più di questi filtri permette quindi all'utente di restringere l'analisi ad uno specifico insieme di stazioni e di inquinanti. Per default, tutte le variabili sono settate su "All" così da fornire una visione complessiva dei dati disponibili.

Nella parte superiore della dashboard è presente, oltre alle variabili, il time picker, il quale permette di selezionare l'intervallo temporale di interesse. Avendo deciso di visualizzare solamente misure valide e verificate, la quantità di dati disponibile può variare in funzione dell'intervallo selezionato. In particolare, nei periodi più recenti il volume dei dati potrebbe risultare ridotto in quanto le misurazioni presenti non risultano ancora verificate.

La dashboard presenta, inoltre, una serie di pannelli. Il primo pannello (Figura 4.2) è costituito da una mappa geografica realizzata mediante il pannello Geomap, che permette di visualizzare la distribuzione delle stazioni di monitoraggio.

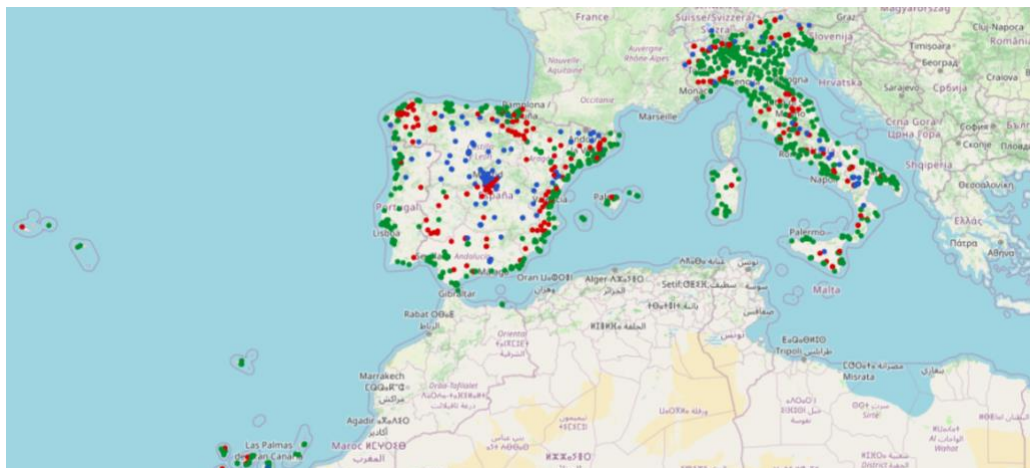


Figura 4.2: Rappresentazione geografica delle stazioni di monitoraggio presenti su AirQuality mediante il pannello Geomap. In verde le stazioni in pianura, in rosso quelle in collina e in blu quelle in montagna.

La mappa mostra le stazioni che, in funzione dell'intervallo temporale e dei filtri selezionati, dispongono di misure valide, verificate e aventi valori positivi entro i limiti stabiliti nella Tabella 3.5. Le stazioni vengono inoltre differenziate con uno specifico colore in base alla fascia altimetrica di appartenenza. In particolare, sono stati arbitrariamente scelti il verde per la pianura, il rosso per la collina e il blu per la montagna. Passando con il cursore sopra ad uno specifico pallino viene visualizzata una finestra contenente le principali informazioni geografiche associate alla stazione. Nella Figura 4.3 è mostrato, a titolo di esempio, le informazioni relative alla stazione di monitoraggio di Piazza Minerva a Pavia.

altitude_stations	
station_name	PAVIA - P.ZZA MINERVA
altitude	77
latitude	45.2
longitude	9.15
Zona Geografica	Northern Italy

Figura 4.3: Finestra informativa contenente le informazioni geografiche associate alla stazione di monitoraggio.

A fianco e sotto la mappa sono presenti diversi grafici a torta, utilizzati per rappresentare la distribuzione delle stazioni rispetto ai principali attributi disponibili nel database. In Figura 4.4, vengono mostrate le suddivisioni delle

stazioni in base all'area_type, a station_type e alla altitude_zone. Le distribuzioni sono calcolate considerando esclusivamente le stazioni che hanno restituito valori validi e verificati negli ultimi cinque anni e che rispettano i filtri impostati dall'utente tramite le apposite variabili.

È inoltre presente un banner che indica l'intervallo temporale per il quale risultano disponibili misurazioni all'interno del database.



Figura 4.4: Suddivisione del numero di stazioni per area_type, altitude_zone e station_type. In basso, range temporale coperto da AirQuality.

Nella parte inferiore della dashboard viene rappresentato l'andamento temporale delle concentrazioni degli inquinanti selezionati (Figura 4.5). Per ciascun giorno viene calcolata la concentrazione media considerando esclusivamente le misurazioni, valide e verificate, relative alle stazioni che soddisfano i filtri precedentemente impostati.

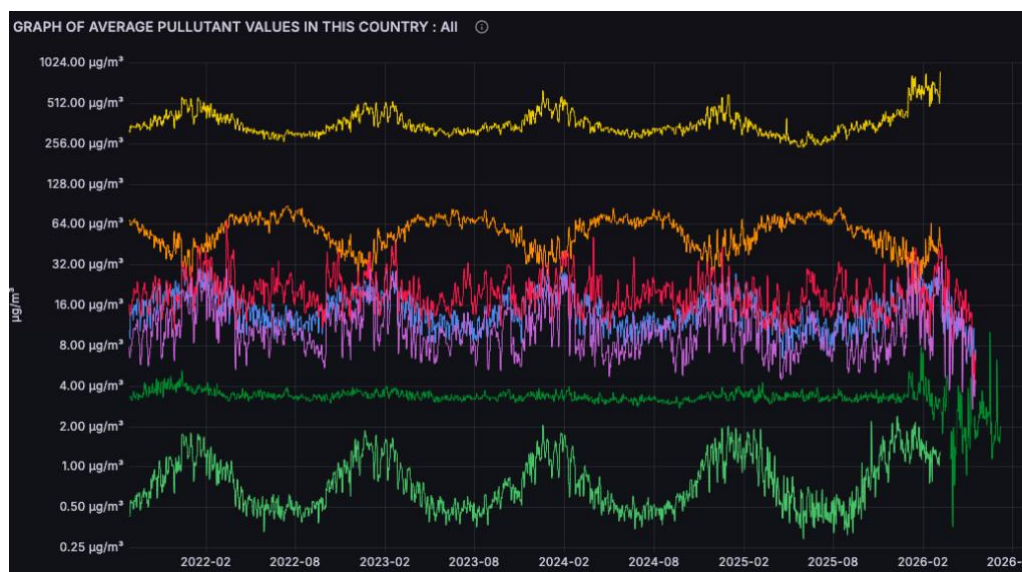
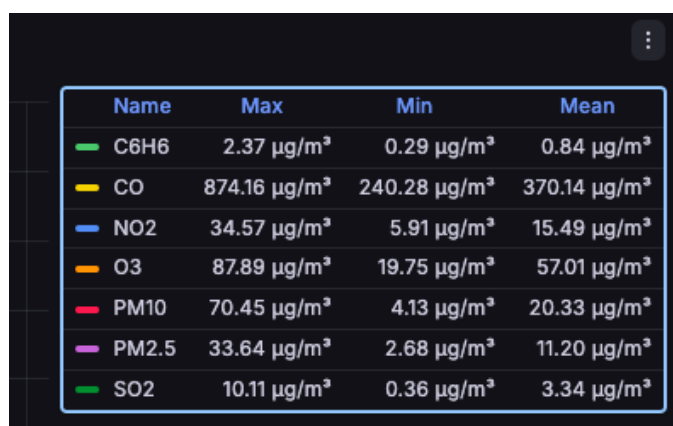


Figura 4.5: Andamento temporale della concentrazione media giornaliera degli inquinanti misurati dalle stazioni di monitoraggio che rispettano i vincoli impostati dall'utente attraverso le variabili di selezione. In questo caso, tutte le variabili sono impostate su All. La legenda relativa agli inquinanti rappresentati, insieme ai rispettivi valori medio, massimo e minimo, è riportata in Figura 4.6.

Per facilitare la rappresentazione simultanea di inquinanti caratterizzati da intervalli di concentrazione differenti, il grafico è impostato su una scala logaritmica. Le concentrazioni dei diversi inquinanti, considerando le singole misurazioni acquisite da AirQuality, sono espresse in $\mu\text{g}/\text{m}^3$. Il CO, invece, è misurato in mg/m^3 su AirQuality; pertanto, per consentirne la rappresentazione insieme agli altri inquinanti, le relative misure sono state convertite in $\mu\text{g}/\text{m}^3$ mediante la query SQL attraverso cui viene realizzato il pannello, moltiplicando i valori per 1000.

Nella Figura 4.6 è riportata la legenda degli inquinanti, nella quale, per ciascun inquinante, sono indicati il colore associato e i valori massimo, minimo e medio calcolati nell'intervallo temporale selezionato.



Name	Max	Min	Mean
C6H6	2.37 $\mu\text{g}/\text{m}^3$	0.29 $\mu\text{g}/\text{m}^3$	0.84 $\mu\text{g}/\text{m}^3$
CO	874.16 $\mu\text{g}/\text{m}^3$	240.28 $\mu\text{g}/\text{m}^3$	370.14 $\mu\text{g}/\text{m}^3$
NO2	34.57 $\mu\text{g}/\text{m}^3$	5.91 $\mu\text{g}/\text{m}^3$	15.49 $\mu\text{g}/\text{m}^3$
O3	87.89 $\mu\text{g}/\text{m}^3$	19.75 $\mu\text{g}/\text{m}^3$	57.01 $\mu\text{g}/\text{m}^3$
PM10	70.45 $\mu\text{g}/\text{m}^3$	4.13 $\mu\text{g}/\text{m}^3$	20.33 $\mu\text{g}/\text{m}^3$
PM2.5	33.64 $\mu\text{g}/\text{m}^3$	2.68 $\mu\text{g}/\text{m}^3$	11.20 $\mu\text{g}/\text{m}^3$
SO2	10.11 $\mu\text{g}/\text{m}^3$	0.36 $\mu\text{g}/\text{m}^3$	3.34 $\mu\text{g}/\text{m}^3$

Figura 4.6: Legenda degli inquinanti, con indicazione del relativo colore e dei valori medio, minimo e massimo.

Infine, nella parte inferiore della dashboard, sono riportati il numero totale di stazioni presenti nel database e, in base ai filtri e all'intervallo di tempo selezionati, il numero complessivo di misure, la percentuale di misure valide e verificate e il numero di stazioni attive.

4.2 Dashboard di analisi per area geografica

La seconda dashboard, denominata `stazioni_per_zona`, è stata progettata con l'obiettivo di analizzare le concentrazioni degli inquinanti nelle diverse aree geografiche considerate.

A differenza delle Dashboard Home, dove viene fornita una visione generale della rete di monitoraggio, questa schermata è maggiormente orientata alla

valutazione delle concentrazioni in specifiche aree territoriali e al confronto con le AQG.

Le variabili selezionabili come filtro sono due: Pollutants e Country.

Per ogni inquinante selezionato, viene generata una riga (row) contenente l'andamento temporale della concentrazione media giornaliera nelle diverse aree geografiche del Paese scelto.

All'interno dello stesso grafico sono riportate le AQG e gli Interim Targets relativi all'inquinante selezionato, permettendo di confrontarne visivamente l'andamento con le relative soglie imposte dalla WHO.

In Figura 4.7 viene mostrato come esempio l'andamento del PM_{2.5} nelle tre sotto-regioni italiane, con la relativa legenda (Figura 4.8).



Figura 4.7: Andamento temporale delle concentrazioni medie giornaliere di PM_{2.5} negli ultimi 5 anni, suddiviso per macroregione (Nord, Centro e Sud Italia). La legenda è riportata nella Figura 4.8.

Name ^	Max	Mean
TARGET 1	75 $\mu\text{g}/\text{m}^3$	75 $\mu\text{g}/\text{m}^3$
TARGET 2	50 $\mu\text{g}/\text{m}^3$	50 $\mu\text{g}/\text{m}^3$
TARGET 3	37.5 $\mu\text{g}/\text{m}^3$	37.5 $\mu\text{g}/\text{m}^3$
TARGET 4	25 $\mu\text{g}/\text{m}^3$	25 $\mu\text{g}/\text{m}^3$
AQG VALUE	15 $\mu\text{g}/\text{m}^3$	15 $\mu\text{g}/\text{m}^3$
Central Italy	39.2 $\mu\text{g}/\text{m}^3$	11.8 $\mu\text{g}/\text{m}^3$
Northern Italy	47.3 $\mu\text{g}/\text{m}^3$	14.9 $\mu\text{g}/\text{m}^3$
Southern Italy	83 $\mu\text{g}/\text{m}^3$	13.0 $\mu\text{g}/\text{m}^3$

Figura 4.8: Legenda della Figura 4.7. Codifica dei colori per le aree geografiche (Nord, Centro, Sud) e per le soglie normative (IT, AQG), integrata con i valori medi e massimi per macroregione

Sono stati aggiunti ulteriori pannelli che riportano, per ciascuna area, il numero di giorni nei quali la concentrazione supera il valore di riferimento selezionato (AQG o Interim Targets) e la relativa percentuale rispetto al numero complessivo di giorni presenti nell'intervallo di tempo selezionato.

Sulla base di tale percentuale è stato inoltre introdotto un indicatore sintetico della qualità dell'aria, rappresentato attraverso differenti categorie:

- Good, se la percentuale è sotto al 2%;
- Moderate, se la percentuale è sotto al 20%;
- Unhealthy, se la percentuale è sotto al 50%;
- Very Unhealthy, se la percentuale è sotto il 75%;
- Hazardous se è superiore al 75%.

È opportuno precisare che i criteri percentuali utilizzati nel presente lavoro sono stati elaborati ad hoc e non si basano su classificazioni ufficiali della WHO o di altre istituzioni. Al contrario, la nomenclatura e la codifica cromatica delle fasce di qualità dell'aria seguono le linee guida dell'EPA, ente responsabile del protocollo AQI qui applicato [30].

Attraverso la funzionalità Value Mapping, Grafana consente di associare in modo univoco un determinato colore a una specifica etichetta. Il colore può essere selezionato sia tra quelli predefiniti disponibili nella piattaforma, sia,

come in questo caso, mediante l'inserimento del relativo codice RGB. In tal modo, a ciascuna fascia individuata è stato associato il colore corrispondente alla classificazione di riferimento dell'EPA.

I pannelli ottenuti considerando le concentrazioni medie giornalieri di PM_{2.5} negli ultimi 5 anni per ciascuna macroregione sono riportati nella Figura 4.9.



Figura 4.9: Pannello contenente le fasce di classificazione associate da ciascuna macroregione considerando i valori di PM_{2.5} rilevati negli ultimi 5 anni

La medesima serie di pannelli è stata inoltre realizzata sostituendo la suddivisione geografica con una classificazione basata sull'altitudine, raggruppando le stazioni nelle categorie pianura, collina e montagna in funzione della quota altimetrica della stazione.

Questa rappresentazione ha permesso di evidenziare eventuali differenze nelle concentrazioni dei diversi inquinanti associate alle diverse fasce altimetriche.

I due insiemi di pannelli consentono quindi di mettere in evidenza alcune caratteristiche generali della distribuzione dei vari inquinanti.

In particolare, per quasi tutti gli inquinanti il Nord Italia presenta concentrazioni medie superiori rispetto alle altre macroaree italiane, mentre nelle isole spagnole si osservano generalmente concentrazioni inferiori rispetto alla penisola iberica. Analogamente, la suddivisione in base all'altitudine mostra concentrazioni maggiori nelle aree di pianura rispetto alle zone collinari e montuose, coerentemente con la maggiore presenza di centri urbani e sorgenti emissive nelle aree a bassa quota.

A titolo di esempio, in Figura 4.10 è riportato l'andamento della concentrazione dell'O₃ negli ultimi 5 anni nelle tre fasce altimetriche. L'andamento dell'O₃ risulta maggiormente influenzato dalla componente stagionale, coerentemente con i processi che ne regolano la formazione. La rappresentazione temporale permette dunque di evidenziare in modo immediato tale comportamento.



Figura 4.10: Andamento temporale delle concentrazioni medie giornaliere di O_3 negli ultimi 5 anni, suddiviso per fascia altimetrica (pianura in viola, collina in azzurro e montagna in blu).

Le dashboard descritte costituiscono la base per la successiva dashboard dedicata al singolo paziente, nella quale le informazioni relative alla qualità dell'aria vengono integrate con i dati geografici e temporali associati al soggetto.

4.3 Dashboard per l'analisi della qualità dell'aria nel CAP selezionato

Successivamente, è stata sviluppata una dashboard finalizzata a visualizzare la qualità dell'aria in corrispondenza dei luoghi presenti nella tabella `postcode_places`, utilizzando le misure rilevate dalle stazioni di monitoraggio limitrofe.

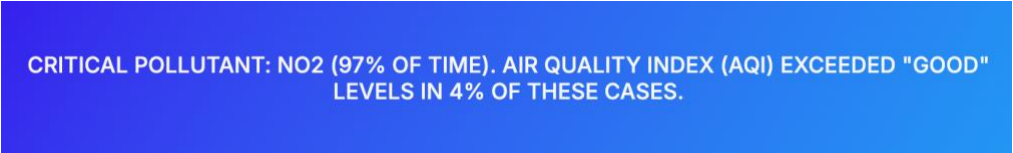
È stata creata in Grafana una variabile che dà la possibilità all'utente di selezionare un CAP tramite un apposito menù a tendina. Una volta selezionato, una seconda variabile mostra le 5 stazioni più vicine, considerando un raggio massimo di 30 km. Qualora all'interno del raggio siano presenti meno di 5 stazioni, vengono mostrate soltanto quelle disponibili. Le stazioni visualizzate sono quelle che risultano attive nell'intervallo selezionato, ovvero quelle che restituiscono misure valide e verificate per l'inquinante selezionato attraverso la variabile `pollutants`. Una volta effettuate le scelte, nella parte superiore della dashboard viene mostrato l'effettivo intervallo temporale per il quale sono presenti misure. Infatti, il periodo selezionato dall'utente può risultare ridotto

qualora, nelle porzioni iniziali o finali dell'intervallo, le stazioni considerate non dispongano di misure valide e verificate.

La schermata comprende inoltre una rappresentazione geografica del CAP selezionato e delle relative stazioni più vicine. Per ciascuna stazione vengono riportate anche informazioni aggiuntive, tra cui la distanza dal CAP di riferimento. Tale distanza viene calcolata tramite una query SQL che recupera le coordinate geografiche della stazione e del relativo CAP e applica la formula di Harvesine (3.2), ottenendo la distanza tra i due punti sulla superficie terrestre. Sono inoltre presenti pannelli descrittivi delle stazioni selezionate, nei quali vengono riportati il numero di misure disponibili e gli inquinanti misurati. La dashboard permette inoltre di visualizzare l'andamento della qualità dell'aria, misurata tramite AQI, attraverso diversi pannelli, relativi alla distribuzione stagionale e mensile, all'andamento orario e all'evoluzione osservata nelle ultime due settimane a partire dalla prima data per la quale sono disponibili misure.

In generale, per tutti gli inquinanti fortemente dipendenti dall'emissione antropica, quali il CO, il NO₂, il PM₁₀ e il PM_{2.5}, si osservano valori di AQI tendenzialmente più elevati nei periodi autunnali e invernali. Tale comportamento può essere ricondotto alla maggiore incidenza delle emissioni derivanti dai sistemi di riscaldamento e alle condizioni atmosferiche tipiche di questi mesi, che possono dunque favorire l'accumulo degli inquinanti negli strati più bassi dell'atmosfera.

A completamento della schermata è presente un banner che identifica l'inquinante avente la situazione peggiore nel luogo e nel periodo selezionato, riportando anche la percentuale di giorni nei quali viene superato il valore stabilito dalle AQG (Figura 4.11).



CRITICAL POLLUTANT: NO₂ (97% OF TIME). AIR QUALITY INDEX (AQI) EXCEEDED "GOOD" LEVELS IN 4% OF THESE CASES.

Figura 4.11: Banner contenente l'indicazione dell'inquinante caratterizzato dalla maggiore frequenza di superamento della soglia Good presso la stazione di Piazza Minerva a Pavia negli ultimi cinque anni. È inoltre riportata la percentuale di misurazioni in cui tale soglia risulta superata.

È possibile, inoltre, analizzare l'andamento orario delle concentrazioni dei vari inquinanti. Particolarmente interessante risulta il comportamento dell'O₃ il quale, oltre ad avere picchi nei mesi estivi, presenta valori generalmente più elevati durante le ore diurne, in accordo con la sua forte dipendenza dalla radiazione solare.

Per gli altri inquinanti si osserva tendenzialmente un andamento orario generalmente quasi costante.

Nella Figura 4.12 è possibile osservare un andamento differente. Il SO₂, le cui concentrazioni sono solitamente ridotte e dipendenti da precise fonti di emissione, mostra picchi nelle ore di maggiore attività antropica.

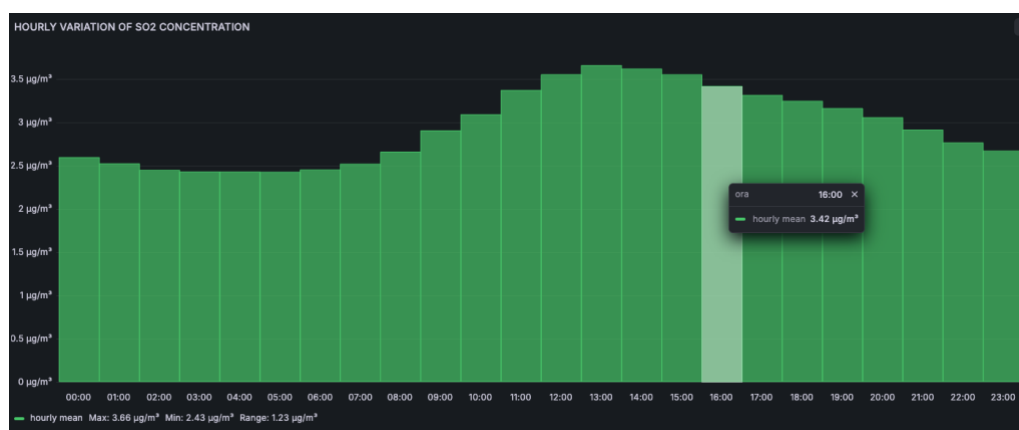


Figura 4.12: Andamento temporale delle concentrazioni medie orarie di SO₂ negli ultimi 5 anni misurate presso la stazione di Via Folperti a Pavia.

4.4 Dashboard Paziente

L'obiettivo della dashboard Paziente è integrare le informazioni relative al paziente, alle sue visite e alle sue ricadute con una descrizione della qualità dell'aria nel comune di residenza, utilizzando le stime delle concentrazioni degli inquinanti precedentemente ottenute con gli opportuni metodi.

A differenza delle dashboard presentate nelle sezioni 4.1, 4.2, 4.3, nelle quali erano presenti anche le misurazioni relative a Spagna e Portogallo, questa schermata e la successiva sono state sviluppate considerando soltanto il territorio italiano. Infatti, come precedentemente descritto, per ragioni computazionali e di memoria, le stime delle concentrazioni degli inquinanti sono state effettuate sui CAP di residenza dei pazienti e sui CAP dei capoluoghi delle province italiane.

Nella parte superiore della dashboard sono presenti le variabili che consentono al personale clinico di selezionare il paziente e l'inquinante di interesse. Ciascun paziente è identificato attraverso un codice univoco. In ottica futura, tale variabile potrebbe essere associata anche al nome del paziente, mantenendo comunque adeguate metodologie di protezione dei dati personali.

Una volta selezionato il paziente, la variabile `date_relapse` permette di scegliere una delle date di ricaduta associate al paziente e presenti nel database AirQuality nella vista `view_patients`. La data selezionata costituisce il riferimento temporale per le successive analisi, che vengono effettuate generalmente considerando i due anni precedenti alla data considerata.

Come descritto negli scorsi paragrafi (pag. 71), il time picker di Grafana non viene aggiornato automaticamente in funzione della data selezionata nella variabile `date_relapse`. Pertanto, è stata necessaria l'introduzione di una nuova variabile che, tramite interazione dell'utente, aggiorna il time picker con il corretto intervallo temporale, così da garantire la coerenza temporale dei dati mostrati rispetto alla data di ricaduta selezionata dall'utente.

La variabile `date_relapse` permette inoltre di inserire manualmente una data non associata a una ricaduta presente nel database. In questo caso, Grafana non fornisce un selettore basato su calendario e di conseguenza la data deve essere inserita manualmente nel formato YYYY-MM-DD. In questo modo è possibile quindi analizzare periodi temporali differenti da quelli associati ad eventi di ricaduta presenti nel database.

Per rendere la schermata il più possibile semplice e immediata, non è stata fornita al personale clinico la possibilità di selezionare direttamente il metodo di stima. Le concentrazioni visualizzate vengono infatti ottenute utilizzando, per ogni inquinante, il metodo che ha mostrato le migliori prestazioni complessive nella fase di confronto.

I metodi di stima selezionati per ciascun inquinante sono riportati nella Tabella 4.1.

Tabella 4.1: Metodo selezionato per ciascun inquinante

INQUINANTE	PRIMA SCELTA
SO ₂	LUR
PM ₁₀	IDW
O ₃	IDW
NO ₂	LUR
CO	IDW
C ₆ H ₆	LUR
PM _{2.5}	IDW

Per ciascun pannello inserito, è stato implementato un controllo che verifica la disponibilità della stima giornaliera mediante il metodo preferenziale. Qualora tale stima non sia disponibile, viene utilizzato il metodo alternativo. Questa strategia permette di minimizzare la presenza di valori nulli, che si verificano solamente quando nessuno dei due metodi è in grado di generare una stima per il luogo e l'istante considerato.

Nella Figura 4.13 è mostrata la prima sezione della dashboard, riportante le informazioni relative alla ricaduta selezionata.

Duration (days)	Brainstem	Relapse Recovery	Corticosteroid Treatment	Inpatient Treatment
15	Si	Parziale	Si	Si
CRITICAL POLLUTANT: PM2.5 (87.4% OF TIME, 646 / 739 DAYS). AIR QUALITY INDEX (AQI) EXCEEDED "GOOD" LEVELS IN 76.5% OF THESE CASES. DAILY PM2.5 VALUES EXCEEDED WHO AIR QUALITY GUIDELINES (AQG) IN 55.6% OF THESE DAYS				
Municipality(CAP)	Province	Region	Inhabitants	
Albuzzano (27010)	Pavia	Lombardia	358	

Figura 4.13: Informazioni relative alla qualità dell'aria e alla ricaduta nel comune di residenza del paziente n. 37735, con i relativi dati comunali.

In questi pannelli vengono indicati il distretto anatomico interessato, la durata dell'evento e l'esito del recupero, oltre ad eventuali trattamenti in atto sul paziente. La disponibilità di tali informazioni è tuttavia limitata.

Nella stessa parte della schermata è inoltre presente un banner che identifica l'inquinante avente la situazione più critica nei due anni precedenti alla data della

ricaduta. Per tale inquinante vengono riportati il numero di giorni nei quali l'AQI supera il livello "Good" e la percentuale di giorni nei quali viene superato il valore delle AQG.

Nella sezione successiva vengono riportate le informazioni relative al luogo di residenza del paziente, tra cui il Comune, il CAP e il numero di residenti.

Scorrendo la dashboard, nella parte sinistra è presente una mappa che rappresenta l'area limitrofa al luogo di residenza, mostrando il CAP scelto, i capoluoghi di provincia e le stazioni di monitoraggio più vicine.

Accanto alla mappa sono presenti ulteriori pannelli dedicati alle informazioni cliniche e ambientali del paziente, riportati nella Figura 4.14.

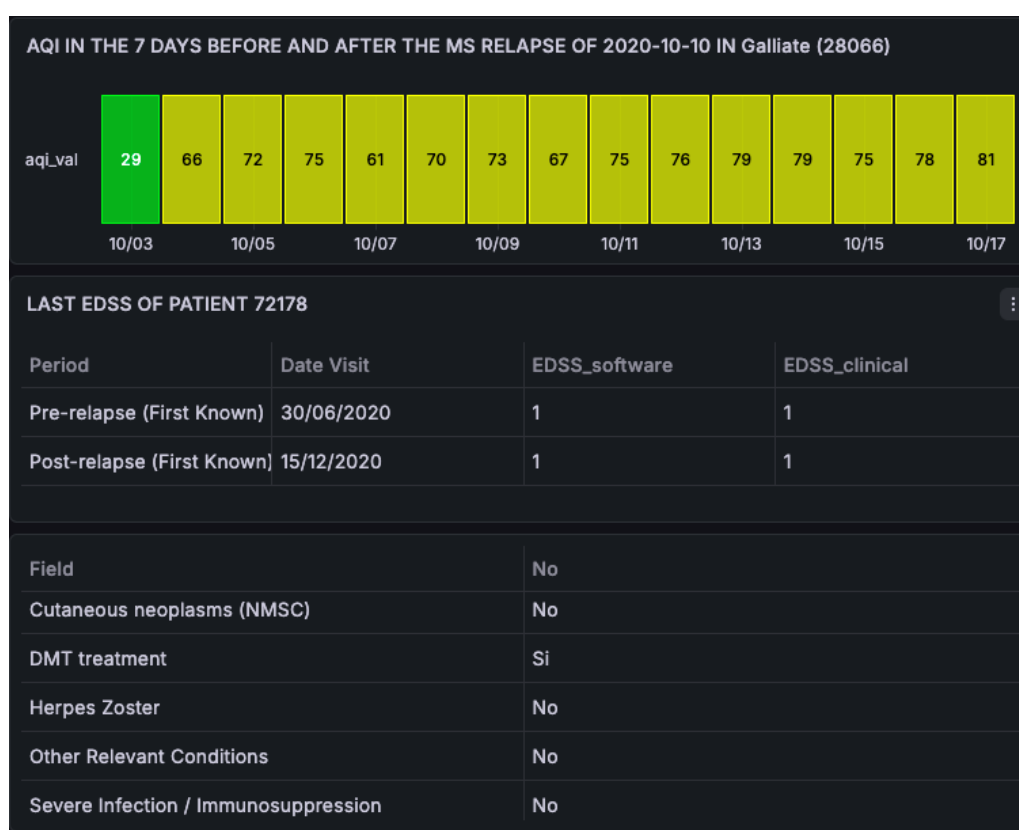


Figura 4.14: Dashboard integrata per il monitoraggio del singolo paziente. L'interfaccia visualizza, dall'alto verso il basso: l'andamento dell'AQI nel comune di residenza nei 7 giorni antecedenti e successivi alla data selezionata; l'evoluzione dell'EDSS prima e dopo la data di riferimento; il quadro clinico aggiornato all'ultima visita effettuata.

Tra questi è presente un pannello relativo all'AQI ottenuto nell'intervallo temporale di 7 giorni intorno alla data di relapse. La scelta dell'intervallo temporale è stata effettuata facendo riferimento alla letteratura recente, nella quale l'esposizione agli inquinanti atmosferici nei 7 giorni precedenti viene

considerata come esposizione a breve termine ed è stata associata al rischio di relapse nella sclerosi multipla [70].

Un ulteriore pannello è dedicato all'EDSS, parametro utilizzato per descrivere lo stato di disabilità associato al paziente. Sono riportate entrambe le versioni presenti nel database, ovvero EDSS_software, ottenuto mediante applicazione diretta della scala, ed EDSS_clinical, assegnato dal medico sulla base della visita condotta.

Il pannello mostra i valori disponibili prima e dopo la ricaduta, consentendo di osservare l'eventuale variazione del livello di disabilità a seguito della ricaduta. L'ultimo pannello di questa sezione mostra le informazioni cliniche raccolte nell'ultima visita antecedente alla ricaduta che presenta non nulli. Nello specifico, il pannello riporta l'eventuale presenza di infezioni gravi, herpes o neoplasie e il trattamento DMT. Tali informazioni possono essere utili al medico per disporre di un quadro clinico più completo in prossimità dell'evento di ricaduta.

Nella parte inferiore della dashboard sono presenti diverse sezioni espandibili dedicate, in ordine, all'AQI, all'EDSS e all'inquinante selezionato nell'apposito menù a tendina.

La prima sezione è dedicata all'AQI associato al comune di residenza del paziente. Sono presenti tre rappresentazioni relative, rispettivamente, alla distribuzione stagionale, all'andamento mensile e alla distribuzione dei livelli di AQI nell'anno precedente e successivo alla data della ricaduta.

Nel presente lavoro, è stato utilizzato l'AQI dell'EPA, convertendo le concentrazioni giornaliere nella metrica richiesta [29]. L'O₃ non è stato incluso nel calcolo dell'AQI poiché le stime disponibili nel presente lavoro sono giornaliere, mentre il calcolo richiede una metrica temporale diversa. Non avendo individuato un corretto metodo di conversione della stima nel formato richiesto, si è preferito non introdurre una conversione potenzialmente errata.

Il primo pannello mostra il valor medio dell'AQI nelle quattro stagioni (Figura 4.15).

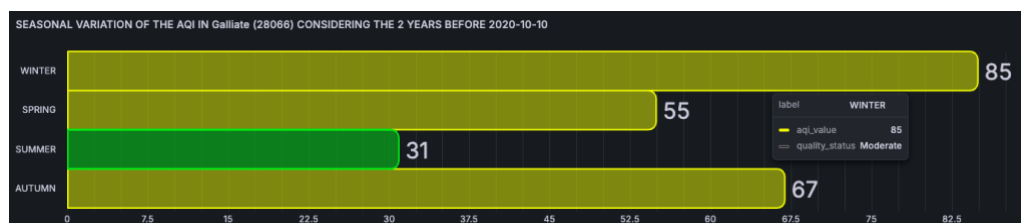


Figura 4.15: Andamento stagionale dell'AQI in Galliate nei due anni precedenti il 10-10-2020

Tendenzialmente, si osservano valori maggiori in autunno e in inverno, mentre durante il periodo estivo la qualità dell'aria risulta mediamente migliore per gli inquinanti associati alle emissioni antropiche.

Un secondo grafico, costruito secondo la stessa logica, riporta il valore medio dell'AQI per ciascun mese, permettendo di evidenziare con maggior precisione eventuali periodi caratterizzati da una maggiore esposizione agli inquinanti.

L'ultimo pannello della riga espandibile mostra un grafico a torta che illustra la distribuzione dei giorni nelle diverse classi di qualità dell'aria previste dall'AQI. La seconda riga estendibile è dedicata all'EDSS e ha lo scopo di mostrare l'evoluzione nel tempo dei valori disponibili per il paziente.

In particolare, vengono visualizzati nel primo pannello gli storici di EDSS_software e EDSS_clinical, insieme ai valori relativi ai singoli FS, permettendo di visualizzare il peggioramento della disabilità nel dettaglio.

Nel pannello successivo, attraverso la Time Series, è possibile osservare l'evoluzione temporale del livello di disabilità e di confrontare tale andamento con le date delle varie ricadute.

Nella legenda vengono inoltre riportate informazioni sintetiche, quali il valore massimo, il primo e l'ultimo valore disponibile di EDSS. La Figura 4.16 mostra l'evoluzione temporale dell'EDSS per uno dei pazienti in cura presso l'Istituto Neurologico Nazionale Casimiro Mondino di Pavia.



Figura 4.16: Andamento temporale della disabilità del paziente n. 37092, misurata attraverso l'EDSS_software (in verde) e l'EDSS_clinical (in giallo).

Le ultime due sezioni sono dedicate all'analisi degli specifici inquinanti selezionati mediante l'apposita variabile. Sono state realizzate due righe distinte: una dedicata agli inquinanti utilizzati per il calcolo dell'AQI e una dedicata agli altri inquinanti.

Per la prima tipologia di inquinanti, la sezione mostra innanzitutto il numero dei giorni nei quali, nell'anno precedente e successivo alla data della ricaduta, viene oltrepassato il valore fissato dalle AQG. Successivamente viene riportata l'evoluzione temporale della concentrazione dell'inquinante, confrontata con gli IT e AQG disponibili (Figura 4.17).

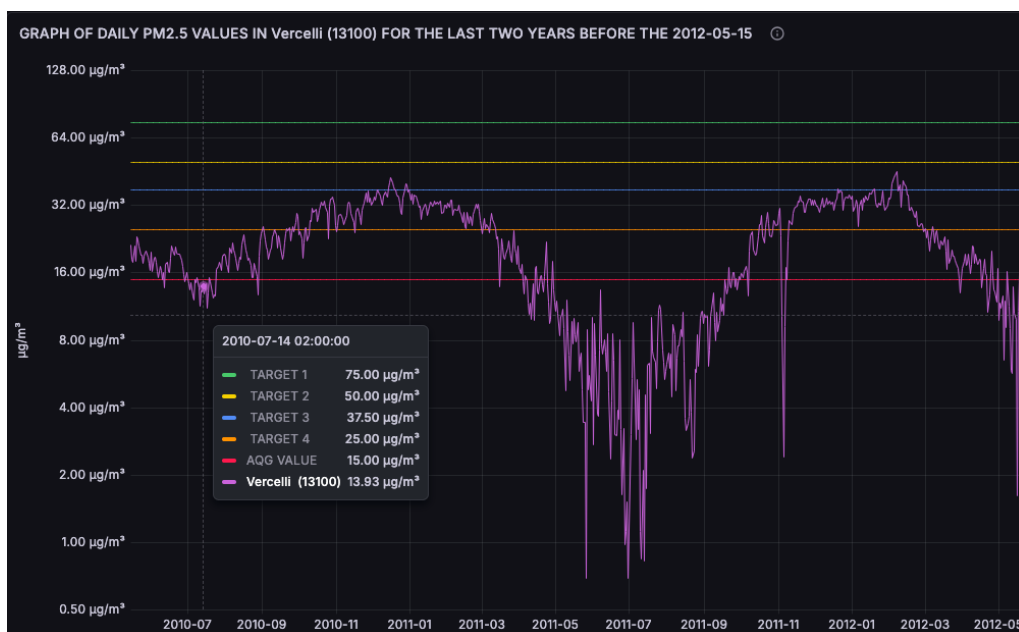


Figura 4.17: Andamento temporale della concentrazione media giornaliera di $PM_{2.5}$ a Vercelli dal 15/5/2010 al 15/5/2012.

Sono inoltre presenti pannelli dedicati all'AQI specifico per l'inquinante selezionato, comprendenti la distribuzione stagionale e mensile e l'andamento nella settimana precedente e successiva alla ricaduta.

Un ulteriore grafico a torta mostra la distribuzione dei livelli di AQI osservati nel CAP di residenza. Tali informazioni, riportate nella Figura 4.18, vengono aggiornate dinamicamente in funzione dell'inquinante selezionato.



Figura 4.18: A sinistra, l'AQI stimato nei sette giorni antecedenti e successivi alla data relapse selezionata. A destra è riportata la suddivisione dei livelli di AQI osservati nei due anni precedenti rispetto alla data selezionata.

La riga successiva è invece dedicata agli inquinanti per i quali non è possibile calcolare l'AQI. Pertanto, vengono mostrati esclusivamente i grafici relativi alle concentrazioni e alla loro evoluzione nel tempo, senza riportate le statistiche basate sull'AQI.

In generale, la dashboard paziente integra quindi informazioni cliniche, geografiche e ambientali, fornendo una rappresentazione sintetica della qualità dell'aria nel comune di residenza e nel periodo scelto dal medico.

L'obiettivo non è stabilire un rapporto causale tra l'esposizione agli inquinanti e l'insorgenza della ricaduta, ma fornire al medico uno strumento per individuare eventuali coincidenze temporali e condizioni ambientali particolarmente critiche per lo specifico paziente.

In questo modo, il medico può utilizzare tali informazioni come supporto alla valutazione clinica e, se opportuno, fornire indicazioni volte a limitare l'esposizione all'aria aperta durante periodi aventi livelli di inquinamento particolarmente elevati.

4.5 Dashboard per l'analisi della qualità dell'aria a livello provinciale e regionale

L'ultima dashboard sviluppata, denominata `cap_ita`, ha l'obiettivo di estendere l'osservazione delle concentrazioni degli inquinanti oltre al comune di residenza del paziente, permettendo di analizzare il contesto provinciale e regionale.

Tale aspetto è motivato dalla possibilità che il paziente, durante determinati periodi, possa trascorrere periodi più o meno lunghi in un comune diverso da quello di residenza, ad esempio per motivi lavorativi, risultando quindi esposto a una differente condizione ambientale.

La schermata introduce innanzitutto la variabile `cap`, che permette di selezionare il CAP di interesse tra tutti i CAP italiani. Come descritto nei precedenti capitoli, per ragioni legate al costo computazionale e all'occupazione di memoria non è stato possibile effettuare la stima specifica per tutti i CAP italiani.

La dashboard utilizza pertanto le stime precedentemente ottenute per i CAP di residenza dei pazienti e quelle relative ai capoluoghi di provincia.

Oltre alla variabile `cap`, si aggiunge la variabile `focus`, che permette di modificare il livello geografico dell'analisi. In particolare, è possibile concentrarsi sul contesto provinciale, considerando le stime disponibili all'interno della provincia associata al CAP selezionato, oppure sul contesto regionale, estendendo l'analisi alle stime disponibili nell'intera regione del CAP selezionato. Queste variabili si aggiungono a quelle già descritte nelle dashboard precedenti, ovvero `pollutants`, `patient` e `date_relapse`.

Il primo pannello proposto in questa dashboard è una heatmap che rappresenta la distribuzione delle concentrazioni stimate nei diversi CAP disponibili. Sebbene la rappresentazione sia limitata alle località per le quali sono state effettuate le stime, essa permette di evidenziare alcune tendenze territoriali. In particolare, come si può osservare nella Figura 4.19, la Pianura padana rappresenta, in generale, l'area avente livelli di inquinamento maggiore.



Figura 4.19: Mappa raffigurante i valori di AQI associati ai diversi CAP italiani.

Ai lati della mappa sono riportate informazioni relative al CAP selezionato, insieme all'andamento dell'AQI nei sette giorni antecedenti e successivi alla data scelta e alla distribuzione dei livelli di AQI osservati nel contesto geografico selezionato. Attraverso la variabile focus è pertanto possibile passare dalla visualizzazione riferita al contesto provinciale a quella relativa al contesto regionale.

Le ulteriori visualizzazioni mantengono una struttura analoga a quelle già introdotte nella dashboard del paziente. In particolare, vengono riportati gli andamenti mensili e stagionali dell'AQI. La possibilità di confrontare i valori provinciali con quelli regionali consente di valutare se la zona analizzata presenti livelli di qualità dell'aria inferiori o superiori rispetto al contesto regionale.

Ad esempio, nel caso in Figura 4.20, la provincia di Lodi ha valori di AQI che risultano tendenzialmente inferiori rispetto alla media regionale. Quest'ultima risulta invece influenzata dai valori più elevati misurati nelle aree maggiormente urbanizzate, tra tutte Milano e provincia.

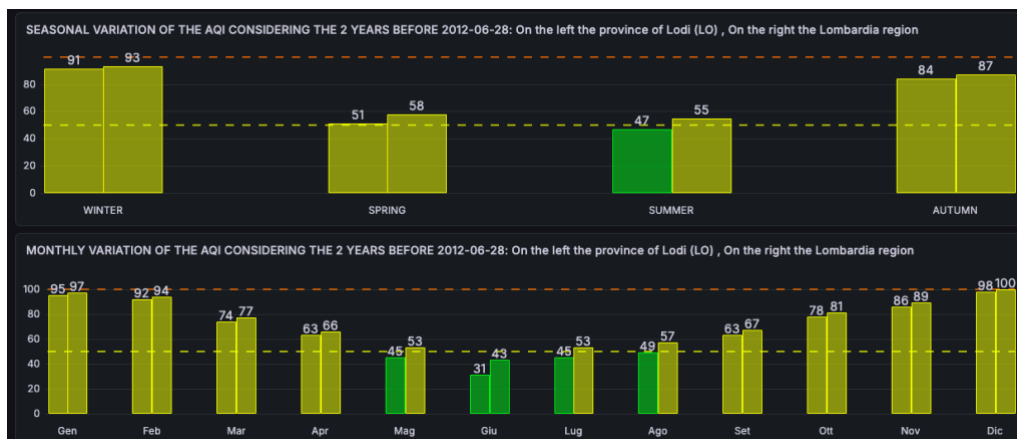


Figura 4.20: Andamento stagionale e mensile dell'AQI misurato a livello provinciale (a sinistra) e regionale (a destra) nei due anni precedenti alla date_relapse selezionata.

Infine, analogamente alla precedente dashboard, sono presenti specifiche sezioni dedicate ai singoli inquinanti. Per ognuno di essi vengono mostrati gli andamenti temporali e le relative statistiche riferite al contesto provinciale e regionale e, se presente la sua stima, a livello comunale.

Questo permette di analizzare più nel dettaglio le differenze territoriali nella concentrazione dell'inquinante scelto nell'apposito menù a tendina.

Nella Figura 4.21 è mostrato l'andamento del PM_{2.5} a Pavia, nella provincia di Pavia e in Lombardia. I relativi valori minimi, massimi e medi sono riportati nella Figura 4.22.



Figura 4.21: Andamento temporale della concentrazione media giornaliera di PM_{2.5} stimata nel comune di Pavia, nella relativa provincia e a livello regionale (Lombardia).

Name	Max ^	Min	Mean
 AQG VALUE	15.00 µg/m³	15.00 µg/m³	15.00 µg/m³
 TARGET 4	25.00 µg/m³	25.00 µg/m³	25.00 µg/m³
 TARGET 3	37.50 µg/m³	37.50 µg/m³	37.50 µg/m³
 Provincial mean of Pavia	41.04 µg/m³	0.56 µg/m³	17.67 µg/m³
 Pavia	41.50 µg/m³	0.19 µg/m³	18.45 µg/m³
 Regional mean	42.54 µg/m³	4.31 µg/m³	21.38 µg/m³
 TARGET 2	50.00 µg/m³	50.00 µg/m³	50.00 µg/m³
 TARGET 1	75.00 µg/m³	75.00 µg/m³	75.00 µg/m³

Figura 4.22: Valori minimi, massimi e medi di PM_{2.5} stimati nel comune di Pavia, nella provincia di Pavia e in Regione Lombardia.

In generale, questa dashboard consente quindi di ampliare l'analisi della sola area di residenza del paziente ad un contesto geografico più esteso, fornendo un termine di paragone utile per interpretare i livelli di inquinamento locale rispetto a quelli misurati su larga scala.

5 CONCLUSIONI E PROSPETTIVE FUTURE

Il presente lavoro ha avuto come obiettivo lo sviluppo di un'interfaccia grafica interattiva volta a supportare il personale medico nella gestione del paziente affetto da SM, fornendo informazioni relative alla qualità dell'aria nelle diverse zone di esposizione del paziente, con particolare attenzione ai periodi in cui si verificano episodi di ricaduta. A tal fine, sono stati adottati due modelli di stima spazio-temporale dei principali inquinanti atmosferici, attraverso i quali è stato possibile ottenere una stima della concentrazione media giornaliera di ciascun inquinante e, di conseguenza, della qualità dell'aria anche in aree geografiche prive di stazioni di monitoraggio nelle immediate vicinanze.

I valori stimati sono stati integrati nelle dashboard opportunamente sviluppate in Grafana, insieme alle informazioni anagrafiche e cliniche associate a ciascun paziente. Lo strumento realizzato consente pertanto di fornire al personale medico una rappresentazione immediata e integrata del quadro clinico e ambientale del paziente, in modo da poter monitorare in modo sistematico condizioni di qualità dell'aria in concomitanza dei periodi di insorgenza delle ricadute. Tale approccio può fornire un supporto aggiuntivo alla valutazione dell'esposizione ambientale e, nei periodi caratterizzati da condizioni di maggiore inquinamento, contribuire alla formulazione di raccomandazioni volte a limitare l'esposizione all'aria aperta, laddove ritenuto opportuno dal medico. La realizzazione di questa piattaforma ha permesso di integrare in un'unica interfaccia informazioni ambientali, anagrafiche e cliniche, rendendo possibile l'analisi delle condizioni di qualità dell'aria nelle diverse aree di esposizione del paziente e il loro confronto con i periodi di insorgenza delle ricadute.

Il progetto presenta margini di sviluppo, sia per quanto riguarda la componente back-end sia per quella front-end. Per quanto concerne i modelli di stima, un possibile sviluppo futuro consiste nell'ottimizzazione delle prestazioni attraverso l'incremento del numero e della risoluzione dei regressori utilizzati, con particolare riferimento alle variabili meteorologiche. In particolare, l'estensione dell'insieme di variabili atmosferiche considerate e la transizione da una risoluzione temporale giornaliera a una oraria potrebbe consentire di aumentare la sensibilità dei modelli rispetto alle oscillazioni ambientali,

permettendo di ottenere stime delle concentrazioni degli inquinanti su base oraria e identificando con maggiore accuratezza le fasce della giornata caratterizzate dalle maggiori criticità e permetterebbe di includere nel calcolo dell'AQI anche l'O₃, attualmente escluso poiché la sua valutazione richiede valori medi orari anziché giornalieri.

Anche gli altri regressori utilizzati nei modelli potrebbero essere soggetti a variazioni e aggiornamenti nel tempo. È opportuno effettuare un aggiornamento periodico dei regressori statici, in particolare per quanto riguarda il dataset CLC, attualmente fermo al 2018 e per il quale ci si aspetta una versione più recente nel prossimo futuro.

La disponibilità di serie storiche e di variabili meteorologiche potrebbe inoltre essere sfruttata per sviluppare modelli in grado di effettuare una previsione futura delle concentrazioni degli inquinanti. In questo modo, sarebbe possibile fornire una stima delle condizioni di qualità dell'aria previste nei giorni o nelle ore successive. L'integrazione di tale funzionalità potrebbe portare, in prospettiva, alla realizzazione di un sistema in grado di generare avvisi relativi alle condizioni di qualità dell'aria e al potenziale rischio associato all'esposizione all'aperto, fornendo al medico un ulteriore elemento di supporto nella valutazione delle condizioni ambientali

Un ulteriore ambito di sviluppo riguarda l'analisi delle possibili associazioni tra le concentrazioni degli inquinanti atmosferici e parametri clinici quali l'EDSS e l'insorgenza delle ricadute, al fine di individuare eventuali correlazioni tra esposizione ambientale e andamento della malattia. Un'analisi condotta su un campione più ampio e attraverso metodologie statistiche appropriate potrebbe quindi contribuire a valutare il possibile ruolo dell'inquinamento atmosferico come fattore associato alla progressione della sclerosi multipla e all'insorgenza delle ricadute, pur senza consentire, di per sé, di stabilire un rapporto causale.

Per quanto riguarda la componente front-end, l'utilizzo di Grafana ha permesso di realizzare un'interfaccia interattiva e funzionale, ma ha evidenziato alcune limitazioni della piattaforma. Tra queste rientrano l'assenza di una variabile di tipo calendario, la difficoltà nell'aggiornamento dinamico del time range in funzione del valore assunto da una variabile e alcune limitazioni nella

personalizzazione grafica di determinati elementi dell'interfaccia, tra cui le variables. Tali limitazioni non compromettono il funzionamento, ma possono ostacolare una successiva evoluzione e personalizzazione in funzione delle esigenze specifiche del progetto.

Le dashboard potrebbero essere ulteriormente arricchite attraverso la visualizzazione delle eventuali correlazioni individuate per il singolo paziente e mediante l'integrazione dei risultati dei modelli predittivi.

In conclusione, il lavoro realizzato costituisce quindi una prima implementazione di uno strumento integrato per la visualizzazione delle informazioni cliniche e ambientali associate al paziente con sclerosi multipla. Gli sviluppi futuri proposti potrebbero consentire di trasformare la dashboard da uno strumento prevalentemente descrittivo e di consultazione a un sistema predittivo e di supporto alle decisioni, capace di integrare informazioni ambientali, cliniche e previsionali e di fornire al medico un quadro sempre più completo delle possibili condizioni di esposizione del paziente.

APPENDICE A- Confronto dei modelli per l'imputazione dei dati in zone con bassa densità di stazioni

A.1 Confronto modelli per il SO₂

Tabella A.1: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il SO₂

METODO	GLOBALE	URBAN
EWMA	0.4479	0.4479
IDW_vicino	1.9291	1.9291
LOCF	0.4566	0.4566
LUR_ibrido	2.3235	2.3235
LUR_puro	2.3235	2.3235
LUR_vicino	1.83	1.83
Linear	0.3543	0.3543

A.2 Confronto modelli per il PM₁₀

Tabella A.2: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il PM₁₀

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	13.0108	13.1962	12.5807	13.0379
IDW_vicino	9.8778	9.2245	8.5794	12.9343
LOCF	12.2359	12.4757	11.9607	11.8871
LUR_ibrido	16.4514	16.428	16.0738	17.0104
LUR_puro	16.4514	16.428	16.0738	17.0104
LUR_vicino	9.9372	9.4518	9.4383	11.8026
Linear	9.3619	9.629	9.1061	8.8943

A.3 Confronto modelli per l'O₃

Tabella A.3: Confronto nell'imputazione di dati in termini di RMSE dei modelli per l'O₃

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	11.229	11.3345	11.4657	10.8753

IDW_vicino	19.847	18.5251	18.9609	22.0931
LOCF	11.0628	11.184	11.3746	10.6157
LUR_ibrido	20.1739	20.3726	18.5865	21.2261
LUR_puro	20.1739	20.3726	18.5865	21.2261
LUR_vicino	18.1859	16.2237	20.3086	18.4431
Linear	8.4393	8.5549	8.7156	8.0305

A.4 Confronto modelli per il NO₂

Tabella A.4: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il NO₂

METODO	GLOBALE	RURAL
EWMA	1.3334	1.3334
IDW_vicino	11.6916	11.6916
LOCF	1.4054	1.4054
LUR_ibrido	12.6065	12.6065
LUR_puro	12.6065	12.6065
LUR_vicino	4.0791	4.0791
Linear	1.099	1.099

A.5 Confronto modelli per il CO

Tabella A.5: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il CO

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	0.1274	0.0995	0.1349	0.1469
IDW_vicino	0.2933	0.2675	0.2688	0.3457
LOCF	0.1181	0.0863	0.131	0.1346
LUR_ibrido	0.3833	0.2652	0.3075	0.5787
LUR_puro	0.3833	0.2652	0.3075	0.5787
LUR_vicino	0.2915	0.2792	0.2662	0.3402
Linear	0.102	0.0636	0.1045	0.1323

A.6 Confronto modelli per il C₆H₆

Tabella A.6: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il C₆H₆

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	0.4417	0.4889	0.2149	0.3117
IDW_vicino	0.9686	0.9934	1.0367	0.6137
LOCF	0.43	0.4743	0.2191	0.3092
LUR_ibrido	0.7787	0.8274	0.5429	0.7027
LUR_puro	0.7787	0.8274	0.5429	0.7027
LUR_vicino	0.9779	1.0055	0.955	0.768
Linear	0.3253	0.3593	0.1644	0.2305

A.7 Confronto modelli per il PM_{2.5}

Tabella A.7: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il PM_{2.5}

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	7.4202	7.7872	6.9013	6.8057
IDW_vicino	6.9639	6.898	5.6529	8.4581
LOCF	6.7196	7.022	6.2577	6.2606
LUR_ibrido	10.8149	11.1789	10.5478	9.8986
LUR_puro	10.8149	11.1789	10.5478	9.8986
LUR_vicino	7.4569	7.6626	6.8198	7.4937
Linear	5.0615	5.2922	4.6929	4.7308

APPENDICE B- Confronto dei modelli per l'imputazione dei dati in zone con alta densità di stazioni

B.1 Confronto modelli per il SO₂

Tabella B.1: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il SO₂

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	1.2782	1.005	1.5774	1.2484
IDW	2.3728	1.979	2.8496	2.2617
IDW_vicino	2.3728	1.979	2.8496	2.2617
LOCF	1.2638	0.9947	1.5558	1.2429
LUR_ibrido	2.4144	1.9976	2.9587	2.3212
LUR_puro	2.4828	2.1731	3.0112	2.1493
LUR_vicino	2.4144	1.9976	2.9587	2.3212
Linear	0.9978	0.7926	1.2252	0.9725

B.2 Confronto modelli per il PM₁₀

Tabella B.2: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il PM₁₀

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	13.7438	13.7556	14.1085	12.5689
IDW	7.3559	6.9214	7.804	8.3573
IDW_vicino	7.3559	6.9214	7.804	8.3573
LOCF	13.3249	13.0419	14.0386	12.8057
LUR_ibrido	7.9475	7.6619	8.3647	8.3945
LUR_puro	17.6589	17.3866	18.6989	16.4007
LUR_vicino	7.9475	7.6619	8.3647	8.3945
Linear	10.3799	10.0603	11.0514	10.1781

B.3 Confronto modelli per l'O₃

Tabella B.3: Confronto nell'imputazione di dati in termini di RMSE dei modelli per l'O₃

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	11.339	11.5272	11.262	11.0834
IDW	12.006	11.6789	10.5946	14.3778
IDW_vicino	12.006	11.6789	10.5946	14.3778
LOCF	11.1844	11.4265	11.0811	10.8598
LUR_ibrido	14.8067	14.409	14.9487	15.3366
LUR_puro	21.7803	22.9888	21.9573	18.981
LUR_vicino	14.8067	14.409	14.9487	15.3366
Linear	8.5467	8.7598	8.4399	8.2833

B.4 Confronto modelli per il NO₂

Tabella B.4: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il NO₂

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	5.7839	6.5399	5.333	3.0953
IDW	9.4072	9.3593	8.2952	11.237
IDW_vicino	9.4072	9.3593	8.2952	11.237
LOCF	5.8973	6.6703	5.4547	3.0795
LUR_ibrido	7.8117	8.5141	7.4435	5.5389
LUR_puro	10.0867	11.2815	9.2077	6.4253
LUR_vicino	7.8117	8.5141	7.4435	5.5389
Linear	4.6045	5.2098	4.2613	2.3896

B.5 Confronto modelli per il CO

Tabella B.5: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il CO

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	0.1252	0.1377	0.1034	0.0765
IDW	0.2378	0.2523	0.2155	0.1826

IDW_vicino	0.2378	0.2523	0.2155	0.1826
LOCF	0.1119	0.1234	0.0911	0.0693
LUR_ibrido	0.2425	0.2603	0.2125	0.1666
LUR_puro	0.2824	0.2993	0.2572	0.2029
LUR_vicino	0.2425	0.2603	0.2125	0.1666
Linear	0.0838	0.0924	0.0675	0.054

B.5 Confronto modelli per il C₆H₆

Tabella B.6: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il C₆H₆

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	0.5143	0.5344	0.519	0.1632
IDW	0.7154	0.7548	0.67	0.4201
IDW_vicino	0.7154	0.7548	0.67	0.4201
LOCF	0.5175	0.5299	0.5417	0.1609
LUR_ibrido	0.6993	0.7527	0.605	0.3562
LUR_puro	0.8171	0.8405	0.8301	0.4296
LUR_vicino	0.6993	0.7527	0.605	0.3562
Linear	0.4019	0.4077	0.4299	0.1241

B.7 Confronto modelli per il PM_{2.5}

Tabella B.7: Confronto nell'imputazione di dati in termini di RMSE dei modelli per il PM_{2.5}

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
EWMA	7.5942	7.9181	7.1259	7.3817
IDW	4.7114	4.662	4.967	4.0342
IDW_vicino	4.7114	4.662	4.967	4.0342
LOCF	6.9717	7.1313	6.6534	7.1696
LUR_ibrido	5.4494	5.2492	6.0615	4.5329
LUR_puro	11.2355	11.2448	11.5276	10.3235
LUR_vicino	5.4494	5.2492	6.0615	4.5329
Linear	5.326	5.424	5.0947	5.5645

APPENDICE C- Confronto dei modelli in comuni privi di stazioni all'interno di una zona geografica con bassa densità di stazioni

C.1 Confronto modelli per il SO₂

Tabella C.1: Confronto modelli in termini di RMSE per il SO₂

METODO	GLOBALE	URBAN
IDW_vicino	1.9291	1.9291
LUR_ibrido	2.3235	2.3235
LUR_puro	2.3235	2.3235
LUR_vicino	1.83	1.83
Media_3_Vicini	2.241	2.241
Media_5_Vicini	1.9867	1.9867
Mediana_3_Vicini	1.7407	1.7407
Mediana_5_Vicini	1.2994	1.2994
Stazione_Più_Vicina	1.9291	1.9291

C.2 Confronto modelli per il PM₁₀

Tabella C.2: Confronto modelli in termini di RMSE per il PM₁₀

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW_vicino	9.8778	9.2245	8.5794	12.9343
LUR_ibrido	16.4514	16.428	16.0738	17.0104
LUR_puro	16.4514	16.428	16.0738	17.0104
LUR_vicino	9.9372	9.4518	9.4383	11.8026
Media_3_Vicini	8.7674	8.0947	7.4533	11.8156
Media_5_Vicini	8.6565	7.8208	7.8168	11.6019
Mediana_3_Vicini	8.9188	8.2206	7.7864	11.8679
Mediana_5_Vicini	8.8649	8.0654	8.0642	11.7125
Stazione_Più_Vicina	9.8778	9.2245	8.5794	12.9343

C.3 Confronto modelli per l'O₃

Tabella C.3: Confronto modelli in termini di RMSE per l'O₃

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW_vicino	19.847	18.5251	18.9609	22.0931
LUR_ibrido	20.1739	20.3726	18.5865	21.2261
LUR_puro	20.1739	20.3726	18.5865	21.2261
LUR_vicino	18.1859	16.2237	20.3086	18.4431
Media_3_Vicini	16.5329	13.7105	15.6046	20.0961
Media_5_Vicini	15.7752	12.7962	14.1839	19.887
Mediana_3_Vicini	17.0437	14.4857	15.6636	20.722
Mediana_5_Vicini	16.2997	13.3085	14.2947	20.7134
Stazione_Più_Vicina	19.8254	18.5251	18.9609	22.0026

C.4 Confronto modelli per il NO₂

Tabella C.4: Confronto modelli in termini di RMSE per il NO₂

METODO	GLOBALE	RURAL
IDW_vicino	11.6916	11.6916
LUR_ibrido	12.6065	12.6065
LUR_puro	12.6065	12.6065
LUR_vicino	4.0791	4.0791
Media_3_Vicini	11.1932	11.1932
Media_5_Vicini	9.4098	9.4098
Mediana_3_Vicini	10.049	10.049
Mediana_5_Vicini	8.2313	8.2313
Stazione_Più_Vicina	9.7431	9.7431

C.5 Confronto modelli per il CO

Tabella C.5: Confronto modelli in termini di RMSE per il CO

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW_vicino	0.2933	0.2675	0.2688	0.3457

LUR_ibrido	0.3833	0.2652	0.3075	0.5787
LUR_puro	0.3833	0.2652	0.3075	0.5787
LUR_vicino	0.2915	0.2792	0.2662	0.3402
Media_3_Vicini	0.3452	0.208	0.3282	0.4737
Media_5_Vicini	0.3666	0.2248	0.35	0.4997
Mediana_3_Vicini	0.3707	0.2376	0.3387	0.5118
Mediana_5_Vicini	0.3867	0.245	0.364	0.527
Stazione_Più_Vicina	0.2933	0.2675	0.2688	0.3457

C.6 Confronto modelli per il C₆H₆

Tabella C.6: Confronto modelli in termini di RMSE per il C₆H₆

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW_vicino	0.9686	0.9934	1.0367	0.6137
LUR_ibrido	0.7787	0.8274	0.5429	0.7027
LUR_puro	0.7787	0.8274	0.5429	0.7027
LUR_vicino	0.9779	1.0055	0.955	0.768
Media_3_Vicini	0.8003	0.7857	1.0112	0.4863
Media_5_Vicini	0.8123	0.8491	0.8089	0.4687
Mediana_3_Vicini	0.8105	0.814	0.9544	0.4835
Mediana_5_Vicini	0.8097	0.8602	0.7239	0.4852
Stazione_Più_Vicina	0.9686	0.9934	1.0367	0.6137

C.7 Confronto modelli per il PM_{2.5}

Tabella C.7: Confronto modelli in termini di RMSE per il PM_{2.5}

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW_vicino	6.9639	6.898	5.6529	8.4581
LUR_ibrido	10.8149	11.1789	10.5478	9.8986
LUR_puro	10.8149	11.1789	10.5478	9.8986
LUR_vicino	7.4569	7.6626	6.8198	7.4937
Media_3_Vicini	6.2218	5.9544	5.2676	7.9017

Media_5_Vicini	6.0806	5.8321	5.3398	7.5304
Mediana_3_Vicini	6.4251	6.2811	5.4643	7.8114
Mediana_5_Vicini	6.2902	6.1129	5.5394	7.5757
Stazione_Piu_Vicina	6.9639	6.898	5.6529	8.4581

APPENDICE D- Confronto dei modelli in comuni privi di stazioni all'interno di una zona geografica con alta densità di stazioni

D.1 Confronto modelli per il SO₂

Tabella D.1: Confronto modelli in termini di RMSE per il SO₂

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW	2.3728	1.979	2.8496	2.2617
IDW_vicino	2.3728	1.979	2.8496	2.2617
LUR_ibrido	2.4144	1.9976	2.9587	2.3212
LUR_puro	2.4828	2.1731	3.0112	2.1493
LUR_vicino	2.4144	1.9976	2.9587	2.3212
Media_3_Vicini	2.4541	2.1517	2.8765	2.2608
Media_5_Vicini	2.3228	1.996	2.7661	2.128
Mediana_3_Vicini	2.3947	2.078	2.8213	2.2242
Mediana_5_Vicini	2.3458	1.994	2.8072	2.1678
Stazione_Piu_Vicina	2.786	2.3975	3.1553	2.9552

D.2 Confronto modelli per il PM₁₀

Tabella D.2: Confronto modelli in termini di RMSE per il PM₁₀

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW	7.3559	6.9214	7.804	8.3573
IDW_vicino	7.3559	6.9214	7.804	8.3573
LUR_ibrido	7.9475	7.6619	8.3647	8.3945
LUR_puro	17.6589	17.3866	18.6989	16.4007
LUR_vicino	7.9475	7.6619	8.3647	8.3945
Media_3_Vicini	7.2515	6.8397	7.6299	8.3293
Media_5_Vicini	7.215	6.7233	7.6	8.6367
Mediana_3_Vicini	7.4062	7.0034	7.8764	8.197
Mediana_5_Vicini	7.3246	6.8192	7.7889	8.6084

Stazione_Piu_Vicina	8.3694	7.8222	9.1238	9.0834
---------------------	--------	--------	--------	--------

D.3 Confronto modelli per l'O₃

Tabella D.3: Confronto modelli in termini di RMSE per l'O₃

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW	12.006	11.6789	10.5946	14.3778
IDW_vicino	12.006	11.6789	10.5946	14.3778
LUR_ibrido	14.8067	14.409	14.9487	15.3366
LUR_puro	21.7803	22.9888	21.9573	18.981
LUR_vicino	14.8067	14.409	14.9487	15.3366
Media_3_Vicini	12.4041	11.9213	10.8093	15.2255
Media_5_Vicini	12.1373	11.4483	10.6465	15.1473
Mediana_3_Vicini	12.5358	11.328	11.4397	15.8788
Mediana_5_Vicini	12.3283	11.1929	11.0824	15.7222
Stazione_Piu_Vicina	14.6855	14.5809	12.752	17.3416

D.4 Confronto modelli per il NO₂

Tabella D.4: Confronto modelli in termini di RMSE per il NO₂

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW	9.4072	9.3593	8.2952	11.237
IDW_vicino	9.4072	9.3593	8.2952	11.237
LUR_ibrido	7.8117	8.5141	7.4435	5.5389
LUR_puro	10.0867	11.2815	9.2077	6.4253
LUR_vicino	7.8117	8.5141	7.4435	5.5389
Media_3_Vicini	9.835	9.7306	9.1982	11.1897
Media_5_Vicini	9.5116	9.4814	8.6449	10.9617
Mediana_3_Vicini	9.9999	10.0456	9.1953	11.133
Mediana_5_Vicini	9.6983	9.8845	8.5774	10.8304
Stazione_Piu_Vicina	11.3493	11.2318	10.4625	13.1203

D.5 Confronto modelli per il CO

Tabella D.5: Confronto modelli in termini di RMSE per il CO

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW	0.2378	0.2523	0.2155	0.1826
IDW_vicino	0.2378	0.2523	0.2155	0.1826
LUR_ibrido	0.2425	0.2603	0.2125	0.1666
LUR_puro	0.2824	0.2993	0.2572	0.2029
LUR_vicino	0.2425	0.2603	0.2125	0.1666
Media_3_Vicini	0.2401	0.2559	0.2131	0.1876
Media_5_Vicini	0.2333	0.2493	0.2075	0.1743
Mediana_3_Vicini	0.252	0.2682	0.226	0.1928
Mediana_5_Vicini	0.2398	0.2572	0.2116	0.1744
Stazione_Piu_Vicina	0.284	0.2947	0.2757	0.218

D.6 Confronto modelli per il C₆H₆

Tabella D.6: Confronto modelli in termini di RMSE per il C₆H₆

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW	0.7154	0.7548	0.67	0.4201
IDW_vicino	0.7154	0.7548	0.67	0.4201
LUR_ibrido	0.6993	0.7527	0.605	0.3562
LUR_puro	0.8171	0.8405	0.8301	0.4296
LUR_vicino	0.6993	0.7527	0.605	0.3562
Media_3_Vicini	0.7079	0.7408	0.6637	0.5112
Media_5_Vicini	0.7091	0.7372	0.6747	0.5318
Mediana_3_Vicini	0.7108	0.7509	0.6512	0.4882
Mediana_5_Vicini	0.7203	0.7594	0.6642	0.4944
Stazione_Piu_Vicina	0.8467	0.8895	0.8073	0.4697

D.7 Confronto modelli per il PM_{2.5}

Tabella D.7: Confronto modelli in termini di RMSE per il PM_{2.5}

METODO	GLOBALE	URBAN	SUBURBAN	RURAL
IDW	4.7114	4.662	4.967	4.0342
IDW_vicino	4.7114	4.662	4.967	4.0342
LUR_ibrido	5.4494	5.2492	6.0615	4.5329
LUR_puro	11.2355	11.2448	11.5276	10.3235
LUR_vicino	5.4494	5.2492	6.0615	4.5329
Media_3_Vicini	4.7316	4.8202	4.696	4.3665
Media_5_Vicini	4.9239	4.9159	4.9692	4.8112
Mediana_3_Vicini	4.8801	4.8919	4.9664	4.5099
Mediana_5_Vicini	4.9931	4.9997	5.0141	4.8865
Stazione_Piu_Vicina	5.3376	5.1115	5.739	5.0907

BIBLIOGRAFIA

- [1] «Global status report on neurology». World Health Organisation, 12 ottobre 2025. [Online]. Disponibile su: <https://www.who.int/publications/i/item/9789240116139>
- [2] «Home», Brainteaser. Consultato: 1 agosto 2026. [Online]. Disponibile su: <https://brainteaser.health/>
- [3] K. Kapica-Topczewska, A. Kułakowska, J. Kochanowicz, e W. Broła, «Epidemiology of multiple sclerosis: global trends, regional differences, and clinical implications», *Neurologia i Neurochirurgia Polska*, vol. 59, fasc. 4, pp. 375–384, 2025, doi: 10.5603/pjnns.103955.
- [4] «Barometro 2026 ESTRATTO WEB». Associazione Italiana Sclerosi Multipla, maggio 2026. [Online]. Disponibile su: <https://www.aism.it/sites/default/files/2026-05/Barometro%202026%20ESTRATTO%20WEB.pdf>
- [5] D. Tafti, M. Ehsan, e K. L. Xixis, «Multiple Sclerosis», in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2026. Consultato: 29 luglio 2026. [Online]. Disponibile su: <http://www.ncbi.nlm.nih.gov/books/NBK499849/>
- [6] M. A. Schilling, «The geographic association of multiple sclerosis and amyotrophic lateral sclerosis», *Sci Rep*, vol. 15, p. 34665, ott. 2025, doi: 10.1038/s41598-025-18755-8.
- [7] J. F. Kurtzke, «Geography in multiple sclerosis», *J. Neurol.*, vol. 215, fasc. 1, pp. 1–26, mar. 1977, doi: 10.1007/BF00312546.
- [8] B. J. Wade, «Spatial Analysis of Global Prevalence of Multiple Sclerosis Suggests Need for an Updated Prevalence Scale», *Mult Scler Int*, vol. 2014, p. 124578, 2014, doi: 10.1155/2014/124578.
- [9] M. J. Mohammadi, K. Zarea, N. Hatamzadeh, A. Salahshouri, e A. Sharhani, «Toxic Air Pollutants and Their Effect on Multiple Sclerosis: A Review Study», *Front Public Health*, vol. 10, p. 898043, lug. 2022, doi: 10.3389/fpubh.2022.898043.
- [10] S. Şen, «Neurostatus and EDSS Calculation with Cases», *Noro Psikiyatr Ars*, vol. 55, fasc. Suppl 1, pp. S80–S83, 2018, doi: 10.29399/npa.23412.
- [11] «Sclerosi multipla | Società svizzera SM». Consultato: 3 settembre 2026. [Online]. Disponibile su: <https://www.multiplesklerose.ch/it/sulla-sm/sclerosi-multipla/>
- [12] IQAir, «2025_IQAir_World_Air_Quality_Report_3.26.2026». IQAir, 2025. [Online]. Disponibile su: <https://www.iqair.com/it/world-air-quality-report>
- [13] IQAir, «Rapporto mondiale sulla qualità dell'aria | IQAir», IQAir. Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://www.iqair.com/it/world-air-quality-report>
- [14] «World Air Quality Report | IQAir USA». Consultato: 3 settembre 2026. [Online]. Disponibile su: <https://www.iqair.com/world-air-quality-report>
- [15] B. Rackow, H.-H. König, M. Wall, e C. Konnopka, «The interaction between air pollution, weather conditions, and health risks: a systematic

- review», *Science of The Total Environment*, vol. 996, p. 180080, set. 2025, doi: 10.1016/j.scitotenv.2025.180080.
- [16] «WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide». Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://www.who.int/publications/i/item/9789240034228>
- [17] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, e E. Bezirtzoglou, «Environmental and Health Impacts of Air Pollution: A Review», *Front Public Health*, vol. 8, p. 14, feb. 2020, doi: 10.3389/fpubh.2020.00014.
- [18] «QUALITÀ DELL'ARIA AMBIENTE: PARTICOLATO (PM_{2.5}) | Indicatori ambientali». Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://indicatoriambientali.isprambiente.it/en/air-quality/ambient-air-quality-particulate-pm2.5>
- [19] «QUALITÀ DELL'ARIA AMBIENTE: PARTICOLATO (PM₁₀) | Indicatori ambientali». Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://indicatoriambientali.isprambiente.it/it/qualita-dellaria/qualita-dell-aria-ambiente-particolato-pm10>
- [20] «QUALITÀ DELL'ARIA AMBIENTE: OZONO TROPOSFERICO (O₃) | Indicatori ambientali». Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://indicatoriambientali.isprambiente.it/it/qualita-dellaria/qualita-dellaria-ambiente-ozono-troposferico-o3>
- [21] «QUALITÀ DELL'ARIA AMBIENTE: BIOSSIDO DI AZOTO (NO₂) | Indicatori ambientali». Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://indicatoriambientali.isprambiente.it/it/qualita-dellaria/qualita-dellaria-ambiente-biossido-di-azoto-no2>
- [22] «Anidride solforosa | IQAir Italia». Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://www.iqair.com/it/newsroom/sulphur-dioxide>
- [23] «Monossido di carbonio - CO - Arpa Sicilia», Liferay. Consultato: 30 luglio 2026. [Online]. Disponibile su: http://qualitadellaria.arpa.sicilia.it:8080/web/guest/inquinante_6?p_p_auth=1yy9cyOg&p_p_id=49&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&_49_struts_action=%2Fmy_sites%2Fview&_49_groupId=20182&_49_privateLayout=false/inquinante_6
- [24] «Benzene - Arpa Sicilia», Liferay. Consultato: 30 luglio 2026. [Online]. Disponibile su: http://qualitadellaria.arpa.sicilia.it:8080/web/guest/inquinante_62?p_p_auth=OSp3XydY&p_p_id=49&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&_49_struts_action=%2Fmy_sites%2Fview&_49_groupId=20182&_49_privateLayout=false/inquinante_62
- [25] «Benzene | Arpa Piemonte». Consultato: 30 luglio 2026. [Online]. Disponibile su: <https://www.arpa.piemonte.it/scheda-informativa/benzene>
- [26] «Standards for Air Quality Indices in Different Countries (AQI)». Consultato: 31 luglio 2026. [Online]. Disponibile su: <https://atmotube.com/blog/standards-for-air-quality-indices-in-different-countries-aqi>
- [27] «Live Air Quality Map | IQAir USA». Consultato: 31 luglio 2026. [Online]. Disponibile su: <https://www.iqair.com/air-quality-map>

- [28] «European Air Quality Index». Consultato: 31 luglio 2026. [Online]. Disponibile su: <https://airindex.eea.europa.eu/AQI/index.html>
- [29] «ETC HE Report 2024/17: EEA's revision of the European air quality index bands», Eionet Portal. Consultato: 31 luglio 2026. [Online]. Disponibile su: <https://www.eionet.europa.eu/etcs/etc-he/products/etc-he-products/etc-he-reports/etc-he-report-2024-17-eeas-revision-of-the-european-air-quality-index-bands>
- [30] «AQI Technical Assistance Document May 2026». Consultato: 31 luglio 2026. [Online]. Disponibile su: <https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-quality.pdf>
- [31] A. A. Toubasi e T. N. Al-Sayegh, «Air Pollution and the Risk and Progression of Multiple Sclerosis: A Systematic Review and Meta-Analysis», *Annals of Clinical and Translational Neurology*, vol. 13, fasc. 7, pp. 1443–1451, 2026, doi: 10.1002/acn3.70330.
- [32] «Download Service». Consultato: 18 agosto 2026. [Online]. Disponibile su: <https://eeadmz1-downloads-webapp.azurewebsites.net/>
- [33] R. C. Cornes, G. Van Der Schrier, E. J. M. Van Den Besselaar, e P. D. Jones, «An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets», *JGR Atmospheres*, vol. 123, fasc. 17, pp. 9391–9409, set. 2018, doi: 10.1029/2017JD028200.
- [34] «Table publisher». Consultato: 18 agosto 2026. [Online]. Disponibile su: https://discomap.eea.europa.eu/App/AQViewer/index.html?fqn=Airquality_Dissem.b2g.measurements
- [35] «DiscoMap API - Swagger UI». Consultato: 4 settembre 2026. [Online]. Disponibile su: http://localhost:8000/docs#/F0%9F%94%A7%20Database%20Optimization/compress_old_chunks_api_v1_compression_compress_post
- [36] «Database dei comuni italiani gratuito, completo e sempre aggiornato», Garda Informatica. Consultato: 19 agosto 2026. [Online]. Disponibile su: <https://www.gardainformatica.it/database-comuni-italiani>
- [37] K. de Hoogh *et al.*, «Spatial PM_{2.5}, NO₂, O₃ and BC models for Western Europe – Evaluation of spatiotemporal stability», *Environment International*, vol. 120, pp. 81–92, nov. 2018, doi: 10.1016/j.envint.2018.07.036.
- [38] S. Dharmalingam *et al.*, «Developing air pollution concentration fields for health studies using multiple methods: Cross-comparison and evaluation», *Environmental Research*, vol. 207, p. 112207, mag. 2022, doi: 10.1016/j.envres.2021.112207.
- [39] «Popolazione residente». Consultato: 19 agosto 2026. [Online]. Disponibile su: <https://demo.istat.it/app/?i=POS&l=it>
- [40] L. V. Rueda-Torres, J. Warthon-Ascarza, e S. Pacsi-Valdivia, «Adjustment Criteria for Air-Quality Standards by Altitude: A Scoping Review with Regulatory Overview», *Int J Environ Res Public Health*, vol. 22, fasc. 7, p. 1053, giu. 2025, doi: 10.3390/ijerph22071053.
- [41] «Principali statistiche geografiche sui comuni». Consultato: 19 agosto 2026. [Online]. Disponibile su:

<https://www.istat.it/classificazione/principali-statistiche-geografiche-sui-comuni/>

- [42] «Codici delle unità amministrative». Consultato: 19 agosto 2026. [Online]. Disponibile su: <https://www.istat.it/classificazione/codici-dei-comuni-delle-province-e-delle-regioni/>
- [43] «lista comune italia istat - Cerca con Google». Consultato: 19 agosto 2026. [Online]. Disponibile su: https://www.google.com/search?q=lista+comune+italia+istat&sca_esv=cf5c3a640caff83a&rlz=1C5CHFA_enIT894IT894&sxsrf=APpeQnsF_Bj4a6rS8_DuXY2bxtCHw9N2XQ%3A1787144229731&ei=JaiFavymLO7_7_UPIYup0AM&biw=1440&bih=778&ved=0ahUKEwi8zeaH36yWAXXu_7sIHZVFCjoQ4dUDCBA&uact=5&oq=lista+comune+italia+istat&gs_lp=Egxnd3Mtd2l6LXNlcuAIGWxpc3RhIGNvbXVvZSBpdGFsaWEgaXN0YXQyBRAhGJ8FMgUQIRifBTIFECEYnwUyBRAhGJ8FMgUQIRifBTIFECEYnwUyBRAhGJ8FMgUQIRifBTIFECEYnwUyBRAhGJ8FSMQKULUDWMIcAF4AZABAJgBygGgAYEIqgEFMC41LjG4AQPIAQD4AQGYAg egAsoIwgIKEAAYRxjWBBiwA8ICDhAAGOQCGNYEGLAD2AEBwgIXEC4Y3AYYuAYY2gYY2AIYyAMYsAPYAQHCAGYQABgWGB7CAggQABgWGB4YCsICCBAAGIAEGKIEwgIEECEYFZgDAIgGAZAGD7oGBggBEAEYCZIHBT EuNS4xoAf2NbIHBT AuNS4xuAe0CMIHBTItNi4xyAcygAgB&sclient=gws-wiz-serp
- [44] «Tuttitalia.it - Guida ai Comuni, alle Province e alle Regioni d'Italia». Consultato: 19 agosto 2026. [Online]. Disponibile su: https://www.tuttitalia.it/#google_vignette
- [45] «ERA5 hourly data on single levels from 1940 to present». Consultato: 20 agosto 2026. [Online]. Disponibile su: <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview>
- [46] «ERA5-Land hourly data from 1950 to present». Consultato: 20 agosto 2026. [Online]. Disponibile su: <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview>
- [47] «CORINE Land Cover 2018 (vector/raster 100 m), Europe, 6-yearly — Copernicus Land Monitoring Service». Consultato: 20 agosto 2026. [Online]. Disponibile su: <https://land.copernicus.eu/en/products/corine-land-cover/clc2018>
- [48] «THE REVISED AND SUPPLEMENTED CORINE LAND COVER NOMENCLATURE». Consultato: 20 agosto 2026. [Online]. Disponibile su: https://land.copernicus.eu/content/corine-land-cover-nomenclature-guidelines/docs/pdf/CLC2018_Nomenclature_illustrated_guide_20190510.pdf
- [49] «IT:Key:highway - OpenStreetMap Wiki». Consultato: 20 agosto 2026. [Online]. Disponibile su: <https://wiki.openstreetmap.org/wiki/IT:Key:highway>
- [50] P. Dadv *et al.*, «Traffic-Related Air Pollution and Birth Weight: The Roles of Noise, Placental Function, Green Space, Physical Activity, and Socioeconomic Status (FRONTIER)», *Res Rep Health Eff Inst*, vol. 2026, p. 236, feb. 2026.

- [51] Y. M. Idir, O. Orfila, V. Judalet, B. Sagot, e P. Chatellier, «Mapping Urban Air Quality from Mobile Sensors Using Spatio-Temporal Geostatistics», *Sensors (Basel)*, vol. 21, fasc. 14, p. 4717, lug. 2021, doi: 10.3390/s21144717.
- [52] A. Fassò *et al.*, «Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy», *Sci Data*, vol. 10, p. 143, mar. 2023, doi: 10.1038/s41597-023-02034-0.
- [53] J. D. Marshall, E. Nethery, e M. Brauer, «Within-urban variability in ambient air pollution: Comparison of estimation methods», *Atmospheric Environment*, vol. 42, fasc. 6, pp. 1359–1369, feb. 2008, doi: 10.1016/j.atmosenv.2007.08.012.
- [54] L. Bao *et al.*, «Association analysis of maternal exposure to air pollution during pregnancy and offspring asthma incidence», *Reprod Health*, vol. 22, p. 29, feb. 2025, doi: 10.1186/s12978-025-01967-6.
- [55] J. Kim, J. Han, K. Park, e S. Seok, «Improved IDW Interpolation Application Using 3D Search Neighborhoods: Borehole Data-Based Seismic Liquefaction Hazard Assessment and Mapping», *Applied Sciences*, vol. 12, fasc. 22, p. 11652, gen. 2022, doi: 10.3390/app122211652.
- [56] M. Jerrett *et al.*, «Spatial Analysis of Air Pollution and Mortality in California», *Am J Respir Crit Care Med*, vol. 188, fasc. 5, pp. 593–599, set. 2013, doi: 10.1164/rccm.201303-0609OC.
- [57] *Direttiva 2008/50/CE del Parlamento europeo e del Consiglio, del 21 maggio 2008, relativa alla qualità dell'aria ambiente e per un'aria più pulita in Europa*, vol. 152. 2008. Consultato: 22 agosto 2026. [Online]. Disponibile su: <http://data.europa.eu/eli/dir/2008/50/oj>
- [58] M. Eeftens *et al.*, «Development of Land Use Regression Models for PM2.5, PM2.5 Absorbance, PM10 and PMcoarse in 20 European Study Areas; Results of the ESCAPE Project», *Environ. Sci. Technol.*, vol. 46, fasc. 20, pp. 11195–11205, set. 2012, doi: 10.1021/es301948k.
- [59] L. Huang, C. Zhang, e J. Bi, «Development of land use regression models for PM2.5, SO2, NO2 and O3 in Nanjing, China», *Environmental Research*, vol. 158, pp. 542–552, ott. 2017, doi: 10.1016/j.envres.2017.07.010.
- [60] «Home :: Corine Land Cover classes». Consultato: 22 agosto 2026. [Online]. Disponibile su: <https://land.copernicus.eu/content/corine-land-cover-nomenclature-guidelines/html/>
- [61] «Data Dictionary - Vocabulary». Consultato: 7 settembre 2026. [Online]. Disponibile su: <https://dd.eionet.europa.eu/vocabularyconcept/aq/areaclassification/urban/view?vocabularyFolder.workingCopy=false&facet=HTML+Representation>
- [62] «Data Dictionary - Vocabulary». Consultato: 7 settembre 2026. [Online]. Disponibile su: <https://dd.eionet.europa.eu/vocabularyconcept/aq/areaclassification/suburban/view?vocabularyFolder.workingCopy=false&facet=HTML+Representation>
- [63] «Data Dictionary - Vocabulary». Consultato: 7 settembre 2026. [Online]. Disponibile su:

- <https://dd.eionet.europa.eu/vocabularyconcept/aq/areaclassification/rural/view?vocabularyFolder.workingCopy=false&facet=HTML+Representation>
- [64] «Data Dictionary - Vocabulary». Consultato: 7 settembre 2026.
[Online]. Disponibile su:
<https://dd.eionet.europa.eu/vocabularyconcept/aq/areaclassification/rural-nearcity/view?vocabularyFolder.workingCopy=false&facet=HTML+Representation>
- [65] «Data Dictionary - Vocabulary». Consultato: 7 settembre 2026.
[Online]. Disponibile su:
<https://dd.eionet.europa.eu/vocabularyconcept/aq/areaclassification/rural-regional/view?vocabularyFolder.workingCopy=false&facet=HTML+Representation>
- [66] «Data Dictionary - Vocabulary». Consultato: 7 settembre 2026.
[Online]. Disponibile su:
<https://dd.eionet.europa.eu/vocabularyconcept/aq/areaclassification/rural-remote/view?vocabularyFolder.workingCopy=false&facet=HTML+Representation>
- [67] M. Vazifehdan *et al.*, «Longitudinal Missing Data Imputation for Predicting Disability Stage of Patients with Multiple Sclerosis», 22 gennaio 2025, *arXiv*: arXiv:2501.12927. doi: 10.48550/arXiv.2501.12927.
- [68] «Full-stack observability for the agentic era | Grafana Labs», Grafana Labs. Consultato: 29 agosto 2026. [Online]. Disponibile su:
<https://grafana.com/index/>
- [69] A. Istat, *APPI Comuni e popolazione per zona altimetrica.xlsx*. doi: 10.1481/Istat.RapportoTerritorio.2020.APP1.
- [70] L. Angelici *et al.*, «Effects of particulate matter exposure on multiple sclerosis hospital admission in Lombardy region, Italy», *Environ Res*, vol. 145, pp. 68–73, feb. 2016, doi: 10.1016/j.envres.2015.11.017.