Applicazione del modulo Machine Learning sull'area urbana torinese

Autori
A. D'Ausilio, U. Giuriato, C. Silibello

Riferimento
ARIANET R2023.19

Dicembre 2023

INDICE

1. Introduzione

Nel precedente studio (Relazione R2022.25- Utilizzo del modello Machine Learning XGBoost a supporto della valutazione dello stato di Qualità dell'aria sulla Regione Piemonte) sono state presentate mappe regionali ad 1 km di risoluzione spaziale relative agli standard di qualità dell'aria per l'anno 2021 dei seguenti inquinanti: biossido di azoto (NO_2), ozono (O_3), PM_{10} e $PM_{2.5}$. Tali mappe sono state prodotte mediante l'applicazione di modelli ML a valle di simulazioni modellistiche realizzate con il CTM FARM a 4 km di risoluzione spaziale.
I promettenti risultati ottenuti hanno suggerito l'applicazione di tali metodologie per l'aumento della risoluzione spaziale a 200 m di analoghe simulazioni modellistiche realizzate con il CTM FARM sull'area torinese (risoluzione spaziale originaria pari a 1 km).
Queste attività (a scala regionale e sull'area urbana torinese) sono state effettuare nell'ambito del progetto SPoTT 2.
Nel presente studio è stato applicato l'algoritmo di *Machine Learning* denominato *Random Forest* (RF) al fine di poter produrre mappe di qualità dell'aria sull'area urbana torinese, per l'anno 2019, alla risoluzione spaziale di 200 m.

2. Spazializzazione dei dati della rete di monitoraggio sull'area urbana torinese

Nel presente studio sono stati considerati i seguenti inquinanti: biossido di azoto (NO_2), ozono (O_3), PM_{10}, $PM_{2.5}$ ed il dominio, rappresentato nella figura seguente, alla risoluzione spaziale di 200m.

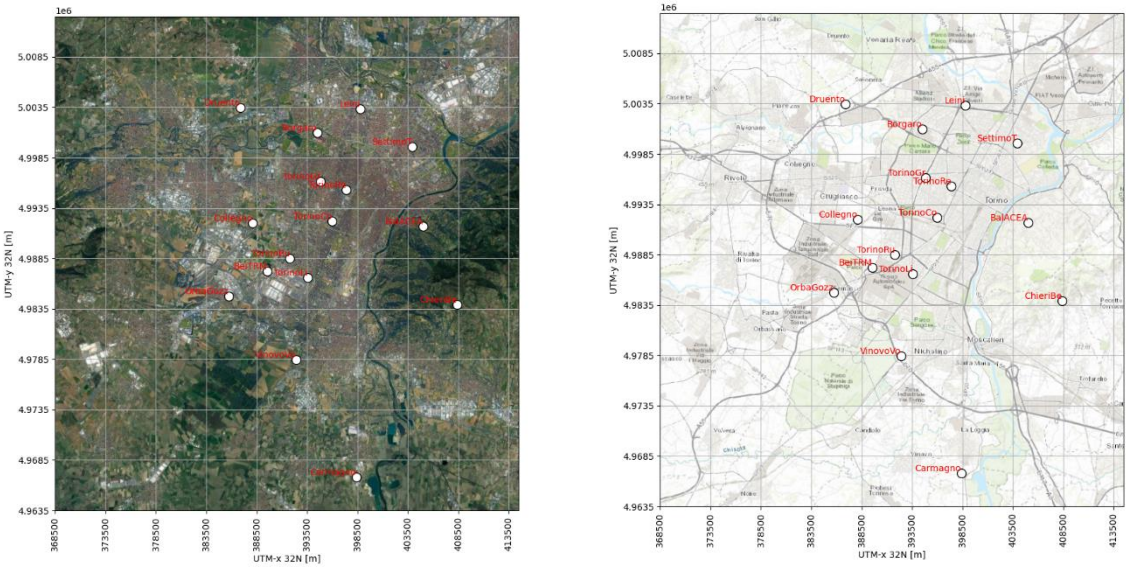


Figura 1 – Dominio di studio. Sinistra: Mappa satellitare con dettaglio della rete di monitoraggio. Destra: Mappa topografica con dettaglio della rete di monitoraggio. Sistema di coordinate UTM WGS84 - zona 32 [m].

La Tabella 1 presenta la suddivisione della rete di monitoraggio per inquinante.

Tabella 1. Dati anagrafici della rete di monitoraggio.

Nome Stazione	x [m]	y[m]	NO_2	O_3	PM_{10}	$PM_{2.5}$
Baldissero T. (ACEA)	404980	4991698	●	●	●	
Beinasco TRM – Aldo Mei	389527	4987219	●		●	●
Borgaro T. - Caduti	394518	5001005	●	●	●	●
Carmagnola	398375	4966819	●		●	
Chieri - Bersezio	408380	4983914	●	●		●
Collegno - Francia	388061	4991997	●		●	
Druento – Parco la Mandria	386869	5003485	●	●	●	
Leini (ACEA) - Grande Torino	398765	5003348	●	●	●	●
Orbassano - Gozzano	385702	4984737	●	●		
Settimo T. - Vivaldi	403942	4999584	●		●	●
Torino Consolata	395961	4992226	●		●	
Torino Grassi	394836	4996153			●	
Torino Lingotto	393571	4986609	●	●	●	●
Torino Rebaudengo	397361	4995339	●		●	●
Torino Rubino	391781	4988521	●	●	●	●
Vinovo Volontari	392417	4978446	●	●		

La rappresentatività spaziale di ciascuna centralina in un'area di raggio pari a 1 km è presentata in Figura 2 utilizzando fotografie da satellite. Dalle immagini emergono diverse coperture spaziali:

- Le stazioni di *Baldissero T. (ACEA)* e di *Druento- La Mandria* risultano le uniche centraline a coprire prevalentemente aree boschive e rurali.
- Le stazioni di *Vinovo Volontari* e *Chieri-Bersezio* presentano una copertura sia di territorio rurale che urbano e sono localizzate in periferia.
- Le stazioni di *Torino Rebaudengo*, *Collegno-Francia*, *Torino Grassi* sono concentrate in contesti urbani, posizionate lungo vie primarie e secondarie di traffico.
- Le restanti centraline mostrano una combinazione di territorio urbano, industriale ed agricolo.

Infine, è importante evidenziare che non sono presenti stazioni in prossimità di aree aeroportuali.

Vinovo Volontari

Orbassano - Gozzano

Chieri - Bersezio

Collegno - Francia

Torino Rebaudengo

Settimo Torinese

Carmagnola

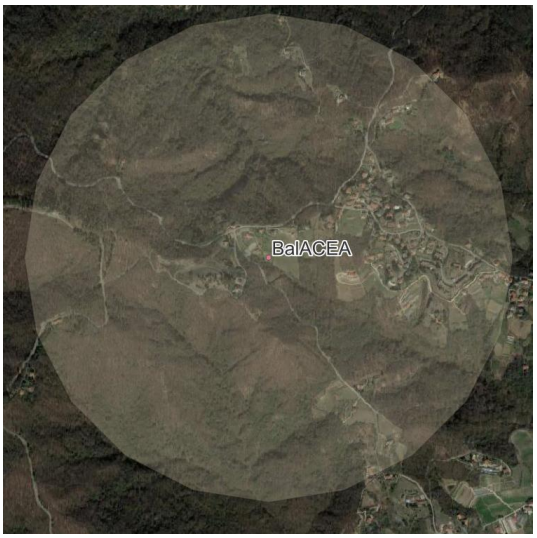
Torino Grassi

Borgaro T. Caduti

Druento – Parco La Mandria

Torino Rubino

Baldissero T. (ACEA)



Torino Consolata

Beinasco TRM – Aldo Mei

Leini (ACEA) - Grande Torino

Torino Lingotto

Figura 2. Dettaglio delle centraline ottenuto mediante utilizzo di immagini satellitari. Il dettaglio è ingrandito per mostrare l'area circostante a ciascuna centralina entro un raggio di 1 km (Map data ©2015 Google)

2.1. Calibrazione ed applicazione dell'algoritmo di *machine learning* RF_{GRID}

I campi di concentrazione di inquinanti prodotti da FARM sono stati elaborati dall'algoritmo *machine learning* denominato *Random Forest* (Breiman, 2001), nel seguito RF, per produrre corrispondenti mappe di qualità dell'aria sul dominio di studio, alla risoluzione spaziale di 200 m nelle due direzioni x: Ovest-Est e y: Sud-Nord nel sistema di coordinate UTM WGS84 - zona 32. I parametri della griglia di calcolo relativa a tale dominio sono i seguenti. Centro della cella più a SW del dominio: x = 368600 m, y = 4963600 m (longitudine 7.34, latitudine 44.81). Centro della cella più a NE del dominio: x = 414600 m, y = 5012600 m (longitudine 7.90, latitudine 45.26). Passo di griglia dx = 200 m, dy = 200 m. Numero di punti in direzione x: 230, in direzione y: 245.

L'algoritmo RF, che consiste in un insieme di alberi decisionali (da cui "foresta"), è adatto sia a problemi di classificazione (separazione dei dati in classi) che di regressione, ed è stato sviluppato per risolvere gli errori di *over-fitting* (che si verifica quando un modello raggiunge buoni risultati sui dati di addestramento ma scarse performance sui dati di controllo) e di alta varianza tipici di un singolo albero decisionale. Ogni albero, viene costruito con un *bootstrap* dei dati di input e ad ognuno viene assegnato casualmente un sottoinsieme di predittori (Liaw e Wiener, 2002). L'*output* finale dell'insieme è calcolato facendo la media degli output di ogni singolo albero.

L'applicazione dell'algoritmo ML-RF si articola in due fasi: la prima finalizzata al suo addestramento, ovvero alla sua capacità di riprodurre le concentrazioni osservate nei siti di monitoraggio (fase di training) sulla base di un insieme di predittori. La seconda, detta di inferenza o generalizzazione, è volta a stimare le concentrazioni nelle celle della griglia dove non sono disponibili osservazioni.

L'elenco dei predittori spaziali e spazio-temporali, scelti per catturare le peculiari fluttuazioni temporali e spaziali dei campi di concentrazione, è di seguito descritto. Tutti i predittori sono stati interpolati dalle loro risoluzioni originali a quella target (200 m) mediante analisi geospaziale. In particolare, è stata utilizzata un'interpolazione di tipo *nearest neighbours*, in modo da non introdurre artefatti interpolativi.

2.1.1. Predittori statici

Di seguito sono riportati i predittori statici considerati:

- copertura del suolo, basati sul database Corine Land Cover (CLC), riferiti all'anno 2018 (Corine Land Cover 2018) e definiti come percentuale di ciascuna cella della griglia coperta dalle seguenti otto classi mostrate in Figura 3:
  1. Tessuto urbano (*CRNurfa*);
  2. Unità industriali, commerciali e di trasporto (*CRNinco*);
  3. Aeroporti (*CRNairp*);
  4. Altre superfici artificiali (*CRNoart*);
  5. Aree agricole (*CRNagri*);
  6. Foreste (latifoglie, conifere o miste, *CRNfore*);
  7. Prati, aree arbustive, pianura brulla, spiagge naturali (*CRNnatu*);
  8. Masse d'acqua (*CRNwate*);

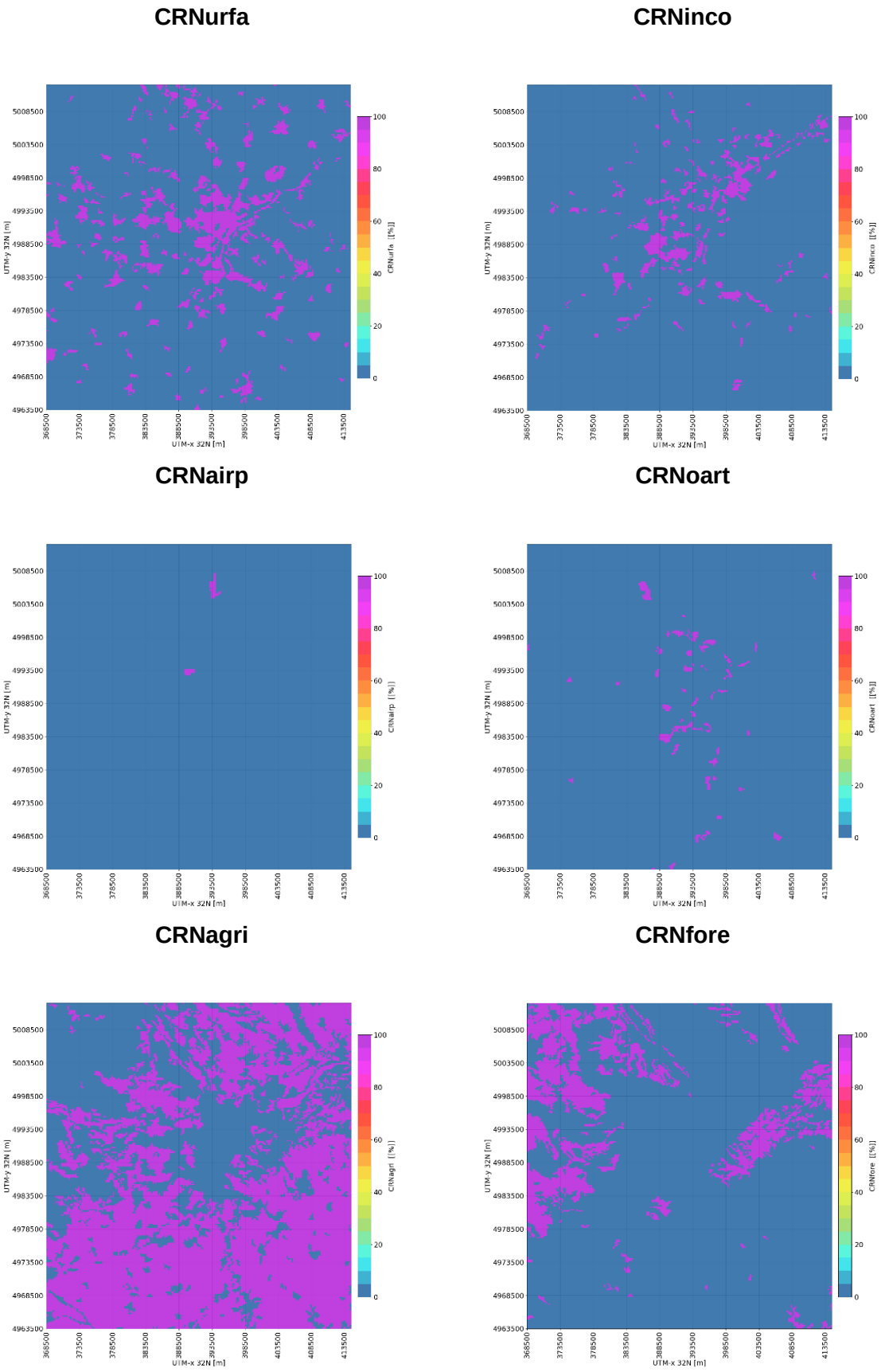


Figura 3. Predittori di copertura del suolo secondo Corine Land Cover 2018. Risoluzione 200m

- inventario globale della distribuzione spaziale e della densità della superficie impermeabile costruita (Impervious Surface Area, **ISA**). Esempi di ISA includono strade, parcheggi,

edifici, vialetti, marciapiedi e altre superfici artificiali, derivata da https://ngdc.noaa.gov/eog/dmsp/download_global_isa.html, dati NOAA relativi all'anno 2020;

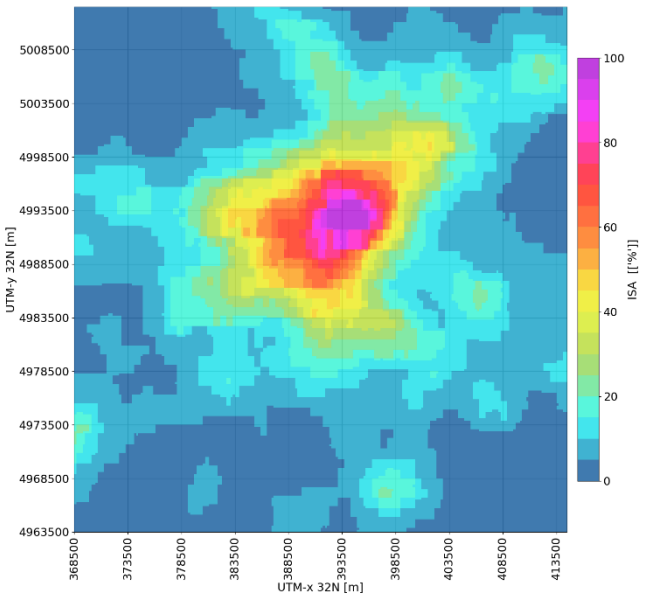


Figura 4. Impervious Surface Area, ISA.

- Immagini satellitari notturne mediante le quali sono state prodotte mappe globali della luce notturna (Light At Night, LAN), raccolti dal Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) nell'anno 2015.

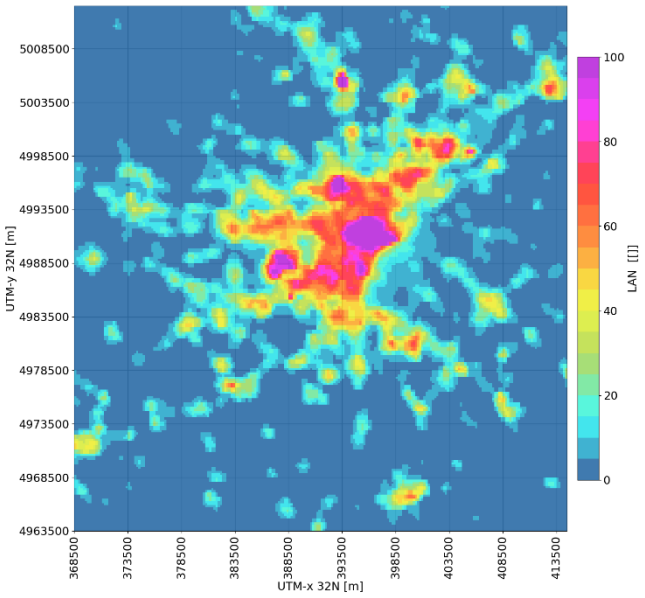


Figura 5. Light at Night, LAN

- densità di popolazione europea, EUPOP:

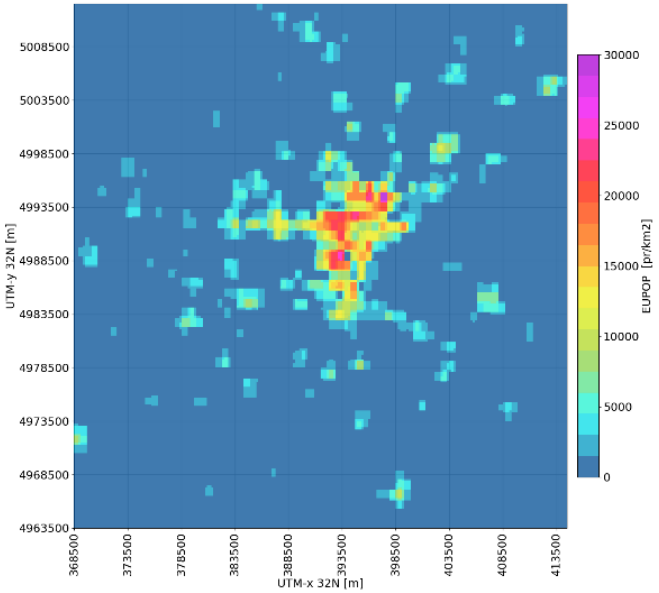
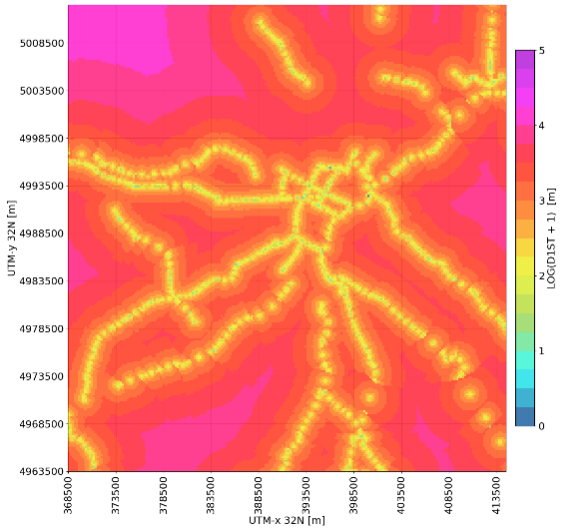


Figura 6. Densità abitativa, EUPOP

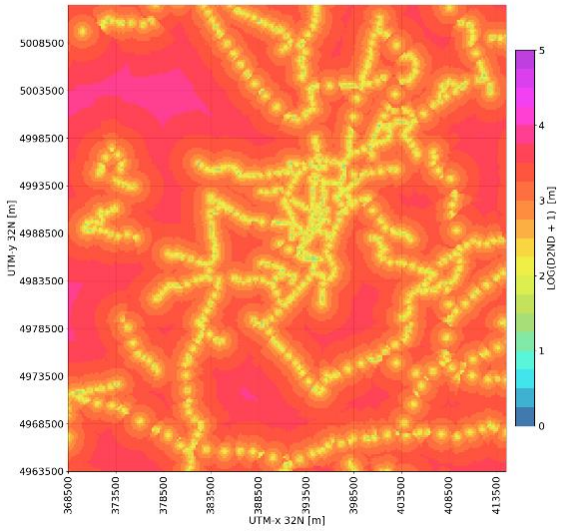
- statistiche di vario tipo definite all'interno della cella ed a distanza del baricentro della cella, considerando in particolare quattro tipi di strade: autostrada, primarie, secondarie e locali. La rete stradale deriva dal progetto (Open Street Map – OSM, OpenStreetMap):
  - Distanza del centro della gridbox dall'autostrada più vicina (**DMTW**);
  - Distanza del centro della gridbox dalle principali strade primarie (**D1ST**);
  - Distanza del centro della gridbox dalle principali strade secondarie (**D2ND**);
  - Distanza del centro della gridbox dalle principali strade locali (**D3RD**);

L'approccio adottato per mitigare l'impatto dei valori estremi della distanza dalle strade è stato l'applicazione della funzione logaritmica in base dieci della distanza, Log(d + 1).

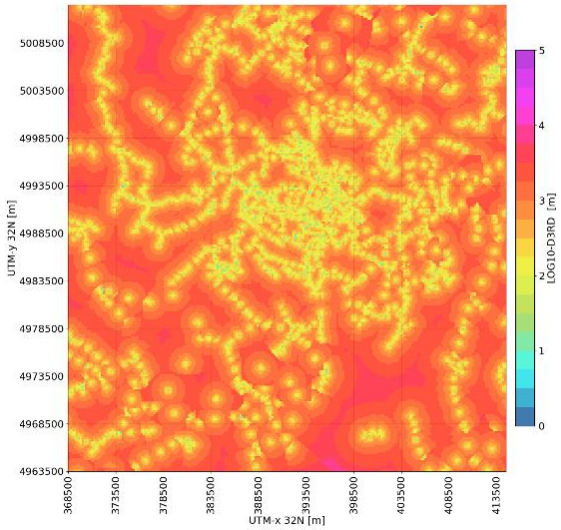
D1ST- Distanza strade primarie



D2ND- Distanza strade secondarie



D3RD - Distanza strade locali



DMTW – Distanza autostrade

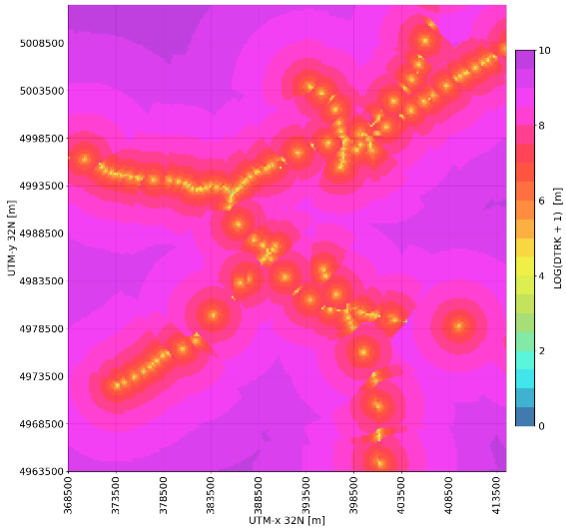


Figura 7. Predittori di distanza da strade primarie (D1ST), strade secondarie (D2ND), strade locali (D3RD), autostrade (DMTW).

- Elevazione media (**ELEV**) ottenuta dal Servizio Copernicus Land Monitoring Service (CLMS) - European Digital Elevation Model;

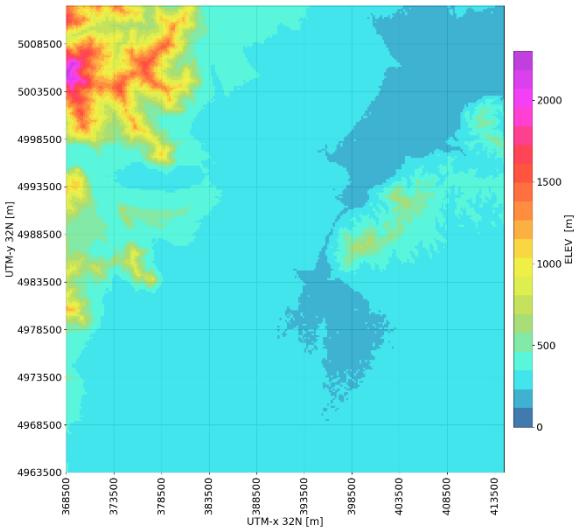


Figura 8. Predittore di elevazione media EU-DEM.

2.1.2. Predittori mensili

È stato considerato l'indice di superficie fogliare (Leaf Area Index, **LAI**), definito come metà dell'area totale degli elementi verdi della chioma per unità di superficie orizzontale del suolo. Il valore derivato dal satellite corrisponde al LAI verde totale di tutti gli strati della chioma, compreso il sottobosco che può rappresentare un contributo molto significativo, in particolare per le foreste. In pratica, il LAI quantifica lo spessore della copertura vegetale (dati raccolti dal satellite PROBA-V destinato all'osservazione della vegetazione, da cui il -V del nome, lanciato a metà del 2013, https://land.copernicus.eu/global/products/lai).
La mappa di LAI dell'anno 2019 è presentata in Figura 9.

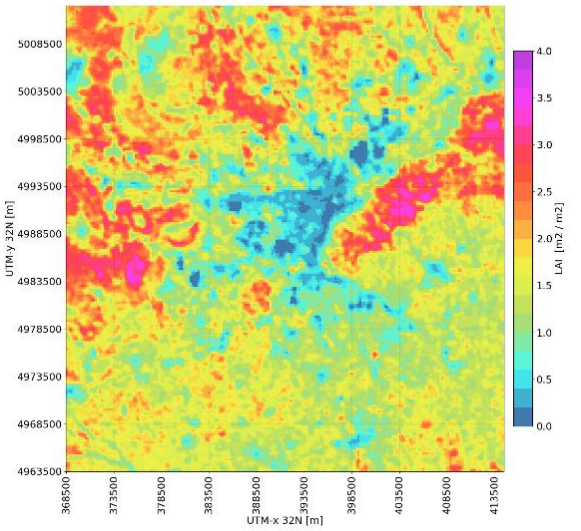


Figura 9. Mappa annuale (2019) del predittore mensile LAI.

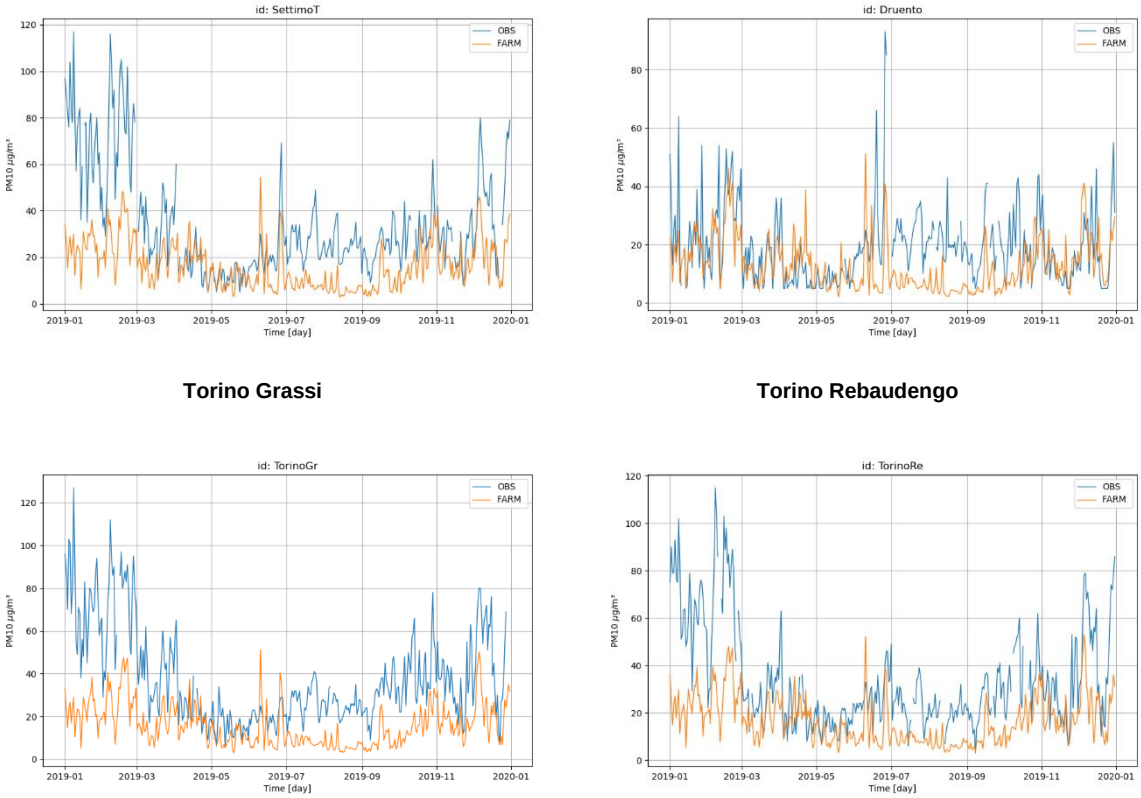
2.1.3. Predittori orari e giornalieri

Sono stati utilizzati i campi di concentrazione prodotti da simulazioni FARM di PM_{10}, $PM_{2.5}$, NO_2, O_3 disponibili in media oraria alla risoluzione spaziale di 1km sulla griglia regionale di cui alla Figura 1. Per i modelli di PM_{10}, $PM_{2.5}$, i campi di concentrazione dei suddetti campi è stata aggregata su base giornaliera mentre per i modelli di O_3 e NO_2 sono stati utilizzati i campi orari. Una novità rispetto agli studi precedenti consiste nell'uso di più campi FARM come predittori, non limitandosi esclusivamente al campo corrispondente all'inquinante target, vista l'alta correlazione tra osservazioni di ogni specie e i valori di FARM per ogni specie (vedi sezione 2.1.5).

In Figura 10 sono riportati, a titolo di esempio, i confronti tra le serie temporali di diverse stazioni per il PM_{10} e il relativo campo di FARM filtrato sulla cella della stazione. Sebbene le serie risultino correlate, si evidenzia una generale sottostima delle predizioni di FARM, indice di un bias sistematico nel modello CTM particolarmente visibile nei mesi invernali. Il presente lavoro si pone tra gli obiettivi quello di verificare se il modulo di machine learning sia in grado di colmare tale bias.

I predittori FARM sono identificati nel modello come c_PM10, c_PM25, c_NO2 e c_O3

Settimo Torinese | **Druento – Parco La Mandria**



Torino Grassi | **Torino Rebaudengo**

Figura 10. Confronto tra profili di concentrazione di PM_{10} generati dal modello FARM (arancio) e le misurazioni osservate (blu) alle stazioni di monitoraggio. Confronto tra le serie temporali osservate alle stazioni e relativo campo di concentrazione di FARM per il PM_{10}.

2.1.4. Altri predittori

Il giorno dell'anno, o giorno giuliano, che va da 1 (il primo gennaio) a 365 (il 31 dicembre, 366 negli anni bisestili), viene spesso usato come indicatore della stagione. In questo lavoro, il giorno giuliano non viene usato direttamente come predittore, perché distingue maggiormente tra il 31 dicembre e il primo gennaio, che sono date consecutive, che tra il primo gennaio e il primo luglio, che hanno stagionalità opposte. Si è scelto di usare invece il seno e il coseno del giorno giuliano moltiplicato per 2π/365 (*julianday_x, julianday_y*), perché questa coppia di indicatori è ciclica.
Analogamente sono stati definiti il seno e il coseno dei giorni della settimana moltiplicati per 2π/7 (*dayofweek_x, dayofweek_y*), di modo di invitare il modello a tenere conto della differenza tra giorni feriali e giorni festivi.
Infine, nel caso di variabili orarie si è scelto di usare il seno e il coseno dell'ora del giorno moltiplicato per 2π/24 (*hours_x*, *hours_y*).

2.1.5. Rete osservativa di $PM_{2.5}$

La rete osservativa di $PM_{2.5}$ è composta da un numero inferiore di stazioni rispetto a quella di PM_{10} (Tabella 1). In particolare, sono assenti le stazioni *di Baldissero T. (ACEA), Collegno - Francia, Druento – Parco La Mandria, Torino Consolata e Torino Grassi*.
La mancanza di due stazioni rurali unita alla disparità nella distribuzione tra le due reti rischia di restituire un quadro poco rappresentativo per il $PM_{2.5}$ nell'area di studio. Per superare questa discrepanza, è stato utilizzato un modello di RF preliminare al fine di ampliare la rete osservativa di $PM_{2.5}$ a partire da PM_{10}. Questo approccio è descritto nel lavoro di (Stafoggia et al., 2020). Il modello utilizza le osservazioni di PM_{10} e $PM_{2.5}$ nelle stazioni co-localizzate per stimare le concentrazioni di $PM_{2.5}$ nei siti di monitoraggio e per i giorni in cui sono disponibili solo dati per il PM_{10}. Ciò è stato ottenuto addestrando il modello descritto nell'equazione:

$$\frac{PM_{2.5,obs}}{PM_{10,obs}} \sim RF\left(PM10_{obs}, \frac{c\_PM_{2.5}}{c\_PM_{10}}\right)$$

La decisione di utilizzare il rapporto tra le due misure come variabile target trova giustificazione nella necessità di addestrare il modello tenendo conto della natura delle due quantità, garantendo cioè che il rapporto resti inferiore a 1. Nelle sezioni successive si riferirà a tale modello come $RF_{ST.}$
I passaggi principali per la generazione di mappe di $PM_{2.5}$ possono essere così elencati:

- Allenamento del modello RF_{ST} utilizzando le stazioni dove sono presenti sia misure di PM_{10} che di $PM_{2.5}$. Il modello prodotto consente di stimare il rapporto di $PM_{2.5}/PM_{10}$ alle stazioni dove è misurato solo il PM_{10}.
- Tramite il modello prodotto si stima il rapporto di $PM_{2.5}/PM_{10}$ alle stazioni dove è misurato solo il PM_{10}.
- La concentrazione target di $PM_{2.5}$ alle stazioni dove è misurato solo il PM_{10} si ottiene moltiplicando le osservazioni di PM_{10} con il rapporto stimato.
- Le concentrazioni di $PM_{2.5}$ stimate sono integrate nel database delle osservazioni da utilizzare nell'addestramento del modello RF_{GRID}

2.1.6. Predittori e rete osservativa

Il modello di RF utilizzato per produrre mappe ad una risoluzione di 200 m e indicato come RF_{GRID} è descritto come di seguito:

$$Inquinante_{i,j} \sim RF\left(P_{d1\{i,j\}}, .., P_{dm\{i,j\}}, P_{s1_{\{i\}}}, .., P_{sn\{i\}}\right)$$

dove la concentrazione di ciascun inquinante (PM_{10}, $PM_{2.5}$, NO_2, O_3) misurata in un punto griglia *i* al tempo *j* è allenata con predittori dinamici P_d alla stessa posizione e allo stesso tempo, e con i predittori statici P_s alla stessa posizione presentati in 2.1.1.

Un primo indicatore per comprendere quanto possono essere importanti i predittori nello spiegare il valore di un osservabile target, o in generale le relazioni tra le diverse variabili di un dataset, è l'indice di correlazione di Pearson. Nella Figura 11 si riporta la *heatmap* del coefficiente di correlazione tra le principali variabili considerate. Si evince come esista una chiara correlazione tra le osservazioni dei vari inquinanti e i relativi campi di FARM, oltre alle osservazioni e ai campi di FARM degli altri inquinanti. Già da ciò, ci si può quindi aspettare che FARM costituirà un predittore preferenziale nei modelli di Machine Learning. Si evidenzia, inoltre, una certa correlazione tra le concentrazioni degli inquinanti e seno/coseno del giorno giuliano, Leaf Area Index e distanza dalle strade, in particolare quelle secondarie. Quest'ultima osservazione può essere dovuta al fatto che la maggior parte delle stazioni sono infatti posizionate in prossimità di strade secondarie. Inoltre, per l'NO_2 esiste una correlazione non trascurabile con i predittori Impervious Surface Area e Light At Night. In Figura 6 si riportano (a titolo esemplificativo) i joint plot tra la variabile osservata di PM_{10} e i campi di FARM, nei quali vengono confrontate le distribuzioni di ciascuna coppia di variabili. La somiglianza tra le distribuzioni è un'ulteriore indicazione della correlazione esistente tra le concentrazioni misurate e i campi di FARM. Considerazioni analoghe valgono per gli altri inquinanti.

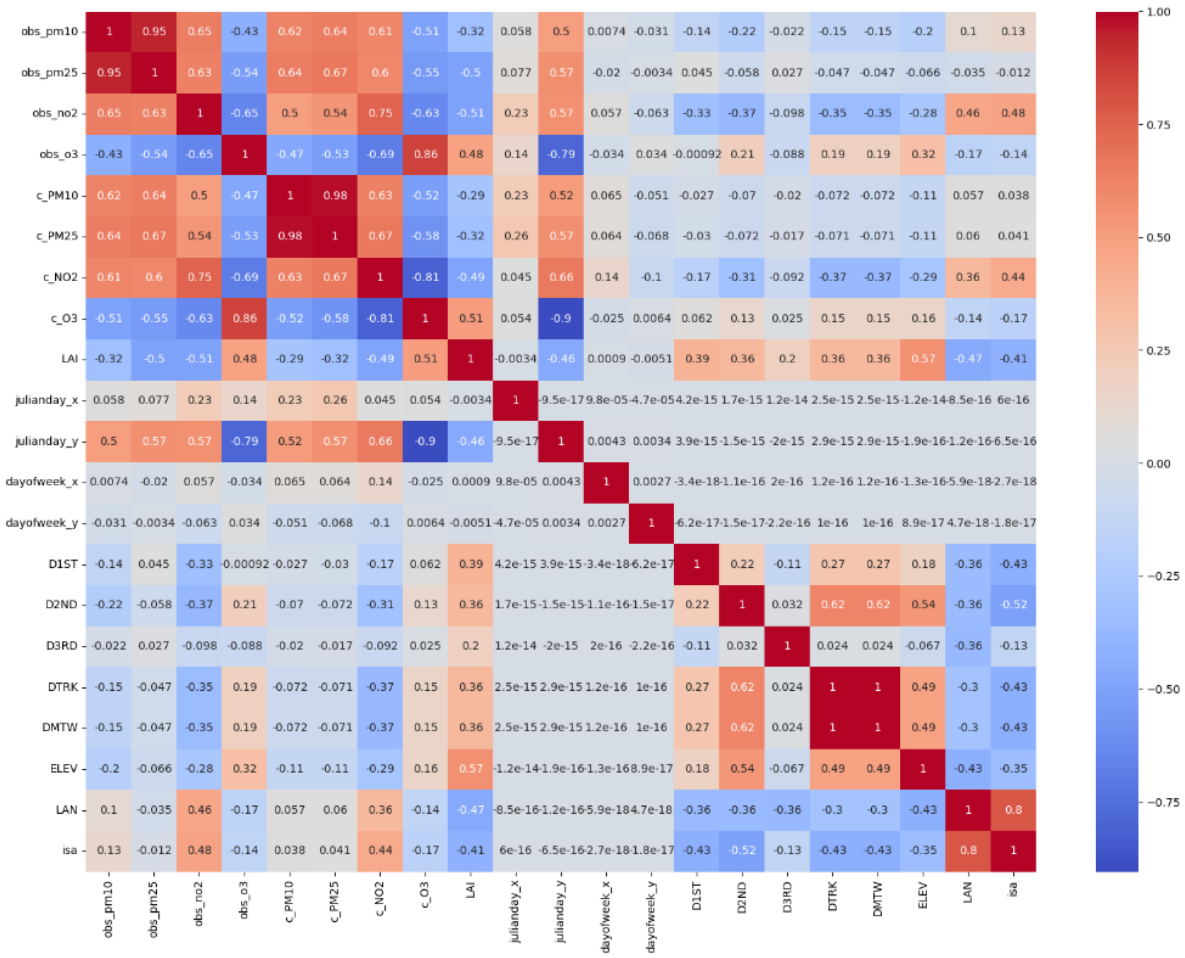


Figura 11. Heatmap del coefficiente di correlazione di Pearson tra le variabili del dataset. Valori vicini a 1 indicano alta correlazione lineare, valori vicini a -1 indicano alta anti-correlazione lineare, valori prossimi allo 0 indicano assenza di correlazione lineare.

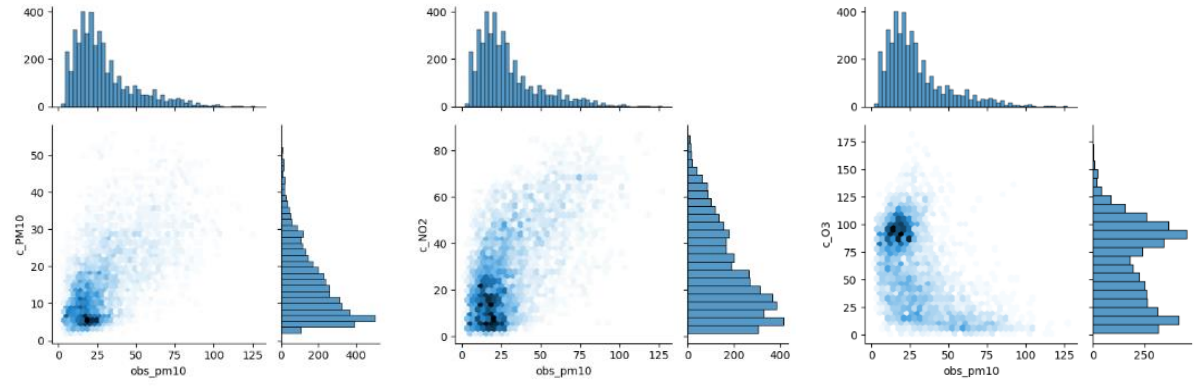


Figura 12. Joint plots tra osservazioni di PM_{10} e campi di FARM (PM_{10}, NO_2 e O_3). Le concentrazioni sono espresse in µg/m3. L'intensità del colore negli scatter plot esagonali è proporzionale alla densità di punti nella regione corrispondente.

Infine, è importante sottolineare che l'efficacia di un predittore non è condizionata esclusivamente dalla sua importanza nello spiegare il campo di concentrazione nel dataset di addestramento. Infatti, la calibrazione è effettuata sulle osservazioni di una definita rete di stazioni localizzate nello spazio. La distribuzione di valori di un predittore sul dataset di

addestramento (quello filtrato sulla posizione delle stazioni) costituisce un campione, cioè un sottoinsieme, della distribuzione complessiva del predittore su tutto il territorio di studio. Pertanto, un predittore sarà tanto più efficace nella regressione del campo di concentrazione, quanto più il campione "visto" dalle stazioni riesce a riprodurre l'intervallo di valori del predittore sul grigliato target. Per valutare indicativamente la somiglianza tra le distribuzioni di un predittore rispettivamente sulla rete osservativa e sul grigliato target, esse sono state confrontate graficamente. Una rappresentazione si trova in Figura 13, in cui vengono confrontate le distribuzioni di vari predittori.

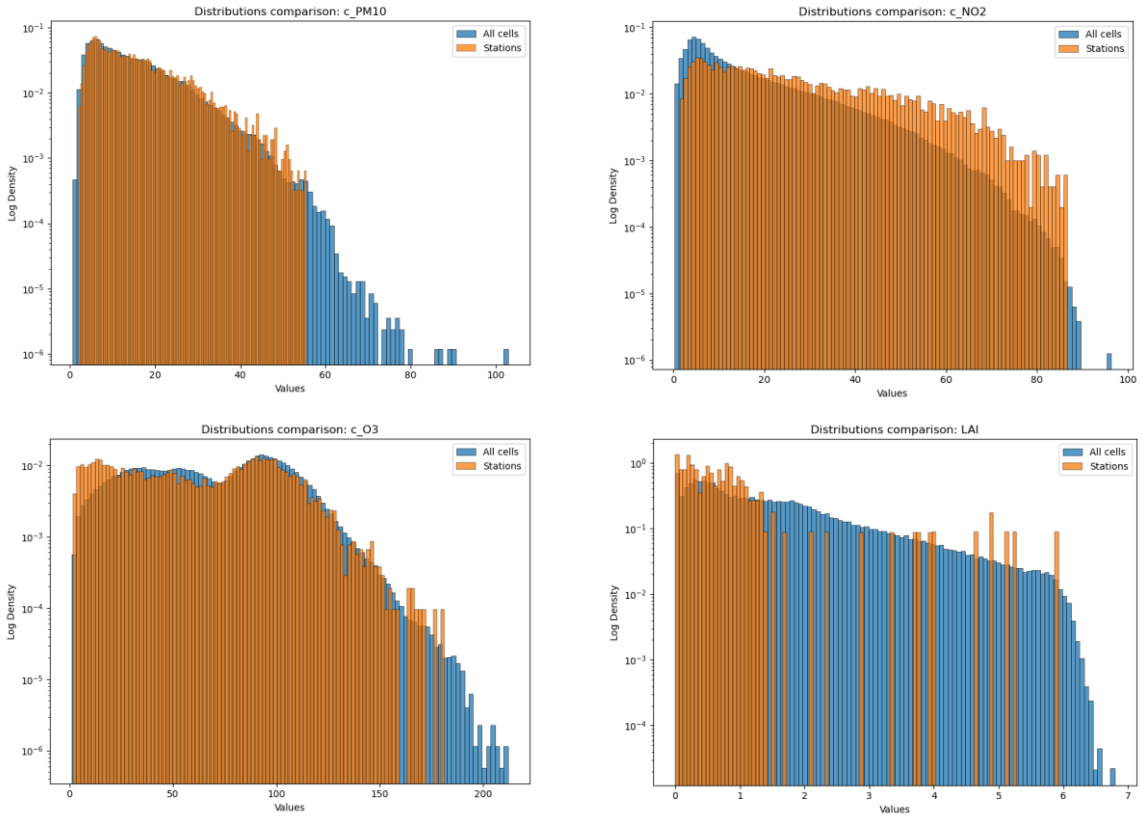


Figura 13. Confronto tra la densità di distribuzione logaritmica tra alcuni predittori campionati sulla rete osservativa e su tutte le celle del dominio di studio. Le distribuzioni sono normalizzate a 1 e sono mostrate in scala logaritmica per accentuare le differenze nelle code.

Si evince, per esempio, che il campo di concentrazione di PM_{10} sul grigliato target presenta dei valori più alti rispetto all'estremo superiore della distribuzione campionata dalle stazioni. È lecito aspettarsi quindi che il comportamento del modello RF_{GRID} allenato sulle stazioni, in fase di inferenza non sarà accurato nell'utilizzare il campo di FARM nelle celle in cui esso assume tali valori. Viceversa, il campo di FARM di NO_2 presenta distribuzioni simili se campionato sulle stazioni o su tutte le celle. Questo suggerisce una maggiore efficacia di tale predittore in fase di inferenza, qualora il modello RF_{GRID} apprenda che esso ha un peso alto nello spiegare la variabile target. Per il predittore LAI, il campione sulla rete osservativa, non è sufficientemente rappresentativo della distribuzione sull'intero dominio eccetto per un intervallo ridotto della distribuzione campionata.

Una volta addestrato il modello e apprese le importanze dei predittori, sarà possibile tentare di dare una stima a posteriori (quantitativa e aggregata) dell'affidabilità della predizione, sulla base della "copertura" dei predittori su ogni cella del grigliato target (vedi sezione 2.3).

2.1.7. Ottimizzazione

Le osservazioni di una stazione di misura sono associate al punto cella del grigliato target in cui cade la stazione. L'ottimizzazione è effettuata usando i valori dei predittori associati alle grid-box che contengono stazioni di misura. In questo modo, ad ogni valore osservato viene associato l'insieme completo dei predittori validi per quella posizione spaziale (quella della stazione di misura) e quel tempo (data e ora dell'osservazione oraria, data dell'osservazione giornaliera).

La messa a punto del modello RF_{GRID} è stata effettuata in *python*, utilizzando il modulo *ensemble.RandomForestRegressor* della la libreria "*scikit-learn*" (Pedregosa, 2011)

Il modello ML-RF possiede degli iper-parametri, ossia parametri esterni che definiscono il tipo di RF da utilizzare e che devono essere fissati prima dell'addestramento. Essi sono stati ottimizzati tramite la procedura di cross-validation grid search, di seguito brevemente descritta.

1. Per ogni iperparametro, è stata definito un set di valori da testare
2. Per ogni combinazione di iper-parametri è stata effettuata una cross-validazione a 5-fold, cioè
   - Il dataset è stato diviso in una serie di sottoinsiemi uguali (fold), contenenti ciascuno il 20% dei dati.
   - Per ciascun fold, il modello RF_{GRID} è stato allenato sull'unione dei fold esclusi e testato sul fold selezionato, misurandone lo score (RMSE, vedi sezione seguente)
   - Lo score complessivo per la combinazione di iper-parametri testata viene calcolato come media dello score sui diversi fold
3. Viene scelta la combinazione di iper-parametri che da luogo allo score migliore.

Gli iper-parametri ottimizzati sono i seguenti

- *n_estimators*: il numero di alberi decisionali nella "foresta". Più alberi possono migliorare la performance, ma aumentano il tempo di addestramento, oltre ad indurre un possibile over-fitting.
- *max_depth***:** Indica la massima profondità dell'albero. Limitarlo può prevenire l'over-fitting.
- *min_samples_split***:** È il numero minimo di campioni richiesti per dividere un nodo interno degli alberi decisionali. Aiuta a evitare divisioni che portano a nodi troppo specifici.
- *min_samples_leaf***:** Rappresenta il numero minimo di campioni richiesti per essere in una foglia (i.e. un nodo terminale di un albero). Aiuta a controllare la dimensione delle foglie dell'albero ed evitare over-fitting.
- *max_features***:** Indica il numero massimo di predittori considerati per dividere un nodo. Controlla la variazione e la diversità tra gli alberi.
- *max_leaf_nodes:* Rappresenta il numero massimo di foglie che un albero può avere durante la crescita. Limitare questo parametro può impedire la crescita eccessiva dell'albero, contribuendo a evitare l'over fitting e a semplificare la struttura dell'albero di decisione.

L'ottimizzazione è stata effettuata separatamente per ogni inquinante. I valori "ottimali" ottenuti sono riportati in Tabella 2.

Tabella 2.: Iper-parametri del modello RF_{GRID} per gli inquinanti in esame

Inquinante	n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features	max_leaf_nodes
PM_{10}	800	40	2	2	0.85	1200
$PM_{2.5}$	400	30	2	2	0.85	700
NO_2	1500	75	10	5	0.75	1800
O_3	1500	100	5	5	0.75	1800

Per il modello RF_{ST} i valori sono presentati in Tabella 3.

Tabella 3. Iper-parametri del modello RF_{ST} per $PM_{2.5}$/PM_{10}

Target	n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features	max_leaf_nodes
$PM_{2.5}$/ PM_{10}	200	30	5	5	0.85	1000

2.2. Scores e validazione

Per valutare il successo degli algoritmi nell'avvicinare la stima ai dati è stato usato il Root-Mean-Square Error (RMSE), ovvero la radice quadrata dello scarto quadratico medio tra stima e osservazioni. Tale metrica è stata usata anche per selezionare gli iper-parametri dei modelli, come descritto nella sezione precedente.

La procedura di validazione utilizzata è la cross-validazione innestata, che consiste nell'eseguire una cross-validazione a 5-fold "esterna" e, per ciascuna fold, un'altra cross-validazione "interna" per l'ottimizzazione degli iper-parametri. Lo score complessivo è quindi la media degli score di ciascuna fold del loop esterno. Sebbene sia più esosa in termini di tempi di calcolo, tale procedura garantisce una maggiore robustezza dei risultati, rispetto ad una cross-validazione standard.

Nella Tabella 4 sono riportati la media e la Deviazioni Standard di RMSE relativi a FARM e prodotti dall' algoritmo RF_{GRID} con la procedura di validazione.

Se lo score di validazione rimane su livelli sensibilmente inferiori rispetto a FARM, significa che il *machine learning* riesce a usare in modo efficace le informazioni contenute nei predittori. Chiaramente non ci si aspetta che l'RMSE di FARM sia migliore, dato che la calibrazione di RF_{GRID} è stata effettuata sui dati osservati, i quali non entrano nei calcoli del modello CTM.

È da notare che lo score di validazione può presentare delle limitazioni sia di tipo spaziale, sia di tipo temporale. Le stazioni, infatti, non sono distribuite in modo omogeneo sul territorio, sebbene sembrino coprire in modo sufficiente l'urbanizzato di Torino, e un'estrazione casuale potrebbe dare risultati (falsamente) migliori rispetto a un campionamento dei dati per sotto-aree spaziali. Inoltre, se un modello ottimizzato viene usato con campi FARM di un anno diverso, o in previsione, ci si può aspettare che l'RMSE peggiori.

Tabella 4. Media e Deviazioni Standard di RMSE [µg m^{-3}] per i Modelli RF_{GRID}

MODELLO		**NO_2**	**O_3**	**PM_{10}**	**$PM_{2.5}$**
FARM	Media	21.2	29.5	20.5	14.0
	Dev. Std	0.42	0.93	0.74	0.64
RF – validazione	Media	8.2	12.0	7.2	5.9
	Dev. Std	0.23	0.22	0.21	0.52

2.3. Importanza dei predittori

La classifica dei predittori per importanza stimata contestualmente all'ottimizzazione dell'algoritmo RF_{GRID} per ciascun inquinante è presentata nelle Tabelle 5-8. L'importanza relativa dei predittori è definita in modo da avere somma 1. L'importanza cumulativa è ottenuta sommando le importanze relative dopo aver ordinato i predittori secondo la loro importanza relativa, dal più importante al meno importante.

Tabella 5: Importanza e importanza relativa per il modello ML-RF allenato per predire concentrazioni di PM_{10}

Predittore	Imp (%)	Imp Cum. (%)
c_PM10	0.38	0.38
julianday_x	0.16	0.54
julianday_y	0.16	0.7
c_NO2	0.08	0.78
c_O3	0.05	0.83
LAI	0.04	0.87
D1ST	0.02	0.89
dayofweek_x	0.02	0.91
D2ND	0.02	0.93
EUPOP	0.01	0.94
ELEV	0.01	0.95
LAN	0.01	0.96
dayofweek_y	0.01	0.97

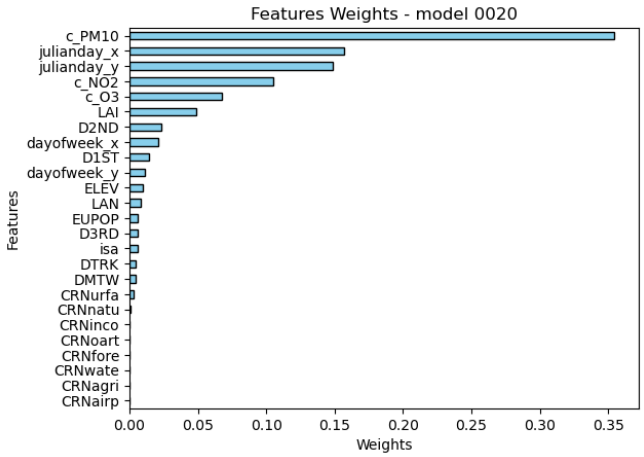


Tabella 6: Importanza e importanza relativa per il modello ML-RF allenato per predire concentrazioni di $PM_{2.5}$

Predittore	Imp (%)	Imp Cum. (%)
julianday_y	0.25	0.25
julianday_x	0.23	0.48
c_O3	0.22	0.70
c_PM25	0.12	0.82
c_NO2	0.03	0.85
dayofweek_x	0.02	0.87
LAI	0.02	0.89
D1ST	0.01	0.90
ELEV	0.01	0.91
EUPOP	0.01	0.92
dayofweek_y	0.01	0.93
D2ND	0.01	0.94

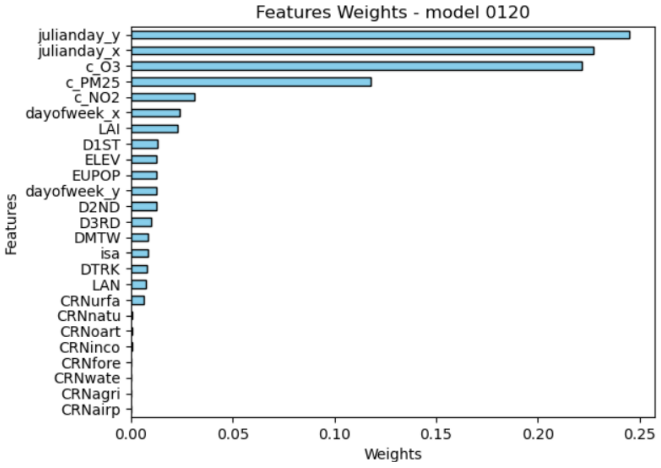


Tabella 7: Importanza e importanza relativa per il modello ML-RF allenato per predire concentrazioni di O_3

Predittore	Imp (%)	Imp Cum. (%)
c_O3	0.48	0.48
julianday_y	0.15	0.63
hours_x	0.1	0.73
hours_y	0.07	0.8
julianday_x	0.06	0.86
c_PM10	0.02	0.88
LAI	0.02	0.9
c_NO2	0.02	0.92
ELEV	0.01	0.93
D2ND	0.01	0.94
dayofweek_x	0.01	0.95
CRNfore	0.01	0.96

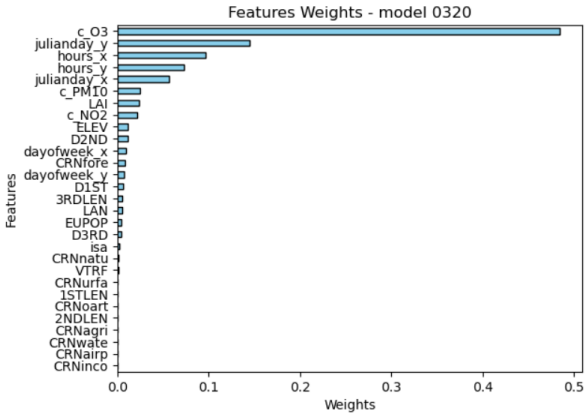
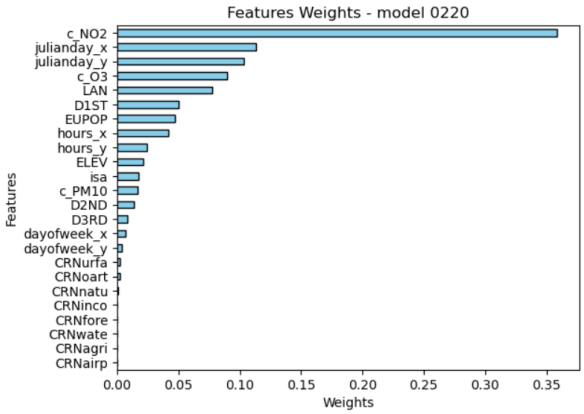


Tabella 8: Importanza e importanza relativa per il modello ML-RF allenato per predire concentrazioni di NO_2

Predittore	Imp (%)	Imp Cum. (%)
c_NO2	0.36	0.36
julianday_x	0.11	0.47
julianday_y	0.1	0.57
c_O3	0.09	0.66
LAN	0.08	0.74
D1ST	0.05	0.79
EUPOP	0.05	0.84
hours_x	0.04	0.88
hours_y	0.02	0.9
ELEV	0.02	0.92
isa	0.02	0.94
c_PM10	0.02	0.96



Per ogni specie analizzata è stata calcolata una mappa di "copertura", che fornisce una stima quantitativa e aggregata dell'affidabilità del modello RF_{GRID} sul grigliato target. La mappa viene calcolata giorno per giorno per il particolato e ora per ora per gli inquinanti gassosi. Per ogni cella del grigliato target e per ogni predittore, si assegna un valore pari all'importanza del predittore se questo appartiene al range di distribuzione campionata dalla rete osservativa e zero viceversa. Quindi, per ogni cella, si sommano tali valori. Al più, se tutti i valori di ogni predittore nella cella sono stati "visti" anche dalla rete osservativa, l'indice di copertura sarà 1, mentre in caso contrario avrà un valore inferiore ad 1 e proporzionale all'importanza dei predittori campionati. La mappa complessiva per ogni inquinante si ottiene mediando su tutte le mappe nel periodo.

A titolo d'esempio, la Figura 14 mostra di copertura per la predizione di PM_{10}, in cui si evince che nella zona del parco della Mandria a nord ovest l'affidabilità del modello RF_{GRID} è leggermente inferiore rispetto a Torino. Ciò riflette l'intuizione che, trattandosi di una zona non urbana, la rete osservativa non campiona completamente i valori assunti in loco dai predittori più importanti (c_PM10 e c_NO2 di FARM, LAI).

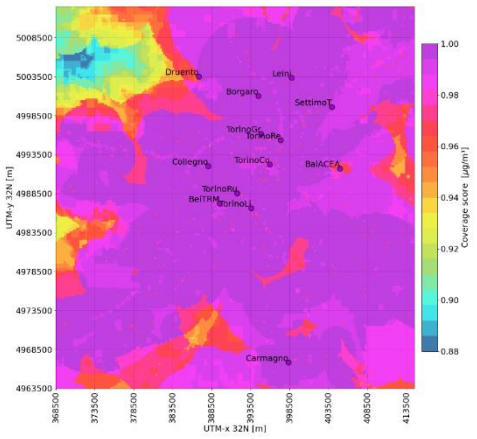


Figura 14. **PM_{10}: Mappa di copertura dei predittori sul grigliato target rispetto al campionamento rete osservativa per il modello RF_{GRID} allenato**

3. Valutazione della qualità dell'aria

Di seguito sono presentate le mappe che illustrano gli standard di qualità dell'aria definiti dal D. Lgs. n. 155, per gli inquinanti considerati nello studio, ossia: PM_{10}, $PM_{2.5}$, NO_2, ed O_3. Questi standard sono generati tramite il sistema modellistico con una risoluzione spaziale di 1 km e ottenuti mediante l'utilizzo del modello RF_{GRID} per PM_{10}, NO_2, ed O_3, e tramite RF_{ST} ed RF_{GRID} per $PM_{2.5}$, producendo mappe ad una risoluzione di 200m.

3.1. PM_{10}

Per l'inquinante in esame, il modello di RF_{GRID} svolge una doppia funzione: da un lato, incrementa la risoluzione spaziale da 1km del modello FARM a 200m e, dall'altro, utilizza il machine learning per integrare i dati osservati e correggere il bias negativo del modello.
Il confronto tra le mappe annuali prodotte dal modello ML-RF e quelle del modello FARM è riportato in Figura 15 per l'intero dominio e in Figura 16 con un focus dettagliato sull'area urbana di Torino.
Nella mappa prodotta dal RF_{GRID}, si osservano medie annuali di PM_{10} comprese tra 25 e 40 µg/m^3 sia nell'area urbana di Torino che nei vari centri abitati disseminati nell'area circostante. Le concentrazioni nelle zone collinari a est e settentrionali del Parco La Mandria variano invece tra 10 e 20 µg/m^3. Inoltre, come descritto nella Sezione 2.2, l'area nord-ovest è caratterizzata da una minore copertura spaziale e pertanto risulta più sensibile agli errori del modello.
Nel dettaglio dell'area urbana di Torino, con una sovrapposizione della mappa degli edifici, si nota come il modello attribuisca valori di concentrazione più elevati nelle vicinanze delle strade secondarie. Questo fenomeno è strettamente legato alle stazioni di *Torino Rebaudengo*, *Collegno - Francia* e *Torino Grassi*, come precedentemente documentato nella Sezione 2, e che correla i valori osservati di PM_{10} sia al predittore di concentrazione di FARM, *c_NO₂*, che al predittore di distanza dalle strade secondarie *D2ND* e primarie *D1ST*.
Secondo i risultati prodotti dal modello RF_{GRID}, il valore limite giornaliero di 50 µg/m^3, da non superare più di 35 volte in un anno è stato superato nel 27.5 % del dominio in esame, come riportato in Figura 17.

Concentrazioni medie annuali di PM_{10}

Valore limite per la protezione della salute:

40 μg m^{-3}

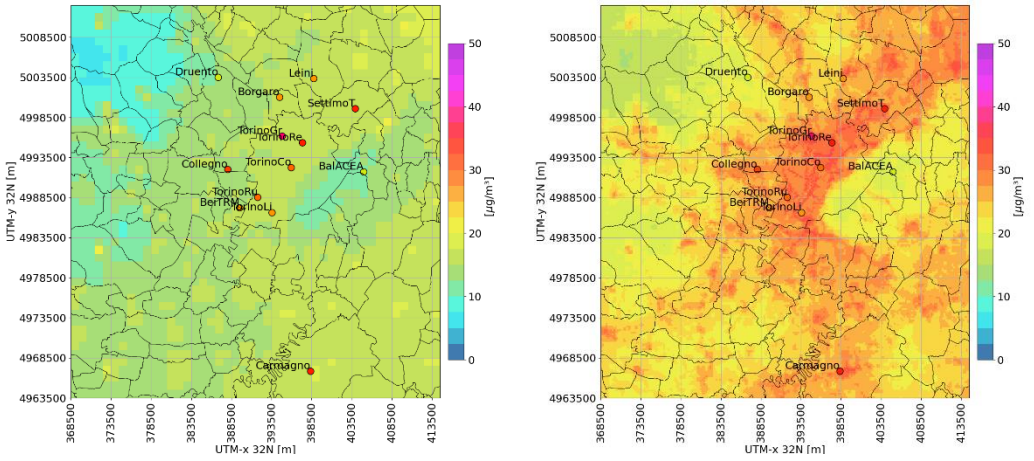


Figura 15. Mappa di concentrazione (a sx campi prodotti da FARM, a dx i campi prodotti dal modello RF_{GRID}). Le linee ricalcano i confini comunali dell'area metropolitana di Torino.

Concentrazioni medie annuali di PM_{10}.

Valore limite per la protezione della salute:

40 μg m^{-3}.

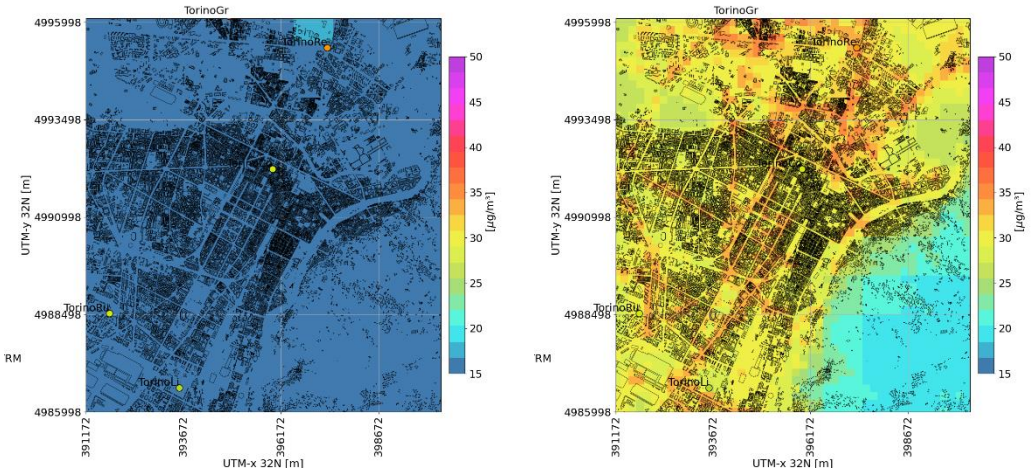


Figura 16 Dettaglio di Torino alle concentrazioni medie annuali (a sx campi prodotti da FARM, a dx i campi prodotti dal modello RF_{GRID})

Numero di superamenti del valore limite per le concentrazioni medie giornaliere di PM_{10}.

Valore limite per la protezione della salute:

50 μg m^{-3}

da non superare più di 35 volte per anno civile.

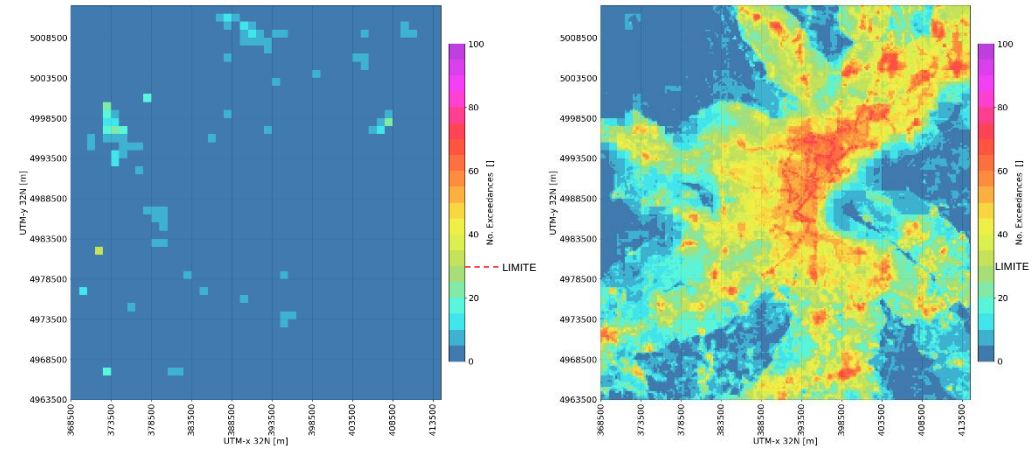
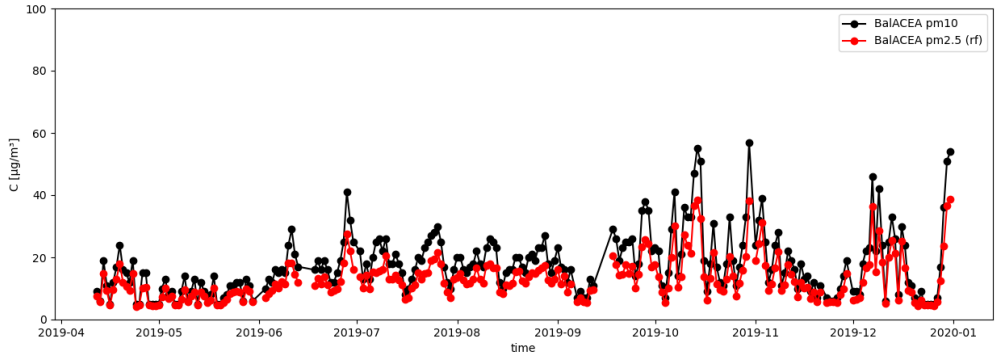


Figura 17. Mappa regionale relativa al numero di superamenti del valore limite per la protezione della salute (a sx campi prodotti da FARM, a dx i campi prodotti dal modello RF_{GRID})

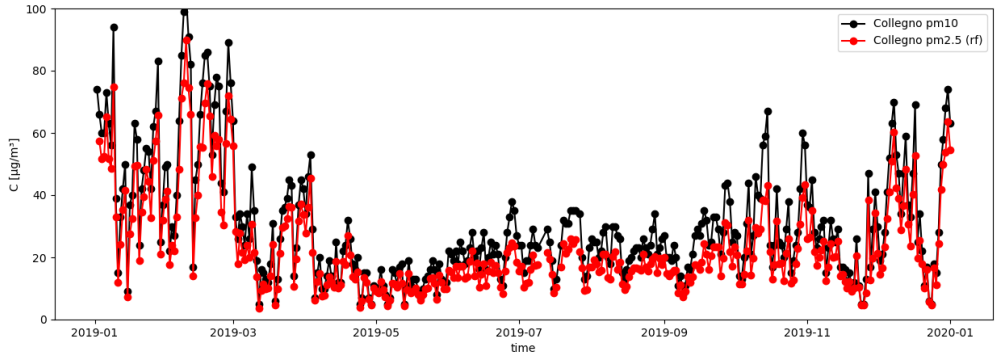
3.2. $PM_{2.5}$

Il modello RF_{ST} è utilizzato per stimare un profilo di concentrazione del $PM_{2.5}$ alle stazioni di Baldissero T. (ACEA), Collegno, Druento – Parco La Mandria, Torino Consolata e Torino Grassi I risultati sono presentati in Figura 18.

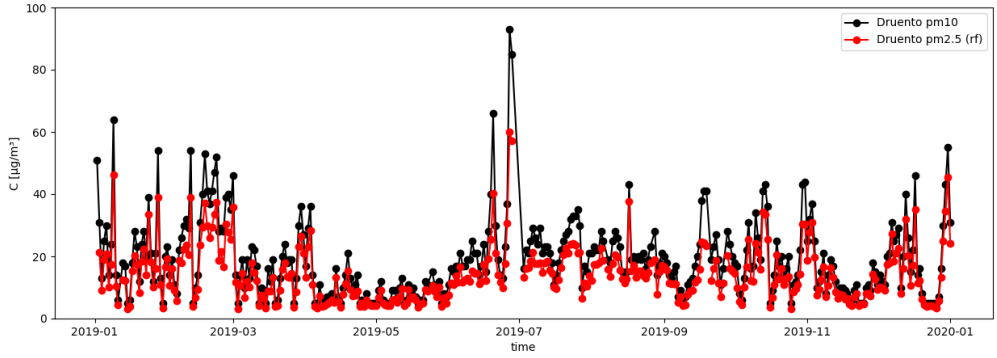
Baldissero T. (ACEA)



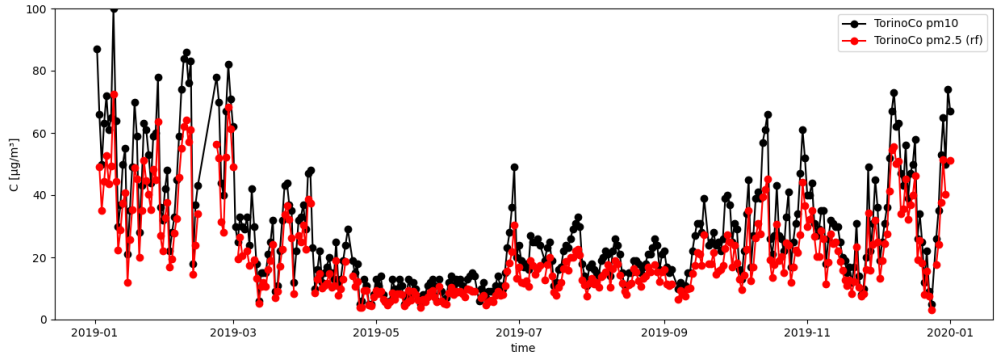
Collegno – Francia



Druento – Parco La Mandria



Torino Consolata



Torino Grassi

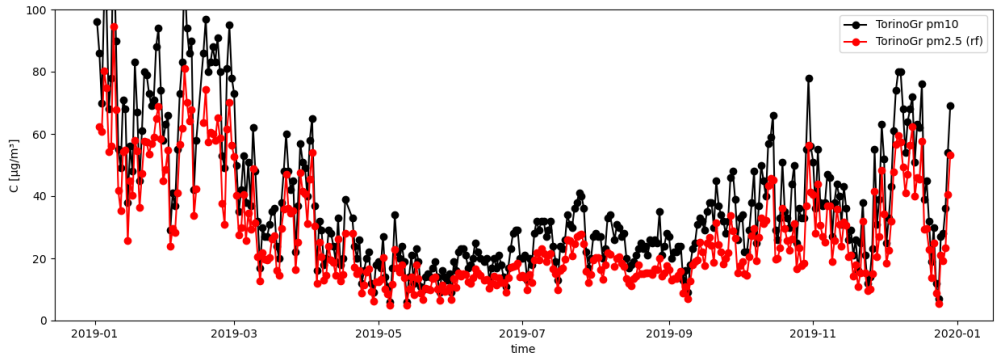


Figura 18. Profilo di concentrazione delle stazioni di PM_{10} e delle rispettive $PM_{2.5}$ surrogate.

Le concentrazioni stimate vengono poi integrate nel dataset di addestramento del modello RF_{GRID}. È fondamentale sottolineare che bisogna garantire sempre il rapporto tra $PM_{2.5}$ e PM_{10} inferiore ad uno. Qualora le stazioni utilizzate siano le stesse e, in ciascuna di esse, $PM_{2.5}$ risulti sempre minore di PM_{10} (utilizzando quindi dati validati), non è possibile che il modello RF possa estrapolare valori che contraddicono questo vincolo, a patto che si utilizzi lo stesso set di predittori. In altre parole, il modello dovrebbe rispecchiare la coerenza delle relazioni osservate nei dati di addestramento. L'utilizzo del predittore c_$PM_{2.5}$/c_PM_{10} nel modello RF_{ST} rappresenta pertanto un elemento essenziale per assicurare la coerenza e l'affidabilità delle stime del modello.

I risultati osservati per il $PM_{2.5}$ non si discostano da quanto precedentemente analizzato per il PM_{10}. Anche in questo caso, infatti, il modello RF_{GRID} agisce da operatore di de-bias sulle sottostime del modello rispetto alle misure osservate. Il confronto delle concentrazioni medie annuali è presentato in Figura 19.

È interessante notare che l'inclusione delle stazioni surrogate ha un impatto positivo sulla capacità del modello di produrre una mappa coerente con quella presentata per il PM_{10} in Figura 15. Dall'analisi comparativa tra i due modelli evidenziata in Figura 20, emerge chiaramente che in assenza delle stazioni surrogate, il modello mostra una tendenza ad attribuire concentrazioni di $PM_{2.5}$ più elevate nelle zone periferiche di Torino rispetto alle aree urbane. Questo risultato è in netto contrasto con la mappa di PM_{10} presentata in Figura 15.

Questa discrepanza può essere attribuita direttamente all'assenza delle stazioni rurali di Druento – Parco La Mandria e Baldissero T. (ACEA) da un lato, e dall'alto livello di $PM_{2.5}$ osservati in Leini (ACEA) - Grande Torino nella zona urbana, dall'altro.

Concentrazioni medie annuali di $PM_{2.5}$

Valore limite per la protezione della salute:

25 μg m^{-3}

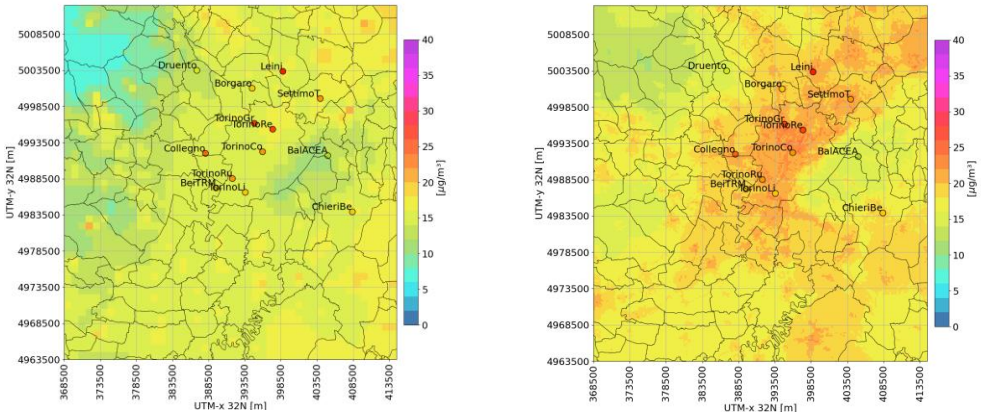


Figura 19. Concentrazioni medie annuali di $PM_{2.5}$. Anno: 2019. Sinistra: modello FARM. Destra: modello RF_{ST}+RF_{GRID}. Le linee ricalcano i confini comunali dell'area metropolitana di Torino e dintorni. La rete osservativa è colorata con le concentrazioni medie annuali di ciascuna stazione.

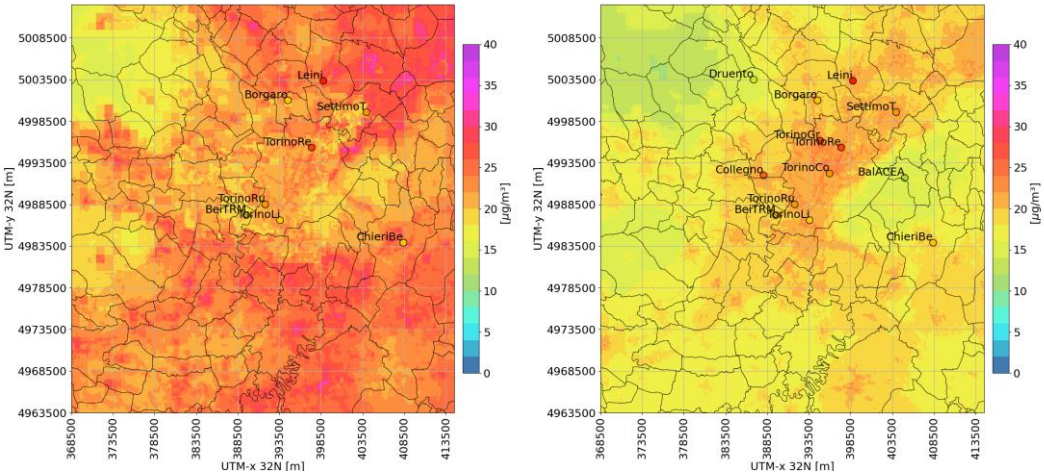


Figura 20 Concentrazioni medie annuali di $PM_{2.5}$. Anno: 2019. Sinistra: modello RF_{GRID}. Destra: modello RF_{ST} + RF_{GRID}. Le linee ricalcano i confini comunali dell'area metropolitana di Torino e dintorni. La rete osservativa è colorata con le concentrazioni medie annuali di ciascuna stazione.

3.3. NO_2

La Figura 21 presenta le mappe delle concentrazioni medie annuali di NO_2, sia per il modello FARM che per il RF_{GRID}. Dal confronto emerge che il RF_{GRID} mostra livelli più elevati di NO_2, attribuibili a valori misurati più elevati alle stazioni di Torino Rebaudengo, in corso Vercelli e Torino Consolata. Ciò si riflette in un aumento delle concentrazioni lungo le principali arterie stradali come corso Grosseto, via Regina Margherita e corso Vittorio Emanuele, come evidenziato nel dettaglio dell'area urbana in Figura 22.

Concentrazioni medie dei massimi giornalieri di NO_2.

Valore limite per la protezione della salute:

40 μg m^{-3}

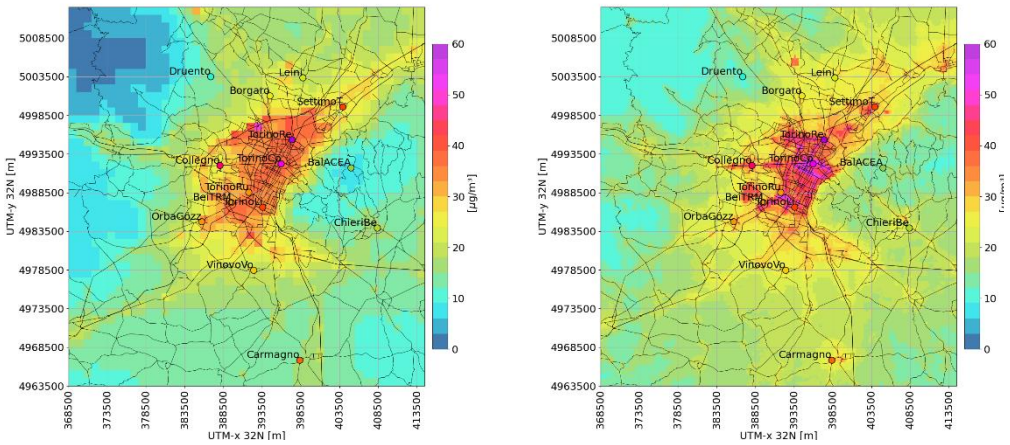


Figura 21. Concentrazioni medie annuali di NO_2. Anno: 2019. Sinistra: modello FARM. Destra: modello RF_{GRID}. Le linee ricalcano la rete di traffico dell'area metropolitana di Torino. La rete osservativa è colorata con le concentrazioni medie annuali di ciascuna stazione.

Concentrazioni medie annuali di NO_2.

Valore limite per la protezione della salute:

40 μg m^{-3}

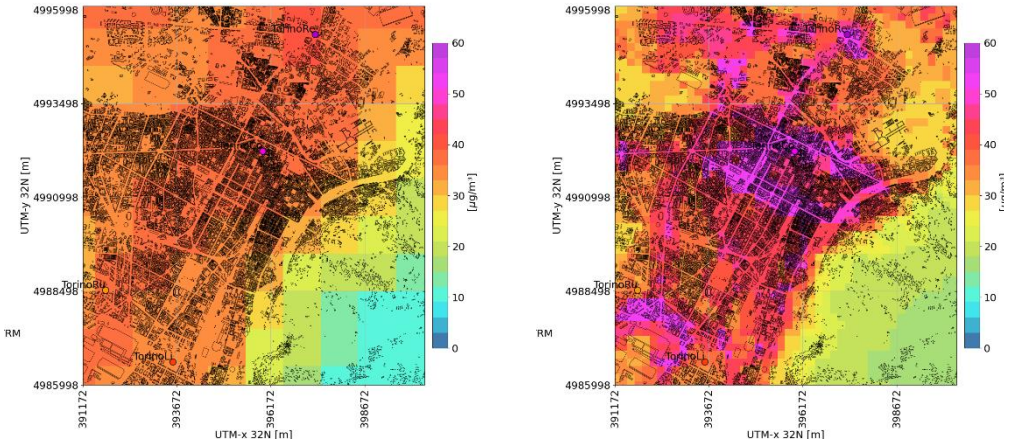


Figura 22. Concentrazioni medie annuali di NO2 dettaglio area urbana Torino. Anno: 2019. Sinistra: modello FARM. Destra: modello RF_{GRID}). Mappa degli edifici in overlay. La rete osservativa è colorata con le concentrazioni medie annuali di ciascuna stazione.

3.4. O_3

La Figura 23 presenta il 93.2 percentile del massimo giornaliero su 8 ore di O_3, sia per il modello FARM che per il RF_{GRID}. La sovrapposizione della rete stradale alle concentrazioni dei due modelli fornisce una visualizzazione chiara della dinamica dell'ozono in relazione all'attività antropica nelle aree urbane, che essendo più ricche di NO (emissione da traffico, riscaldamento e altre sorgenti antropiche) portano ad un consumo di ozono per effetto del cosiddetto "ozone titration". In generale, FARM mostra una tendenza a stimare concentrazioni di O_3 più elevate rispetto al modello RF_{GRID}.

Concentrazioni ≥ 120 μg /m^3 indicano un numero di superamenti superiori o pari a 26. L'intera area di interesse supera o raggiunge tale limite.

93.2 percentile delle concentrazioni medie massime giornaliere di 8 ore

Valore obiettivo per la protezione della salute:

120 μg m^{-3}

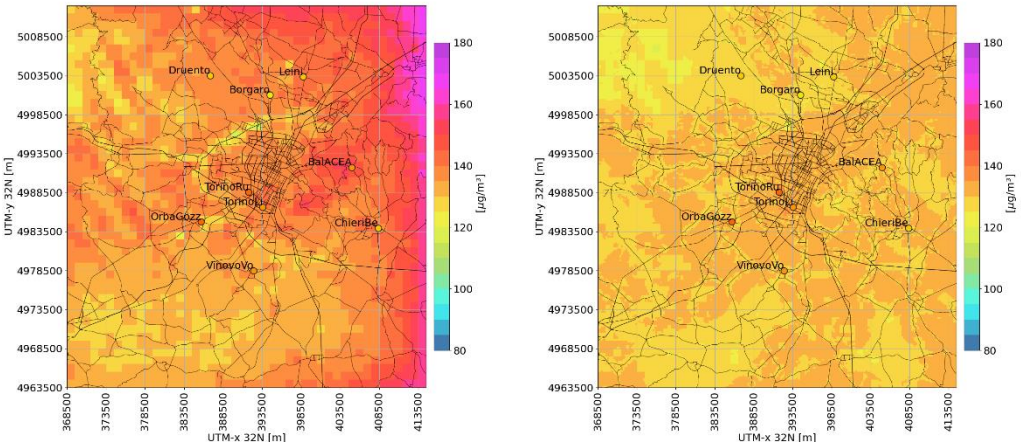


Figura 23. 93.2 percentile delle concentrazioni medie massime giornaliere di O_3 su 8 ore. Anno: 2019. Sinistra: modello FARM. Destra: modello RF_{GRID}. Le linee ricalcano la rete di traffico dell'area metropolitana di Torino. La rete osservativa è colorata con le concentrazioni medie annuali di ciascuna stazione.

4. Conclusioni

Nel presente studio sono mostrati i risultati di elaborazioni effettuate, mediante l'applicazione del modello RF_{GRID} al fine di poter produrre mappe relative agli standard di qualità dell'aria sull'area urbana di Torino per l'anno 2019 dei seguenti inquinanti: biossido di azoto (NO_2), ozono (O_3), PM_{10} e $PM_{2.5}$. Per quest'ultimo è stato utilizzato un modello di RF, RF_{ST} per stimare i profili di concentrazione alle stazioni dove sono disponibili misure di PM_{10} ma non di $PM_{2.5}$.
Tale modulo è un'evoluzione del modello precedentemente utilizzato per la produzione di mappe ad 1 km di risoluzione sul dominio regionale. In questo studio, sono stati utilizzati i campi di concentrazione prodotti dal sistema di previsione di qualità dell'aria messo a punto da ARPA Piemonte (http://www.sistemapiemonte.it/ambiente/srqa/) alla risoluzione spaziale di 1 km. Tali campi, unitamente ad altri predittori spaziali e spazio-temporali, scelti per catturare le peculiari fluttuazioni temporali e spaziali dei campi di concentrazione, sono stati elaborati dall'algoritmo Machine learning Random Forest (Breiman, 2001) per produrre corrispondenti mappe di qualità dell'aria sul territorio urbano di Torino, alla risoluzione spaziale target di 200 m.
Tramite la procedura di cross-validazione innestata il modello è stato validato su dati osservati temporaneamente esclusi dall'addestramento, evidenziando una buona capacità di generalizzazione.
I campi di concentrazione prodotti dal sistema di previsione e dal modulo RF_{GRID} sono stati quindi utilizzati per produrre le mappe relative agli standard di qualità dell'aria, definiti dal D. Lgs. n. 155, per gli inquinanti considerati nello studio. L'esame di tali mappe ha evidenziato significative differenze in termini di concentrazioni massime e minime stimate e conferma l'utilità dell'approccio qui descritto a supporto delle attività istituzionali svolte da ARPA.
I risultati dell'analisi evidenziano l'efficacia complessiva della data fusion operata dal modello *machine learning* nel colmare il bias di FARM rispetto alle concentrazioni osservate alle stazioni di PM_{10} e O_3. La mappa ottenuta con RF_{GRID} riflette un miglioramento notevole, presentando concentrazioni mediamente più elevate, soprattutto nell'area urbana di Torino con picchi a 200 m di risoluzione dislocati lungo le aree di traffico.
La predominanza del campo di FARM come primo predittore per importanza si evidenzia nella somiglianza dei pattern di concentrazione a larga scala con la mappa ottenuta tramite RF_{GRID}. Per quanto riguarda il $PM_{2.5}$, l'integrazione delle stazioni surrogate stimate tramite RF_{ST} nei dati di addestramento in RF_{GRID} produce un campo di concentrazione più in linea con quello ottenuto dallo stesso modello per il PM_{10}. Inoltre, l'utilizzo del rapporto delle concentrazioni $c\_PM_{2.5}/c\_PM_{10}$, combinato all'utilizzo dello stesso set di predittori per i due modelli, consente di vincolare i due campi alla relazione $PM_{2.5} < PM_{10}$.
Per l'inquinante NO_2, la data fusion ha portato a stime di concentrazioni più elevate nelle aree urbane ad alto traffico di Torino, un effetto dovuto all'utilizzo dei predittori statici di distanza dalle strade.
Infine, se si analizza l'ozono, le mappe di concentrazioni medie massime giornaliere su 8 ore mostrano una riduzione intuitiva dell'O_3 nelle aree urbane, le quali, a causa dell'abbondanza di NO proveniente da attività antropiche, portano a un consumo di ozono per effetto del

fenomeno cosiddetto “ozone titration”. In generale, si osserva che FARM tende a stimare concentrazioni di O_3 più elevate rispetto al modello $RF_{GRID.}$

I moduli che implementano l’algoritmo di questo studio, unitamente ai predittori utilizzati, sono stati forniti ad ARPA. Il personale ARIANET fornirà il supporto necessario al loro utilizzo mediante sessioni di formazione che prevedono l’applicazione pratica dei moduli forniti.

5. BIBLIOGRAFIA

Arianet, 2019. Valutazione modellistica dello stato di Qualità dell'aria sulla Regione Sardegna per l'anno 2018. ARIANET R2019.05.

ARPAS, 2019. Relazione annuale sulla qualità dell'aria in Sardegna per l'anno 2019. Giugno 2019.

Breiman, L., 2001. Random Forests. *Machine Learning*, **45**, 5–32. https://doi.org/10.1023/A:1010933404324.

Pedregosa, T. et al., 2011. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, **12** , 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News, 2/3, 18–22. URL: https://journal.r-project.org/archive/r-news.html.

Silibello, C., Carlino, G., Stafoggia, M. et al., 2021. Spatial-temporal prediction of ambient nitrogen dioxide and ozone levels over Italy using a Random Forest model for population exposure assessment. *Air Qual Atmos Health*, **14**, 817–829 (2021). https://doi.org/10.1007/s11869-021-00981-4.

Stafoggia, M., Johansson, C., Glantz, P., Renzi, M., Shtein, A., Hoogh, K. de, Kloog, I., Davoli, M., Michelozzi, P., & Bellander, T. (2020). *A random forest approach to estimate daily particulate matter, nitrogen dioxide, and ozone at fine spatial resolution in Sweden*. Atmosphere, 11(3).