simularia

Comparazione di Modelli Machine Learning

Discesa di scala da simulazioni con modello FARM.

A cura di
Giuseppe Carlino

Milano, 30/01/2020

I risultati mostrati sono stati ottenuti grazie alla collaborazione con il **progetto BEEP**.

BEEP
BIGDATA in Epidemiologia ambiEntale ed occuPazionale

In particolare, grazie ai contributi, discussioni, commenti di:

- Claudio Gariazzo (INAIL)
- Massimo Stafoggia, Simone Bucci, Matteo Renzi (DEP Lazio)
- Camillo Silibello et al. (Arianet)
- Rossella Prandi (Simularia)

INAIL
DIPARTIMENTO MEDICINA EPIDEMIOLOGIA IGIENE DEL LAVORO AMBIENTALE (INAIL-DIMEILA)

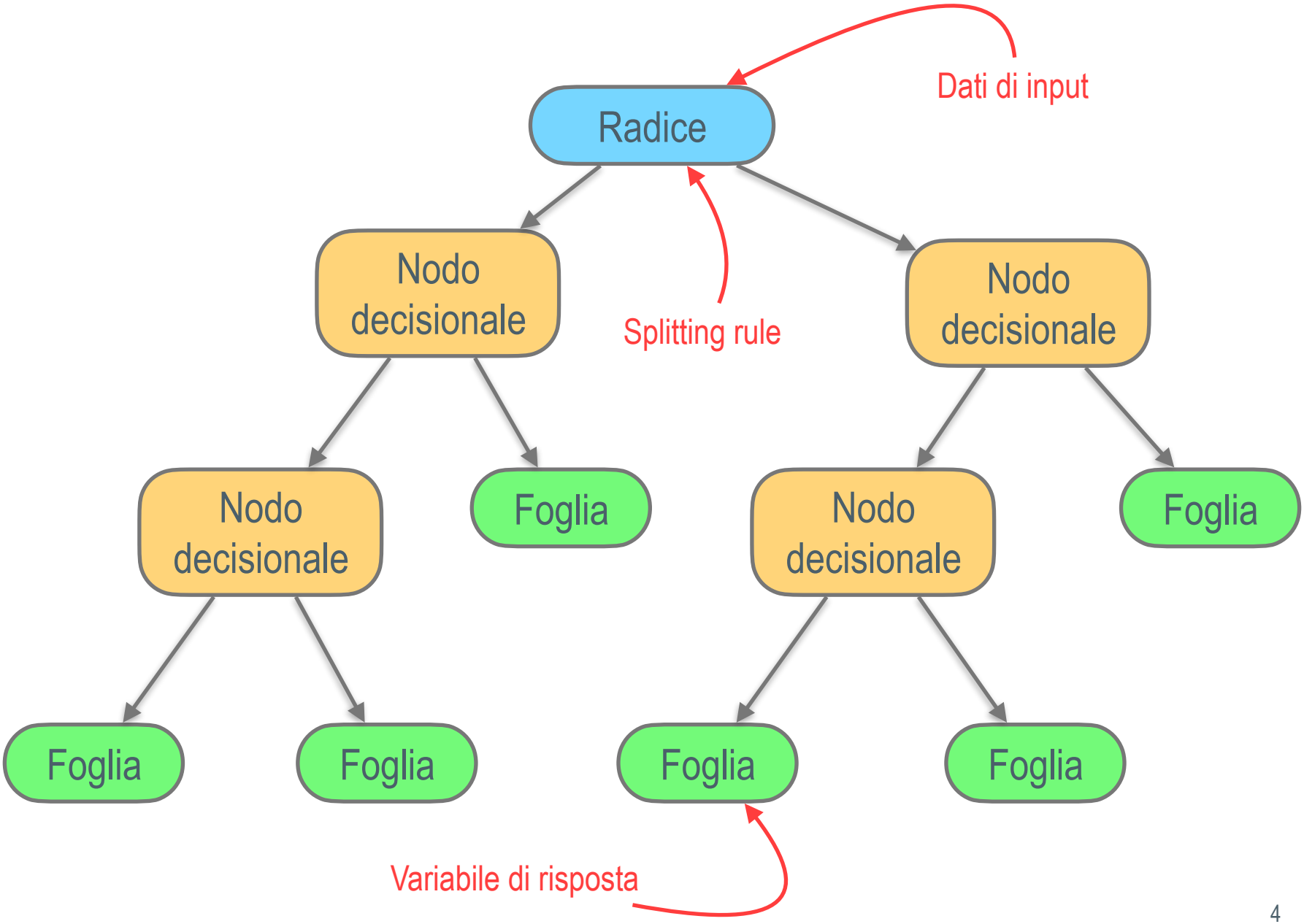
DEP Lazio
Dipartimento di Epidemiologia
Servizio Sanitario Regionale
Regione Lazio

ARIANET

Modelli ad Albero Decisionale

- Gli algoritmi basati sugli **Alberi Decisionali** (AD) sono una classe di modelli Machine Learning (ML) supervisionato.
- Sono tra i più popolari perché intuitivi da capire e da interpretare e sono considerati molto efficaci per **dati tabulari/strutturati**.
- L'obiettivo è addestrare un modello sulla base di osservazioni da usare per **predire una variabile di risposta** dato un insieme di valori di input:
- **Alberi di classificazione** se la variabile di risposta è discreta.
- **Alberi di regressione** se la variabile di risposta è continua.
- L'AD classifica le osservazioni riordinandole dal nodo radice alle foglie attraverso una **bipartizione ricorsiva** delle variabili di input secondo delle **regole di splitting**. Le foglie forniscono la classificazione della **variabile di risposta**.

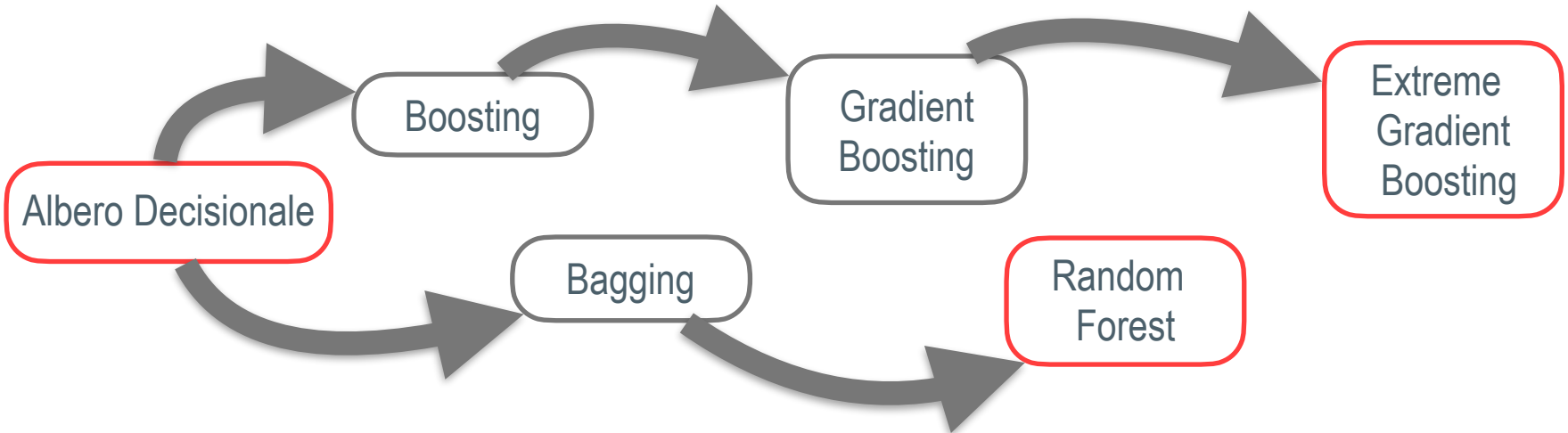
Modelli ad Albero Decisionale



Ensemble di Alberi Decisionali

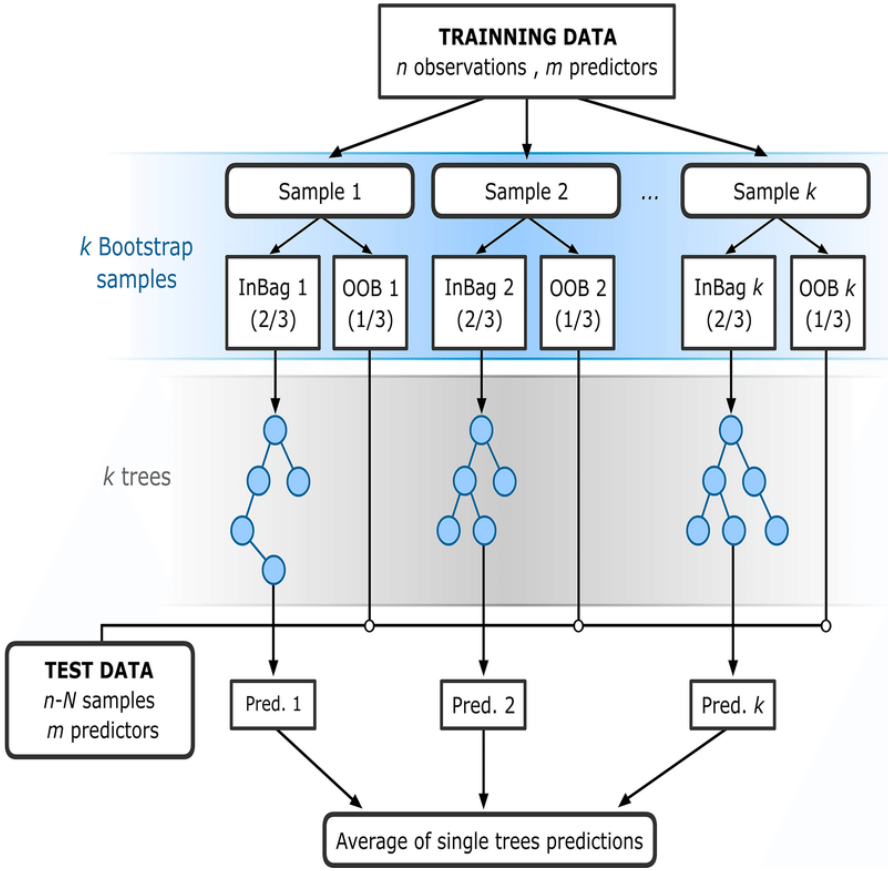
- Problema di **overfitting** degli AD: nel caso estremo avremo un nodo terminale per ogni osservazione del set di dati di apprendimento.
- Gli **ensemble di modelli** forniscono soluzioni più accurate:
  - Ensemble **sequenziali**: i singoli modelli sono costruiti in sequenza premiando i modelli precedenti con performance migliore;
  - Ensemble **paralleli**: i singoli modelli sono costruiti in parallelo e l'accuratezza del modello è migliorata mediando i risultati.

Due (tra tanti) modi di costruire ensemble **omogenei** di modelli:



Random Forest

- Ensemble di AD paralleli relativamente poco correlati tra di loro, costruiti grazie ai seguenti meccanismi:
  - **Bagging**: per ogni AD campionamento delle osservazioni con sostituzione.
  - Ad ogni biforcazione dei nodi viene selezionato un **sottoinsieme casuale** delle variabili di input.
- Il risultato finale si ottiene per votazione (classificazione) o media (regressione) di tutte le risposte dei singoli AD.
- Funziona anche con variabili di input correlate tra di loro.
- Gestisce variabile di input di tutti i tipi (continue, discrete).
- Minimizza l'overfitting e la varianza.



Rodriguez-Galiano, Victor et al. *Modelling interannual variation in the spring and autumn land surface phenology of the European forest*. Biogeosciences 13(11):3305-3317 (2016).

XGBoost: extreme gradient boosting

- XGBoost è una delle ultime evoluzioni più di successo dei modelli basati sugli alberi decisionali.
- Autori: Tianqi Chen e Carlos Guestrin (arXiv 2016).
- Su github: 4025 commits, 389 contributors; 7300 forks.

- **Boosting**: i modelli dell'ensemble sono costruiti sequenzialmente minimizzando gli errori dei modelli precedenti aumentando (boosting) il peso dei modelli che hanno performance migliore.
- **Gradient boosting**: la minimizzazione della funzione di costo è effettuata con l'algoritmo gradient descent.
- **Extreme gradient boosting**: introduce molteplici ottimizzazioni per migliorare la scalabilità per grandi moli di dati e la velocità dell'algoritmo.

Dati di input

Dominio di calcolo:

Risoluzione spaziale	200 m
Risoluzione temporale	1 giorno
Estensione temporale	1 anno (2015)
Numero complessivo di celle	201781125
Numero di predittori	37

Addestramento

NO_2 **28715** misure da **85** centraline (2015).

PM10 **24508** misure da **74** centraline (2015).

Predittori spazio-temporali:

Concentrazioni medie giornaliere di NO_2 (PM10) modellate da **FARM** (intero anno solare 2015)

Normalized Difference Vegetation Index (risoluzione originale 300m, valori decadali)

Giorno giuliano e Giorno della settimana

Mese

Predittori spaziali:

Popolazione residente

Altitudine

Imperviousness: superficie impenetrabile (risoluzione originale 100 m)

Uso suolo CORINE 22 classi (risoluzione originale 100 m)

Codice regione, provincia e comune

Flussi di traffico veicolare da **Open Transport Map** separato per strade principali e secondarie.

Performance dei modelli

- Addestramento effettuato in ambiente **R**:
  - **`caret`** per la regolazione fine dei parametri.
  - **`ranger`** per il training di Random Forest.
  - **`xgboost`** per il training di XGBoost.

	NO_2		PM10	
	R^2	RMSE	R^2	RMSE
Random Forest	0.87	7.88	0.84	7.55
XGBoost	0.92	6.04	0.90	6.12

- Cross validazione “by observations”: misure dalla stessa stazione vengono campionate contemporaneamente nei dataset di testing e training, quindi la performance è sovrastimata.

Importanza delle variabili

- L'importanza misura l'influenza di ciascun predittore sulla variabile di risposta.
- È calcolata in modo esplicito per ogni singolo albero decisionale e mediata su tutti gli alberi dell'ensemble.

NO_2

Random Forest		XGBoost	
Concent. NO_2 FARM	100.0	Concent. NO_2 FARM	100.0
Traffico strade princ.	49.5	Giorno giuliano	53.4
Giorno giuliano	40.5	NDVI	31.2
Imperviousness	34.8	Imperviousness	25.4
NDVI	33.2	Popolaz. residente	23.1
Popolaz. residente	24.2	Traffico strade sec.	22.7
Traffico strade sec.	23.2	Codice regione	22.5
Altitudine	20.2	Altitudine	22.0
Codice comune	18.5	Mese	17.6
Mese	18.0	Traffico strade princ.	15.7

PM10

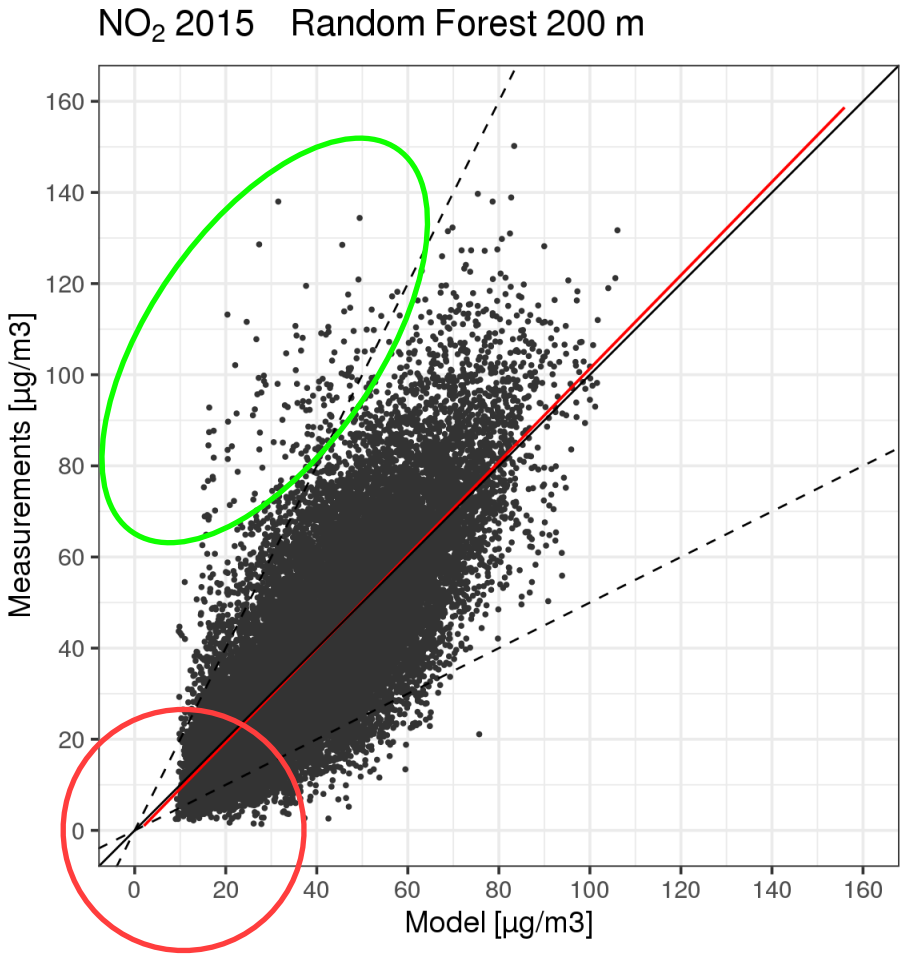
Random Forest		XGBoost	
Conc. PM10 FARM	100.0	Conc. PM10 FARM	100.0
Giorno giuliano	44.7	Giorno giuliano	66.7
Codice regione	29.1	Codice regione	19.3
Altitudine	13.6	Altitudine	12.8
Mese	12.6	Mese	10.0
Codice comune	9.6	Codice provincia	9.9
NDVI	7.0	NDVI	7.9
Codice provincia	6.7	Giorno settimana	6.7
Giorno settimana	6.1	Codice provincia	4.9
Imperviousness	4.3	Imperviousness	4.1

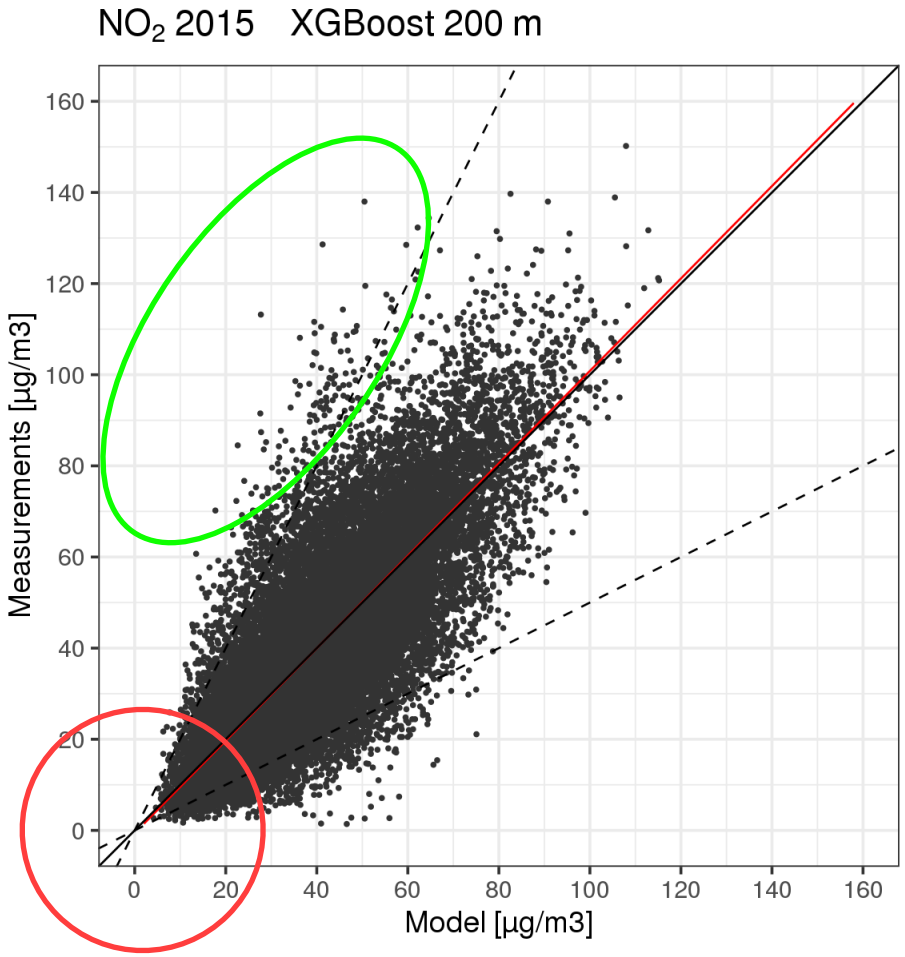
10-fold Cross-Validation “by monitor”

La CV by monitor permette di stimare meglio la performance del modello quando andremo a predire le concentrazioni nelle celle dove non ci sono misure:

1. Le **centraline di misura** vengono suddivise in 10 gruppi a caso.
2. Il dataset di addestramento viene suddiviso nei corrispondenti **10 gruppi**.
3. Il modello ML viene addestrato su **9 gruppi (training)** e si predice sul decimo (testing).
4. Si ripete **10 volte** cambiando gruppo di testing.
5. Si confrontano le misure delle centraline con le predizioni effettuate nel testing.
6. Si riporta la performance del modello in termini di R^2, RMSE, pendenza e intercetta del fit lineare.

10-fold Cross-Validation “by monitor”

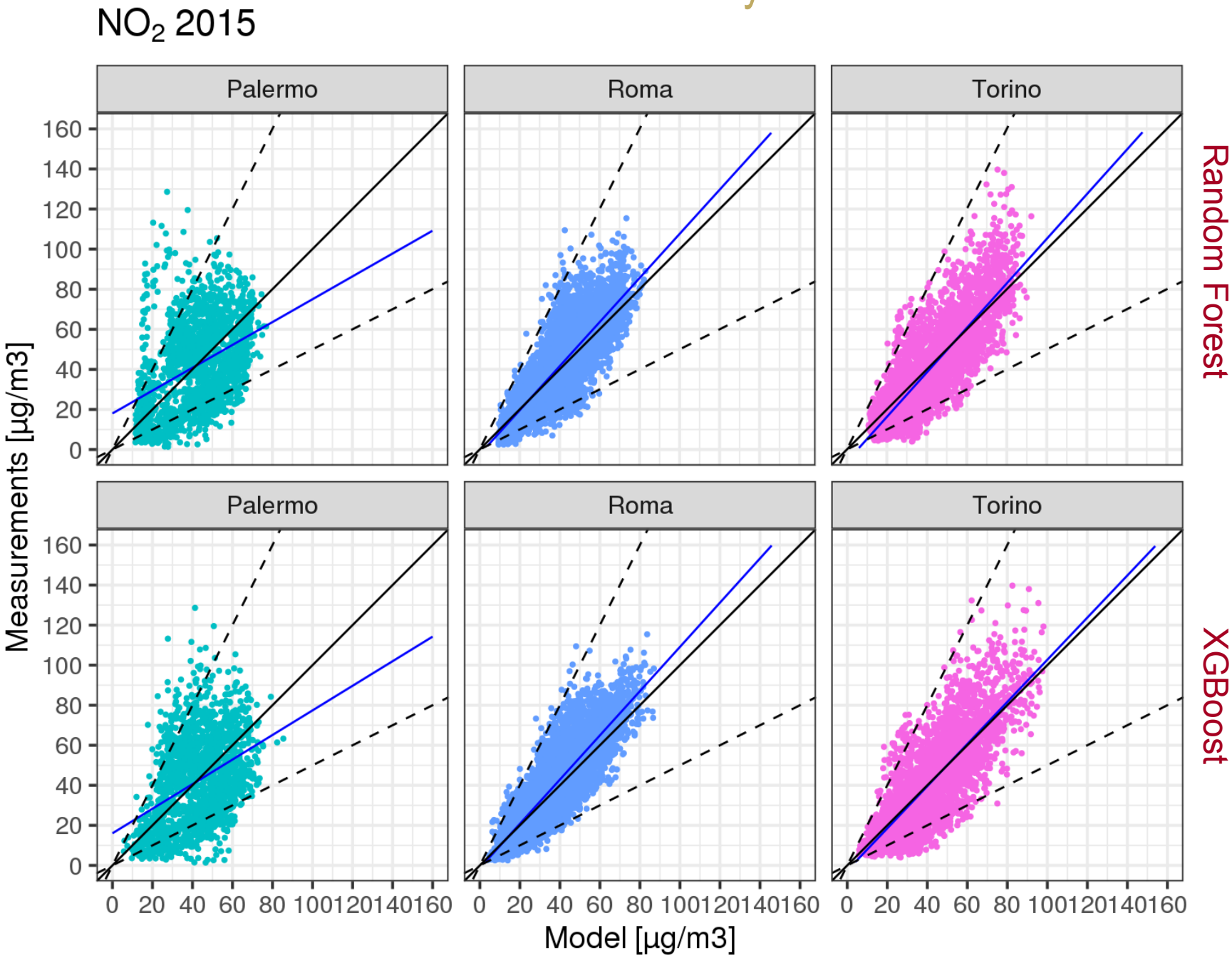




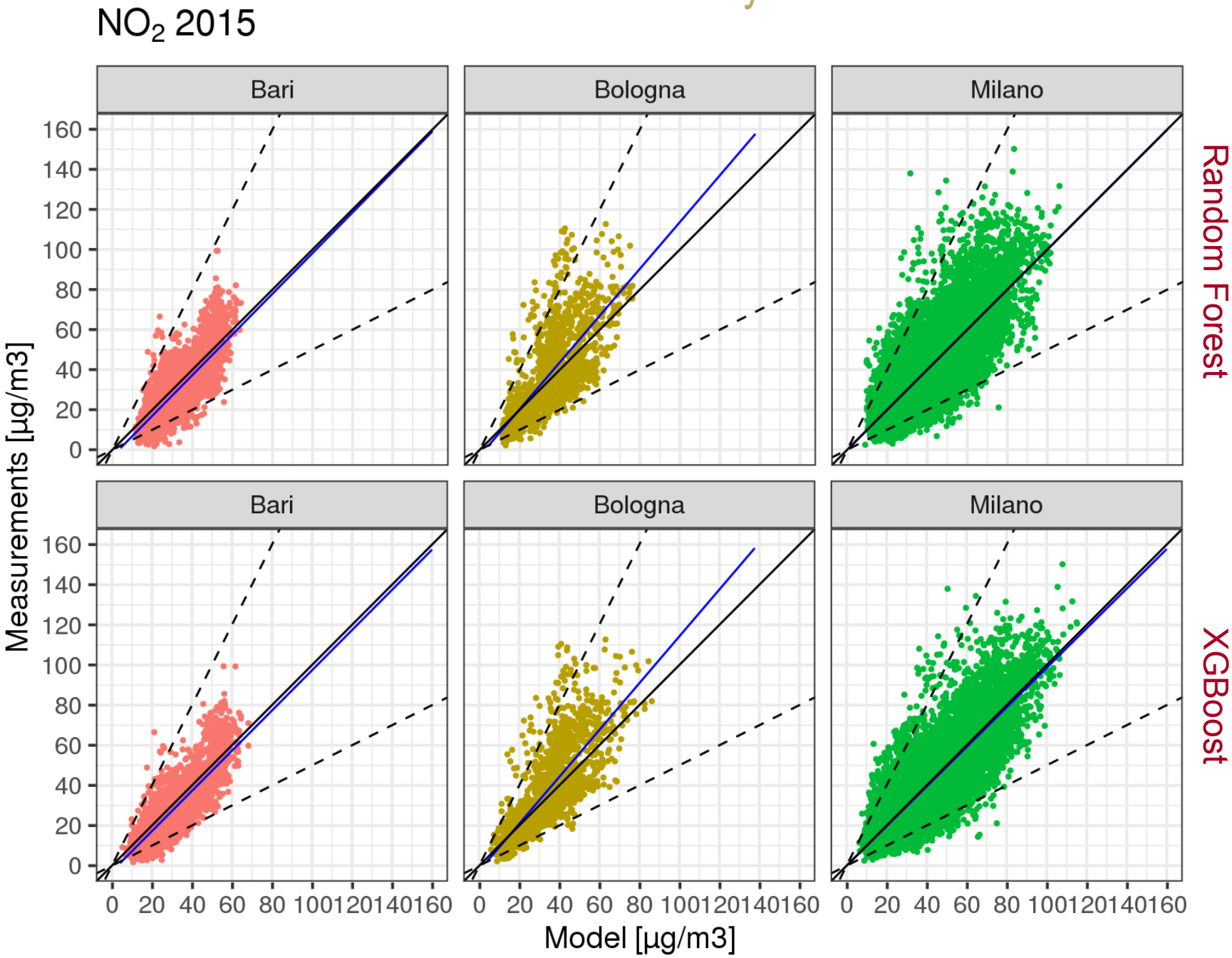
Scatterplot complessivo con tutti le coppie di punti misure/predizione.

	R^2	RMSE	Slope	Intercept
RF	0.62	13.48	1.02	-1.08
XGB	0.65	12.82	1.01	-0.59

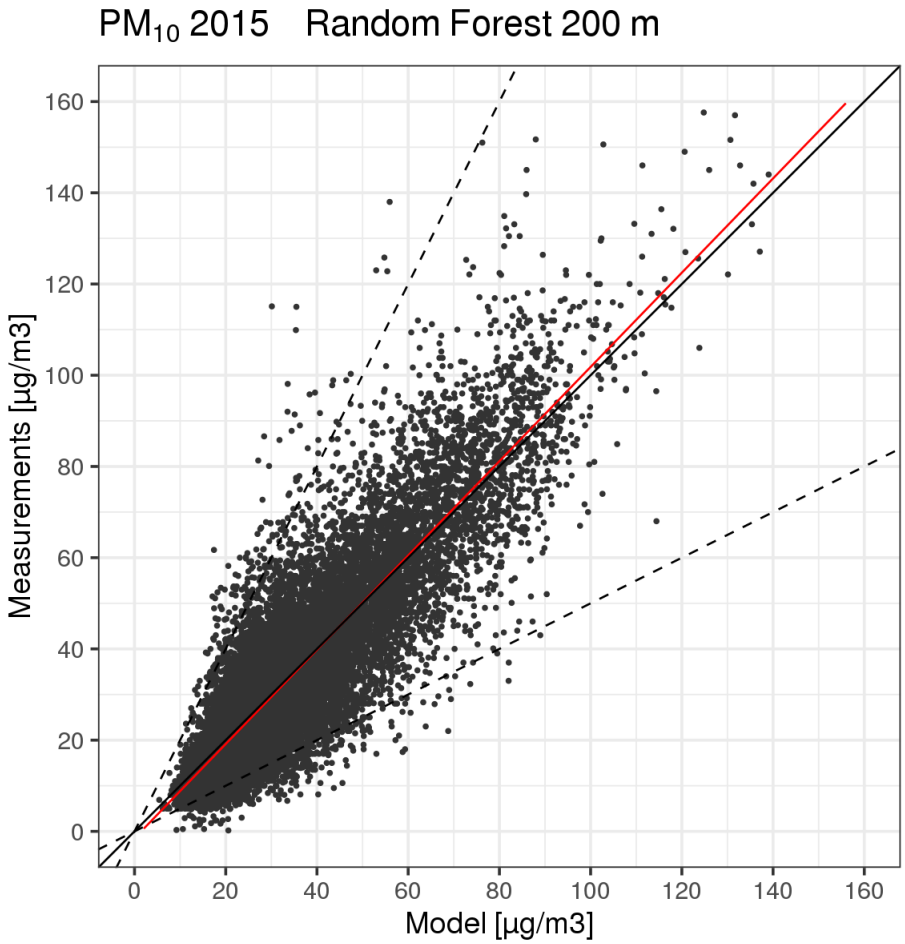
10-fold Cross-Validation “by monitor”

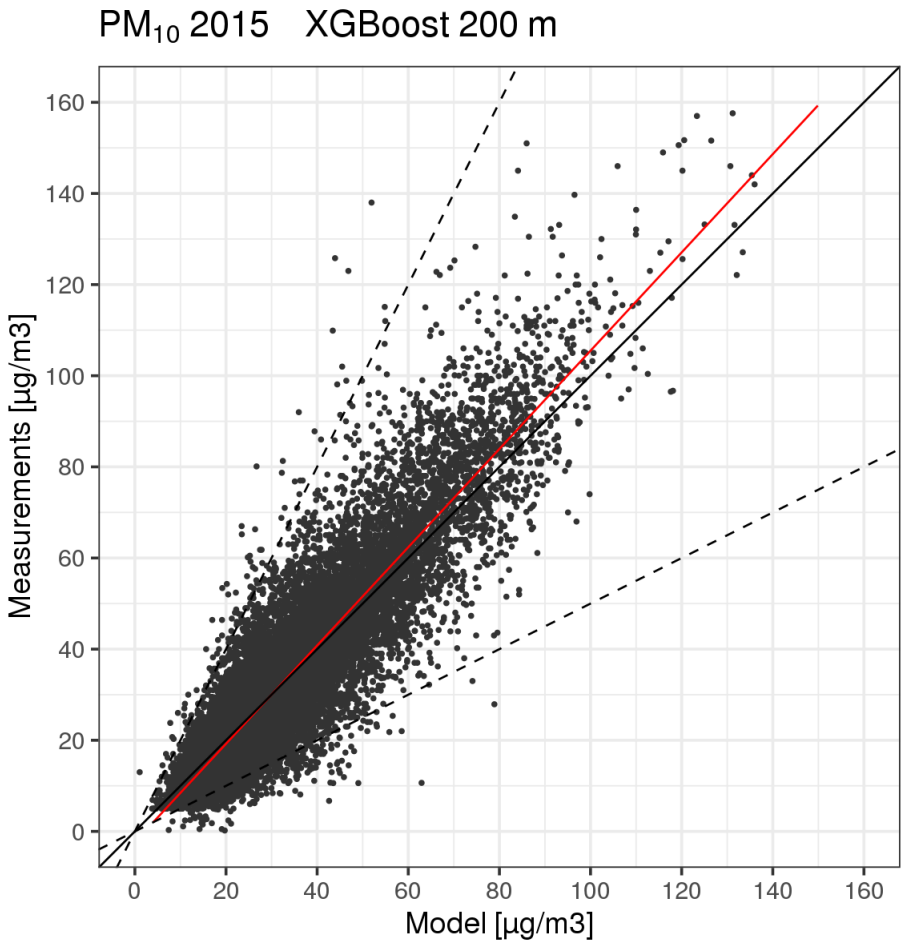


10-fold Cross-Validation “by monitor”



10-fold Cross-Validation “by monitor”

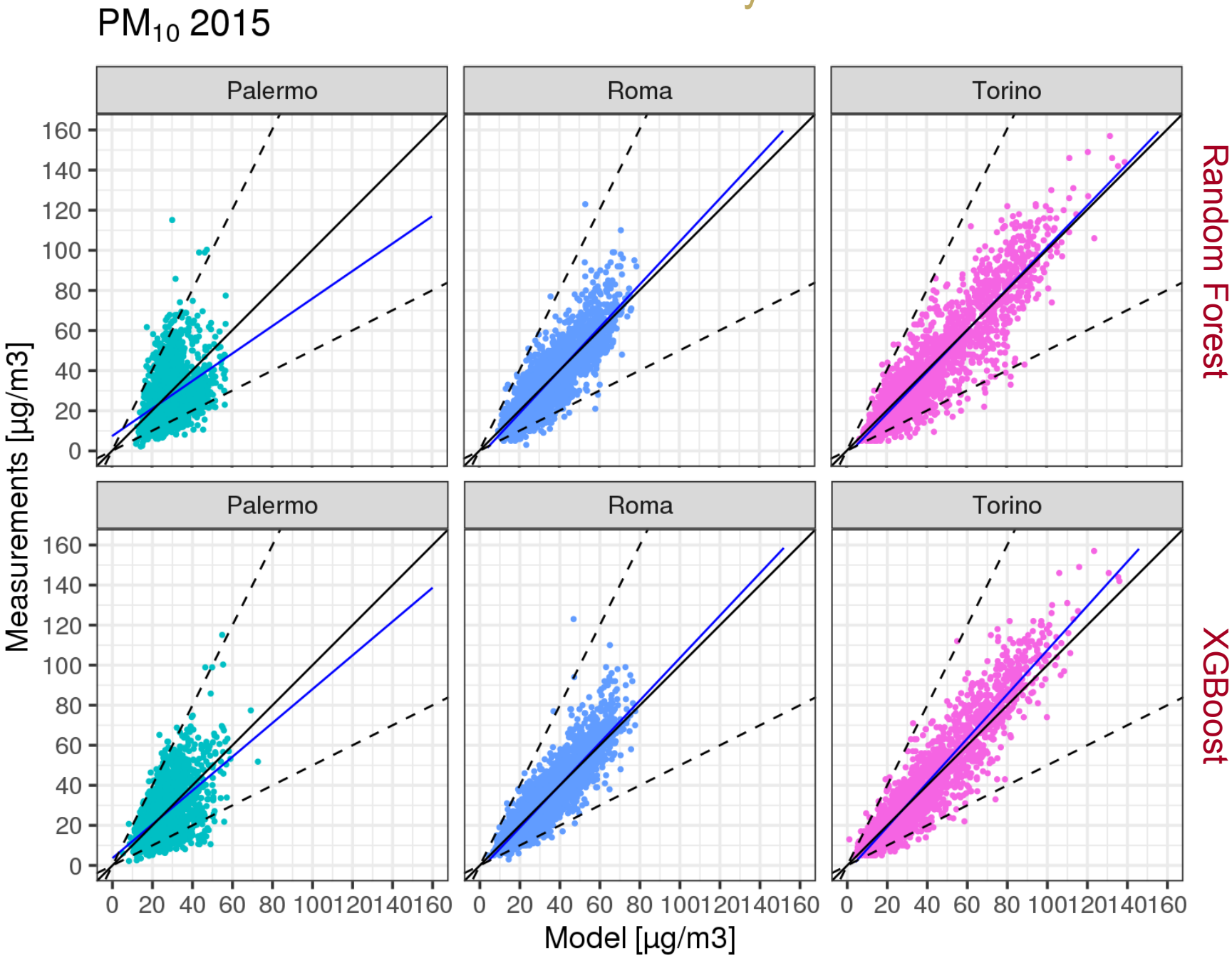




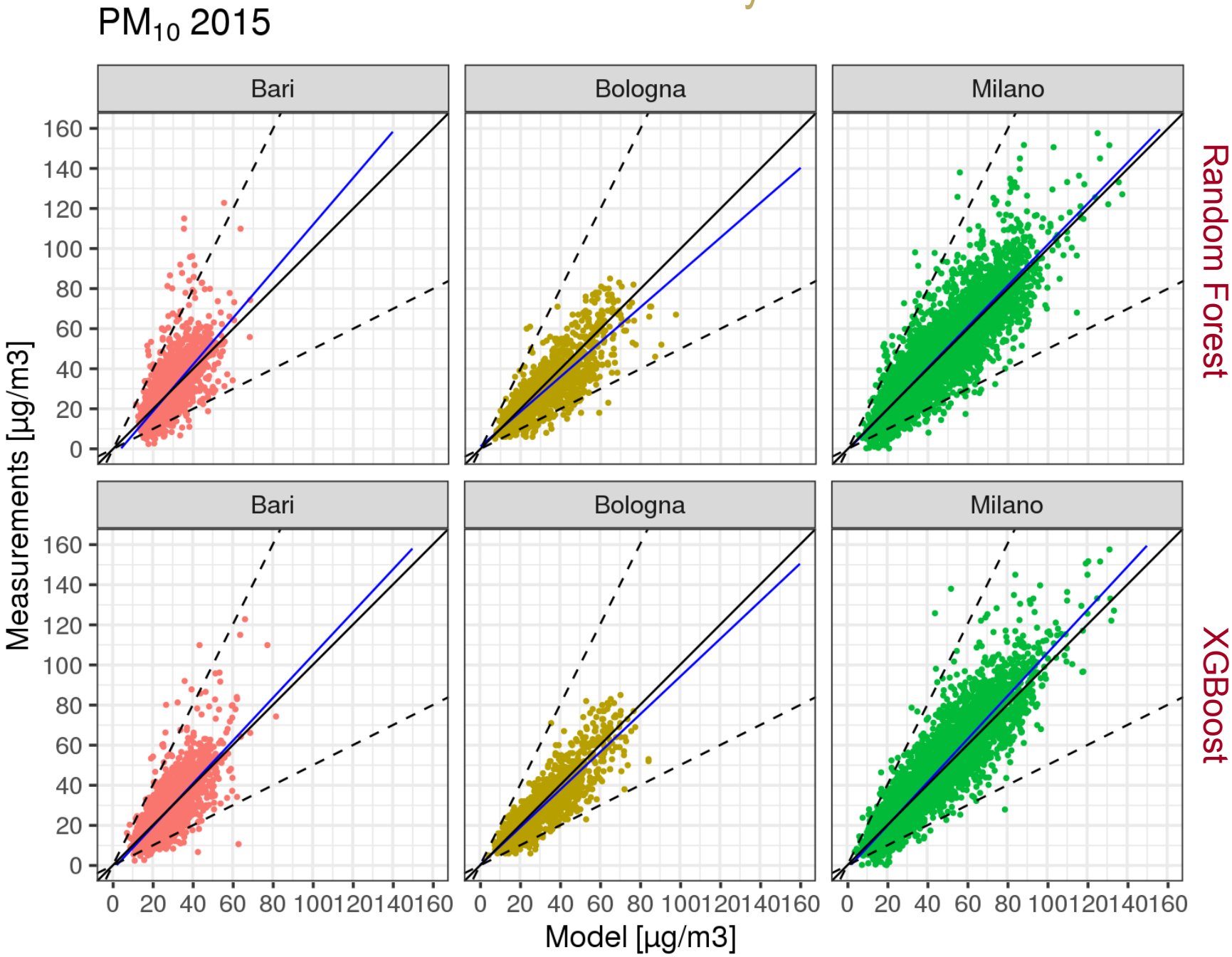
Scatterplot complessivo con tutti le coppie di punti misure/predizione.

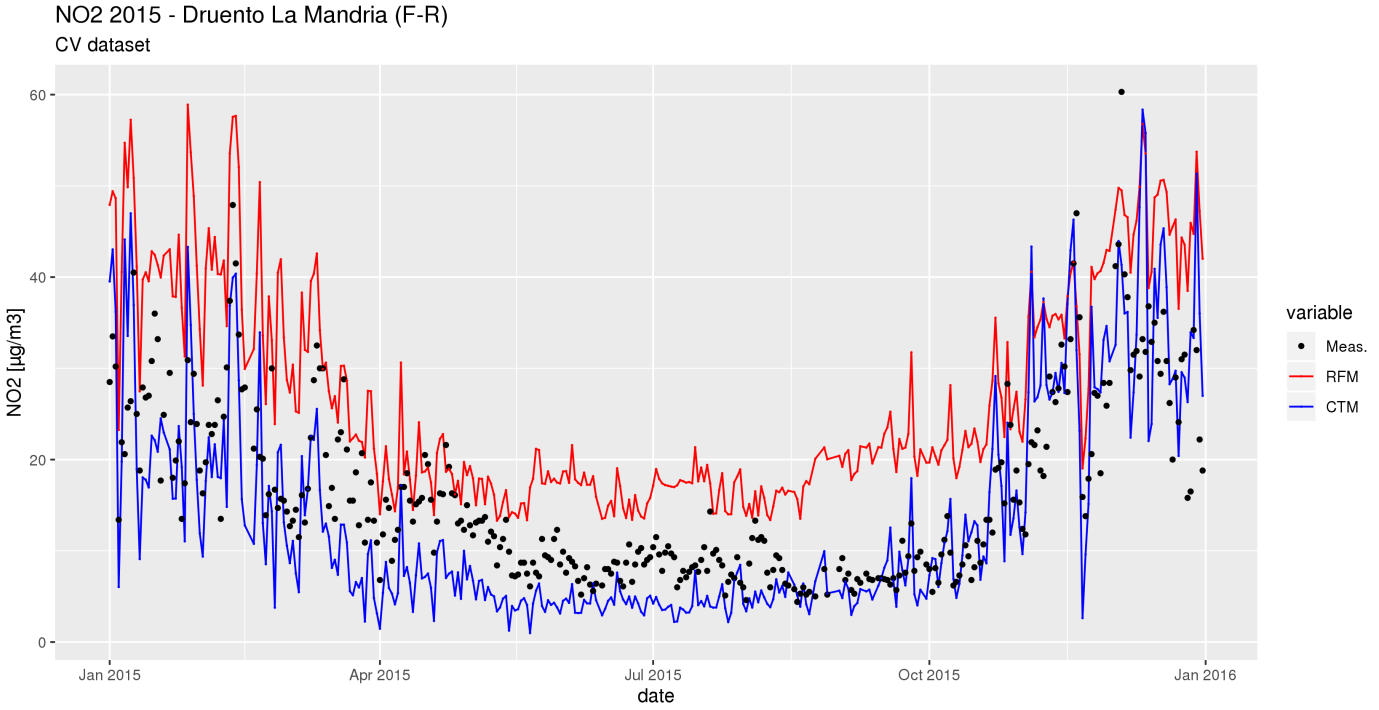
	R^2	RMSE	Slope	Intercept
RF	0.76	9.32	1.03	-1.52
XGB	0.81	8.31	1.08	-2.35

10-fold Cross-Validation “by monitor”

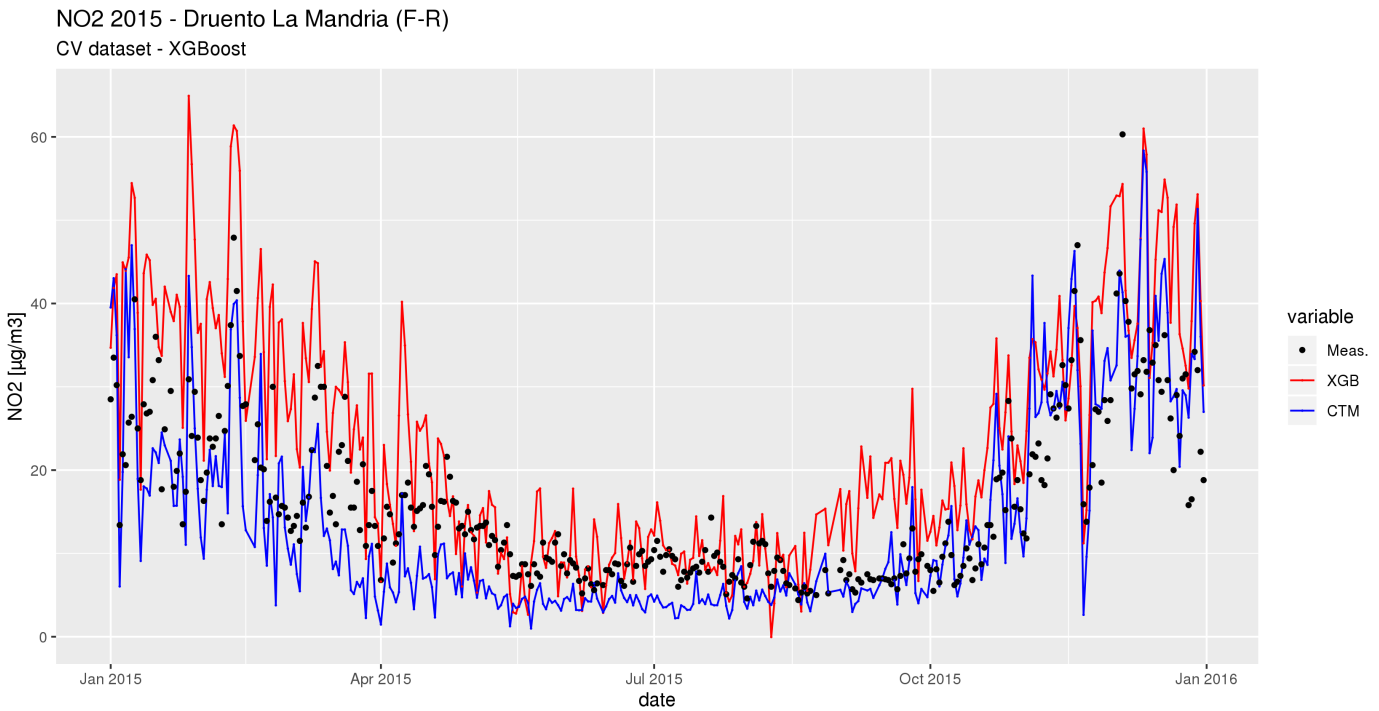


10-fold Cross-Validation “by monitor”

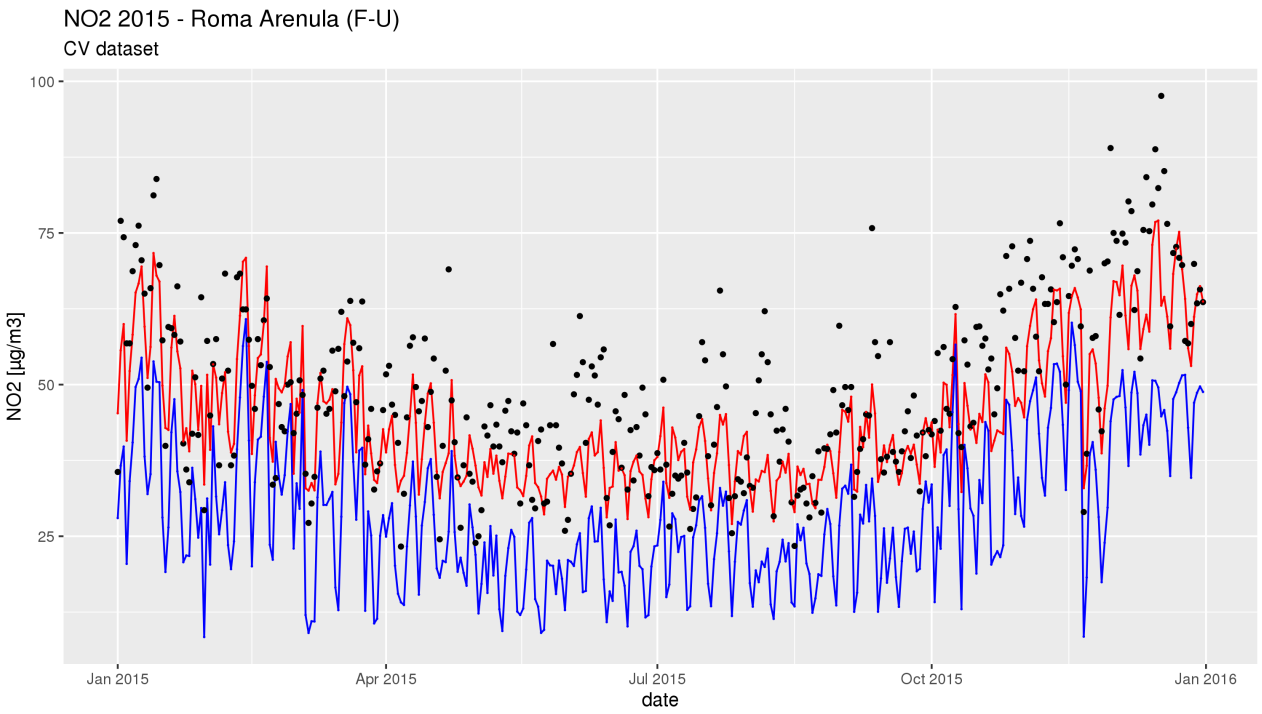




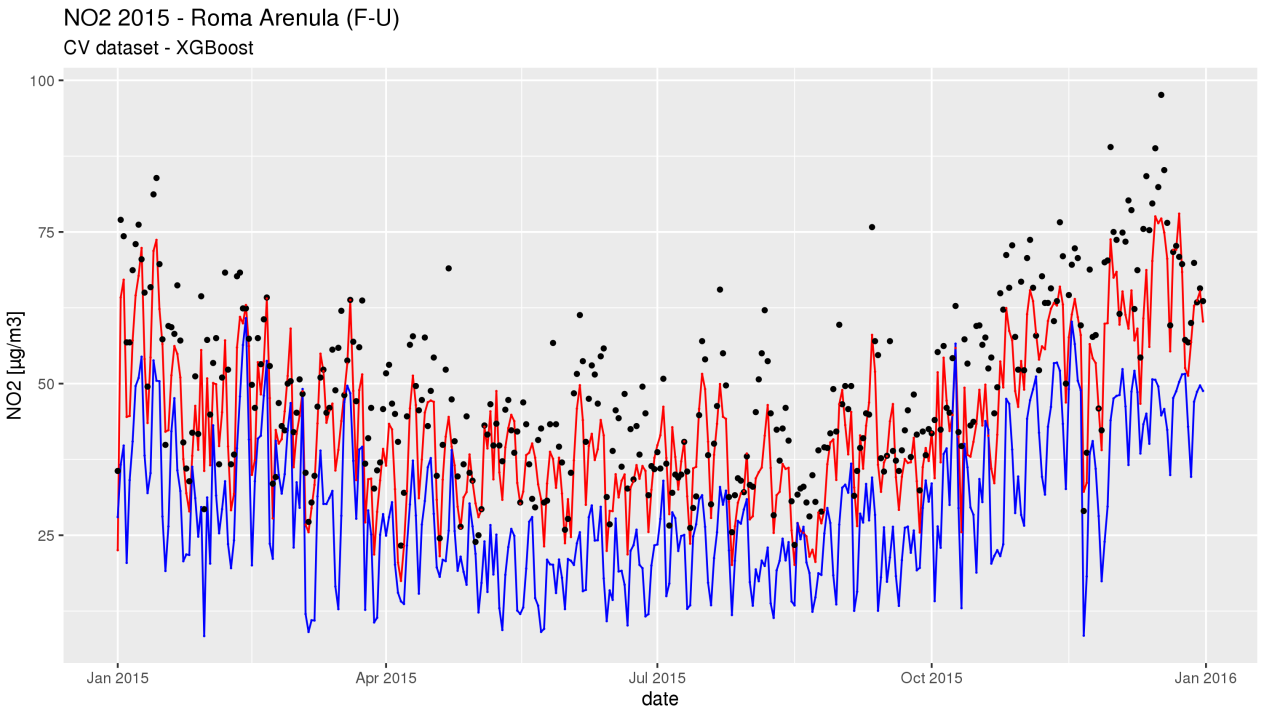
Random Forest



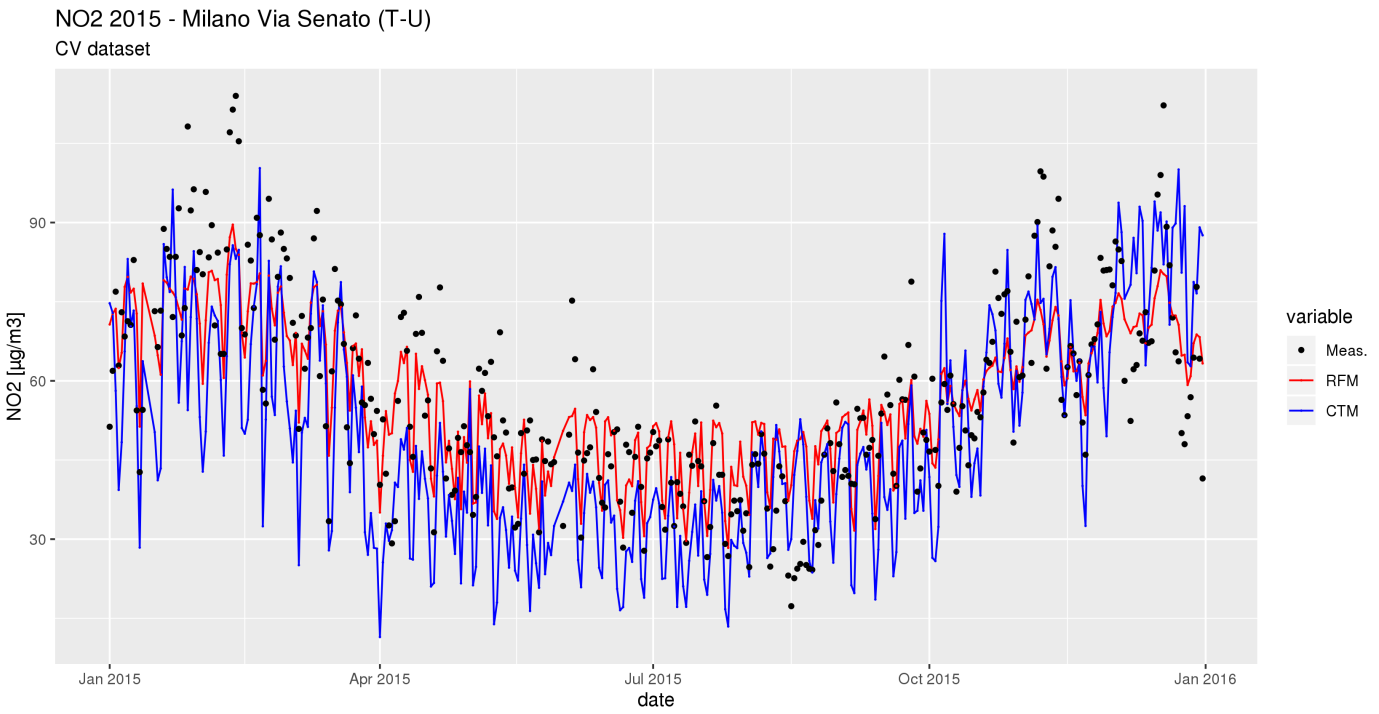
XGBoost



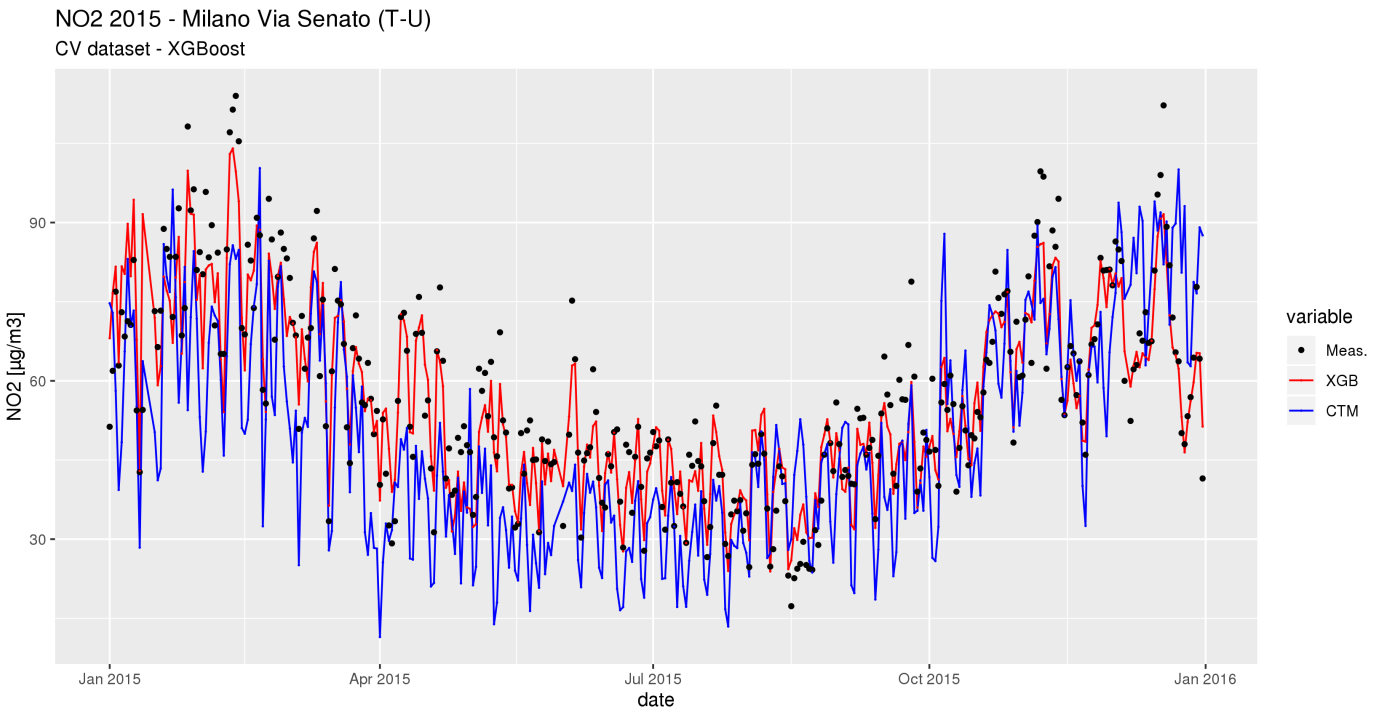
Random Forest



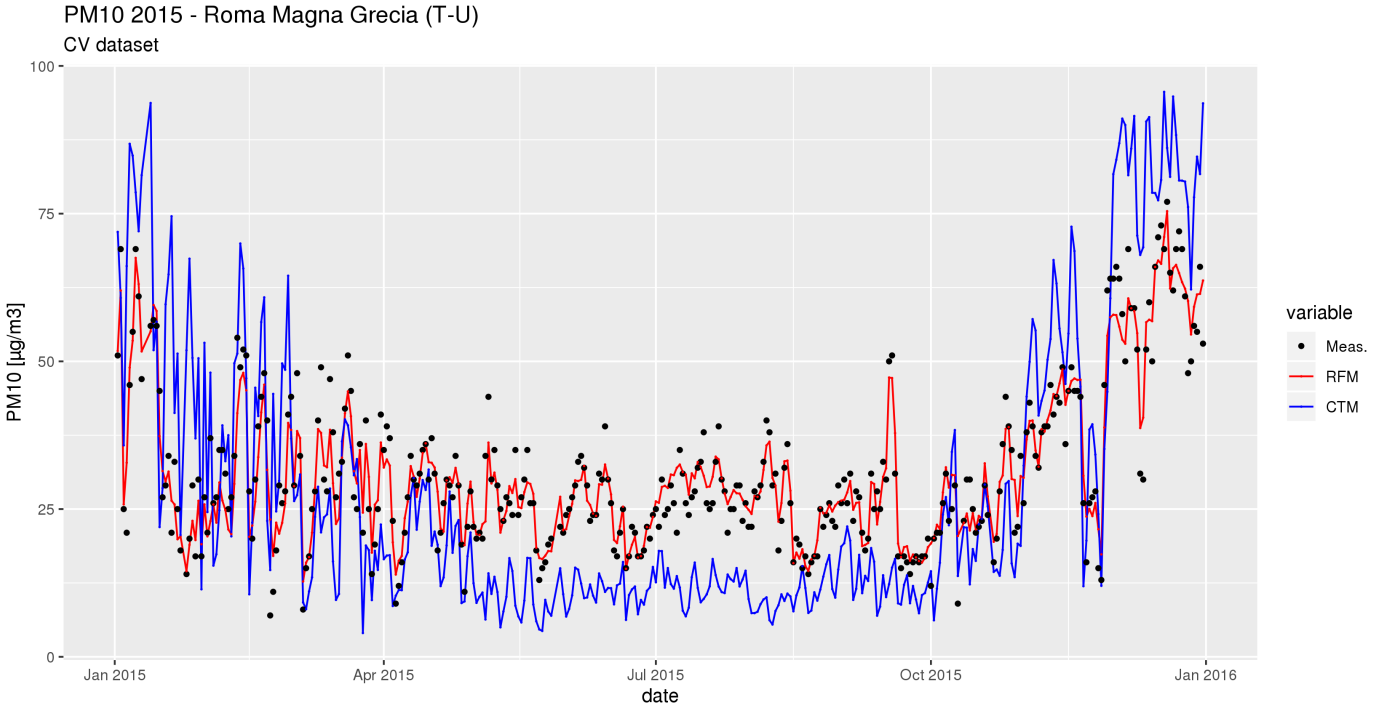
XGBoost



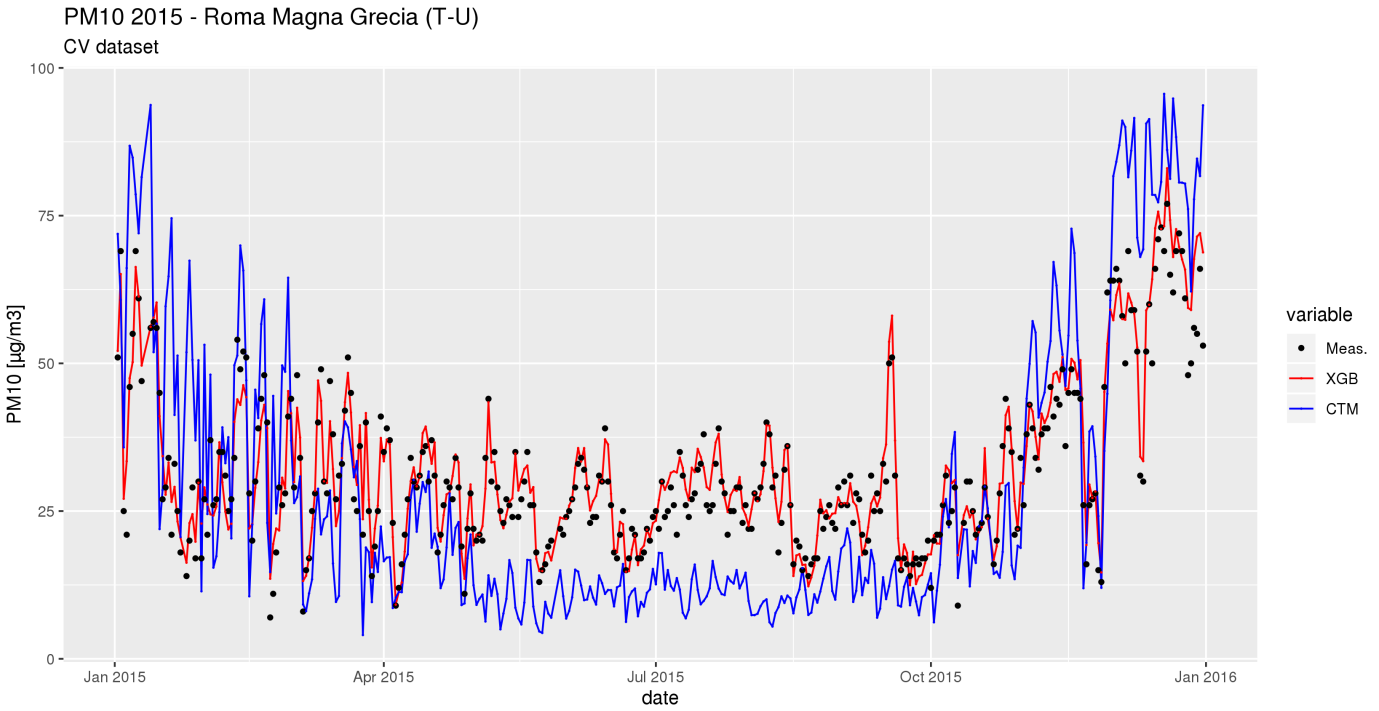
Random Forest



XGBoost



Random Forest



XGBoost

Considerazioni finali

- I risultati ottenuti con i modelli ML sono **utilizzati con successo** negli studi in campo epidemiologico (progetto BEEP).
- La **qualità dei dati di input** è fondamentale per l'addestramento e quindi per conseguire risultati accurati.
- Gli output del modello di dispersione (FARM) sono i **predittori più significativi**. Sono anche più semplici da usare rispetto ai dati satellitari.
- La costruzione dei set di dati input è la parte più onerosa della metodologia ma è quella da cui ci attendiamo un **miglioramento significativo** delle performance.
- La scelta del modello ML è meno importante, tuttavia XGBoost fornisce **risultati migliori** rispetto a Random Forest (soprattuto per NO_2.)
- I modelli ML sono molto meno onerosi in termini di CPU rispetto ai modelli di dispersione ma sono più **avidi di memoria**.