XI giornata sulla modellistica in aria(net)
Milano, 11 aprile 2024

MaL-Air [Machine Learning for Air]

Data fusion con metodi Machine Learning

Esempio di Applicazione sull'area urbana torinese

Umberto Giuriato, Alessandro D'Ausilio, Camillo Silibello

ARIANET s.r.l. - via Benigno Crespi, 57 – 20159 Milano, ITALY - ph. +39-02-27007255 - www.aria-net.it

Data Fusion and Downscaling of Air Quality Deterministic models in the Turin area with Random Forest

U. Giuriato[1], A. D'Ausilio[1], C. Silibello[1], R. De Maria[2], S. Bande[2], C. Cascone[2], M. Maringo[2]

[1]ARIANET srl, 20159 Milano, Via Benigno Crespi 57, Italy.
[2]ARPA Piemonte, Regional Environmental Protection Agency of Piemonte, 10135 Torino, Italy.

14th International Conference on Air Quality - Science and Application
Helsinki. 13 May - 17 May 2024

ARIANET
A suez company

PIEMONTE
Arpa
Agenzia Regionale per la Protezione Ambientale

Introduction and motivation

Supervised learning can be successfully employed to spatialize concentrations measured by an ***observation network***

CTMs concentration fields play the role of ***predictors***, allowing the ***data fusion*** between deterministic models and actual observations

The combination with other High-Resolution predictors leads to the ***downscaling*** of concentration maps

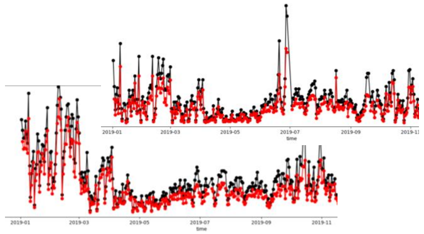
A large and ***representative sensor network*** is fundamental to improve the learning process and enhance ***generalization capabilities***

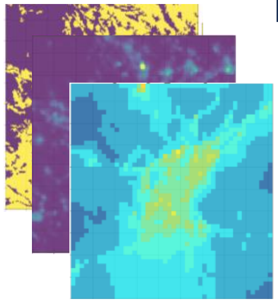
Since in real-case scenarios this is not always the case, we need to engineer a way to

- increase the representativeness of training set
- quantify the uncertanties

MaL-Air [Machine Learning for Air]

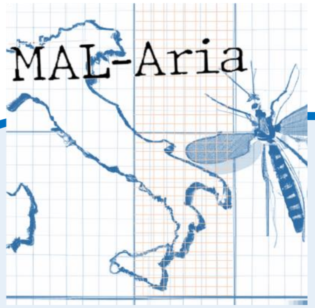
The ARIANET solution for data fusion with supervised learning



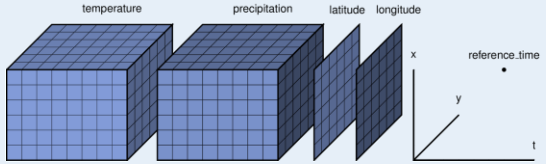


Preprocess outside MaL-Air:

- **Observations csv**: station concentration time series in any time aggregation
- **Predictors netcdf**:
  - FARM simulations
  - LAI
  - Land use
  - Other static predictors



MaL-Air



MAL-Air pre-processor:

- **Extract** predictors at the station locations
- **Resample** predictors and observations to target time aggregation
- **Combine** predictors and observations in the training dataset
- **Feature Engineering** for manipulation of variables

Supervised learning

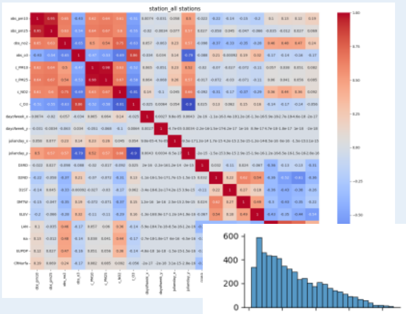
Random Forest

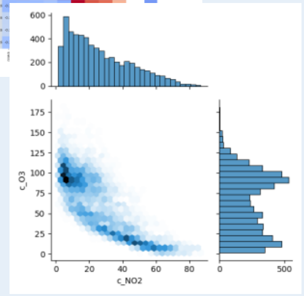
Linear Models

MLP (To Add)

MAL-Air core:
Base Model class containing all methods for ML models

- Performs ML tasks
  - **Training**
  - **Inference**
  - **Validation**
  - **Hyperparameters tuning**





MAL-Air support analysis tool:

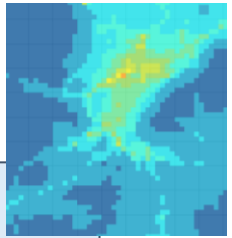
- Principal Component Analysis and clustering for outlier detection
- Correlation maps
- ADF Test
- Joint plots
- Distributions
- Coverage score

The app is modular!
Since supervised learning tasks all require a matrix predictors X samples, every new model implemented could be a class inheriting the main attributes and methods from a base class

Supervised training with Random Forest

- Training is performed on the monitoring station time series.
- The trained **Random Forest** model is then applied to each cell of the domain for each day to ***infer concentration maps***

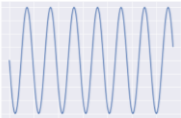
PREDICTORS



Spatio-temporal
CTM concentration fields
Leaf Area Index (LAI)

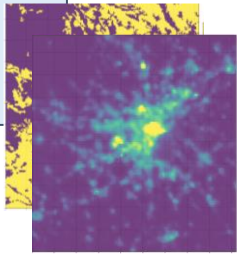
`.netcdf`

Temporal (homogeneous)
Periodic functions of
Julian day, day of week



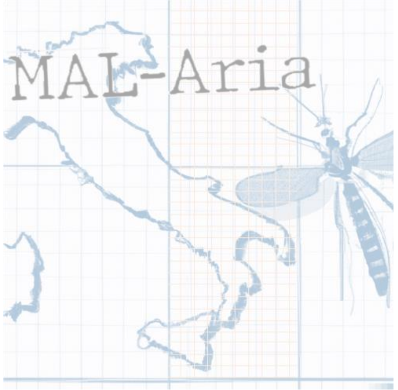
Spatial (stationary)
Distance from roads
Population density
Land use
……….

`.netcdf`



PRE-PROCESSING

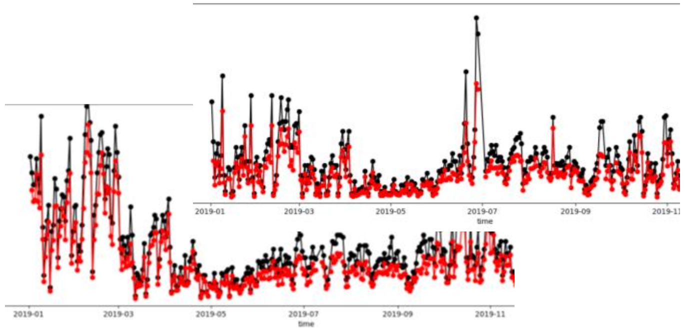
- ***Extraction*** *of predictors fields at stations locations*
- ***Resampling*** *in time*
- *Combination of predictors and target in the* ***training set***



TARGET

Concentrations measured by the sensors

`.csv`



Supervised training with Random Forest

- Training is performed on the monitoring station time series.
- The trained **Random Forest** model is then applied to each cell of the domain for each day to ***infer concentration maps***

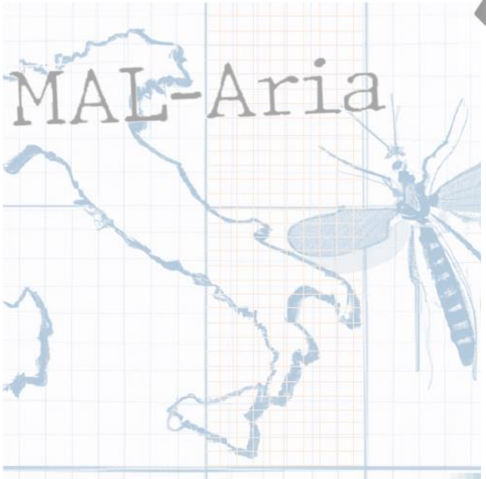
PREDICTORS

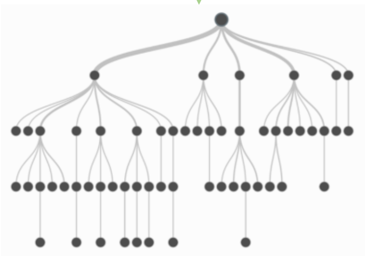
Spatio-temporal
CTM concentration fields
Leaf Area Index (LAI)

Temporal (homogeneous)
Periodic functions of
Julian day, day of week

Spatial (stationary)
Distance from roads
Population density
Land use
……….

PRE-PROCESSING training set





RANDOM FOREST TRAINING

TARGET

Concentrations measured by the sensors

TRAINED MODEL

(Hyperparameter tuning and cross-validation also in this phase)

Supervised training with Random Forest

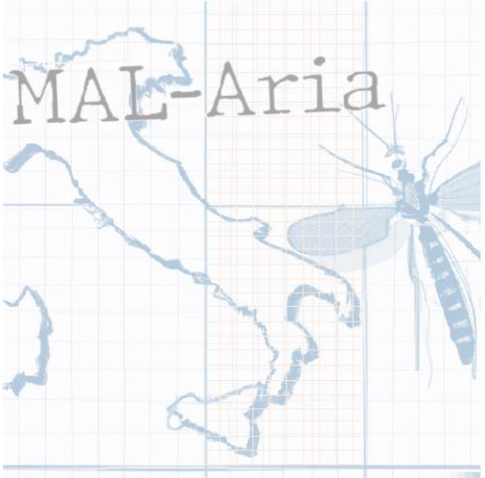
- Training is performed on the monitoring station time series.
- The trained **Random Forest** model is then applied to each cell of the domain for each day to ***infer concentration maps***

PREDICTORS

Spatio-temporal
CTM concentration fields
Leaf Area Index (LAI)

Temporal (homogeneous)
Periodic functions of
Julian day, day of week

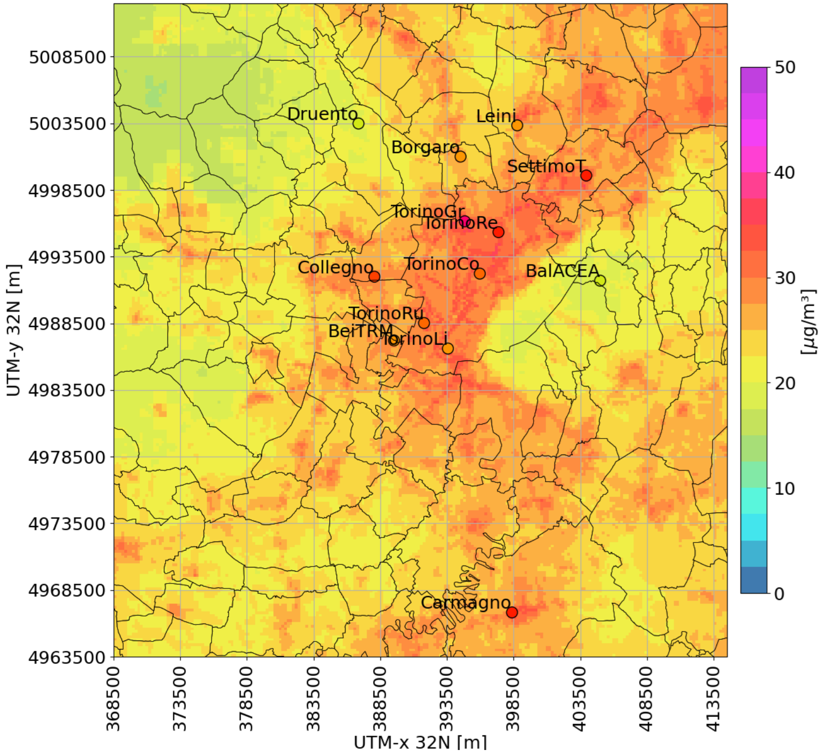
Spatial (stationary)
Distance from roads
Population density
Land use
……….



TRAINED MODEL

RANDOM FOREST INFERENCE

Data fusion maps between models and observation data

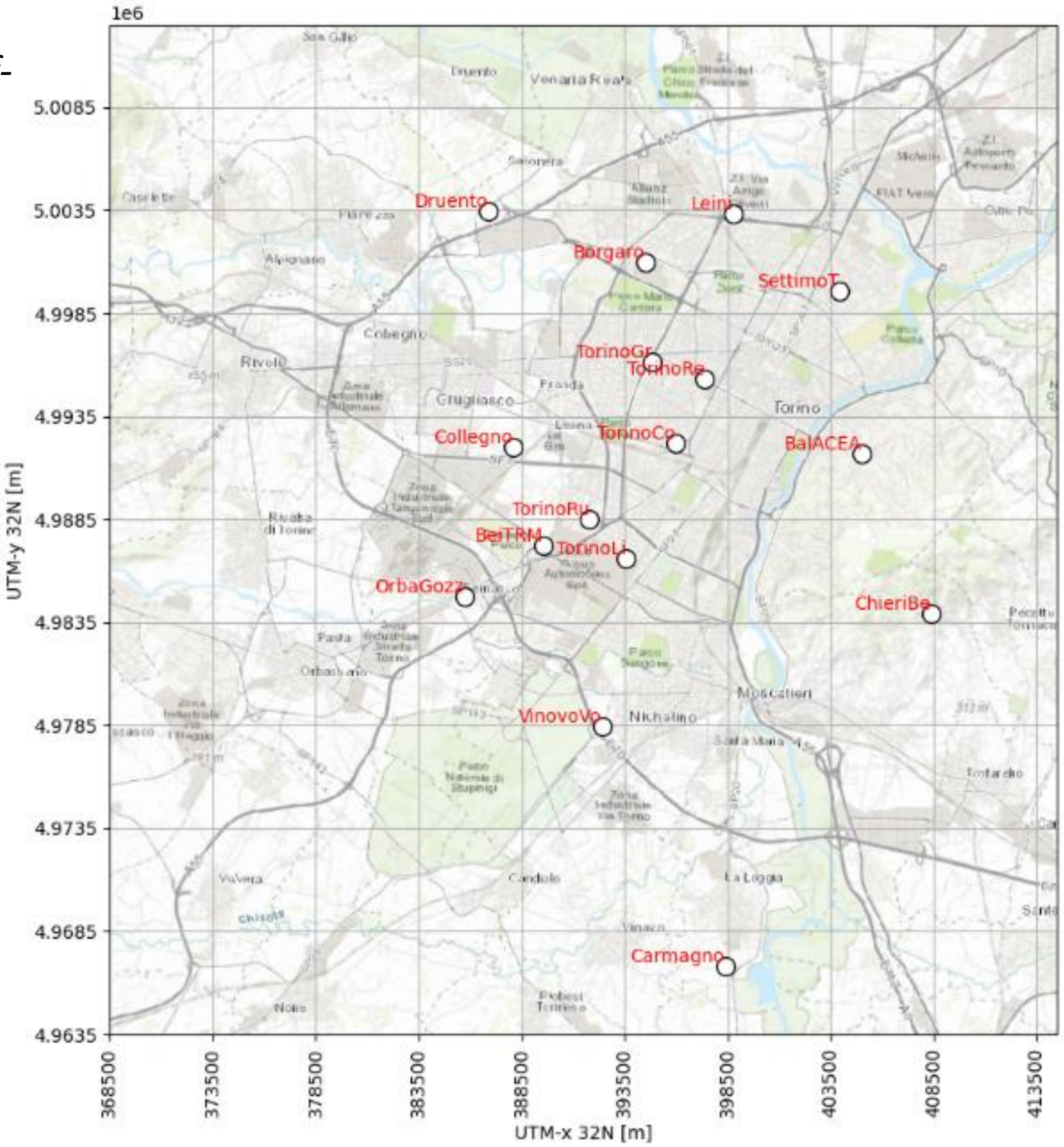


Use case: SPoTT project

Surveillance on POpulation health around the Turin waste-of-energy plant

- Pollutants under study: ***PM10*** and ***PM2.5***
- Year of study*: **2019***
- Target resolution: ***200m***
- ***14*** monitoring stations
  - ***8*** measure PM25, ***13*** measure PM10
  - Majority of ***urban*** stations, just **2** ***rural*** stations

Nome Stazione	PM_{10}	$PM_{2.5}$
Baldissero T. (ACEA)	●	
Beinasco TRM – Aldo Mei	●	●
Borgaro T. - Caduti	●	●
Carmagnola	●	
Chieri - Bersezio		●
Collegno - Francia	●	
Druento – Parco la Mandria	●	
Leini (ACEA) - Grande Torino	●	●
Settimo T. - Vivaldi	●	●
Torino Consolata	●	
Torino Grassi	●	
Torino Lingotto	●	●
Torino Rebaudengo	●	●
Torino Rubino	●	●

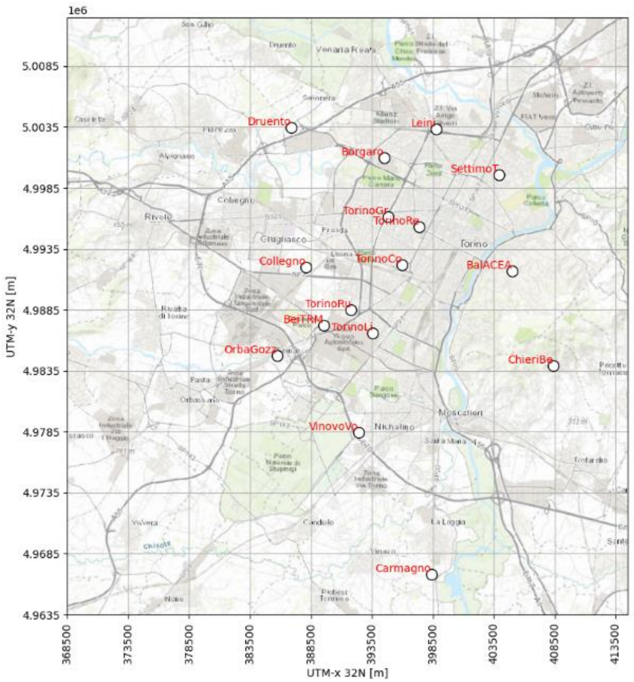


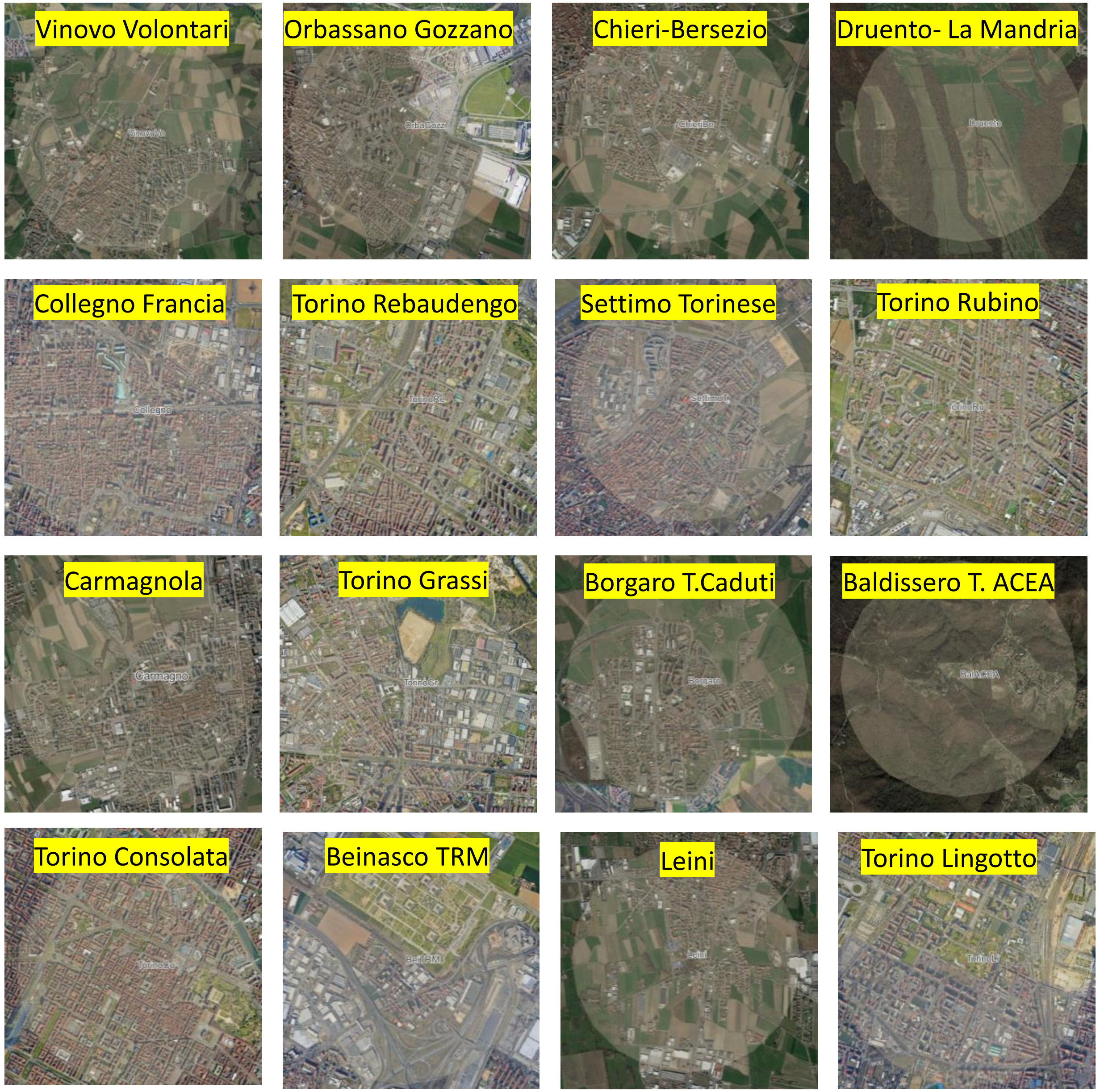
Use case: SPoTT project

Surveillance on POpulation health around the Turin waste-of-energy plant

- Pollutants under study: ***PM10*** and ***PM2.5***
- Year of study*: **2019***
- Target resolution: ***200m***
- ***14*** monitoring stations
  - ***8*** measure PM25, ***13*** measure PM10
  - Majority of ***urban*** stations, just **2** ***rural*** stations

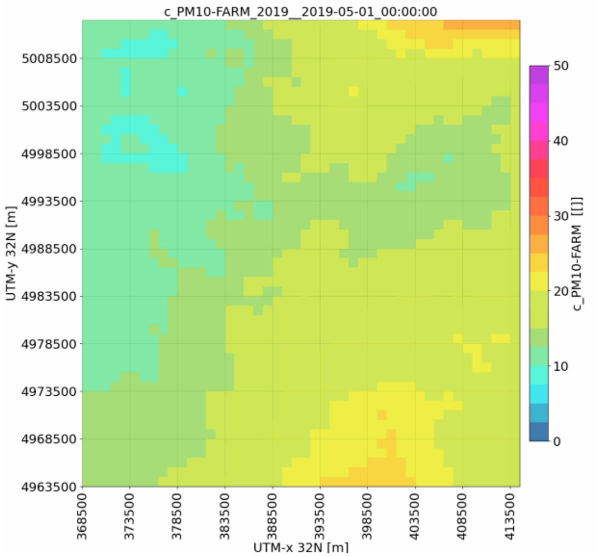
Nome Stazione	PM_{10}	$PM_{2.5}$
Baldissero T. (ACEA)	●	
Beinasco TRM – Aldo Mei	●	●
Borgaro T. - Caduti	●	●
Carmagnola	●	
Chieri - Bersezio		●
Collegno - Francia	●	
Druento – Parco la Mandria	●	
Leini (ACEA) - Grande Torino	●	●
Settimo T. - Vivaldi	●	●
Torino Consolata	●	
Torino Grassi	●	
Torino Lingotto	●	●
Torino Rebaudengo	●	●
Torino Rubino	●	●



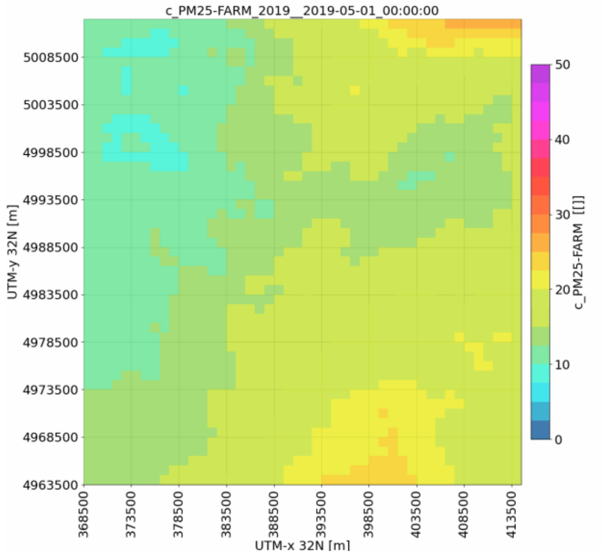


Spatio-temporal Predictors: FARM and Leaf Area Index

FARM: PM_{10} – daily mean



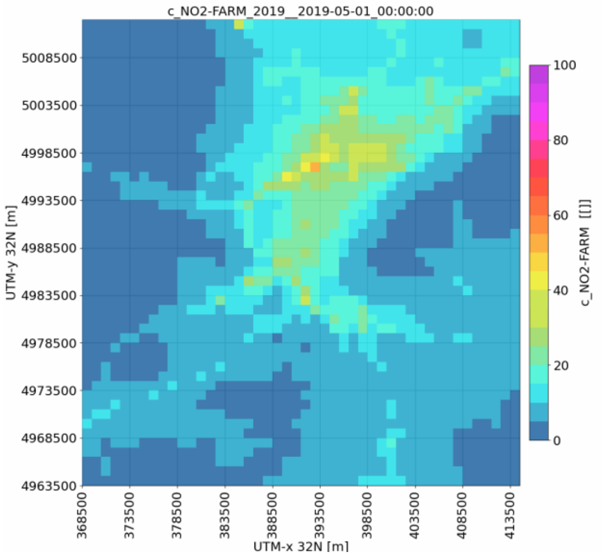
FARM: $PM_{2.5}$ – daily mean



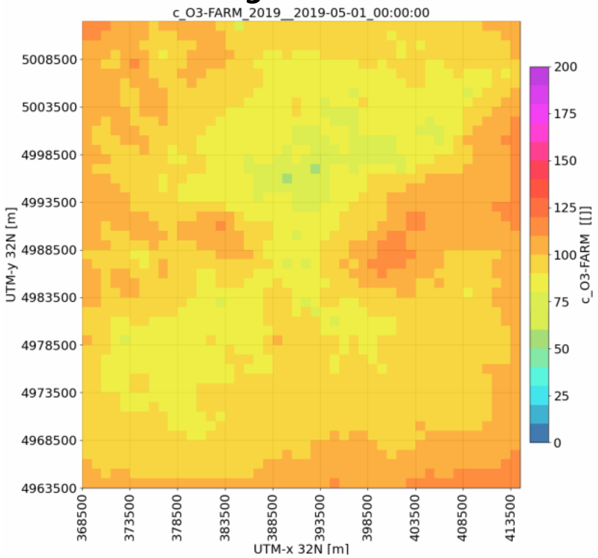
FARM simulations have been performed on the domain of interest at the resolution of **1km**

Leaf Area Index satellite images are monthly sampled at the resolution of **200m**

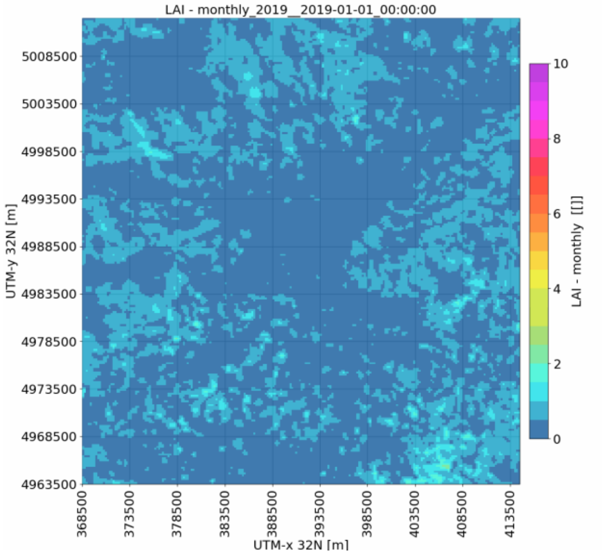
FARM: NO_2 – daily mean



FARM: O_3 – daily mean



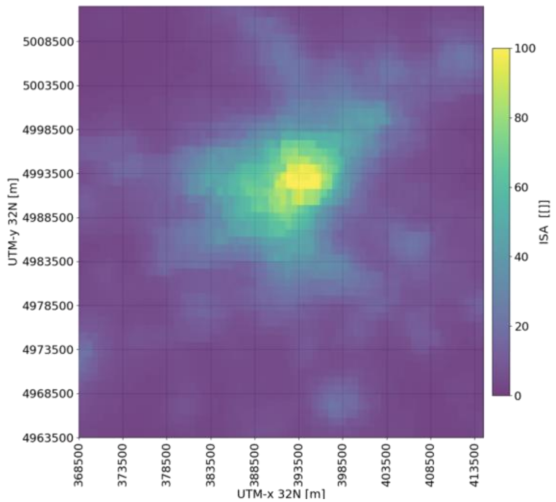
LAI - monthly



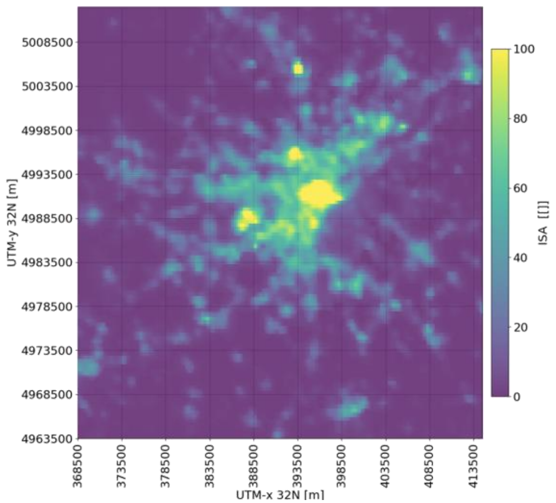
Stationary (Spatial) Predictors

All the stationary predictors are at 200m resolution

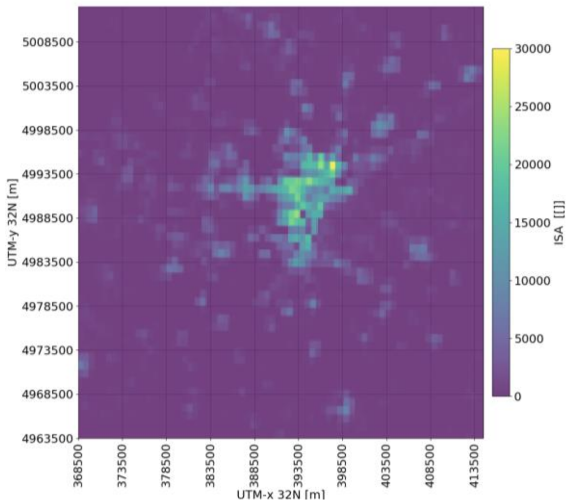
Impervious Surface Area (ISA)



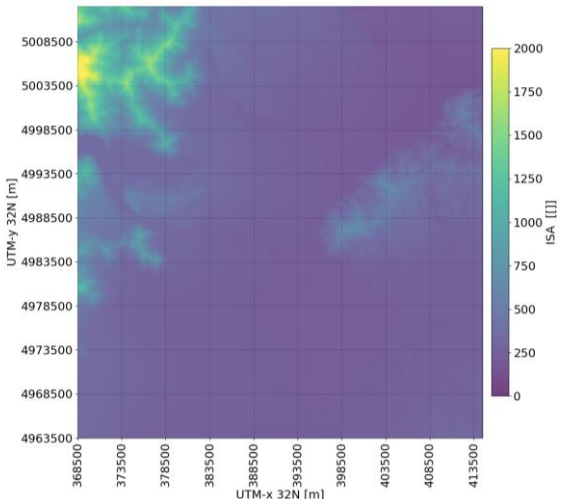
Light at Night (LAN)



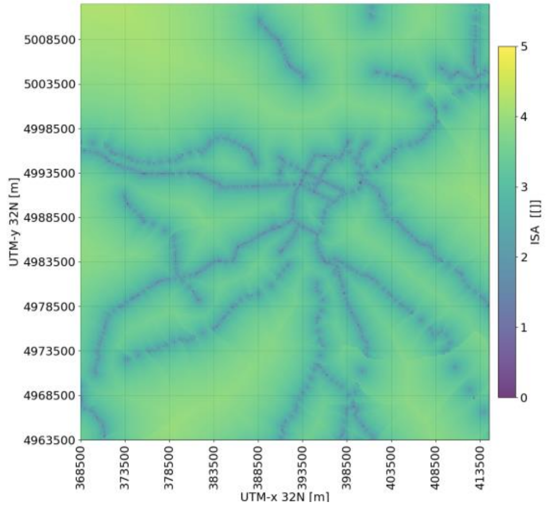
Population Density



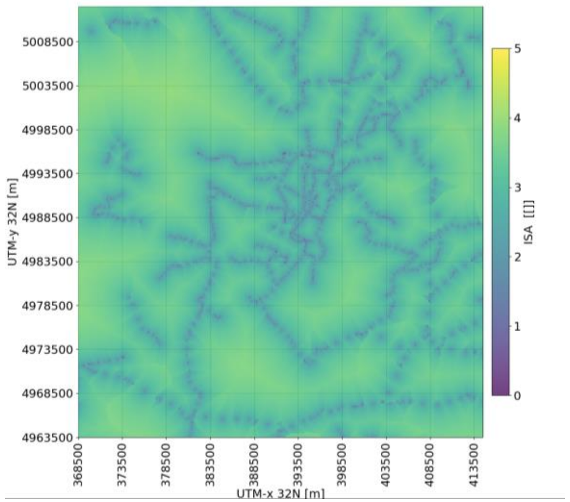
Elevation



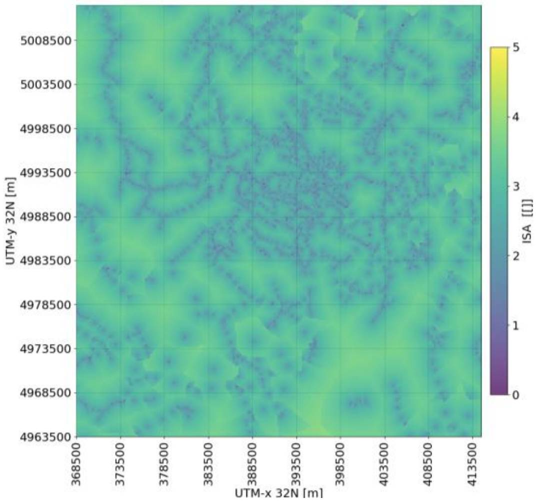
Distance Primary Roads



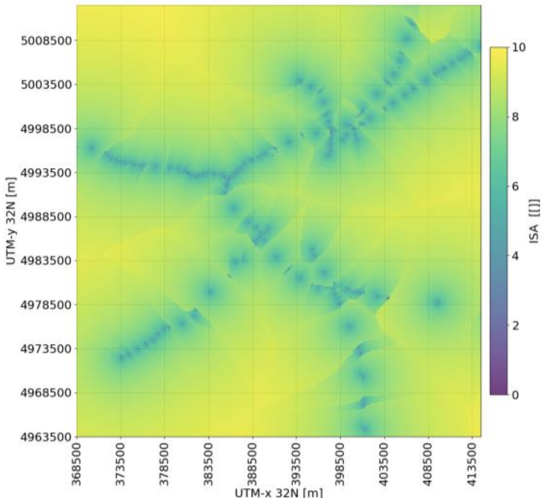
Distance Secondary Roads



Distance Tertiary Roads

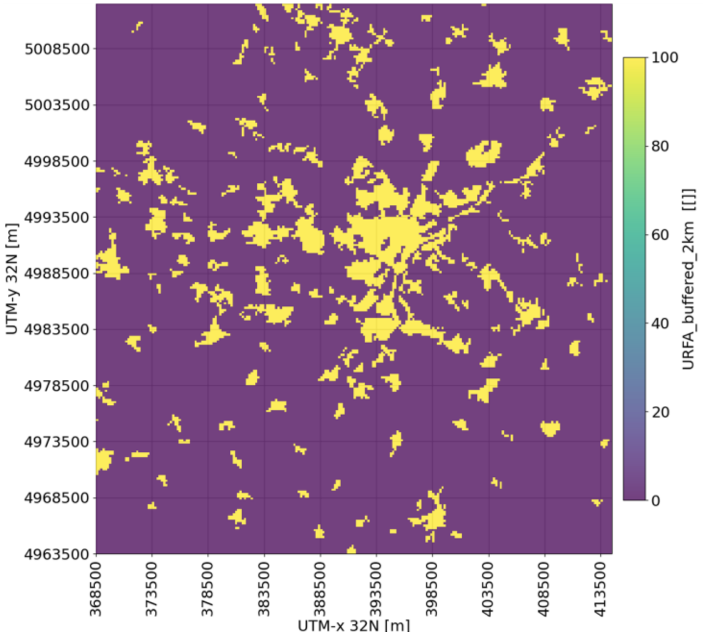


Distance Motorways

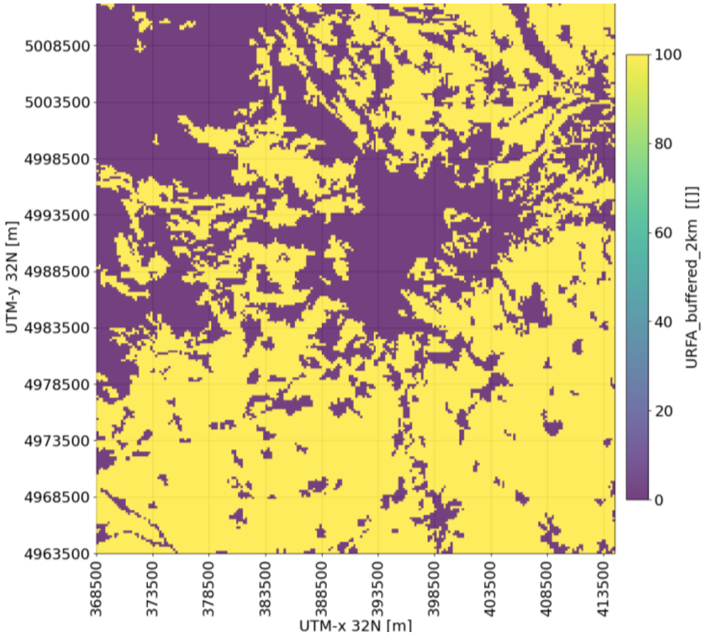


Stationary (Spatial) Predictors: Corine Land Cover

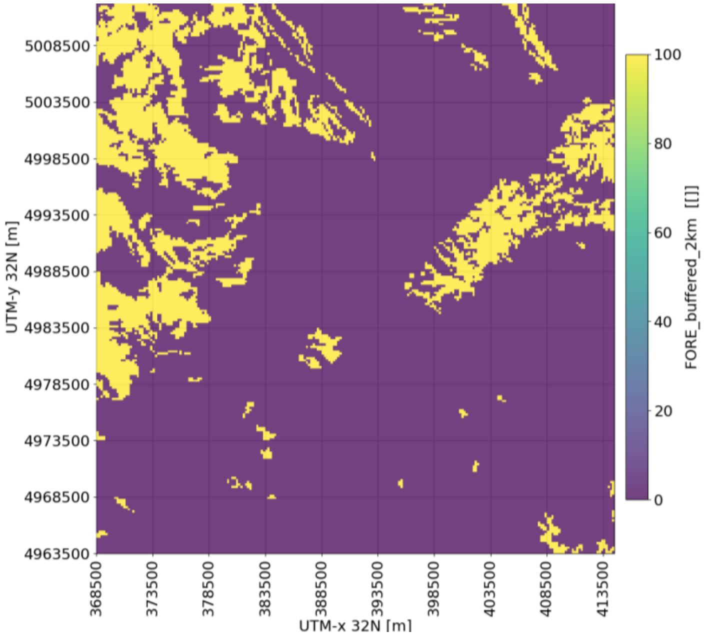
Urban-Fabric



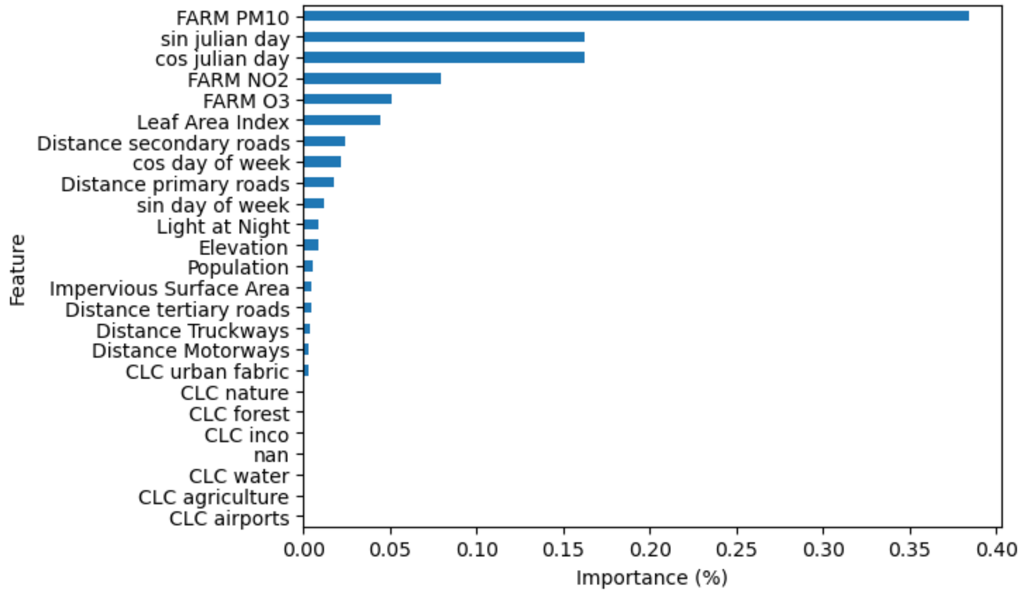
Agriculture



Forest



PM_{10} annual mean concentration



Feature importances

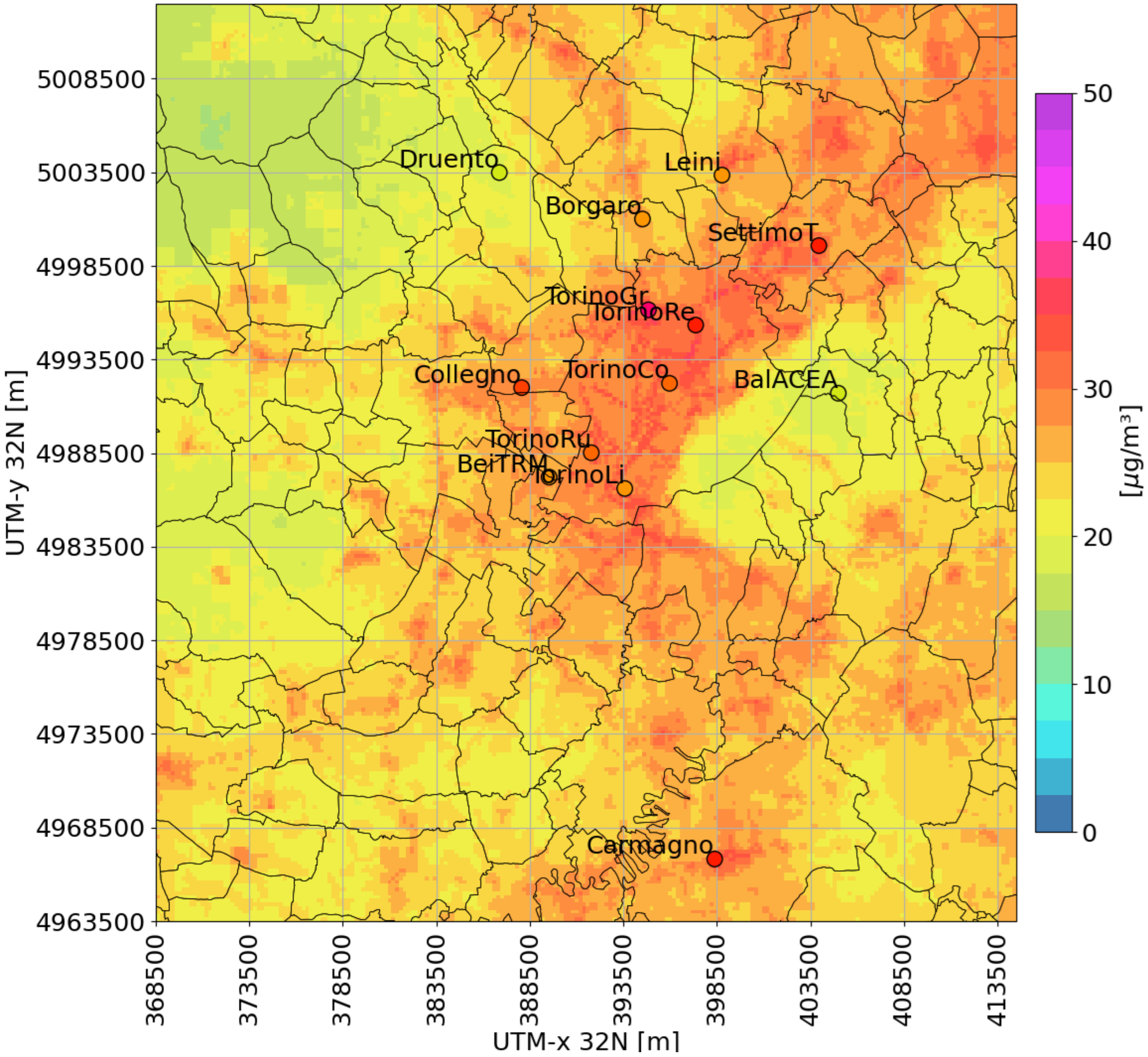
Feature	Importance (%)	Cumulative importance (%)
FARM PM10	0.38	0.38
cos julian day	0.16	0.54
sin julian day	0.16	0.7
FARM NO2	0.08	0.78
FARM O3	0.05	0.83
Leaf Area Index	0.04	0.87
Dist primary roads	0.02	0.89
cos day of week	0.02	0.91
Dist second roads	0.02	0.93
Population	0.01	0.94
Elevation	0.01	0.95
Light at Night	0.01	0.96
sin day of week	0.01	0.97

Hyperparameters

Hyperparameter	Value
N_trees	800
Max depth	400
Min samples split	2
Min samples leaf	2
Max features	0.85
Max leaf nodes	1200

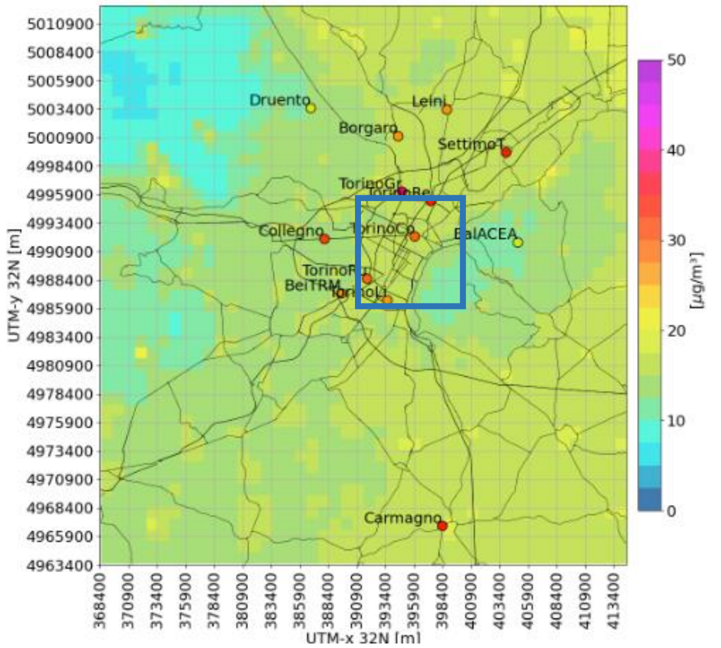
Cross-validation RMSE

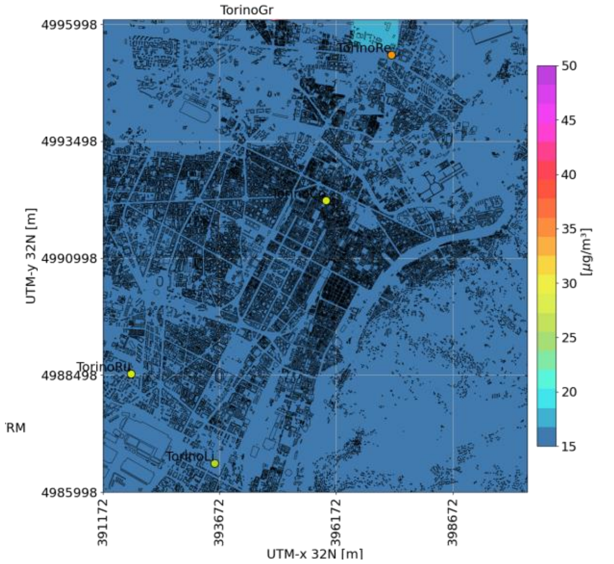
Model	Value
FARM	20.5 ± 0.74
Random Forest	7.2 ± 0.21



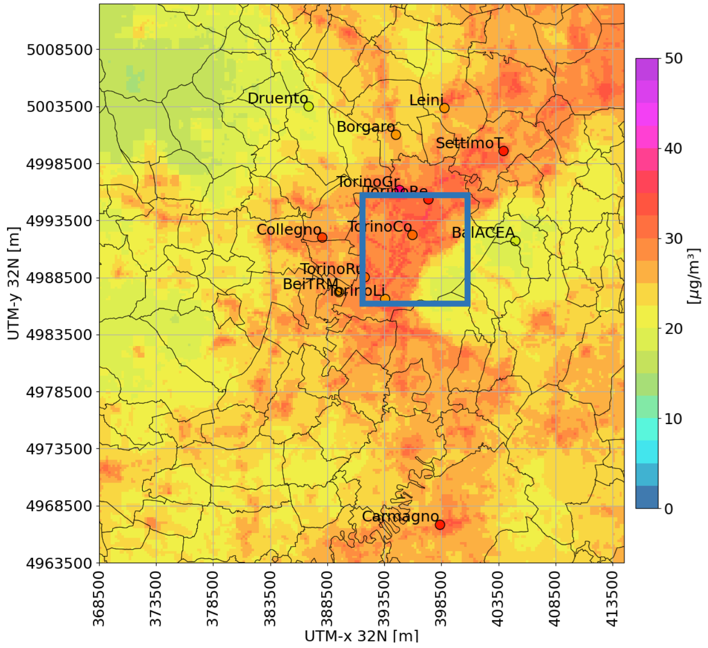
PM_{10} annual mean concentration

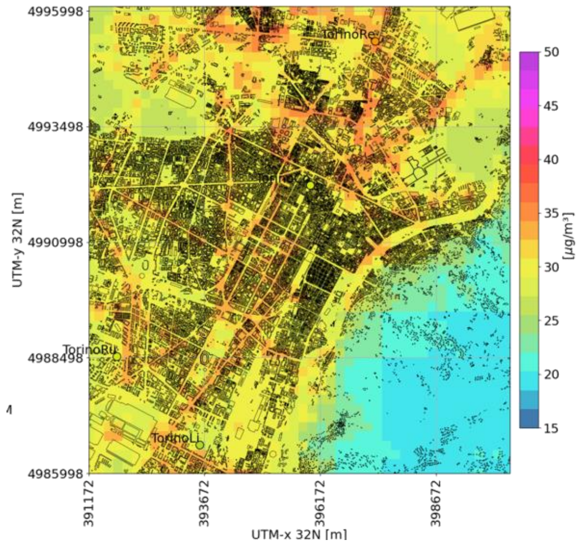
FARM

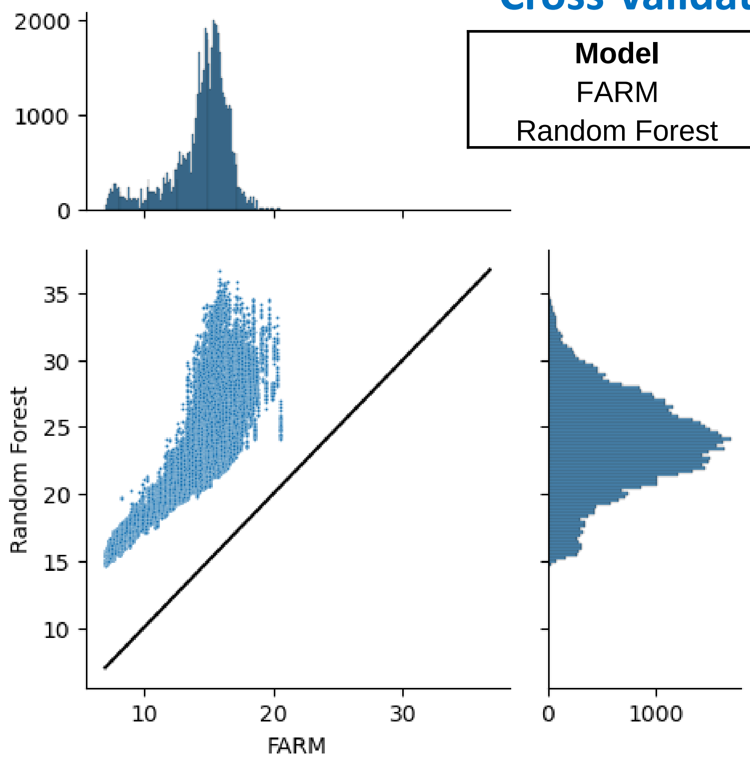




Random Forest







Cross-validation RMSE

Model	Value
FARM	20.5 ± 0.74
Random Forest	7.2 ± 0.21

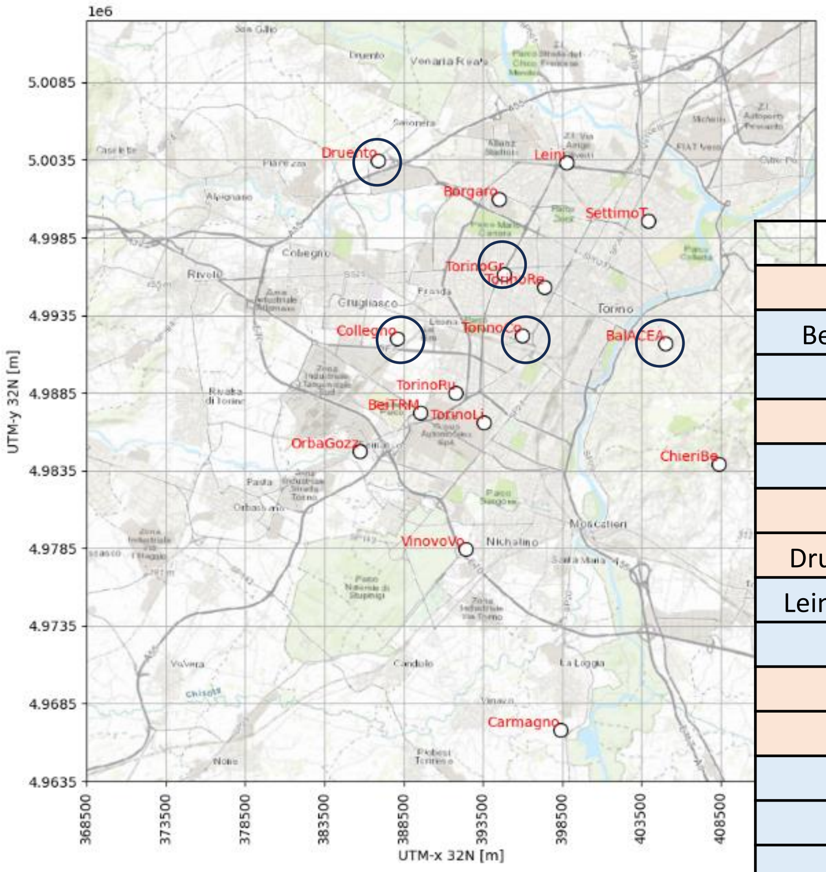
- Globally, Random Forest inference ***fixes the bias*** of FARM predictions with respect to observations
- ***Downscaling*** due to stationary predictors is an ***increment of concentration on roads*** at the local scale

Data augmentation: imputing $PM_{2.5}$ time series

A suez company

Lack of representativeness of observation network

- Only **8** stations measuring $PM_{2.5}$
- Lack of rural stations **Druento**, **Baldissero** and some **traffic** stations in Turin's urban area



Nome Stazione	PM_{10}	$PM_{2.5}$
Baldissero T. (ACEA)	●	
Beinasco TRM – Aldo Mei	●	●
Borgaro T. - Caduti	●	●
Carmagnola	●	
Chieri - Bersezio		●
Collegno - Francia	●	
Druento – Parco la Mandria	●	
Leini (ACEA) - Grande Torino	●	●
Settimo T. - Vivaldi	●	●
Torino Consolata	●	
Torino Grassi	●	
Torino Lingotto	●	●
Torino Rebaudengo	●	●
Torino Rubino	●	●

It is possible to increase the training set size for $PM_{2.5}$ by **learning information from PM_{10} stations**

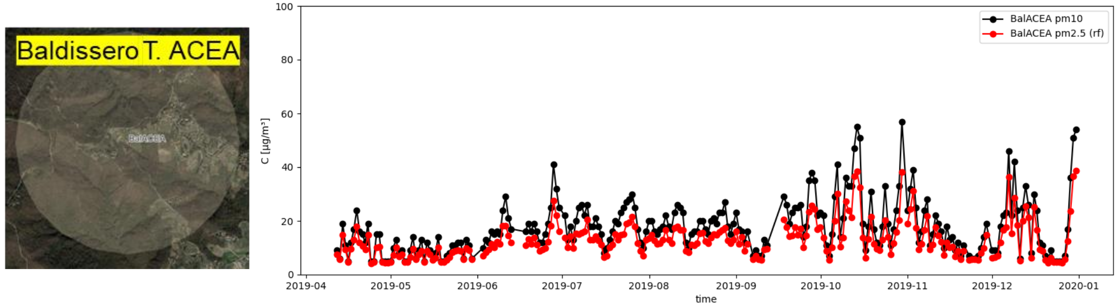
- Preliminary random forest model (RF_{st}) is trained at stations where both species are sampled
- Then it is applied to infer timeseries of the ratio **$PM_{2.5}$ / PM_{10}** at stations where only PM_{10} data are available

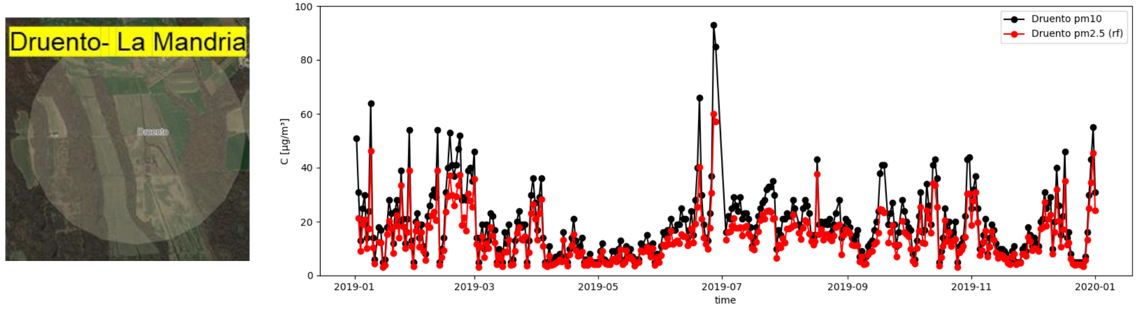
$$\frac{PM_{2.5,obs}}{PM_{10,obs}} \sim RF_{st}\left(PM10_{obs}, \frac{PM_{2.5,FARM}}{PM_{10,FARM}}\right)$$

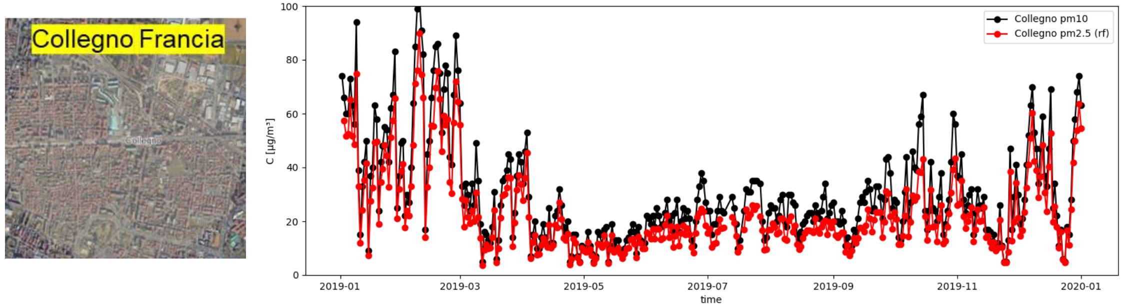
Using **$PM_{2.5}$ / PM_{10}** as target forbids unphysical values with **$PM_{2.5} > PM_{10}$**

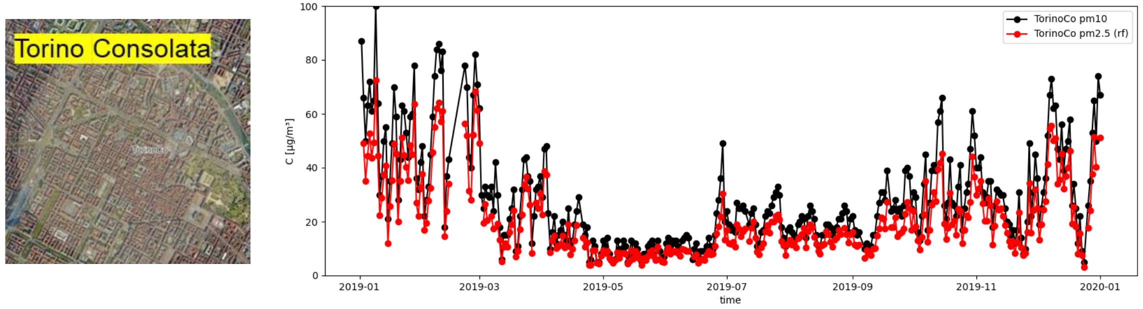
Stafoggia M, Johansson C, Glantz P, Renzi M, Shtein A, de Hoogh K, Kloog I, Davoli M, Michelozzi P, Bellander T. A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in Sweden. *Atmosphere*. 2020; 11(3):239. https://doi.org/10.3390/atmos11030239

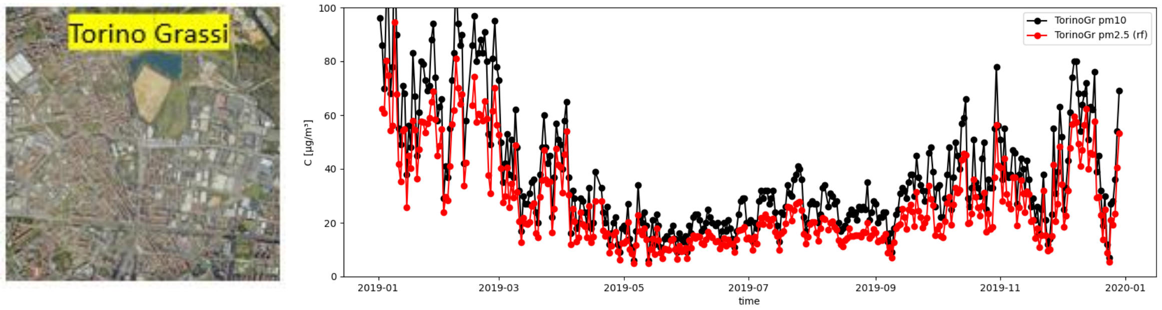
Data augmentation: imputing $PM_{2.5}$ time series











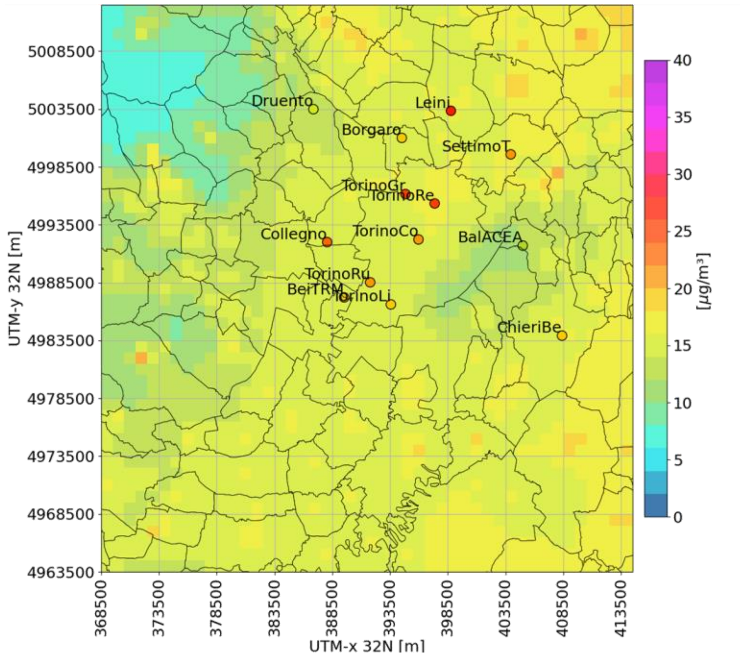
PM_{10}

$PM_{2.5}$

The “*surrogate*” $PM_{2.5}$ observation time series are highly correlated with PM_{10} ones, with small variability in their ratio

$PM_{2.5}$ annual mean concentration

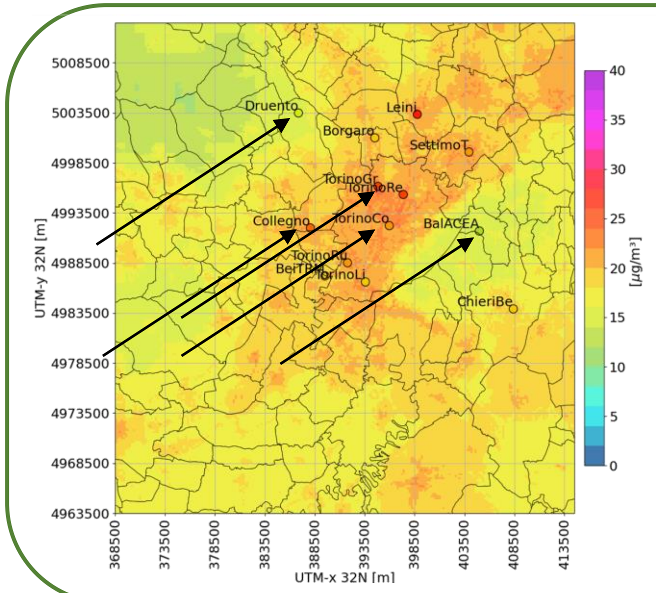
FARM

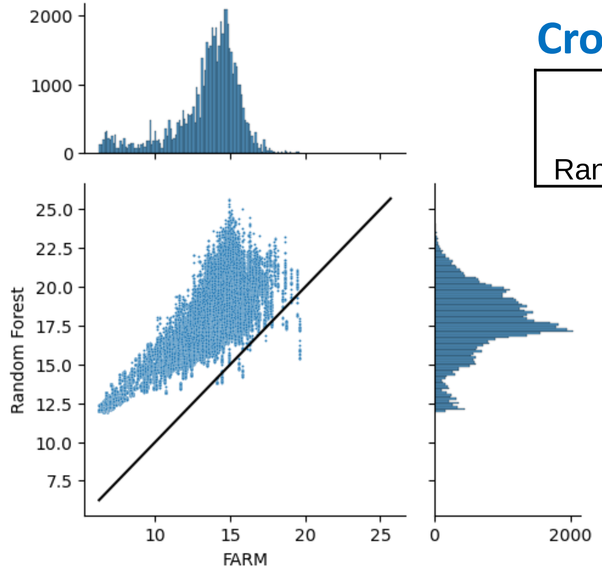


Augmenting data based on FARM PM ratio and PM_{10} observations:

- Thins out the difference with FARM field, still healing the bias
- Keeps similarity with PM_{10} map, without concentration increase outside urban area

Random Forest WITH augmented data

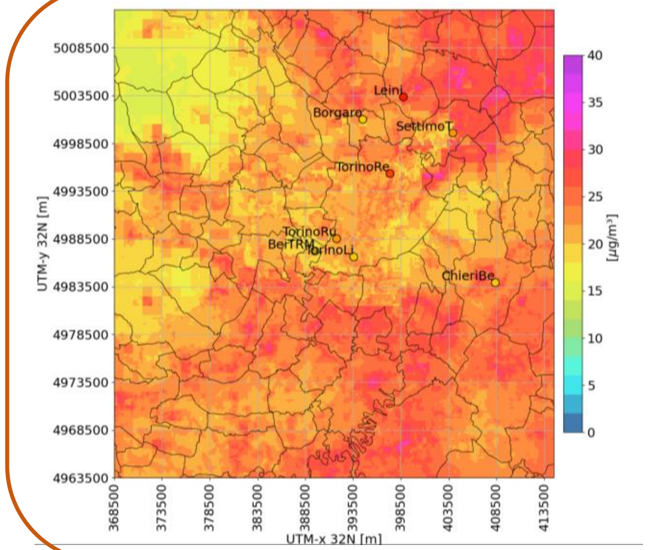


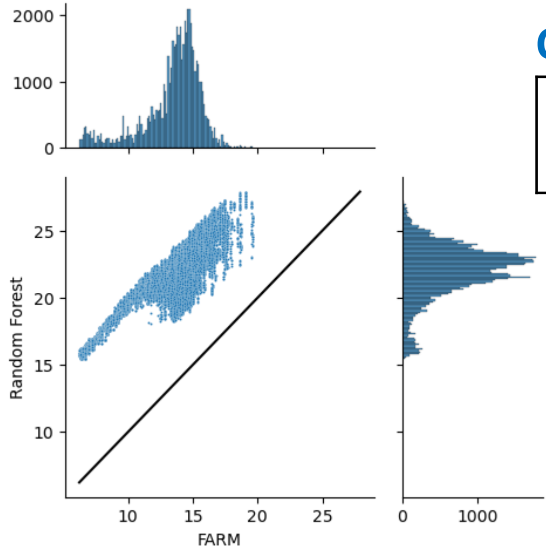


Cross-validation RMSE

Model	Value
FARM	14.0 ± 0.64
Random Forest	5.9 ± 0.52

Random Forest WITHOUT augmented data





Cross-validation RMSE

Model	Value
FARM	14.0 ± 0.64
Random Forest	6.02 ± 0.48

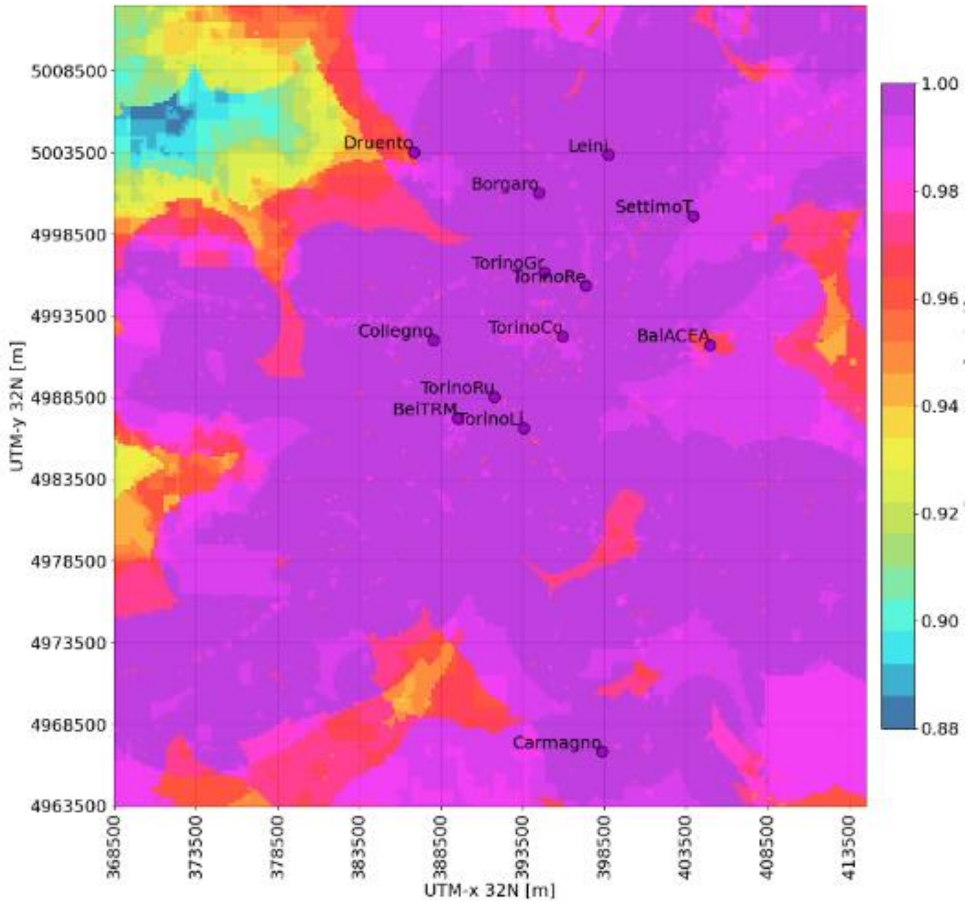
BONUS: Coverage Score

A **Coverage Score** can be defined to quantify how much the predictors' distribution in the target domain matches the one *"seen"* by the observation network.

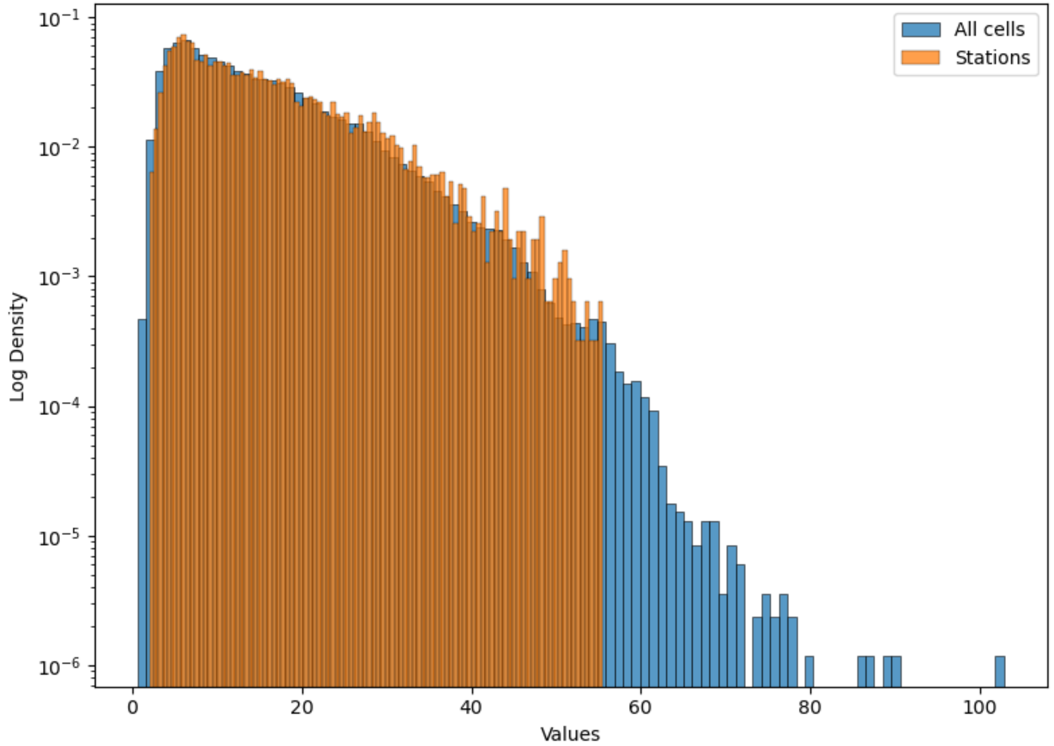
In every cell ij: $$C_{ij} = \sum_{k=1}^{N_{\text{predictors}}} \text{importance}_k \cdot Q_{ij}$$

Where Q_{ij} is:
- *1 if the values belong to the distribution sampled by the observation network*
- *0 otherwise*

Mean Coverage map for PM_{10} - Turin

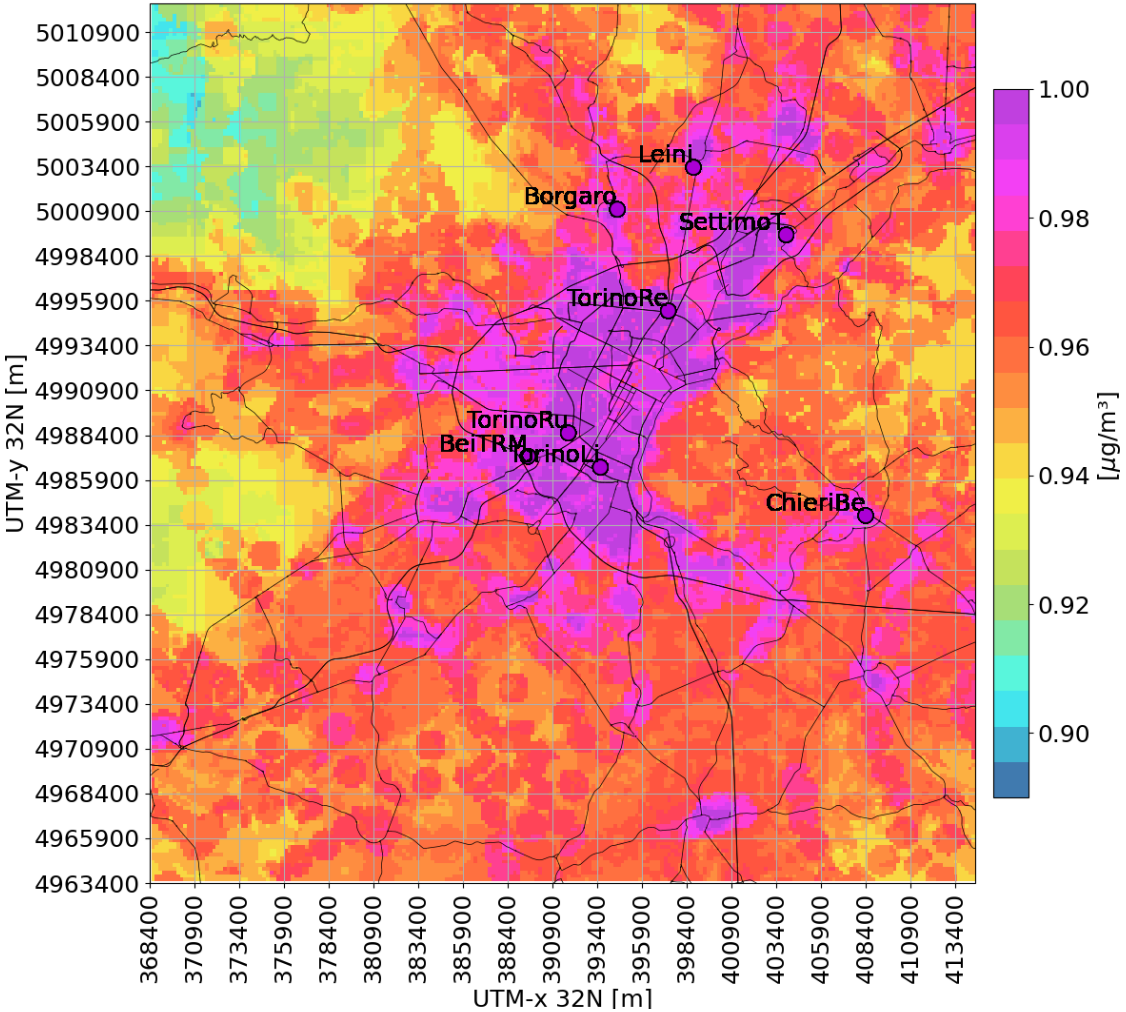


Distribution of FARM PM10 values (stations vs all cells)

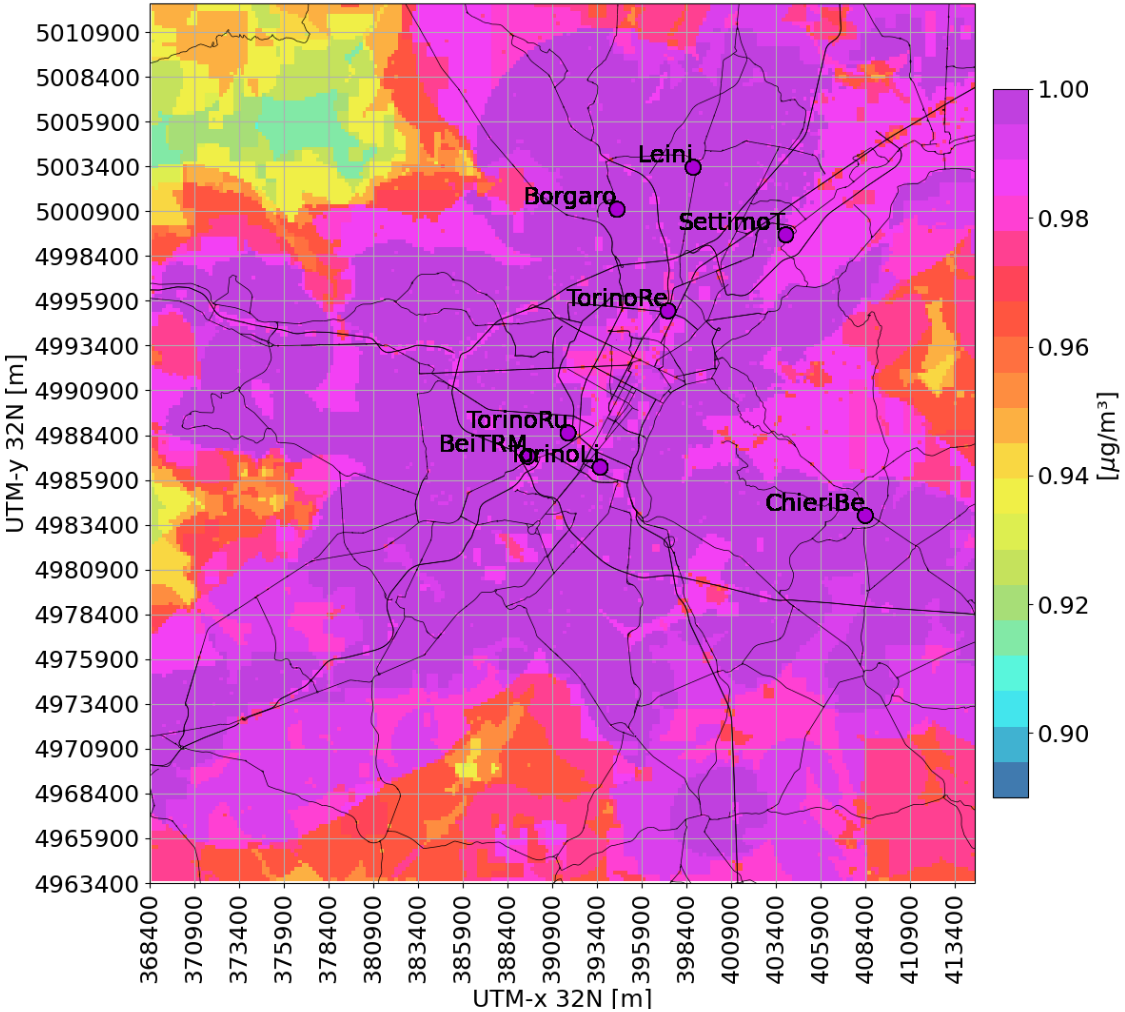


BONUS: Coverage Score for $PM_{2.5}$

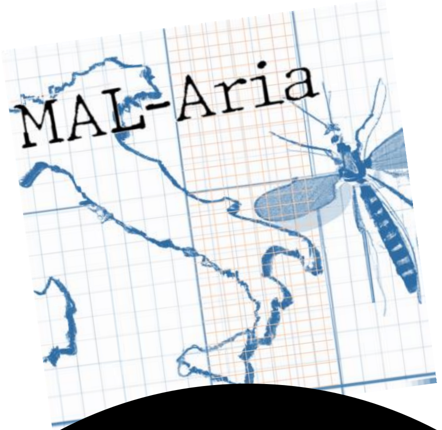
Mean Coverage map for $PM_{2.5}$
WITHOUT augmented data



Mean Coverage map for $PM_{2.5}$
WITH augmented data



Adding the *"surrogate"* stations improves the coverage in regions outside Turin, increasing the reliability of the model.



Thank you for your attention !

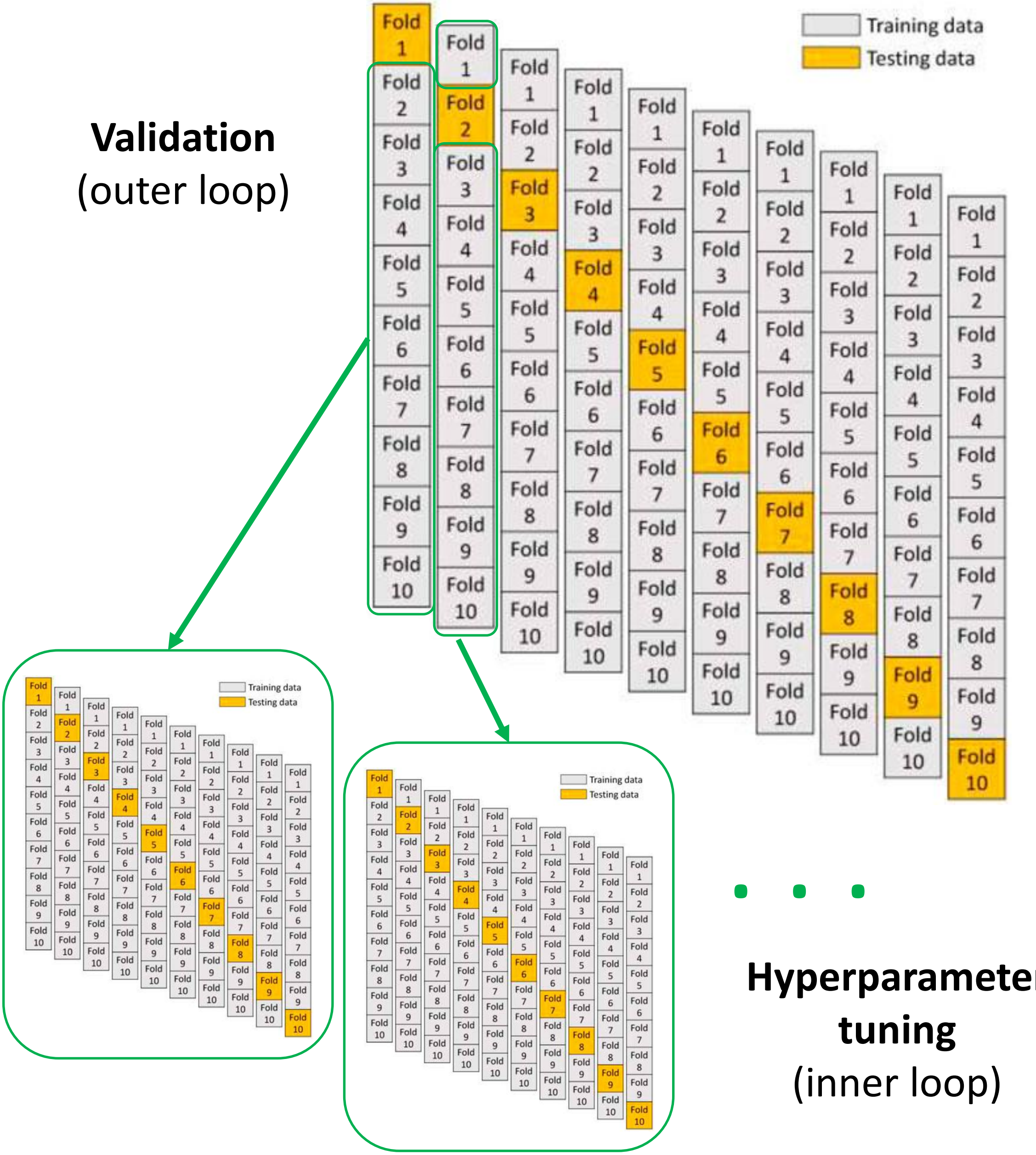
Umberto Giuriato

Alessandro D’Ausilio

Camillo Silibello

Nested K-fold cross validation

Validation
(outer loop)



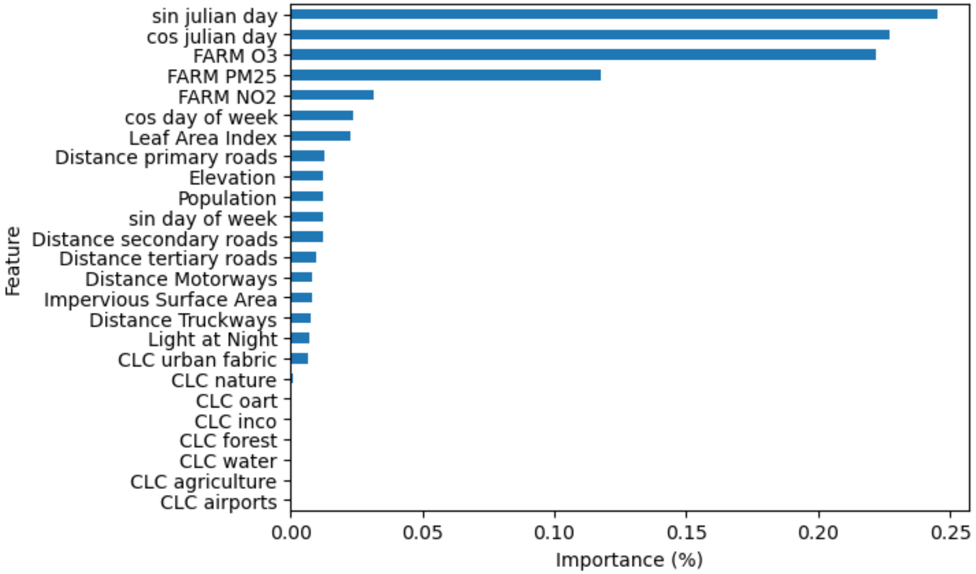
Hyperparameter tuning
(inner loop)

Validation of the Random Forest training is performed with a K-fold nested cross-validation

The evaluation metric used is ***RMSE***

- Split dataset into K equal folds
- One fold is used as **test** set and K-1 remaining as **training**
    - In each inner training fold, perform an extra K-fold splitting and use it to perform **hyperparameter tuning**
- Train the model on the training set for each iteration independently
- Validate the model on the test set for each iteration
- The final score is the average obtained from all K iterations to get the final score

PM$_{2.5}$ annual mean concentration



Feature importances

Feature	Importance (%)	Cumulative importance (%)
sin julian day	0.25	0.25
cos julian day	0.23	0.48
FARM O3	0.22	0.70
FARM PM25	0.12	0.82
FARM NO2	0.03	0.85
cos day of week	0.02	0.87
Leaf Area Index	0.02	0.89
Dist primary roads	0.01	0.90
Elevation	0.01	0.91
Population	0.01	0.92
sin day of week	0.01	0.93
Dist second roads	0.01	0.94

Hyperparameters

Hyperparameter	Value
N_trees	400
Max depth	30
Min samples split	2
Min samples leaf	2
Max features	0.85
Max leaf nodes	700

Cross-validation RMSE

Model	Value
FARM	14.0 ± 0.64
Random Forest	5.9 ± 0.52