XIII giornata della modellistica in aria(net)
Milano, 25/03/2026

Machine Learning for Air Quality and Environmental Modelling

Overview of **ARIANET solutions**

Umberto Giuriato – Alessandro D'Ausilio

Is Machine Learning the Death-Of-Modelling?

ML refers to statistical techniques to automatically detect patterns in data by optimizing a performance metric (Murphy, 2022).

Unsupervised learning

Finding structure in unlabeled data

Meteorological-day clustering for faster long-term dispersion simulations

Road-segment classification for urban dispersion kernels and digital twins

Supervised learning

Learning from labeled data

Inverse modelling to rescale emission factors from observations

Mass-consistent concentration downscaling driven by geophysical proxies

Data fusion and concentration time-series forecasting

At *ARIANET*, we use machine learning **not to replace modelling**,
but to make **traditional models more efficient**
and to **integrate external data sources** into our modelling framework.

Machine Learning at ARIANET

Unsupervised learning

Finding structure in unlabeled data

Meteorological-day clustering for fast long-term dispersion simulations

Road-segment classification for urban dispersion kernels and digital twins

Supervised learning

Learning from labeled data

Inverse modelling to rescale emission factors from observations

Mass-consistent concentration downscaling driven by geophysical proxies

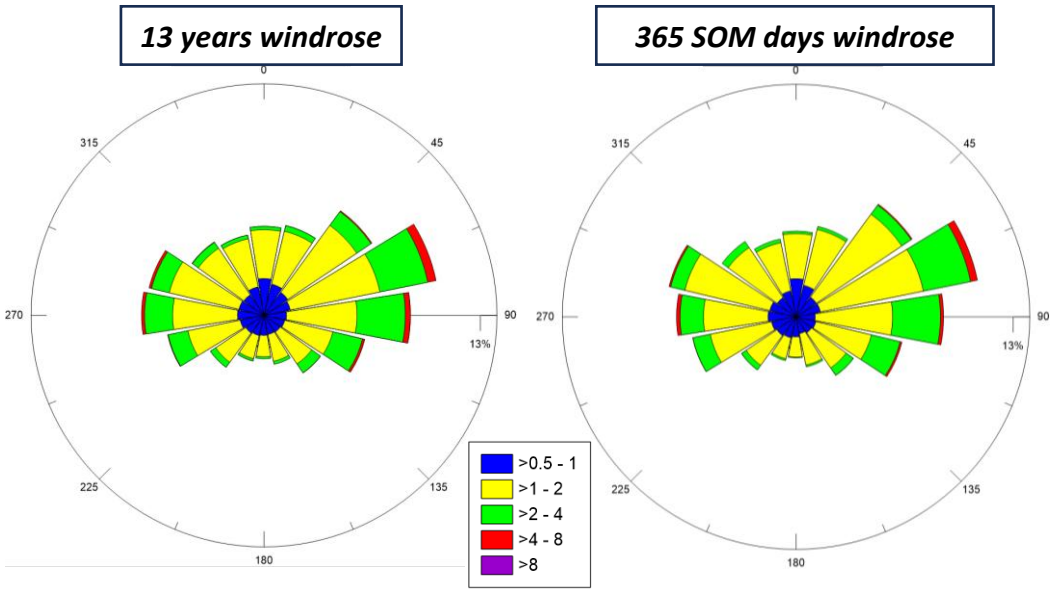
Data fusion and concentration time-series forecasting

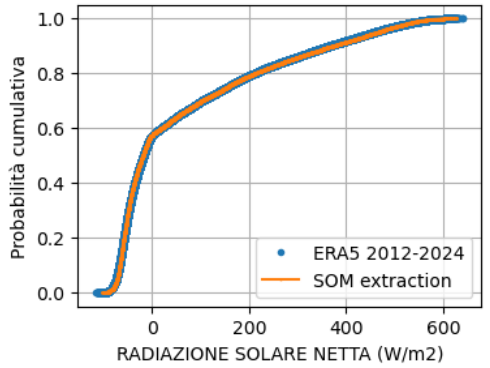
Meteorological days clustering for fast long-term dispersion simulations

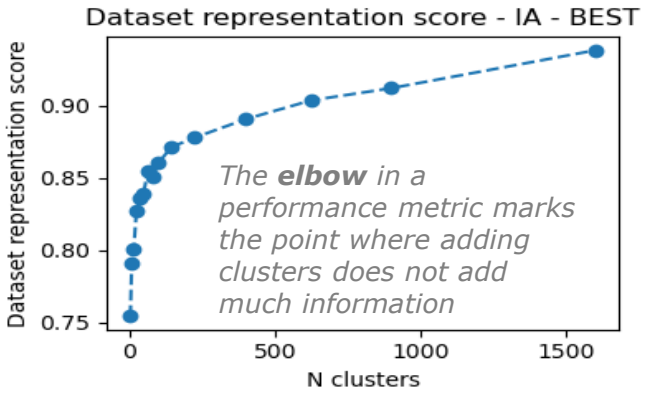
Need: Air quality assessments on long periods of time (e.g. in VIS for industrial plants or landfills). Need to fasten the generation of long-term statistics of concentration fields

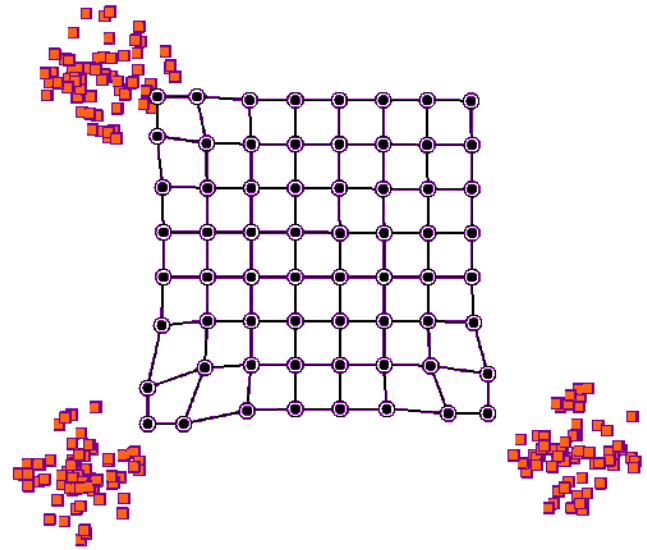
Solution: The *Self Organizing Maps (SOMs)* algorithm automatically separates the days into groups and find a representant for each group, based on similarity of hourly meteorological variables (*wind speed*, *wind direction*, *cloud cover*, *net radiation*)

Workflow: Run dispersion (*SPRAY*) only for representative days, then reconstruct long-term statistics from their frequency

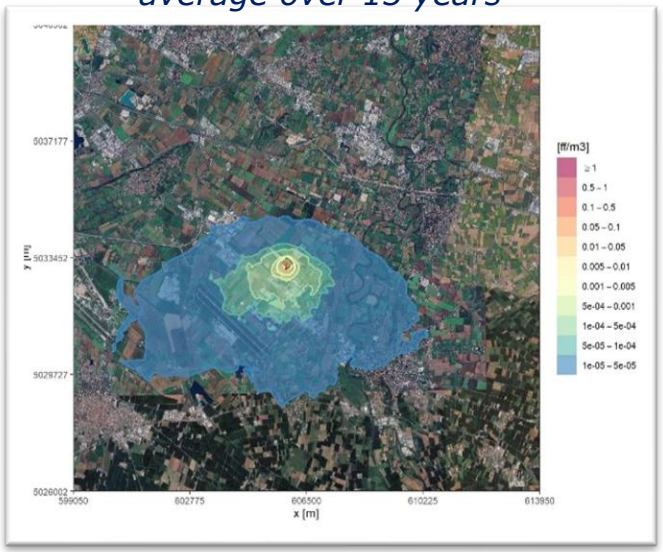








Output concentration average over 13 years



Some Studies delivered with SOMs #Ecologja

VIS Ecoeternit Landfill - VIS Tecnoinerti Landfill
VIS Acqua&Sole plant - AQ Assessment Cargill plant

Machine Learning at ARIANET

Unsupervised learning

Finding structure in unlabeled data

Meteorological-day clustering for faster long-term dispersion simulations

Road-segment classification for urban dispersion kernels and digital twins

Supervised learning

Learning from labeled data

Inverse modelling to rescale emission factors from observations

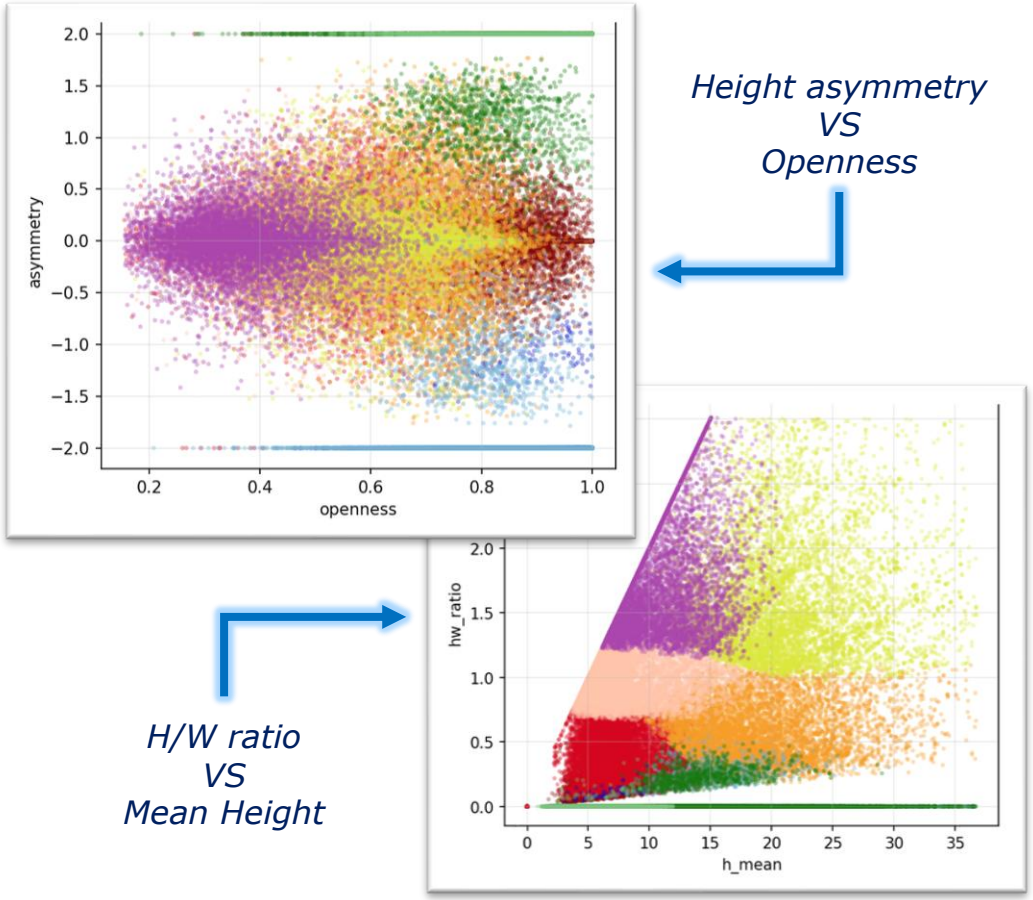
Mass-consistent concentration downscaling driven by geophysical proxies

Data fusion and concentration time-series forecasting

Unsupervised learning for roads clustering in an urban environment

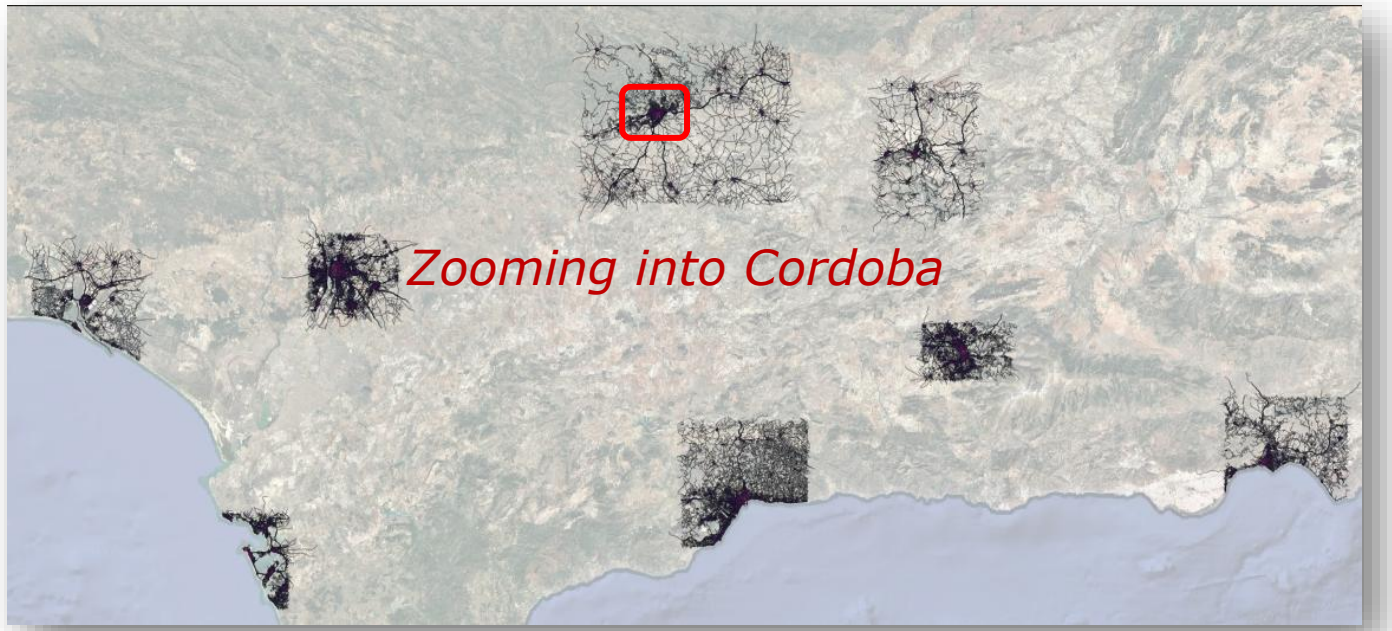
Goal: Building representative dispersion kernels in urban environment to be employed in air quality systems

Task: We leverage clustering algorithms (eg *K-Means*, *Gaussian mixtures, SOMs, HDBSCAN*) to classify road segments in the features space that characterize urban environment: *H/W ratio*, *Height asymmetry*, *Openness (1 - % built soil)*, …



Height asymmetry VS Openness

H/W ratio VS Mean Height



Shallow canyon Very low builtup	*Shallow canyon Low buildings*	*Medium-shallow canyon Medium builtup*	
Medium canyon Short buildings	*Deep canyon Low buitup*	*Deep canyon High builtup*	
Left asymmetric canyon Tall buildings	*Left asymmetric canyon Short buildings*	*Right asymmetric canyon Short buildings*	*Right asymmetric canyon Tall buildings*

Some projects involved

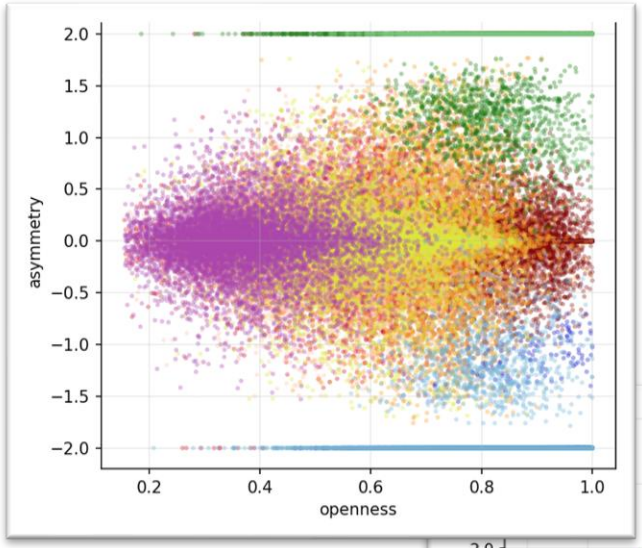
Andalucia Air Quality Forecasting system

CALLIOPE – Digital Twin of the city of Taranto

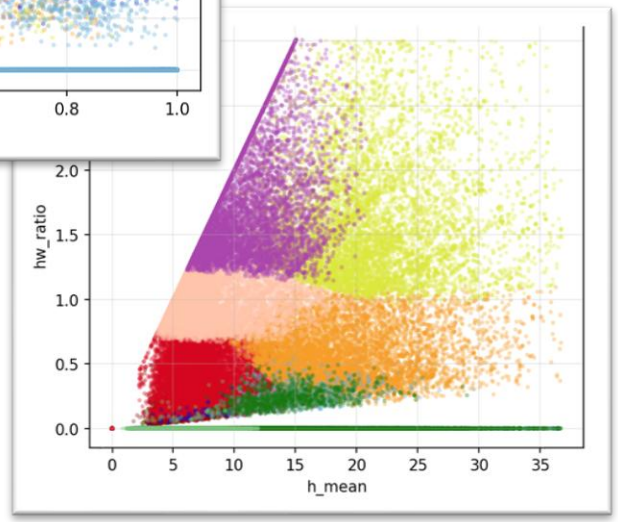
Unsupervised learning for roads clustering in an urban environment

Goal: Building representative dispersion kernels in urban environment to be employed in air quality systems

Task: We leverage clustering algorithms (eg *K-Means*, *Gaussian mixtures, SOMs, HDBSCAN*) to classify road segments in the features space that characterize urban environment: *H/W ratio*, *Height asymmetry*, *Openness (1 - % built soil)*, …



Height asymmetry VS Openness



H/W ratio VS Mean Height

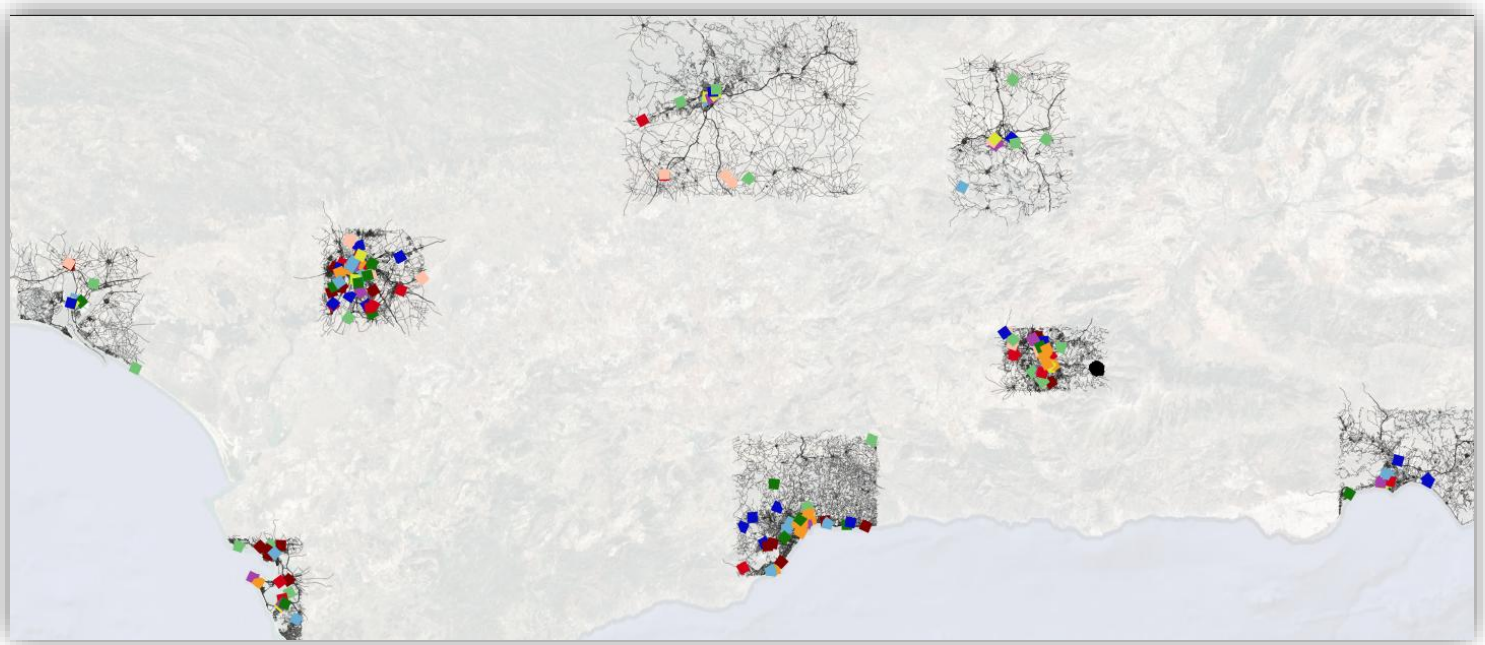
Shallow canyon Very low builtup	*Shallow canyon Low buildings*	*Medium-shallow canyon Medium builtup*	
Medium canyon Short buildings	*Deep canyon Low buitup*	*Deep canyon High builtup*	
Left asymmetric canyon Tall buildings	*Left asymmetric canyon Short buildings*	*Right asymmetric canyon Short buildings*	*Right asymmetric canyon Tall buildings*

Some projects involved

Andalucia Air Quality Forecasting system

CALLIOPE – Digital Twin of the city of Taranto

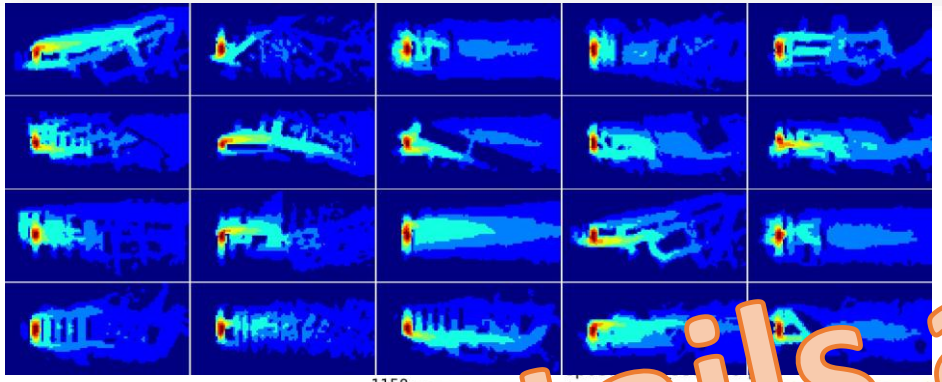
Unsupervised learning for roads clustering in an urban environment

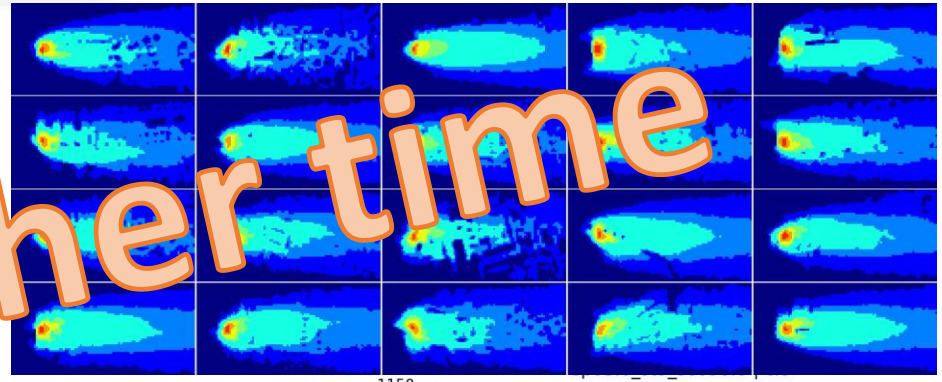


In each cluster we identify an *ensamble* of representative roads

For each ensamble, we build a set of dispersion kernels averaging *stationary PMSS concentrations*, in classified combinations of *Windspeed, Stability, Wind-road orientation*

Example of kernels building for two road types

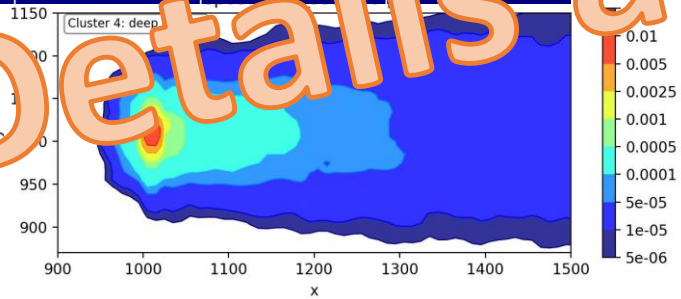




Details another time

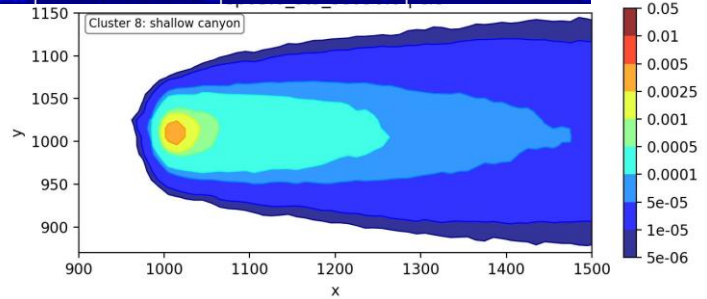
Deep canyon

*Windspeed **3 m/s***
*PGT **C***
*Wind-road angle **90°***



Shallow canyon

*Windspeed **3 m/s***
*PGT **C***
*Wind-road angle **90°***



Machine Learning at ARIANET

Unsupervised learning
Finding structure in unlabeled data

Meteorological-day clustering for faster long-term dispersion simulations

Road-segment classification for urban dispersion kernels and digital twins

Supervised learning
Learning from labeled data

Inverse modelling to rescale emission factors from observations

Mass-consistent concentration downscaling driven by geophysical proxies

Data fusion and concentration time-series forecasting

RemiScaler: Regularized inverse modelling for Emission reScaling

Goal: estimate source-specific emission rescaling factors by combining observed concentrations with modelled source-labelled concentrations.

Working assumptions

- Modelled concentrations are fully source-apportioned (typical of Lagrangian dispersion models).
- No dominant atmospheric chemistry breaks the approximate linearity between emissions and concentrations.

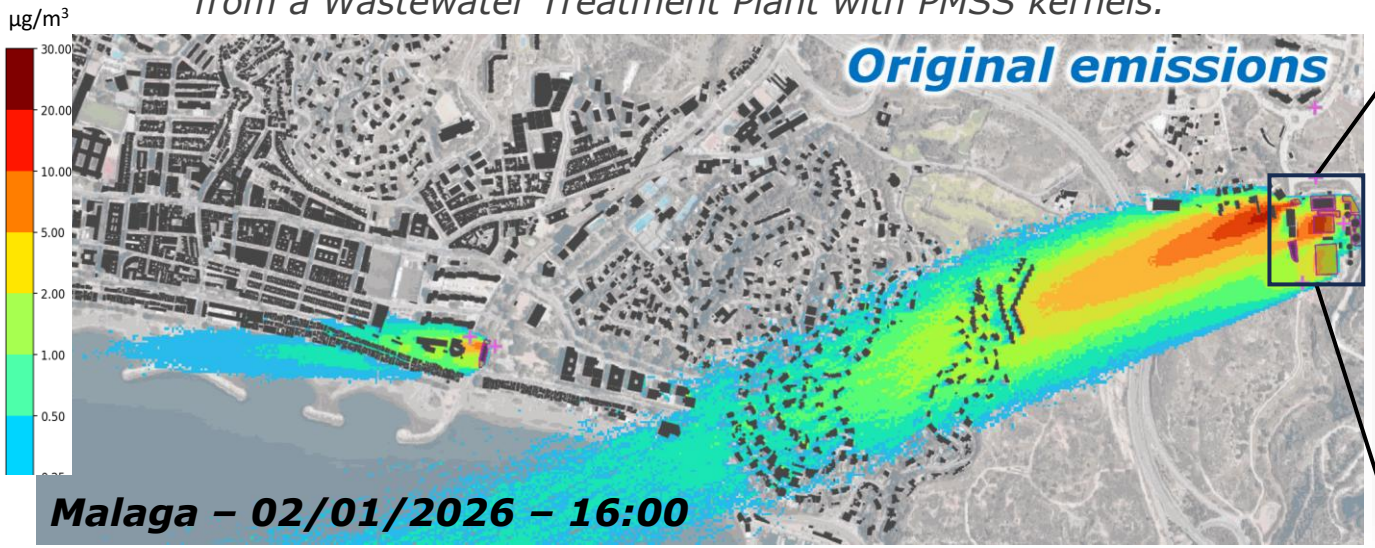
Inverse model

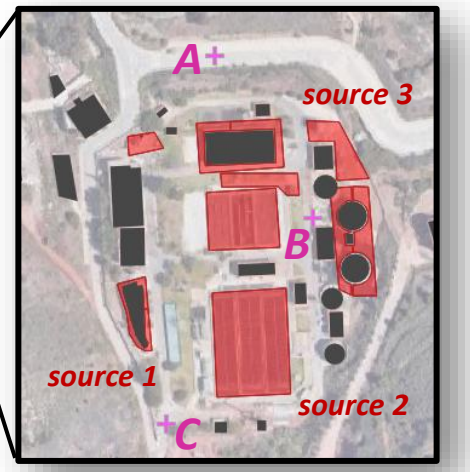
Ridge regression with non-negative constraints and *Tikhonov* regularization

$$\hat{\mathbf{k}} = \operatorname{argmin}_{\mathbf{k}} \left[\left\| W^{1/2}(\mathbf{Hk} - \mathbf{c_{obs}}) \right\|^2 + \lambda \|\mathbf{r} \odot (\mathbf{k} - \mathbf{1})\|^2 \right]$$

H: Modeled concentration transfer matrix
$\mathbf{c_{obs}}$: Observed concentrations - **W**: Observations weights matrix
k: Emission scaling factors
λ: Ridge hyperparameter - **r**: Source inertias to rescaling

Example application *Forecasting system of H_2S dispersion from a Wastewater Treatment Plant with PMSS kernels.*





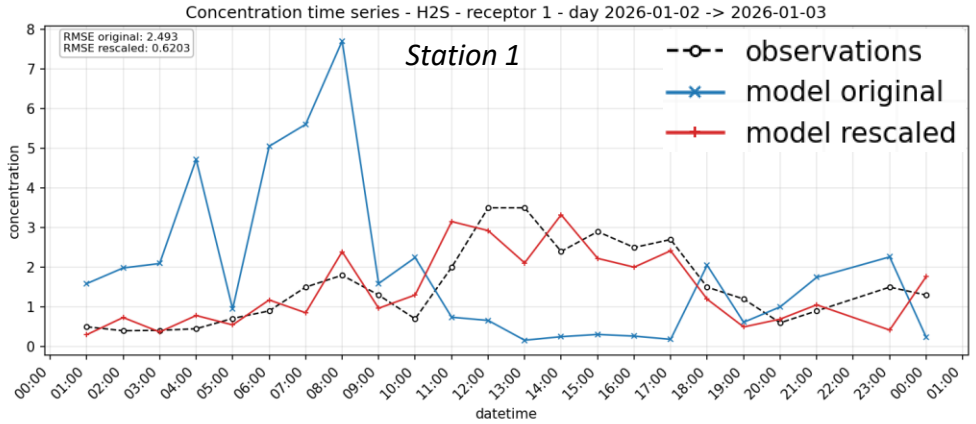
src 1	src 2	...	src n	H matrix
H_{11}	H_{12}	...	H_{1n}	*Obs point A, Time 0*
H_{21}	H_{22}	...	H_{2n}	*Obs point A, Time 1*
H_{31}	H_{32}	...	H_{3n}	*Obs point B, Time 0*
H_{41}	H_{42}	...	H_{4n}	*Obs point C, Time 0*
...	...	...	...	...
H_{m1}	H_{m2}	....	H_{mn}	*Obs / time m*

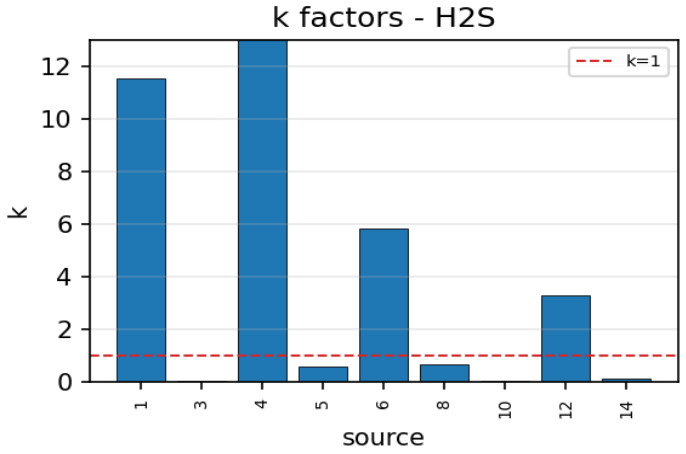
RemiScaler: Regularized inverse modelling for Emission Scaling

Why regularization is needed

The inverse problem is *ill-posed*: different emission combinations can reproduce the same concentrations*.
Regularization stabilizes the solution and keeps scaling factors close to prior values.

Without regularization

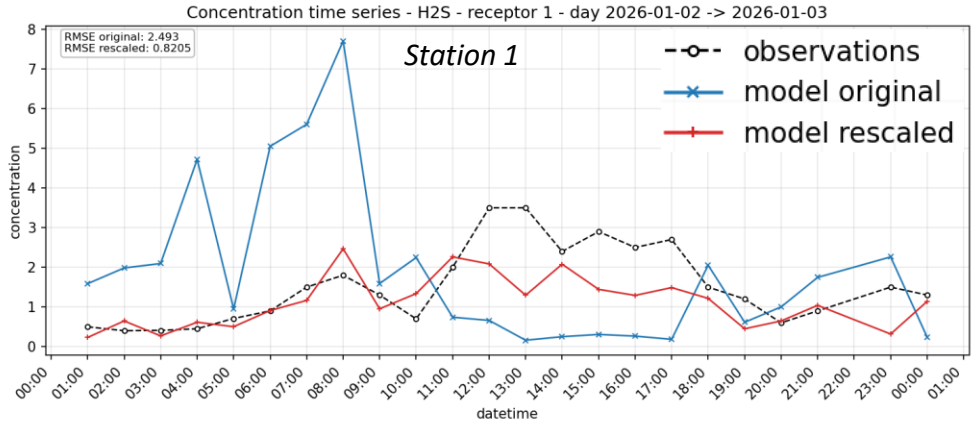


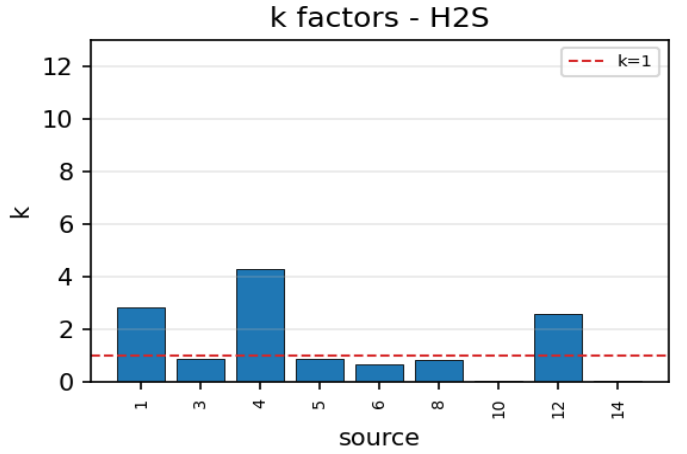


RMSE original emi
2,49 µg/m³

RMSE rescaled emi
0,62 µg/m³

With regularization λ=0,5





RMSE original emi
2,49 µg/m³

RMSE rescaled emi
0,82 µg/m³

** Practical lever: more observations in time help constrain the solution and reduce degeneracy between sources.*

Machine Learning at ARIANET

Unsupervised learning
Finding structure in unlabeled data

Meteorological-day clustering for faster long-term dispersion simulations

Road-segment classification for urban dispersion kernels and digital twins

Supervised learning
Learning from labeled data

Inverse modelling to rescale emission factors from observations

Mass-consistent concentration downscaling driven by geophysical proxies

Data fusion and concentration time-series forecasting

Mass-consistent concentration downscaling with Random Forest

Idea – Use coarse-grid concentration together with fine-grid proxies (*roads*, *land use*, *population*, …) to **redistribute** concentration from 1 km to 100 m

Training – A **Random Forest** is trained on concentration using geophysical proxies as features

$$C_{ML} = RF(proxies)$$

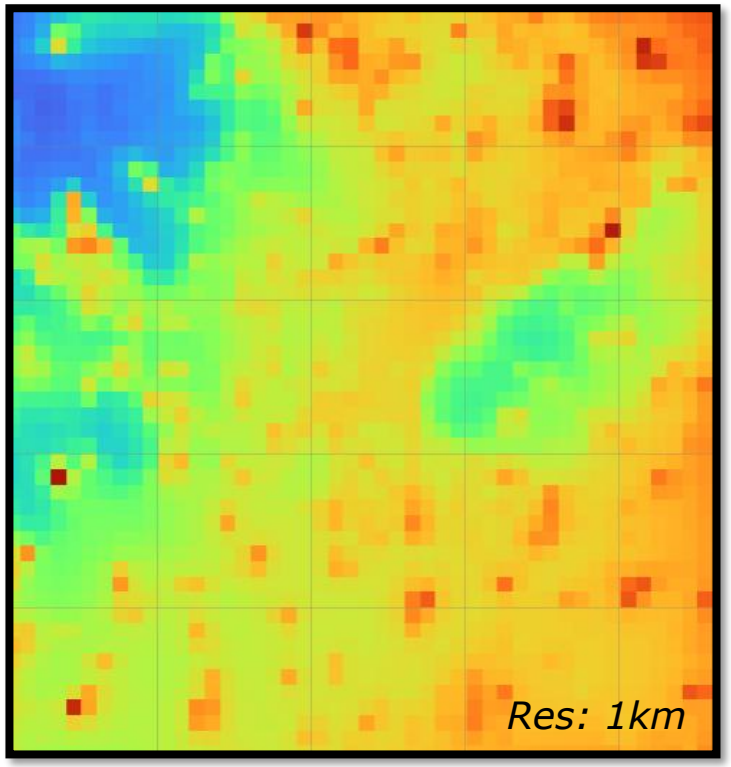
Mass constraint – Within each 1 km FARM cell, the average of the 100m predictions matches the original coarse-cell concentration

$$\hat{C}_{ij}^{ML} = \frac{C_k^F N_i N_j}{\sum_{ij} C_{ij}^{ML}} C_{ij}^{ML}$$

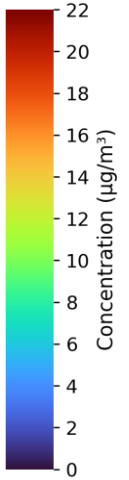
Interpretation – ML does not model concentration at 100 m; it learns **redistibution weights** on the fine grid, while preserving the mass of the original FARM cell

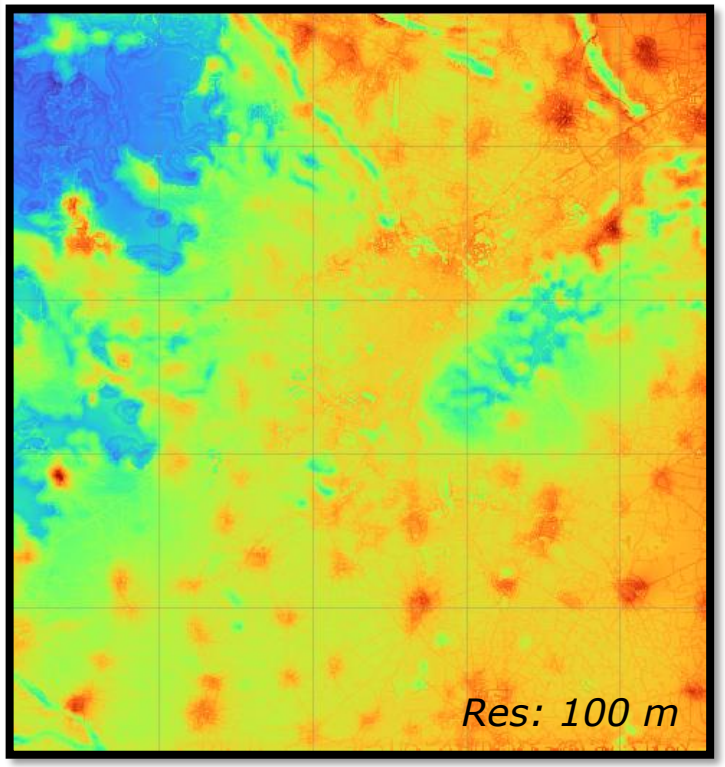
Some projects involved

MISTRAL Horizon: Emission scenarios platform on Rybnik
Andalucia Air Quality Forecasting system



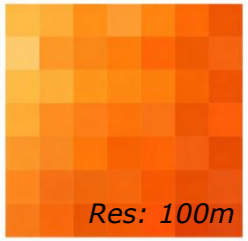
Annual average PM10 concentration over Turin





FARM cell

Mass-consistent downscaling



Machine Learning at ARIANET

Unsupervised learning
Finding structure in unlabeled data

Meteorological-day clustering for faster long-term dispersion simulations

Road-segment classification for urban dispersion kernels and digital twins

Supervised learning
Learning from labeled data

Inverse modelling to rescale emission factors from observations

Mass-consistent concentration downscaling driven by geophysical proxies

Data fusion and concentration time-series forecasting

Mal-Air: Random Forest for Data Fusion

Risoluzione target: 1km

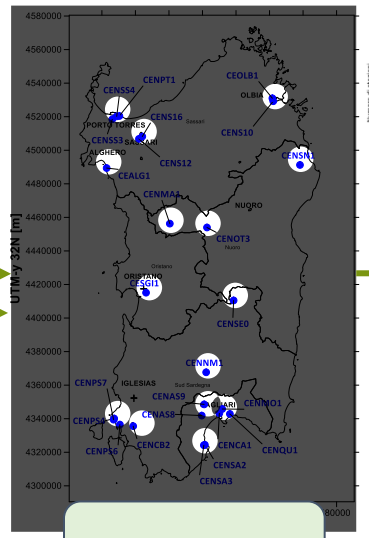
Osservazioni
Serie temporali orarie/giornaliere degli inquinanti
PM10, NO2, PM2.5

Predittori - Proxies

Spazio-termporali
FARM/CAMS (# km)
Leaf Area Index (LAI)
Meteorological (ERA5)

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

Spaziali (stazionari)
Length roads in buf.
Population density
Land use
…



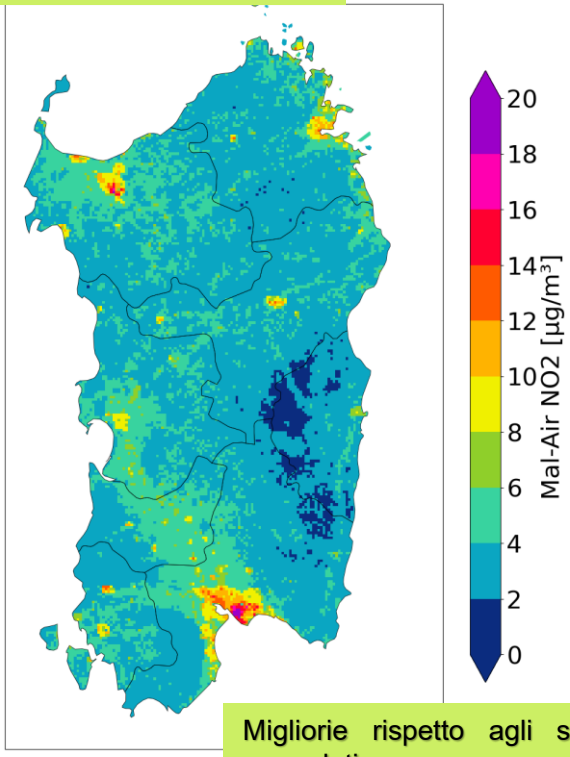
Allenamento (2024)

Pre-processore

Random Forest

Inferenza (2024)

7 modelli (PM10, PM2.5 NO2, O3, SO2, C6H6, CO)



Migliorie rispetto agli studi precedeti
- Proxy meteo ad alta risoluzione
- Prodotti satellitari
- Approccio multi predittore CAMS applicato a tutti gli inquinanti

Allenamento
Inferenza

Dati di Input

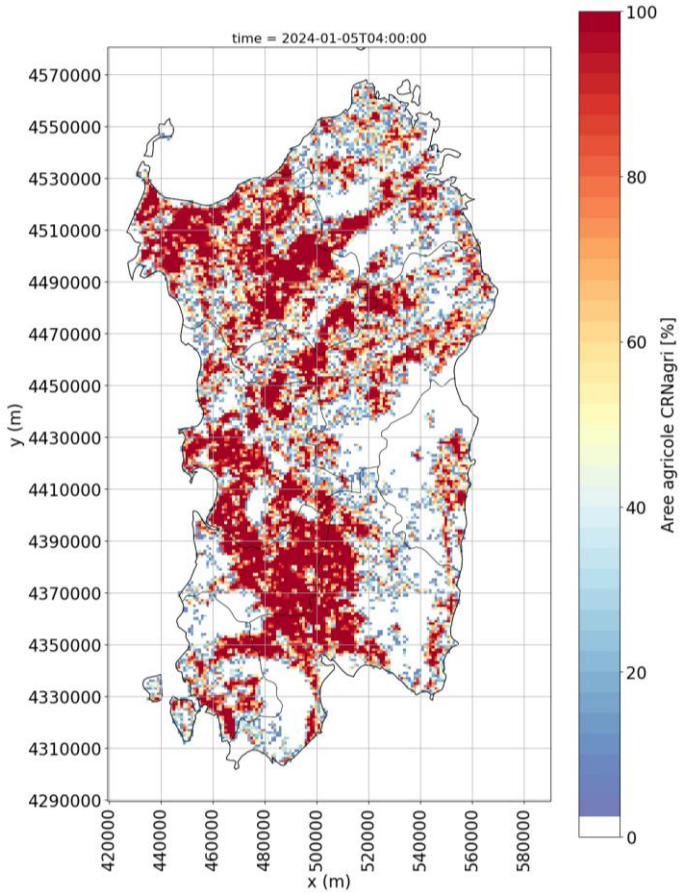
Modello / processo

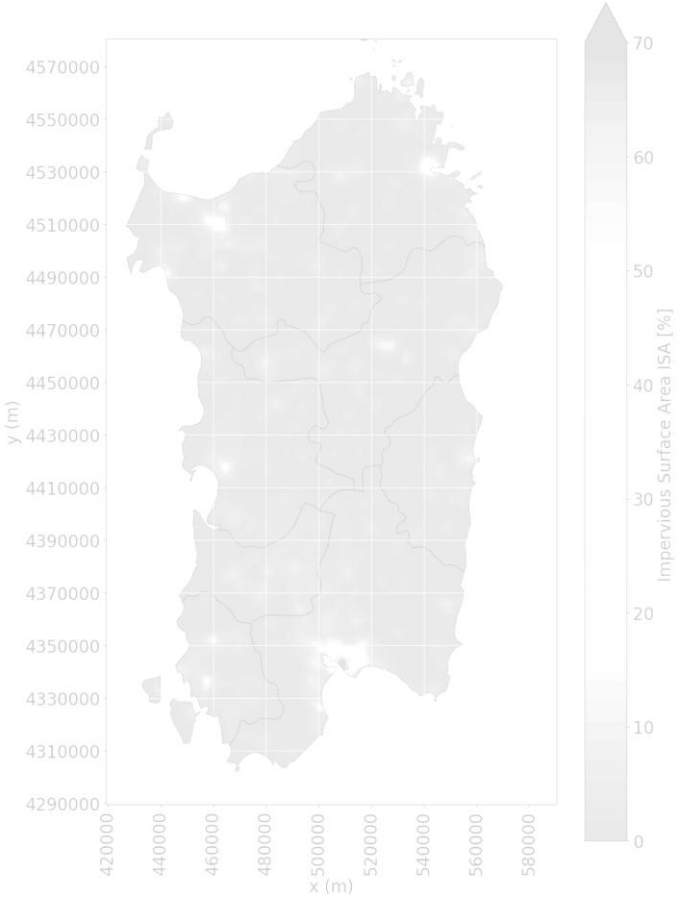
Some projects involved:
Valutazione modellistica dello stato di Qualità dell'aria sulla Regione Sardegna per l'anno 2024; SPOTT; CALLIOPE

PREDITTORI

Classi di uso del suolo – Corine Land Cover 2018
Aree urbane (tessuto urbano continuo e discontinuo)
Aree industriali e commerciali
Aeroporti e infrastrutture aeroportuali
Altre superfici artificiali
Aree agricole
Aree forestali
Prati, aree arbustive e superfici naturali aperte (pianure brulle, spiagge naturali)
Risoluzione orig: 100 m

Predittore statico

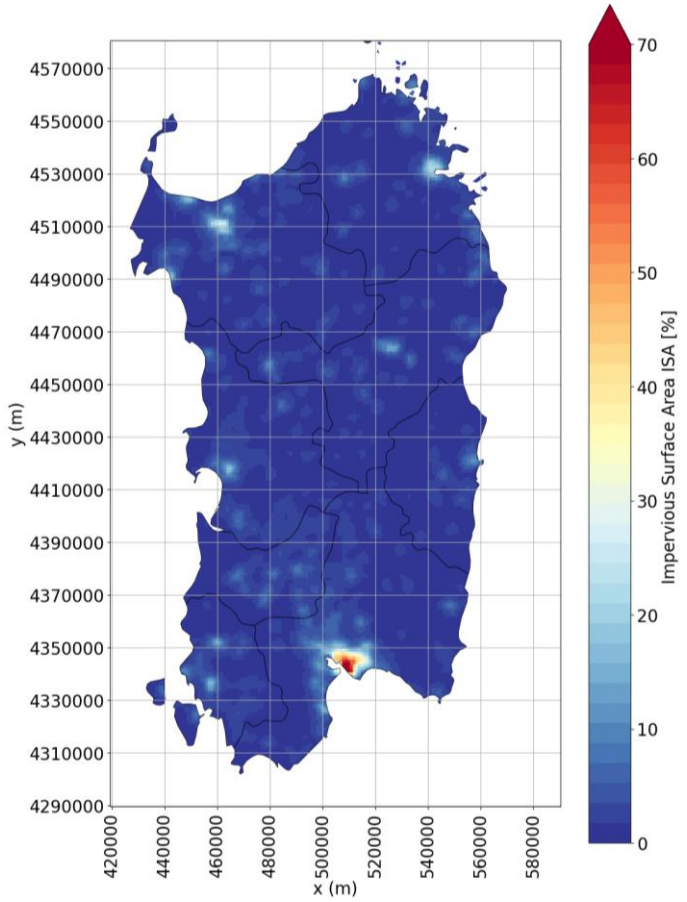


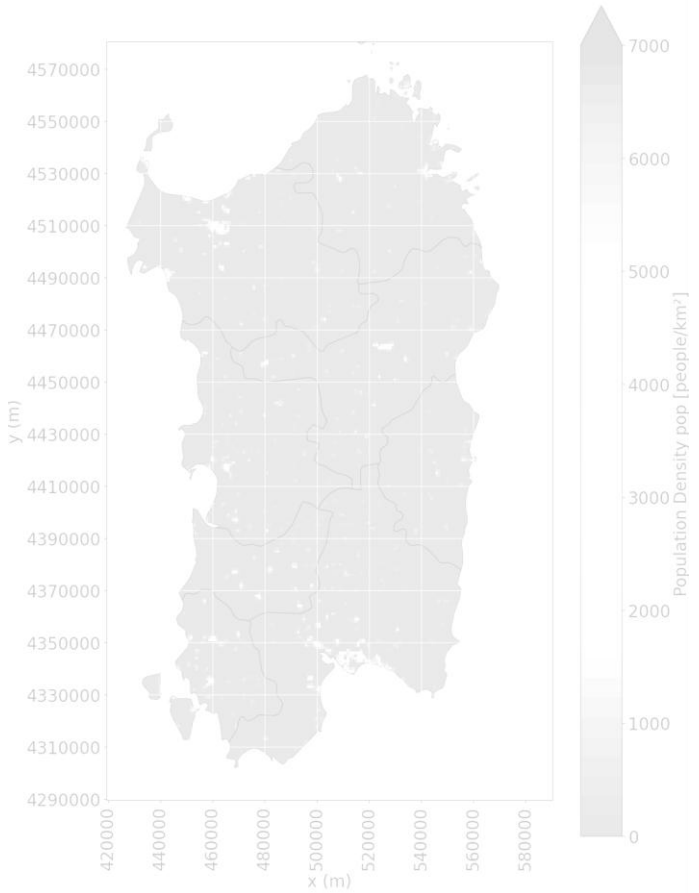


PREDITTORI

Inventario globale della distribuzione spaziale e della densità della superficie impermeabile costruita

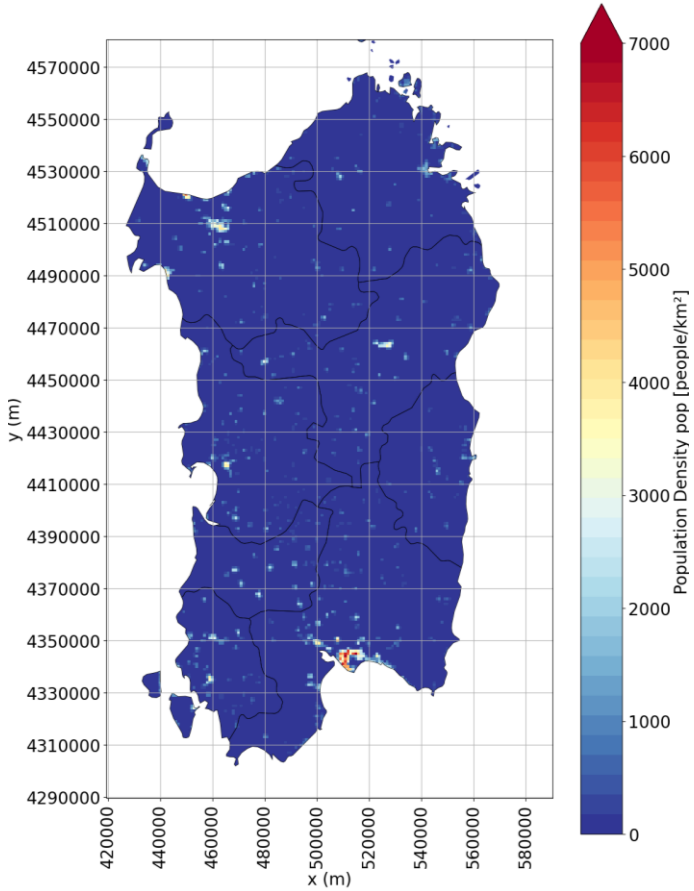
Predittore Statico

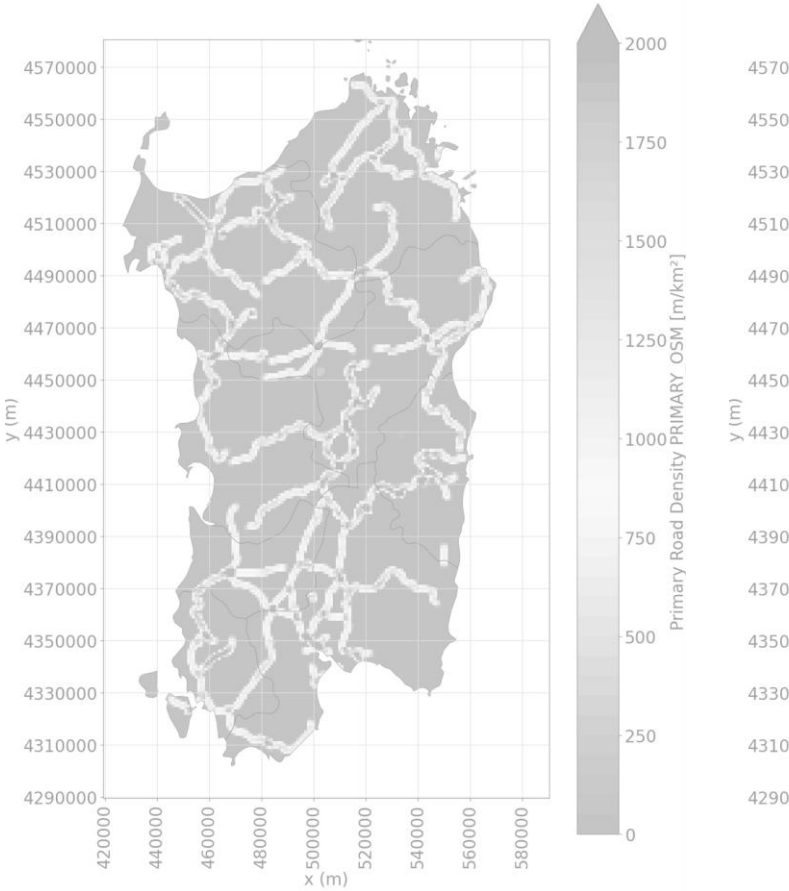




PREDITTORI

Aree agricole CRNagri [%]
Impervious Surface Area ISA [%]
y (m)
x (m)


Population Density pop [people/km²]
y (m)
x (m)

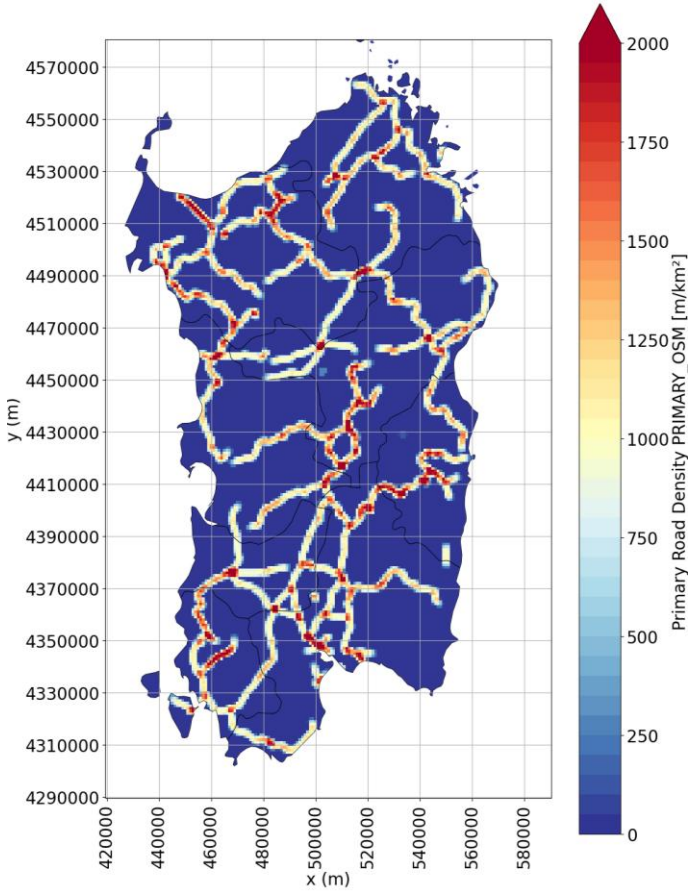

Primary Road Density PRIMARY_OSM [m/km²]
y (m)
x (m)

PREDITTORI

Per ogni cella della griglia è stata stimata la densità stradale, espressa come metri di strada per chilometro quadrato [m/km²], distinguendo le seguenti tipologie di infrastrutture

- **PRIMARY**
- **SECONDARY**
- **TERTIARY**
- **TRUNK-MOTORWAY**

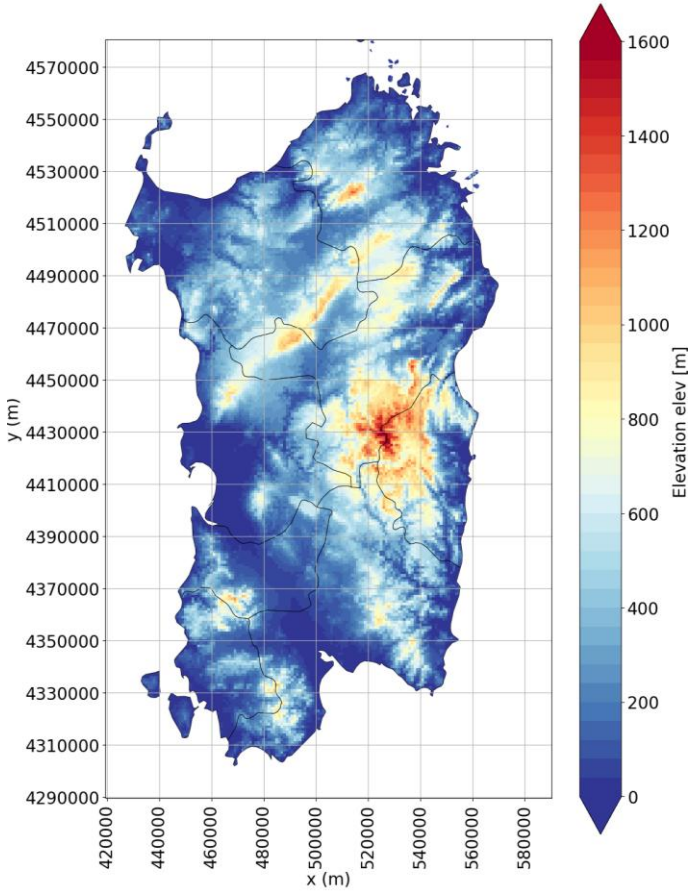
Fonte: OpenStreetMap
Predittore Statico

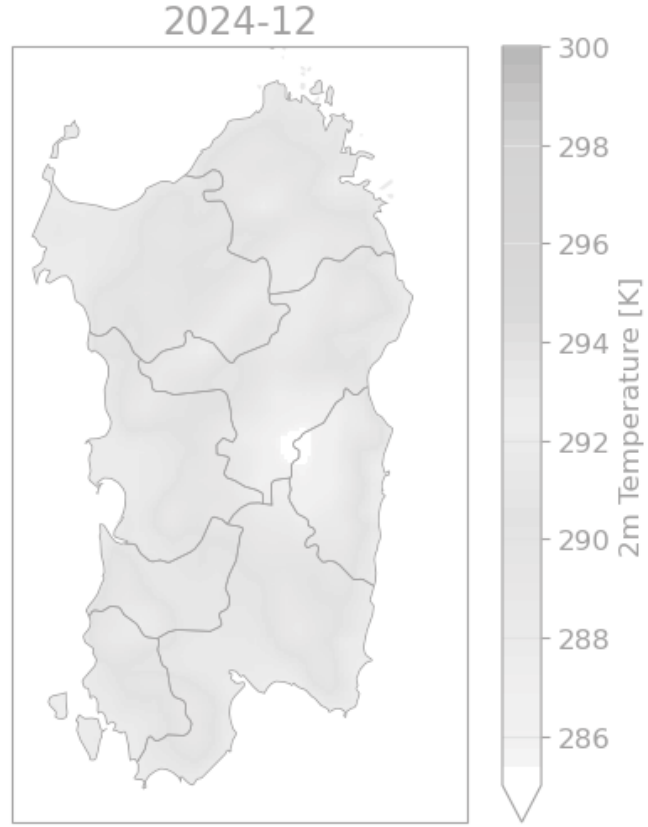


PREDITTORI

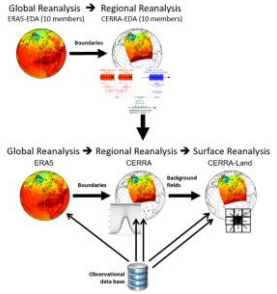
Elevazione media (ELEV) ottenuta dal Servizio Copernicus Land Monitoring Service (CLMS) - European Digital Elevation Model

Predittore Statico





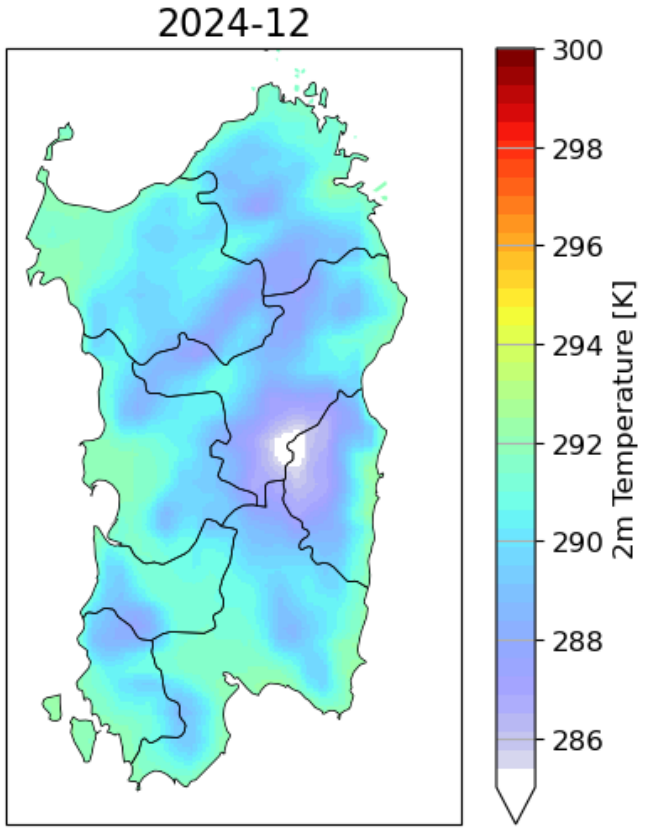
PREDITTORI

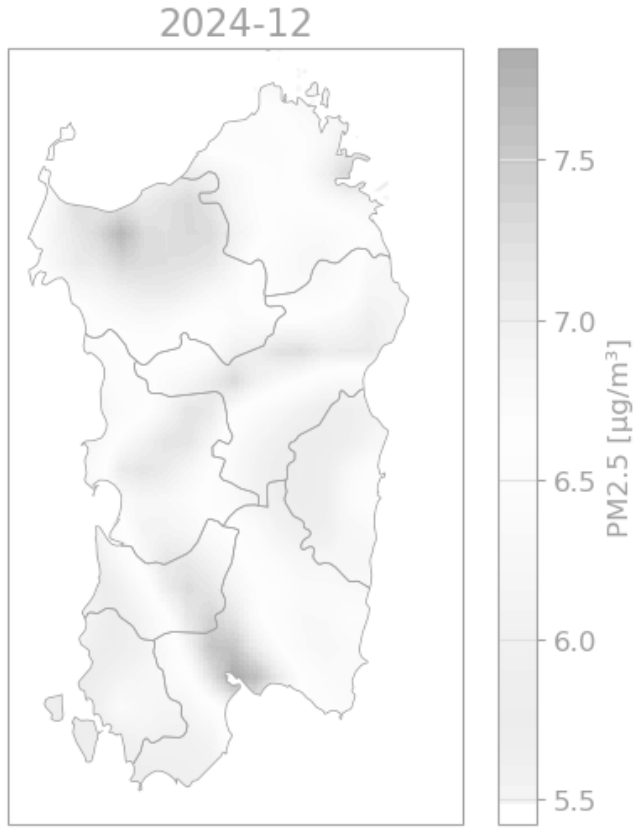


CERRA – [Copernicus regional reanalysis for Europe (CERRA) | Copernicus]

Dataset sviluppato con il sistema numerico **HARMONIE-ALADIN**, forzato ai bordi da **ERA5**

- **Configurazione utilizzata nello studio**
- **Prodotto**: *CERRA sub-daily regional reanalysis on single levels*
- **Risoluzione spaziale**: ~5.5 km
- **Risoluzione temporale**: 3 ore
- **Periodo**: fino a novembre 2024 *(dicembre 2024: campi ERA5)*
- **Variabili impiegate**
- Vento a 10 m (velocità e direzione, espressa come componenti seno/coseno), t2m

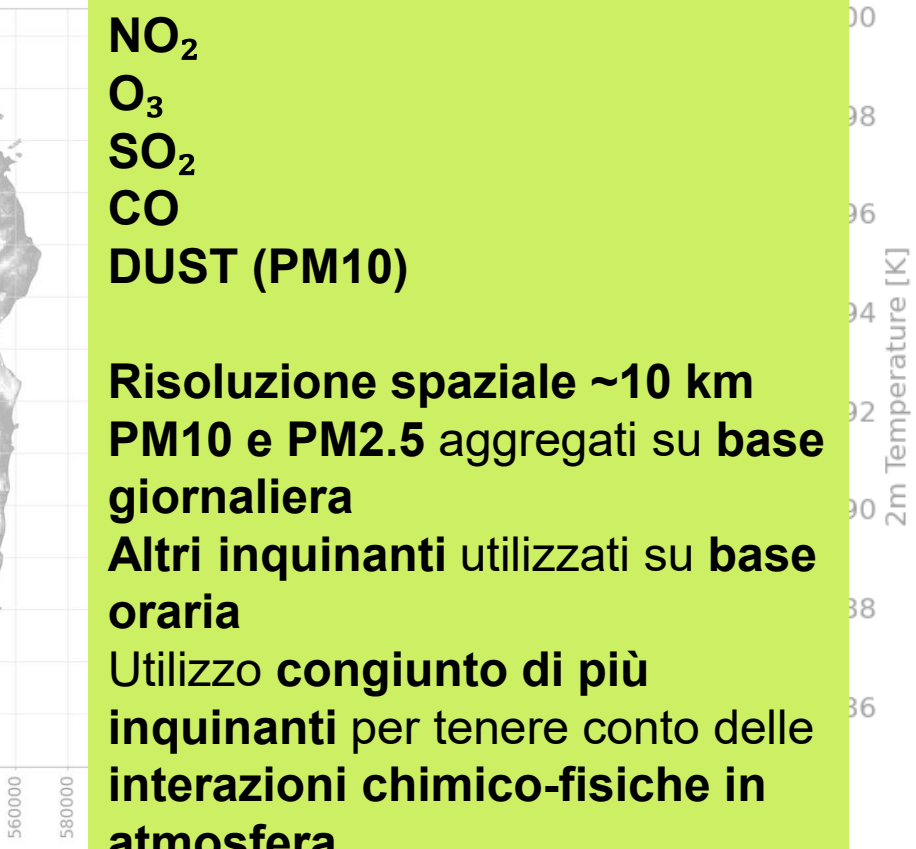


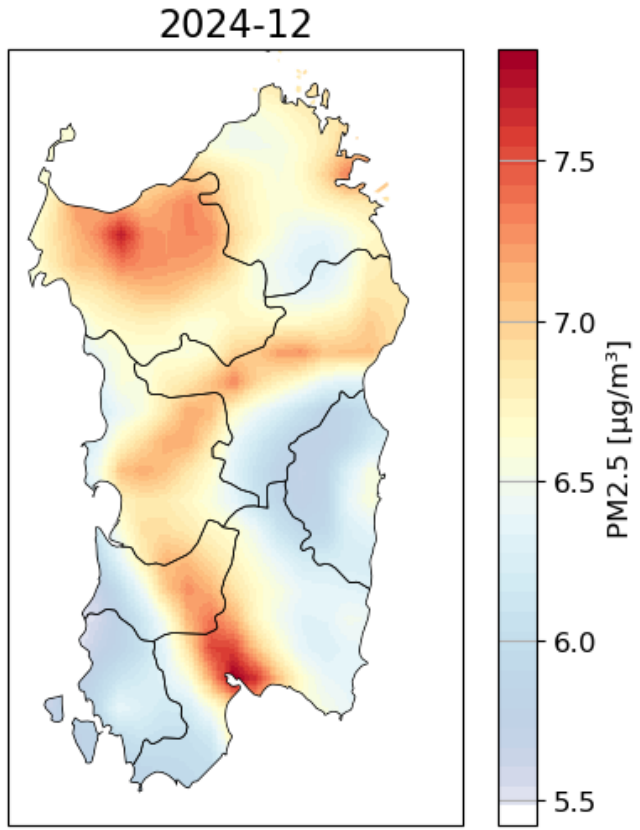


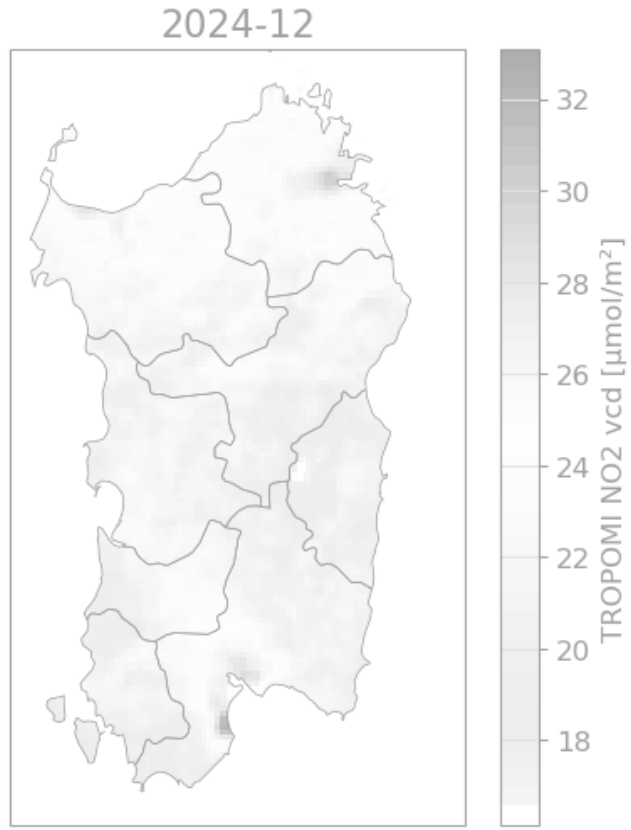
PREDITTORI

Campi di concentrazione **orari** di:
PM10
PM2.5
NO_2
O_3
SO_2
CO
DUST (PM10)

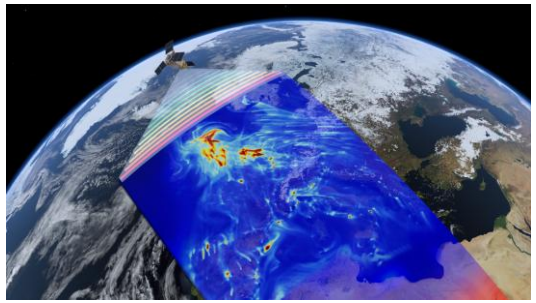
Risoluzione spaziale ~10 km
PM10 e PM2.5 aggregati su **base giornaliera**
Altri inquinanti utilizzati su **base oraria**
Utilizzo **congiunto di più inquinanti** per tenere conto delle **interazioni chimico-fisiche in atmosfera**







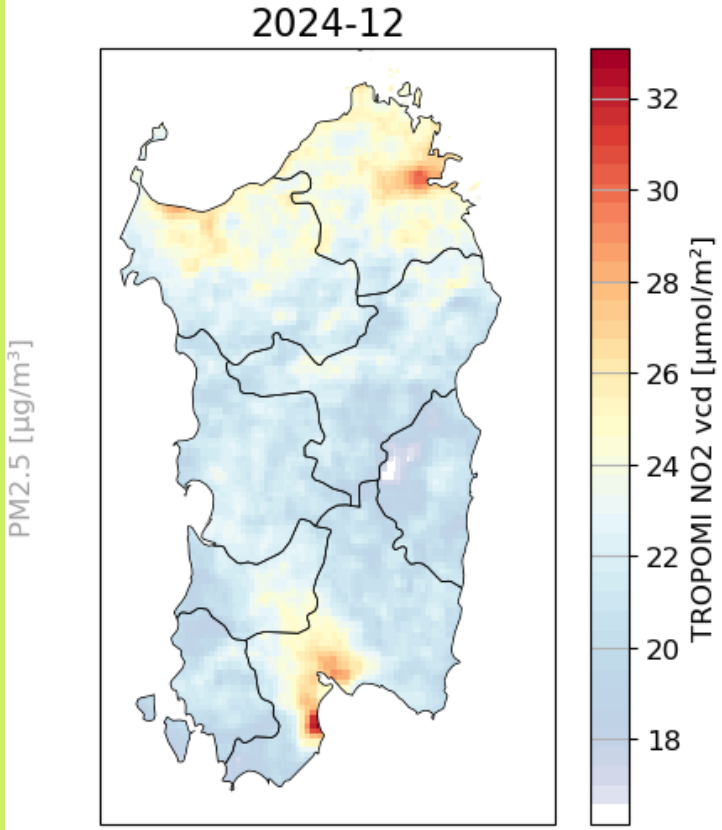
PREDITTORI



NO_2 troposferico da Sentinel-5P / TROPOMI (Copernicus)

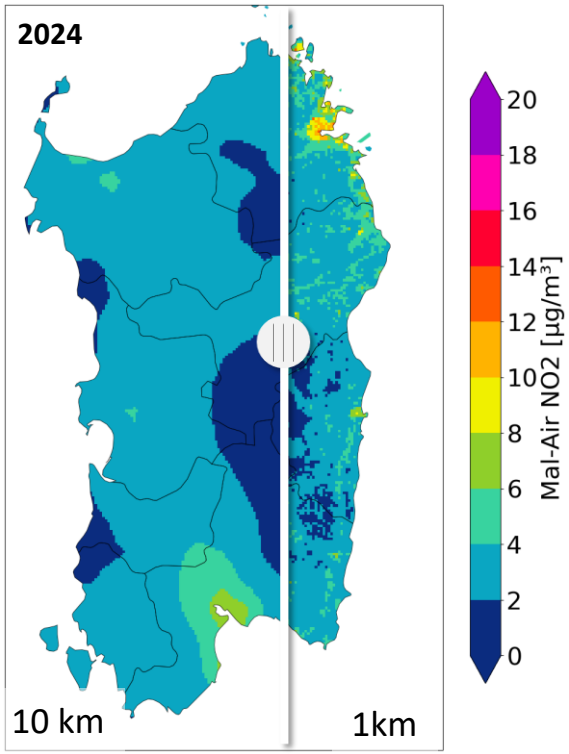
- Colonna troposferica di **NO_2** (µmol/m²), **media mensile – anno 2024**
- **Risoluzione spaziale ~5.5 × 3.5 km²** (al nadir)
- Evidenzia i principali **hotspot emissivi**:
  - aree urbane e industriali
  - principali assi viari
- Media calcolata su **tutte le osservazioni mensili** disponibili sulla regione
- **Orario di passaggio fisso** del satellite (~11:00–15:00 locali):
  - rappresenta la **condizione tipica diurna**
  - **non** una media sull'intero ciclo giornaliero
- Dati disponibili dal **Copernicus Data Space**

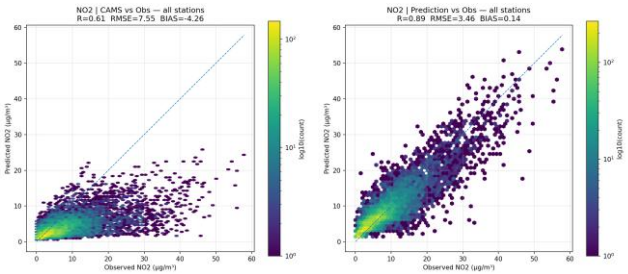
Predittore dinamico



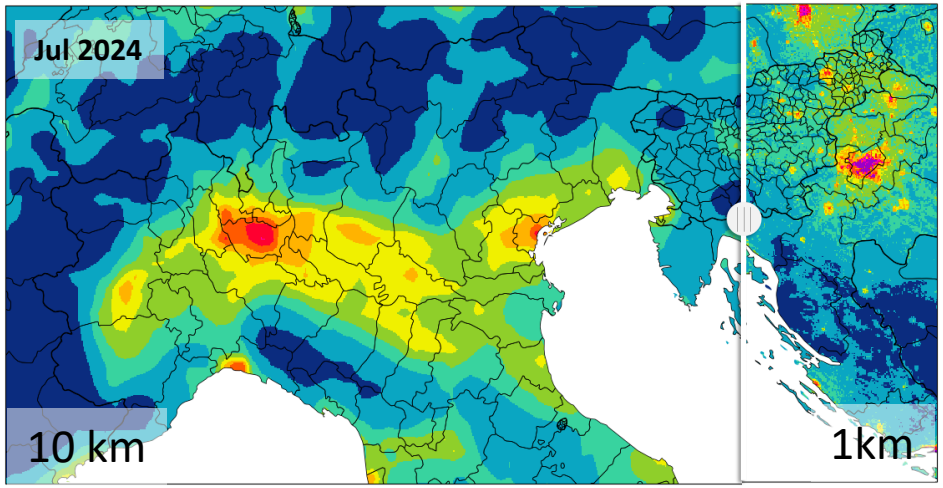
Mal-Air: Random Forest for Data Fusion

CAMS rean Mal-Air

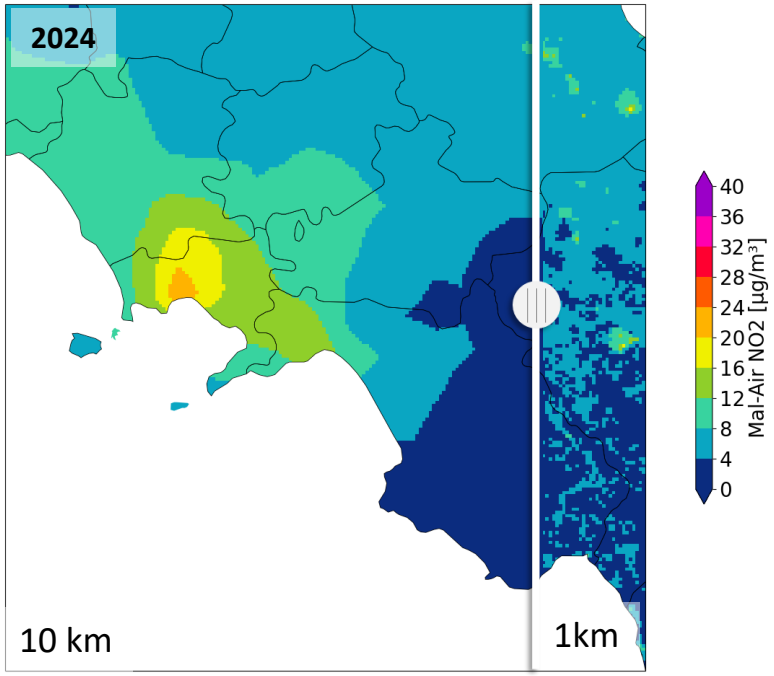


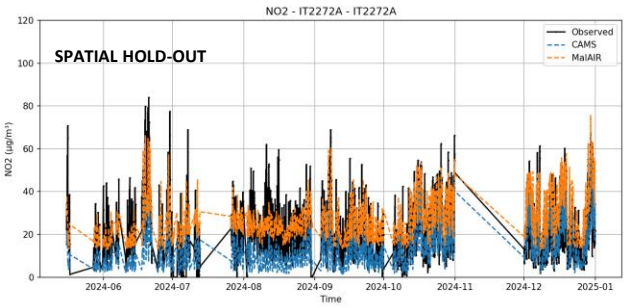


CAMS rean Mal-Air



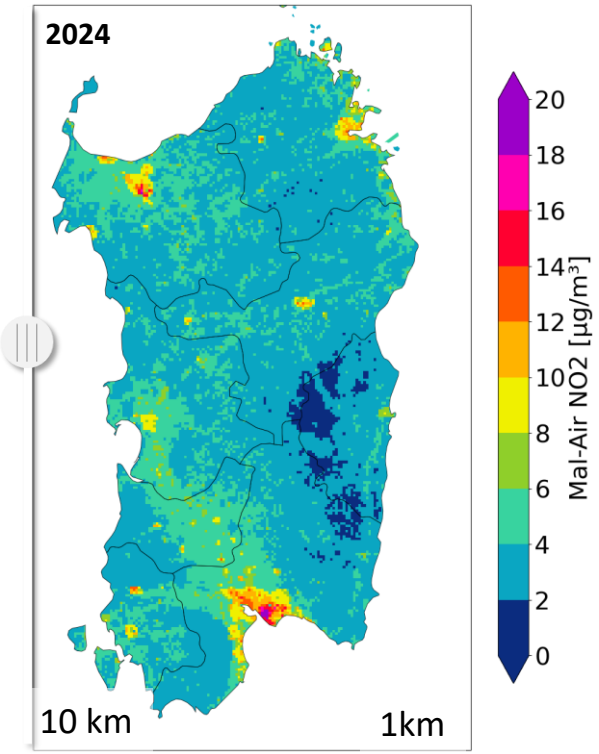
CAMS rean Mal-Air



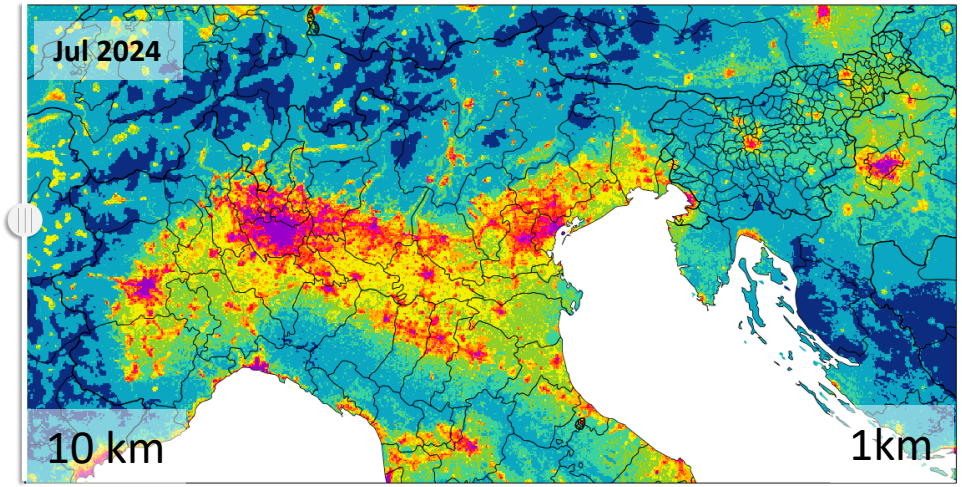


Mal-Air: Random Forest for Data Fusion

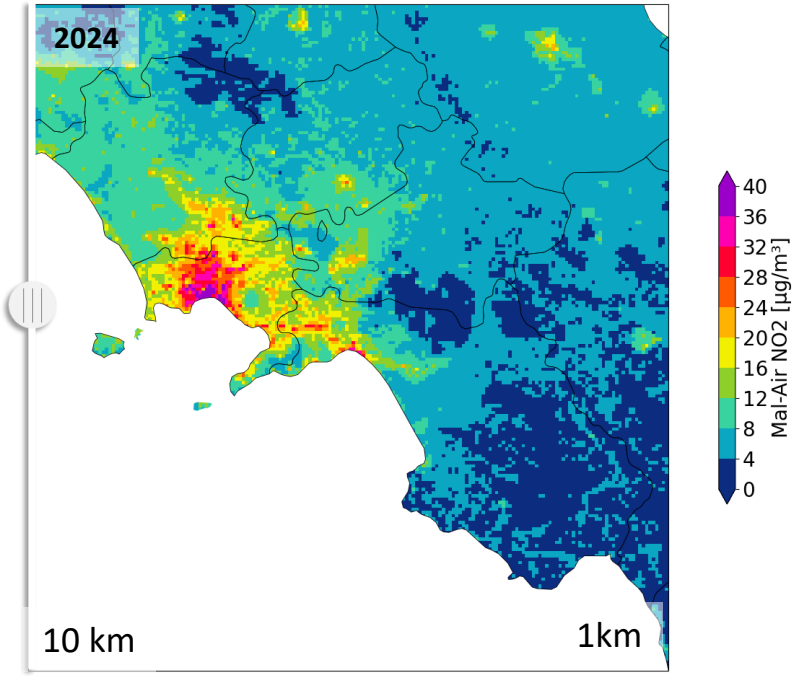
Mal-Air

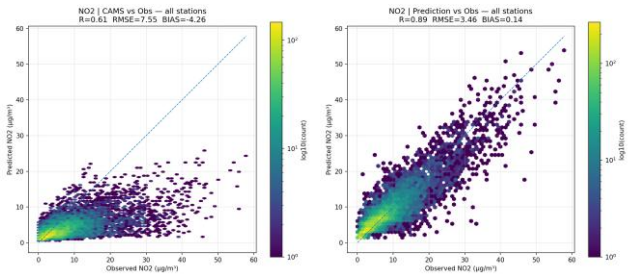


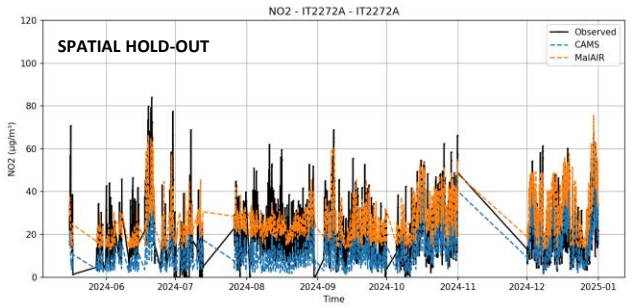
Mal-Air



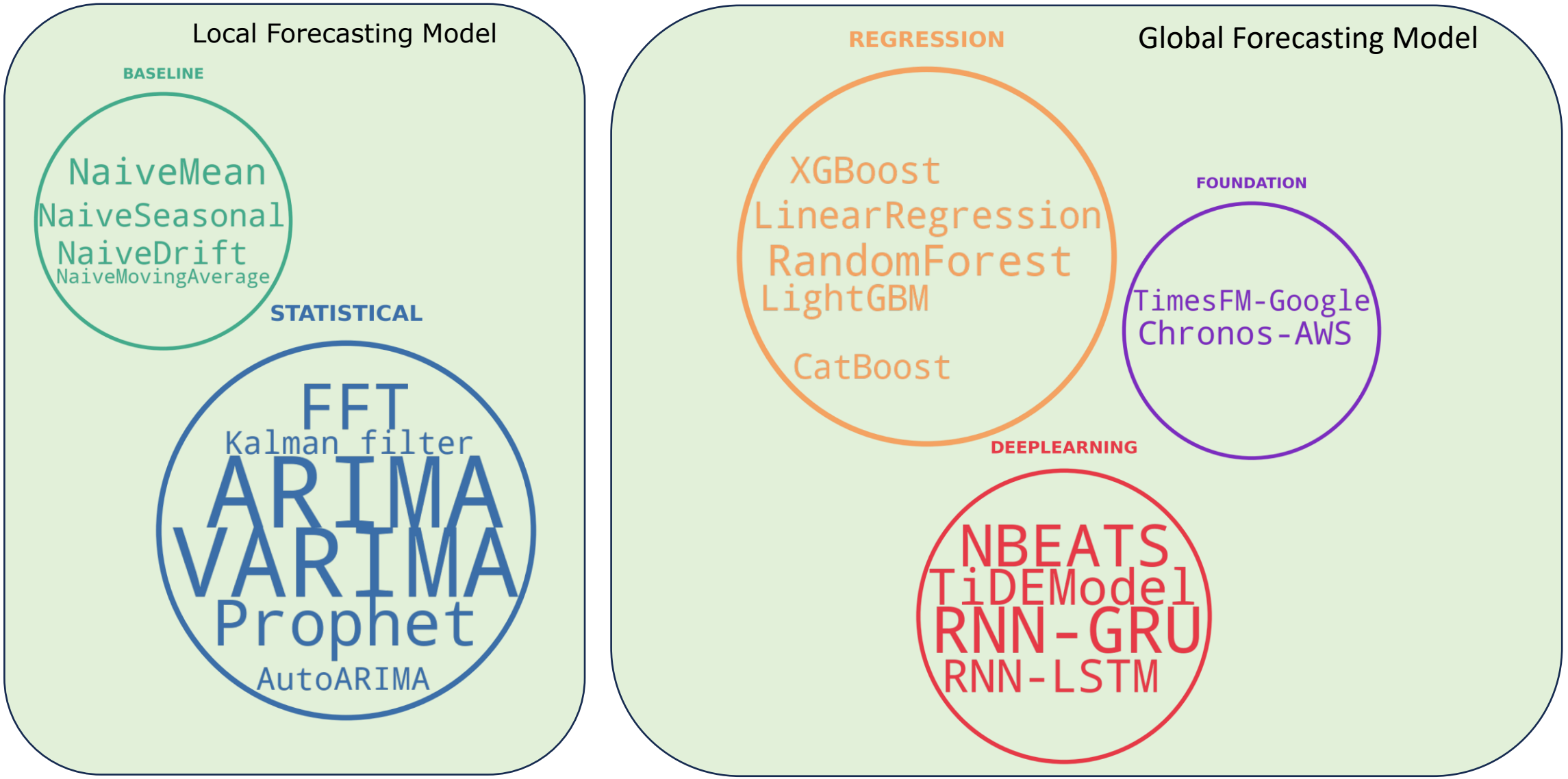
Mal-Air







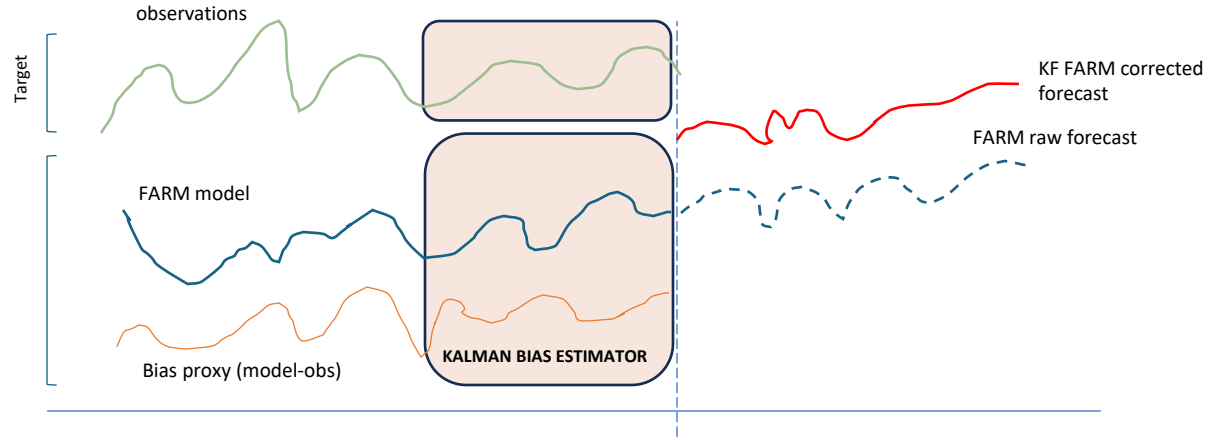
Concentration time series forecasting



Concentration time series forecasting with Neural Networks

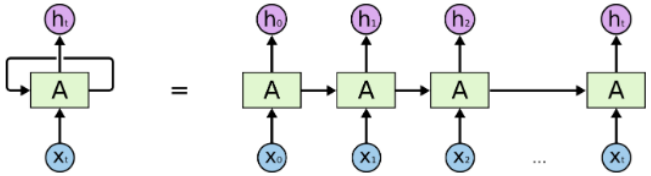
Kalman Filtering

linear, sequential, Gaussian assumptions

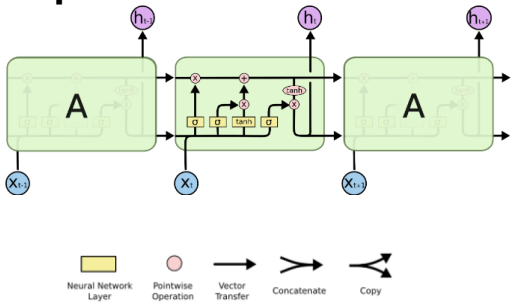


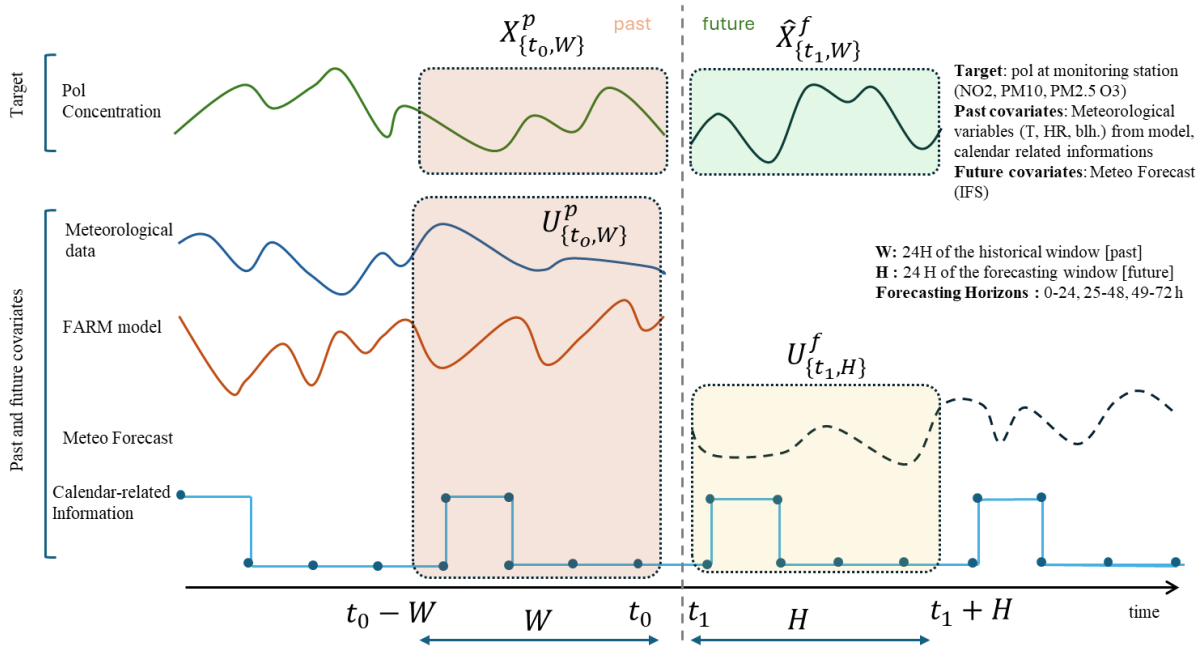
Recurrent Neural Networks

nonlinear, data-driven, multi-feature



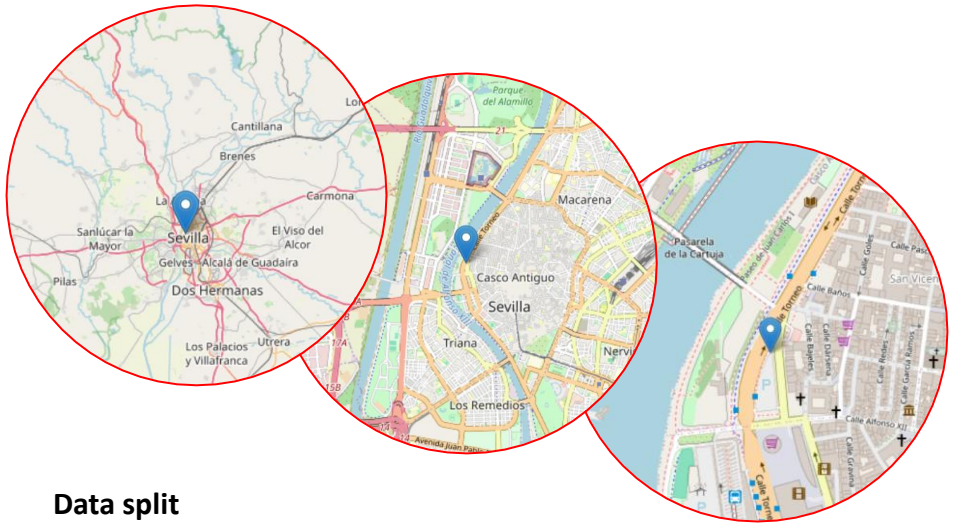
The problem of Long-Term Dependencies - LSTM





Recurrent Neural Network - example

ES0890 – TORNEO Sevilla [Traffic]



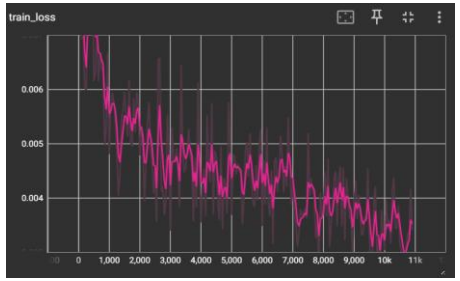
Data split

- Training : 2023 -> Aug 2024
- Validation : Aug 2024 -> Dec 2024
- Test : 2025 -> Jan 2026

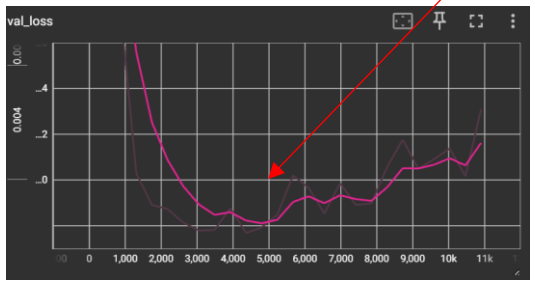
Input

- Future covariates (+48h): CAMS reg fc (NO2, PM10)
- Past covariates (-24h): CAMS reg fc (NO2, PM10); EEA obs
- Calendar related info: hours_x, hours_y

Loss on training set

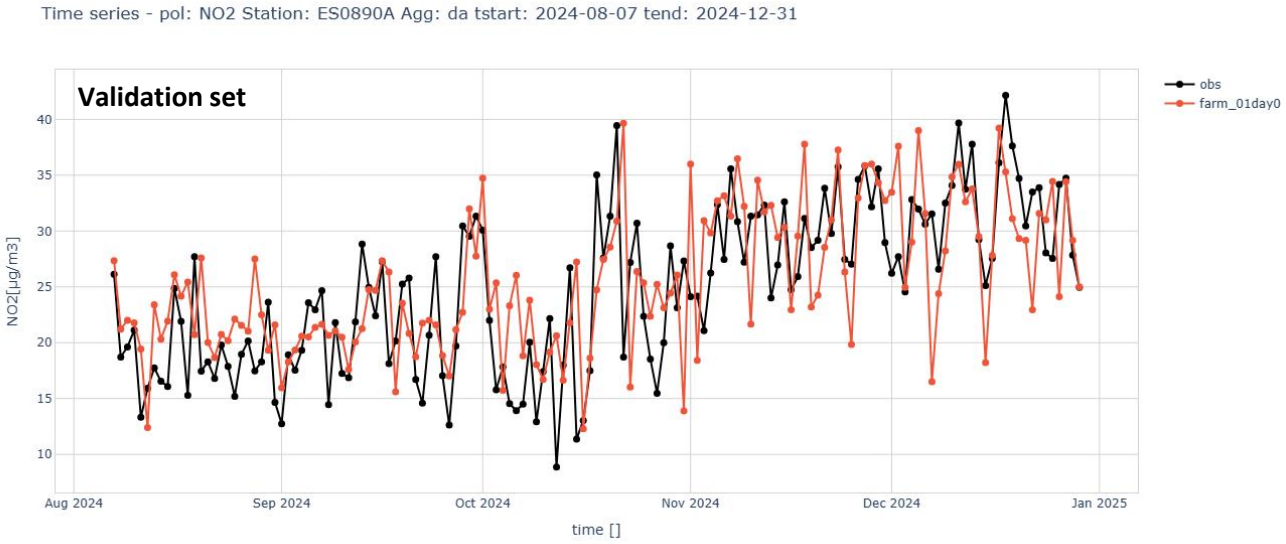


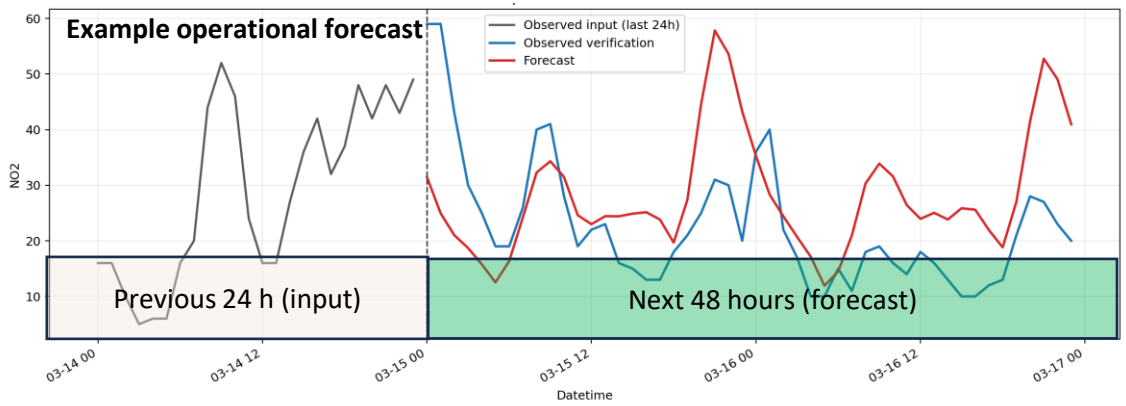
Loss on validation set



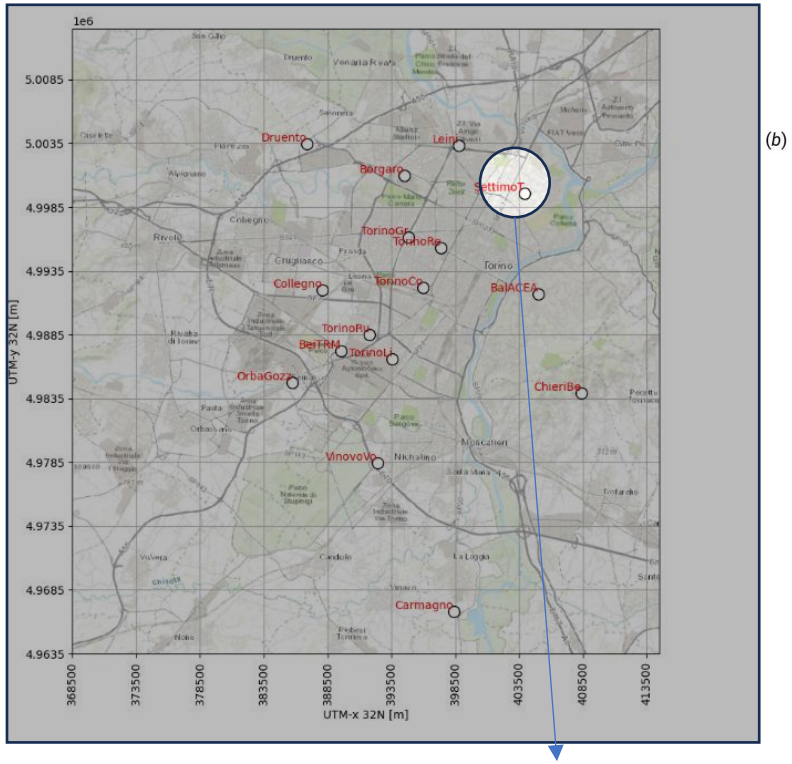
Model : **BlockRNNModel –LSTM**

Loss **: MSE**

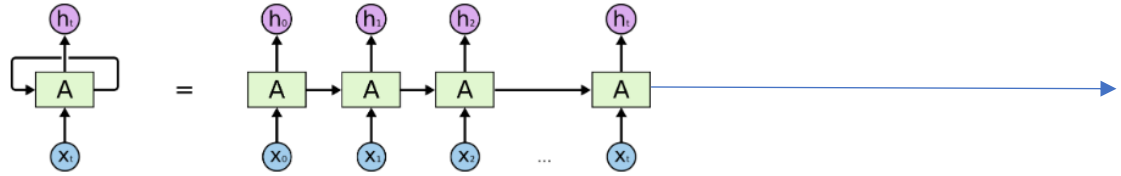




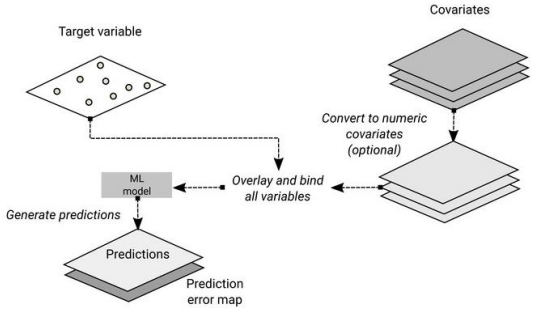
Recurrent Neural Network + Data fusion

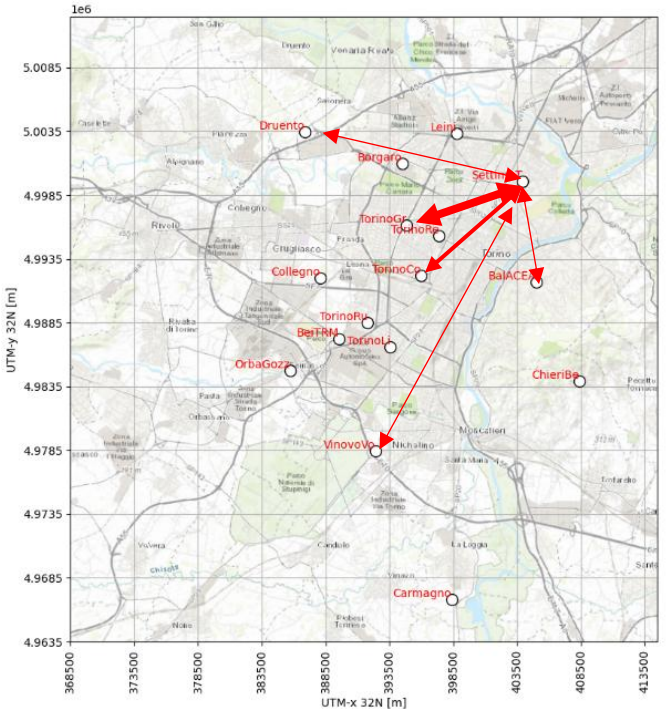


(b)



Supervised learning (Random Forest)





Predictors accounting for geographical proximity (and generic distance-based relations) between observations (to mimic spatial correlation used in Kriging methods)

Sekulić, A.; Kilibarda, M.; Heuvelink, G.B.M.; Nikolić, M.; Bajat, B. Random Forest Spatial Interpolation. *Remote Sens.* **2020**, *12*, 1687. https://doi.org/10.3390/rs12101687