Environ Health Perspect

DOI: 10.1289/EHP9752

Supplemental Material

Deep Ensemble Machine Learning Framework for the Estimation of $PM_{2:5}$ Concentrations

Wenhua Yu, Shanshan Li, Tingting Ye, Rongbin Xu, Jiangning Song, and Yuming Guo

Table of Contents

Table S1. The model hyper-parameters and computing cost information for study models

Model	Main Hyper-parameters	R package	Time cost for 10,000 cases
GLM	alpha = 1, nlambda = 100	glmnet	44.1 secs
GBM	n.trees = 100, interaction.depth = 1,n.minobsinnode = 10, shrinkage = 0.1, bag.fraction = 0.5,	gbm	14.15 mins
SVM	gamma = 0.1, tolerance=0.001, epsilon =0.1	e1071	2.97 mins
RF	mtry=4, num.trees = 500	ranger	30.4 secs
XGBoost	ntrees = 1000, max_depth = 4, shrinkage = 0.1, minobspernode = 10	XGBoost	5.00 mins
SL	The same as GBM,SVM, RF, and XGBoost	SuperLearner	6.35 mins
DEML	The same as GLM, GBM,SVM, RF, and XGBoost	deeper	11.98 mins

Note: SVM: support vector machine; RF: Random Forest; XGBoost: extreme gradient boosting; GBM: gradient boosting Machine; Date Fusion GLM: the generalized linear regression model with the data fusion method; SL: Super Learner algorithm; DEML: the three-stage stacked deep ensemble framework method. The CPU time cost was calculated using Intel(R) Core(TM) i7-1165G7 @ 2.80GHz with 4 physical cores paralleling computing.

Table S2. The variables information and sources

Variable type	Variable name	unit	Resolution	Variable Resources
Response variable	$PM_{2.5}$	ug/m^3	-	the Italian National Institute for Environmental Protection and Research (ISPRA) (https://www.isprambiente.gov.it/en/databases)
independent variable	PM_{10}	ug/m^3	-	the Italian National Institute for Environmental Protection and Research (ISPRA) (https://www.isprambiente.gov.it/en/databases)
Position	longitude		-	the Italian National Institute for Environmental Protection and Research (ISPRA) (https://www.isprambiente.gov.it/en/databases)
Position	latitude		-	the Italian National Institute for Environmental Protection and Research (ISPRA) (https://www.isprambiente.gov.it/en/databases)
Climate	dewpoint Temperature	K	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	daily average Temperature	K	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	daily maximum Temperature	K	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	daily minimum Temperature	K	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	Relative humidity	%	hourly data from 2015-2019,Italy, 0.1 x 0.1	Calculated through ‘humidity’ R package
Climate	Total Evaporation	mm	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	U-wind(east-west component of the wind) at 10m	m/s	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	V-wind(north-south component	m/s	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY

	of the wind) at 10m			
Climate	Surface Pressure	Pa	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	solar radiation	J	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
Climate	Total Precipitation	m	hourly data from 2015-2019,Italy, 0.1 x 0.1	GEE ERA5-land- ECMWF climate reanalysis https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY
land use	landcover classification		yearly from 2015-2019, Italy, 100 meters	GEE Copernicus Global Land Cover Layers: CGLS-LC100 collection 3 https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_Landcover_100m_Proba-V-C3_Global
elevation	elevation	m	90 meters	GEE SRTM Digital Elevation Data Version 4 the Shuttle Radar Topography Mission (SRTM) project https://srtm.csi.cgiar.org
populaiton density	population density		yearly from2015-2019, Italy, 100 meters	GEE WorldPop Global Project Population Data: Estimated Residential Population per 100x100m Grid Square https://www.worldpop.org/
satelliate	AOD 470nm		daily data from 2015-2019, Italy, 1km	GEE MCD19A2.006: Terra & Aqua MAIAC Land Aerosol Optical Depth Daily 1km https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MCD19A2_GRANULES#bands

Table S3. The seasonal distribution of daily average $PM_{2.5}$ ($\mu g/m^3$) from 2015 to 2019 in Italy

Season	Mean	SD	Median	IQR
Spring	16.83	10.66	14.00	12.49
Summer	14.16	6.67	13.05	8.10
Autumn	19.27	13.49	15.90	14.86
Winter	32.10	18.86	29.22	26.22

Note: $PM_{2.5}$: particles with a diameter of < 2.5 μm; SD: Standard deviation; IQR: The Interquartile Range by using the 75^{th} minus 25^{th} percentile of $PM_{2.5}$ concentrations in the study period. Spring: from March to May; Summer: from June to August; Autumn: from September to November; Winter: from December to February.

Table S4. The distribution comparison between observed $PM_{2.5}$ concentration and estimated $PM_{2.5}$ by RF, XGBoost, SL, and DEML in Italy from 2015 to 2019

		Mean	SD	Min	Q1	Q2	Q3	Q4	Cor
Observed PM2.5		20.33	14.70	1.20	10.08	15.90	25.89	98.11	NA
Estimated PM2.5	RF	20.43	12.24	1.60	11.89	16.53	25.39	86.02	0.89
	XGBoost	20.38	11.21	0.27	12.44	17.31	25.53	79.83	0.81
	SL	20.44	11.80	1.73	12.14	16.77	25.43	84.35	0.89
	DEML	20.34	13.56	0.65	11.11	16.06	25.33	92.10	0.91

Note: $PM_{2.5}$: particles with a diameter of < 2.5 μm; SD: Standard deviation; Q1-Q4: The 25^{th} , 50^{th} , 75^{th} , and 100^{th} percentile of $PM_{2.5}$ concentrations in the study period. RF: Random Forest; XGBoost: extreme gradient boosting; SL: Super Learner method; DEML: the three-stage stacked deep ensemble framework method; Cor: the Spearman's rank correlation coefficient.

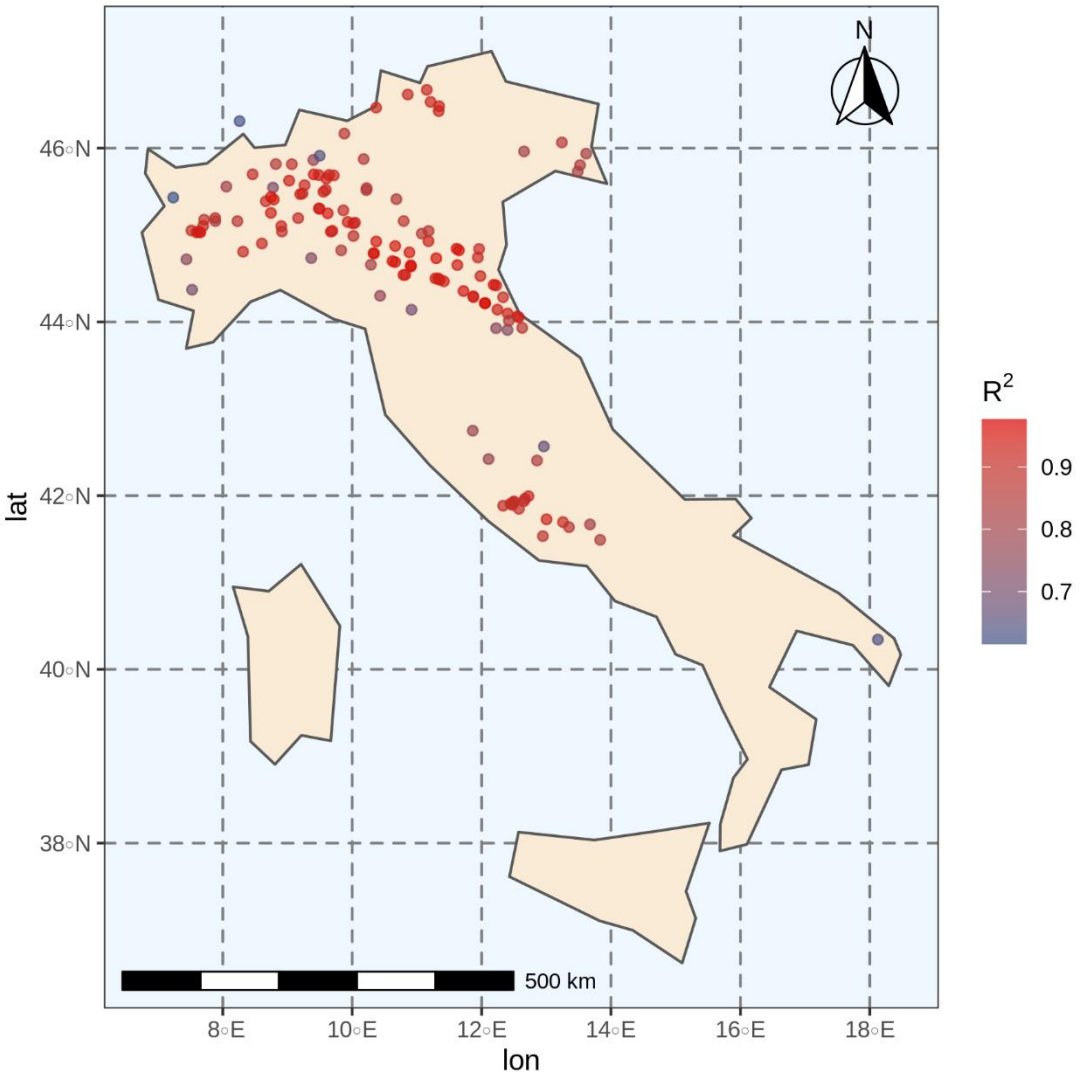

Figure S1. R^2 for each monitoring station based on predicted and observed $PM_{2.5}$ from 2015 to 2019 in Italy

Note: R^2 is the R-squared for 10-fold cross-validation. The points denote the position of monitor stations.

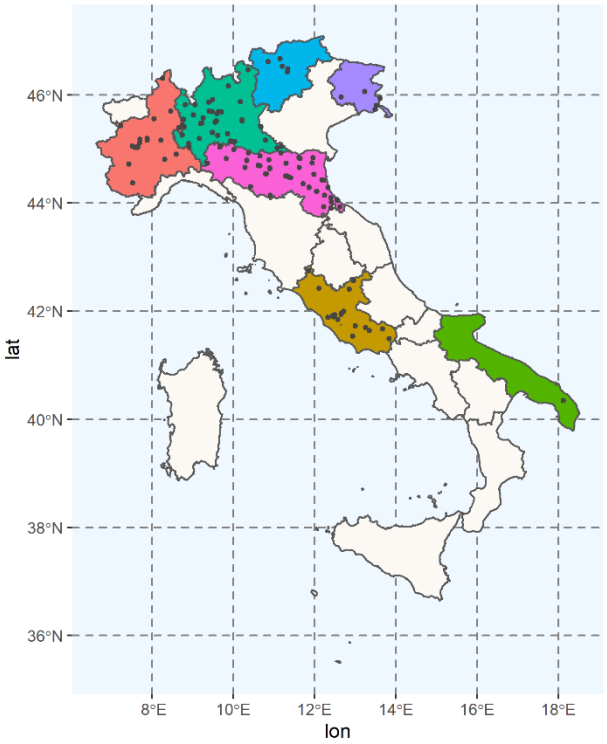

Figure S2. Italian regional distribution in Cluster-based spatial Cross-validation

Note: 7 out of 20 Italian regions were involved in the cluster-based spatial Cross-validation analysis. Dark points denote the position of monitor stations.

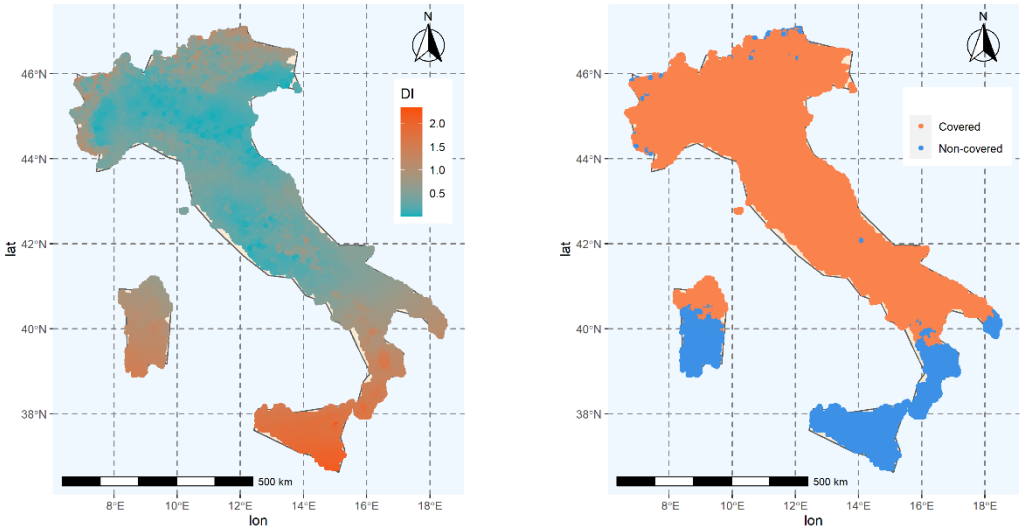

Figure S3. The dissimilarity index (DI) of the spatial locations in Italy with the ground stations in a specific day

Note: The dissimilarity index (left) is the normalized and weighted minimum distance to the nearest training data point divided by the average distance within the training data; The dissimilarity and uncertain area (right): the covered area is the region where the DI value is lower than 1, which means the spatial difference to the nearest station point is smaller than the

average dissimilarity of all station data. The variable importance was calculated by the RF as the variables' weights; We randomly selected a specific day on 2015-07-15. RF: Random Forest.

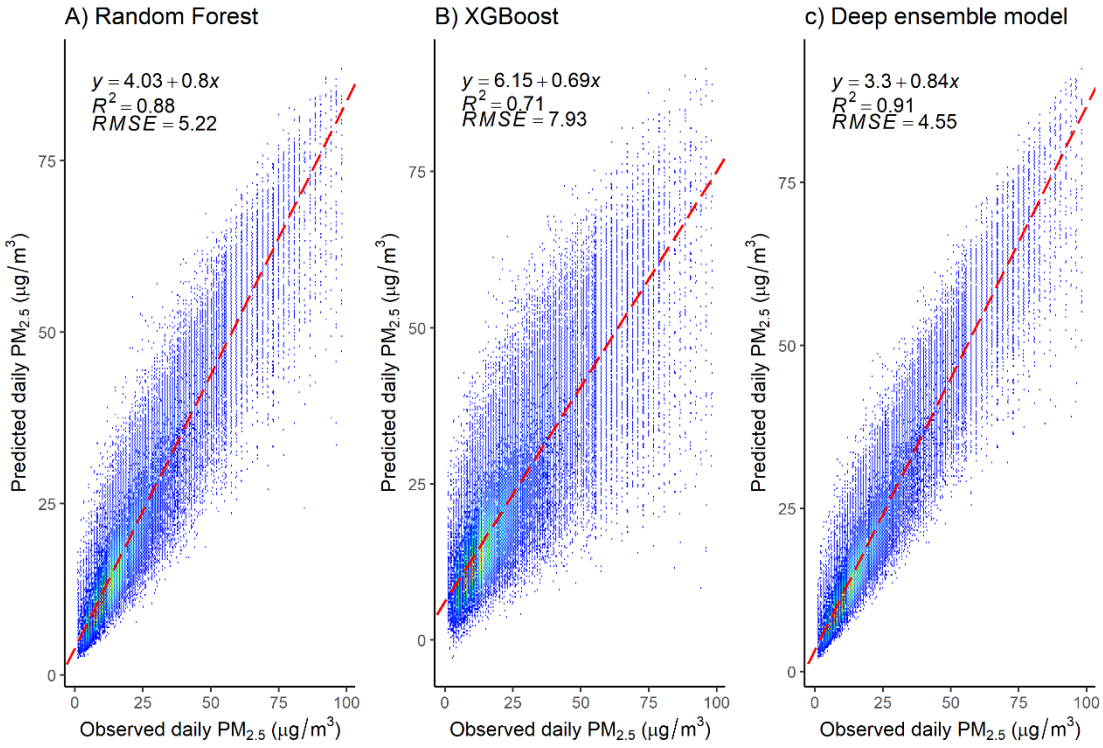

Figure S4. $PM_{2.5}$ imputation performance of the RF, XGBoost, and DEML by PM_{10} from 2015 to 2019 in Italy

Note: $PM_{2.5}$: particles with a diameter of < 2.5 µm; PM_{10}: particles with a diameter of < 10 µm; R^2 is the R-squared for 10-fold cross-validation; RMSE: the root-mean-square error (µg/m^3); RF: Random Forest; XGBoost: extreme gradient boosting. DEML: the three-stage stacked deep ensemble framework method.

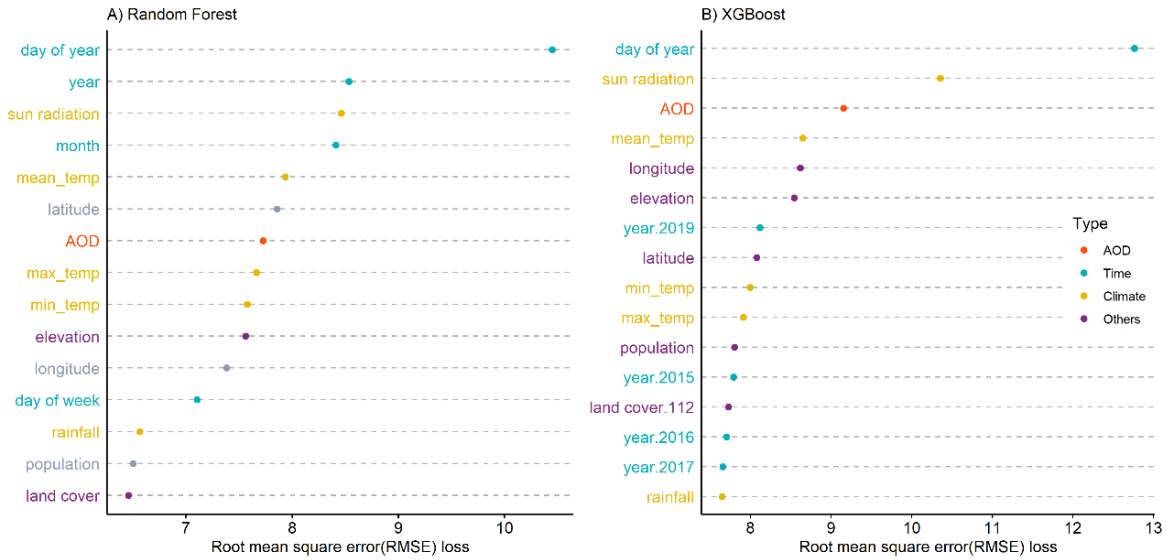

Figure S5. The evaluation of the importance of the explanatory variables in $PM_{2.5}$ estimation by RF and XGboost from 2015 to 2019 in Italy

Note: The loss of root mean square error was calculated to measure how much a model's performance change if the effect of a selected explanatory variable is removed. 10 permutations were selected to repeat the process 10 times to compute the mean values of RMSE loss. The RF and XGBoost models were used to calculate the variable importance. RF: Random Forest; XGBoost: extreme gradient boosting. The legend indicate the type of variables: Time included day of year, month, year, and day of week; climate included mean/max/min daily temperature, sun radiation, and rainfall; Others included longitude, latitude, elevation, land cover, and population density. $PM_{2.5}$: particles with a diameter of < 2.5 μm.