BC proxy with RF, SVR and ANN(MLP)

Bertille Temple

July 14, 2023
SANS research group, UPC

Processing details

The data from winter 2018 and winter 2019 are merged and shuffled, without distinguishing the year. The same is done for each season. This means that, whatever the data set or sub data set under study, the data is shuffled, split into training (75%) and testing (25%) and cross-validation is applied to the training set with k=10.

Regarding the training and the validation, the RMSE is computed by making the mean of the RMSE for each fold of the cross validation.

RH missing values represents 26% of the whole dataset.

The column labeled "scale 90% of BC values" on the tables indicates the range within which 90% of the Black Carbon values are located.

Contents

1 Relative Humidity column dropped

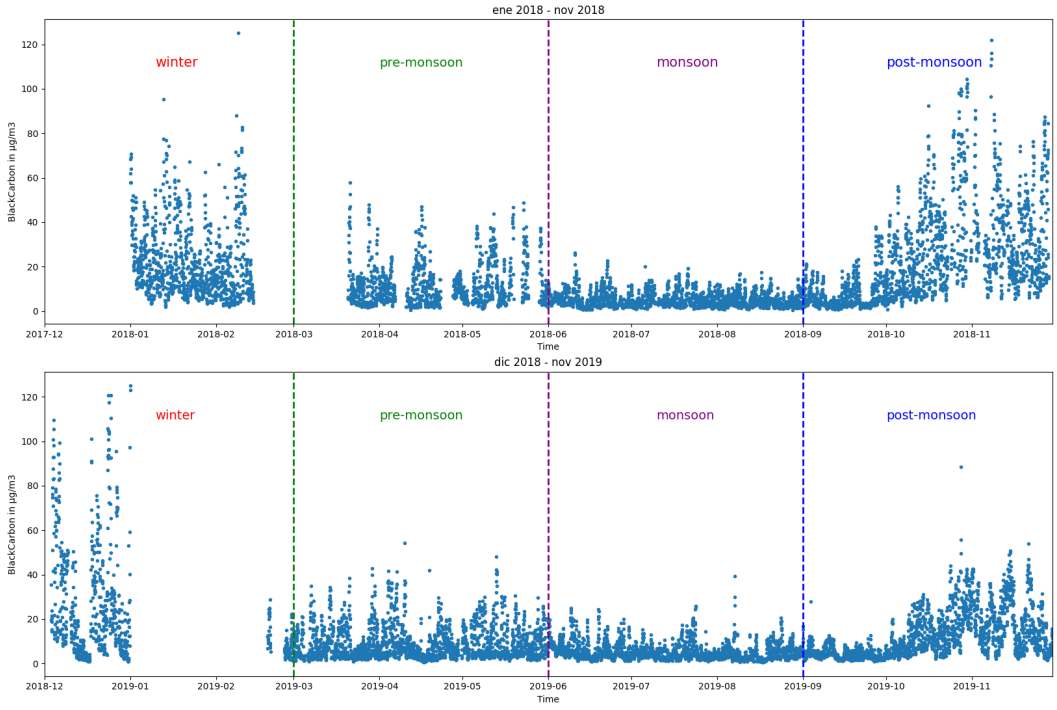


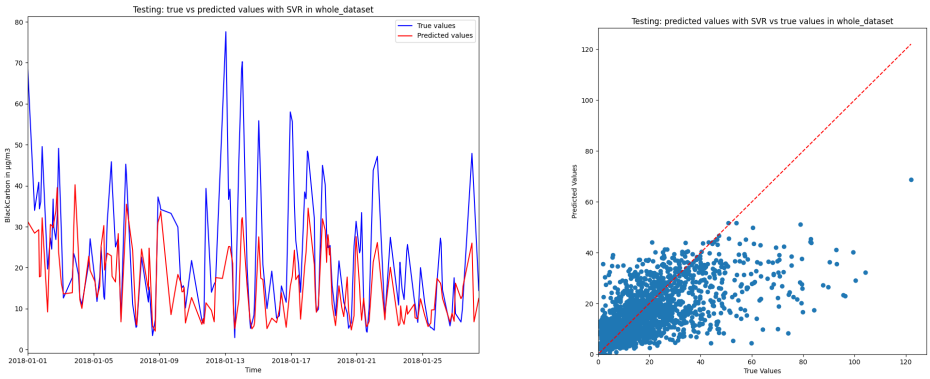
Figure 1: Seasons spliting.

1.1 Support Vector Regression (SVR)

The first approach consists of **dropping** the Relative Humidity column to avoid losing 26% of the whole dataset.
Features used: $PM_{2.5}$, PM_{10}, NO_x, O_3
Parameters are printed as follow: [C, Gamma, Epsilon]

season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]
scale 90 % of BC values	[0, 41]	[1, 60]	[0, 28]	[1, 13]	[1, 43]
BC mean	13	22	10	5	13
RMSE train	9.85	14,81	5,93	3.01	7.31
RMSE validation	10	14.9	6,06	3.09	7.79
RMSE test	9.8	15.13	6,35	3.16	7.7
MAE train	5.42	8.92	3.71	2.04	4.34
MAE validation	5.53	9.19	3.86	2.13	4.68
MAE test	5.37	8.9	3.95	2.04	4.56
R^2 train	0.53	0.37	0,5	0.35	0.77
R^2 test	0.52	0.37	0,49	0.37	0.76

Table 1: SVR metrics **without RH**



(a) (b)

Figure 2: BC concentration from the testing predicted with SVR on all data

Similar plots were made for each season. To find them please go to the images folder approach_1_SVR.

The results are poor. When Relative Humidity is excluded, we avoid throwing away 25% of the original dataframe but we loose a key feature.
Let us try a second approach with keeping RH as a feature.

2 Relative Humidity as a feature, missing values dropped

The second approach consists of **keeping** the Relative Humidity column to avoid losing a key feature to predict black carbon, and to drop all the rows with missing values in RH. Features used: $PM_{2.5}$, PM_{10}, NO_x, O_3, RH

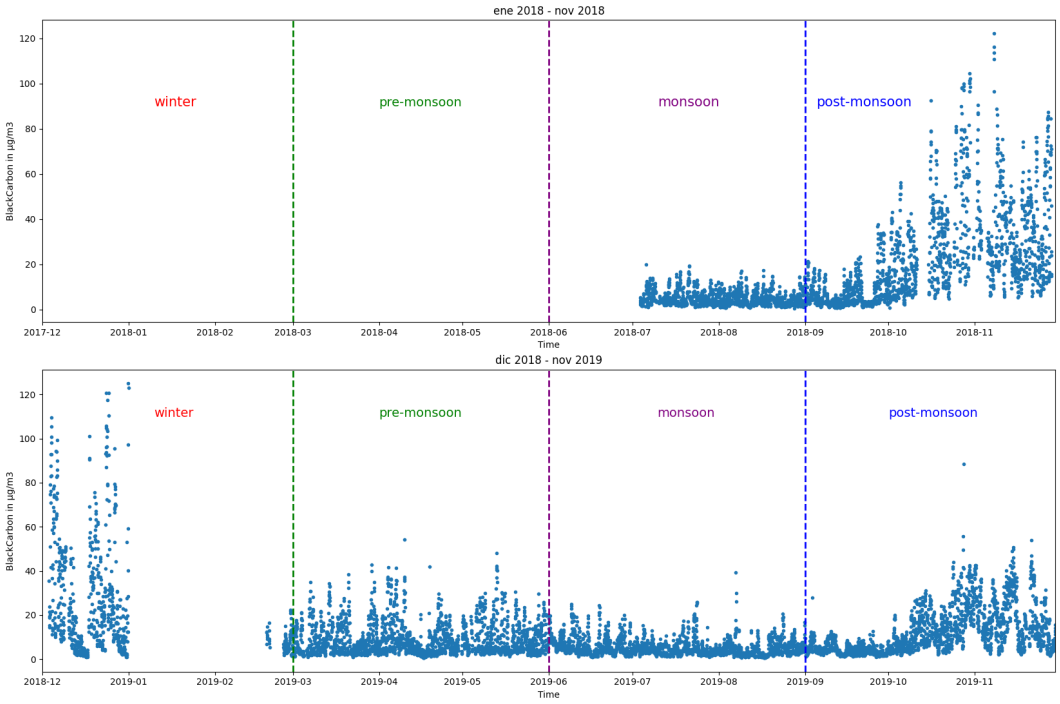


Figure 3: Seasons spliting

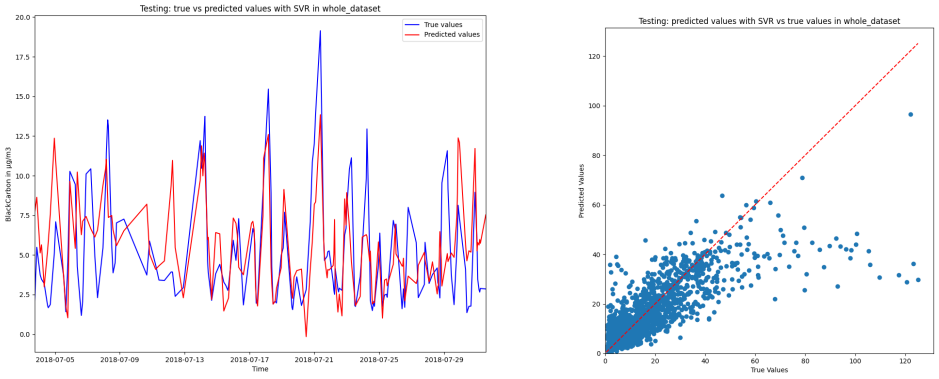
Removing nan values from RH column discards the data from the first 6 months of 2018.

2.1 Support Vector Regression (SVR)

Parameters are printed as follow: [C, Gamma, Epsilon]

season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]	
scale 90 % of BC values	[0, 41]	[1, 66]	[0, 25]	[1, 13]	[1, 41]
BC mean	12	22	9	5	13
RMSE train	7.76	14.08	4.07	2.5	6.05
RMSE validation	7.99	15.21	4.36	2.62	6.72
RMSE test	8.3	11.81	3.58	2.26	5.48
MAE train	4.09	7.8	2.5	1.66	3.42
MAE validation	4.26	8.34	2.78	1.77	3.84
MAE test	4.12	6.48	2.23	1.52	3.06
R^2 train	0.72	0.5	0.72	0.57	0.84
R^2 test	0.67	0.58	0.78	0.63	0.88

Table 2: SVR metrics **with RH. Missing values of RH dropped**



(a) (b)

Figure 4: BC concentration from the testing predicted with SVR on all data

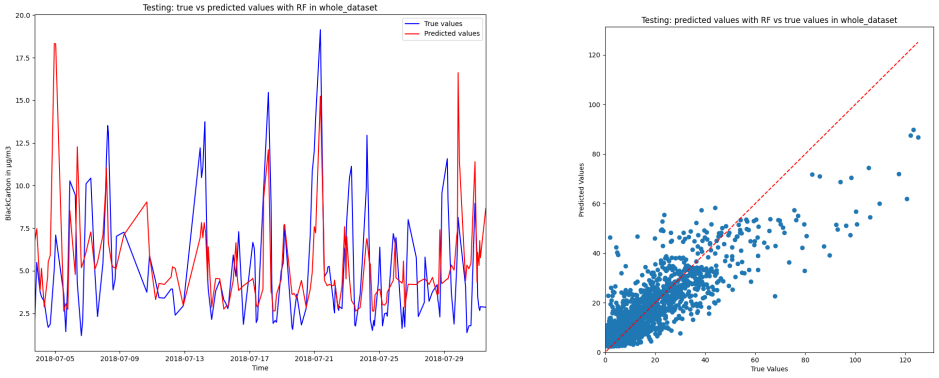
Similar plots were made for each season. To find them please go to the images folder approach_2_SVR.

2.2 Random Forest (RF)

Parameters are printed as follow: [n_estimators, max_features, max_depth]

season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[500, 20, 5]	[100, 20, 5]	[100, 5, 5]	[100, 20, 5]	[100, 20, 5]
scale 90 % of BC values	[0, 41]	[1, 66]	[0, 25]	[1, 13]	[1, 41]
BC mean	12	22	9	5	13
RMSE train	7.69	11.9	3.96	2.47	5.39
RMSE validation	8.32	14.51	4.53	2.72	6.89
RMSE test	7.22	8.29	3.31	2.17	4.72
MAE train	4.42	7.09	2.63	1.77	3.44
MAE validation	4.68	8.64	2.99	1.92	4.08
MAE test	4.25	5.33	2.3	1.53	3.05
R^2 train	0.72	0.67	0.73	0.58	0.87
R^2 test	0.75	0.79	0.81	0.66	0.91

Table 3: RF metrics **with RH. Missing values of RH dropped**



(a) (b)

Figure 5: BC concentration from the testing predicted with RF on all data

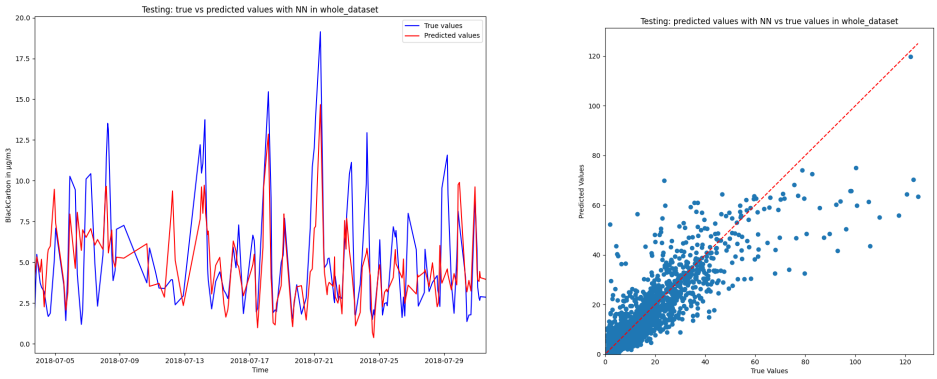
Similar plots were made for each season. To find them please go to the images folder approach_2_RF.

2.3 Multi-Layer Perceptron (MLP)

Parameters are printed as follow: [nb_neurons, alpha]. We use Adam optimizer and ReLu as activation function.

season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[(100, 100, 50), 0.001]	[(100, 100, 50), 0.01]	[(100, 100, 50), 0.001]	[(100, 100, 50), 0.01]	[(100, 100, 50), 0.01]
scale 90 % of BC values	[0, 41]	[1, 66]	[0, 25]	[1, 13]	[1, 41]
BC mean	12	22	9	5	13
RMSE train	7.37	14.51	4.39	2.38	6.15
RMSE validation	7.81	14.76	4.53	2.54	6.59
RMSE test	7.01	11.22	3.26	1.67	4.98
MAE train	4.09	8.7	2.88	1.73	3.68
MAE validation	4.29	8.94	2.97	1.16	3.89
MAE test	3.99	6.81	2.25	0.76	3.03
R^2 train	0.81	0.6	0.79	0.75	0.89
R^2 test	0.77	0.62	0.82	0.8	0.9

Table 4: MLP metrics **with RH. Missing values of RH dropped**



(a) (b)

Figure 6: BC concentration from the testing predicted with MLP on all data

Similar plots were made for each season. To find them please go to the images folder approach_2_NN.

The results are better than in the first approach, specifically with MLP. SVR and RF have results in summer (when RH is high) quite lower than in others seasons, MLP has good results in summer.

3 Relative Humidity as a feature, missing values imputed

The third approach consists of **keeping** the Relative Humidity column to avoid losing a key feature to predict black carbon, and to impute the missing values with RH data from a neigborhd city (where ??) to avoid losing 26%of the whole dataset.
Features used: $PM_{2.5}$, PM_{10}, NO_x, O_3, RH (imputed)

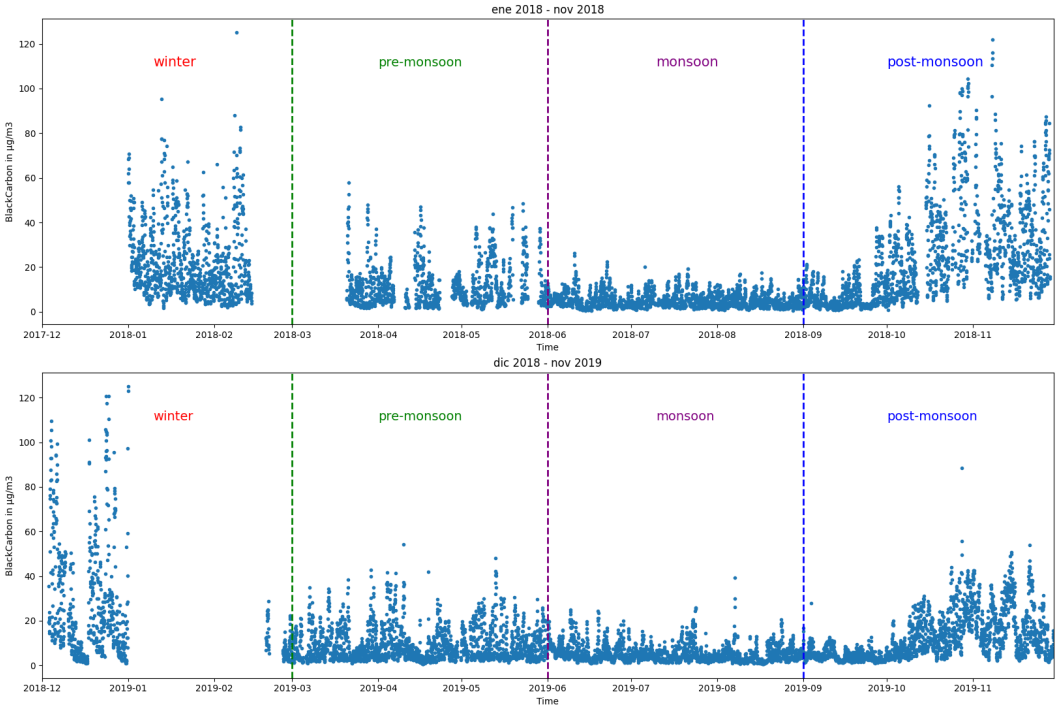


Figure 7: Seasons spliting

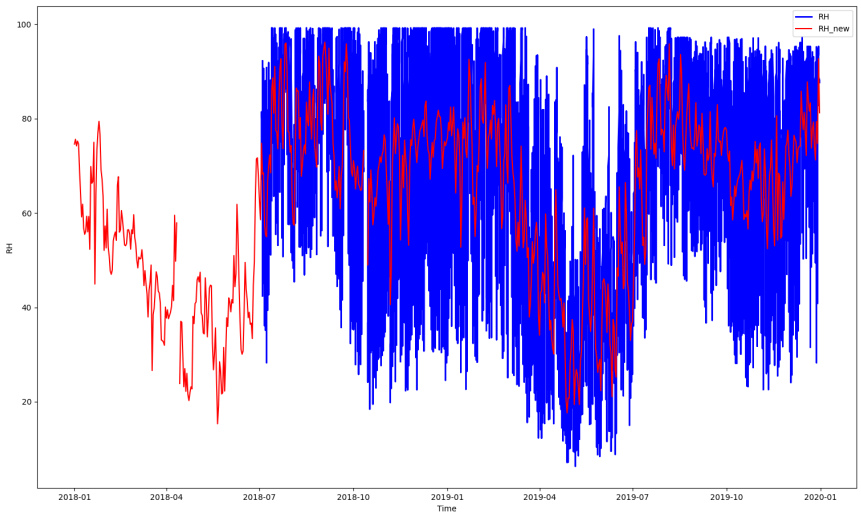


Figure 8: RH values, inplace values in blue are by hour. new RH values in red are by day

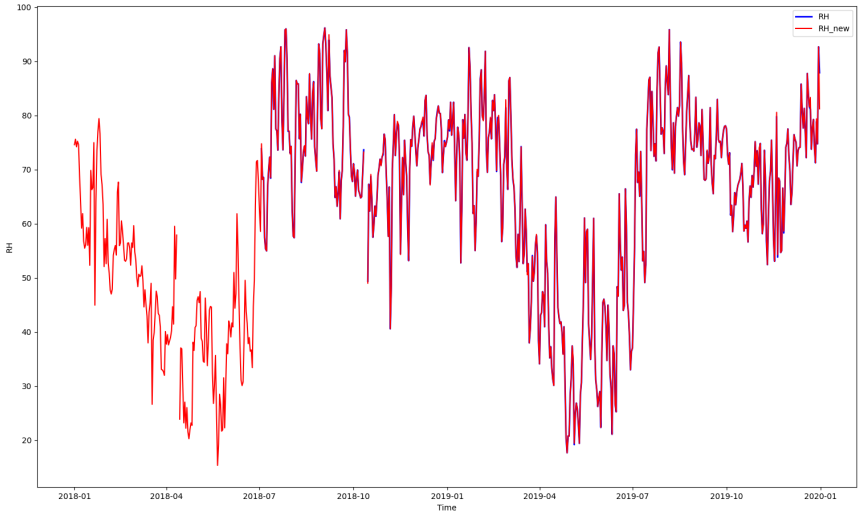


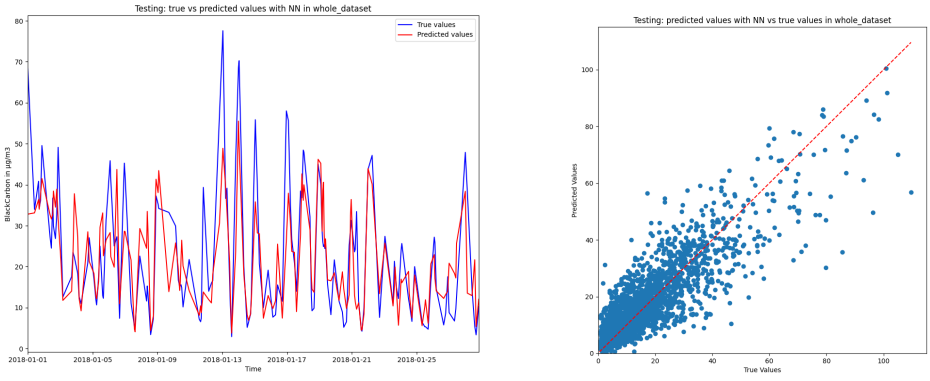
Figure 9: RH values, the mean is done by day for the inplace values

The plot above shows that the new values of RH are equal or close to the RH values already inplace when the mean is made by day. RH measures in Delhi are taken every hour, whereas the new RH measures are taken every day. We will input the inplace values missing (from January 2018 to July 2018) by the new ones. Post-monsoon results should not be affected by the imputation as the imputed data concern winter, pre-monsoon, and summer 2018.

3.1 Multi-Layer Perceptron (MLP)

season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[(100, 100, 50), 0.001]	[(100, 100, 50), 0.001]	[(100, 100, 50), 0.0001]	[(100, 100, 50), 0.01]	[(100, 100, 50), 0.001]
scale 90 % of BC values	[0, 41]	[1, 60]	[0, 28]	[1, 13]	[1, 42]
BC mean	13	22	10	5	13
RMSE train	8.25	13.41	5.55	2.44	6.11
RMSE validation	8.69	13.85	5.82	2.62	6.64
RMSE test	6.95	13.37	4.9	2.32	5.59
MAE train	4.75	8.75	3.72	1.7	3.69
MAE validation	4.94	9.04	3.91	1.83	3.91
MAE test	4.15	8.54	3.47	1.45	3.36
R^2 train	0.75	0.58	0.68	0.69	0.89
R^2 test	0.74	0.51	0.69	0.66	0.88

Table 5: metrics



(a) (b)

Figure 10: BC concentration from the testing predicted with MLP on all data

Similar plots were made for each season. To find them please go to the images folder approach_3_RF.

When there are missing values in RH column, Relative Humidity is imputed. We avoid throwing away 25% of the original df and keep a key feature. The results are actually worse or similar than without imputation(second approach). Also, we tried with SVR and RF and it gave the same conclusion.

4 Relative Humidity and Solar radiation as features, missing values dropped

The fourth approach consists of **keeping** the Relative Humidity, to concatenate the Solar Radiation (SR) feature and to drop the missing values(RH **is not** imputed)
Features used: $PM_{2.5}$, PM_{10}, NO_x, O_3, RH, SR.

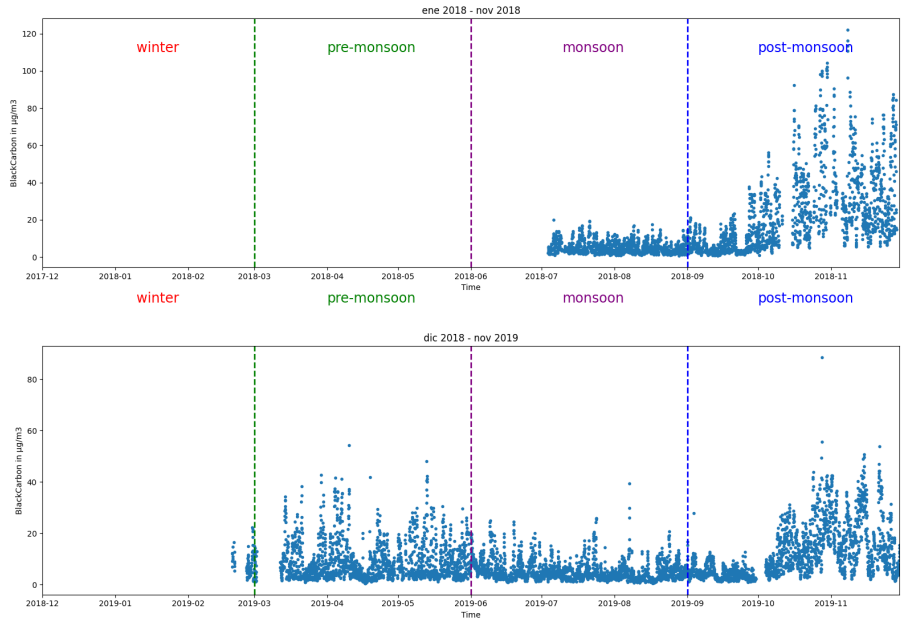


Figure 11: Seasons spliting

4.1 SVR

season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]	[100, 0.1, 1]
scale 90 % of BC values	[0,37]	[1, 39]	[0,26]	[1, 13]	[1, 42]
BC mean	11	17	9	5	13
RMSE train	5.36	3.51	3.52	2.23	5.52
RMSE validation	5.71	4.11	3.98	2.37	6.08
RMSE test	5.75	3.01	3.48	2.01	6.26
R^2 train	0.83	0.9	0.79	0.66	0.87
R^2 test	0.8	0.93	0.81	0.71	0.83

Table 6: SVR metrics

4.2 RF

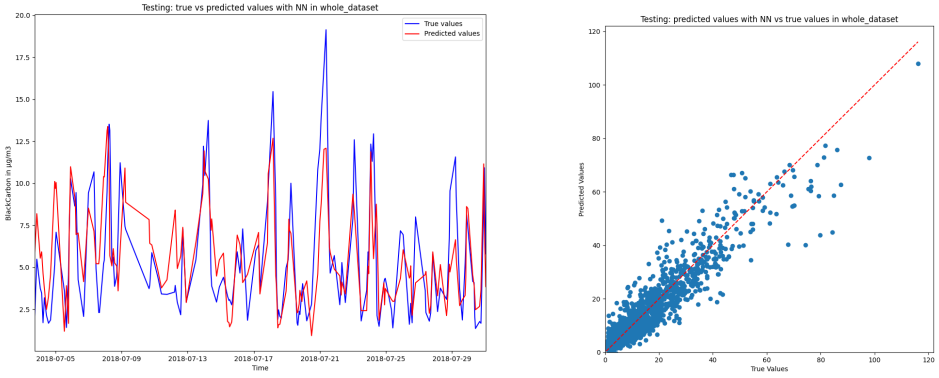
season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[100, 5, 5]	[500, 5, 5]	[100, 5, 5]	[500, 10, 5]	[500, 5, 5]
scale 90 % of BC values	[0,37]	[1, 39]	[0,26]	[1, 13]	[1, 42]
BC mean	11	17	9	5	13
RMSE train	5.42	3.16	3.59	2.35	4.86
RMSE validation	5.91	4.51	4.25	2.63	6.11
RMSE test	5.26	2.49	3.08	2.09	5.12
R^2 train	0.83	0.92	0.78	0.62	0.9
R^2 test	0.83	0.95	0.85	0.68	0.89

Table 7: RF metrics

4.3 MLP

season	all	winter	pre-monsoon	summer	post-monsoon
parameters	[(100, 100, 50), 0.001]	[(100, 100, 50), 0.0001]	[(100, 100, 50), 0.0001]	[(100, 100, 50), 0.001]	[(100, 100, 50), 0.0001]
scale 90 % of BC values	[0,37]	[1, 39]	[0,26]	[1, 13]	[1, 42]
BC mean	11	17	9	5	13
RMSE train	5.04	4.91	4.06	2.25	5.88
RMSE validation	5.39	5.08	4.34	2.42	6.29
RMSE test	4.65	4.57	3.17	1.45	6.06
R^2 train	0.9	0.9	0.7	0.82	0.91
R^2 test	0.87	0.84	0.84	0.85	0.84

Table 8: MLP metrics



(a) (b)

Figure 12: BC concentration from the testing predicted with MLP on all data

Similar plots were made for each season. To find them please go to the images folder approach_4_RF. Adding the feature Solar Radiation results in better results, specifically in the whole dataset. The results are much better in winter than before, yet it should be pointed out that solar radiation has a significant amount of missing values in winter. At the end, the winter dataset contains only 756 rows.