Black Carbon proxy in Delhi taking into account the seasonality

teacher: Jose M. Barcelo Ordinas
master students: Bertille Temple and Iva Bokšić
with the help of the phd student Juan Antonio Paredes Ahumada

June, 2023
SANS group, FIB, UPC

Introduction

The phd student J. Antonio Paredes Ahumada from the research group Statistical Analysis of Networks and Systems predicted Black Carbon concentration in Barcelona in the year 2018 using the following features: NO2, N, PM 2.5, O3, T, PM1, RH, PM10. The results obtained were pretty good with a R2 of 0.83 and a RMSE of 0.48. When trying to apply the same proxy to predict Black Carbon with data from the Indian cities Delhi and Agra, the results turn out to be very poor, with a R2 below 0.5. To explain such poor results, 2 hypothesis were made: the first hypothesis is that the poor predictions obtained are the results of ignoring the seasonality in the model. The second hypothesis is that in the Indian data, N-CPC is not measured(contrarily to Barcelona data) whereas it might be one of the key feature to predict Black Carbon. In this report, we will study the first hypothesis and predict Black Carbon in Dehli for the year 2018 and 2019 by applying Random Forest, Support Vector Regression and Neural Network, with taking care of including the seasonality in the analysis. Agra dataset could be studied in a further work.

Contents

1 Pre-processing

1.1 Missing values

Feature	Date	Hrs.	BC	PM10	PM2.5	RH
missing values(%)	0	0	13	5	5	26

Feature	WD	WS	Temp	RF	NOx	Ozone
missing values(%)	99	99	100	100	3	5

Table 1: Percentage of missing values in Delhi dataset for 2018 and 2019

First, we remove the columns: 'WD', 'WS', 'Temp' and 'RF' as they do not contain measures(99% or 100% of the measures are missing).
Then, **we remove the rows with missing values which is 37% of the original dataframe in total**, it is significant. The table above points out that the missing values represents 13% of the dataset for Black Carbon, and 26% for Relative Humidity(!). In this report, we will limit ourselves to remove all the rows with missing values, but it could be tested in a further work to impute the Relative Humidity values.

1.2 Seasons splitting

The seasons are split as follow:

- Winter: December, January, and February (DJF): 1458 rows
- Pre-monsoon: March, April, and May (MAM): 2085 rows
- Summer: June, July, and August (JJA): 3518 rows
- Post-monsoon: September, October, and November: 2666 rows

The data from winter 2018 and winter 2019 is merged together and shuffled, without distinguishing the year. The same is done for every season.

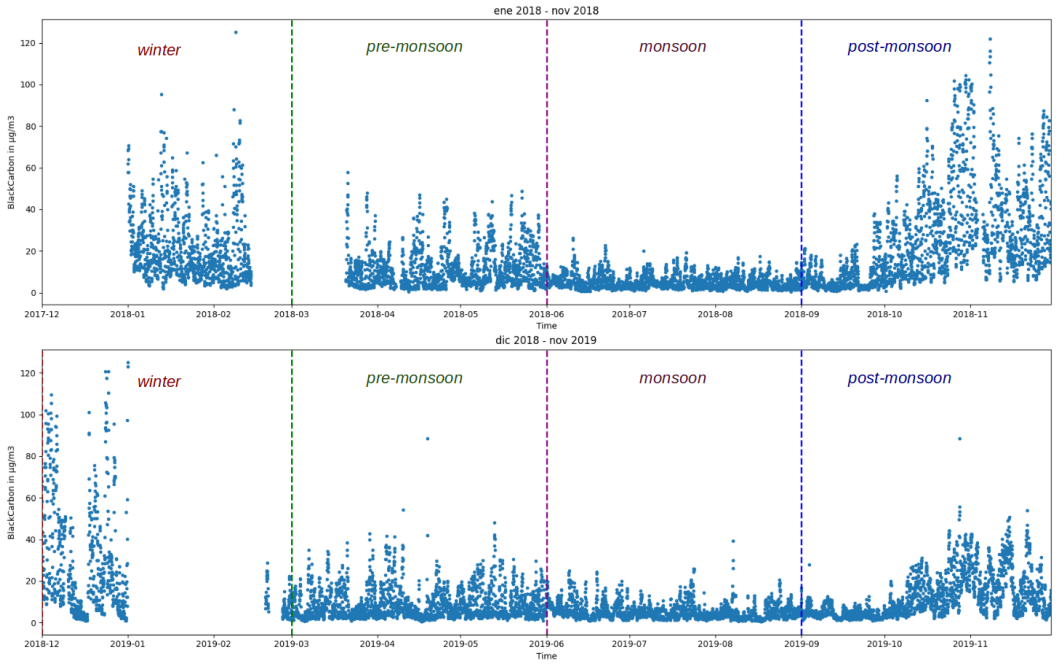

Figure 1: Seasons spliting

Note that, after removing the rows with missing values, the first 6 months from 2018 were discarded.

2 Random forest

First let us predict Black Carbon concentration with Random Forest in the whole dataset. We expect to obtain similar results to Juan, which is R2 around 0.5. Then, we predict Black Carbon for 4 seasonal subsets, it means that we train 4 different models. Whatever the dataset under study, data is scaled, shuffled, split into training(75%) and testing (25%), and cross validation is performed on the training with k = 10.

2.1 Whole dataset

Features used: PM2.5, PM10, RH, NOx and Ozone

n range	[50, 100, 300, 500, 1000]
features range	[1, 3, 5, 10, 15]
max_depth range	[5, 10, 15, 20, 30]
best n	500
best features	10
best depth	5
RMSE train	0.53
R2 train	0.72
RMSE test	0.50
RMSE test in the original unit	7.24µg/m3
R2 test	0.75

Table 2: Metrics of Random Forest applied to the whole dataset with 5 features

On the one hand, $RMSE_{test}$ is high (≥ 0.5), which means that the average difference between the true and predicted values is quite high, around 7.24 in µg/m3. In the whole year, Black Carbon takes value between 0 and 100,120. So the model does not predict Black Carbon accurately when Random Forest is applied to the whole dataset. On the other hand, the model generalises well to unseen data and R2 is high.
The difference between Juan and our result can be explained by the fact that we have RH and PM10 as features, and Juan does not. We think that RH is a key feature to predict BC.

2.2 Seasonal subsets

Features used: PM2.5, PM10, RH, NOx and Ozone

season	winter	pre-monsoon	summer	post-monsoon
best parameters	[10, 1, 3]	[500, 30, 3]	[500, 15, 3]	[500, 20 3]
RMSE train	0.51	0.37	0.59	0.22
RMSE validation	0.54	0.42	0.63	0.26
RMSE test	0.50	0.51	0.74	0.42
unscaled RMSE test in µg/m3	15.15	4.05	2.80	6.51
R2 test	0.50	0.67	0.46	0.80

Table 3: Metrics of Random Forest applied to 4 seasonal datasets with 5 features

The model overfits significantly for pre-monsoon, summer and post-monsoon. We tried to remove PM10 but it did not reduce the overfitting. Also, removing PM10 resulted in an increase of all RMSE test, and in a drop of almost all R2. [R2 went to 0.56 for winter, 0.63 for pre-monsoon, 0.41 for summer, and 0.62 for post monsoon without PM10]

RF conclusion With Random Forest, we obtained quite good results without spliting the dataset into seasons, with a R2 of 0.75, but with a mitigated accuracy of 7.24µg/m3. This leads us to think it might not be necessary to take into account the temporality nor N-CPC to predict Black Carbon in Delhi. Applying the model to seasonal subset gave similar or poorer results in term of R2 than when it is applied to the whole dataset, and it gave a bit better results regarding the error. The error in winter is high, and R2 is 0.50. Figure 2 (Seasons splitting) shows that the data from winter is really bad, so we think none of the winter models will output good predictions.

3 Support Vector Regression

First we predict Black Carbon concentration with SVR in the whole dataset. Then, we predict Black Carbon for 4 seasonal subsets, which means that we train 4 different models. Whatever the dataset under study, data is scaled, shuffled, split into training(75%) and testing (25%), and cross validation is performed on the training with k = 10.

3.1 Whole dataset

Features used: PM2.5, PM10, RH, NOx and Ozone

kfold	10
cs	[1, 10, 100]
gammas	[0.001, 0.01, 0.1]
epsilons	[0.01, 0.1, 1]
best C	10
best gamma	0.1
best epsilon	0.1
RMSE train	0.53
R2 train	0.72
RMSE test	0.59
RMSE test in the original unit	8.60µg/m3
R2 test	0.65

Table 4: Metrics of SVR applied to the whole dataset with 5 features

The conclusions are the same as with Random Forest. The model generalises well but has poor performance to accurately predict Black Carbon. Moreover, R2 is 0.1 lower than with Random Forest. Yet, note that the model is trained with a smaller set of hyper parameters than in RF. The SVR model might give better result if trained with a gridsearch narrower.

3.2 Seasonal subsets

Features used: PM2.5, PM10, RH, NOx and Ozone

season	winter	pre-monsoon	summer	post-monsoon
best parameters	[100, 0.1, 0.01]	[10, 0.1, 0.1]	[100, 0.1, 0.01]	[10, 0.1, 0.01]
RMSE train	0.44	0.27	0.41	0.16
RMSE validation	0.48	0.31	0.47	0,19
RMSE test	0.64	0.45	0.65	0.38
unscaled RMSE test in µg/m3	13.02	3.50	2.48	5.68
R2 test	0.63	0.76	0.58	0.84

Table 5: Metrics of SVR applied to 4 seasonal subsets with 5 features

The predictions seem better for pre-monsoon and post-monsoon. But the models overfit significantly(despite cross validation and feature selection).

SVR conclusion Spliting into seasons improve result in terms of R2, specifically for pre-monsoon and post-monsoon. But there is a lot of overfitting. As with RF, winter results have a high error.

4 Neural network LSTM

4.1 Whole dataset

5 features used: PM2.5, PM10, RH, NOx and Ozone

nb of layer	1
optimizer	Adam
regularization	L2
learning rate	0.001
epochs	50
unscaled RMSE test	8.28 µg/m3
scaled RMSE train	0.53
scaled RMSE test	0.54
R2 train	0.68
R2 test	0.66

Table 6: Metrics of LSTM applied to the whole dataset with 5 features

4.2 Season, date and time as features

10 features used: PM2.5, PM10, RH, NOx, Ozone, Day, Month, Year, Hour, Season_Encoded. Season_Encoded is a categorical variable which can take 4 values, one for each season.

nb of layer	1
optimizer	Adam
regularization	L2
learning rate	0.001
epochs	50
unscaled RMSE train	6.39 µg/m3
unscaled RMSE test	6.89 µg/m3
scaled RMSE train	0.42
scaled RMSE test	0.47
R2 train	0.81
R2 test	0.76

Table 7: Metrics of LSTM applied to the whole dataset with 10 features

LSTM conclusion It seems that taking into account Day, Month, Year, Hour, Season_Encoded as features result in an an increase of the R2 of 0.1. With the dataset of 10 features, we looked for different combinations of parameters before selecting the best combination. For the dataset of 5 features, we only reused the best combinations previously found. We think that if we conduct a parameters research for the dataset with 5 features too, we could find better parameters that would give a R2 higher than 0.66.

Conclusion

We systematically applied cross validation for RF and SVR but it might have not been necessary when the model is run with the whole data as it contains approximatley 11 000 rows once the rows with missing values have been removed. When predicting Black Carbon Concentration in Delhi on seasonal subsets, SVR has better results than RF. Both predict better Black Carbon in pre-Monsoon and post-Monsoon, when Black Carbon concentration is quite high. In Summer, when Black Carbon is low, the prediction results are quite poor with a R2 of 0.53 for SVR and R2 of 0.46 for RF. It might be related to the fact that in Summer, during the Monsoon, Relative Humidity is really high and PM sensors are sensitive to humidity. As a result, the PM measures might be less reliable in this season and the quality of the PM data is poorer(recall that PM is a key feature to predict BC). In any case, the results should be taken with precaution as there is overfitting in the model applied to seasonal subsets, despite cross validation and feature selection.

Random Forest applied on the whole dataset is quite promising, when tuned properly and without season splitting. Yet, the big disadvantage of RF is the computational time it takes to tune the hyper parameters with the cross validation and the grid search. As we said, the big size of the dataset might allow to avoid doing cross validation and save us a precious amount of time in the RF hyper parameters tuning phase. SVR and LSTM give good results on the whole dataset too, even if they could be improved with a finer tuning of the hyper parameters. It seems that we can have good predictions of Black Carbon without measuring N-CPC. In a further work, it could be tested to train the models on the year 2018 and test them on the years 2019 (forecasting), and it could be tested to work with Agra data too.