Feature selection and extraction

Table of Contents

We have the total data saved in "**data10_Real_clean**" set. Firstly, we obtain a new feature : **MIS (wind-waves misalignment)**

1 New feature: Misalignment

```
data10_Real_clean = calculateMis(data10_Real_clean)
```

```
data10_Real_clean = 69282x16 table
```

...

	YY	MM	DD	hh	mm	WDIR	WSPD	GST
1	2010	2	16	23	50	319	7.0000	8.5000
2	2010	2	17	0	50	320	6.6000	7.8000
3	2010	2	17	1	50	327	6.5000	7.5000
4	2010	2	17	2	50	327	5.2000	6.5000
5	2010	2	17	3	50	320	4.1000	5.4000
6	2010	2	17	4	50	13	2.0000	2.5000
7	2010	2	17	5	50	32	1.8000	2.4000
8	2010	2	17	6	50	326	1.9000	2.8000
9	2010	2	17	7	50	2	2.1000	2.7000
10	2010	2	17	8	50	27	3.9000	4.6000
11	2010	2	17	9	50	1	1.1000	1.7000
12	2010	2	17	10	50	341	4.2000	5.1000
13	2010	2	17	11	50	14	2.4000	3.2000
14	2010	2	17	12	50	354	0.6000	1.2000
15	2010	2	17	13	50	352	4.0000	5.2000
16	2010	2	17	14	50	355	2.6000	3.5000
17	2010	2	17	15	50	18	1.2000	2.0000

	YY	MM	DD	hh	mm	WDIR	WSPD	GST
18	2010	2	17	16	50	59	0.4000	1.1000
19	2010	2	17	17	50	325	2.9000	3.7000
20	2010	2	17	18	50	315	3.3000	4.1000
21	2010	2	17	19	50	311	3.5000	4.2000
22	2010	2	17	20	50	310	4.1000	4.9000
23	2010	2	17	21	50	310	4.0000	5.3000
24	2010	2	17	22	50	299	4.2000	5.0000
25	2010	2	17	23	50	316	3.6000	5.1000
26	2010	2	18	0	50	311	3.3000	4.7000
27	2010	2	18	1	50	326	3.5000	4.6000
28	2010	2	18	2	50	339	1.5000	2.8000
29	2010	2	18	3	50	340	1.7000	3.7000
30	2010	2	18	4	50	314	2.5000	4.2000
31	2010	2	18	5	50	298	2.2000	3.6000
32	2010	2	18	6	50	305	3.4000	4.6000
33	2010	2	18	7	50	335	2.7000	4.3000
34	2010	2	18	8	50	358	2.0000	3.6000
35	2010	2	18	9	50	357	1.2000	2.8000
36	2010	2	18	10	50	2	2.8000	4.1000
37	2010	2	18	11	50	344	2.6000	4.2000
38	2010	2	18	12	50	349	2.4000	4.2000
39	2010	2	18	13	50	327	1.0000	2.5000
40	2010	2	18	14	50	320	1.9000	3.7000
41	2010	2	18	15	50	34	1.6000	2.6000
42	2010	2	18	16	50	41	1.2000	2.3000
43	2010	2	18	17	50	183	1.3000	2.2000
44	2010	2	18	18	50	149	1.5000	2.4000
45	2010	2	18	19	50	268	1.9000	2.7000
46	2010	2	18	20	50	262	2.6000	4.6000
47	2010	2	18	21	50	272	1.8000	2.7000
48	2010	2	18	22	50	246	2.0000	2.8000
49	2010	2	18	23	50	296	1.0000	1.8000
50	2010	2	19	0	50	327	0.3000	0.9000
51	2010	2	19	1	50	291	1.0000	2.2000

	YY	MM	DD	hh	mm	WDIR	WSPD	GST
52	2010	2	19	2	50	325	1.0000	1.5000
53	2010	2	19	3	50	300	1.7000	2.4000
54	2010	2	19	4	50	185	0.1000	0.6000
55	2010	2	19	5	50	316	1.3000	2.7000
56	2010	2	19	6	50	324	0.9000	2.4000
57	2010	2	19	7	50	13	0.3000	0.7000
58	2010	2	19	8	50	257	0.7000	1.7000
59	2010	2	19	9	50	252	0.5000	1.3000
60	2010	2	19	10	50	128	0.3000	0.8000
61	2010	2	19	11	50	180	0.1000	0.3000
62	2010	2	19	12	50	212	3.2000	3.9000
63	2010	2	19	13	50	234	3.7000	4.5000
64	2010	2	19	14	50	226	4.8000	5.8000
65	2010	2	19	15	50	216	5.0000	6.0000
66	2010	2	19	16	50	208	4.9000	6.0000
67	2010	2	19	17	50	172	4.0000	5.0000
68	2010	2	19	18	50	179	6.4000	7.8000
69	2010	2	19	19	50	177	7.7000	9.3000
70	2010	2	19	20	50	166	7.2000	8.5000
71	2010	2	19	21	50	163	8.0000	9.2000
72	2010	2	19	22	50	161	8.8000	10.0000
73	2010	2	19	23	50	161	8.3000	9.9000
74	2010	2	20	0	50	160	9.2000	11.1000
75	2010	2	20	1	50	144	9.8000	11.8000
76	2010	2	20	2	50	141	6.5000	7.9000
77	2010	2	20	3	50	147	6.2000	7.5000
78	2010	2	20	4	50	151	2.9000	4.4000
79	2010	2	20	5	50	128	0.5000	1.5000
80	2010	2	20	6	50	134	2.3000	4.4000
81	2010	2	20	7	50	136	5.1000	7.0000
82	2010	2	20	8	50	142	4.8000	6.3000
83	2010	2	20	9	50	121	3.6000	5.2000
84	2010	2	20	10	50	138	3.5000	4.3000
85	2010	2	20	11	50	110	2.2000	3.5000

	YY	MM	DD	hh	mm	WDIR	WSPD	GST
86	2010	2	20	12	50	32	2.6000	3.6000
87	2010	2	20	13	50	355	2.5000	3.3000
88	2010	2	20	14	50	344	2.5000	3.3000
89	2010	2	20	15	50	327	2.6000	3.5000
90	2010	2	20	16	50	335	3.6000	4.9000
91	2010	2	20	17	50	350	3.3000	4.3000
92	2010	2	20	18	50	323	3.2000	4.1000
93	2010	2	20	19	50	314	4.7000	5.7000
94	2010	2	20	20	50	316	5.1000	6.1000
95	2010	2	20	21	50	306	6.3000	7.3000
96	2010	2	20	22	50	312	7.0000	8.7000
97	2010	2	20	23	50	302	6.2000	7.7000
98	2010	2	21	0	50	307	8.1000	9.8000
99	2010	2	21	1	50	314	8.2000	10.2000
100	2010	2	21	2	50	310	7.5000	9.6000

⋮

2 Inital data sets

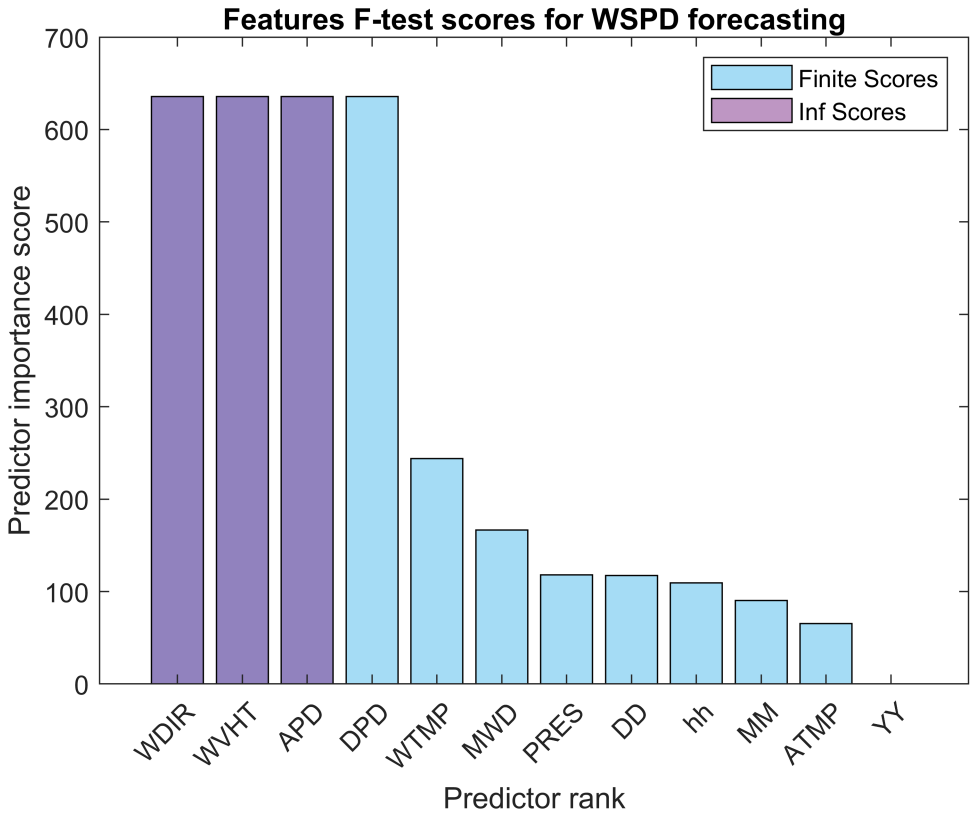
```
data2018 = data10_Real_clean(data10_Real_clean.YY == 2018,:);
data2019 = data10_Real_clean(data10_Real_clean.YY == 2019,:);
dataRealTime = data10_Real_clean(data10_Real_clean.YY == 2020,:);
data10_18 = data10_Real_clean(data10_Real_clean.YY ~= 2019 & data10_Real_clean.YY ~= 2020,:);
```

Now we are going to study the significance of features respect to the variable we want to predict, by applying the **F-score technique** to data from 2018.

3 F-test score - Features significance study

For Wind speed forecasting

```
[idx, scores,b1] = fSelection(data2018(:,{'YY','MM','DD','hh','WDIR','WSPD','WVHT','DPD','APD',
```




```
idx = 1×12
     5     6     8     7    12     9    10     3     4     2    11     1
scores = 1×12
         0   90.2098  117.2723  109.5398  635.5853  635.5853  635.5853  635.5853 ···
b1 =
  Bar (Finite Scores) with properties:

    BarLayout: 'grouped'
     BarWidth: 0.8000
    FaceColor: [0.3010 0.7450 0.9330]
    EdgeColor: [0 0 0]
    BaseValue: 0
        XData: [1 2 3 4 5 6 7 8 9 10 11 12]
        YData: [635.5853 635.5853 635.5853 635.5853 243.8173 166.5324 117.9960 117.2723 109.5398 90.2098 65.2230 0]

  Show all properties
```

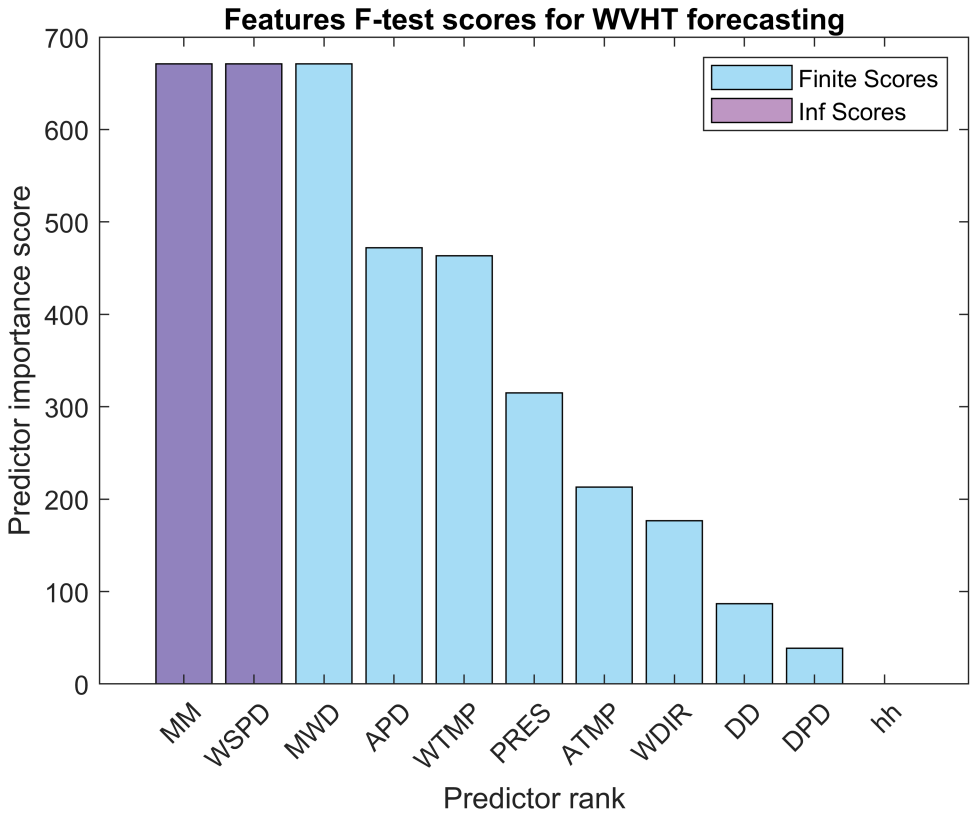
We observe every feature except YY gives a *F*-test score higher that 1.3, especially WDIR and WVHT (and therefore DPD and APD) are highly significant in relation to Wind speed

For Waves forecasting

Let's explore features F-test scores for Significant Wave height (WVHT) forecasting

Let's apply Feature selection method

```
[idx2, scores2,b2] = fSelection(data2018(:,{'MM','DD','hh','WDIR','WSPD','WVHT','DPD','APD','MW
```



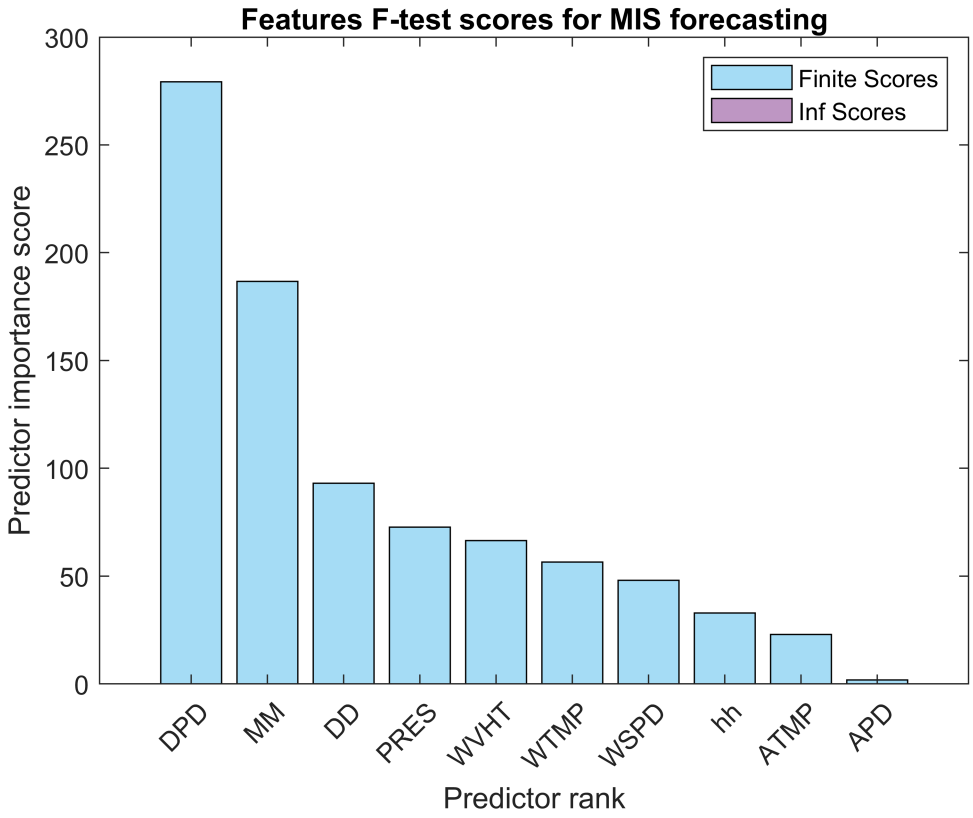

```
idx2 = 1×11
     1     5     8     7    11     9    10     4     2     6     3
scores2 = 1×11
  671.1943   87.0205    0.0556  176.6359  671.1943   38.7265  471.9295  671.1943 ···
b2 =
  Bar (Finite Scores) with properties:

    BarLayout: 'grouped'
     BarWidth: 0.8000
    FaceColor: [0.3010 0.7450 0.9330]
    EdgeColor: [0 0 0]
    BaseValue: 0
        XData: [1 2 3 4 5 6 7 8 9 10 11]
        YData: [671.1943 671.1943 671.1943 471.9295 463.2801 314.8759 212.9943 176.6359 87.0205 38.7265 0.0556]

  Show all properties
```

For Misalignment forecasting

```
[idx, scores,b1] = fSelection(data2018(:,{'MM','DD','hh','WSPD','WVHT','DPD','APD','PRES','ATMP
```




```
idx = 1×10
     6     1     2     8     5    10     4     3     9     7
scores = 1×10
  186.7756   93.1225   32.9219   48.1088   66.4262  279.3149    1.7921   72.7828 ···
b1 =
  Bar (Finite Scores) with properties:

    BarLayout: 'grouped'
     BarWidth: 0.8000
    FaceColor: [0.3010 0.7450 0.9330]
    EdgeColor: [0 0 0]
    BaseValue: 0
        XData: [1 2 3 4 5 6 7 8 9 10]
        YData: [279.3149 186.7756 93.1225 72.7828 66.4262 56.5766 48.1088 32.9219 22.9956 1.7921]

  Show all properties
```

4 Data combination for time series forecasting

Wind speed forecasting

To obtain the data sets with data from 1, 2,..., 6 hours before : use the files .m listed above

Data from 1 hour-before : 1hbefore_WSPD.m

Data from 2 h-before: 2hbefore_WSPD.m

...

Data from 6h-before: 6hbefore_WSPD.m

For wind speed forecasting with NAR and NARX

```
targetsWind2018 = data2018.WSPD;
targetsWind2019 = data2019.WSPD;
targetsWindRealTime = dataRealTime.WSPD;
predictorsWind2018 = table2array(data2018(:,{'WSPD','ATMP','PRES','WDIR'}));
predictorsWind2019 = table2array(data2019(:,{'WSPD','ATMP','PRES','WDIR'}));
predictorsWindRealTime = table2array(dataRealTime(:,{'WSPD','ATMP','PRES','WDIR'}));
```

Significant Waves height forecasting

For Significant Waves Height prediction in the frequency domain we select features in the Regression learner toolbox directly from the initial data sets

On the other hand, for Feedforward neural networks we separate predictors and targets:

```
predictorsWaves2018a = table2array(data2018(:,{'WSPD','MWD','WDIR'}))';
predictorsWaves2018b = table2array(data2018(:,{'WSPD','MWD','WDIR','PRES','WTMP'}))';
targetsWaves2018 = data2018.WVHT;
```

```
predictorsWaves2019a = table2array(data2019(:,{'WSPD','MWD','WDIR'}))';
predictorsWaves2019b = table2array(data2019(:,{'WSPD','MWD','WDIR','PRES','WTMP'}))';
targetsWaves2019 = data2019.WVHT;
```

```
predictorsWavesRealTimea = table2array(dataRealTime(:,{'WSPD','MWD','WDIR'}))';
predictorsWavesRealTimeb = table2array(dataRealTime(:,{'WSPD','MWD','WDIR','PRES','WTMP'}))';
targetsWavesRealTime = dataRealTime.WVHT;
```

Misalignment Forecasting

To obtain the data sets with data from 1 hour before we execute 1hoursbefore_MIS.m

Other tries of modeling with more hours before values as predictors have been made giving similar or even poorest results, so this models haven't been include in the memory and thus data sets with more than 1 hour before predictors have not been included in the repository.

For NAR and NARX training and testing:

```
predictorsMis2018a = table2array(data2018(:,{'MIS','WSPD'}));
predictorsMis1018a = table2array(data10_Real_clean(data10_Real_clean.YY ~= 2019 & data10_Real_c
predictorsMis2019a = table2array(data10_Real_clean(data10_Real_clean.YY == 2019,{'MIS','WSPD'})
predictorsMisRealTimea = table2array(data10_Real_clean(data10_Real_clean.YY == 2020,{'MIS','WSP
```

```
targetMis2018 = table2array(data10_Real_clean(data10_Real_clean.YY == 2018,{'MIS'}));
targetMis1018 = table2array(data10_Real_clean(data10_Real_clean.YY ~= 2019 & data10_Real_clean.
targetMis2019 = table2array(data10_Real_clean(data10_Real_clean.YY == 2019,{'MIS'}));
targetMisRealTime = table2array(data10_Real_clean(data10_Real_clean.YY == 2020,{'MIS'}));
```