

Supplemental material to:
“The matching pursuit algorithm revisited:
A variant for big data and new stopping rules”

Fangyao Li^a, Christopher M. Triggs^a, Bogdan Dumitrescu^b, and Ciprian Doru Giurcăneanu^a

^aDept. of Statistics, Univ. of Auckland, New Zealand
^bUniv. Politehnica of Bucharest, Romania

Contents

1 Additional information on the simulation study	2
1.1 MISE tables	2
1.1.1 Experiments for $n = 20$	2
1.1.2 Experiments for $n = 100$	4
1.1.3 Experiments for $n = 10000$	6
1.2 Median number of iterations	8
1.2.1 Experiments for $n = 20$	8
1.2.2 Experiments for $n = 100$	10
1.2.3 Experiments for $n = 10000$	12
1.3 Median number of predictors	14
1.3.1 Experiments for $n = 20$	14
1.3.2 Experiments for $n = 100$	16
1.3.3 Experiments for $n = 10000$	18
1.4 Score plots	20
2 Additional information on experiments with air pollution data	23
2.1 Location of the four sites where the concentrations of the air pollutants are measured	23
2.2 Detailed information on the experiments with air pollution data collected from Patumahoe site	24
2.2.1 The number of iterations and the number of predictors	24
2.2.2 Statistics on how many times each predictor is selected in 100 runs	27
2.2.3 How many times each predictor from the constrained set is selected in 100 runs: The case when the total number of predictors is $p_n = 68$ (ConSet) versus the case when the total number of predictors is $p_n = 1464$ (FullSet)	28
2.2.4 Statistics for EgMDL ₁ * on how many times each predictor from the FullSet ($p_n = 1464$) not included in the ConSet ($p_n = 68$) is selected in 100 runs	30

1.4 Score plots

In this section, we plot graphs similar to those in [1, Figs. 1-4]. The scores presented in each graph are computed by aggregating the scores obtained from all the experiments that have a common parameter. For example, we show in Figure 1 the scores aggregated from all the experiments in which SNR is high ($\zeta^2 = 8$).

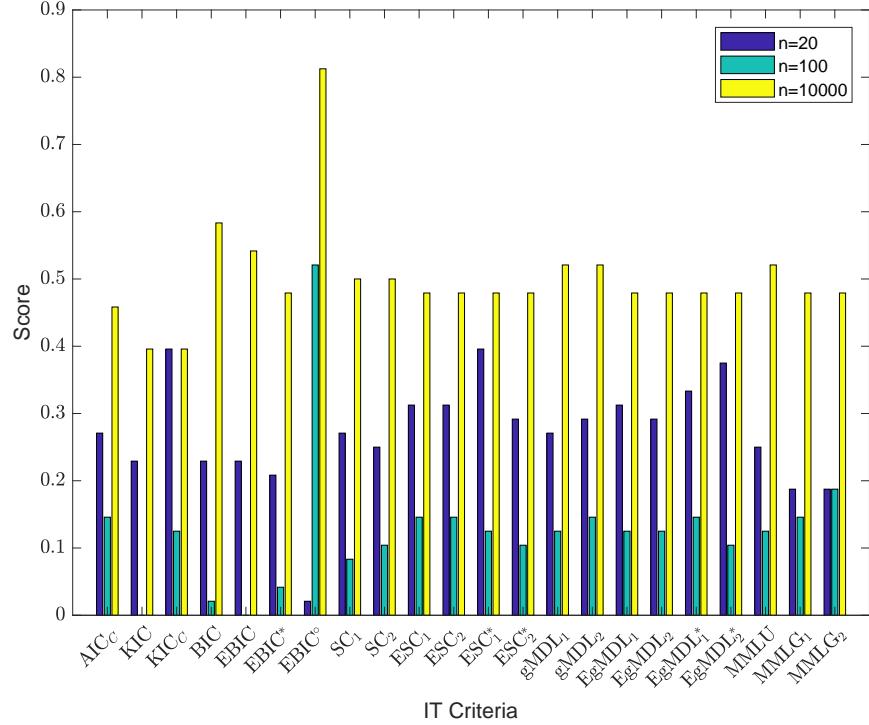


Figure 1: Scores aggregated from all experiments in which SNR is high ($\zeta^2 = 8$).

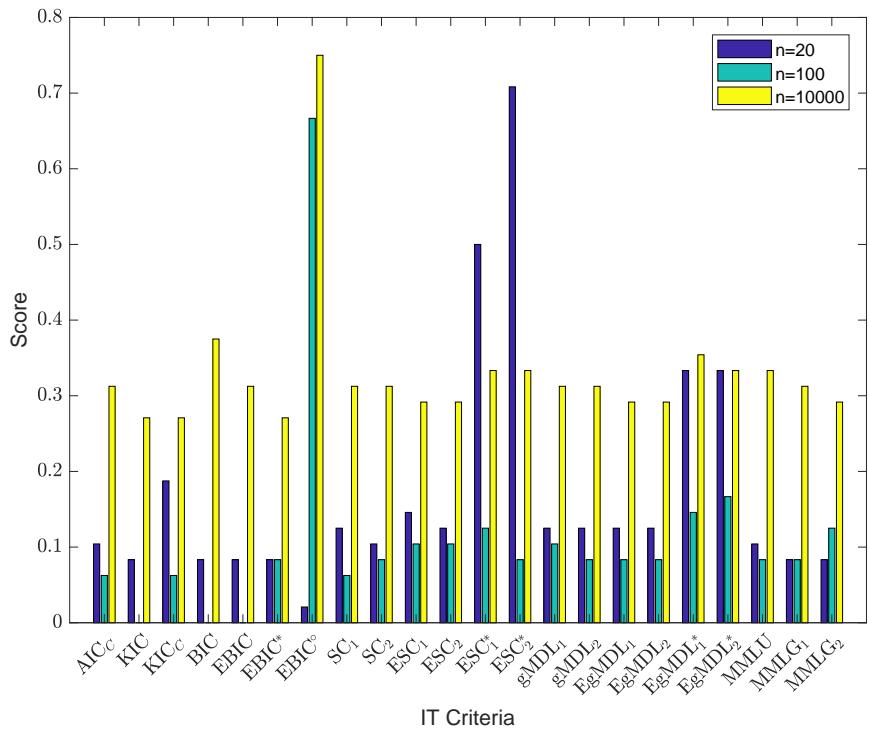


Figure 2: Scores aggregated from all experiments in which $\omega = 0$.

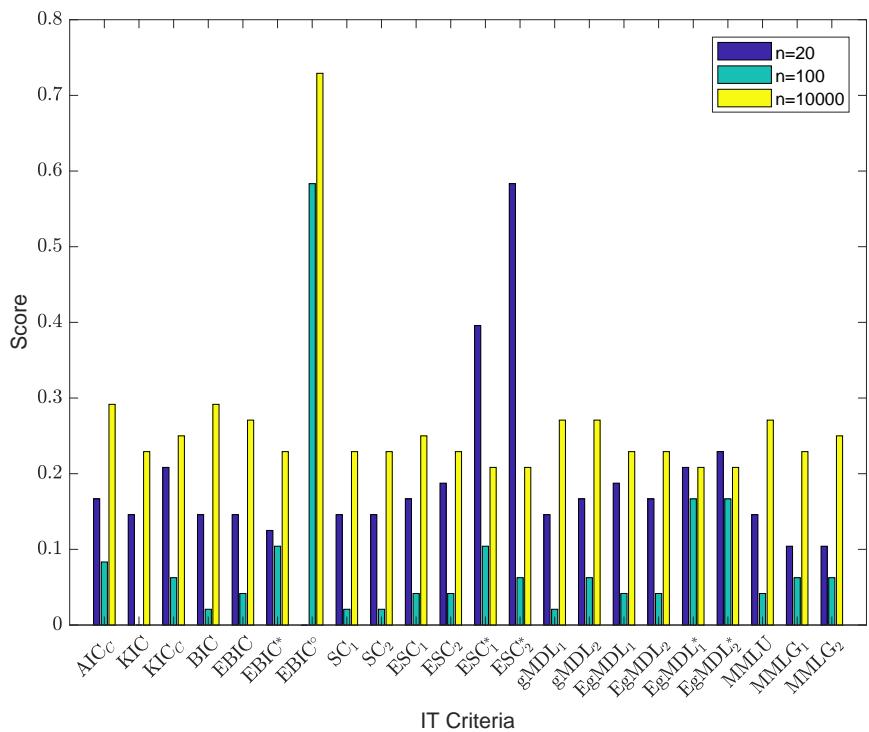


Figure 3: Scores aggregated from all experiments in which $\omega = 0.75$.

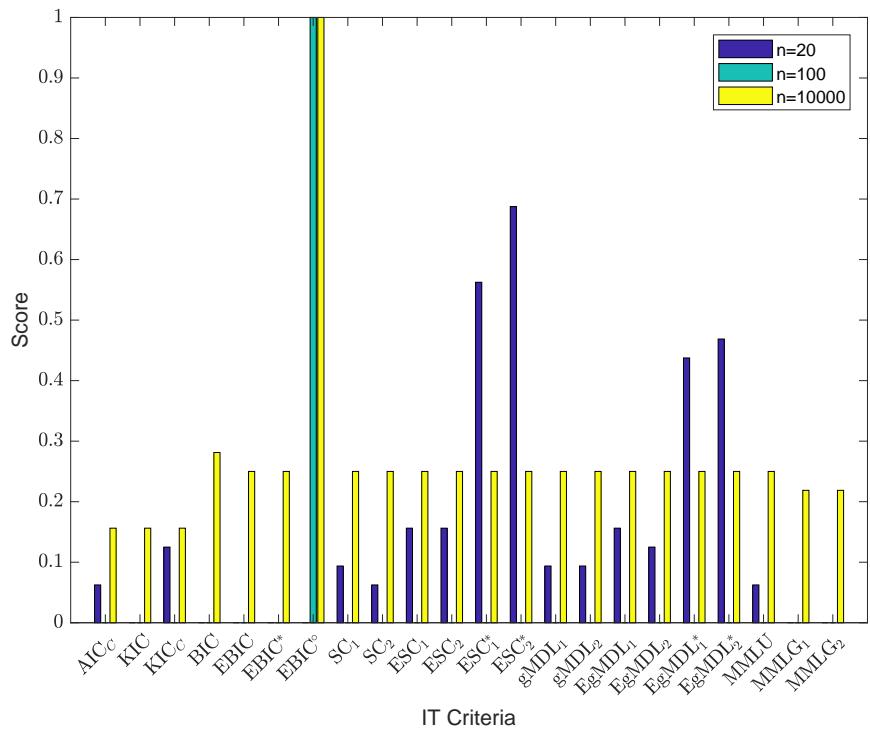


Figure 4: Scores aggregated from all experiments done for Case 2.

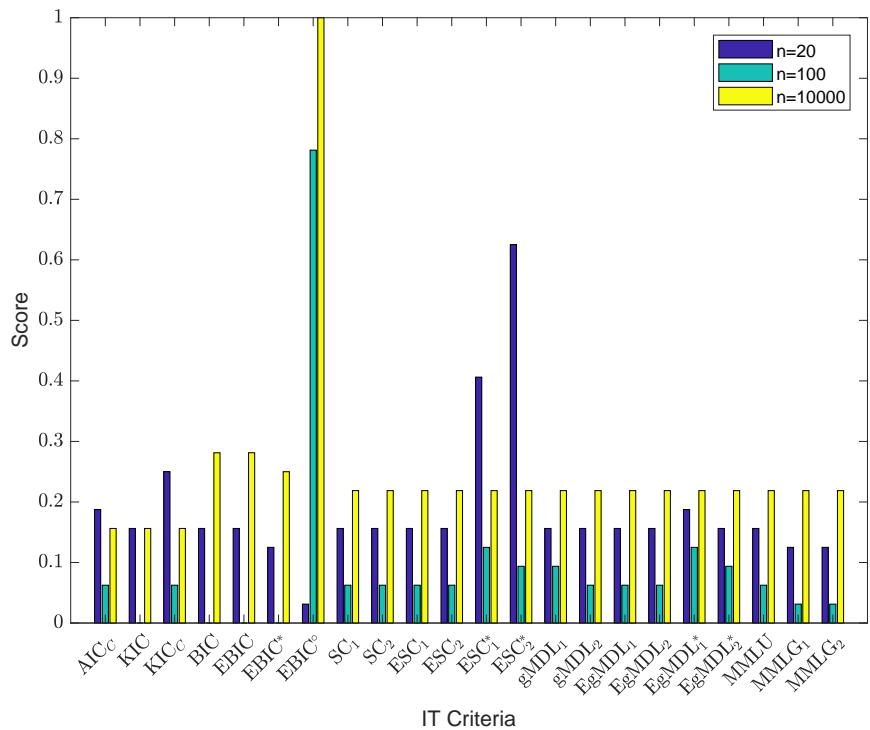


Figure 5: Scores aggregated from all experiments done for Case 3.

2 Additional information on experiments with air pollution data

2.1 Location of the four sites where the concentrations of the air pollutants are measured

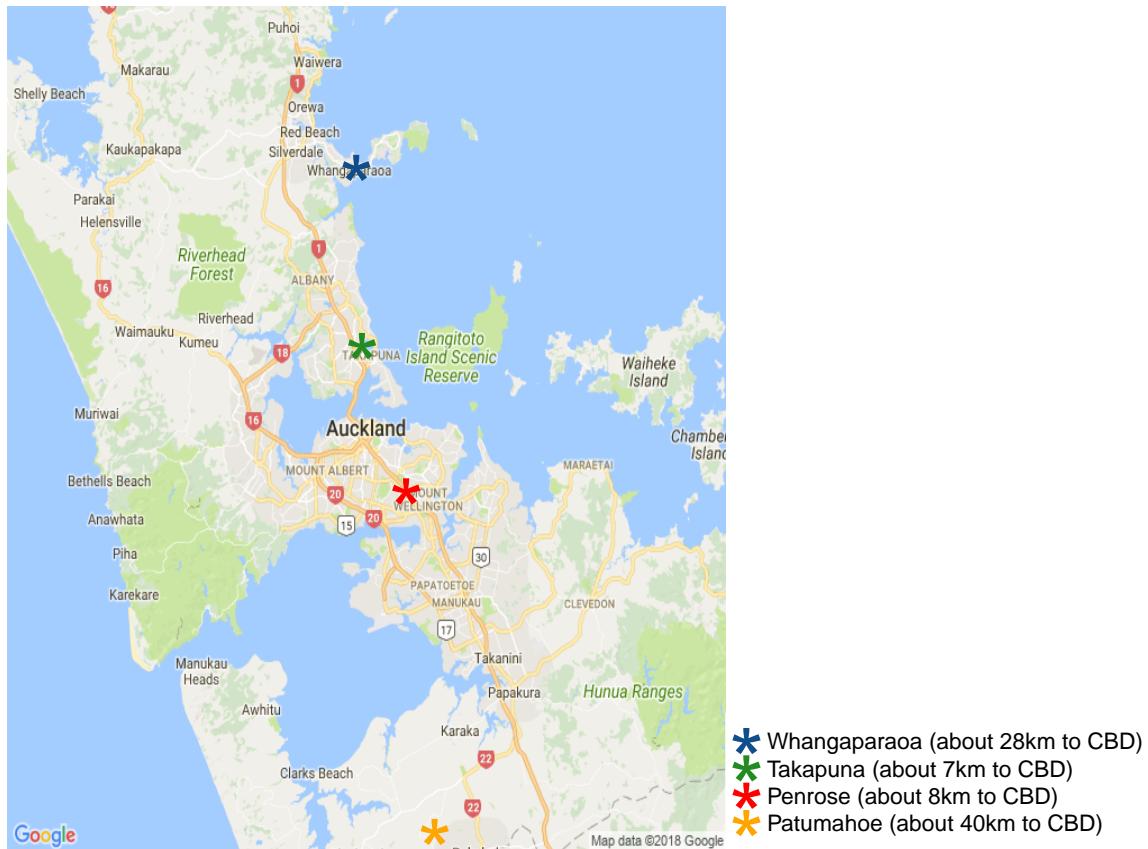


Figure 6: The distances shown in the legend are measured from Auckland central business district (CBD).

2.2 Detailed information on the experiments with air pollution data collected from Patumahoe site

2.2.1 The number of iterations and the number of predictors

In Tables 37-38 we report, for all stopping rules, the statistics computed from 100 data frames (see [1, Sec. 5.2] for the definition of the data frame). The number of iterations and the number of predictors produced by five stopping rules (AIC_C , $EBIC^\circ$, ESC_2 , ESC_2^* , CV) for each data frame are shown in Figures 7-8.

Quantiles	FullSet			ConSet		
	0.025	0.50	0.975	0.025	0.50	0.975
AIC_C	9562.0	14834.5	19998.0	96.0	239.5	487.0
KIC	20000.0	20000.0	20000.0	74.0	119.0	287.0
KIC_C	4602.0	7474.0	13474.0	71.0	113.0	246.0
BIC	19999.0	20000.0	20000.0	50.0	64.5	103.0
$EBIC$	19999.0	20000.0	20000.0	45.0	61.0	96.0
$EBIC^*$	44.0	62.0	94.0	37.0	47.5	62.0
$EBIC^\circ$	27.0	32.0	40.0	32.0	37.0	44.0
SC_1	2298.0	4687.5	10849.0	48.0	61.5	128.0
SC_2	2258.0	4912.5	11677.0	48.0	61.5	128.0
ESC_1	667.0	1609.5	4148.0	44.0	59.0	104.0
ESC_2	563.0	1546.0	5588.0	44.0	59.0	104.0
ESC_1^*	43.0	62.0	94.0	37.0	47.0	62.0
ESC_2^*	41.0	59.5	94.0	37.0	47.0	62.0
$gMDL_1$	2298.0	4687.5	10849.0	49.0	66.0	128.0
$gMDL_2$	2258.0	4912.5	11677.0	49.0	66.0	128.0
$EgMDL_1$	667.0	1609.5	4148.0	45.0	60.5	117.0
$EgMDL_2$	563.0	1552.5	5588.0	45.0	60.5	110.0
$EgMDL_1^*$	44.0	63.5	94.0	37.0	47.0	62.0
$EgMDL_2^*$	44.0	62.0	94.0	37.0	47.0	62.0
MMLU	2298.0	4962.0	11677.0	48.0	61.5	126.0
MMLG ₁	8195.0	17609.0	20000.0	45.0	61.0	136.0
MMLG ₂	9154.0	17959.0	20000.0	45.0	61.0	136.0
CV	111.0	4875.0	20000.0	94.0	241.5	6515.0

Table 37: Summary table for the number of iterations.

Quantiles	FullSet			ConSet		
	0.025	0.50	0.975	0.025	0.50	0.975
AIC _C	513.0	590.5	654.0	18.0	27.0	36.0
KIC	623.0	646.0	678.0	14.0	20.0	29.0
KIC _C	397.0	475.5	571.0	14.0	19.0	27.0
BIC	623.0	646.0	678.0	7.0	13.0	17.0
EBIC	623.0	646.0	678.0	6.0	11.0	16.0
EBIC*	12.0	20.0	31.0	5.0	9.0	12.0
EBIC°	4.0	6.0	10.0	4.0	6.0	8.0
SC ₁	305.0	392.0	542.0	7.0	12.0	19.0
SC ₂	302.0	401.5	555.0	7.0	12.0	19.0
ESC ₁	149.0	252.0	375.0	6.0	11.0	17.0
ESC ₂	136.0	240.5	423.0	6.0	11.0	17.0
ESC ₁ *	12.0	20.0	32.0	5.0	8.0	12.0
ESC ₂ *	11.0	19.5	31.0	5.0	8.0	12.0
gMDL ₁	305.0	392.0	542.0	7.0	13.0	20.0
gMDL ₂	302.0	401.5	555.0	7.0	13.0	19.0
EgMDL ₁	149.0	252.0	375.0	6.0	11.0	18.0
EgMDL ₂	136.0	242.5	423.0	6.0	11.0	18.0
EgMDL ₁ *	12.0	21.0	32.0	5.0	8.0	12.0
EgMDL ₂ *	12.0	20.0	31.0	5.0	8.0	12.0
MMLU	305.0	408.0	555.0	7.0	12.0	18.0
MMLG ₁	494.0	615.5	654.0	7.0	12.0	20.0
MMLG ₂	513.0	618.0	657.0	7.0	12.0	20.0
CV	42.0	397.5	660.0	16.0	31.5	66.0

Table 38: Summary table for the number of predictors.

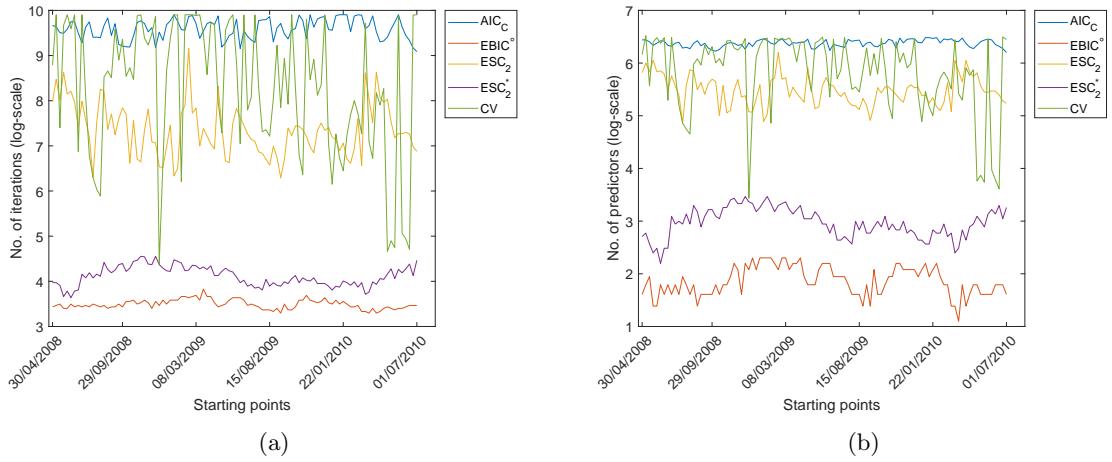


Figure 7: Results obtained for each data frame when the total number of predictors is $p_n = 1464$ (FullSet). For each plot, the starting points of the data frames are marked on the horizontal axis.

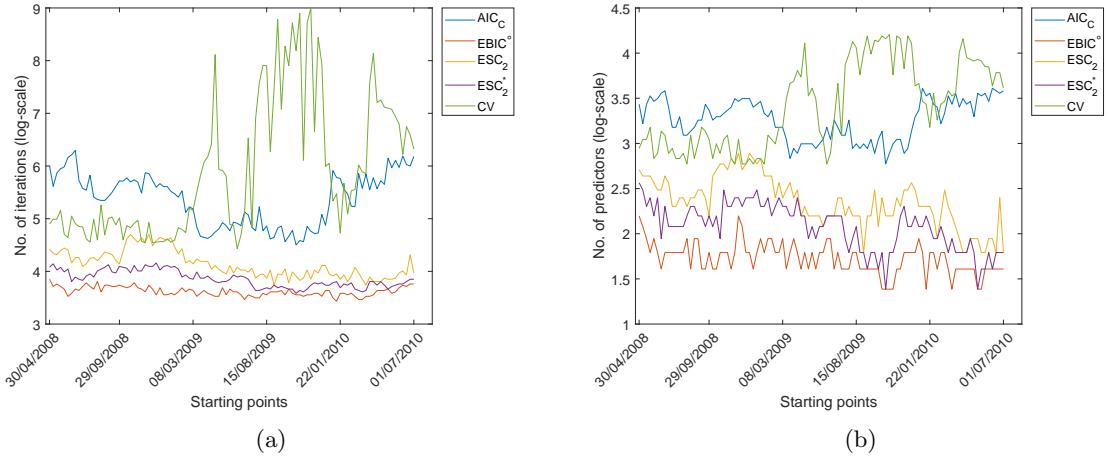


Figure 8: Results similar to those reported in Figure 7: The total number of predictors is $p_n = 68$ (ConSet).

2.2.2 Statistics on how many times each predictor is selected in 100 runs

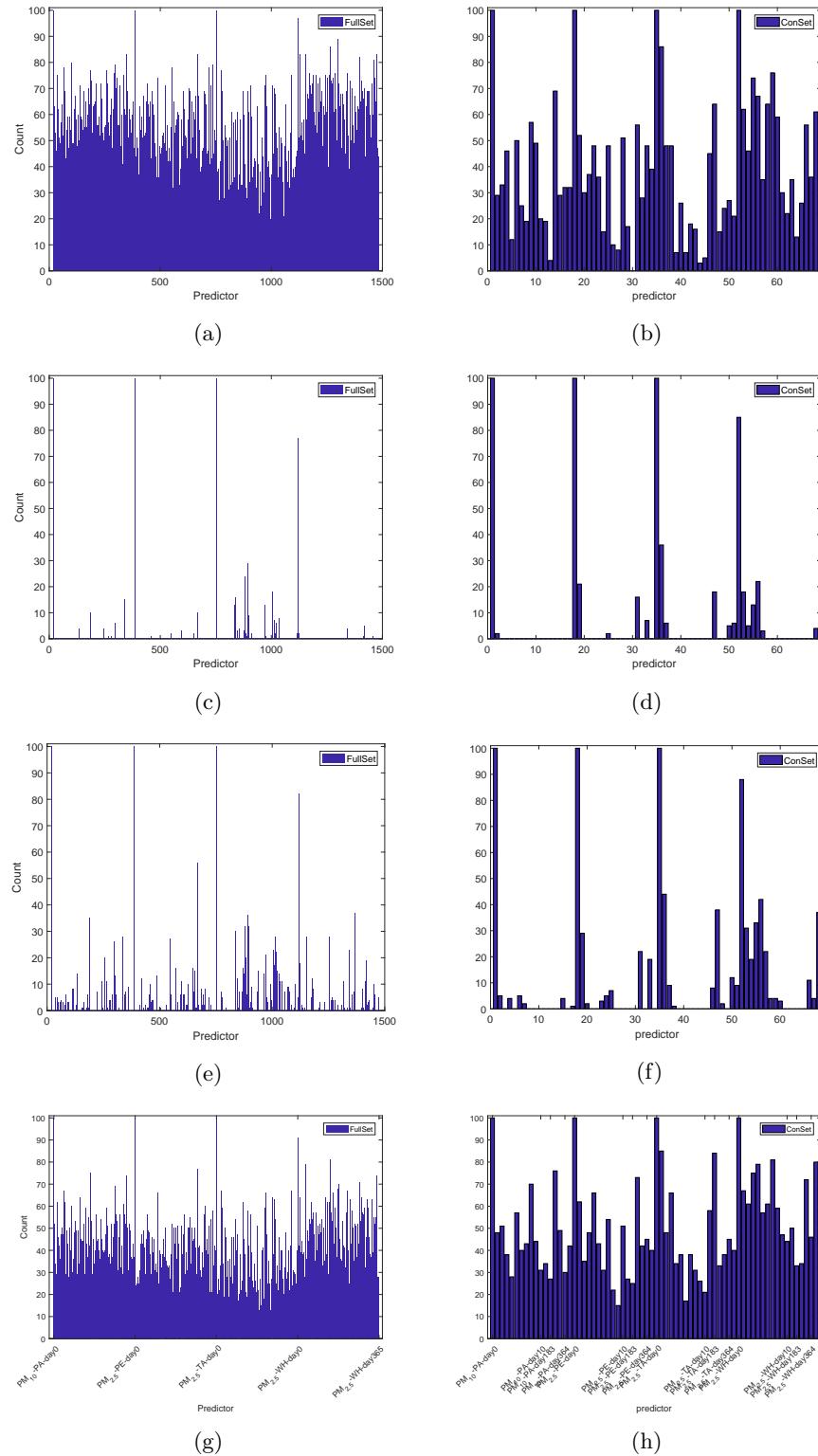


Figure 9: Results obtained with AIC_C [panels (a)-(b)], EBIC^o [Panels (c)-(d)], EgMDL₁^{*} [panels (e)-(f)] and CV [panels (g)-(h)]. On the horizontal axes, the predictors are enumerated from left to right, starting from the day t when the prediction is done and going back to day $t - 365$. Day t is marked on the graphs as day0.

- 2.2.3 How many times each predictor from the constrained set is selected in 100 runs: The case when the total number of predictors is $p_n = 68$ (ConSet) versus the case when the total number of predictors is $p_n = 1464$ (FullSet)**

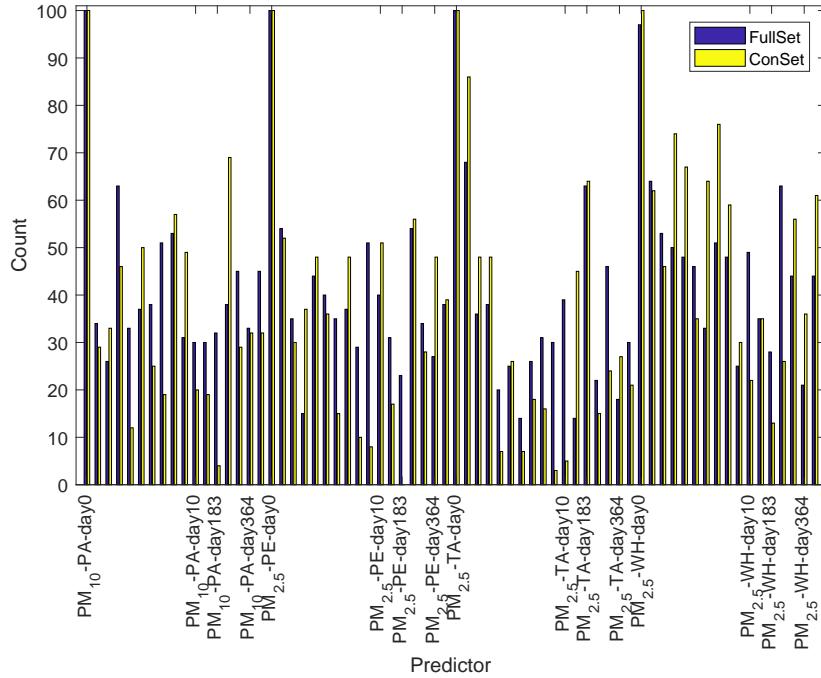


Figure 10: Results obtained when AIC_C is applied.

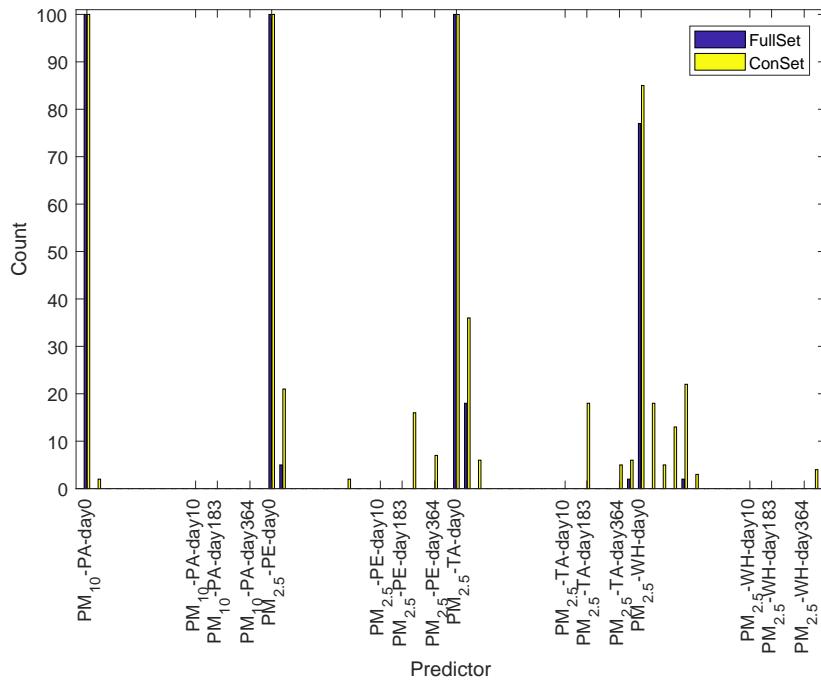


Figure 11: Results obtained when $EBIC^o$ is applied.

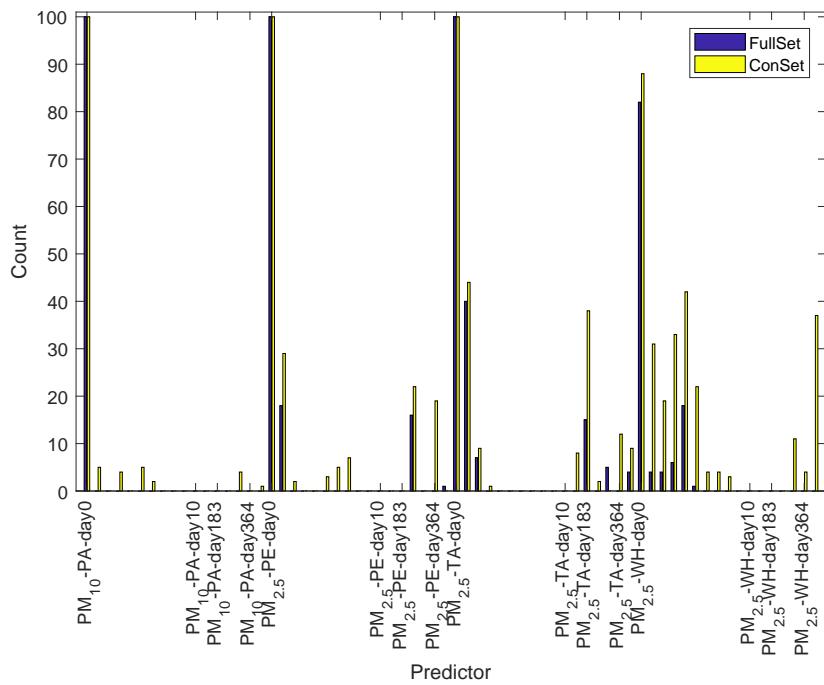


Figure 12: Results obtained when EgMDL₁* is applied.

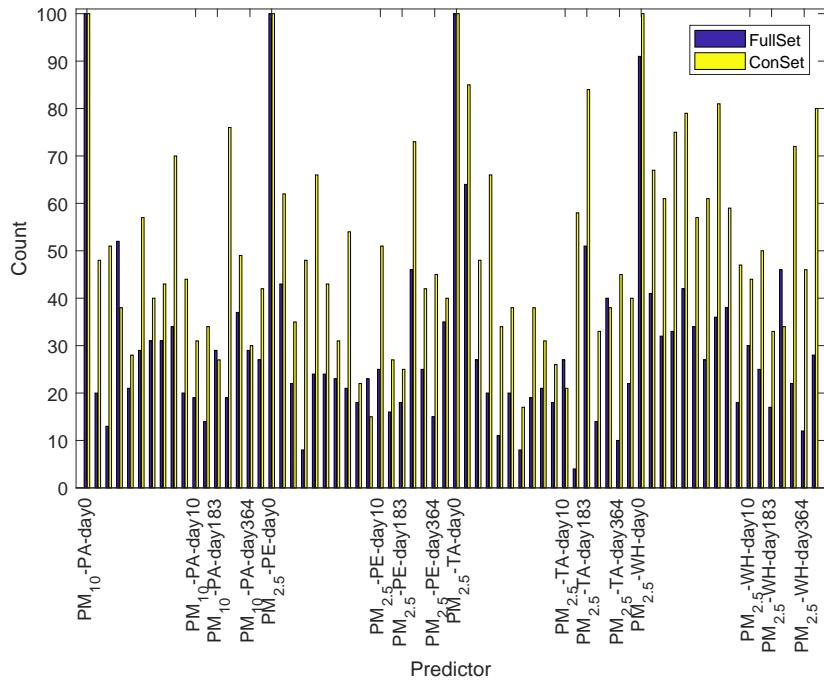


Figure 13: Results obtained when CV is applied.

2.2.4 Statistics for EgMDL₁^{*} on how many times each predictor from the FullSet ($p_n = 1464$) not included in the ConSet ($p_n = 68$) is selected in 100 runs

Predictor	Count
PM2.5-PE-day280	56
PM2.5-WH-day249	37
PM2.5-TA-day140	36
PM10-PA-day167	35
PM2.5-TA-day128	32
PM2.5-TA-day145	32
PM2.5-TA-day84	30
PM10-PA-day317	28
PM2.5-TA-day261	28
PM2.5-WH-day36	28
PM2.5-WH-day134	28
PM2.5-PE-day161	27
PM10-PA-day277	26
PM2.5-TA-day251	23
PM2.5-WH-day224	23
PM2.5-TA-day267	22
PM2.5-TA-day217	21
PM10-PA-day233	20
PM2.5-TA-day129	20

Table 39: List of predictors which have been selected at least twenty times in one hundred runs.

References

- [1] F. Li, C.M. Triggs, B. Dumitrescu and C.D. Giurcăneanu. The matching pursuit algorithm revisited: A variant for big data and new stopping rules. 2018.