Water Resources Research

Supporting Information for

Water demand forecasting accuracy and influencing factors at different spatial scales using a Gradient Boosting Machine

M.Xenocristou[1], C. Hutton[2], J. Hofman[3], and Z. Kapelan[4]

[1, 4] Centre for Water Systems, University of Exeter, North Park Road, EX4 4QF Exeter, U.K.

[2] Wessex Water, Claverton Down Road, BA2 7WW Bath, U.K.

[3] Water Innovation and Research Centre, University of Bath, BA2 7AY Bath Avon, U.K.

[4] Delft University of Technology, Stevinweg 1, 2628CN Delft, Netherlands

Contents of this file

Figure S1
Tables S1 to S4

Introduction

This supplement contains supporting information regarding the household composition of each spatial aggregation, i.e. the variation in types of households among the groups and days in the data (Figure S1). In addition, it also includes detailed information regarding the model hyperparameter values that were selected as a result of the model tuning process (Tables S1-S4).

The automated machine learning (automl) module of h2o was used to tune a range of GBM models trained on different aggregations of properties and different input configurations. H2o tuned the models for 9 hyperparameter values using a random grid search. Table S1 shows the chosen parameters for each one of the models that were trained with only 7 days of past consumption as input, for 9 different aggregations of properties. Tables S2 to S4 show the hyperparameter values for each one of Models 1-8 for each spatial aggregation of properties. Table S2 refers to aggregation at the District level, Table S3 at the area level, while Table S4 represents the Network level. The 'auto' tag under the histogram type means that the cuts to be tested for splitting at each node of the decision trees were chosen by dividing the variable range in equal steps, which in this case were 20. As it can be seen from the above tables, when the learning rate of the algorithm decreases, the number of trees increases, as the model requires more trees in order to converge to a solution when the trees have smaller contributions to the final result.

These values are provided for guidance only and as a good starting point for the hyperparameter values but they do not replace the need for tuning the model based on the corresponding dataset.

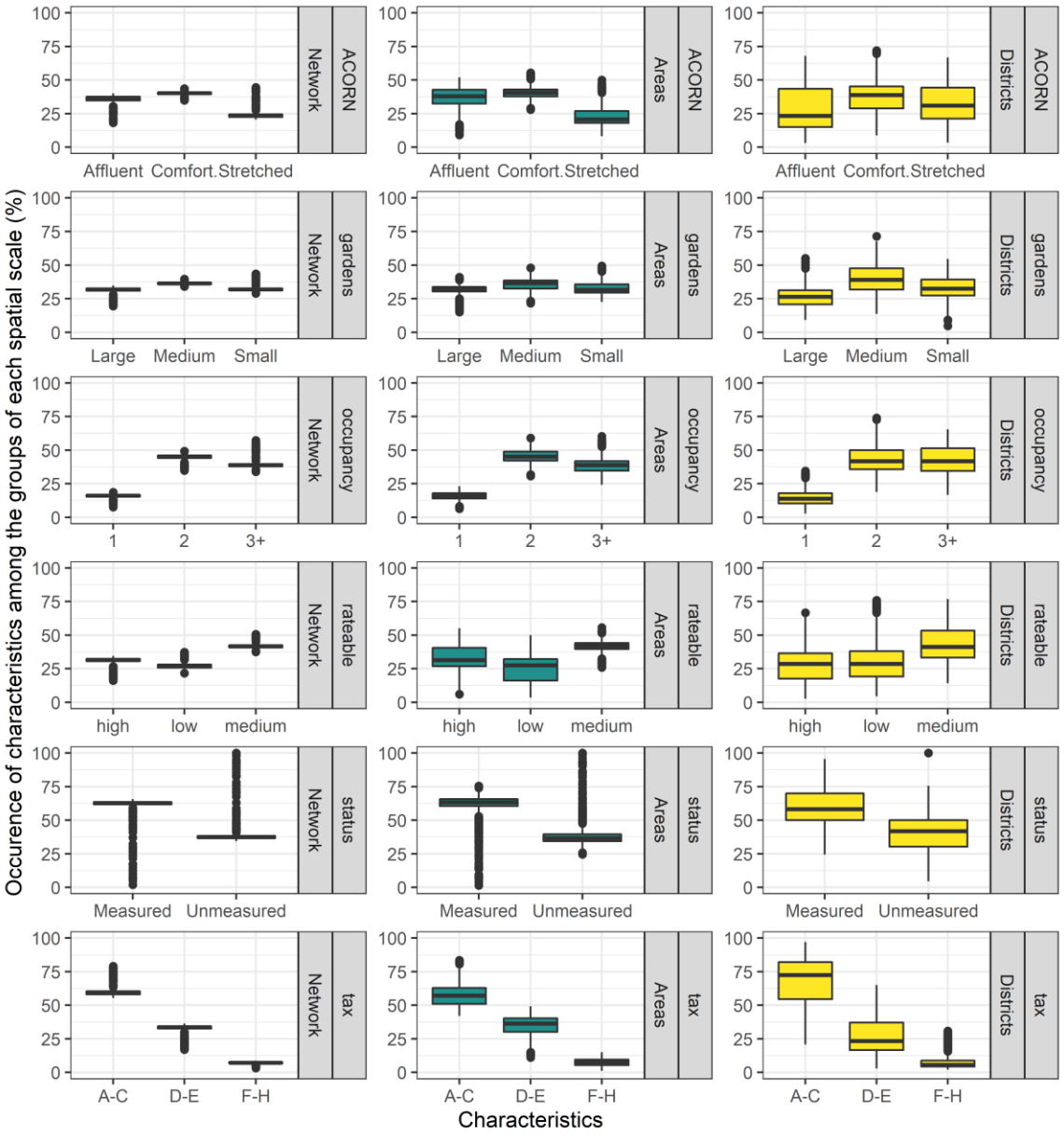

Figure S1. Group composition in terms of household types among the groups for each level of spatial aggregation.

Hyperparameters	5	10	20	40	80	120	200	400	600
Ntrees	44	54	329	3545	45	605	30	34	758
Max_depth	6	3	5	5	15	8	10	10	16
Learn_rate	0.1	0.1	0.01	0.001	0.1	0.005	0.1	0.1	0.005
Sample_rate	0.8	0.8	0.7	0.9	0.8	0.5	0.8	0.8	0.9
Col_sample_rate	0.8	1	1	0.4	0.8	0.7	0.8	0.8	1
Col_saple_rate_per_tree	0.8	1	0.7	1	0.8	0.7	0.8	0.8	0.4
Histogram_type	auto	auto	auto	auto	auto	auto	auto	auto	auto
Min_split_imrpovement	1e-05	1e-05	1e-05	1e-04	1e-05	1e-05	1e-05	1e-05	1e-04
Min_rows	1	5	10	100	100	15	10	10	30

Table S1: Hyperparameters for the GBM models for different group sizes.

Hyperparameters	1	2	3	4	5	6	7	8
Ntrees	81	388	342	352	104	81	28	104
Max_depth	3	5	5	5	3	3	3	3
Learn_rate	0.1	0.01	0.01	0.01	0.08	0.1	0.1	0.08
Sample_rate	0.8	0.7	0.7	0.7	1	0.8	0.8	1
Col_sample_rate	1	1	1	1	0.4	1	1	0.4
Col_saple_rate_per_tree	1	0.7	0.7	0.7	0.4	1	1	0.4
Histogram_type	auto	auto	auto	auto	auto	auto	auto	auto
Min_split_imrpovement	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05
Min_rows	5	10	10	10	5	5	5	5

Table S2: Hyperparameters for the GBM models 1-8 for the District level.

Hyperparameters	1	2	3	4	5	6	7	8
Ntrees	56	446	46	52	59	47	55	36
Max_depth	10	5	15	15	10	8	3	8
Learn_rate	0.1	0.01	0.1	0.1	0.1	0.1	0.1	0.1
Sample_rate	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8
Col_sample_rate	0.8	1	0.8	0.8	0.8	0.8	1	0.8
Col_saple_rate_per_tree	0.8	0.7	0.8	0.8	0.8	0.8	1	0.8
Histogram_type	auto	auto	auto	auto	auto	auto	auto	auto
Min_split_imrpovement	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05
Min_rows	10	10	100	100	10	10	5	10

Table S3: Hyperparameters for the GBM models 1-8 for the Area level.

Hyperparameters	1	2	3	4	5	6	7	8
Ntrees	118	62	5708	98	42	85	42	35
Max_depth	8	3	12	15	8	15	3	3
Learn_rate	0.05	0.1	0.001	0.1	0.1	0.1	0.1	0.1
Sample_rate	1	0.8	0.9	0.8	0.8	0.8	0.8	0.8
Col_sample_rate	1	1	0.7	0.8	0.8	0.8	1	1
Col_saple_rate_per_tree	1	1	0.7	0.8	0.8	0.8	1	1
Histogram_type	auto	auto	auto	auto	auto	auto	auto	auto
Min_split_imrpovement	1e-04	1e-05	1e-04	1e-05	1e-05	1e-05	1e-05	1e-05
Min_rows	5	5	100	100	10	100	5	5

Table S4: Hyperparameters for the GBM models 1-8 for the Network level.