Gradient Boosting Machine for Phosphorus Removal Prediction in Multi-Soil-Layering (MSL) system operated in a rural area

***Sofyan Sbahi**, Naaila Ouazzani, Abdessamed Hejjaj, Aberrahman Lahrouni, and Laila Mandi*

Cadi Ayyad University, Faculty of science Semlalia, Marrakesh, Morocco (sofyansbahi@gmail.com)

Introduction

Simulating the efficiency of wastewater treatment plants is critical in the management of its effluents

Challenge :

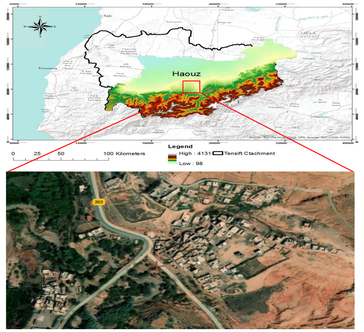
- Linear models that use a high dimensional dataset may be unable to build accurate and interpretable models.

To address this complexity :

- Multiple linear regression (MLR) # Gradient boosting machine algorithm (GBM)

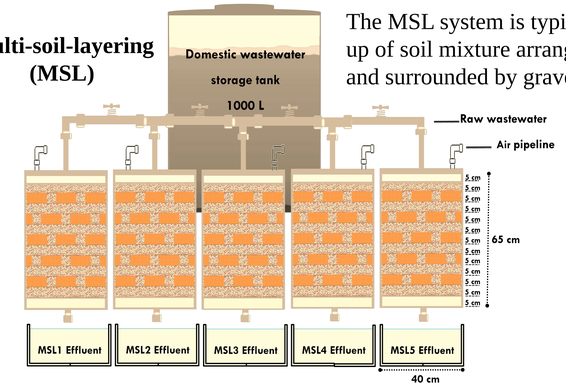
MATERIAL AND METHODS

Study area and WWTP



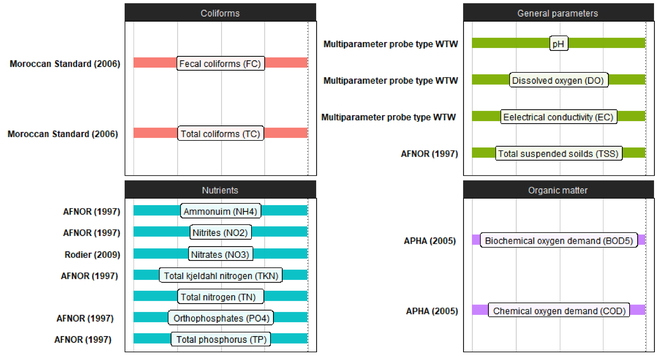
Talat Marghen village in Al-Haouz Province, Morocco

Multi-soil-layering (MSL)

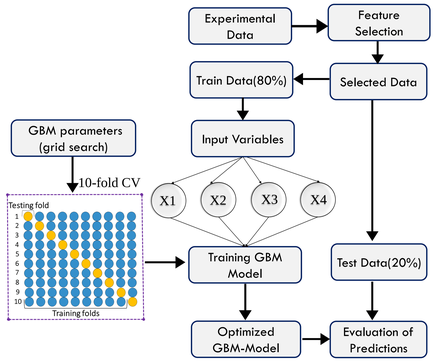


The MSL system is typically made up of soil mixture arranged in bricks and surrounded by gravel particles

… Water quality indicators



Methodology



MODELING

Pseudo-code of the GBM algorithm

1- $F_0(x) = \underset{\gamma}{\arg\min} \sum_{i=1}^{N} = L(y_i, \gamma)$

2- From the first tree (m=1) to the last tree (m=M)

3- For $i = 1, 2, \ldots, N$ compute

4- $\tilde{n}_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$

5- R_{jm}, For $j = 1, 2, \ldots . J_m$

6- For $j = 1, 2, \ldots . J_m$ compute

7- $\gamma_{jm} = \underset{\gamma}{\arg\min} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

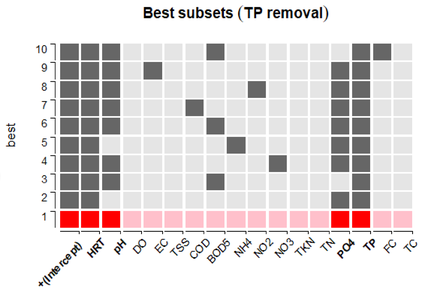
Friedman (2002): Boosting learning and editing error of decision trees is used to develop the GBM algorithm

Feature Selection

Bayesian information criterion (BIC)

Relevant predictors :

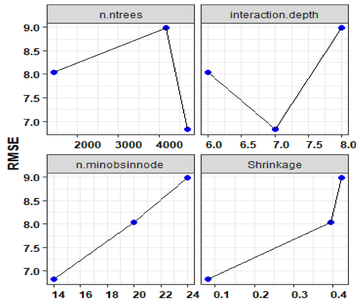
- Hydraulic retention time (HRT)
- Orthophosphates (PO_4–P)
- pH
- Total phosphorus (TP)



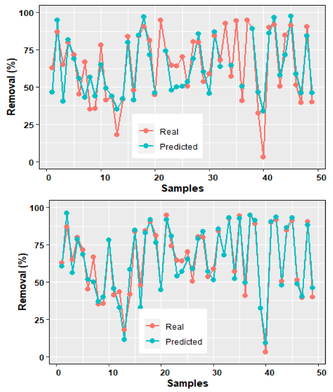
TUNING RESULTS

- n.ntrees = 4685,
- interaction.depth = 7,
- shrinkage (λ= 0.082)
- n.minobsinnode = 14

have a significant effect on reducing the training error (RMSE= 6.64 %) of the GBM model for TP removal



RESULTS of VALIDATION



MLR model

GBM model

CONCLUSION

The selected input variables have been shown to have significantly ($p < 0.05$) impact the removal of TP in the MSL system. In comparison, the GBM model has proved to be more accurate than the linear model (R^2 for the output variable exceeded 0.93). Thus, the GBM model is capable of investigating the nonlinear relationship between the removal of TP and the affecting parameters.

ACKNOWLEDGEMENT

The authors would like to thank the National Center for Studies and Research on Water and Energy (Cadi Ayyad University) and the Pole of competences of Water and Environment (PC2E) for their support of this research.