Application of machine learning to predict the performance of a municipal wastewater treatment plant anaerobic digester

Keywords: Process optimization, Supervised learning, Partial Least Squares Regression, Multilayer perceptron, Wastewater treatment

Mark McCormick

Use of machine learning

Motivation

- The study subject is a complex system comprising physical, chemical, and biological processes defined by 26 electro-mechanical, and 15 physical-chemical parameters
- What is needed are methods for system identification and predictive modeling
- Models can be used for monitoring and to adjust operational parameters to improve technical and economic performance
- Anaerobic digester performance can be evaluated in terms of energy efficiency and methane production

Methods

Data collection: The data comprised 365 days of online physical and mechanical measurements, and the results from laboratory analysis of samples from the anaerobic digester located at the wastewater treatment plant in Yverdon-les-Bains.

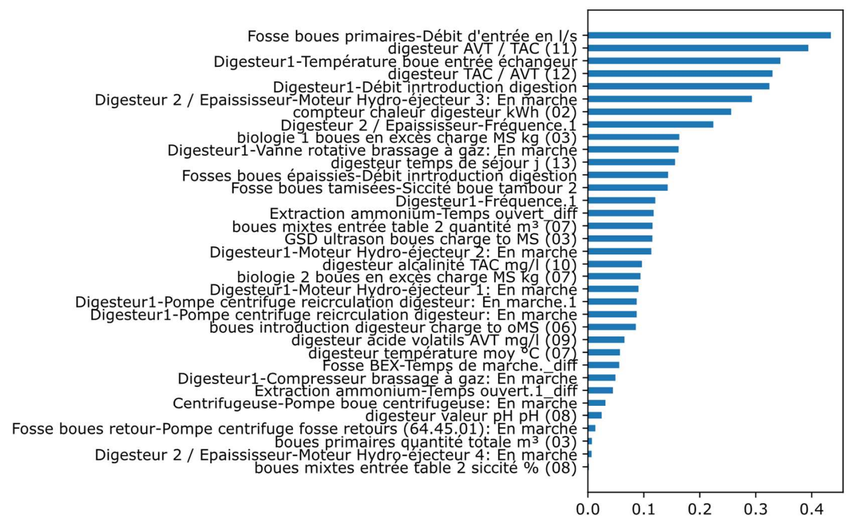
Data cleaning and transformation: A single dataset for model building was obtained by fusing separate files, resampling, and by interpolation and mean filling of missing data points (10% of the data). The dataset was spit into training and testing sets and loaded into 3 different machine learning algorithms[1,2]. The leave-one-out method and the non-use of an initialization seed were used to introduce randomness during training of 100 replicate models.

1. Multiple Linear Regression model (LM): seeks to minimize the residual sum of squares between the observed and the predicted response variables. The use of this model requires an orthogonal experimental design.
2. Partial Least Squares Regression model (PLSR): seeks to maximize the correlation between the predictors and the response variable. The number of components was set to 10. PLSR is not a predictive model.
3. Multilayer Perceptron model (MLP): seeks to minimize the residual sum of squares between the observed and the predicted response variables. The MLP model architecture had 1 dense layer with 24 nodes.

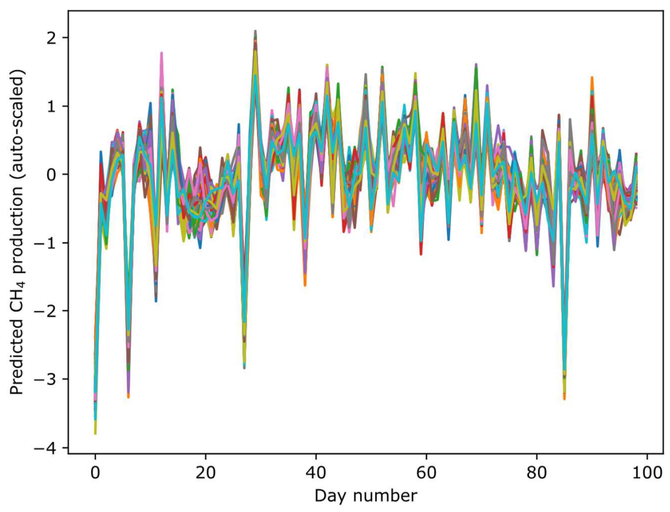
Results and Discussion

The ranking of the absolute value of PLSR coefficients shows that the parameters related to the feed rate and temperature, and the ratio of volatile fatty acids/total alkalinity are the most strongly correlated with methane production. Increasing the correlation of the parameters related to mixing, and consequently electrical efficiency, is a possible objective for optimisation. The MLP model has a higher coefficient of determination (R^2 = 0,61) than the LM model (R^2 = 0.50). This shows the higher accuracy of the MLP predictions.

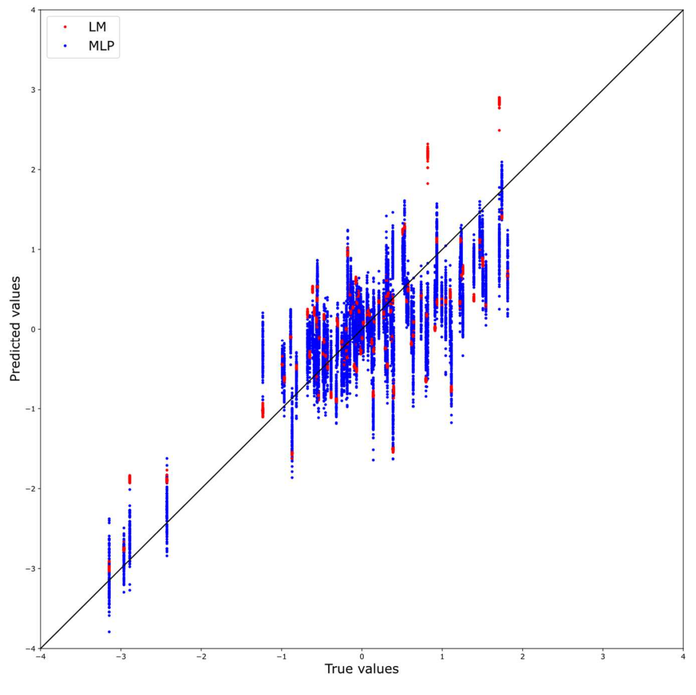
Model	R^2 (mean)	σ
Multiple Linear Regression (LM)	0.497	0.004
Partial Least Squares Regression (PLSR)	0.507	0.002
Multi-layer Perceptron (MLP)	0.608	0.031



PLSR: Digestion parameters ranked by coefficient absolute value



MLP: Predicted CH_4 production during 100 consecutive days



MLP (blue), LM (red): True vs predicted CH_4 production

Perspectives

Conclusions

- The PLSR algorithm can be used to extract information from operational data
- The most important predictors of methane production are candidates for process optimisation
- The ratio AVT/TAC is a more important predictor than the concentration of AVT in the digester
- The Multi Layer Perceptron model most accurately predicts methane production

Mark McCormick
Environmental engineer, M.Sc.
PhD, Information systems
mark.mccormick@alumnil.unil.ch
Tél: +41 78 604 52 42
www.markmccormick.ch

UNIL | Université de Lausanne

Neuro Heuristic Research Group

References and acknowledgements
1. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
2. Chollet, F. et. al. Keras; 2015.

Thank you to Marcel Pürro and Julien Ming from the Yverdon-les-Bains WWTP for their collaboration.