Using Random Forest to predict the BOD5 of an aerated lagoon at a pulp and paper mill

Presenter: L. M. ALMEIDA*

L. M. ALMEIDA[1], K. P. O. ESQUERRE[1], B. B. S. SILVA[1]

[1]Polytechnic School - Federal University of Bahia

Prof. Aristídes Novis St, 2 - Federação, Salvador - BA, Brazil

*lucasmascalmeida@gmail.com

INTRODUCTION

The aim of wastewater treatment systems is to reduce the pollution load of the effluent, in view of the adequate final disposal [1].

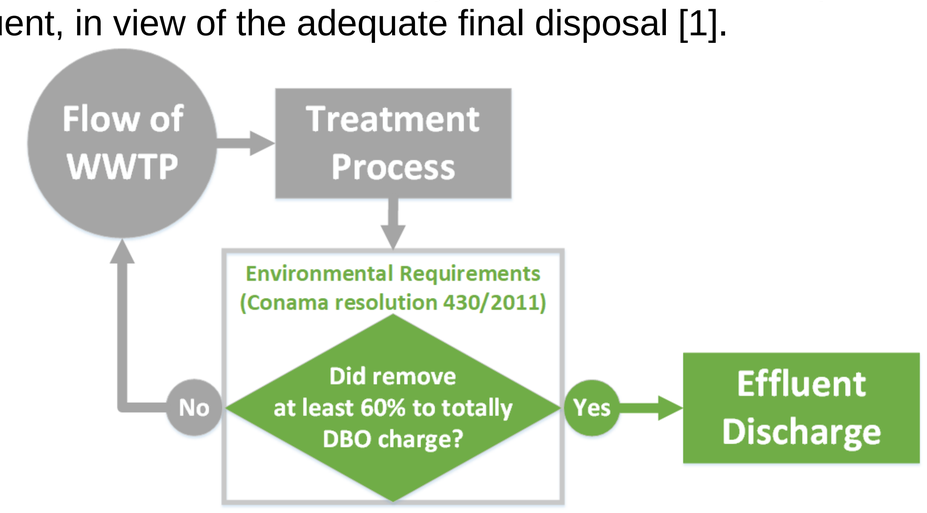

Figure 1: Illustration of the case study

Figure 1 shows how a WWPT should work at Brazil. Legally at Brazil, by the resolution 430/2011 of Conama, the effluents can be discharge after the correct treatment that follow some conditions, one condition is the charge of BOD should be removed at least 60% to totally charge [2].

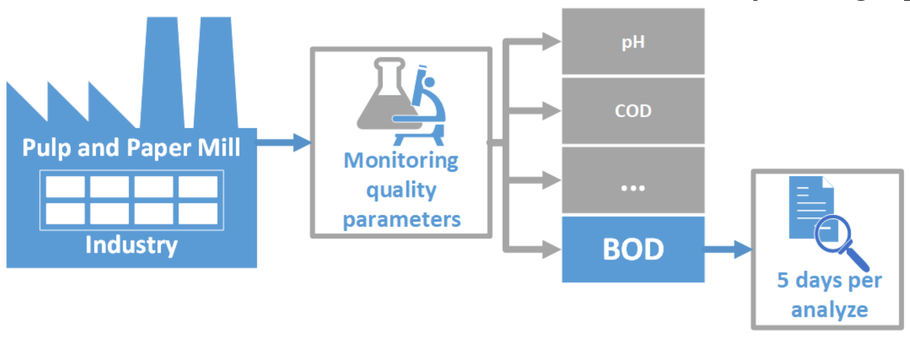

Figure 2: Monitoring quality parameters

Figure 2 shows some important effluent parameters of pulp and paper mill industry. The effluent parameters are very important because they translate the principal proprieties of effluent [3]. The principal effluent parameters of pulp and paper mill industry are BOD [1]. However, the obtain of these parameters can consume much time, mainly for BOD that demand five days per analyze [4].

AIM OF WORK

The aim of this work is to propose a monitoring alternative for BOD5 by the predictive method called Random Forest.

MATERIALS AND METHODS

Data set: This work was made on base in information of monitoring pulp and paper mill industry. The data are referent the window time from September 1998 to July 2000, with 1537 observation and 23 variables.

Parameter	Description	Mean	S.D	NA (%)
BODin	Inlet wastewater BOD (mg/L)	245.12	46.35	12.8
BODout	Outlet wastewater BOD (mg/L)	85.19	25.45	12.6
CODin	Inlet wastewater COD (mg/L)	561.39	104.16	12.8
CODout	Outlet wastewater COD (mg/L)	315.44	73.55	12.5
pH	pH	7.45	1.21	10.4
NAm	Inlet ammonia concentration (mg/L)	2.45	1.77	57.1
NN	Inlet nitrate concentration (mg/L)	1.43	0.88	81.8
P	-	1.41	1.00	83.1
Col	Color (mg/L)	464.40	123.52	10.3
T	Temperature (°C)	45.45	3.07	37.3
Cond	Conductivity (µS/cm a 25°C)	1530.46	378.03	10.6
RF	Rainfall (mm/dia)	4.82	11.50	23.6
Pulp	Pulp production (t/dia)	884.74	159.26	13.7
Pap	Paper production (t/dia)	1042.71	94.09	12.9
FR	Inlet flow rate (m³/dia)	67358.61	11592.81	7.0
SS	Inlet total suspended solids (mg/L)	149.20	85.74	63.0

Table 1: Description and some statistics for the main variables

Table 1 presents the descriptions of the variables. For NA (%) values higher than 35% the variable was excluded from the development.

Software: All work was developed by R. The main packages are the caret and tidyverse.

```
1  Select the number of models to build, m
2  for i = 1 to m do
3      Generate a bootstrap sample of the original data
4      Train a tree model on this sample
5      for each split do
6          Randomly select k (< P) of the original predictors
7          Select the best predictor among the k predictors and partition the data
8      end
9      Use typical tree model stopping criteria to determine when
       a tree is complete (but do not prune)
10 end
```

Algorithm 1: Random Forest Algorithm

Random Forest: A general random forests algorithm for a tree-based model can be implemented as shown in Algorithm 1 [5]. A range between 64 and 128 trees is enough for a good result by the Random Forest method [6]. For this work the quantity of trees is 128 trees.

Model evaluation: For prediction models the RMSE was utilized for choose the best model (mtry) of Random Forest models and the R^2 was utilized for evaluate the quality of model selected.

RESULTS

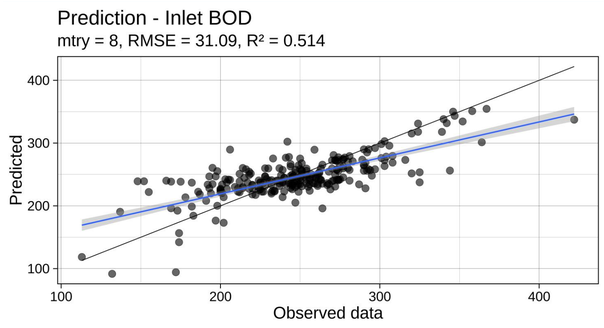

Figure 3: Checking Graph

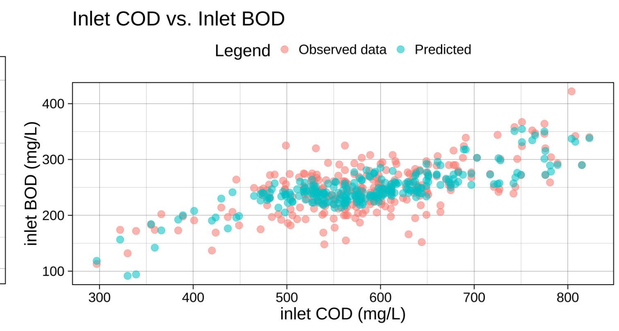

Figure 4: Observed/Predicted

Figure 3 exhibits the checking graph for Inlet BOD model. The model presents a RMSE = 31.09 and R^2 = 0.514. Figure 4 presents the comparison between the observed data and the prediction data (CODin vs BODin observed/predicted).

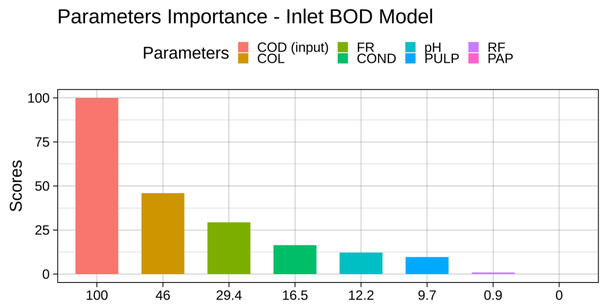

Figure 5: Importance Graph

Model	R^2
Random Forest	0.514
Multiple Linear Regression	0.492

Table 2: Comparing models

Figure 5 presents the main parameters to model develop. The positive highlight is the CODin. Table 2 shows the comparison between the result of this work and a similar work called *Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches* [7].

CONCLUSION

Although of the difficulties in relation the data missing problems and the complexity of modeling a bio process vulnerable to external factors, as rains and temperature variation, the prediction model of Inlet BOD (RMSE = 31.09 and R^2 = 0.514) presentation a satisfactory performance, and suggest that applications of machine learning algorithms, such as Random Forest, can be good tools for monitoring important variables in industrial processes.

For the next steps this work can be expand to estimate the efficiency of the Wastewater Treatment Plant. The same method can be applied to build the BOD outlet model and so the efficiency estimated. With this structure is possible proposing a classification model in that every value that showed an efficiency higher than 60% is classified how be able to final disposal (positive class) and for efficiency less than 60% how not be able to final disposal (negative class).

REFERENCES

[1] Morais, J. T. G. (2011) Análise de Componentes Principais Integrada a Redes Neurais Artificiais Para Predição de Matéria Orgânica. [s.l.] Universidade Federal da Bahia.

[2] Brasil. Resolução No 430. [s.l.] Ministério do Meio Ambiente.

[3] Von Sperling, M. (2014) Introdução à qualidade das águas e ao tratamento de esgotos. Edição 4 ed. Belo Horizonte MG: UFMG.

[4] Yu, P. (2019) A Real-time BOD Estimation Method in Wastewater Treatment Process Based on an Optimized Extreme Learning Machine. Applied Sciences, v. 9, n. 3, pp. 523.

[5] Kuhn, M. & Johnson, K. (2013) Applied Predictive Modeling. New York: Springer.

[6] Mayumi O T Santoro, P P Baranauskas J A 2012 How Many Trees in a Random Forest? Lecture notes in computer science v 7376 pp 154 168.

[7] Oliveira-Esquerre, K. P. (2004) Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches. Chemical Engineering Journal, v. 104, n. 1–3, pp. 73–81.

ACKNOWLEDGMENTS

The present work was supported by the “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq), the “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES), the “Programa de Pós-Graduação em Engenharia Industrial” (PEI) and the GAMMA (Growing with Applied Multivariate Analysis) group of the Federal University of Bahia.

SUPPORT: