Brazilian Journal
of Chemical
Engineering

ISSN 0104-6632
Printed in Brazil
www.abeq.org.br/bjche

Vol. 34, No. 01, pp. 53 - 63, January - March, 2017
dx.doi.org/10.1590/0104-6632.20170341s20150475

MACHINE LEARNING TECHNIQUES APPLIED TO LIGNOCELLULOSIC ETHANOL IN SIMULTANEOUS HYDROLYSIS AND FERMENTATION

J. Fischer, V. S. Lopes, S. L. Cardoso, U. Coutinho Filho* and V. L. Cardoso*

Faculty of Chemical Engineering, Uberlândia Federal University, P.O. Box 593, Av. João Naves de Ávila 2121,
Campus Santa Mônica, Bloco 1K, CEP: 38400-902 Uberlândia - MG, Brazil.
E-mail: janaffischer@hotmail.com; vv_lopes@hotmail.com; saulo_lcardoso@hotmail.com
*E-mail: ucfilho@feq.ufu.br; vicelma@ufu.br

(Submitted: July 27, 2015 ; Revised: December 30, 2015 ; Accepted: January 4, 2016)

Abstract - This paper investigates the use of machine learning (ML) techniques to study the effect of different process conditions on ethanol production from lignocellulosic sugarcane bagasse biomass using *S. cerevisiae* in a simultaneous hydrolysis and fermentation (SHF) process. The effects of temperature, enzyme concentration, biomass load, inoculum size and time were investigated using artificial neural networks, a C5.0 classification tree and random forest algorithms. The optimization of ethanol production was also evaluated. The results clearly depict that ML techniques can be used to evaluate the SHF (R^2 between actual and model predictions higher than 0.90, absolute average deviation lower than 8.1% and RMSE lower than 0.80) and predict optimized conditions which are in close agreement with those found experimentally. Optimal conditions were found to be a temperature of 35 °C, an SHF time of 36 h, enzymatic load of 99.8%, inoculum size of 29.5 g/L and bagasse concentration of 24.9%. The ethanol concentration and volumetric productivity for these conditions were 12.1 g/L and 0.336 g/L.h, respectively.

Keywords: Lignocellulosic ethanol; Machine learning; Simultaneous hydrolysis and fermentation; Crude enzyme complex.

INTRODUCTION

One of the most promising methods to obtain renewable energy in an environmentally sustainable way is to produce it from cheap and abundant biomass sources like lignocellulosic materials. Bioethanol production from waste crop and crop residues could potentially surpass 491 GL/year worldwide. Under such circumstances, ethanol production from lignocellulosic biomass is a promising technology, and several techniques have been proposed to reduce the recalcitrance of the lignocellulosic matrix structure, reduce the cost of enzyme production and improve enzymatic hydrolysis and fermentation (Chen *et al*. 2014; Wu *et al*. 2011; Karlsson *et al*. 2014; Prado *et al*. 2014)

Although companies and academia have made a lot of progress, enzymatic hydrolysis remains one of the critical bottle-necks as a result of the large amounts of enzyme required for hydrolysis, the complexity of mass transfer and the large number of chemical reactions with the generation of inhibition products (Khare *et al*. 2015; Goldbeck *et al*. 2013). The combination of hydrolysis and fermentation in a simultaneous process represents one strategy that can lower capital cost, facilitate the recovery of the product and reduce contamination and inhibition (Ohgren *et al*. 2007; Ask *et al*. 2012; Saha *et al*. 2011). There-

*To whom correspondence should be addressed fore, a large number of studies have been conducted to evaluate the effects of different biomasses, solid loading, inhibition and hydrolysis conditions on the feasibility of ethanol production by simultaneous saccharification and fermentation (SSF) (Cuevas *et al*. 2015; Asada *et al*. 2015; Narra *et al*. 2015; Gu *et al*. 2014; Chong *et al*. 2014).

There is great interest in using machine learning (ML) procedures like artificial neural networks (ANNs), classification trees (CTs) and random forests (RFs) in the context of achieving feasible production of ethanol from lignocellulosic biomasss by SSF, but few studies using ANN to describe the reduction in cost of enzyme production and improve the steps of enzymatic hydrolysis and fermentation are available (Vani *et al*. 2015; Das *et al*. 2015; Giordano *et al*. 2013; Gitifar *et al*. 2013), and no study on methodologies other than ANN has yet been reported. Consequently, the aim of this study was to use the ability of ML techniques (ANN, RF and CT) to model the effects of temperature, time, biomass and inoculum size on ethanol fermentation by SSF. The optimization of ethanol production was also evaluated.

MATERIALS AND METHODS

Raw Materials

All the ethanol fermentations were performed using the enzyme complex (Enz) produced in situ by extraction of the enzyme content provided by solid state cultivation (SSC) and exploded sugarcane bagasse (Bag) with a severity factor (SF) of 3.4 donated by the Centro de Tecnologia Canavieira (CTC, Brazil) which contained about 50% water, 30.0% cellulose, 7.3% hemicellulose, 11.2% lignin and 1.5% ash (content analysis performed as described in Browning, 1967). The SF of 3.4 was chosen from a previous study where Bag samples with SFs of 3.4 and 3.8 were tested, and the best result was obtained using the sample with an SF of 3.4 (data not shown here). The Enz was produced using the same Bag (SF of 3.4) and rice bran (RB), as described below. The RB was purchased from Cocal Foods (Uberlândia-MG, Brazil). The raw materials were stored at -18 °C and subsequently milled and sieved through a 1.8 mm mesh prior to their use as samples in the experiments.

Microorganisms, Fermentations and Enzyme Complex

The SSF was performed using *Saccharomyces cerevisiae* Y904 (AB Brasil, Pederneiras-SP, Brazil) and an enzyme complex obtained from SSC using a strain of *Aspergillus niger* reported in a previous study (Fischer *et al*. 2014). The conditions used in SSC, enzyme production and SSF are described in Table 1.

Experimental Strategy and Overview of Proposed ML Methods

To model the effects of process variables (time, load of bagasse, enzyme, temperature and inoculum concentrations) on ethanol production and find the optimized conditions of SSF, a total of 17 experimental runs with different sampling times were performed, and a total of 1560 experimental points expressing the evolution of cells and SSF products were collected. The operational conditions used in the runs are presented in Table 2.

Table 1: SSC, SSF and enzyme complex production.

Process-Step	Description
SSC	*A. niger* cells were produced by submerged fermentation at 30 ± 2 °C, agitated at 150 rpm in a rotatory shaker using Czapec medium composed of (g/L): $NaNO_3$ (2.0), K_2HPO_4 (1.0), $MgSO_4$ (0.5), KCl (0.5), $FeSO_4$ (0.01) and glucose (20.0). After 48 h of submerged fermentation the cells were harvested by centrifugation at 8000 g for 10 min and the cell pellets were washed twice, re-suspended in sterile water and used to start the SSC (1.0×10^8 spores/g of solid medium). The SSC was done in a 0.25 L conical flask reactor at 30 ± 2 °C containing 40 g of solid medium (composed of 40% Bag and 60% RB) and 40 g of water containing the harvest cells.
Enz production	Forty (40) g of solid fermented medium was mixed with 50 mL of 1.0% (w/w) aqueous Tween 80 at 30 °C in a 500 mL closed Duran bottle. The mixture was agitated for 10 min and the extracted slurry was filtered to collect the Enz (liquid fraction).
SSF	The SSF was performed in a 0.25 L conical flask reactor at variable temperatures (30 to 40 °C) in a medium containing K_2HPO_4 (5.0 g/L), $MgSO_4$ (1.0 g/L), NH_4Cl (1.0 g/L), KCl (5.0 g/L), yeast extract (1.0 g/L) and variable concentrations of Bag (15 to 25%) and Enz (60 to 100% vol/vol).

Table 2: SSF operating conditions.

Run	Code	T (°C)	Enz (%)	Inoc (g/L)	Bag (%)
1	A1	40	100	35	25
2	A2	40	100	25	15
3	A3	30	60	25	25
4	A4	40	60	25	25
5	A5	30	100	25	25
6	B1	30	100	35	15
7	B2	40	60	35	15
8	B3	30	60	25	15
9	B4	40	60	25	15
10	B5	40	60	35	25
11	C1	40	100	35	15
12	C2	30	60	25	25
13	C3	30	60	35	15
14	C4	30	100	25	15
15	C5	40	100	25	25
16	D1	30	100	35	25
17	D2	35	80	30	20

ANN Model

The ANN model containing three layers was implemented and used to find optimal conditions employing R software and the library AMORE (http://cran.r-project.org/web/packages/AMORE/) as follows. First, the values of variables and responses were normalized using z-score standardization (calculated for each data set of variables and response by subtracting its mean value and dividing the result by the standard deviation). Second, the data set was categorized into two random subsets: a training data set (2/3 of the original experimental data set) and a test data set (1/3 of the original experimental data set). Third, a total of 500 ANN were tested, varying the number of hidden neurons and transfer functions (purelin, sigmod and tansig) in layers to optimize the ANN for both data sets (training and validation) and achieve a coefficient of determination (R^2) close to 1 and a reduction of the root mean squared error (RMSE) and the absolute average deviation (AAD) calculated according to Equations (2), (3) and (4), respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i^{calc} - Y_i^{\exp}\right)^2}{\sum_{i=1}^{n}\left(Y_i^{calc} - Y_m\right)^2} \tag{2}$$

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{calc} - Y_i^{\exp}\right)^2\right)^{\frac{1}{2}} = \sqrt{MSE} \tag{3}$$

$$AAD = \frac{1}{n}\sum\left|\frac{Y_i^{calc} - Y_i^{calc}}{Y_i^{calc}}\right|, \tag{4}$$

where n is the number of points, Y_i^{calc} is the predicted value, $Y_i^{\exp}$ is the experimental value, Y_m is the average value of all experimental data and MSE is the mean square error. Third, the connection weights of the ANN were used to calculate the effect of features (variables of the process) on ethanol production, as described in Equation (5) (Garson 1991):

$$I_j = \frac{\sum \frac{\left|W_{jm}^{ih}\right|}{\sum\left|W_{km}^{ih}\right|}\cdot\left|W_{mn}^{hO}\right|}{\sum\sum \frac{\left|W_{km}^{ih}\right|}{\sum\left|W_{km}^{ih}\right|}\cdot\left|W_{mn}^{hO}\right|} \tag{5}$$

where I_j is the relative importance of the j^{th} input variable on ethanol concentration, N_i and N_h are the number of input and hidden neurons, Ws are the connection weights, the subscripts *i*, *h* and *O* refer to the input, hidden and output layers, respectively, and the subscripts *k*, *m* and n represent the input, hidden and output neurons, respectively. Fourth, the optimized conditions related to ethanol production were determined using the ANN and an *R* script for ant colony optimization (ACO) written as described in Dorigo *et al.* (1996). The ACO was used with different randomly initiated input variables to secure the solution corresponding to the best multi-objective optimizations. Accordingly, the optimal ethanol concentration for the optimal volumetric ethanol productivity and the optimal ethanol conversion for the lowest time were found. The volumetric productivity was calculated as the ratio between the ethanol concentration and time of fermentation, and the ethanol conversion was defined as described in equation 6 (Naveen *et al.* 2011):

$$\text{Ethanol conversion} = \frac{Et - Et_0}{0.51 \times f \times Dry\_Bag \times 1.11} \tag{6}$$

where Et is the ethanol concentration at time *t*, Et_0 is the initial ethanol concentration, 0.51 is the conversion factor for glucose to ethanol based on the stoichiometry of yeast, *f* is the glucan fraction of dry biomass, *Dry_Bag* is the dry biomass and 1.11 is the conversion factor for glucan to glucose.

Random Forest (RF) Model

RF is a non-parametric ML algorithm derived from a classification and regression tree and performs very well when compared with ANN and other ML methodologies. RF characteristics include robustness to noise, tuning simplicity and ability to handle high dimensional non-linear problems (Breiman, Friedman and Stone 1984; Breiman 2001; Seyedosseini and Tasdizen 2015; Liaw and Wiener 2002). In this work, the use of RF was performed using the RF library for R language (Random Forest) and was used to describe the SSF and predict the influence of variables using the available measure of increase of node purity according to MSE (IncMSE) present in the software. To ensure good predictive performance of the RF, a total of 1000 RFs containing different numbers of trees and variables in each of the branches (parameters of the method) were evaluated. The evaluation of the optimal RF model was conducted using the same division of experimental points (two random data sets, containing 2/3 for training and 1/3 for test) and the same criteria described in the ANN methodology (i.e., reduce RMSE and AAD as much as possible, and obtain an R^2 as close as possible to 1 in both the training and test data sets).

C5.0 Model

C5.0 has become the industry standard for producing decision trees. It is based on the concept of entropy of information for recursively separate observations in branches to construct a tree based on rules that are logically understood (Mistikoglu *et al.* 2015; Lantz 2013). In this work, C5.0 was used as follows: (a) the ethanol concentration was described in three classes: low, if it was below the first quartile, high, if it was equal to or greater than the fourth quartile and medium, if it was between the first and fourth quartile; (b) C5.0 script was written using the default library for R (http://cran.r-project.org/web/packages/C50/) for ranking the variables of the process based on their ability to partition the data and find the rules for the correct classification of ethanol production.

Analytical Methods

The cell concentrations in SSF and those required to begin SSC were determined by counting in a Neubauer chamber and by estimation from the optical density at 600 nm, respectively. The estimation methodology used a correlation determined a priori between the optical density and number of colonies obtained in using a spread plate methodology after 48 h of incubation at 40 °C. The inoculating plates contain Czapek with agar and the same nutrient concentrations described above. The sugars and ethanol concentrations were determined by high performance liquid chromatography (HPLC; Shimadzu LC-20A) equipped with a refractive index detector, a Supelcogel Ca column operated at 80 °C and deionized water (pH 7.0) as the mobile phase at a flow rate of 0.5 mL/min.

RESULTS AND DISCUSSION

Table 3 presents some descriptive statistics of the experimental runs, and Figures 1 and 2 show the results of the sugars and metabolites of fermentation detected during the SSF runs, respectively. The inspection of this table and the figures reveals that the yeast cell growth was found to be low, and the accumulation of arabinose and glycerol was lower when compared with the production of xylose, ethanol and acetic acid. These results suggest active utilization of glucose and that the concentrations of the metabolites and pentoses found are likely to have a significant impact on microorganism viability and ethanol production. Loss of viability of cells during fermentation was not observed (data not presented in figures). According to the literature, acetic acid can inhibit the cell metabolism as a result of an increase in the ATP required for cell maintenance (Mariorella *et al*., 1983; Narendranath *et al*., 2001; Sousa *et al*., 2012), xylose can inhibit the pathway of glucose-phosphorylating enzymes (Fernandez *et al*., 1985), arabinose can positively affect the enzymatic hydrolysis of lignocellulosic biomass by reduction of crystallinity (Fengcheng *et al*., 2013) and glycerol is essential for balancing the redox potential in the absence of oxygen and osmoregulation of the cell (Neivoig *et al*., 1997). The high concentrations of inhibitors found suggest the choice of the configuration of the fermentation in two steps described as separated hydrolysis and fermentation (SHF) as more favourable than one step described as SSF. However, the process operation in SSF or SHF modes is an open question. Although SSF has been widely described as more favourable than SHF because it results in an improved ethanol yield by reducing product inhibition and a reduction in cost as there is no need for separate reactors (Narra *et al*. 2015; Ask *et al*. 2012), both configurations have advantages and disadvantages. According to the literature, the accumulation of glucose that inhibits cellulase activity (Gosh *et al*., 1982; Alfani *et al*., 1990; Ohgren *et al*., 2007) is not present in the latest generation of commercial enzymes, which work equally well in SSF and SHF (Pachos *et al*., 2015). On the other hand, the suboptimal temperatures in SSF are expected to be minimized by using thermotolerant microorganisms (Narra *et al*., 2015; Naveen *et al*., 2011).

Table 3: Descriptive analysis of the runs.

Variables	Code	Min	1st quartile	Mean	Median	3rd quartile	Max
Time (h)	**Time**	**0**	**18**	**29.3**	**34**	**42**	**48**
Glucose (g/L)	Glu	0	0.25	0.48	0.37	0.42	4.1
Arabinose (g/L)	Ara	0.56	0.87	1.18	1.12	1.39	2.56
Xylose (g/L)	Xyl	1.32	6.97	8.74	10.8	10.8	14.1
Acetic acid (g/L)	Acet	1.51	5.37	6.83	7.09	8.51	11.8
Glycerol (g/L)	Gly	0.62	1.39	1.71	1.66	2.05	3.04
Ethanol (g/L)	Et	0	5.86	7.09	7.42	8.43	12.4
$\log_{10}$(cel/mL)	Cel	8.20	8.95	9.06	9.08	9.13	9.38
Temperature (°C)	TC	30	30	35.02	35	40	40
Enzyme (%)	Enz	60	60	80	100	82.2	100
Inoculum (g/L)	Inoc	25	25	30	30	35	35
Bagasse (%)	Bag	15	15	19.5	15.0	25	25

Note: the data represent the average of measurements in duplicate

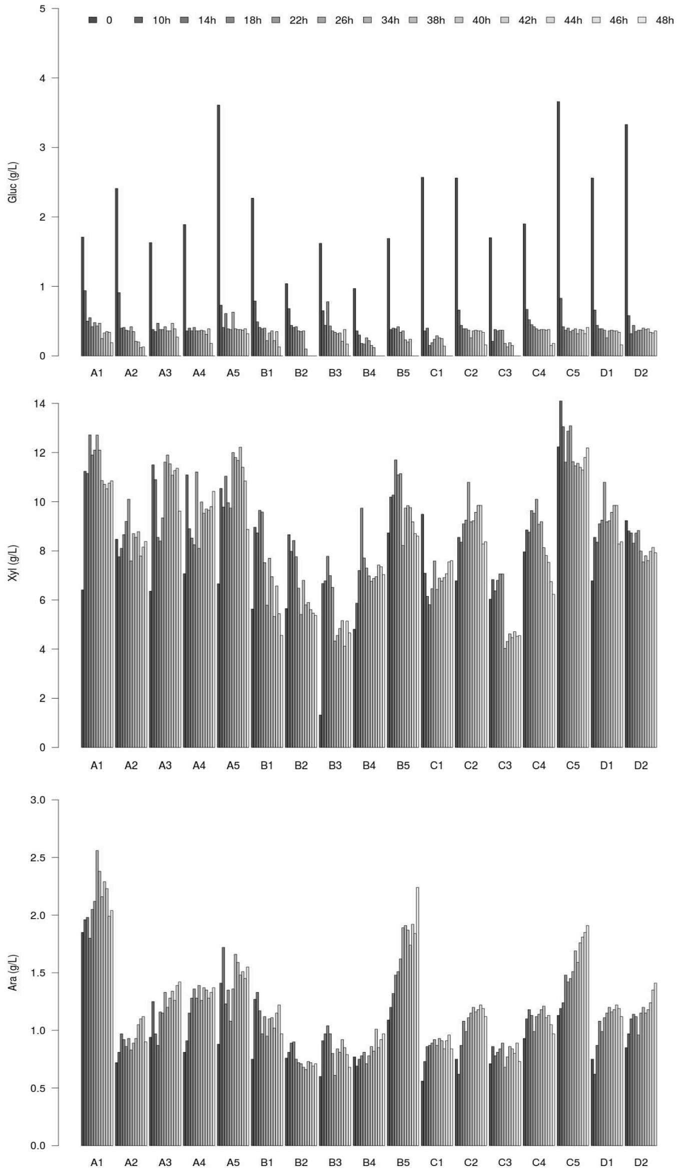

Figure 1: Dynamics of carbohydrates during SSF experiments. Codes of the runs are presented in Table 2 and the data represent the average of measurements in duplicate.

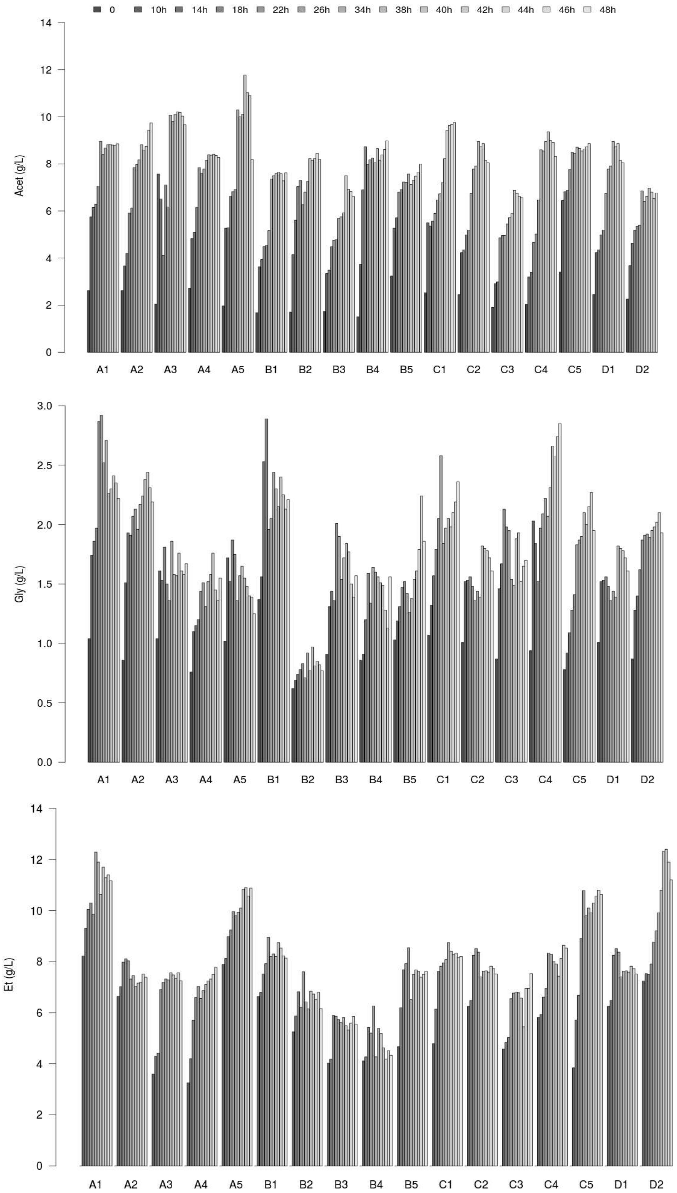

Figure 2: Metabolic products formed during SSF experiments. Codes of the runs are presented in Table 2 and the data represent the average of measurements in duplicate.

Table 4 presents the Pearson correlation coefficients among the variables obtained. The correlations obtained are important to better evaluate the process and select the variables with high predictive power for modelling ethanol production using ML methodologies. A high correlation between the pairs ethanol-glycerol, inoculum-cell concentration, xylose-arabinose, xylose-bagasse and arabinose-bagasse is observed. The bivariate correlations also show that: a) there is not a high correlation between cell concentration and variables distinct from the inoculum; b) ethanol is associated with increased time, cell, xylose, arabinose, acetic acid and glycerol and decreased glucose. According to the correlation coefficients, the variables time, acetic acid, glucose, xylose and arabinose cannot be chosen simultaneously to describe the ML models because they are highly correlated, indicating redundant information. Consequently, the ML models used in this study have time, temperature and the concentrations of enzyme, inoculum and bagasse as dependent variables to describe the ethanol concentration.

Table 5 summarizes the best models found to describe ethanol using ANN and RF, and Figure 3 presents details of the RF adjustment. Based on the low values of RMSE and AAD found, the deviations of the models from the experimental results are satisfactory, and, according to the high values of R^2, the accuracy of the RF and ANN to predict future outcome is also satisfactory. According to the values, it is possible to say that RF and ANN fitted well to the experimental data. In addition, it should be noted that ANN produced lower values of AAD than RF to describe ethanol. For this reason, ANN was selected for additional studies.

Table 4: Correlation matrix in the SSF.

	Time	Glu	Xil	Ara	Acet	Gly	Et	Cel	TC	Enz	Inoc	Bag
Time	1.00											
Glu	-0.64*	1.00										
Xil	-0.02	0.05*	1.00									
Ara	0.24*	-0.11	0.69*	1.00								
Acet	0.83*	-0.59*	0.35*	0.38*	1.00							
Gly	0.44*	-0.37*	0.24*	0.41*	0.42*	1.00						
Et	0.64*	-0.61*	0.42*	0.56*	0.68*	0.65*	1.00					
Cel	-0.05*	0.05*	0.27	0.35*	0.05	0.28*	0.23*	1.00				
TC	0.01	-0.08*	0.07	0.13	0.11	-0.19*	-0.03	-0.37	1.00			
Enz	0.01	0.13*	0.39*	0.35*	0.15*	0.46*	0.34*	0.02	0.12*	1.00		
Inoc	-0.01	-0.04	-0.01	0.17	-0.03	0.13*	0.10	0.61*	0.00	-0.12*	1.00	
Bag	-0.01	0.15*	0.70*	0.67*	0.20*	0.08*	0.30	0.41*	-0.12*	0.02	0.12	1.00

Note: * Statistically significant correlations ($P < 0.05$). Codes of the variables are presented in Table 3.

Table 5: The best ANN and RF using time, temperature and concentrations of bagasse, enzyme and inoculum as dependent variables.

ML	Description	Train data set	Validation data set

ANN	Number of hidden layer = 1
	Number of hidden neurons = 7
	TF = purelin (OL) and tansig (HL)
		$R^2 = 0.92$
		RMSE = 0.68
		AAD = 5.22%
			$R^2 = 0.90$
			RMSE = 0.78
			AAD = 8.04%

RF	Number of trees = 109
	mtry = 3 (nodes in each tree)
		$R^2 = 0.92$
		RMSE = 0.77
		AAD = 8.32%
			$R^2 = 0.91$
			RMSE = 0.87
			AAD = 9.32%

Note: TF transfer function, OL output layer, HL hidden layer

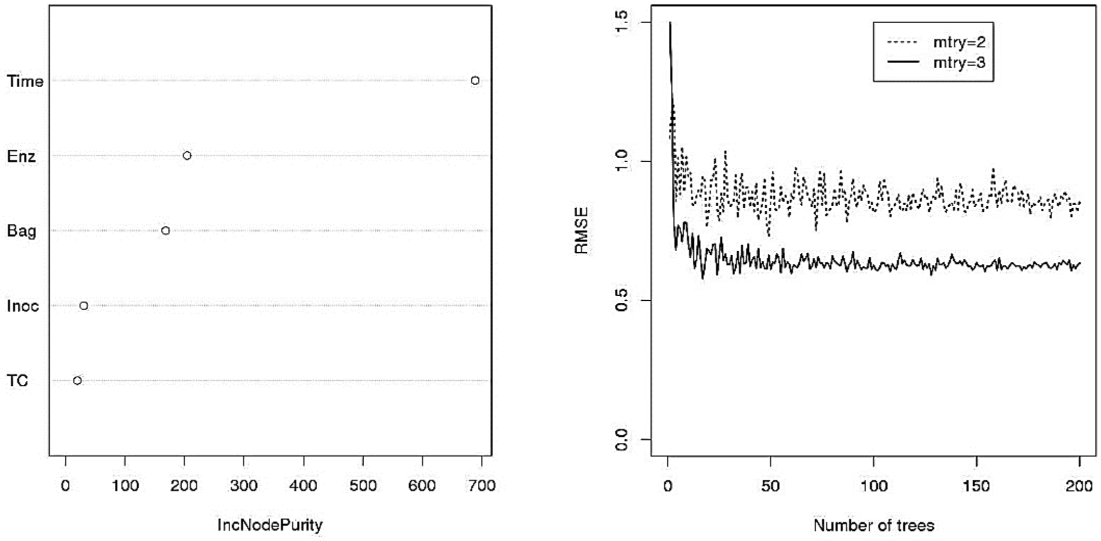

Figure 3: Results of the adjustment of the RF methodology: the importance of effects predicted using RF (right), the worst and best results according to the parameter mtry (left). Codes of the variables are presented in Table 3.

According to the ANN, the importance of time, enzyme, bagasse, inoculum and temperature calculated by Equation 4 were, 50.1, 18.3, 17.7, 8.1 and 5.9%, respectively. The RF prediction of the importance of variables on ethanol production provided by IncMSE (an available measure present in the RF algorithm that represents the increase of MSE when each predictor is replaced in turn by a random noise) is the same observed using the ANN (Figure 3, left). These results indicate that all of the variables used are important to describe the ethanol concentration during SSF. The poor correlation between these input variables and ethanol described by bivariate correlations (Table 3) and the importance of these variables described by ANN and RF suggest that highly nonlinear interactive effects are found in SSF. Complex interactive effects between the input variables are expected in SSF using a high a load of solid. Bellido *et al*. (2011), studying the inhibition effects in ethanol production from wheat straw using *Scheffersomyces stipitis*, found a synergistic inhibition effect between acetic acid and furaldehydes. Pietrzak and Kawa-Rygielska (2015), in a study using a high concentration of solid biomass and saccharification of starch, found lower dynamics of ethanol production caused by the synergistic stressing action of sugars and ethanol.

According to the C5.0 methodology, different combinations of variables are able to yield a high production of ethanol (Table 6), and the importance of time, enzyme, bagasse, inoculum and temperature were 100, 100, 99.1, 90.3 and 68.9%, respectively. The percentages indicate the number of times each variable was used to describe the rules of classification presented in Table 5. Despite the fact that C5.0 is related to a qualitative analysis, the results found are very close to the results obtained using ANN and RF, suggesting that this methodology, which has not been used before in fermentation studies, can be very useful to describe SSF and other kinds of fermentation. The comparison between C5.0, RF and ANN show that: a) all variables tested are important to describe ethanol production; b) the relative importance of bagasse and enzyme are nearly the same in value and rank; and c) the relative importance of time, inoculum and temperature are nearly in the same order as that found using ANN and RF.

The optimum values predicted by ACO-ANN for simultaneous optimization of volumetric productivity and ethanol concentration found a volumetric productivity, an ethanol concentration and a conversion of 0.345 g/L·h, 12.1 g/L and 0.29 g/g, respectively, at the set input conditions of 99.8% enzyme, 35 °C, 29.5 g/L of inoculum, bagasse concentration of 24.9% and 36 h of SSF. The experimental validation under optimized conditions determined that the volumetric productivity and ethanol concentration were 0.336 g/L·h and 12.1 g/L, respectively, which is in close agreement with the ACO-ANN results. In terms of the error of these results, it is important to note that, according to the theory of error propagation, the magnitude of errors in inoculum size, enzymatic loading and bagasse concentration are 0.15 g/L, 0.14% and 1 g/L, respectively. The comparison of these results with the literature results demonstrates that an optimization goal was found since a high concentration of ethanol was obtained at optimized conditions. Das *et al*. (2015), studying ethanol production by different microorganisms (*Scheffersomyces stipitis*, *Candida shehatae* and *Saccharomyces cerevisiae*) using hyacinth as lignocellulosic biomass and a commercial enzyme, found *S. stipitis* to be the best microorganism, with an optimal ethanol concentration of 10.4 g/L (ethanol conversion of 0.104 g/g) after 36 h. Asada *et al*. (2015), using thermotolerant yeast *S. cerevisiae* BA11, commercial enzyme and cedar lignocellulosic biomass, obtained their best results of 9.96 g/L of ethanol (conversion not reported) in a batch process of 24 h and 26.5 g/L (conversion of 0.741 g/g) after 60 h in a fed-batch process using the same yeast and detoxification to reduce inhibition effects. Swain and Khrishnan (2015), studying ethanol production by *S. cerevisiae* and *Candida tropicalis* using commercial enzyme in a SHF (72 h of hydrolysis and 18 h of fermentation) and rice straw, found *C. tropicalis* to be the best microorganism, with an optimal ethanol concentration and conversion of 26.2 g/L and 0.992 g/g, respectively.

Table 6: Decision tree by entropy analysis using C5.0 (error = 12.2% found using C5.0).

Rule	Description of Rules	Ethanol
1	Time (h) > 14 & Enz (%) > 60	High
2	Bag (%) > 20	High
3	Time (h) > 34 & Enz (%) ≤ 60 & Bag (%) 20 & Inoc (g/L)> 30	Med
4	Time (h) > 44 & Enz (%) ≤ 60 & Inoc (g/L) > 30	Med
5	24 < Time (h) ≤ 26 & Enz (%) ≤ 60 & Bag (%) ≤ 20 & Inoc (g/L) > 30	Med
6	Time (h) ≤ 14 & Enz (%) > 60 & Bag (%) ≤ 20	Med
7	Enz (%) ≤ 60 & Inoc (g/L) ≤ 30 & Bag (%) ≤ 20	Small
8	Time (h) ≤ 44 & Enz (%) ≤ 60 & Bag (%) ≤ 20	Small
9	24 < Time (h) ≤ 34 & Enz (%) ≤ 60	Small

Note: Low: Ethanol (g/L) ≤ 1st quartile. High: Ethanol (g/L) ≥ 3rd quartile

The optimum values predicted by ACO-ANN for the optimization of ethanol conversion were 0.45 g/g and 11.5 g/L of ethanol at the set input conditions of 86.0% enzyme, 33.7 °C, 34.2 g/L of inoculum, bagasse concentration of 15.1% and 33.7 h of SSF. This value represents a 1.5-fold increase in ethanol conversion compared to that observed in the first optimization. These results also suggest that the ML model proposed is in good agreement with the expected results, which demonstrate that higher ethanol concentrations can be reached without achieving a very high ethanol conversion (Pachos *et al*. 2015).

Although the potential ML to predict and optimize the lignocellulosic ethanol production was evaluated in a study using traditional microorganisms for both the production of the enzyme complex and ethanol, it could be used directly to optimize other situations. This is important because the production of lignocellulosic ethanol continues to face technical and economic challenges as it seeks to find a cost-effective process with ethanol concentration and volumetric productivities higher than 4% and 1 g/lh, respectively (Petersen *et al*., 2015; Jin *et al*., 2012; Kang *et al*., 2015; Raele *et al*., 2014), which will be possible in the future by several strategies, including fermentations using a single genetic modified yeast available to ferment both C5 and C6 sugars (He *et al*., 2015; Lever, 2015; Baeyens *et al*., 2015) and using yeast strains which combine thermotolerance and higher ethanol productivity (Narra *et al*., 2015; Hasunuma and Kondo, 2012).

CONCLUSIONS

The ML methodologies were successfully able to predict the effects of temperature, bagasse load, inoculum size and enzyme load without requiring the knowledge of the kinetics and the inhibition process. In addition, it was shown that the RF and ANN mathematical models are effective in evaluating the production of ethanol. The temperature of 35 °C, SSF time of 36 h, enzymatic load of 99.8%, inoculum size of 29.5 g/L and bagasse concentration of 24.9% was considered to be the optimum for the simultaneous optimization of volumetric productivity and concentration of ethanol, which were found to be 0.336 g/L·h and 12.1 g/L, respectively.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the National Council for Scientific and Technology Development-Brazil (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES) and the Minas Gerais State Research Foundation-Brazil (FAPEMIG) for their financial support. The authors also gratefully acknowledge the *Centro de Tecnologia Canavieira* (CTC, Brazil).

REFERENCES

Asada, C., Sasaki, C., Takamatsu, T., and Nakamura, Y., Conversion of steam-exploded cedar into ethanol using simultaneous saccharification, fermentation and detoxification process. Bioresour. Technol., 176, 203-209 (2015).

Badal, C., Saha, B. C., Nichols, N. N., Qureshi, N. and Cotta, M. A., Comparison of separate hydrolysis and fermentation and simultaneous saccharification and fermentation processes for ethanol production from wheat straw by recombinant *Escherichia coli* strain FBR5. Appl Microbiol. Biotechnol., 92, 865-874 (2011).

Baeyens, J., Kang, Q., Appels, L., Dewil, R., Lv, R. and Tan, R., Challenges and opportunities in improving the production of bio-ethanol. Prog. Energy Combust. Sci., 47, 60-88 (2015).

Bellido, C., Bolado, S., Coca, M., Lucas, S., González-Benito, G. and García-Cubero, M. T., Effect of inhibitors formed during wheat straw pretreatment on ethanol fermentation by *Scheffersomyces stipites.* Bioresour. Technol., 102, 10868-10874 (2011).

Breiman, L., Friedman, J. and Stone, C. J., Classification and Regression Trees. Chapman & Hall, New York (1984).

Breiman, L., Random Forest. Machine Learn. 45, 5-32 (2001).

Brett, L., Machine learning with R. 1st Ed., Pact Publishing, London (2013).

Browning, B. L., Methods of Wood Chemistry. New York: Interscience, 377 p. (1967).

Chen, H., Li, G. and Li, H., Novel pretreatment of steam explosion associated with ammonium chloride preimpregnation. Bioresour. Technol., 153, 154-159 (2014).

Chong, B., Harrison, M. D. and O'Hara, I. M., Stability of endoglucanases from mesophilic fungus and thermophilic bacterium in acidified polyols. Enzyme Microb. Technol., 61-62, 55-60 (2014).

Cuevas, M., Sanchez, S., Garcia, J. F., Baeza, J., Parra, C. and Freer, J., Enhanced ethanol production by simultaneous saccharification and fermentation of pretreated olive stones. Renew. Energy, 74, 839-847 (2015).

Das, S., Bhattacharya, A., Haldar, S., Ganguly, A., Gu, Sai, Ting, Y. P. and Chatterjee, P. K., Optimization of water hyacinth biomass for bio-ethanol: Comparison between artificial neural network and response surface methodology. Sustainable Mater. Technol., 3, 17-28 (2015).

Dorigo, M., Maniezzo, V. and Colorni, A., Ant system: Optimization by a colony of cooperating agents IEEE Transactions on System. Man and Cybernetics-Part B: Cybernetics, 26, 29-41 (1996).

Fernandez, R., Herrero, P. and Moreno, F., Inhibition and inactivation of glucose-phosphorylating enzymes from saccharomyces cerevisiae by D-xylose. J. Gen. Microbiol., 131, 2705-2709 (1985).

Fischer, J., Lopes, V. S., Queiroz, E. F., Coutinho Filho, U. and Cardoso, V. L., Second generation ethanol production using crude enzyme complex produced by fungi collected in Brazilian Cerrado (Brazilian Savanna). Chem. Eng. Trans., 38, 487-492 (2014).

Garson, G. D., Interpreting neural-network connection weights. AI Expert. 6, 47-55 (1991).

Giordano, P. C., Beccaria, A. J., Héctor C. Goicoechea, H. C. and Olivieri, A. C., Optimization of the hydrolysis of lignocellulosic residues by using radial basis functions modelling and particle swarm optimization. Biochem. Eng. J., 80, 1-9 (2013).

Goldback, R., Ramos, M. M., Pereira, G. A. G. and Maugeri-Filho, F., Cellulase production from a new strain of *Acremonium strictum* isolated from the Brazilian biome using different substrates Bioresour. Technol., 128, 797-803 (2013).

Gitifar, V., Eslamloueyan, R. and Sarshar, M., Experimental study and neural network modelling of sugarcane bagasse pretreatment with H_2SO_4 and O_3 for cellulosic material conversion to sugar. Bioresour. Technol., 148, 47-52 (2013).

Gu, H., Zhang, J., and Bao, J., Inhibitor analysis and adaptative evolution of Saccharomyces cerevisiae for simultaneous saccharification and ethanol fermentation from industrial waste corncob residues. Bioresour. Technol., 157, 6-13 (2014).

Karlsson, H., Börjesson, P., Hansson, P. A. and Ahlgren, S., Ethanol production in biorefineries using lignocellulosic feedstock–GHG performance, energy balance and implications of life cycle calculation methodology. Journal of Cleaner Production, 83, 420-427 (2014)

Hasunuma, T. and Kondo, A., Consolidated bioprocessing and simultaneous saccharification and fermentation of lignocellulose to ethanol with thermotolerant yeast strains. Process Biochem., 47, 1287-1294 (2012).

He, Q., Hemme, C., Jiang, H., He, Z. and Zhou, J., Mechanism of enhanced cellulosic bioethanol fermentation by co-cultivation of *Clostridium* and *Thermoanaerobacter* spp. Bioresour. Technol., 102, 9586-9592 (2011).

Jin, M., Gunawan, C., Balan, V., Yu, X. and Dale, B. E., Continuous SSCF of AFEX pretreated corn stover for enhanced ethanol productivity using commercial enzymes and *Saccharomyces cerevisiae* 424A (LNH-ST). Biotechnol. Bioeng., 110, 1302-1311 (2012).

Kang, E. K., Chung, D. P., Kim, Y., Chung, B. W. and Choi, G. W., High-titer ethanol production from simultaneous saccharification and fermentation using a continuous feeding system. Fuel, 145, 18-24 (2015).

Khare, S. K., Pandey, A. and Larroche, C., Current perspectives in enzymatic saccharification of lignocellulosic biomass. Biochem. Eng. J., 102, 38-44 (2015).

Lever, M., Modelling the energy performance of a farm-scale cellulose to ethanol process with on-site cellulase production and anaerobic digestion. Renew. Energy, 74, 893-902 (2015).

Li, F., Ren, S., Zhang, W., Xu, Z., Xie, G., Chen, Y., Tu, Y., Li, Q., Zhou, S., Li, Y., Tu, F., Liu, L., Wang, W., Jiang, J., Qin, J., Li, S., Li, Q., Jing, H., Zhou, F., Gutterson, N. and Peng, L., Arabinose substitution degree in xylan positively affects lignocellulose enzymatic digestibility after various $NaOH/H_2SO_4$ pretreatments in Miscanthus. Bioresour. Technol., 130, 629-637 (2013).

Liaw, A. and Wiener, M., Classification and regression by random forest. R. News, 2, 18-22 (2002).

Maiorella, B., Blanch, H. W. and Wilke, C. R., By-product inhibition effects on ethanolic fermentation by *Saccharomyces cerevisiae*. Biotech Bioeng., 25, 103-121 (1983).

Mistikoglu, G., Gerek, I. H., Erdis, E., Usmen, P. E. M., Cakan, H. and Kazan, E. E., Decision tree analysis of construction fall accidents involving roofers. Expert Syst. Appl., 42, 2256-2263 (2015).

Narra, M., James, J. P. and Balasubramanian, V., Simultaneous saccharification and fermentation of delignified lignocellulosic biomass at high solid loadings by a newly isolated thermotolerant *Kluyveromyces* sp. for ethanol production. Bioresour. Technol., 179, 331-338 (2015).

Narendranath, N. V., Thomas, K. C. and Ingledew, W. M., Effects of acetic acid and lactic acid on the growth of *Saccharomyces cerevisiae* in a minimal medium. J. Ind. Microbiol. Biotechnol., 26, 171-177 (2001).

Naveen, K. P., Hasan, K. A., Mark, R. W., Danielle, D. B. and Ibrahim, M. B., Simultaneous saccharification and and fermentation of Kanlow switchgrass by thermotolerant *Kluveromyces marxianus* IMB3: The effect of enzyme loading, temperature and higher solids. Bioresour. Technol., 102, 10618-10624 (2011).

Ohgren, K., Bura, R., Lesnicki, G., Saddler, J. and Zacchi, G. A., Comparison between simultaneous saccharification and fermentation and separate hydrolysis and fermentation using steam-pretreated corn stover. Process Biochem., 42, 834-839 (2007).

Pachos, T., Xiros, C. and Christakopoulos, P., Simultaneous saccharification and fermentation by cocultures of *Fusarium oxysporum* and *Saccharomyces cerevisiae* enhances ethanol production from liquefied wheat straw at high solid content. Ind Crops Prod., 76, 793-802 (2015).

Petersen, A. M., Melamu, R., Konoetze, J. H. and Gorgens, J. F., Comparison of second-generation process for the conversion of sugarcane bagasse to liquid biofuels in terms of energy efficiency, pinch point analysis and life cycle analysis. Energy Convers. Manage., 91, 292-301 (2015).

Pietrzak, W. and Kawa-Rygielska, J., Simultaneous saccharification and ethanol fermentation of waste wheat–rye bread at very high solids loading: Effect of enzymatic liquefaction conditions. Fuel, 147, 236-242 (2015).

Raele, R., Boaventura, J. M. G., Fischmann, A. A. and Sarturi, G., Scenarios for the second generation ethanol in Brazil. Technological Forecasting & Social Change, 87, 205-223 (2014).

Swain, M. R. and Krishnan, C., Improved conversion of rice straw to ethanol and xylitol by combination of moderate temperature ammonia pretreatment and sequential fermentation using *Candida tropicalis*. Ind Crops Prod., 77, 1039-1046 (2015).

Seyedhosseini, M. and Tasdizen, T., Disjunctive normal random forest. Pattern Recognit., 48, 976-983 (2015).

Sousa, M. J., Ludovico, P., Rodrigues, F., Leão, C. and Côrte-Real, M., Stress and Cell Death in Yeast Induced by Acetic Acid. Cell Homeostasis and Stress Response, Chapter 4, Edited by Paula Bubulya, Croatia (2012).

Vani, S., Sukumaran, R. and Savithri, S., Prediction of sugar yields during hydrolysis of lignocellulosic biomass using artificial neural network modelling Bioresour. Technol., 188, 128-135 (2015).

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranov, F., Leo, G. A. D. and Torriceli, P., Application of a Random Forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice lagoon. Italy. Ecol. Modell., 222, 1471-1478 (2011).

Wu, H., Mora-Pale, M., Miao, J., Doherty, T. V., Linhardt, R. J. and Dordick, J. S., Facile pretreatment of lignocelluosic biomass at high loadings in room temperature ionic liquids. Biotechnol. Bioeng., 108, 12, 2865-2875 (2011).