J Control Autom Electr Syst (2017) 28:203–216
DOI 10.1007/s40313-016-0291-x

Predicting Ethanol Concentration in Industrial Sugarcane Fermentation Based on Knowledge Discovery in Databases

Márcio José da Cunha[1] · **Glauco A. P. Caurin**[2]

Received: 3 May 2016 / Revised: 17 August 2016 / Accepted: 13 November 2016 / Published online: 28 November 2016

Abstract At the present time, the amount of data stored in the sugar and alcohol industries is considered extensive and continuous. In the production of sugar and alcohol, stored information is not always analyzed. This is due to the amount of data, the diversity of sectors in the production process, along with the difficulty in knowing whether such data can be considered valid for any kind of analysis. This work proposes the use of the *Knowledge Discovery in Databases (KDD)* as an alternative tool for applying data from manufacturing process pertinent to the sugar and alcohol industries. The experiments were conducted with real data obtained from fermentation process during the harvest period. The contribution of this work is the identification of a KDD based on a knowledge structure, which can be used for prediction and simulation activities from the sugar and alcohol production process.

Keywords KDD · Sugar and alcohol production · Fermentation process · Process optimization

1 Introduction

The biofuel industries have shown interest in adapting their production processes to meet increasing requirements of economic efficiency, as well as the need to make the most sustainable processes. This adaptation occurs mainly through process automation, resulting in an increase in the number of sensors, actuators and equipment ready at all stages of production, and correspondingly increase the data generated in the process. The gain efficiency through data analysis is not a trivial task. In most cases, the data are stored in dedicated repository (data warehouse) for post-analysis (Lydon 2015). Generally, this database has many records in its tables; however, not always are such records considered useful or provide some kind of knowledge or information relevant to the process. One way to obtain such knowledge is through mechanisms that make the activity regarding knowledge discovery both simultaneous and automatic. In this context, methods using the Knowledge Discovery process in databases (KDD) for data analysis is produced in an intelligent and automated mode (Choudharya et al. 2009). Thus, various tools and methods have been proposed in order to extract some information concerning such amounts of data.

For the sugar and alcohol sector, research is focused on production improvements, through the construction of computational decision-making and simulation tools in order to optimize the production planning, sugarcane crop and manufacturing process for sugar and alcohol. Such surveys do not use KDD.

This work carried out a research investigation to an ethanol production plant, and an immediate observation made was the fact that most of the data are not used to optimize the production process, and such data are simply eliminated. The rationale behind this practice may be attributed to a lack of data storage and analysis culture in the sugar and ethanol industry. In this context, the development of a KDD method represents the main contribution of this work. The KDD method provides a production forecast for a sugar and alcohol production plant. An approximate characterization of the production is achieved using data from historical oper-

✉ Márcio José da Cunha
mjcunha@ufu.br

Glauco A. P. Caurin
gcaurin@sc.usp.br

[1] Faculdade de Engenharia Elétrica, Universidade Federal de Uberlândia, Uberlândia, Brazil

[2] Dep. de Engenharia Mecânica - EESC, Universidade de São Paulo, São Paulo, Brazil

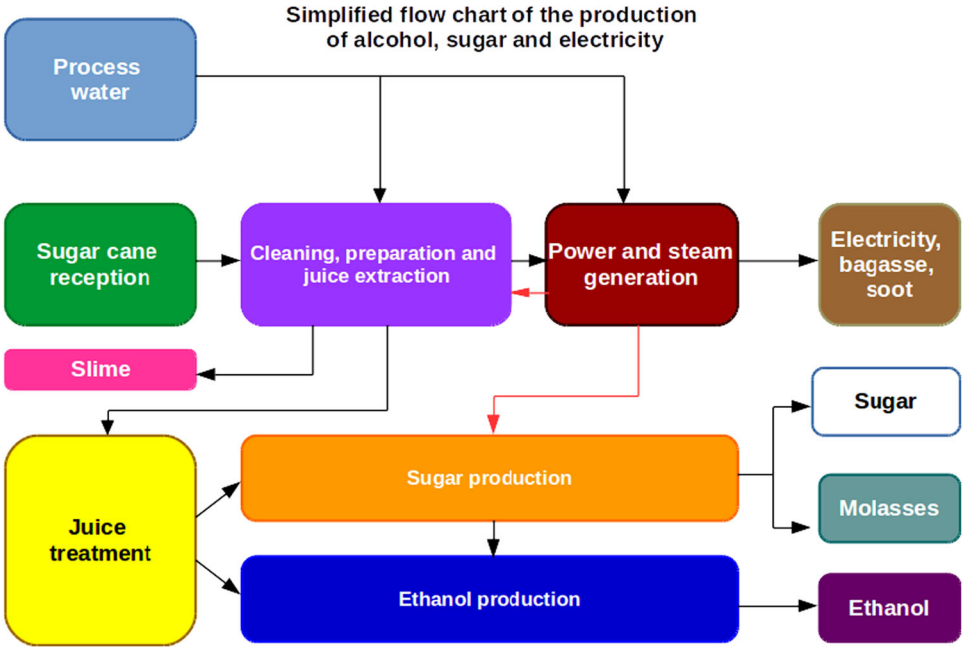

Fig. 1 Product flow within a manufacturing plant alcohol

ation and production. The work also includes an analysis of how feasible it would be to apply a KDD process into daily production tasks. We are especially interested in systemic modeling and in the prediction future behavior of such process.

Among the different processes that compose the ethanol production, this work adopted on the fermentation process as an implementation challenge and example, due to its complexity, variability, lack of comprehensive and well-accepted models and tools for prediction and decision-making. Other processes that may be subject to future investigation include extraction, destination and evaporation.

Preliminary research have been published in order to check the acceptance of the research, along with the improvements that could be applied (Cunha et al. 2012a, b). It is important to note that in this article were implemented improvements which deal with the processing of data and new statistical analysis of the results.

Following this introduction, Sect. 2 will present a compact view of the sugar and ethanol production as well as explanation of the KDD basic concepts. This section gives special attention to data acquisition and preprocessing, completing the set of methods adopted in this work. In Sect. 3, computer simulations experiments are described followed by conclusion in Sect. 4.

2 Materials and Methods

This section deals with the description of the materials and methods used in this article, the industrial process for production of ethanol and the KDD process. In Sect. 2.1, the production flow is described, along with the reasons for having chosen such a process. Section 2.2 describes the KDD process, the processing steps, data mining and the model validation, as well as a state of the art of its application in industry and related work.

2.1 Industrial Ethanol Production Process

Ethanol production is an agro-industrial activity. When restricting our studies to only the industrial part of the ethanol manufacturing process, the existence of different production strategies is observed, these being generated from different countries or production regions. In summary, the industrial ethanol production part comprises the following steps: receiving and washing cane, cane preparation, juice extraction, broth transport, the fermentation with the introduction of yeast, filtration and centrifugation, along with distillation. For a detailed analysis of each stage, it is recommended that the interested individual reads (Amorim 2005).

The diagram in Fig. 1 illustrates the flow of products at each of these steps. Improvements are developed for each of these areas incessantly, in order to increase productivity, reduce losses and search for new and more efficient solutions. The introduction of new sensors, actuators, controllers and fieldbus technology has enabled the automation of virtually all this structure in order to offer a very broad field of the production activities involved.

From the point of view of control and process automation, the fermentation process is still a great challenge, not only technologically, but also scientifically. In this article, we

will focus attention on the pattern known as fed-batch, the most adopted strategy in Brazil for the production of ethanol from sugarcane. This mixture must have fermentation yeast added as the vat components are being added. It is a very productive method as the yeast is at a lower risk of becoming inactive, compared to a single batch process. For a more complete description of the different forms of production, refer to Amorim (2005).

When we are dealing with fermentation and process analysis inside an industrial vat, even the process of modeling is an open problem. First, it is important to emphasize that this study has its emphasis placed on real industrial processes and not systems in a small-scale laboratory. Several reasons help explain the complexity of this challenge:

- The composition is not pure and varies greatly due to the weather during harvest (dry or rainy season), the transport distance and time spent in transporting cane to the industry. Crop residues such as land, straw, more increase this variation
- The Brazilian production standard with centrifuging of wine post-fermentation and reuse of yeast by introducing unique features of evolution into the ecosystem within the fermenter, the domain changes and survival over the period of a production crop, which has its duration set at approximately 8 months
- The greater or lesser presence of microorganisms that compete or cooperate with the sugar in the yeast biomass, CO_2 and alcohol
- The variation of the external temperature
- Constructive geometry and dimensions of the fermentation vat
- Sugar feedback policy
- The fermentation time
- Strategy of adding yeast
- The recent techniques of crop irrigation with treated vinasse creates a complex dynamic feedback between agriculture and industry, involving microorganisms previous harvests
- Difficulty in measuring
- Difficult to measure reliably and accurately the fundamental quantities such as the effective volume of vinasse inside the vat
- Difficulty of measuring expeditiously levels of alcohol, sugars and other analytically

This whole scenario creates a unique multidisciplinary opportunity for scientific cooperation in different areas, in the search for solutions that enhance the understanding of this short-term and long-term process for creating new forms of control that eliminate the ad hoc feature control processes, performance evaluation and consumption of raw materials that currently dominate the industry.

2.2 The KDD Process

The KDD process is described through a sequence of steps, interdependent, applied to find new patterns of knowledge in a database, previously unknown (Fayyad and Shapiro 1996). The knowledge is evident when new patterns identified are used in further analysis (Donauera et al. 2015). Figure 2 shows the KDD process, where the dashed lines indicate the KDD interaction. As observed in the Fig. 2, if it is necessary to run a step again, it can be executed, independent of its execution order. This operation can be performed until the pattern, identified by KDD, is considered as a valid standard.

2.2.1 The Identification of Patterns by KDD

The description of the KDD process is defined as:

- **Data selection**: The process starts from the choice of the dataset that will be analyzed by the KDD. After this choice, a dedicated database is created (data warehouse). This dedicated database is created for the data at the plant containing information related to different sectors, such as the administrative sector, sales and others
- **Preprocessing**: This step is applied with the aim of detecting and eliminating possible noise found in the data, outliers and the values that have records with zero information. These irregularities in the data can occur due to several factors, such as operational failures relating to data manipulation operations, physical failures, such as power outages at the time the data was being requested. According to Fayyad and Shapiro (1996), well-performed treatment and preprocessing reduce processing costs in future steps, such as data mining
- **Transformation**: This step is dedicated to the treatment of data so that they are suitable for the processing of the KDD core, which is the data mining step
- **Data Mining**: This step uses mining techniques, such as neural networks, decision trees and genetic algorithms, to identify patterns that describe the behavior of a process under analysis
- **Validation**: This step is based on mathematical and statistical criteria and verifies whether the pattern discovered by KDD is assessed and validated.

2.2.2 KDD Applied to Industrial Process and Related Work

The KDD application is indicated for various industry segment applications, for example, segments that target quality and production control, engineering applications and equipment maintenance.

Some processes involving quality improvement and production require data collection and analysis to solve problems

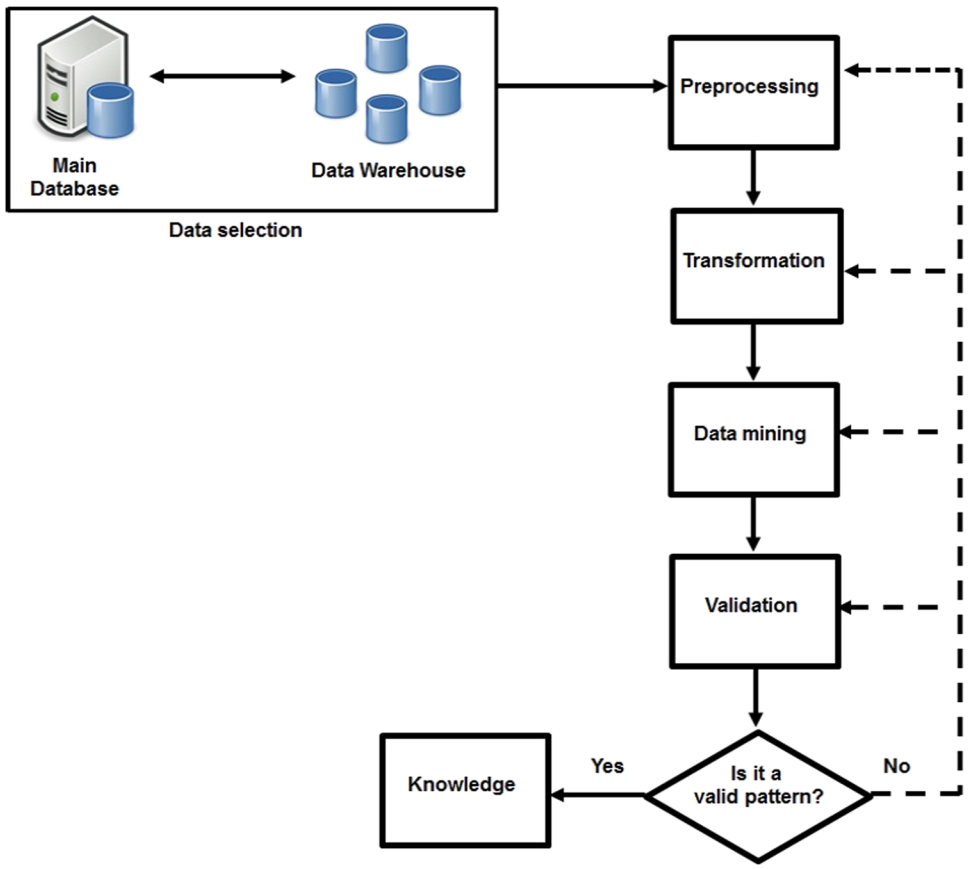

Fig. 2 The KDD process adopted by Fayyad and Shapiro (1996)

in the manufacturing of products/services. Generally, the manufacturing process involves several input and output variables, which produce the modeling and optimization actions (Koksal et al. 2011). To resolve this problem, the KDD process can be used with problems involving multivariable processes, and be applied in various stages of the process cycle of a product (Bingru et al. 2009). In related searches for this kind of application, it was observed that KDD is used as an auxiliary tool for standardizing the quality and production of products preventing, for example, additional production costs (Harding et al. 2006; Browne et al. 2006).

Still further into the application of quality and production control, surveys were identified related to fault diagnosis, analysis of defects inherent to the production processes, identification of functional parameters and forecast production, production quality (Donauera et al. 2015; Choudharya et al. 2009; Koksal et al. 2011).

In the segment of engineering applications, the KDD is used in the creation of new computational tools used by engineers and technicians. Such tools have in their design the prior knowledge of the projects developed by the engineering sector and are constantly updated from new projects that run. This produces agility, reliability and lower hours of engineering costs (Harding et al. 2006).

There are related searches for specific application areas, such as in power generation industries. In one of the studies found, KDD is applied to the identification of patterns that are used in pooling analysis of electrical networks, and also in situations where monitoring of power generation conditions is needed (McDonald and Steele 2006).

In sugar and alcohol environments, the research is focused on improvements in the production of sugar and alcohol, proposing the construction of computational decision-making and simulation tools aimed at optimizing the production planning, sugarcane crop and in the manufacturing process of sugar and alcohol. These surveys do not use KDD.

In Agudelo (2012), a sensor software used for controlling fermentation processes was developed, using information relating to biomass concentration, substrate and other secondary measures (turbidity, pH, CO_2, flow). This model was designed using a hybrid neural responsible for describing the fermentation kinetics.

In other papers, the use of bagasse from sugarcane surplus was evaluated for the production of electricity, or used in ethanol production, with possible applications in the conventional sugar and alcohol production process (Albarelli 2013). In Batista (2008), a computational simulation tool for the procedures used in beverage production processes distilled

drinks in continuous processes is highlighted. Composition profiles were surveyed such as, temperature, pressure, flow and collecting information for creating equipment used in the distillation process. From this tool, the new beverage production patterns were found, with the goal being to have a better quality standard manufacturing process of distilled drinks.

In Dias (2008), simulation tools were developed that analyze production processes, from information obtained from the sugar cane harvest and processed bagasse, in order to analyze the energy consumption at each part of the plant process. Some points were investigated for possible improvements that could be deployed. For this study, the authors considered both the conventional process of producing ethanol from sugarcane, as the bagasse hydrolysis process. In ethanol simulation situations, Marquini et al. (2007) purposed an industrial system of distillation columns for the production of hydrous ethanol fuel, where they optimized the steam consumption that was used by the system, through a binary mixture of ethanol-water. In Decloux and Coustel (2005), simulations were performed by way of a speaker system for producing neutral alcohol, which was considered a wine containing ethanol, water and four contaminants. This mixture increases the complexity of the simulation and the speaker system. In this work, the authors simulated the production of an intermediate alcohol, and also simulated the steps of purification, by the addition of three columns used in the production of neutral alcohol.

In Batista (2012), a simulation tool responsible for hydrous and neutral alcohol production systems was developed in order to optimize the alcohol and cachaça manufacturing process in a continuous system.

In other works, research was developed that identified models that represent the dynamic behavior of processes. In Bergamasco (2003), a computational tool based on a mathematical model, which used that information concerning nitrogen management used in fertilizer applied to sugarcane crops, was developed. These models simulate different scenarios in which the fertilizer could be applied, in order to be have a maximum crop yield, improving the allocation of resources and varieties of inputs. In Hahn (1994), a simulation system that improved daily decisions making process was developed, which was related to the operational planning of transport of raw materials for the sugar and alcohol industries.

Other research is inserted into the decision-making process in planning cases related to the harvest (Bocca et al. 2015) earnings forecast of agricultural systems, crop planning (Grunow et al. 2007; Higgins 2002) and in planning the harvest schedule. A well-executed harvest plan optimizes the supply of raw materials, as well as other benefits such as increasing the amount of sugarcane available for crushing.

2.3 Data Acquisition

Initially, an analysis was performed on how to access the information contained in the fermentation process. What was found is that, currently, the ethanol production plants have a communication structure responsible for providing the information in the process for the various sectors of the industry. In the process, there is a variety of field devices (sensors, actuators and positioners) interconnected by a fieldbus. In most Brazilian mills, the network protocol used is the Profibus and its extensions (Profibus DP, Profibus PA, PROFINET) (Profibus 2015).

In the above network protocol, there is a layer of software responsible for providing equipment with data, for computer applications of existing control in the plant (supervisory systems, historians and setters). This standard is the OPC Classic (*OLE for Process Control Classic*) (OPC 2015). The OPC has in its design, the client–server information exchange architecture.

Access and storage of processed information occurs at this point of the structure. By means of an ODBC (*Open DataBase Connectivity*) connection provided by the supervisory system, data is stored in a dedicated database (data warehouse) of the fermentation process. The schematic of the data acquisition process used in this work is shown in Fig. 3.

The acquisition module, that was developed in C#, is responsible for acquiring information used in this work. The information was obtained from an industrial alcoholic fermentation tank of a sugar and alcohol industry, under typical operating conditions. Traditionally, the tank information was stored in a central database, containing operational information of the ethanol manufacturing process. The data acquisition module retrieves the data and stores it in an external database belonging to the central bank of the plant. The DBMS (*Database Management System*) used in this work was the SQL Server 2012 Community Edition. Each variable used in KDD was recorded in a table of the data warehouse of the fermentation process with the information regarding the time of purchase and the present value. The example of the structure of wort temperature table created to data warehouse is shown in Table 1. In the data warehouse, the database was divided according to their acquisition year.

Traditionally, the harvest period for most of the sugar and alcohol production plants in Brazil begins between the months of March and April, and extends into the middle of October and November. This period is related to the natural cultivation of sugarcane. In this period, however, there may be situations in which the production process can be paralyzed, due to the thunderstorms and other related weather conditions. An example of these conditions are prolonged seasonal rains, which cause the sugarcane cutting areas to

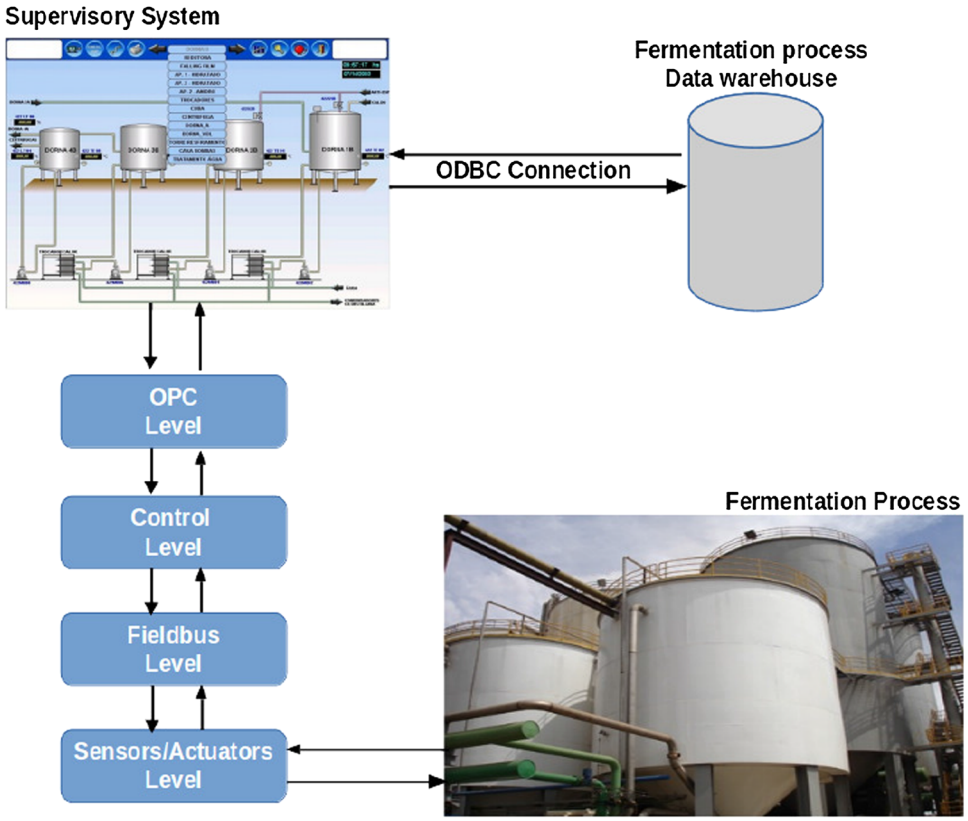

Fig. 3 Communication architecture of a typical plant of ethanol production

become inaccessible to trucks responsible for transporting sugar cane to the sugarcane industry. Another factor that can cause stoppages is technical issues, which appear generally when technical adjustments need to be made, or that there is a point status of failure, be it human or generated by equipment. At end of the harvest, the sugar and alcohol goes into a period of preventive maintenance, which ensures that the plant has a new production cycle (Amorim 2005). Considering the information for the period of the season and the fact that the plants take on average 2 months to be considered ready for operation, in this work, information regarding the harvest period of a plant located in Brazil, during the months of May–September, for the production during the years 2008 and 2009.

2.4 Preprocessing

The preprocessing data module is responsible for performing the necessary treatment of the fermentation process data in order to minimize the influence of possible invalid data (with noise, zero data) in the subsequent data mining step. This module was developed in MATLAB and is divided into three parts: data filtering, data interpolation and data normalization.

Data Filtering

In this step, the data are subjected to the filtering stage, where only the values that are within a range of + or −10% of the average value are considered valid. However, following this criterion, the values that were outside this range, for example, those values well above the range, would be discarded. In order to evaluate the contribution of these values in the description of the fermentation process, the following rule was adopted: The values over 10% of the average have their value changed to the upper limit operation, and those that are 10% below the average have their value changed to the lower limit.

Data Transformation

In the data transformation step, modules were developed that are responsible for the scale of variation variables values across a specific operating range and filter them, so as to obtain a better computational performance during the data mining step (Fayyad and Shapiro 1996). Normalization follows the following equation:

$$y_i = y_{\min}(x_i - x_{\max}) + y_{\max}\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad i = 1 \ldots N \tag{1}$$

where

- N: size of the dataset
- x: the dataset to be normalized
- $x_{\max}$: the maximum dataset value to be normalized
- $x_{\min}$: the minimum dataset value to be normalized
- y: the desired value to be normalized
- $y_{\max}$: the maximum desired value to be normalized
- $y_{\min}$: the minimum desired value to be normalized

This is can be more observed in Figs. 9 and 10. Figure 9 shows the raw data, where it is possible to see some anomalies, like noise. After execution of the module, the data were normalized to desired values and filtered, as shown in Fig. 10. The desired values can be configured in a holistic manner, i.e., the desired values can be configured according to the situation and analysis.

Data Interpolation

The tables in the data warehouse fermentation process have different amounts of stored records, due to the behavioral characteristics of the process variables, so that each variable sampling rate is different from each other. Thus, it is necessary to possess an equal amount of records without the behavioral properties of variables being affected. To meet this requirement, this work applied linear interpolation data rules, so that all data had the same amount of points (Stephens 1998). The Fig. 11 shows the wort temperature filtered and interpolated.

2.4.1 Data Mining

The choice of data mining tool following criteria relating to the operational characteristics of the fermentation process, for example, the fermentation process has characteristic of being a multivariable system and being a nonlinear system, as well as having some kind of application in simulation and prediction cases. The chosen data mining tool is the NARX neural network. This type of network is based on an important class of discrete-time and nonlinear system, the NARX model (Nonlinear autoregressive with exogenous inputs). This model is described by Eq. 2:

$$\begin{aligned} y(n+1) = {} & f(y(t-1), y(t-2), \ldots, y(t-n_y), \\ & u(t-1), u(t-2), \ldots, u(t-n_u)) \end{aligned} \tag{2}$$

where $u(t)$ and $y(t)$ are the experimental data and input output system, and is responsible for representing the amount of necessary information (memory) for processing.

The function f is a nonlinear function, and mapping of their nonlinearity is unknown. When this mapping can be represented by a network type structure multilayer perceptron (MLP), the resulting network structure is a NARX neural network. The network training can be conducted in two ways, the serial-parallel and parallel. In the serial-parallel, the output covariates are formed only by collected values of the real system, according to Eq. 3:

$$\begin{aligned} \hat{y}(n+1) = {} & \hat{f}\left[y_{\text{sp}}(n); u(n)\right] \\ & \rightarrow \hat{f}[y(t-1), y(t-2), \ldots, y(t-n_y), \\ & \quad u(t-1), u(t-2), \ldots, u(t-n_u)] \end{aligned} \tag{3}$$

In the parallel mode, the generated outputs by the network are supplied and included in the output vector of the regressor, as shown in Eq. 4:

$$\begin{aligned} \hat{y}(n+1) = {} & \hat{f}\left[y_p(n); u(n)\right] \\ & \rightarrow \hat{f}[\hat{y}(t-1), \hat{y}(t-2), \cdots, \hat{y}(t-n_y), \\ & \quad u(t-1), u(t-2), \cdots, u(t-n_u)] \end{aligned} \tag{4}$$

Thus, with the NARX neural network, one can predict situations where the experimental data are used at the end of the process for calculating subsequent values of interaction, and can also be simulated situations where the experimental data are used in a specific set of interactions, and the new calculated interactions are used for the calculation process of subsequent values (Menezes and Barreto 2008; Ljung 2002).

According to Menezes and Barreto (2008), the NARX neural network has many applications in real problems, such as water treatment plants, investments in oil refineries and the prediction of time series.

2.4.2 Model Validation

To evaluate and validate the performance of the KDD model, the following statistical criteria were calculated (Stephens 1998; Fayyad and Shapiro 1996; Han and Kamber 2006):

- **Coefficient of determination** (R^2): This criteria indicates how much the observed output can be explained by the output generated by the KDD model, according to Eq. 5:

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \bar{P})^2}{\sum_{i=1}^{n}(P_i - \bar{P})^2 + \sum_{i=1}^{n}(O_i - P_i)^2} \tag{5}$$

where

  - P_i: the ethanol concentration degree predicted by the KDD model

- $\bar{P}$: square error calculated by the prediction and the observed ethanol degree
- O_i: the observed ethanol degree
- n: number of samples

– **Coefficient of correlation** (R): This criteria quantifies the model global description. For a high value of R, there is a significant correlation between the observed values and the values generated by the KDD model. The calculation is performed according to Eq. 6:

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})}{(\sigma_o)(\sigma_p)} \quad (6)$$

where

- P_i: the ethanol concentration degree predicted by the KDD model
- $\bar{P}$: square error calculated by the prediction and the observed ethanol degree
- O_i: the observed ethanol degree
- $\bar{O}$: the average value of the concentration ethanol degree observed
- σ_o: standard deviation of the observed ethanol concentration degree
- σ_p: standard deviation of the predicted ethanol concentration degree
- n: number of samples

– **Mean square error** (MSE): This value is the sum of the differences between the values generated by the model and the observed values being weighed by the number of terms as shown in Eq. 7:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2 \quad (\%) \quad (7)$$

where

- P_i: the ethanol concentration degree predicted by the KDD model
- O_i: the observed ethanol degree n: number of samples

– **Mean percentage error** (MPE): This value represents the percentage of error between the sum of the differences between the value generated by KDD and the observed and weighed value by the number of terms. Equation 8 shows how the MPE is calculated:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2 \times 100 \quad (8)$$

where

Table 1 Structure of wort temperature table corresponding to 2008

Variable name	Data type	Is primary key
Date and time	Datetime	Yes
Process value	Float	No

- P_i: the ethanol concentration degree predicted by the KDD model
- O_i: the observed ethanol degree n: number of samples

The validation routines and the indicatives shown in Eqs. 5, 6, 7, and 8 were implemented in Matlab.

3 Experimental Results

This section shows the experimental results obtained in this paper. Here, some analyses of the KDD application are performed on the sugar and alcohol production, that justify its application.

3.1 Preprocessing Data

Considering the information for the period of the season, cited in Sect. 2.3. Data were classified in input and output quantities (variables), according to the mapping realized during the fermentation process. Quantities under consideration in this work are: total reducing sugar concentration (TRS) [%(w/w)], pH, total soluble solids concentration [°Brix], alcohol degree by volume [%], yeast growth rate [%], yeast viability [%], cane must temperature [°C] and tank level [%]. The output variable is the behavior of the ethanol concentration [%]. Figures 4 and 5 show the typical behavior of these variables under regular fermentation conditions in the month of May.

In a fermentation process, yeasts extract energy from sugar and releases ethanol and CO^2. Therefore, sugar concentration measure (TRS) is commonly used to evaluate the must quality for ethanol production. Usually in industrial processes, sugar concentration is given in percentage of weight (mass) of all available sugars (sucrose, fructose and glucose) per total weight (mass) of the solution [% (w/w)]. In the literature, it is usual to find values ranging from 18 to 22%. In Table 2, real data usage for this work is observed, which may assume values in the interval between 7 and 27 %. Total soluble solids concentration measured in degrees Brix is an alternative form of assessing sugar concentration in the cane must. The °Brix corresponds to the percentage in mass of sucrose in a specific solution. The °Brix is a standard well-established measure in the sugar and ethanol industrial.

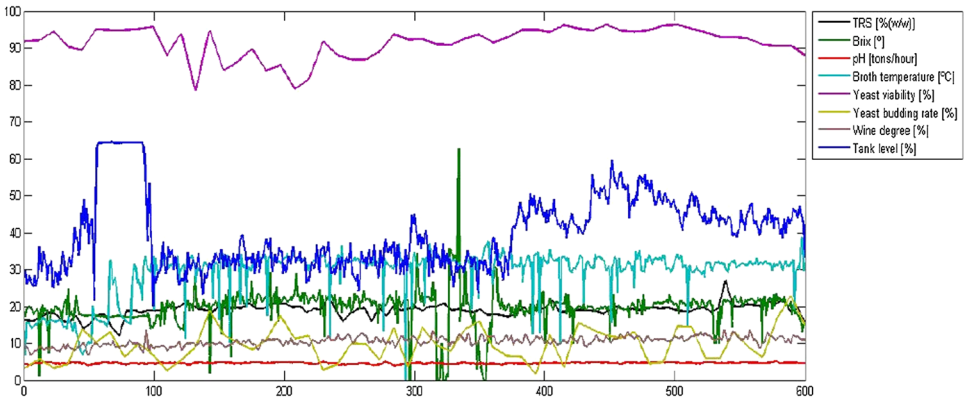

Fig. 4 Representative behavior of the chosen input variables adopted to describe the fermentation process. For visualization purposes, the data were interpolated but not filtered, which explains the appearance of some outliers, particularly in the temperature values. The data correspond to May 2008. The axis x corresponds to the sampled interpolated points and the axis y corresponds to the behavior of the variables

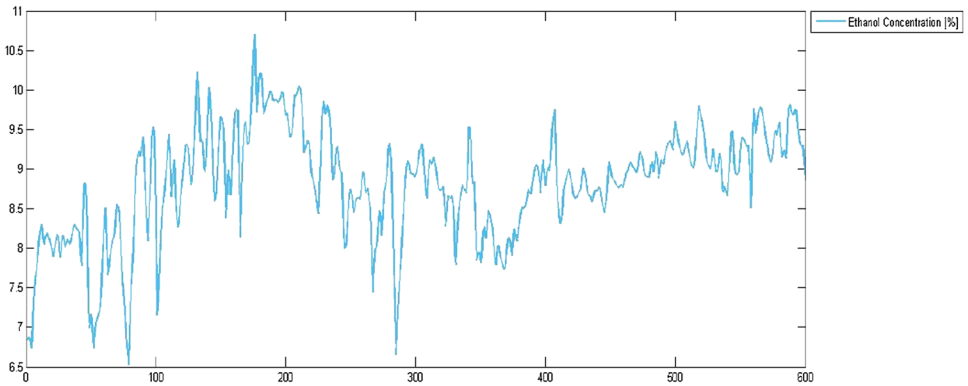

Fig. 5 Representative behavior of the chosen output variable, the ethanol degree, adopted to describe the fermentation process. The data correspond to May 2008. The x-axis corresponds to the sampled interpolated points, and the y-axis corresponds to the behavior of the variables

As any biological agent yeasts may be active (alive) or dead cells. The ratio of active cells in percentage is called yeast viability. In Brazil, the majority of industries adopt Melle–Boinot process. In this kind of process, up to 90–95% of the yeasts are recycled after every fed-batch fermentation cycle. Therefore, it is important that the yeasts reproduce (5–10%) during fermentation cycle to compensate cell loss due to the cell centrifugation process. Yeast growth rate gives a measure of this reproductive capacity. Nevertheless, in industrial processes yeast is stressed to limit cell growth to the minimal necessary amount, since biomass production competes with ethanol production.

Following the steps of the KDD preprocessing process, the data were passed through appropriate treatment. After being filtered, the data are subjected to the linear interpolation process, which considered an average of 20 points per day for all variables used, resulting in a set of 600 dots per month. Continuing the analysis of the number of records, only records of the must temperature variable showed blank records. A total of 37 records obtained in 2008, and 70 in 2009. The possible cause of this failure is related to specific communication problems between filed devices and the supervisory systems.

Recorded data corresponding to the quantities are discussed over the last three paragraphs, and which is used in the experiments are presented in a compact form in Table 2. Numeric values in the table cells corresponding to the production of the years 2008 and 2009 are separated by a "/" symbol. For each quantity, the number of available registries, maximal and minimal values, average (AVG) and standard deviation (SD) are presented in separated columns.

To evaluate the effects of the data processing and the interpolation step, the mean relative error between the data was calculated as shown in Table 3. One notes that there are situations where the percentage error is high, as observed in the third item on the table. This high error occurs due to the distance of experimental data collected from its normal operating range. According to these errors, we can identify whether these distances are above (positive errors) or below (negative errors) the variable average value.

3.2 Identification and Analyses of the KDD Performance

At this stage, five different topologies were analyzed and implemented on the NARX neural networks for data mining, from the definition of subsets: N1: 2–10 (two input regressors and two output regressors, and ten neurons in the hidden layer), N2: 5–10, N3 5–5, N4: 3–5, N5: 10–5. The learning algorithm was the gradient descent back-propagation with the tangent sigmoid activation function at the hidden layer and a linear transfer function at the output layer (Matlab) and the maximum training epochs were 800.

In total, 600 points in May 2008 were analyzed. These data were divided into a training set containing 420 records

Table 2 General information of the raw data concerning the fermentation kernel variables over the period of May–September 2008/2009

Variable	Number of registers	Max	Min	AVG	SD
TRS [% w/w]	589/468	27.0/23.4	12.0/7.1	19.2/17.6	0.8/1.7
Soluble solids [° Brix]	1763[a]/186098[b]	26.6/31.6	12.9/0	19.8/14.8	0.8/9.5
pH	1756/1435	5.3/5.1	3.8/3.8	4.5/4.6	0.1/0.1
Broth temperature [°C]	61646/37927	74.3/72.4	−*/10.7	32.4/29.9	4.1/4.5
Yeast viability [%]	295/246	98.0/99.0	67.3/68.5	90.9/90.3	4.8/4.4
Yeast budding rate [%]	295/246	30.4/26.3	0.6/4.3	13.0/13.1	4.9/4.4
Cane juice flow [ton/h]	1757/1434	14.3/19.0	3.0/1.0	8.4/9.7	1.7/2.5
Tank level [%]	43443/37942	64.4/59.7	0**/0**	32.8/32.0	7.4/9.3
Ethanol concentration [% v/v]	1756/1435	10.8/10.8	6.5/4.8	9.2/9.0	0.3/0.6

Subtitle:
[a] Data from manual acquisition
[b] Data from supervisory system
* Blank data
** Minimum tank level corresponds to the beginning of the fermentation

Table 3 Error calculated on the raw and processed data. Data concern May to September 2008/2009

Variable	Error on the average (%)
TRS [% w/w]	+1.0/ − 3.4
Soluble solids [°Brix]	+0.5/ − 7.4
pH	+0.0/ + 13.0
Broth temperature [°C]	+1.2/ − 0.6
Yeast viability [%]	+1.6/ + 2.2
Yeast budding rate [%]	−2.3/ − 6.1
Cane juice flow [ton/h]	−3.5/ + 6.1
Tank level [%]	+0.6/ − 5
Ethanol concentration [% v/v]	−1.0/ + 1.1

(70%) selected at random and a set of 180 records (30%). The values of the input and output variables were normalized to a mean value distributed on the [−1,1] range. During the training phase, the MSE for predicting the concentration of ethanol was used as a stop criterion (Table 4).

According to the results, the NARX N3:5–5 model was the best in the ethanol prediction task, with an average error of 0.08%, obtained during the validation step. Thus, this network structure was incorporated into the KDD model and validated. Figure 6 shows the ethanol concentration values that were predicted by the KDD model for August 2008.

The KDD values predicted by the model relate satisfactorily with data from experimental testing of the concentration of ethanol, due to grouping of data. In order to simplify the analysis of the predicted values for subsequent months, Fig. 7 shows the correlation of the subsequent months with the concentration of ethanol for the months of May, June, July, August and September 2008, and of May by September 2009.

In another analysis, the predictive values for the KDD model under specific fermentation conditions were verified. Tests were conducted to predict the behavior of ethanol concentration in specific situations and random operations. The KDD were presented to the data model for the first 4 h of fermentation, data for the first 8 h of fermentation, first hour fermentation data and data relating to the first and fourth hours of fermentation.

Table 5 shows the average percentage error between the estimated degree of concentration predicted by the model

Table 4 Mean square error for the testing data

Tr	Val	Top	2008				2009				
			Jun	Jul	Aug	Sep	May	Jun	Jul	Aug	Sep
0.04	0.04	N1:2–10	0.72	0.64	0.71	0.50	0.51	0.87	0.91	0.62	0.65
0.03	0.04	N2:5–10	0.19	0.22	0.45	0.38	0.25	0.19	0.10	0.35	0.50
0.01	0.03	N3:5–5	0.06	0.09	0.11	0.10	0.06	0.05	0.05	0.10	0.07
0.03	0.04	N4:3–5	0.06	0.08	0.09	0.21	0.08	0.06	0.09	0.28	0.36
0.02	0.04	N5:10–5	0.12	0.14	0.17	0.23	0.17	0.12	0.07	0.21	0.28

Subtitle:
Tr Training, *Val* validation, *Top* topology

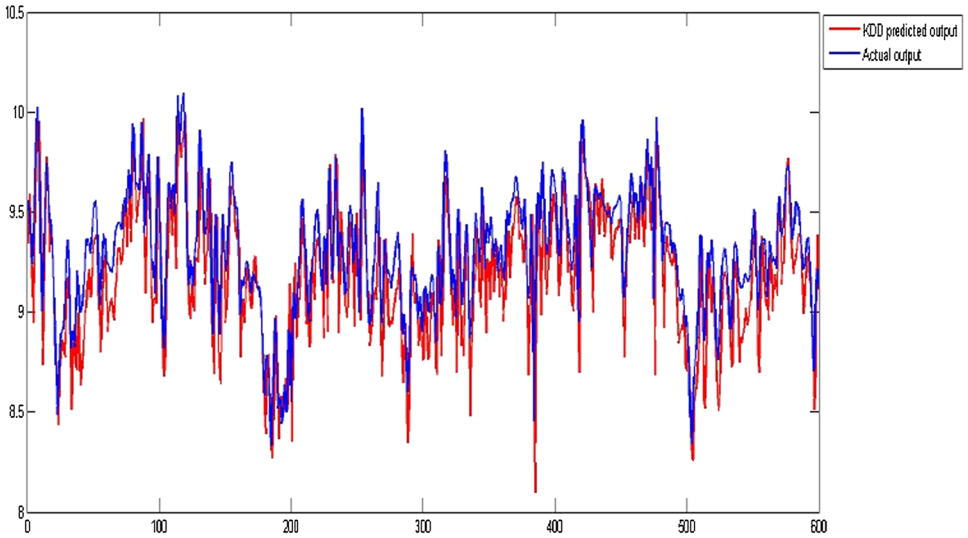

Fig. 6 Comparison between the ethanol concentration predicted by the KDD model and the actual value for the month of August of 2008. The x-axis corresponds to the sampled interpolated points and the y-axis corresponds to the behavior of the variable

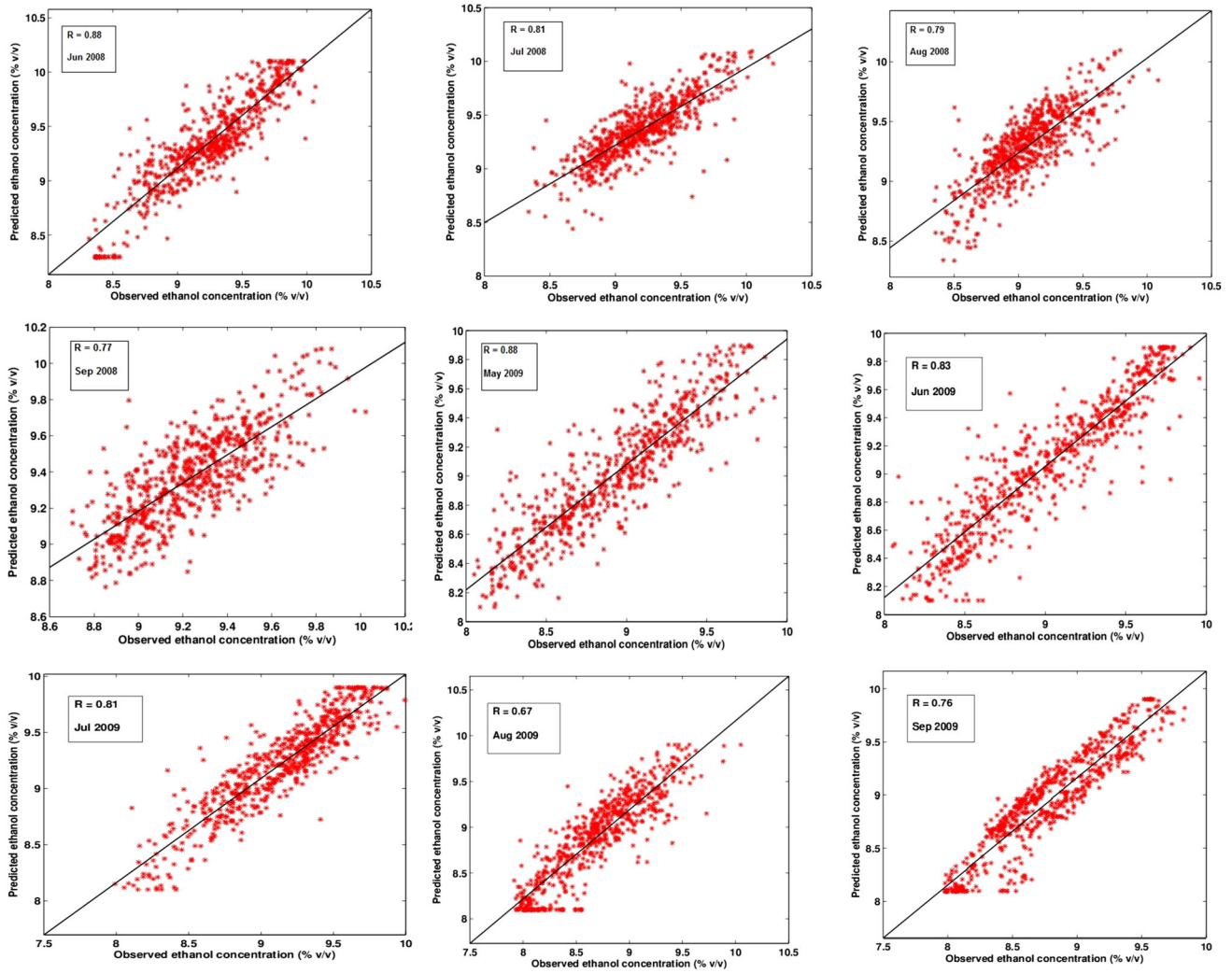

Fig. 7 Scatter plot on test data for the best KDD model with the selected NARX topology (5 regressors and 5 hidden neurons), as evaluated by the best MSE

Table 5 Mean square error for the specific situation of fermentation

Instant of fermentation (h)	2008				2009				
	Jun	Jul	Aug	Sep	May	Jun	Jul	Aug	Sep
$1st-4th$	0.02	0.03	0.02	0.01	0.05	0.04	0.06	0.04	0.05
$1st-8th$	0.01	0.01	0.03	0.02	0.03	0.02	0.05	0.04	0.05
$1st$	0.05	0.07	0.09	0.07	0.07	0.03	0.03	0.09	0.03
$1st$ and $4th$	0.08	0.08	0.09	0.10	0.07	0.06	0.06	0.07	0.07

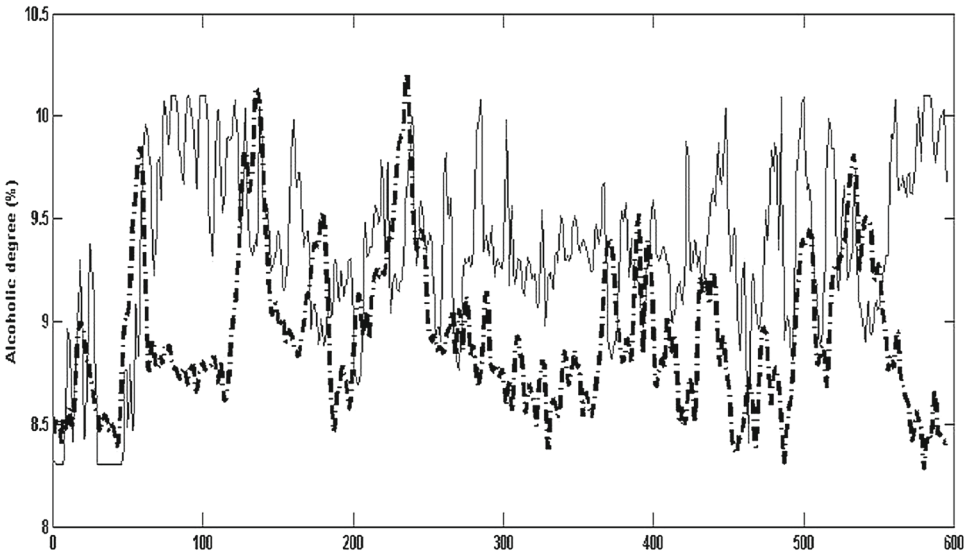

Fig. 8 Comparison between the ethanol degree of simulation obtained by KDD model and the observed data. The data corresponds to the month of August 2008. The x-axis corresponds to the sampled interpolated points and the y-axis corresponds to the behavior of the variable

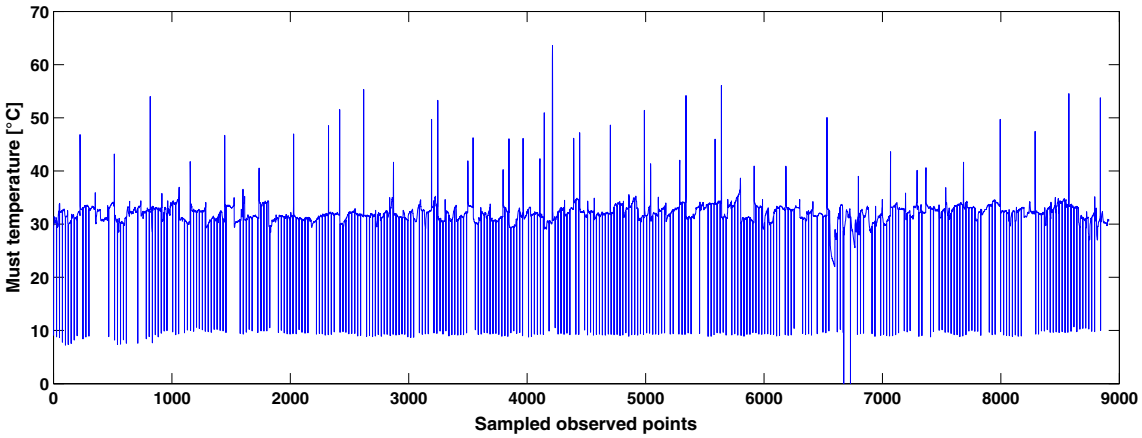

Fig. 9 Observed raw data for must temperature corresponds to the August of 2008

and KDD-grade ethanol observed under specific fermentation conditions.

Simulation tests analyzed the ability of the model to simulate possible KDD behavior of ethanol concentration in a given month, based on prior knowledge (the first 4 h of fermentation) of the process. The result of this test can be observed in Fig. 8, where the dashed line indicates the change in the value simulated by the KDD model and the continuous line the sampled values of the fermentation process (Fig. 9).

The analysis of the performance values, shown in Table 6, shows that the results are interesting and differ from expected behavior for the KDD model in simulation situations. The expectation was to obtain results that degrade with time; however, the error remains in a substantially acceptable range. Considerations regarding the error significance should take into consideration two respects: Firstly, the adopted sensors provide values accurately, which means that the model has a large standard deviation for this type of process, a sufficient

Table 6 Mean percentage error calculated by the difference between the ethanol concentration and the simulated output generated by the KDD model

	2008				2009				
	Jun	Jul	Aug	Sep	May	Jun	Jul	Aug	Sep
Mean percentage error (%)	6.00	6.00	5.30	4.12	7.10	6.62	6.72	4.80	8.61

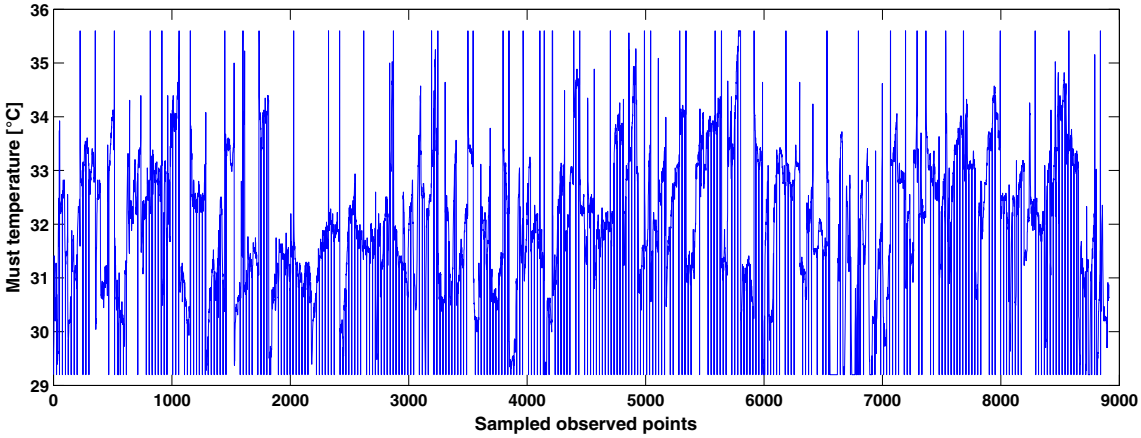

Fig. 10 Observed filtered data for must temperature corresponds to the August of 2008

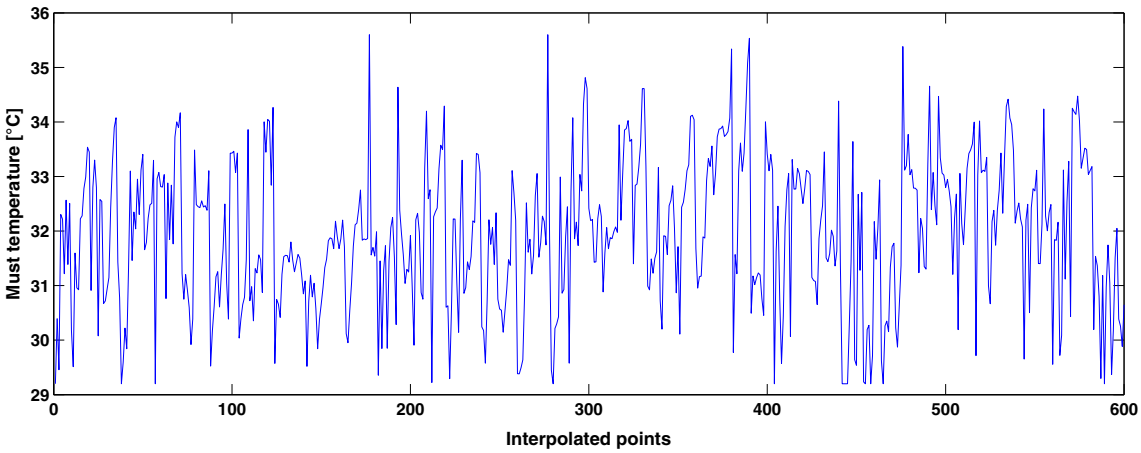

Fig. 11 Observed filtered and interpolated data for must temperature corresponding to the August of 2008

limit can be achieved accurately. Another aspect is that a miscalculation in the concentration of ethanol may represent an error in a certain amount of liters of alcohol per batch. Thus, this quantity of liters can represent a loss in the production of ethanol during the harvest (Figs. 10, 11).

4 Conclusion

In this paper, the authors present a structured roadmap to transform data into knowledge from the monitoring of the alcoholic fermentation process in sugar and alcohol. The proposed approach uses a methodology based on the extraction of knowledge in databases, KDD. To illustrate the feasibility of the proposal, actual data extracted from a production of a plant process were used. It implemented an intelligent system based on NARX neural network, which was able to perform both simulations, as estimates of ethanol production from input signals and preprocessed output. The experiments tried to reproduce logistical limitations by capturing similar data to that under actual operating conditions. In this context, it seems relevant and not trivial the fact that the errors observed in the system outputs during the simulations with the system over the following months did not degrade, and increase value. These results deserve more attention and deeper study with new data collections and the use of larger data vol-

umes. It is important to emphasize that the application of the method in other production units would require adjustment parameters and a new network of learning steps. In this context, even the network architecture should be tested. In this study, the network with a two hidden layer architecture with 5 neurons in the hidden layer proved to be more effective. The authors hope that further work will test whether this architecture itself is the most appropriate for the problem in question or whether the architecture will also be only a local production feature function. The characterization and mapping of the fermentation process will be applied, to provide a better understanding of the fermentation process operation, where it will be possible to analyze the influence and behavior of each variable process. This mapping will be applied in over the whole sugar and alcohol process, such as the energy generation, destilation sector and others.

References

Agudelo, W. E. H. (2012). Development and implementation of a software sensor for monitoring online of bioprocessos. Master's thesis, State University of Campinas, School of Chemical Engineering.

Albarelli, J.Q. (2013). Sugar and ethanol production of first and second generation: Simulation, energy integration and economic analysis. Ph.D. thesis, State University of Campinas, School of Chemical Engineering.

Amorim, H. V. (2005). *Alcoholic Fermentation: Science and Technology*. Piracicaba: Fermentec.

Batista, F. R. M. (2008). Study of the alcoholic continuous distillation process: Simulation of industrial plants for production of hydrated alcohol, neutral alcohol and cachaça. Master's thesis, State University of Campinas, Faculty of Food Engineering.

Batista, F. R. M. (2012). Computational simulation applied to the improvement of the bioethanol purification process. Ph.D. thesis, State University of Campinas, Faculty of Food Engineering.

Bergamasco, A. F. (2003). System decision support for the management of nitrogen fertilizers in sugar-cane harvested without burning. Master's thesis, State University of Campinas, Faculty of Agricultural Engineering.

Bingru, Y., Wei, H., Zhun, Z., & Huabin, Q. (2009). KAAPRO: An approach of protein secondary structure prediction based on KDD in the compound pyramid prediction model. *Expert Systems with Applications*, *36*, 9000–9006. doi:10.1016/j.eswa.2008.12.029.

Bocca, F. F., Rodrigues, L. H. A., & Arraes, N. A. M. (2015). When do i want to know and why? Different demands on sugarcane yield predictions. *Agricultural Systems*, *135*, 48–56. doi:10.1016/j.agsy.2014.11.008.

Browne, W. L., Yao, L., Postlethwaite, I., Lowes, S., & Mar, M. (2006). Knowledge-elicitation and data-mining: Fusing human and industrial plant information. *Engineering Applications of Artificial Intelligence*, *19*, 345–359. doi:10.1016/j.engappai.2005.09.005.

Choudharya, A. K., Oluikpeb, P., Hardinga, J., & Carrillob, P. (2009). The needs and benefits of text mining applications on post-project reviews. *Computers in Industry*, *60*, 728–740. doi:10.1016/j.compind.2009.05.006.

Cunha, M. J., Belini, V. L., & Caurin, G. (2012a). Discovery of behavior in industrial plants: A KDD based proposal. *8th IEEE International Conference on Automation Science and Engineering-CASE*. doi:10.1109/CoASE.2012.6386363.

Cunha, M. J., Belini, V. L., & Caurin, G. (2012b). Discovery of behavior in industrial plants: A KDD based proposal. Conference on Industry Applications (INDUSCON). *10th IEEE/IAS International*, 01–06. doi:10.1109/INDUSCON.2012.6451382.

Decloux, M., & Coustel, J. (2005). Simulation of a neutral spirit production plant using beer distillation. *International Sugar Journal*, *107*, 628–643.

Dias, M. O. S. (2008). Simulation of ethanol production processes from sugar and sugarcane bagasse, aiming process integration and maximization of energy and bagasse surplus. Master's thesis, State University of Campinas, School of Chemical Engineering.

Donauera, M., Pecas, P., & Azevedoa, A. (2015). Identifying nonconformity root causes using applied knowledge discovery. *Robotics and Computer-Integrated Manufacturing*, *36*, 84–92. doi:10.1016/j.rcim.2014.12.012.

Fayyad, U., & Shapiro, G. P. (1996). Data mining and knowledge discovery in databases: An overview. *Communications ACM, Special Issue on Data Mining*, *39*(11), 1–34.

Grunow, M., Gntherb, H., & Westinnerb, R. (2007). Supply optimization for the production of raw sugar. *International Journal of Production Economics*, *110*, 224–239. doi:10.1016/j.ijpe.2007.02.019.

Hahn, M. H. (1994). SISTECH: Transport system simulator of cane sugar. Master's thesis, State University of Campinas, School of Electrical and Computer Engineering.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. California, EUA: Morgan Kaufmann Publishers.

Harding, J. A., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering, ASME Proceedings*, *128*, 969–976.

Higgins, A. J. (2002). Australian sugar mills optimize harvester rosters to improve production. *Interfaces*, *32*, 15–25. doi:10.1287/inte.32.3.15.41.

Koksal, G., Batmaz, I., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, *38*, 13,448–13,467. doi:10.1016/j.eswa.2011.04.063.

Ljung, L. (2002). *System Identification*. Englewood: Prentice Hall.

Lydon, B. (2015). Big data in industrial automation. URL http://www.automation.com/automation-news/article/big-data-in-industrial-automation

Marquini, M. F., Mariani, D. C., Meirelles, A. J. A., Santos, O. A. A., & Jorge, L. M. M. (2007). Simulation and analysis of an industrial system of columns for ethanol distillation. *Acta Scientiarum-Technology*, *29*, 23–28.

McDonald, J. R., & Steele, J. A. (2006). Knowledge discovery in database: Applications in electrical power engineering domain. *IEEE Power Engineering Society Winter Meeting*, 8/1–8/4, doi:10.1049/ic:19971153.

Menezes, J. M. P, Jr., & Barreto, G. A. (2008). Long-term time series prediction with the NARX network: An empirical evaluation. *Neurocomputing*, *71*, 3335–3343. doi:10.1016/j.neucom.2008.01.030.

OPC (2015). OPC foundation. URL https://opcfoundation.org/

Profibus (2015). Profibus international. URL http://www.profibus.com

Stephens, L. J. (1998). *Schaurns Outline of Theory and Problems of Beginning Statistics*. New York: McGraw-Hill.