Mathematical Modeling of the Biochemical Oxygen Demand (BOD) from Water Table in Sugar Cane Plantation Fertirrigated with Vinasse in Brazil

José Antônio Aleixo da Silva1, Marília Regina Costa Castro2, and Lourdinha Florêncio3
1. Universidade Federal Rural de Pernambuco (UFRPE), Brazil
2. Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE), Brazil
3. Universidade Federal de Pernambuco (UFPE), Brazil

Abstract: The uses of the vinasse as fertilizer through fertirrigation for sugar cane plantations constitute a potential high source pollutant of soils and underground waters. Thus, it is necessary to monitor areas fertirrigated with vinasse to evaluate the quality conditions of those waters in order to conduct inferences or diagnoses of the environmental quality. The decomposition of the vinasse organic matter, represented by the biochemical oxygen demand (BOD) is the most important parameter in terms of pollution indicators. The BOD laboratory analysis is not a direct measurement and can last some days to determine it. The adjustment of mathematical models can generate equations to estimate BOD as a function of independent variables that can be determined directly in short time. The data set for this study came from an area of 12 hectares planted with sugar cane, fertirrigated with vinasse. Thirty observation wells were installed with depths of 3 m, distributed in parallel transects 100 m from one to another. Eleven water quality parameters from three collect along the time (three, six and nine months) were analyzed. Simple and multiple linear and non-linear models, the Chapman-Richards and Silva-Bailey models were used to estimate the BOD. The Adjusted Fit Index in percent (AFI%) obtained presented values of adjustments superior to 95%. The Silva-Bailey model adjusted with two independent variables COD (chemical demand of oxygen) and T (age of the plantation) showed the higher AFI% = 98.78%. In the experimental conditions, the results showed that BOD could be estimated mathematically through its relationship with other chemical variables of simplified analytical procedures. The chemical oxygen demand (COD) stood out among the selected variables.

Key words: biochemical oxygen demand (BOD), chemical oxygen demand (COD), vinasse, mathematical models, regression analysis

1. Introduction

The vinasse resultant from the alcohol processing is the main sub-product of the sugar cane industry as function of the great amount of volume produced. Among its main uses, there are fertirrigation, animal feeding and biogas production.

Early scientific studies on vinasse date from 1950, carried out by the researchers of ESALQ. Later, after the Brazilian Pro-alcohol Program, the development of sugar cane industry brought an increased production of alcohol and, consequently, of vinasse production. At that time the current practice was the accumulation and discard of vinasse in the courses of water generating high levels of pollution. Thus, a strong pressure from the environmental agencies began, especially from CETESB of the State of São Paulo, and as result changes were made in the strategy of the sugar mills concerning the disposition of the vinasse in the water and soil [1]. Nowadays, the
control of the BOD pollution is extended along the country.

Several studies about vinasse disposition in the soil have been carried out with a focus on the effects in the pH of the soil, physiochemical properties and their effects on the culture of sugarcane [2-7]. However, only a few of them evaluated the pollutant potential of vinasse to the soil and the water table [7-9].

According to Almeida (1952) and Freire & Cortez (2000), vinasse as a pollutant of the courses of water demands a high amount of oxygen for decomposition of the contained organic matter [10, 11]. Therein it is harmful to aquatic animals such as toads, fish and crustaceans, it drives away the marine fauna that seeks the coast for spawning, destroys fish larvae causing a population unbalance and kills the microflora, microfauna and the submerged aquatic plants. Vinasse is bad smelling due to the anaerobic decomposition of originated gases with toxic characteristics.

Although the technology of the application or disposition of vinasse in the soil has been researched since 1950, only after 1987 those studies were developed to evaluate the possibility of vinasse components causing pollution in the underground waters.

The amount of organic matter, represented by BOD is an important parameter for knowing the degree of pollution of the residual water, as well as to set up the treatment stations of that residue. The larger the degree of organic pollution, the larger is the BOD of the water; similarly, when there is stabilization of organic matter, the BOD decreases [12].

Table 1 shows some data that makes possible to evaluate the pollutant power of vinasse, as result of the BOD in relation to different types of residual waters [13].

<table>
<thead>
<tr>
<th>Residual waters</th>
<th>BOD (mg/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effluent of alcohol distillation (vinasse)</td>
<td>15000-20000</td>
</tr>
<tr>
<td>Sanitary sewers</td>
<td>200-600</td>
</tr>
<tr>
<td>Effluent of canned foods</td>
<td>500-2000</td>
</tr>
<tr>
<td>Effluent of breweries</td>
<td>500-2000</td>
</tr>
<tr>
<td>Effluent of oil processing</td>
<td>15000-20000</td>
</tr>
<tr>
<td>Effluent of dairy products</td>
<td>30000</td>
</tr>
<tr>
<td>Paper Mill</td>
<td>500</td>
</tr>
<tr>
<td>Effluent of slaughterhouse</td>
<td>30000</td>
</tr>
</tbody>
</table>

Table 1  Values of BOD for different types of residual waters [13].

months, in an area of red yellow latosol, through tensiometers and extractors of solution of the soil placed at several depths (up to 1.20 meters). It was found that the risk of pollution of the water tables was small, and there was small risk of the potassium and nitrate polluting the underground water.

1.1 Biochemical Oxygen Demand (BOD)

During the process of decomposition of the present organic matter in the residual waters, there is formation of a range of organic substances. However, it is not necessary to characterize each one of those substances, since their variety in forms and compounds would make laboratory analyses more difficult. This way, direct or indirect methodologies have been adopted to evaluate the present organic matter in the samples [15]. Direct methods measure the consumption of oxygen and are expressed as biochemical oxygen demand (BOD) and chemical oxygen demand (COD). Indirect methods measure the organic carbon [16]. According to Areerachakul & Sanguansintukul (2009) the most frequent way to measure the quality of present organic matter in wastewater is through the determination of BOD [17].

1.2 Regression Analysis

The use of regression analysis has been frequent in estimative processes, especially, when there are variables of difficult determination in terms of methodological procedures or even as regard costs. For instance, the determination in laboratory of BOD that can take five days. Therefore, regression analysis
is a mathematical procedure that can be used in those situations, where it is considered a dependent variable that is more difficult to determine, as a function of independent variable(s) easy to be direct measured.

The relationships between the dependent variable (the response variable) with the independent variable(s) can be expressed as linear or non-linear functions. When that relationship is non-linear, but can be transformed in linear through the use of logarithms or another transformation, the model is called non-linear intrinsically linear. When no transformation allows linearization the model is called non-linear intrinsically non-linear. The estimates of the coefficients of those models are calculated through numeric analysis [18].

When the estimate of the dependent variable can be calculated as a function of a single independent variable, there is a simple linear model or a simple non-linear model. When more than an independent variable is used, the model is multiple linear model or a multiple non-linear model.

In the multiple models, the selection of the independent variables in the final equation can be done through some procedures that are based on the existing correlations among the dependent variable and the independent variable(s) and among the independent variables.

However, some basic requirements such as normal distribution and errors independence, constant variance and non-collinearity among the independent variables should be considered [19]. When some of those requirements are violated, the use of transformations of the data is required [14, 20].

Box and Cox (1964) developed a family of transformations that can be used in any dataset where the dependent variable is positive [21]. The Box and Cox methodology indicates the correct transformation to a particular data set. It is expressed as:

$$W_i = \begin{cases} \frac{(Y_i^\lambda - 1)}{\lambda} & \text{for } \lambda \neq 0 \\ \ln Y_i & \text{for } \lambda = 0 \end{cases}$$

Where:
- $W_i$ = transformed dependent variable
- $Y_i$ = dependent variable in the original form

That family depends directly on the parameter $\lambda$ that is the coefficient of transformation of the data, as well as of the vector of parameters $\beta$ for the model to be adjusted this way:

$$W = X \beta + \varepsilon$$

Where:
- $W = (W_1, W_2, \ldots, W_n)$

The value of $\lambda$ that maximizes the function is [22].

$$L_{\text{max}} \lambda = -\frac{n}{2} \ln \left( \frac{\text{RSS}}{n} \right) + (\lambda - 1) \sum_{i=1}^{n} \ln Y_i$$

Where:
- $\text{RSS} = \text{residual sum of squares}$,
- $n = \text{number of observations}$,
- $\ln = \text{Napierian logarithm}$,
- $Y_i = \text{dependent variable}$.

When the confidence interval for $\lambda$ includes the unit (1,0) the data should not be transformed and the resulting equation should be considered.

Such transformation can be adjusted by iterative mathematical methods this way [23].

$$Y_i = \left[ \lambda \left( \beta_0 + \beta_1 \sum_{i=1}^{n} X_i \right) + 1 \right]^{1/\lambda} + \varepsilon_i$$

Where:
- $Y_i = \text{dependent variable}$,
- $\lambda = \text{coefficient of transformation of the data}$.

The other variables and coefficients as defined before.

Because the water is one of the most important natural resources, contamination can cause serious problems relating to the environment that could compromise especially the health of people. For this reason, the mathematical modelling of water quality is of great importance and several studies have been conducted [17, 24-29].

As the biological processes generally cannot be represented by linear functions, the use of the
nonlinear intrinsically nonlinear functions can improve the estimates of those processes.

2. Material and Methods

2.1 Experimental Area

The territorial space of Pernambuco comprises five geographical mesoregions. Among them there is the mesoregion of the Tropical Rain Forest of Pernambuco, where most of the sugarcane plantations are found.

The experiment took place in an area from Salgado Sugar Mill, located in the municipal district of Ipojuca (coordinates 07°10’ and 07°25’ S, 39°10’ and 39°30’ W) in the state of Pernambuco, 53 km from the Recife, capital of the state. The climate is of the type Ams, according to Köppen, tropical rainy with dry summer, with a rainy station from March to August. The average annual precipitation in the area is 1800 mm and the medium annual temperature is 25.2°C [7].

To monitor an area of 12 hectares with three soil types, two cultivated with sugarcane and another without vegetation where was the pond of vinasse distribution, 30 observation wells were installed with depths of 3 m, distributed in parallel transects, 100 m from one to another. Three samples of 11 quality water parameters of the water table were collected at 3, 6 and 9 months.

2.2 Biochemical Oxygen Demand (BOD)

The concentration of biodegradable organic matter was determined by the test of the standard bottles with incubation to 20°C for 5 days [30].

For each sample the concentration percentages were calculated to proceed to the dilutions already taking into account the values of COD. The concentration of dissolved oxygen was determined by the Winkler method [30].

2.3 Statistical Analysis

Due to the complexity of laboratory routines involved to determine BOD, it was decided to model it as a function of other variables that are easier to determine: Chemical Oxygen Demand (COD), Electric Conductivity (EC), Total Solids Dissolved (TSD), Bicarbonate (BICAR), Chloride (CLORE), Sodium (NA), Potassium (K), Magnesium (MG), Nitrate (NTA), Nitrite (NTI) and Time (T).

The tested models were:

1. Linear model

\[ BOD_i = \beta_0 + \beta_1 \sum_{i=1}^{n} X_i + \epsilon_i \]

2. Non-linear model

\[ BOD_i = \beta_0 \prod_{i=1}^{n} X_i^{\beta_i} + \epsilon_i \] (Chapman-Richards) (2)

\[ BOD_i = \beta_0 \cdot COD_i^{\beta_i} + \epsilon_i \] (Silva-Bailey) (3)

3. Non-linear model with the transformation of Box and Cox (1964) [21]

\[ BOD_i = \frac{[\lambda (\beta_0 + \beta_1 \sum_{i=1}^{n} X_i) + 1]^{1/\lambda}}{\lambda} + \epsilon_i \]

Where:

- \( BOD_i \): Biochemical oxygen demand (dependent variable);
- \( COD_i \): Chemical oxygen demand;
- \( T_i \): time (age of the plantation);
- \( X_i \): other independent variables;
- \( \beta_0, \beta_i \): coefficients of the models;
- \( \lambda \): coefficient of transformation of the data;
- \( \epsilon_i \): random error;
- \( i = 1, \ldots, n \).

Firstly, the coefficient of correlations among the independent variables and BOD was calculated to model the BOD in function of the more correlated independent variable. Then the BOD was estimated as a function of all the other independent variables, with the purpose of improving the precision of the final equations with the possible introduction of more variables in the equations.

The selection and elimination of independent variables in the final equations was processed through
the Stepwise method, by considering the level of 1% of probability to the variable be included in the final equation [31].

To compare linear models with nonlinear models it can be used the Fit Index (FI) [32] that is expressed for:

\[ FI = 1 - \frac{\sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2} \]

Where:
- \( Y_i \) = observed value of the dependent variable;
- \( \hat{Y}_i \) = estimated values by the equation;
- \( \bar{Y} \) = mean of the dependent variable.

As the resulting equations had different number of coefficients it was used the Adjusted Fit Index in percent (AFI%), expressed as:

\[ AFI\% = 100 \left[ 1 - (1 - FI) \left( \frac{n-1}{n-p} \right) \right] \]

Where:
- \( n \) = number of observations;
- \( p \) = number of coefficients in the equation.

All the statistical analysis were performed using the SYSTAT (System of Statistics) version Demo.

### 3. Results and Discussion

The resulting for the coefficient of correlations among all independent variables with BOD showed the following results (Table 2).

#### Table 2  Correlation coefficients among the BOD and the independent variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>BOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>COD</td>
<td>0.978</td>
</tr>
<tr>
<td>EC</td>
<td>-0.011</td>
</tr>
<tr>
<td>STD</td>
<td>-0.002</td>
</tr>
<tr>
<td>BICAR</td>
<td>0.245</td>
</tr>
<tr>
<td>CLORE</td>
<td>-0.218</td>
</tr>
<tr>
<td>NA</td>
<td>0.081</td>
</tr>
<tr>
<td>K</td>
<td>0.011</td>
</tr>
<tr>
<td>CA</td>
<td>0.057</td>
</tr>
<tr>
<td>MG</td>
<td>0.071</td>
</tr>
<tr>
<td>NTA</td>
<td>0.463</td>
</tr>
<tr>
<td>NTI</td>
<td>0.625</td>
</tr>
<tr>
<td>T</td>
<td>0.007</td>
</tr>
</tbody>
</table>

As expected the most correlated independent variable was the COD with a value of 0.978. All the others independent variables had low correlation with BOD. The one that show higher correlation coefficient value after COD was the NTI = 0.625. The independent variables EC, STD, BICAR, CLORE, NA, K, CA and MG, can be considered as non-correlated with BOD.

The adjustment of the full model with all independent variables shows the following result:

#### 3.1 Linear Multiple Model with All Independent Variables

\[ \hat{\text{BOD}} = 39.0476 + 0.1680\text{COD} - 0.0078\text{EC} - 0.0162\text{STD} + 0.0530\text{BICAR} - 0.0159\text{CLORE} + 0.1216\text{NA} - 0.0309\text{K} + 0.0837\text{CA} + 0.4583\text{MG} - 4.4116\text{NTA} + 10.3396\text{NTI} - 5.7621\text{T} \]

\( \text{AFI}\% = 97.79\% \)

The next step was to apply the Stepwise procedure to eliminate independent variables that are non-significant in the final equation.

#### 3.2 Stepwise for the Multiple Linear Model with All Variables

\[ \hat{\text{BOD}} = 48.9945 + 0.1667\text{BOD} - 0.0173\text{CLORE} + 10.6642\text{NITRI} - 7.0563\text{T} \]

\( \text{AFI}\% = 97.42\% \)

Four independent variables were selected to remain in the equation: COD, CLORE, NITRI and T. The loss in precision when compared with the full equation was only 0.36%, which is non-significant.

#### 3.3 Non-linear Model with the Box and Cox Transformation

\[ \hat{\text{BOD}} = [0.9190(-12.1989 + 0.1099\text{COD} - 0.0010\text{CL} + 14.6329\text{NITRI} + 0.8519\text{T}) + 1]^{(0.9180)} \]

\[ \hat{\text{BOD}} = (-10.2108 + 0.1010\text{COD} - 0.0009\text{CL} + 13.4476\text{NITRI} + 0.7829\text{T})^{0.8811} \]

\( \text{AFI} = 96.95\% \)

\( 0.7948 \leq \lambda \leq 1.0432 \)
The input of the parameter $\lambda$ in the model with four variables decrease the precision and the confidence interval for $\lambda$ included the unit, indicating the no transformation must be applied in the model.

As the COD is the independent variable more correlated with the BOD a simple linear, a simple non-linear and the simple linear model with the Box-Cox transformation were adjusted to the data set with this variable.

The resulting equations for the BOD estimation as a function of COD were:

3.4 Linear Model

$$\hat{\text{BOD}}_i = -4.8309 + 0.1847\text{COD}_i$$

AFI% = 95.65%

3.5 Non-linear Model

$$\hat{\text{BOD}}_i = 0.1470\text{COD}^{0.0271}$$

AFI% = 95.67%

3.6 Non-linear Model with the Box and Cox Transformation

$$\hat{\text{BOD}}_i = [0.9713(-0.5715 + 0.1565\text{COD}_i) + 1]^{0.9713}$$

$$\hat{\text{BOD}}_i = (-0.5551 + 0.1520\text{COD}_i)^{0.0295}$$

AFI% = 95.67%

0.8761 $\leq \lambda \leq 1.0634$

The results show that the three equations presented good values of AFI%, once the smallest registered value was (95.65%). The Box and Cox transformation included the unit (1.0) in the confidence interval for $\lambda$, showing that is not necessary to transform the data for the estimate of BOD as function of COD. The simple non-linear equation presented higher AFI% = 95.67%, but when compared with the simple linear equation, the gain in precision was only 0.02% that is no significant. It means that the simple linear model with COD as independent variable is sufficient to estimate the COD in plantations of sugar cane that uses vinasse as fertilizer.

Comparing the equation with the independent variable COD with the full equation, the loss in precision was only 2.14%. With relation to the equation resulting from Stepwise procedure, the loss in precision was 1.77%. However, collecting data of a large number of independent variables involves costs and time consuming. For this reason, it was decided to consider just de independent variable COD to adjust the model of Chapman-Richards.

3.7 Chapman-Richards Model

$$\hat{\text{BOD}}_i = 1271.3632[1 - \exp(-0.0003\text{COD}_i)]^{0.3231}$$

AFI% = 95.95%

Even showing a high precision in terms of AFI%, the non-linear model of Chapman-Richards just increases the precision in 0.3% comparing with the simple linear model.

3.8 Silva-Bailey Model

The Silva-Bailey model is a two parameters model and the time (T) must be included as an independent variable. Thus, the model was adjusted with two independent variables: COD and T. The final equation was:

$$\hat{\text{BOD}}_i = 0.8244\text{COD}_i^{0.9812}\text{exp}^{-0.9400T}$$

AFI% = 98.78

The Silva-Bailey model presented a better adjustment. The increase in precision when compared with the simple linear model just with the independent variable COD was 3.13%. However, it is necessary to have information about the age of the sugarcane plantation that is ease to obtain. In the absence of this information, the Silva-Bailey model cannot be applied.

4. Conclusion

All the models tested to estimate BOD in function of 11 independent variables generated equations with higher precisions, with AFI% superior to 95%. The model of Silva-Bailey with two parameters (COD and
T) presented the higher AFI% = 98.78. The independent variable COD was the most correlated with BOD with a coefficient of correlation of 0.978 indicating that BOD in the conditions of this study can be estimated as function of only COD, using a single linear model.

References


[25] V. A. Misra, P. Chandra and J. B. Shukla, Mathematical modeling and analysis of the depletion of dissolved


