Modeling Anaerobic Digestion with Artificial Neural Networks

Liliana Mafalda Soares Fernandes
Email: liliana.fernandes@ist.utl.pt
Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal; November, 2014

***Abstract*—Artificial neural networks (ANN) are one of the latest tools used to model and predict complex problems, that cannot be treated using conventional solutions. An example of such problems is the anaerobic digestion. This study uses the ANN to model and predict the production of methane in the anaerobic digesters of the WWTP of Guia located in Cascais, Portugal. Operational data of the plant for a period of 12 months was collected and employed in the analysis. The study considered the following digesters operational parameters: Input sludge flow in the digesters, input sludge flow in the Solid Phase of the treatments of the WWTP in study, the input percentage and load of total solids in the digesters. For predicting the production of methane, a model with ANN was built, with one hidden layer containing 30 neurons and a maximum of 600 iterations. The training and testing parts of the construction of the model were performed with the first 9 months of the data. During the construction of the model, the prediction of the testing set had a mean normalized error of 9.84% and a mean coefficient of determination (R^2) of 0.86. The model was then validated with the data that was not used during the training and testing phases of the construction of the model (last 3 months of the data), demonstrating the effectiveness of the model to predict the production of methane, with a $R^2 = 0.79$ and a normalized error of 11.6%.**

Keywords: Biogas, Prediction, Modeling, Artificial neural networks, Anaerobic Digestion

I. Introduction

Solid organic waste removal has become an ecological problem, first brought to light as a result of an increase in public health concerns [1]. The processing of refuse was usually undertaken to reduce the pollution potential and volume to ease the handling and disposal. This perspective has since been adjusted to include the transformation of the waste, which was by the time unwanted, into useful end-products [2]. Recently, the organic fraction of solid waste has been recognized as a valuable resource that can be converted into useful products via microbially mediated transformations [1].

Various methods are available for the treatment of organic waste but anaerobic digestion appears to be the most promising approach [1]. Anaerobic digestion involves a series of metabolic reactions such as hydrolysis, acidogenesis and methanogenesis.

The anaerobic digestion of organic waste in landfills releases the gases methane and carbon dioxide that escape into the atmosphere and pollute the environment. Under controlled conditions the same process has the potential to provide useful products such as biofuel and organic amendment (soil conditioner) and the treatment system does not require an oxygen supply. Further, methane and hydrogen as potential fuels are considered comparatively cleaner than fossil fuel [1].

The biofuel, which is routinely referred as biogas, can be used as fuel to help the WWTP meet its energy needs, and thus, its use has the benefit of reducing the dependency of the WWTP on fossil fuels [1]. Thus, anaerobic digestion represents an opportunity to decrease environmental pollution and, at the same time, to provide biogas and organic fertilizer or carrier material for biofertilizers [1].

Summarizing, the anaerobic digestion helps in reducing the environmental pollution and, at the same time, the products formed during this process (biogas and biofertilizers), can be used to reduce the energy costs of the WWTP and to add value the organic waste, respectively. Given the growing concern over the decrease of the dependence on fossil fuels and also the increasing concern to reuse the compounds formed during the treatment of wastewater, like the biogas, the need for methods to control and optimize the process of biogas production is clear. The core of an control/optimization system is the model describing the process. Despite several attempts to model the processes that occur in a digester, up to now classical mathematical modeling was only possible when severe simplifications of the process representation were performed [3], [4].

The main reason for this situation is that the mechanisms ruling this processes are not sufficiently well understood to formulate reliable non-linear mathematical models [4]. As an alternative, artificial neural networks are claimed to have a distinctive advantage over some other non-linear estimation methods used for bio-processes, because they do not require any prior knowledge about the structure of the relationships that exist between important variables [4].

Neural networks have been used in anaerobic digestion systems to describe trace gases [5], digester start up and recovery [4], advanced control and prediction of biogas [6] and in modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm [7]. In [5] an ANN model was developed to predict trace gases in biogas stream such as hydrogen sulfide and

ammonia. The model was capable of predicting the trace gases successfully even under dynamic conditions. In [4], the authors were interested in the start-up and in the adaptation phase of a bioreactor and in the recovery of the biocoenose after a toxic event. It was showed that the anaerobic digestion of surplus sludge can be effectively modeled by means of a hierarchical system of neural networks and a prediction of biogas production and composition can be made in several time-steps in advance. It was thus concluded that is possible to optimally control the loading rate during the start-up of a non-adapted system and to recover an anaerobic reactor after a period of heavy organic overload. In [6] different ANNs were used to model and control the production of methane from anaerobic continuously stirred tank reactors that were operated under different organic loading rates. It was concluded that the developed models could effectively predict the gas production and composition from the reactors. Finally, in [7] a multi-layer ANN model was trained to simulate the digester operation, and to predict the methane production. The performance of the ANN model was verified and demonstrated the effectiveness of the model to predict the methane production accurately. The developed ANN model was used with a genetic algorithm to optimize the value of the % of methane.

II. Characteristics of the WWTP

The WWTP under study does the collection, treatment and final rejection of urban wastewater from approximately $800,000$ equivalent-inhabitants and covers an area of $220km^2$. The supply to the WWTP is done through a general interceptor with $24,7km$ and the rejection is made through an outfall at $2,7km$ from the coast at a depth of $45m$.

The WWTP comprises two stations at different locations but connected through a $4km$ long conduct. These treatment plants are: the Liquid Phase Treatment Station (LP) where where the treatment of liquid waste occurs. The resulting sludge, from these treatments, is sent to the other treatment station: Solid Phase Treatment Station (SP). Currently this WWTP, provides primary treatment with additional treatment during the months of the bathing season, in order to increase the capacity of removing the pollution load of waste water. In a general way, the sludge that arrives at SP is subject to a thickening process, then to a anaerobic digestion process, followed by a dehydration process. After this, the treated sludge is stored to be transported to the appropriate final destination.

III. Materials and Methods

A. Operational Data

The operational data provided was related to the input sludge flow that arrives to the SP (IFSP), the thickening process, the anaerobic digestion process and the input of the dehydration process. The variables provided related to these processes were: % of Total Solids (pTS), % of Volatile Solids (pVS), Alkalinity, Volatile Fatty Acid (VFA), ratio Alkalinity/VFA and the sludge input and exit flows for the thickening process, the sludge input flow for both digestion (IF) and dehydration processes and the sludge input flow that arrives to the SP (IFSP). For the digestion process the following variables were also provided: biogas flow (BF), pH and temperature (T) both inside of the digesters, and % of methane in the biogas (pmethane). The data provided was related to a period of twelve months, starting in August 2013 and ending on July 2014. The first nine months (August 2013 to April 2014) were used in the training and testing of the models. The last three months (May, June and July 2014), that were not used in the construction of the models, were predicted using the model to validate its accuracy. For future reference, summer months, in this work, correspond to the months of the bathing season, which starts on June 1st and ends on September 30th. The remaining months (October to May) are referred to as winter months.

B. Steps of Data Processing and the Development of Predictive Models

The general steps of the data processing that were done and the major steps that were made in the development of the prediction models are described next.

The first step was to represent all variables that were provided in histograms and in boxplots. During analysis of these representations it was observed that the data related with the input flow in the SP and the data related with digestion had seasonality. That is, the data of the summer months tends to have higher values than the data from the winter months. Was however observed, that there was no seasonality on the data related to thickening and dehydration processes. Because of this, it was decided to drop the data related to the thickening and dehydration processes. The outliers elimination was made by crossing the information of the abnormal conditions on the operations of the WWTP and the information from the boxplots.

A correlation analysis with the Pearson correlation coefficient (defined for example in [8]) was also done. After this analysis we decided to group all variables which represent the same variable but are related to a specific digester (there are, in the WWTP in study, three digesters A, B and C), in one variable, creating a new variable named “Digester”, in order to not lose the information about which digester a given value is related to.

In this step it was also decided to only use in further data processing, the following variables: IFSP, IF, T, pH, pmethane, pTSI, % of Total Solids Output (pTSO) and BF, because the remaining variables had a very few points and as such, was not possible to include them to be used in the next steps of data processing. It was observed that there no variable was correlated with pmethane. As such, in order try to improve the correlation with this variable and also to improve the correlation with the variable pTSI, we decided to calculate two new variables: TS input load in the digesters (TSIL) and methane production (P_methane) resulting from the multiplication of the pTSI and the IF and by the multiplying of the pmethane and the BF, respectively.

Then, with these ten variables, a PCA analysis was made and with the information gained by, both, this and the correlation analysis, the selected variables more related with the variables that we want to predict (P_methane/BF) were: IFSP, IF, pTSI

and TSIL. With these variables the prediction models, using both PLS regression and ANN methods, were constructed for P_methane/BF prediction, using data from August 2013 to April 2014. With the best obtained model from the ANN method, the prediction for the months of May, June and July 2014 was made. We then tried to improved the performance of prediction by going back and re-training and testing this new model with data between August 2013 and June 2014. The validation of this model was made with the data from July 2014.

C. Algorithms and Software libraries

This work was performed with the help of various Python[1] libraries. In particular, the library scikit-learn [9] was used, which implements various algorithms and methods commonly used in machine learning, including, among many others, PLS and ANN. For finding the optimal number of iterations and neurons, an exhaustive search was performed across the space containing all possible combinations of number os iterations and number of nodes. To perform this search we made use of the grid-search implementation provided by scikit-learn, that allows one to vary several parameters simultaneously. For drawing the histograms and plotting the boxplots the matplotlib library [10] was used.

IV. Results

A. Data Pre-Processing

Today's real-world data is highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results [11]. There are a number of data pre-processing techniques like: data cleaning, data integration, data reduction and data transformations. Data cleaning, can be applied to remove noise and correct inconsistencies in the data, e.g., by filling missing values (when possible) and removing outliers. There are several methods to fill the missing values like inserting the mean or the most frequent value. One can also make a model to predict the missing values of this variable by using the other variables as input. However this is dependent on the nature of the values available. On the outliers, usually, the approach taken is to represent the data in boxplots and observe the outliers, which corresponds to the points outside of the boxes in the boxplots representation. Another step in data pre-processing is data integration. This step merges data from multiple sources into coherent data. In the data reduction step, high dimensional data is transformed in a data representation with a much smaller number of dimensions while still containing most of the information from the original source. Strategies to accomplish this are the dimensionality reduction, such as Principal Component Analysis (PCA). Data transformations, such as normalization, may be applied. Data pre processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining [11].

1) Histograms and Boxplots: Like mentioned earlier, in this step all the variables related to entry in the SP, input and output of the thickening, entry and exit of digestion and also for the variables related with entry in the dehydration, were represented in histograms and boxplots. Seasonality in the data related with entry and exit of digestion and with the entry in the Solid Phase was observed, and since this was not observed in the data related with thickening and dehydration processes, it was decided in subsequent analysis not use the variables of these operations and focus only on variables related to input and output of digestion being that it was decided to also use the input flow in FS.

2) Outliers elimination: As mentioned earlier, it is common to fill the missing values. However, this largely depends on the problem at hand. In this case study, several variables have less than thirty values while others variables have 273 values. Predicting about 240 values it seems an impossible task because almost any existing method to fill in missing values would be inappropriate. Also in the case of not fill the missing values and train a model with only thirty points it seems likewise impossible, because certainly the model obtained would not be a good model. Thus, the approach used about the missing values was to delete rows (days) where there is at least one missing value.

The most usual way of removing outliers is to eliminate all of the data points that are outside the limits of the boxes of the boxplots, however, due to having already a short number of points to work, was decided to subjectively eliminate the points outside the boxes, getting started by eliminating the days where the digesters were operating under abnormal conditions. Such as between February and April months where the digesters suffered interventions. Next, the data without these days were represented in boxplots and histograms, like in the Figure 1. In this figure, can be observed that, there are a few outliers for this variable, corresponding to the points outside whiskers (the lines extending vertically f rom the boxes) of the boxplots. One can also observe the seasonality in the data, since for the months of August and September the values are higher than for the remaining months (which are winter months).

The exceptions are the months of March and April that have almost the same height as the summer months. This is probably a consequence of the interventions in the digesters that are affecting the normal behavior on these months. As these months belong to the set of winter months, under normal conditions, they should be at the same level, of the remaining winter months.

In regard to the histograms presented in the same figure, it can be observed that the data can be approximated to a normal distribution.

It was decided,as previously mentioned, not to remove all points outside the boundaries of the whiskers of the boxplots, as shown in Figure 1, to minimize the number of data points lost, as we have few to begin with. An analysis of the outliers, observed in the boxplots, for all variables was made for each month and in each month for each day. The decision to remove or keep the outliers was driven by the following rational: in the

[1]Python is a widely used general-purpose, high-level programming language that is gaining wide adoption in the discipline of data analysis and machine learning

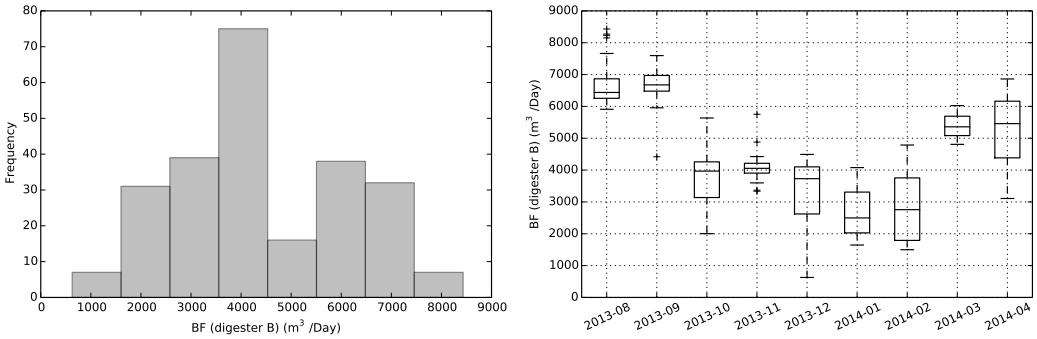

Fig. 1. Histogram and boxplot representations for the biogas flow (BF) for the digester B.

case that only one variable had outliers for the day in question, the decision to eliminate this day from the data was based on whether or not, it was considered that the observed outliers were far away from the whiskers. For the case that more than one variable had outliers, the day in question, was frequent eliminated from the data. However, if the outliers were close to whiskers of the boxes this values were still considered.

B. Data Processing

1) Correlations analysis: After eliminating the outliers, the next step of data processing was see which variables are most related to the variable we want to predict: biogas flow and/or % of methane. For this, we calculated the Pearson correlation, for each pair of variables. From the analysis of these values, it was decided that the variables % of VS, VFA and alkalinity will not be used in building the models, because they have a very few points (at most 30 values). Since there are so few points any existing method of filling missing values would likely be unreliable. Also, since the models will be build with only the days where all variables have values, when building the models, there will be the need to split the data in training and testing sets and such will be impossible to do it if this variables were used.

Thus, the variables that were used in the future work were: the input flows in the SP (IFSP) and in the digesters (IF), the biogas flow (BF), % of TS both input (pTSI) and output values (PTSO), pH, and temperature (T) for each digester, and also the % of methane (P_methane). It was also found that the relationships of the variables, for each digester, with the others variables are similar for each digester, i.e. there is no apparent difference between digesters. To facilitate future work it was decided to join the variables that were specific for each digester in a single variable.

Thus, for the information about which specific digestor some value is referred to, a new variable was created called "Digester". Next, with this variables a correlation matrix was made. One of the things observed in this matrix was that the % of methane was not correlated with any of the others variables. So a new variable was created: production of methane (P_methane) that is the multiplication of the BF with the pmethane in each day. To also improve the correlation with the variable pTSI, a new variable was created, resulting from the multiplication of the pTSI and IF for each day, called total solids loading input (TSLI). From the analysis of Table I, we can observe that, keeping in mind that a Pearson correlation coefficient of -1 means that the variables are not correlated, a coefficient of 1 means that the variables are perfectly correlated and that a coefficient of 0 means that the variables are uncorrelated: P_methane is correlated perfectly with the BF. Is much more correlated with the TSLI than with the variables that gave rise to this variable: pTSI and IF. This variable is also well correlated with IFSP and pSTO. P_methane is not correlated with the other variables (pmethane, pH and T). The biogas flow is perfectly correlated with P_methane, is well correlated TSLI, IF and with IFSP. It is also correlated with pTSI and pTSO, apparently is uncorrelated with the pmethane, pH and T; pmethane apparently is not correlated with any of the others variables; TSLI is correlated with the P_methane, with BF and with IFSP. Is also correlated although not as strongly with the IF, with the pTSI and with pTSO. With the remaining variables there are not significants observed correlations. IF is well correlated with P_methane and with BF. It is also correlated with TSLI, pTSO and with iFSP and is substantially less correlated with pTSI. It is not correlated with pmethane, pTSI, pH and T; pTSI is correlated with P_methane, BF, TSLI, pTSO and with IFSP. Is practically not correlated with others variables; IFSP is well correlated with P_methane, BF, TSLI and pTSO. It is also correlated with IF and with pTSI. Is practically not correlated with other variables; pSTO is well correlated with P_methane, BF, TSLI and with IFSP. Also is correlated with pTSI and IF, as previously mentioned. Is practically not correlated with other variables; pH and T are apparently not well correlated with any of the variables. From this analysis, the conclusions are that the most correlated variables with BF and/or P_methane are: IF, IFSP, pTSI and TSLI. Also was concluded that there is no variable that is correlated with pmethane. Was also observed that the correlations in general are more pronounced with the variables p_methane and TSLI than with the variables that gave rise to them and therefore will be used in future work.

2) PCA analysis: A PCA analysis was made with the ten variables referred before, the analysis was made with the first three PCs that can reduce the dimensionality that would be 10 dimensions (one for each variable) for only three dimensions,

TABLE I. CORRELATION MATRIX

	P_methane	BF	pmethane	TSLI	IF	pTSI	IFSP	pTSO	pH	T
P_methane	1.00	1.00	0.03	0.90	0.79	0.60	0.84	0.75	0.26	0.06
BF	1.00	1.00	-0.04	0.90	0.80	0.59	0.84	0.73	0.26	0.06
pmethane	0.03	-0.04	1.00	0.02	-0.09	0.11	0.07	0.34	0.03	0.11
TSLI	0.90	0.90	0.02	1.00	0.78	0.75	0.86	0.73	0.36	0.12
IF	0.79	0.80	-0.09	0.78	1.00	0.21	0.68	0.51	0.28	-0.11
pTSI	0.60	0.59	0.11	0.75	0.21	1.00	0.57	0.60	0.22	0.26
IFSP	0.84	0.84	0.07	0.86	0.68	0.57	1.00	0.79	0.32	0.15
pTSO	0.75	0.73	0.34	0.73	0.51	0.60	0.79	1.00	0.33	0.16
pH	0.26	0.26	0.03	0.36	0.28	0.22	0.32	0.33	1.00	0.01
T	0.06	0.06	0.11	0.12	-0.11	0.26	0.15	0.16	0.01	1.00

with the ability to describe $79,2\%$ of the total variance present in the original data set.

In the Figure 2 are the loadings plots. In the representation of PC1 versus PC2 (A.1) one can see that the less related variables (meaning furthest from) to the methane production and biogas flow are the pmethane, T, pTSI and pTSO. In the figure A.2 in the Figure 2, that is the zoom of the figure A.1, it can be seen more clearly that the variables related to the P_methane and BF are the IF, TSLI and IFSP. In the representation of PC1 against PC3, which is referred as B.1, one can once again observe that the pmethane and T with the addition, in this case, of the pH are less related variables to the variables of interest and that pTSI in this case is close to the relevant variables. In the zoom of this figure (B.2) is observed more clearly that the variables related to the variable of interest are the TSLI and the IFSP (they are, in this representation, very close to each other). Finally, in the representation of PC2 against PC3 (C.1), the variables IF, pH, pmethane and T are the farthest from the P_methane and BF and again pTSI also in this case is close to the relevant variables. From figure C.2, the closest variables to the variables of interest are the IFSP and the TSLI.

In general, with the loadings plots it was possible to observed that the variables nearest to the biogas flow/methane production were the sludge input flow in the digesters (IF), the sludge input flow at the SP (IFSP), the input % of TS (pTSI) and the load input of TS (TSLI). As noted in both correlations and principal components analysis, the variables referred above are the variables more related with the biogas, and so, was with different combinations from this variables that the predictive models were built for both PLS regression and ANN methods.

C. Development of Prediction Models

1) PLS Model: Prior to building a prediction model using the technique of neural networks, it seemed a good practice modeling the data first with the PLS method, that is an approach of linear regression, to better understand the behavior of the data, before moving to a more complex, non-linear approach as ANN method. Using, as input, the variables referred before, that are most related with the output variable, various PLS models were built. To construct these models (in both PLS and ANN methods), the methodology was: randomly divide the initial data comprising 419 points (no missing values) in training set (70 % of the initial data: 293 points) and test set (30 % of the initial data: 126 points). Then train the model with the training data, providing both input and output variables and after, test the model with the testing test providing only the input variables and repeating this process with a new training an testing sets that were obtained with a new random split of the initial data (in this case study was ten repetitions), and then by cross-validation, select the best model which is the model with the smallest error (an error measure can be, for example, the RMSE (Root Mean Square Error)). It was decided therefore, by cross-validation, the number of components to use (1,2,3 or 4, equal to the number of input variables). However, it was observed that the addition of one more component does not increase significantly the mean fraction of the variance explained by the models. As such it was decided to use only one component in all prediction models made.

The overall mean fraction of the variance explained by the different models with different input variables to predict the methane production, can be observed in the Table II. In this table the characteristics of the models, the error and the R^2 for the test part of the models can also be seen. Note that the NRMSE is the normalized RMSE ($NRMSE = RMSE/(y_{max} - y_{min})$, where y_{max} and y_{min} are the maximum and minimum values, respectively, of the observed data for the output variable). In this table is seen that is possible to built with good accuracy various models varying the input variables, however, the best model still remains as the model with the input variables: IF, IFSP, pTSI and TSLI. As an example in the Figure 3 are represented the methane production observed versus predicted for the model with the input variables: IF, IFSP, pTSI and TSLI, for a random training an testing sets.

As it can be seen, a model with a good accuracy can be built to predict the methane production. It was also built similar models to these, to predict biogas flow, % of methane and also to the methane production but using only the winter months data to train and test the model. It was concluded that with PLS method was possible to achieve good models, with good accuracy, to predict the variables: biogas flow and methane production. The same does not happens to the % methane prediction, as was not possible to built a model with satisfactory mean variance explained. Was also concluded that constructing a model to predict the production of methane only with winter months data does not bring any advantages.

2) ANN Model: Like in the PLS method, in the ANN method various models were built by varying the input variables. In the construction of these models, the following

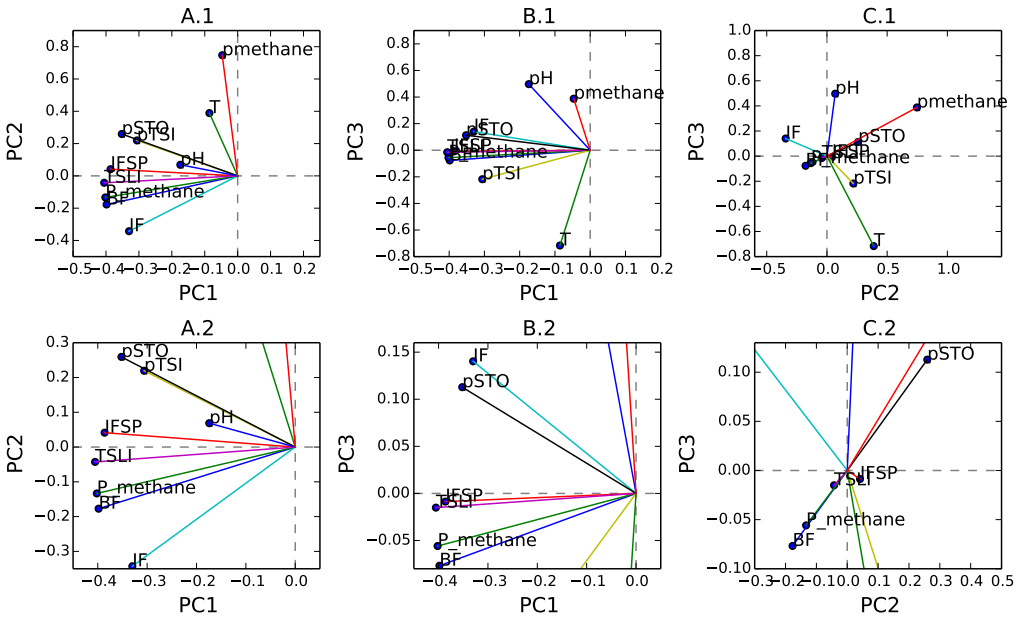

Fig. 2. Loadings plots for the first three PCs in the first line. In the second line are the same plots but with different zooms to better observe the relationships between variables that are too close from each other, in the first representations.

TABLE II. CHARACTERISTICS OF THE VARIOUS MODELS BUILT WITH PLS METHOD TO PREDICT METHANE PRODUCTION. THE ACCURACY MEASURES ARE RELATED TO THE TEST OF THE MODEL.

Input variables	Components	Mean variance explained	Mean NRMSE (%)	Mean R^2
Input flow at the SP (IFSP) Input of the % of TS (pTSI)	1	0,71	15,1	0,71
Input flow at the SP (IFSP) Input flow of the digesters (IF)	1	0,80	12,2	0,80
Input load of TS (TSLI) Input of the % of TS (pTSI) Input flow of the digesters (IF)	1	0,81	12,1	0,81
Input load of TS (TSLI) Input of the % of TS (pTSI) Input flow of the digesters (IF) Input flow at the SP (IFSP)	1	0,85	11,2	0,85

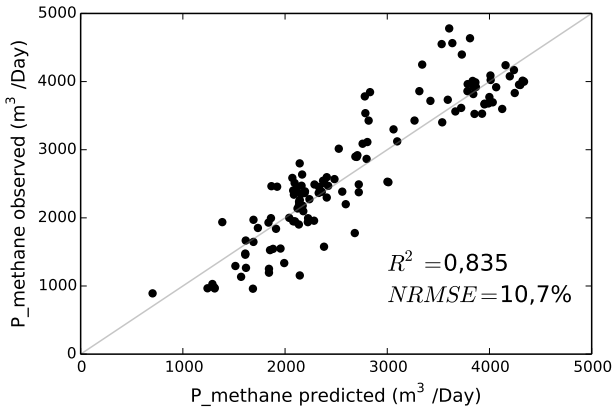

Fig. 3. Methane production observed versus predicted, for the test part of the PLS model built with one component and with following the input variables: IF, IFSP, pTSI and TSLI.

parameters were used: the activation function used was the hyperbolic tangent [12]; the training algorithm used was the limited-memory method (BFGS) [13] from the family of quasi-Newton methods [14]; The value of tolerance (tol) used was 10^{-5}; The number of hidden layers used was one, since this number is usually enough in most cases [12]; The algorithm iterates until convergence (determined by tol) or until it reaches the number of maximum iterations defined. Discovering the optimal number for this variable, as the the optimal number of neurons in the hidden layers, is basically problem-dependent [12]. Although there exists many different approaches such as the pruning algorithm [15], for finding the optimal architecture of an ANN, these methods are usually quite complex in nature and are difficult to implement. Furthermore none of these methods can guarantee the optimal solution for all real forecasting problems. To date, there is no simple clear-cut method for determination of these parameters. Guidelines are either heuristic or based on simulations derived from limited

experiments. Hence the design of an ANN is more of an art than a science [12]. With that said, for each ANN model built in this work, the approach to choose the optimal number of iterations and neurons, was the following: for each possible combination of the number of iterations and the number of neurons, the median error (RMSE) of the testing part (ten repetitions were made for each case) was written down. Then the representation of the median of the error versus the number of iterations was made, for each number of neurons.

All curves observed that corresponded to a certain number of neurons, had the same behavior: there was an interval where the error drops significantly after which the error was almost stable. The number of optimal iteration chosen was the number from which the error was almost stable, was decided to select this number and not the iterations number with the lowest error because were considered that the increase of the complexity was not worth for the little difference between the lowest error and the error for the chosen iteration.

Then median error versus number of neurons was represented only for the number of iterations chosen. We also observed the behavior described above and so, again, the number of the neurons chosen was the number from which the error was almost stable. The use of median instead of the mean results from the fact that as the weights of the networks are initialized randomly, certain network are initiated far from the weights which can be used to accurately predict the outputs. As there are always some of these cases in some of the ten repetitions, they end up having a great influence in the mean, as it is not robust measure. The median is not so easily influenced by these extremes and as such the metric chosen was the median, which is a more robust measure.

Just as in the construction of PLS models, we built several models with different combinations of the variables that were selected as the most related to the output. It was decided to normalize the inputs to be determined whether the models became better. The characteristics (iterations, neurons and the mean NRMSE and R^2 for the testing part of the construction of the model) are presented in Table III. From the analysis of table, it can be seen that generally all models have better results when the inputs are normalized, wherein, in some models the improvements are very small but in models when the IFSP is an input variable these improvements are quite significant. This is due to the magnitude of this variable which is relatively larger than the remaining variables. It can then be seen that all models with the inputs normalized give satisfactory results with errors between $9,86\%$ and $11,9\%$ and R^2 values between $0,802$ and $0,861$. At last, it can also be observed, that the best model obtained, like in the PLS method, is with the input variables: IF, IFSP, pTSI and TSLI, with the input variables normalized.

Next was decided to investigate whether it was possible to improve the performance of this network. To this end, we tried adding to this model the pH and temperature as inputs, both one at a time and together at the same time. It was observed that these variables do not add any significant improvements in the performance of the model. We also tried modifying some parameters in the best model in order to ascertain whether they produce improvements in the performance of the model. These parameters were: the activation function and the number of hidden layers. It was concluded that using the hyperbolic tangent function as the activation function instead of log-sigmoid activation does not result in improvements in the performance of the model. Similarly, the introduction of a second hidden layer was also attempted but did not bring any benefits. After having tried various combinations of different input variables and changing certain parameters it was concluded that the best prediction model with ANN was with first model obtained with the input variables: IF, IFSP, pTSI and TSLI, with the inputs normalized.

This model, that was trained an tested with the data between August 2013 and April 2014 was used to predict the methane production from May, June and July 2014. Recall that this data was not used in the training/testing parts of the model. The results from this predictions are in Table IV. In this table we can see that the prediction of the months individually results in bad performance. We also eliminated the outliers from May 2014, related with the end of the interventions that started in February to try to improve the performance of its prediction. It can be seen that the results without this points are a little better. As already mentioned, in Table IV it is possible to see that the prediction of the months individually results in bad performance but when May and June were predicted together the performance of the prediction was good and even further, the prediction of the three months together have even better results. This may be due to the fact that with more than two months the mean of these values together, compared to the mean values of the individual months, are closer to the mean values of the training set, and so, the model can predict better. In Figure 4 are represented, the observed and the predicted values, for the methane production for the training with the data from August 2013 to April 2014 and the observed and the predicted values for the months of May, June and July 2014.

It can be observed that this values are represented individually for each digester. Note that the information on the digester was left out when the model was trained and validated. It was decided to represent by digester just to be able to have on the abscissa axis a period of time, otherwise, with the all points represented in the same figure, one could not so easily analyze certain behaviors observed during a period of time such as a month. In this figure, it can also be observed that the real and the predicted values are similar.

In the summer months, it was observed both in training and in the validation, that the prediction have the shape of a horizontal line (which corresponds to the mean value of methane production in the summer months on the training set, i.e., the mean of the months of August and September 2013). Despite being a flat line, the observed values, in this months, are not far from the ones predicted and as such the error is not very high. Another thing that can be noticed is that, for the prediction of the data of June and July for the the digester B, at first glance, it seems that the prediction achieved better results when compared to the others digesters for the same period. However, in a more careful analysis, one can observe that in this digester, in these months the observed values have more variance than in the others digesters and thus it seems

likely that the prediction, for this months in this digester, can have a slight difference in appearance from the others.

In conclusion, the prediction of the B digester for June and July is not exactly a straight line, but one can see that despite having highs and lows, the prediction is very close to the line observed for the predictions of the same months for the other digesters, maintaing thus the conclusion that the digesters have a very similar behavior. In an attempt to improve the prediction for a month individually, it was decided to re-train the best ANN model obtained earlier, adding the months of May and June to the training and testing sets, and then validate this model with the month of July. The optimal number of iterations and neurons in the hidden layer was found with the same method described earlier. The results for the prediction of the month of July were, however, not better. The error associated with this prediction, was still substantial (about 22%).

As noted in both pre-processing of the data and in the PCA analysis, the data has seasonality. So, a model with the same variables as the best model obtained earlier, but trained and tested with only data from summer months (August 2013, September 2013 and June 2014) was built and the the month of July 2014 was predicted. However, the results showed that using only the summer months yields predictions with worse results than using all the months (from winter and summer).

V. Conclusion and Recommendations

This work had the goal of building a model to predict the methane production, and was carried out with PLS (Partial Least Squares) and ANN (Artificial Neural Networks) methods. Before building these models a pre-processing, where the outliers were eliminated, was carried out.

Next, a correlations analysis and PCA were performed. In these analysis it was concluded that the variables most related with the output were: the sludge input in the digesters, the sludge input in the Solid Phase of the WWTP in study, the % input of TS in the digesters and the input load of TS also in the digesters. And was with this variables (input variables) that the models were built.

Both training and testing results, for both methods are quite similar: $R^2 = 0,85$ and $NRMSE = 10\%$. We also experimented with varying the input variables, but in both methods the best model still remained the model with this variables. With the ANN model with these variables, and trained with the data from August 2013 to April 2014 the predictions for the months of May, June and July 2014 were carried out. From the results of this predictions, it was concluded that the prediction error associated with the prediction of only one month is about 22%, however if tested with two months, May and June, the error is reduced to about 14% and, even better, predicting three months of data, results in an even further reduced error associated with the prediction, of 12%.

Concluding, one can have a better prediction, when using more data, and it is possible to predict only a month data, by having in mind that this prediction can have an error of about 20%. Thus it can be concluded that, apparently, the input variables that were used cannot describe all the variance present in the output variable, and as such, requires further investigation of other variables that may be related to the output and that could help explain its variance. It was also concluded that the three digesters have very similar behavior and it is possible to build only a model for predicting the methane production/biogas flow, with good results, for the three digesters simultaneously.

It was also noted that a clear seasonality exists in the data with summer months having higher values than the values recorded in the winter months, for all variables. Some adversities in this work were the existence of interventions during the sampling time, which created extra noise and may have influenced the results.

In general it is possible to conclude that good predictive models can be obtained for the production of methane/biogas. It is also possible to conclude that models with good results ($R^2 = 0,790$ and $RMSE = 11,59\%$) can be built to predict data that was not used in the training/testing sets.

A recommendation for future work is to investigate other variables potentially related to the variable of interest, that can be used to better describe the variance in the output. It would also be interesting to evaluate the relationship of certain variables that were available only in small number and could not be used such as alkalinity, VFA and % of VS.

However, in the case of % of VS, the correlation analysis indicated that this variable is closely related to the % of TS so we suspect that its inclusion would not bring much more predictive power than ST. It would, however, be interesting to confirm this suspicion with results.

Although there is not strong evidence that adding more data from summer months will improve the accuracy in summer months, it would also be of interest to investigate its effect. The continuous introduction of new data and re-training/testing is also suggested as future work.

References

[1] A. Khalid, M. Arshad, M. Anjum, T. Mahmood, and L. Dawson, "The anaerobic digestion of solid organic waste," *Waste Management*, vol. 31, no. 8, pp. 1737–1744, 2011.

[2] A. Hilkiah Igoni, M. Ayotamuno, C. Eze, S. Ogaji, and S. Probert, "Designs of anaerobic digesters for producing biogas from municipal solid-waste," *Applied energy*, vol. 85, no. 6, pp. 430–438, 2008.

[3] A. Sulaiman, A. M. Nikbakht, M. Tabatabaei, M. Khatamifar, and M. A. Hassan, "Modeling anaerobic process for wastewater treatment: new trends and methodologies," in *Proceedings of International Conference on Biology, Environment and Chemistry (ICBEC 2010)*, 2010.

[4] P. Holubar, L. Zani, M. Hager, W. Fröschl, Z. Radak, and R. Braun, "Start-up and recovery of a biogas-reactor using a hierarchical neural network-based control tool," *Journal of Chemical Technology and Biotechnology*, vol. 78, no. 8, pp. 847–854, 2003.

[5] D. P. Strik, A. M. Domnanovich, L. Zani, R. Braun, and P. Holubar, "Prediction of trace compounds in biogas from anaerobic digestion using the matlab neural network toolbox," *Environmental Modelling & Software*, vol. 20, no. 6, pp. 803–810, 2005.

[6] P. Holubar, L. Zani, M. Hager, W. Fröschl, Z. Radak, and R. Braun, "Advanced controlling of anaerobic digestion by means of hierarchical neural networks," *Water Research*, vol. 36, no. 10, pp. 2582–2588, 2002.

[7] H. Abu Qdais, K. Bani Hani, and N. Shatnawi, "Modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm," *Resources, Conservation and Recycling*, vol. 54, no. 6, pp. 359–363, 2010.

TABLE III. CHARACTERISTICS OF THE VARIOUS MODELS BUILT WITH ANN METHOD TO PREDICT METHANE PRODUCTION. THE ACCURACY MEASURES ARE RELATED TO THE TEST OF THE MODEL.

Input variables	Normalized Inputs variables	Iterations	Neurons	Mean NRMSE (%)	Mean R^2
Input flow at the SP (IFSP)	X	1500	11	26,7	-0,003
Input of the % of TS (pTSI)	✓	1000	50	11,9	0,802
Input flow at the SP (IFSP)	X	1300	60	24,9	0,124
Input flow of the digesters (IF)	✓	1000	50	10,8	0,834
Input load of TS (TSLI) Input of the % of TS (pTSI)	X	1500	65	11,2	0,827
Input flow of the digesters (IF)	✓	1000	35	10,4	0,849
Input load of TS (TSLI) Input of the % of TS (pTSI)	X	1100	50	18,6	0,494
Input flow of the digesters (IF) Input flow at the SP (IFSP)	✓	600	30	9,84	0,861

TABLE IV. PERFORMANCE MEASURES FOR THE PREDICTION OF THE MONTHS OF MAY, JUNE AND JULY USING THE BEST ANN MODEL OBTAINED.

Month/months to predict	NRMSE (%)	R^2	Nos. of data without missing values
May	26,91	0,123	66
May without outliers	22,56	0,513	45
June	18,14	0,245	66
July	22,50	-0,182	60
May without outliers and June	13,68	0,767	105
May without outliers, June and July	11,59	0,790	162

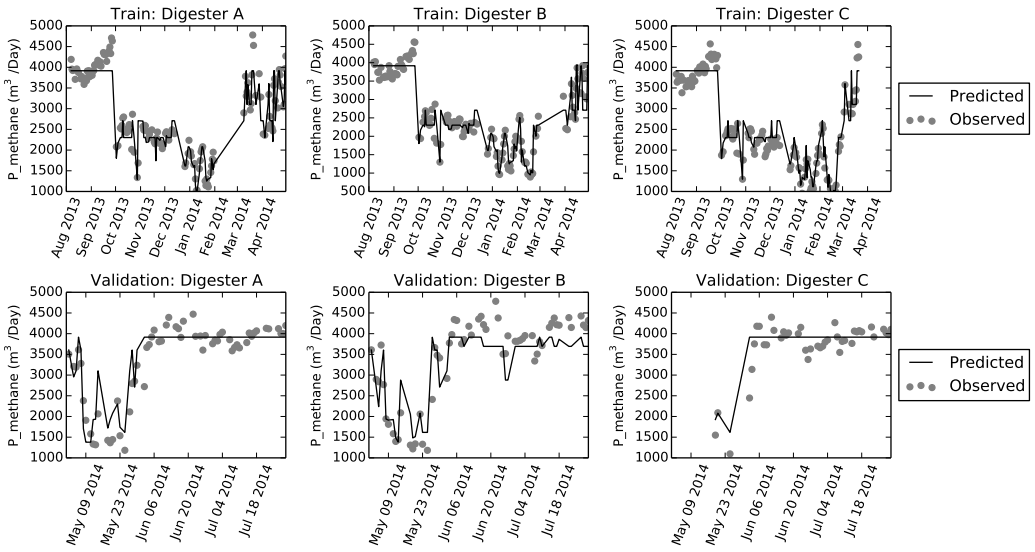

Fig. 4. Values of the methane production, observed and predicted, for the training of the model with data between August 2013 and Abril 2014, and for the validation, with the months of May, June and July 2014 for each digester. The ANN model used had as input variables: TSLI, pTSI, IF and IFSP and was built with 30 neurons and with 600 maximum iterations.

[8] L. Egghe and L. Leydesdorff, "The relation between pearson's correlation coefficient r and salton's cosine measure," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 1027–1036, 2009.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 0090–95, 2007.

[11] J. Han and M. Kamber, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.

[12] G. Zhang, B. Eddy Patuwo, and M. Y Hu, "Forecasting with artificial neural networks:: The state of the art," *International journal of forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[13] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[14] C. Broyden, "Quasi-newton methods and their application to function minimisation," *Mathematics of Computation*, pp. 368–381, 1967.

[15] G. Castellano, A. M. Fanelli, and M. Pelillo, "An iterative pruning algorithm for feedforward neural networks," *Neural Networks, IEEE Transactions on*, vol. 8, no. 3, pp. 519–531, 1997.