COMPARATIVE STUDY OF PCA-ANN HYBRID INFERENTIAL SYSTEMS IN NATURAL GAS PROCESSING UNITS

Leandro Luttiane da Silva Linhares, luttiane@dca.ufrn.br
Fábio Meneghetti Ugulino de Araújo, meneghet@dca.ufrn.br
Computation and Automation Engineering Department
Federal University of Rio Grande do Norte, Brazil

Abstract. *The main product of the Natural Gas Processing Unit (NGPU) studied in this work is the liquefied petroleum gas (LPG). The LPG is ideally formed by propane and butane, however, in practice, this also has in its composition some contaminants, such as pentane and ethane. The LPG quality control is done through its chemical composition, however, chemical compositions are traditionally known as variables of difficult measurement. The instruments used to measure these variables, such as gas chromatographies, are expensive and have long intervals of measurement, what turns difficult the development of more efficient control strategies. A way to reduce this problem is to use secondary process variables to infer the chemical compositions in shorter time intervals. Systems that perform this task are know in the literature as inferential sensors or inferential systems. This paper presents a comparative study of four hybrid inferential systems. These systems use the techniques of principal component analysis and artificial neural networks to estimate, in each minute, the ethane and pentane molar fractions in LPG and the propane molar fraction in the residual gas. In this work, a part of a NGPU process formed by a deethanizer and a debutanizer column is simulated on HYSYS® software.*

Keywords: *hybrid inferential system, principal component analysis, neural networks, natural gas processing.*

1. INTRODUCTION

In natural gas processing units (NGPUs) the raw natural gas passes through an initial preprocessing stage, where the water and oxidizing elements are removed. Then the gas is sent to distillation columns, where it is decomposed into various subproducts such as the residual gas, the natural gasoline (C5+) and the liquefied petroleum gas (LPG). The columns demethanizer, deethanizer, depropanizer and debutanizer are examples of distillations columns that can be found in NGPUs.

The NGPUs are complex processes and its configurations depend on the chemical characteristics of the natural gas which is being processed and on the production goals of the processing unit. The real NGPU adopted as the basis of this work consists of a column deethanizer in series with a debutanizer column. The main product of the process in study in this work is the liquefied petroleum gas (LPG). The mentioned columns were computationally simulated in HYSYS®, a software for chemical processes simulation.

The chemical compositions are rarely used directly as controlled variables in quality control strategies of the subproducts of a distillation column, because these variables are difficult to measure. The measurements of these important indicators of process performance and product quality are often obtained through sample analysis in laboratories. This methodology results in large measurement delay, hindering that the necessary adjustments to maintain the behavior of the process according to the desired occur at the right time. Thus, one can arrive at a situation where the final product will be out of specification, causing an unwelcome economic loss.

There are also devices that can measure the composition analysis on the production line, such as gas chromatographies. However they are expensive to purchase and maintain, and present significant time intervals between the measurements. This last feature is the major restriction to implement more efficient control strategies in distillation columns processes.

According to Zamprogna *et al.* (2005), the inferential systems, also known as inferential sensors or soft sensors, are an attractive way to address the problem of measuring the primary variables of a process, particularly when physical sensors to measure these variables are not available, or when the high costs and/or technical limitations of these devices prevent its use in real time. In these systems, the primary variables of the process are estimated from secondary variables easy to measure, such as temperatures, pressures, levels, flows, among others.

In this work, it is presented an analysis of the use of artificial neural networks (ANN) in conjunction with principal component analysis (PCA) to implement hybrid inferential systems. The neural networks are widely used to develop these systems. Its application to estimate chemical compositions has been reported in different kind of processes (Bo *et al.*, 2003; Chella *et al.*, 2006). Likewise, we can also find works in the scientific literature that present PCA being applied with artificial intelligence techniques to implement inferential sensors (Warne *et al.*, 2004a,b; Linhares *et al.*, 2008).

This paper presents a comparative study of four ways to combine PCA and ANN techniques. From the analysis of different configurations, important observations as advantages and disadvantages of each system can be made regarding the combination of these techniques when used to implement inferential sensors. The structures analysed in this work are called PCA-ANN inferential systems.

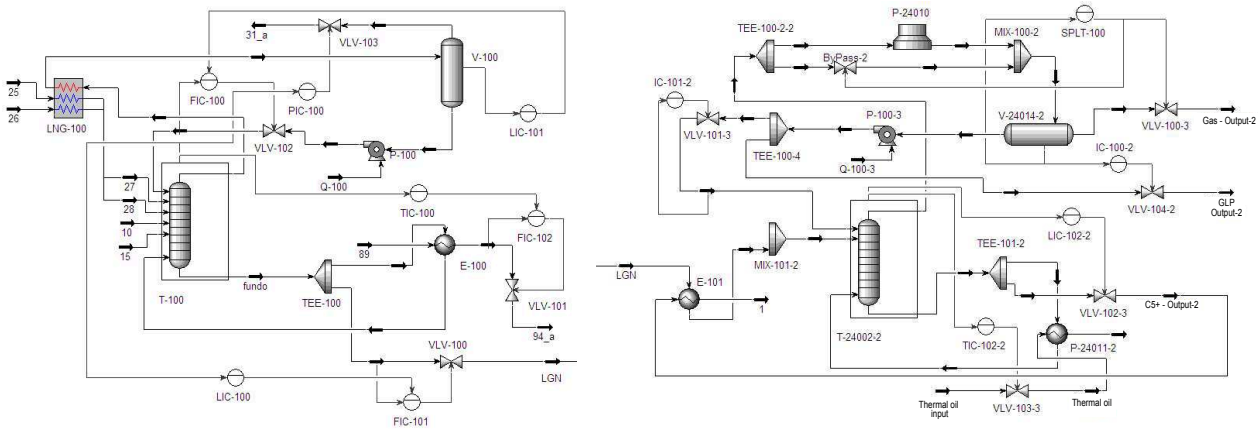
The goal of the inferential systems analyzed in this paper is to estimate the molar fractions of contaminants ethane (C2) and pentane (C5) in LPG and the mole fraction of propane (C3) in the residual gas. The last one represents a loss indicator of the NGPU taken as basis to develop the simulated plant used in this work. The inference of these variables is performed every minute by a multilayer perceptron (MLP) neural network. The PCA is applied to reduce the number of inputs of the ANN without loss of information and performance, reducing the network complexity.

2. PROCESS SIMULATION

Some important stages of a NGPU were simulated using the HYSYS® software to analyse the four PCA-ANN inferential systems proposed. The process simulation was implemented based on a real NGPU formed by a deethanizer column in series with a debutanizer column. After an initial removal process of water and oxidants from the natural gas it is forwarded to these fractional distillation columns where the main products of the process are extracted.

The deetanizer is the first column of the process simulation. It receives the preprocessed natural gas and by distillation gets on its top the residual gas, consisting mainly of methane and ethane. The main product of this column, the liquid natural gas (LNG), feeds the next simulation stage: the debutanizer column. In this last step are extracted the natural gasoline and LPG, respectively, the botton and top products of the debutanizer column.

Figure 1 ilustrates the schematic diagrams of the deethanizer and debutanizer columns as well the PID controllers and others instruments on process simulation.



(a) Simulated deethanizer column. (b) Simulated debutanizer column.

Figure 1. Schematic diagram of the process simulation in HYSYS®.

The LPG is the most important economic product of the NGPU taken as basis of the process simulation, being compounded ideally by propane and butane. However, in practice, the producted LPG has some contaminats in its composition as ethane and pentane. These contaminants must be controlled to maintain the final LPG composition according to quality specification laws and to ensure higher production profits. In this work the inferential systems in study estimate the ethane and pentane molar fractions in LPG as well the estimated propane molar fraction in residual gas. The reduction of C3 loss in residual gas results in a C3 concentration increasing in LNG. As consequence of it the debutanizer column will present as its final product a LPG richer in C3.

3. INFERENTIAL NEURAL MODEL

An inferential system has to adequately represent the dynamic relationships between secondary variables used by the system and the primary variables adopted to be estimated. To achieve this goal, it is necessary that the system describes a dynamic model that represents these relationships with a satisfactory level of accuracy.

In this way, performing inference using ANN can be seen as an identification problem, since the neural network applied have to be able to effectively represent the dynamics between the secondary and primary variables of the process under study. One of the main advantages of using neural structures for identification and/or for inference is its ability to represent even the nonlinear dynamics based only on experimental measured data.

The identification procedure requires an initial model structure selection to be used. In the family of neural networks multilayer perceptron (MLP), the identification models most used are NNFIR, NNARX, NNARMAX, NNOE and NNSSIF, all of them are based on its respective traditional linear model structures. More details about these models can be found in Nørgaard *et al.* (2001).

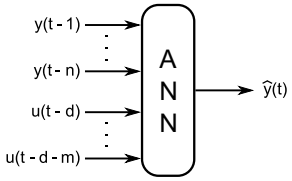


Figure 2. NNARX model structure.

The model used as the basis for the inferential systems proposed in this work is the model NNARX (Neural Network AutoRegressive with eXogenous inputs). Figure 2 shows, in general, this structure, where ANN is the multilayer perceptron, n and m are, respectively, the regressors applied to the input u and output y and d the transport delay.

4. PRINCIPAL COMPONENT ANALYSIS

According to Salahshoor *et al.* (2009), PCA is a useful statistical technique that has found application in different fields to find latent patterns in high dimensional data. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences.

The PCA aims to map a system made up of p correlated variables into uncorrelated linear combinations k ($k < p$), called principal components. Jolliffe (2002) affirms that the PCA main idea is to reduce the dimensionality of a data set formed by correlated variables, keeping as much as possible of the variance of the original data set.

The k principal components that represent the original data set system can be obtained using the covariance or correlation matrixes. The decision on which matrix to use is usually made according to the discrepancy caused by the different measurement units of the original variables (Mingoti, 2005).

Considering the use of the correlation matrix, the PCA calculations can be summarized by the following steps (Salahshoor *et al.*, 2009):

- **Step 1**: Get the experimental data $X = (X_1 \;\; X_2 \;\; \ldots \;\; X_p)'$.
- **Step 2**: Normalize the random variables X_1, X_2, ..., X_p to zero mean and unit variance.
- **Step 3**: Calculate the correlation matrix $\mathbf{S}$.
- **Step 4**: Calculate the eigenvectors $e_1,\ e_2,\ \ldots,\ e_p$ and the respective eigenvalues $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_p$ of the correlation matrix. The coefficients of the j-th main component are the elements of the eigenvector e_j as demonstrated in Eq. 1, while the eigenvalue λ_j represents the variance of this component.
- **Step 5**: Sort and choose the appropriate principal components, forming a feature vector. In general, once eigenvectors are found from the correlation matrix, the next step is to sort them from highest to lowest eigenvalues. This gives the components in order of significance. Now, it is possible to ignore the components of lesser significance.
- **Step 6:** Derive the new data set. This is the final step in PCA transformation. Once the significant components that are going to be kept in the data are selected and hence the feature vector is formed, it is simply needed to take the transpose of the vector and multiply it by the original data set.

The k linear combinations (principal components) chosen to represent the original data set X are directly related to the total variance of the system, being chosen according to the eigenvalues $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_p$. Since $e_1,\ e_2, \ldots,\ e_p$ are the normalized eigenvectors, the j-th principal component (PC) is defined by:

$$\widehat{Y}_j = e_{j1}\, X_1 \;+\; e_{j2}\, X_2 \;+\; \ldots \;+\; e_{jp}\, X_p \tag{1}$$

The PCs are sorted in descending order according to their variances, or sorted from highest to lowest eigenvalues. Jolliffe (2002) says that many of the selection rules used to find the number k of principal components are not strictly accurate. A widely used criterion is to select a number of components k which together represent a percentage γ of the total variance of the problem. Thus, it seeks the smallest integer value of k such that satisfies:

$$\frac{\sum_{i=1}^{k} \widehat{\lambda}_i}{\sum_{j=1}^{p} \widehat{\lambda}_j} \geq \gamma \tag{2}$$

According to Jolliffe (2002), γ corresponds to a cutoff point and is usually selected from a region between 70% and 95%, depending on the application characteristics and requirements.

5. PCA-ANN HYBRID INFERENTIAL SYSTEMS

The quality control in chemical industries requires the implementation of monitoring networks using high cost online measurement devices and, when possible, appropriated models that produce real time estimates of unmeasured variables on the basis of measurement data available (Fortuna *et al.*, 2007). The inferential systems fall in these latter case.

The inferential systems analyzed in this paper combine the ANN and PCA features. These systems differs in the way these techniques are applied to the available process data. The goal of the proposed inferential sensors is to estimate in every minute the ethane and pentane molar fractions in LPG, as well as the propane molar fraction in residual gas. In this way, these information about chemical compositions are considered to our study the primary process variables (VP).

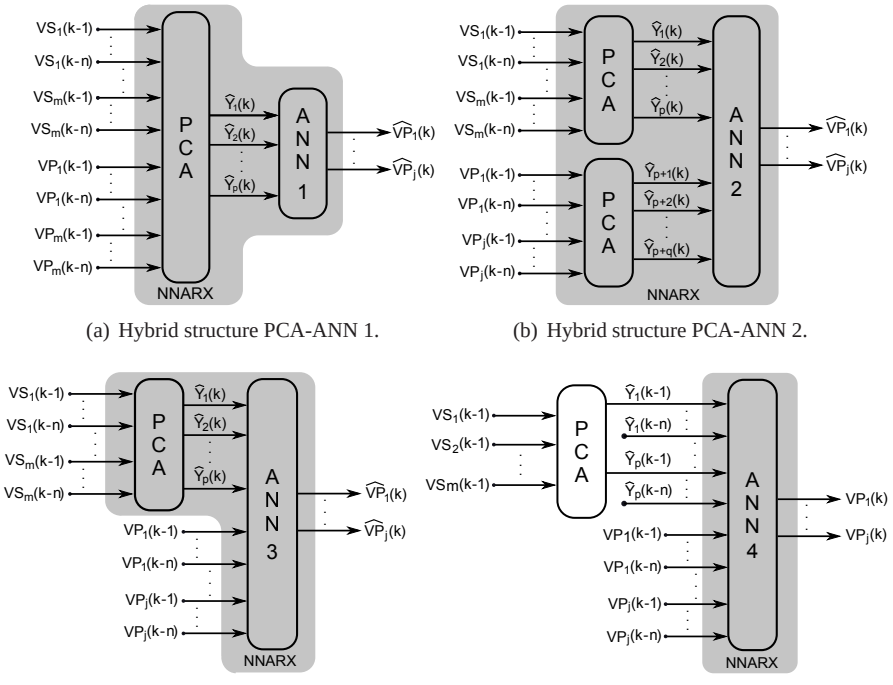
A common practice is to use temperatures of the distillation columns trays to estimate the chemical composition of its main products. However, due to a lack of temperature sensors on the trays of the columns in the process taken as the basis of this work, all PID process variables that somehow affect the primary variables dynamics were chosen as secondary variables of all proposed inferential systems. These secondary variables (VS) adopted are listed in Table 1 along with their respective PID controller and in which column they are obtained.

Table 1. Chosen secondary variables.

j	Secondary variable (VS_j)	Column	PID Controller
1	Top pressure	Deethanizer	PIC-100
2	Reflux flow	Deethanizer	FIC-100
3	Tray 40 temperature	Deethanizer	TIC-100
4	Output flow	Deethanizer	FIC-101
5	Tray 16 temperature	Debutanizer	TIC-102-2
6	Tray 28 liquid volume	Debutanizer	LIC-102-2
7	Reflux flow	Debutanizer	FIC-101-2
8	Condensated level	Debutanizer	LIC-100-2

The number of selected secondary variables and the presence of some of their past values in the identification model NNARX turns the neural network inputs number of the PCA-ANN structures relatively high. To reduce this number the statistical tool principal component analysis is applied. The goal is to reduce the complexity of the inferential systems, without impair the quality of the estimates of the primary variables.

Figure 3 shows the schematics diagrams of the four inferential systems analysed in this work. In shade, it is possible to note the part of the systems configuration that resembles the NNARX identification model. Didactically, we can divide each system in a PCA module and an ANN module.



(a) Hybrid structure PCA-ANN 1.

(b) Hybrid structure PCA-ANN 2.

(c) Hybrid structure PCA-ANN 3.

(d) Hybrid structure PCA-ANN 4.

Figure 3. Schematic diagrams of the PCA-ANN hybrid inferential systems.

The inferential system PCA-ANN 1 is composed of a PCA module that has as its inputs past values of primary and secondary variables. Thus, the goal is to minimize the complexity of the ANN module. Depending on the quality of the inference performed, this configuration allows the system to be used or tested with high order models, since the ANN module input is formed only by the k principal components selected.

In the inferential system PCA-ANN 2, as well as in the previous structure, the goal is to achieche a high reduction of the ANN module complexity. However, with this structure it's possible to perform an analysis of the PCA module associative capacity. As illustrated in Figure 3, the inferential system PCA-ANN 2 is composed by two distinct PCA modules. The PCA module at the top of the diagram is responsible for filtering the information of the secondary variables, while the module at the bottom filters the primary variables information.

The inferential structure PCA-ANN 3 was designed with the aim of analysing the importance of the primary variables past values information for the ANN module. Thus, this inferential system is formed by a single PCA module that has in its inputs secondary variables past values. Compared with the previous structure, the PCA module responsible for filtering information related to the primary variables was removed. The past values of these variables are used directly as inputs of the neural network, turning the ANN module complexity larger than in the first two structures.

The ANN module of the inferential system PCA-ANN 4 is the one that most closely matches the NNARX model. The cited module receives as inputs the past values of the principal components extracted from secondary variables and the past values of the primary variables. The PCA module of this inferential structure only reduces the information of the current secondary variables values, not more their past values. The NNARX model used in this system, in theory, causes a better assimilation of the dynamics between the primary and secondary variables from the network, reducing the estimation error. On the other hand, the complexity of the ANN module will significantly increase in relation to the other structures analysed in this comparative study.

6. SIMULATION RESULTS

Firstly, it's important to choose and define some criteria to evaluate the proposed hybrid inferential systems. The comparison of these structures will be held regarding the PCA module complexity reduction ability, the neural network structure and the reliability of the primary variables inference. These criteria will be analysed, respectively, according to the following: reduction rate(R_r), number of neural network synaptic connections (N_{sc}) and mean squared error (MSE). There is a relationship between R_r and N_{sc}, since the greater the reduction of the ANN inputs provided by the PCA module, the lower the number of ANN synaptic connections.

The reduction rate is used as a way to compare the efficiency of the PCA modules regarding their reduction ability. According to the previously presented inferential structures it's possible to define the reduction rate from the model order and the number of inputs and outputs (principal components) of PCA modules as:

$$R_r = 1 - \frac{k}{n(V_s + V_p)} \tag{3}$$

In Eq. 3, k is the number of principal components, n the model order and V_s and V_p are the numbers of secondary and primary variables, respectively, that made up the PCA module input.

The first practical step to develop the proposed inferential systems is to collect experimental samples of the simulated process. The goal is to use the data set obtained to select the principal components and identify the dynamical relationships between primary and secondary variables.

The experimental data were collected by applying PRS (Pseudo Random Signal) signals on the set points of the PID controllers related to the selected secondary variables (see Tab. 1). With this procedure, it was possible to provide changes in the process variables (PVs) of these controllers and, consequently, in the mole fractions of propane in the residual gas, and of ethane and pentane in LPG. Applying this methodology a set of 3,000 training samples and five other small sets of 400 validation samples were obtained.

Table 2 presents a brief summary of the principal components analysis performed for each inferential systems proposed. The data set of 3,000 samples was used to obtain these results. For each of proposed structure is presented the cut-off limit γ, the model order n, the principal components number k (outputs of the PCA modules), the number of PCA modules inputs - $n(V_s + V_p)$ - and the reduction rate.

The inferential system PCA-ANN 2 has two PCA modules, so on the fourth column of Table 2 are given two values: the first related to the PCA module that adresses the secondary variables data, and the other to the PCA module that treats the primary variables data.

The PCA-ANN 4 is a special case when talking about the model order, since the regressors are applied to the selected principal components and not to the secondary variables and/or primary variabels. The input of this structure is composed of the immediate past values of the secondary variables, thus it is considered the "model" formed by a "first order" PCA module ($n = 1$).

From the presented, it is clear that with the increasing of the model order the reduction efficiency of the module PCA also increases, resulting in a decreasing of the ANN modules complexity. It can also be noted that the selected principal

Table 2. PCA modules comparison.

PCA-ANN	γ	n	k	$n(V_s + V_p)$	R_r
1	95%	4	10	44	0,77
		3	9	33	0,73
		2	9	22	0,59
2	95%	4	9 e 3	32 e 12	0,72 e 0,75
		3	8 e 3	24 e 9	0,67 e 0,67
		2	7 e 3	16 e 6	0,56 e 0,50
3	95%	4	8	32	0,75
		3	8	24	0,67
		2	7	16	0,56
4	95%	1	7	8	0,12
	84%	1	5	8	0,37
	75%	1	4	8	0,50

component numbers for the first three structures, despite the model order under study, are almost the same. This is an indication that the use of high model orders to represent the process dynamics are unnecessary.

Initially, the γ was set at 95%, with the aim of extracting a large amount of information of the original data. With this cutoff value was possible to obtain a good data reduction for the first three PCA-ANN inferential systems. However, this not happened with the fourth structure. Due to this reason, the γ value was reduced to 75 % on the PCA-ANN 4 structure.

In this work, the training algorithm used was the scaled conjugate gradient. In conjunction with this algorithm, we used the early stopping techinique to avoid overfitting or overscaling. The ANN training were performed using the MATLAB® neural network toolbox.

Since there isn't a technique to define with precision the number of neurons and layers a network must have to better solve a problem, several neural networks were trained formed by a single hidden layer and having different neurons numbers. In this way the number of hidden layers was fixed and its number of neurons was defined in a trial and error procedure. We adopted neurons with sigmoid functions in the hidden layer and linear activation functions to the three output neurons of the networks.

After making the proper adjustment of the training data according to each of inferential systems proposed the training procedure was realized. Then the validation data sets were presented to the trained networks. The Table 3 presents the validation results of the best ANN found for each of inferential system under study. In this table are presented the system order (n), the number of hidden neurons (N_{hn}), the validation mean squared error and average percentual error of ethane (E_{et}), propane (E_{pr}) and pentane (E_p) molar fractions. The value in parenthesis is the cutoff point used with the principal component analysis. It is important to note that the outputs of these ANNs are the outputs of the inferential systems proposed, therefore, the results presented in Table 3 are in fact the results of each of the proposed inferential systems.

Table 3. ANN validation results.

System	n	N_{hn}	**MSE**	E_{et}	E_{pr}	E_{pe}
NNARX	2	20	2,67e-08	-0,00	0,07	0,06
	3	40	1,67e-08	0,02	0,06	-0,05
	4	46	1,38e-08	-0,02	0,08	0,03
PCA-ANN 1 (95%)	2	10	7,00e-06	-1,51	0,32	-1,15
	3	15	8,40e-06	-1,66	0,55	-0,06
	4	10	3,01e-06	-0,39	0,84	0,24
PCA-ANN 2 (95%)	2	15	2,69e-06	-0,32	0,80	0,93
	3	12	2,25e-06	-0,56	0,58	0,56
	4	16	1,49e-06	-0,07	0,35	0,43
PCA-ANN 3 (95%)	2	16	5,98e-08	-0,07	0,01	-0,05
	3	22	3,65e-08	0,00	0,06	0,02
	4	20	2,98e-08	-0,02	0,04	-0,07
PCA-ANN 4 (84%)	4	28	1,91e-08	0,01	0,03	-0,08
PCA-ANN 4 (75%)	4	24	3,32e-08	-0,01	0,02	-0,01

Table 3 also presents the results of an inferential system based on a NNARX identification model. From the designing processes of the NNARX system and the PCA-ANN systems 1, 2 and 3 was possible to note that the best representations of the process dynamics were obtained by fourth order models. Due to this only the fourth order models were used with the inferential system PCA-ANN 4.

Table 4 shows a summary of the characteristics of the best fourth order inferential systems obtained and presented in Tab. 3. To ensure the credulity of the comparison of the training time T_t, the training sessions were performed on the same computer whith the same processing conditions. In this table N_ϕ is the number of inputs of the ANN module.

From Tabs. 3 and 4 note that all PCA-ANN provided a reduction of the neural network when compared with the NNARX inferential system.

Table 4. ANN characteristics ($n = 4$).

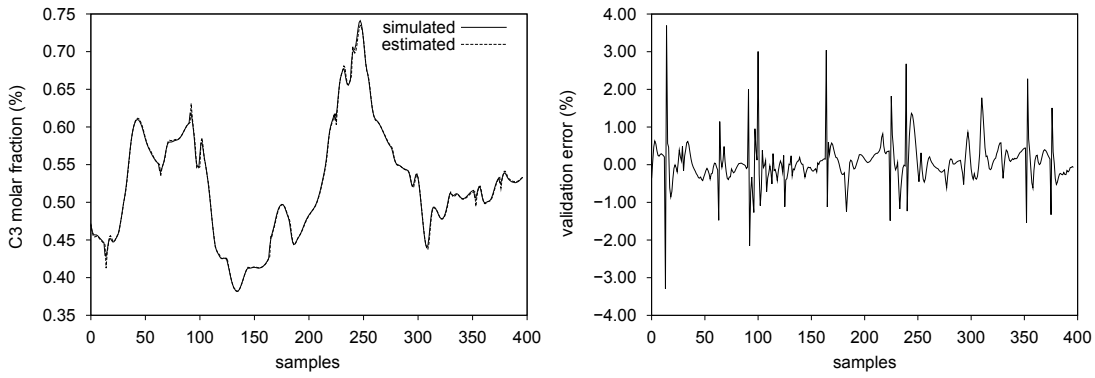
System	k	N_ϕ	N_{hn}	N_{sc}	T_t (s)
NNARX	0	44	46	2162	106.80
PCA-ANN 1 (95%)	10	10	10	130	9.70
PCA-ANN 2 (95%)	9 e 3	12	12	150	13.03
PCA-ANN 3 (95%)	8	20	20	460	28.80
PCA-ANN 4 (84%)	5	32	28	980	60.96
PCA-ANN 4 (75%)	4	28	24	744	42.88

The neural networks of the PCA-ANN 1 and 2 are much less complex than the networks of the others inferential structures. However the estimates provided by these systems are not considered satisfactory. The best estimation results were obtained by the structures PCA-ANN 3 and 4. The performance of these systems was very close to the NNARX inferential system.

Drawing a comparison between the PCA-ANN systems 3 and 4, if we consider only the MSE shown in Table 3, we can note that the structure PCA-ANN 4 ($\gamma = 84\%$) has a slightly better performance. However, the goal is to find an inferential structure that combines efficiency with simplicity, the ANN module complexity must be also taken into account.

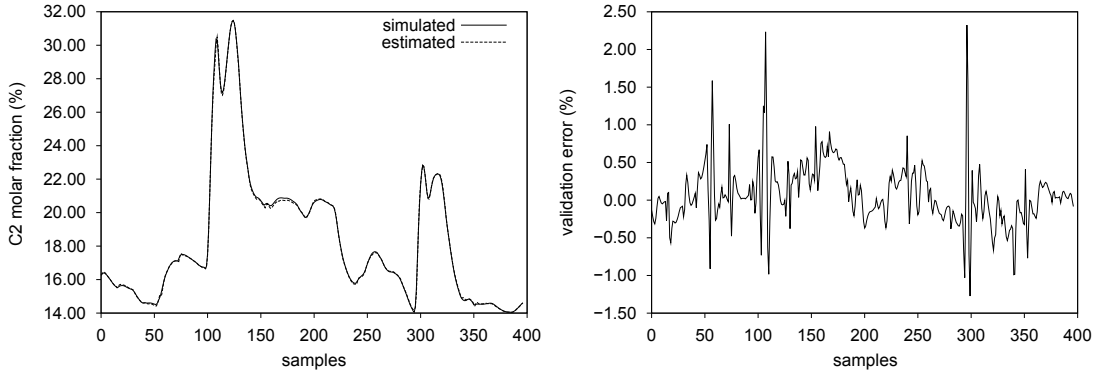
The PCA-ANN 4 ($\gamma = 85\%$) structure has 60% higher number of inputs at its neural network than the PCA-ANN 3 structure. Another note is that the PCA-ANN 4 ($\gamma = 85\%$) has 40% higher number of neurons in its hidden layer than the PCA-ANN 3. As a result of these, the fourth hybrid inferential system proposed had an increasing of 113.04% in the synaptic connections number and consequently a superior training time (111.67%) when compared with the third structure. Thus, the difference between these structures lies in their numbers of synaptic connections, since their performance are similar. Due to its lower complexity the inferential system PCA-ANN 3 was selected as the best structure of this study.

To confirm the functionality of the inferential system selected it was attached to the simulated system formed by the deethanizer and debutanizer columns. The goal is to validate the system comparing the primary variables mole fractions provided by simulation and by the inferential system PCA-ANN 3 in real time. In this validation process, we chose to change the set points of the temperature TIC-100 and TIC-102-2, located, respectively, in deethanizer and debutanizer columns. These procedure can be considered as a routine operation that can be performed either by human operators, as for any control strategy, since these controllers affects directly or the entire process work. Figs. 4–6 shows the comparisons between the estimated primary variables and the primary variables provided by the HYSYS® software.



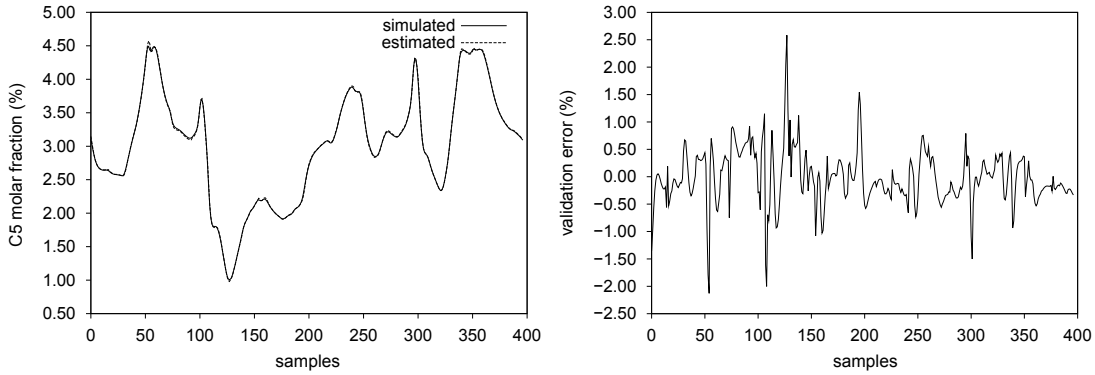
(a) Estimated propane. (b) Estimative error - propane.

Figure 4. Validation of the structure PCA-ANN 3 ($n = 4$ e $N_{hn} = 20$) - Propane estimation.



(a) Estimated ethane. (b) EStimative error - ethane.

Figure 5. Validation of the structure PCA-ANN 3 ($n = 4$ e $N_{hn} = 20$) - Ethane estimation.



(a) Estimated pentane. (b) Estimative error - pentane.

Figure 6. Validation of the structure PCA-ANN 3 ($n = 4$ e $N_{hn} = 20$) - Pentane estimation.

Analysing the validation results illustrated in Figs. 4–6, when abrupt changes occur in primary variables the estimating errors become larger as expected. However, the selected inferential structure (fourth order PCA-ANN 3) is able to represent with efficiency the dynamics of the simulated process. The estimating errors do not exceed 4% for propane, 2.5% for ethane and 3% for pentane, while the module of the mean percentage errors for mole fractions of propane in the residual gas and ethane and pentane in LPG are, respectively, 0.35%, 0.28% and 0.37%.

7. CONCLUSION

According to the results presented, it was noted that the structures of the inferential systems PCA-ANN 1 and 2 didn't had a satisfactory performance when compared with the other structures in study, including even a NNARX inferential system. On the other hand, the systems PCA-ANN 3 and 4 were able to properly estimate the molar fractions of ethane and pentane in LPG and the propane molar in the residual gas.

Looking at the schematic diagrams of the structures studied in this work, we can conclude that the PCA modules perform well its complexity reduction function. However when a PCA module filters the primary variables information, it adversely affects the neural network performance. Therefore, in the compared inferential systems, it is necessary that the primary variables information is used directly at the neural network inputs of the ANN modules to achieve a satisfactory performance.

The structures PCA-ANN 3 and 4 had similar performances and close of the NNARX inferential system. The best results of these hybrid structures were achieved when fourth order models were used. In this condition, it was seen that the PCA-ANN 3 is less complex. So it was the structure selected to be evaluated in real time. The validation of this structure was realized attaching it to the simulated process and confirmed its efficiency.

In this particular process study case it's important to note that the real values of the primary variables are not always available at the input of the inferential systems. So, in a real application in a NGPU, the estimated values of these variables can be used to compensate this lack. Over the time, the use of these estimates lead to an "accumulation" of estimation errors that can deteriorate the performance of the inferential system. A next step of this work is to develop, from measurements of gas chromatographs, an online method to adjust the PCA-ANN system to reduce the negative effect of using the primary variables estimated values. Thus, maintaining the quality of the estimates, it becomes possible to implement control techniques to improve the quality control of NGPUs products, in practice.

8. ACKNOWLEDGEMENTS

The authors would like to thank the the Coordination for the Improvement of Higher Level Personnel (CAPES) by the financial support.

9. REFERENCES

Bo, C., Li, J., Sun, C. and Wang, Y., 2003. "The application of neural network soft sensor technology to an advanced control system of distillation operation". In *Proceedings of the International Joint Conference on Neural Networks*. Portland, Oregon, USA, Vol. 2, pp. 1054–1058.

Chella, A., Ciarlini, P. and Maniscalco, U., 2006. "Neural networks as soft sensors: a comparison in a real world application". In *Proceedings of the International Joint Conference on Neural Networks*. Vancouver, Canada, pp. 2662–2668.

Fortuna, L., Graziani, S., Rizzo, A. and Xibilia, M.G., 2007. *Soft sensors for monitoring and control of industrial processes*. Springer Verlag, London.

Jolliffe, I.T., 2002. *Principal Component Analysis*. Springer-Verlag, New York, USA, 2nd edition.

Linhares, L.L.S., Reboucas, D.L., Araujo, F.U.M. and Maitelli, A.L., 2008. "Sistema de inferência baseado em redes neurais artificiais aplicado em plantas de processamento de gás natural". In *XVII Congresso Brasileiro de Automática*. Juiz de Fora, MG, Brazil.

Mingoti, S.A., 2005. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. UFMG, Belo Horizonte, Brasil.

Nørgaard, M., Ravn, O., Poulsen, N.K. and Hansen, L.K., 2001. *Neural networks for modelling and control of dynamic systems*. Springer-Verlag London Limited, London, England.

Salahshoor, K., Kordestani, M. and Khoshro, M.S., 2009. "Design of online soft sensors based on combined adaptive pca and rbf neural networks". In *IEEE Symposium on Computational Intelligence in Control and Automation*. pp. 89–95.

Warne, K., Prasad, G., Siddique, N.H. and Maguire, L.P., 2004a. "Development of a hybrid pca-anfis measurement system for monitoring product quality in the coating industry". In *IEEE International Conference on Systems, Man and Cybernetics*. The Hague, Netherlands, Vol. 4, pp. 3519–3524.

Warne, K., Prasad, G., Siddique, N.H. and Maguire, L.P., 2004b. "Statistical and computational techniques for inferential model development: a comparative evaluation and a novel proposition for fusion". *Engineering Applications of Artificial Intelligence*, Vol. 17, No. 8, pp. 871–885.

Zamprogna, E., Barolo, M. and Seborg, D., 2005. "Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis". *Journal of Process Control*, Vol. 15, No. 1, pp. 39–52.

10. RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.