Data derived analysis and inference for an industrial deethanizer

Francesco Corona * **Michela Mulas** ** **Roberto Baratti** ***
Jose A. Romagnoli ***,[1]

* *Dept. of Computer and Information Science, Helsinki University of Technology, Finland, (e-mail: francesco.corona@hut.fi).*
** *Dept. of Bio and Chemical Technology, Helsinki University of Technology, Finland, (e-mail: michela.mulas@hut.fi)*
*** *Dept. of Chemical Engineering and Materials, University of Cagliari, Italy, (e-mails: baratti@dicm.unica.it, jose@lsu.edu)*

Abstract: In this paper, we present an application of data derived approaches for analyzing and monitoring an industrial deethanizer column. The discussed methods are used in visualizing process measurements, extracting operational information and designing an estimation model. Emphasis is given to the modeling of the data obtained with standard paradigms like the Self-Organizing Map (SOM) and the Multi-Layer Perceptron (MLP). The SOM and the MLP are classic methods for nonlinear dimensionality reduction and nonlinear function estimation widely adopted in process systems engineering; here, the effectiveness of these data derived techniques is validated on a full-scale application where the goal is to identify significant operational modes and most sensitive process variables before developing an alternative control scheme.

Keywords: Process monitoring, Process supervision, the Self-Organizing Map

1. INTRODUCTION

A modern process plant is under tremendous pressure to maintain and improve product quality and profit under stringent environmental and safety constraints. For efficient operation, any decision-making action related to the plant operation requires the knowledge of the actual state of the process. The availability of easily accessible displays and intuitive knowledge of the states is thus indispensable, with immediate implications for profitability, management planning, environmental responsability and safety.

Due to the advances in measuring and information technology, historical data are available in abundance. Remarkable characteristics of the data acquired in industrial facilities are redundancy and possibly insignificance, not to mention the presence of disturbances that corrupt the measurements. Very often, the amount and quality of the data together with their high-dimensionality can be a limiting factor for the analysis; therefore, it is necessary the availability of effective methods that: i) model the data to extract the structures existing in the measurements, ii) identify and reconstruct the most relevant structures for the scope at hand and, iii) allow for easily interpretable displays where the states' information is presented to the plant operators. Intuitive knowledge of all visited states is invaluable for safe plant operation and trustworthy methods become necessary when considering statistical process monitoring as part of a supervision and control strategy.

In this paper, we discuss the implementation and direct application of a strategy to model, visualize and analyze the information encoded in industrial process data. The approach is based on a classical machine learning method for dimensionality reduction and quantization, the Self-Organizing Map, SOM (Kohonen, 2001). The SOM combines many of the main properties of other general techniques and shares many commonalities with two standard methods for data projection (Principal Components Analysis, PCA (Jolliffe, 2002)) and clustering (K-means, (Hartigan et al., 1979)). In addition, the SOM is also provided with a set of tools that allow for efficient data visualization in high-dimensional settings.

The use of the Self-Organizing Map in the exploratory stage of data analysis is discussed in (Kaski, 1997; Vesanto, 2002) and it is widely employed in many fields. In general terms, the main contributions in applying the SOM on industrial process data are collected by Alhoniemi (2002) and Laine (2003), whereas more domain specific developments can be found in the SOM's bibliography (Oja et al., 2003). Here, the SOM is used as a framework for the identification of the process modes with their time of occurence and present the information on simple displays.

To support the presentation, the analysis is discussed on a full-scale deethanizer where the goal is to identify significant operational modes and most sensitive process variables before developing an alternative control scheme. The study relies on an regression model for estimating an important quality variable (the ethane concentration in the bottom) otherwise difficult to measure in real-time from a set of easily measurable process variables. Inference is based on the Multi-Layer Perceptron Haykin (1998).

[1] On leave from the Department of Chemical Engineering, Louisiana State University, Baton Rouge LA 70803, USA.

2. THE SELF-ORGANIZING MAP

The Self-Organizing Map (Kohonen, 2001) is an adaptive formulation of vector quantization performing in unison:

- a reduction of the data dimensionality by projection; that is, the reduction of the dimensionality of the data by mapping all the observations onto a meaningful subspace with lower dimensionality;
- a reduction of the amount of data by clustering; that is, the retention of the original dimensionality of the data space while reducing the amount of observations by prototyping them by similarity.

The SOM nonlinearly projects vast quantities of high-dimensional data onto a low-dimensional array of few prototypes in a fashion that aims at preserving the topology of the observations. By choosing a conventional bi-dimensional array of prototypes, the main advantage of the map is in a wealth of visualization techniques that allows the analysis of the structures existing in the data.

The following overviews the SOM algorithm and its analogies with other projection and clustering methods. A brief presentation of the most common SOM-based visualization methods for exploratory data analysis is also reported.

Algorithm and properties The basic Self-Organizing Map consists of a low-dimensional and regular array of K nodes, where a prototype vector $\mathbf{m}_k \in \mathbb{R}^p$ is associated with each node k. Each prototype acts as an adaptive model vector for the N observations $\mathbf{v}_i \in \mathbb{R}^p$. During the computation of the SOM, the observations are mapped onto the array of nodes and the model vectors adapted according to:

$$\mathbf{m}_k(t+i) = \mathbf{m}_k(t) + \alpha(t) h_{k,c(\mathbf{v_i})} \Big(\mathbf{v}_i(t) - \mathbf{m}_k(t) \Big). \quad (1)$$

In the learning rule in Equation 1, t denotes the discrete-time coordinate of the mapping steps and $\alpha(t) \in (0,1)$ is the monotonically decreasing learning rate. The scalar multiplier $h_{k,c(\mathbf{v_i})}$ denotes a neighborhood kernel centered at the Best Matching Unit (BMU); that is, at the model vector $\mathbf{m}_c(t)$ that, at time t, best matches with the observation vector $\mathbf{v}_i$. The matching is based on a competitive criterion on the Euclidean metric $d(\mathbf{m}_k(t), \mathbf{v}_i(t))$, for all $k = 1, \ldots, K$. At each step t, the BMU is thus the prototype $\mathbf{m}_k(t)$ that is the closest to observation $\mathbf{v}_i(t)$:

$$c(t) = \underset{k}{\operatorname{argmin}} \Big(d(\mathbf{m}_k(t), \mathbf{v}_i(t))^2 \Big), \ \forall k \text{ and } \forall i. \quad (2)$$

The kernel $h_{k,c(\mathbf{v_i})}$ centered at $\mathbf{m}_c(t)$ is often a Gaussian:

$$h_{k,c(\mathbf{v_i})} = \exp\Big(- \frac{||\mathbf{r}_k - \mathbf{r}_c||^2}{2\sigma^2(t)} \Big), \quad (3)$$

where the vectors $\mathbf{r}_k$ and $\mathbf{r}_c$ represent the geometric location of the nodes on the array and $\sigma(t)$ denotes the monotonically decreasing width of the kernel. The effect of the kernel decreases with the distance from the BMU.

The SOM is computed recursively for each observation. As $\alpha(t) h_{k,c(\mathbf{v_i})}$ tends to zero with t, the set of prototype vectors $\{\mathbf{m}_k\}_{k=1}^K$ are adaptively updated to represent similar observations in $\{\mathbf{v}_i\}_{i=1}^N$, and converge toward their asymptotic limits. The resulting model vectors learn a nonlinear manifold in the original embedding space such that the relevant topological and metric properties of the observations are preserved on the map. Thus, the SOM is to be understood as an ordered image of the original high-dimensional data modeled onto a low-dimensional manifold, where the complex data structures are represented by simple geometric relationships.

A rigorous analysis of the SOM has demonstrated difficult. However, in the case of the basic algorithm with a fixed kernel function, also the SOM algorithm can be understood from the optimization of a cost function:

$$\mathtt{E(SOM)} = \sum_{i=1}^{N} \sum_{k=1}^{K} h_{k,c(\mathbf{v_i})} d(\mathbf{m}_k, \mathbf{v}_i)^2. \quad (4)$$

The cost function in Equation 4 is closely related to the objective optimized with the K-mean algorithm (Lloyd, 1982). The only difference is in the neighborhood function that smoothly weights all the distances between the observations and the prototypes, instead of just the closest one. In that sense, the SOM operates as the conventional clustering method where the width of the kernel is zero. Moreover, there is no need to explicitly specify the number of taxonomies; in fact, the number of prototypes in the SOM can be chosen without any specific concern on the actual number of clusters. The SOM has also neat projection properties. In fact, the cost in Equation 4 closely resembles the objective optimized by Curvilinear Components Analysis CCA (Demartines et al., 1997); CCA is a modification of metric Multi-Dimensional Scaling MDS (Cox et al., 2000) and Principal Components Analysis PCA (Jolliffe, 2002). Similarity is in the decreasing and smoothing nature of the neighborhood function that emphasizes smaller distances in the projection. Conversely, the notion of locality in the SOM does not correspond to the global concern on small distances characterizing CCA.

Data exploration methods In the typical case of projections onto 2D arrays, the SOM offers excellent techniques for data exploration. In that sense, the approach to data analysis with the SOM is mainly visual and focuses on the low-dimensional displays specifically designed for the map.

The data visualization techniques based on the SOM assume that the prototype vectors are representative models for groups of similar observations, and projecting the data onto the low-dimensional array allows for an efficient display of the dominant relationships existing between them. For instance, the displays permit to identify the shape of the data distribution, cluster borders, projection directions and dependencies between variables. The visualizations techniques considered here are i) the component planes and ii) the distance matrix. Such techniques were thoroughly studied by Kaski (1997) and Vesanto (2002).

A component plane shows on the SOM's array the coordinates of the prototype vectors along a specific direction in the data embedding space; that is, each component plane is associated to one original variable and there are as many planes as directions in the embedding. The coordinate values are encoded into gray levels or colors, and the area of each unit on the array is dyed with the color associated to the component value. A component plane thus displays the distribution of the corresponding variable among the prototype vectors. The component planes are useful in order to visually identify possible dependencies between variables. The dependencies between variables can be seen as similar patterns (the colors corresponding to the values

of the variables) in identical locations on the component planes. Such representations can be also used to quantify dependencies. In that sense, the SOM reduces the effect of noise and outliers in the observations and, therefore, may actually make any existing dependence simpler to detect.

A distance matrix visualizes on the SOM's array the average distance between each prototype vector and its adjacent neighbors. In a distance matrix, distances are encoded into gray levels or colors and each unit on the array is dyed with the color associated to the distance with the neighbors. The most widely used distance matrix for the SOM is the Unified Distance Matrix, or U-matrix (Ultsch, 1993). Here, the dominant clustering structure of the observation can be seen as clearly separated areas (large distances) characterized by a homogeneous coloring. In the U-matrix, visualization of the clusters is improved by augmenting the distance matrix with additional entries (nodes) between each prototype vector and each of its neighbors. Unconventional alternatives to the U-matrix are reported by Oja et al. (2003) but not considered here.

3. CASE STUDY

To illustrate the potentialities of topological data analysis using the Self-Organizing Map, the overviewed methods are applied on a set of measurements from a full-scale process. The monitoring problem consists of modeling and analyzing the operational behaviour of an industrial deethanizer, starting from a set of online process measurements. The objective of the deethanizer (in Figure 1) is to separate ethane from the feed stream (a light naphta) while minimizing the ethane extracted from the bottom of the column (an economical constraint for the subsequent unit in the plant). Such a constraint is quantified by the maximum amount of ethane lost from the column bottom; the operational threshold is set be smaller than 2%. The

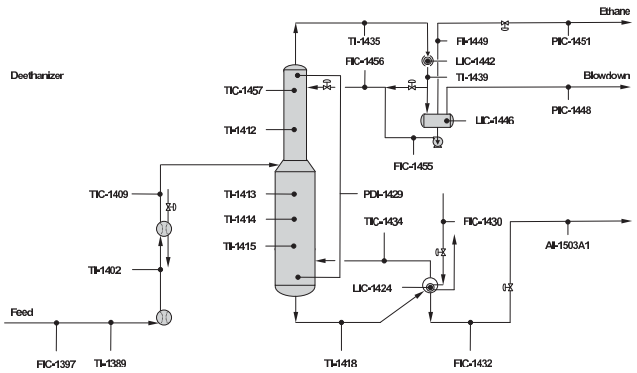


Fig. 1. Deethanizer: Simplified flowsheet.

motivation for choosing this unit is merely illustrative; in fact, the considered deethanizer offers an ample variety of behavior that reflects the operational usage; hence, an interesting groundwork for presentation and discussion.

TAG/Variable	TAG/Variable
FIC-1397/Inlet Flowrate	FIC-1430/Vapor Flowrate
TI-1389/Inlet Temp.	LIC-1424/Reboiler Level
TI-1402/Inlet Temp.	FIC-1432/Bottom Flowrate
TI-1409/Inlet Temp.	FI-1449/Distillate Flowrate
TI-1435/Top Temp.	PIC-1451/Distillate Pressure
TIC-1457/Enriching Temp.	TI-1452/Reflux Temp.
TI-1412/Enriching Temp.	FIC-1455/Bypass Flowrate
TI-1413/Exhausting Temp.	TI-1439/Condensed Temp.
TI-1414/Exhausting Temp.	LIC-1442/Top Drum Level
TI-1415/Exhausting Temp.	PIC-1448/Blowdown Pressure
TI-1418/Bottom Temp.	LIC-1446/Bottom Drum Level
FIC-1456/Reflux Flow.	AI-1503A1/Ethane Conc.
TIC-1434/Vapor Temp.	AI-1503A2/Butane Conc.
PDI-1429/Delta Pressure	

Table 1. Deethanizer: Process variables

In order to analyze the behaviour of the unit, a set of process variables was collected from the plant's distributed control system (DCS). The measurements correspond to three weeks of continuous operation in winter asset and three weeks in summer asset. The data are available as 3-minute averages and 27 process variables (in Table 1) are available for a macroscopic characterization of the unit.

In addition, there are a number of control loops in the process. Briefly, the temperature $TIC-1457$ and the vapor temperature $TIC-1434$ out of the reboiler are controlled by manipulating the reflux flow $FIC-1456$ and the steam rate $FIC-1430$ to the reboiler, respectively; with both loops cascaded with the corresponding flowrates. The distillate pressure $PIC-1451$ is controlled by the distillate flowrate $FI-1449$ and the level in the reboiler $LIC-1424$ by the bottom flowrate $FIC-1432$.

3.1 Analysis and inference

The operational objective of the column is to produce as much ethane as possible (minimizing concentration of propane from the top of the column) while satisfying the constraint on the amount of impurity from the bottom (maximum concentration of ethane in the bottom $\leq 2\%$). With respect to the loss of ethane from the bottom, such considerations led to the definition of 3 operational modes:

- a *normal* status, corresponding to the operation of the column, where the concentration of ethane is within allowable bounds (within the $1.8-2.0\%$ range)
- a *high* status, corresponding to the operation of the column, where the concentration of ethane is exceeding the allowable upper bound (above 2%)
- a *low* status, corresponding to the operation of the column, where the concentration of ethane is below the allowable lower bound (below 1.8%).

The two abnormal conditions have a direct and important economic implication. In fact, when at low status, the process is delivering a product out of specifications whereas, when at high status the product is within the specifications, but an unnecessary operational cost is observed.

To understand under which conditions such modes are experienced, in a recent study (Corona et al., unpublished) we analyzed the clustering structure of the data and visualized the operating conditions of the unit. Starting from a selection of important process variables, we expanded this

subset by incorporating an additional *dummy* indicator, specifically calculated to indicate the status. As such, the new variable was defined as to take values $+1$, -1 or 0, according to the operational status of the process. Value 0 is assigned to the normal operation, whereas values $+1$ and -1 correspond to high and low operations, respectively. Notice that the calculation of the *dummy* variable requires the availability of a real-time measurement for the ethane concentration; such a variable ($AI - 1503A1$) is presently acquired from a continuos-flow chromatograph. The subset of selected process variables augmented by the *dummy* indicator was used to calibrate a SOM over which the resulting component planes and U-matrix were analyzed; the exploration was perfomed as a direct application of the techniques discussed by Alhoniemi (2002). The study allowed us to extract the clustering structure of the data and illustrate on simple displays how it corresponds to the operational modes of the unit.

However, the delay associated with the analytical measurements of the ethane concentration from the bottom of the column can pose severe limitations to the online analysis. Moreover, the existing instrumentation setup available for the unit may benefit from a backup measurement for such an important variable. In this study, we are thus extending the analysis of the operational modes of the deethanizer, by validating the functionality of the approach when replacing the analytical measurements of ethane with online estimates. In that sense, the availability of an inference model would allow the development of a fully automated system to be implemented online in the plant's DCS.

For the purpose, a soft sensor based on the standard Multi-Layer Perceptron MLP (Haykin, 1998) with one hidden layer and sigmoidal activation functions was developed to infere the ethane concentration from the bottom. The estimates are obtained starting from the same input subset of easily measurable process variables used for the SOM and selected according to the guidelines provided by Baratti et al. (1995). The parameters of the MLP (that is, number of hidden nodes, one, and the connection weights) were optimized using the Levemberg-Marquard method and cross-validation. In Figure 2, the response of the soft sensor on a set of independent testing observations is reported for about a week of continuous operation.

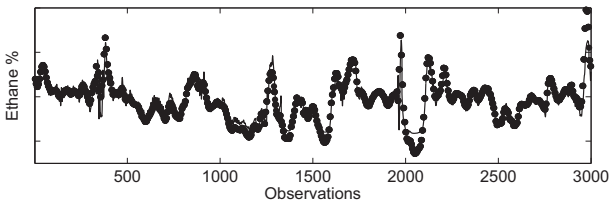


Fig. 2. Ethane concentration from the bottom: Analytical measurements $(\cdot)$ and MLP estimates $(-)$.

Based on the MLP estimates, a bidimensional Self-Orgaznizing Map was calibrated using only the winter data. The map consists of a hexagonal array of prototype vectors initialized in the space spanned by the eigenvectors corresponding to the two largest eigenvalues of the covariance matrix of the data. As usual, the ratio between the two largest eigenvalues was used to calculate the ratio between the two dimensions of the SOM; the resulting map consists of a 70×24 array of 15-dimensional prototype vectors, where the dimensionality of the vectors equals the number of variables used for calibration. On the SOM, we analyzed the clustering structure of the data and visualized the operating conditions of the unit using the U-matrix.

The U-matrix is based on distances between each prototype vector and its immediate neighbors. A common way to visualize it consists of an initial projection of all the distances onto a color axis and the subsequent display with colored markers between each prototype vector. On the display, areas with homogeneous coloring correspond to small within-cluster distances (recognized as clusters), whereas cluster borders are areas with homogeneous coloring but corresponding to large between-cluster distances. The use of the U-matrix in clustering the operational regimes of the deethanizer column is shown in Figure 3.

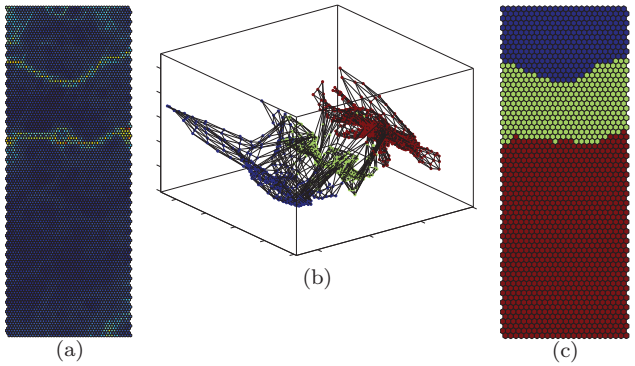
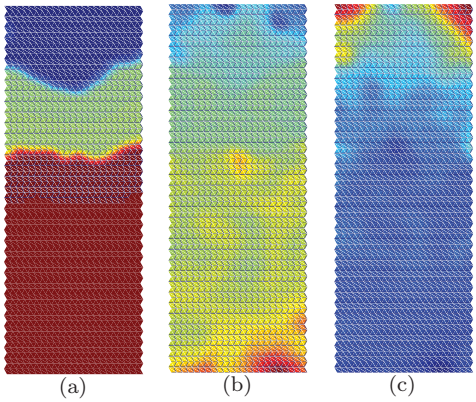


Fig. 3. The U-Matrix (a), the clustered SOM projected onto the 3D principal components space (b) and the SOM colored according to the K-means clustering (c).

In Figure 3(a), distances are depicted with dark blue colors shading toward dark red as the proximity between the prototypes decreases. The visualization permits to clearly recognize the presence of three distinct clusters of prototypes, as well as several other data substructures. An analogous visualization of the grouping is achieved by projecting the map onto a low-dimensional subsapce; in Figure 3(b), a tridimensional principal components space learned by the metric MDS. Indeed, also this visualization permits to illustrate the actual clustering structure of the process measurements and displays a good separateness also in this space of reduced dimensionality. However, to obtain a quantitative characterization of the clustering structure, the prototypes of the SOM should be regarded as a reduced data set and modeled with a standard clustering algorithm. For simplicity, we are here adopting a standard K-means algorithm coupled by the Davier-Bouldin index, a measure of cluster validity to identify an optimal number K of taxonomies from data Milligan et al. (1985). As expected, optimality was found for $K = 3$ clusters, corresponding to the modes of the unit.

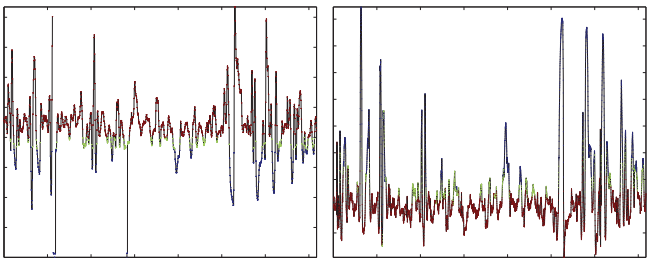
On the SOM, such clusters are located in the lower, middle and upper part of the map. After coloring the SOM according to the cluster membership obtained by using the K-means algorithms, in Figure 3(c), and comparing it with the component plane of the *dummy* variable (and equivalently, the MLP estimated ethane concentation), it is straightforward to associate the three taxonomies to the



(a) (b) (c)

Fig. 4. The component planes for the *dummy* variable (a), the estimated ethane concentration (b) and the temperature $TI-1414$ (c), with a coloring scheme that assigns blue to high values of the variables fading toward red as the values decrease. This scheme differs from the what defined for the clustering with blue and red corresponding to -1 and $+1$, respectively.

three main operational modes of the deethanizer, Figure 4. Specifically, Figure 3(c) shows the clusters on the SOM as distinct regions dyed in blue, green and red with a cloring scheme that assignes those colors to the operational modes ($+1$, 0 and -1, respectively). As expected, a similar structure is also retrieved from the component plane for the *dummy* variable, Figure 4(a). Though apparently less evident, the same structuring is retrieved from the component planes of the estimated ethane concentration (Figure 4(b)) and one of the temperatures in the exhausting section of the dethanizer; namely, $TI-1414$ in 4(c). Looking for similar patterns in similar positions in such components planes allows the visualization of a neat dependence between the ethane composition and such temperature indicator. Such pair of variables shows near identical but reversed component planes, thus highlighting the inherent inverse correlation that exists between them.



(a) AI-1503A1 (b) TI-1414

Fig. 5. The colored time series (3 winter weeks) for the ethane concentration $AI-1503A1$ (a) and the temperature in the enriching section $TI-1414$ (b). The actual values of the variables could not be reported because of the confidentiality agreement.

Information about this dependence can be further enhanced by applying the coloring scheme resulting from clustering directly to the original observations in the time domain. In fact, all points can be dyed using the cluster color of the corresponding Best Matching Unit, as in Figure 5. The figure shows how $TI-1414$ is mostly responsible for the transition between the aformentioned operational

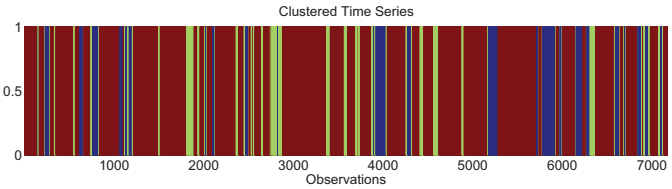
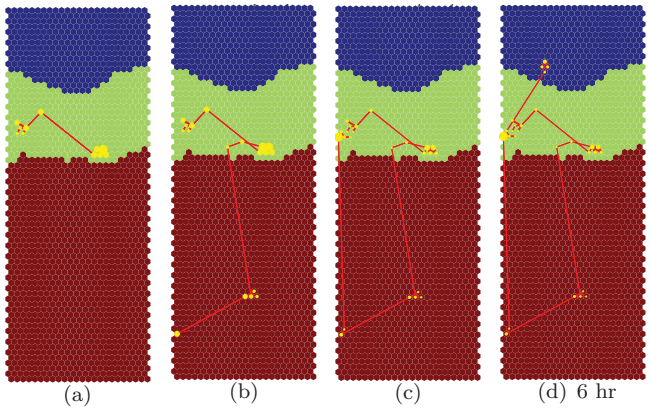


Fig. 6. The temporal evolution of the winter operational modes colored according to the SOM clustering.

conditions. The correspondence with the ethane concentration is observed as clear banded regions and indicates that, in order to maintain the column at optimality (withing the $1.8-2.0\%$ range of ethane from the bottom), such a temperature should be controlled (possibily, within the $52-55°C$ range). A possible variable to manipulate is the steam flowrate $FIC-1430$. However, such a variable is not used in the present control scheme and induces an overall 85% of off-spec operation of the unit, during the given winter period. Such information is obtained by calculating the number of point measurements that falls outside the normality conditions over the total count and pictorially depicted also as clustered time series (in Figure 6).

So far, we have restricted the analysis only to the measurements observed under winter asset. However, it is also possible to directly use the calibrated SOM as a reference model for new and unseen observations; in our setting, the three weeks of data corresponding to the summer operation of the deethanizer column. To validate this idea, the winter SOM was used to explore the behaviour of the deethanizer under summer asset. Again, the summer measurements from $AI-1503A1$ were replaced by the estimates from the soft sensor. The analysis was accomplished by initially projecting the new data onto the calibrated SOM, being the mapping based on a nearest neighbor criterion between the new sample vectors and the prototype vectors of the SOM. In this respect, novelty detection using the SOM is based on finding the BMU. Once the mapping is completed, the inspection is performed for the new data.

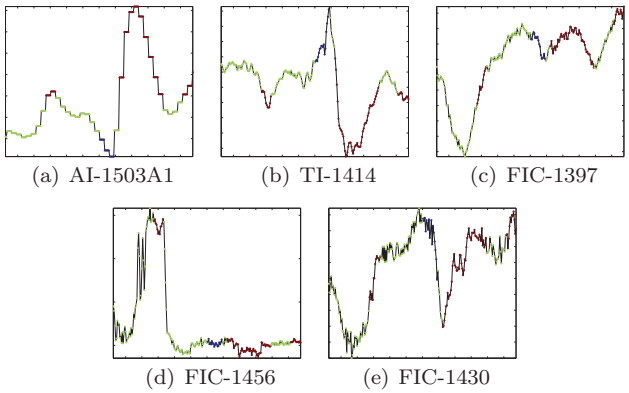


(a) (b) (c) (d) 6 hr

Fig. 7. Trajectory of a selection the summer observations (approximatively, 6hr) displayed on the winter SOM.

The results in extrapolation are presented by illustrating another technique for visualization on the SOM. The approach allows to follow operational changes in the process and tries to provide a simple display for identifying reasons of specific behaviors. For the purpose, the map calibrated on the winter data can be enhanced by the inclusion of

the summer point trajectories followed by the process. The trajectory permits to intuitively indicate the current mode of the process and observe how it has been reached. In Figure 7, the process trajectory is sequentially reported for a small time window corresponding to six hours of continuous summer operation of the deethanizer. The process trajectory on the SOM's domain passes through all the BMUs of each new data vector and it is shown as red line connecting the visited prototypes (the nodes are marked as yellow dots and thicken with the count of visits).



(a) AI-1503A1 (b) TI-1414 (c) FIC-1397

(d) FIC-1456 (e) FIC-1430

Fig. 8. Status transitions on the time domain (approximatively, 6hr), for a set of relevant process variables.

Following the temporal evolution from Figure 7 and 8, the diagrams show a process that is initially operated in the green area, or *normal* condition (as for the ethane in the bottom and reference temperature). As the process has moved further in time, new prototype vectors are visited and added to the trajectory until the column eventually leaves the normality region and crosses the boundary towards the region of high ethane composition (in red). In a similar fashion, all the process variables changed coloring to match the visited modes allowing to appreciate that the change in the operation was mainly due to an abrupt change in the feed flowrate ($FI-1397$), in Figure 8(c), and possibly its composition. In turns, the variation triggered the action on steam to reboiler flowrate ($FIC-1430$), in Figure 8(e), as well as the reflux to control the top temperature ($FIC-1456$), in Figure 8(d). The events initiated a sequence of oscillations around normality that could be reestablished only after several hours.

4. CONCLUSIONS

In this work, we implemented and discussed a strategy to model, visualize and analyze the information encoded in industrial process data. In particular, the proposed strategy was applied to a full-scale distillation column.

From a methodological point of view, the process monitoring problem was casted in a topological framework by using the Self-Organizing-Map. On the SOM, the identification of the process modes was approached as a clustering task rather than classification; that is, in an unsupervised rather than supervised fashion. Moreover, in order to overcome the limitations associated with the time delay and costs of the analytical instrumentation, a software sensor based on a Multi-Layer Perceptron was developed to infer a primary process variable, thus favoring the possibility to directly use such a strategy also for online monitoring.

The application allowed the definition of simple displays capable to present meaningful information on the actual state of the process and also suggested an alternative control strategy for maintaining the unit in normal conditions.

ACKNOWLEDGEMENTS

J. Romagnoli kindly acknowledges Regione Sardegna for the support through the program *Visiting Professor 2008*.

REFERENCES

E. Alhoniemi. Unsupervised pattern recognition methods for exploratory analysis of industrial process data. *Doctoral Dissertation*, Lab. of Computer and Information Science. Helsinki University of Technology, Finland, 2002.

R. Baratti, G. Vacca, and A. Servida. Neural network modeling of distillation columns. *Hydrocarbon Processing*, 74:35–38, 1995.

F. Corona, M. Mulas, R. Baratti, and J.A. Romagnoli. On the topological analysis of industrial process data. PSE 2009 International Symposium on Process Systems Engineering, to appear.

T.F. Cox, and M.A.A. Cox. *Multidimensional scaling, Second edition*. Chapman & Hall, 2000.

P. Demartines, and J. Herault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Transaction on Neural Networks*, 8:148–154, 1997.

S. Haykin. *Neural Networks: A Comprehensive Foundation, Second Edition*. Prentice Hall, 1998.

I.T. Jolliffe. *Principal Components Analysis, Second edition*. Springer, 2002.

S. Kaski. Data exploration using Self-Organizing Maps. *Doctoral Dissertation*, Lab. of Computer and Information Science. Helsinki University of Technology, Finland, 1997.

T. Kohonen. *Self Organizing Maps, Second edition*. Springer, 2001.

A. Hartigan, and M.A.A. Wong. K-means clustering algorithm. *Applied Statistics*,28:100–108, 1979.

S. Laine. Using Visualization, Variable selection and feature extraction to learn from industrial data. *Doctoral Dissertation*, Lab. of Computer and Information Science. Helsinki University of Technology, Finland, 2003.

P. Lloyd. Least squares quantization in PCM. *IEEE Transaction on Information Theory*, 28:129–137, 1982.

G.W. Milligan, and M.C. Cooper. An examination of procedures for determining the number of clusters in a dataset. *Psychometrika*, 50:159–179, 1985.

M. Oja, S. Kaski, and T. Kohonen. Bibliography of Self-Organizing Map (SOM) papers: 1998-2001 addendum. *Neural Computing Surveys*, 3:1–156, 2003.

A. Ultsch. Self-organizing neural networks for visualization and classification. *Information and Classification*, 307–313. Springer, 1993

J. Vesanto. Data exploration process based on the Self-Organizing Map. *Doctoral Dissertation*, Lab. of Computer and Information Science. Helsinki University of Technology, Finland, 2002.