A Decision Support System for consumption optimization in a naphtha reforming plant

Félix Biscarri [a,*], Iñigo Monedero [a], Carlos León [a], Juan Ignacio Guerrero [a], Rocío González [b], Luis Pérez-Lombard [b]

[a] *Escuela Politécnica Superior, Electronic Department, Virgen de Africa, 7, 41011 Sevilla, Spain*
[b] *Escuela Técnica Superior de Ingeniería, Thermotechnics Research Group, Sevilla, Camino de los Descubrimientos s/n, 41092 Sevilla, Spain*

Keywords:
Energy systems engineering
Plant operation
Naphtha distillation Decision Support System Data mining
cost optimization

A B S T R A C T

In a naphtha distillation process, the natural objective is to perform an entire process maximization of the production rate while meeting required product qualities by searching for an optimal operating condition by manipulating the operating variables. The objective of this paper includes performing an energy process optimization. Not only is an adequate production rate met with the required product qualities but the operating cost is also minimized through a data mining approach. The study of the influence of all process attributes in the defined Energy Efficient Indicator (EEI) allows the construction of a multivariate linear model to aid human experts in the recovery of energy losses. A canonical discriminant function carried out the data prediction step. The quality of the Decision Support System framework is illustrated by a case study considering a real database. Also, a commercial software supported by this mining framework is presented.

1. Introduction and bibliographical review

The goal of simulating the performance of an expert is to help human workers solve real-world problems by expertise, a specific domain of knowledge (Shiau, 2011). There are diverse problems which need to be solved in the real world. Thus, the use of an expert system (or a similar artificial intelligence framework) becomes prolific in many fields (Liao, 2005). One of the complex problems for the control in which a computational intelligent approach is amenable is a crude oil distillation unit. In a crude distillation process, the first objective is to perform an entire process optimization including high production rate with a required product quality by searching for an optimal operating condition of the operating variables (Frenkel, 2011; Ouattara et al., 2012). In the previous decade, there was considerable research concerning the optimization of crude distillation processes (Ghashghaee & Karimzadeh, 2011). In Seo, Oh, and Lee (2000), the optimal feed location on both the main column and stabilizer is obtained by solving rigorous "a priori" models and mixed integer nonlinear programming. The sensitivity to small variations in feed composition is studied in Dave, Dabhiya, Satyadev, Ganguly, and Saraf (2003). Julka et al. propose in a two-part paper (Julka, Karimi, & Srinivasan, 2002; Julka, Srinivasan, & Karimi, 2002) a unified framework for modeling, monitoring and management of supply chain from crude selection and purchase to crude refining. In addition to analytical non-linear models, computational intelligence techniques such as neural networks (Gueddar & Dua, 2012; Liau, Yang, & Tsai, 2004) and genetic algorithms (Motlaghi, Jalali, & Ahmadabadi, 2008) are used for the same purpose. Alhajree, Zahedi, Manan, and Zadeh (2011) cite several Artificial Neural Network research studies for the control of processes in petrochemicals and refineries. From cited papers, most of the nonlinear controllers require the feedback of state information for effective control and close monitoring of a process. In practice, however, the complete online information about the present state of the industrial process is rarely available. If the real-world values are not provided to the algorithm on time, the control algorithm becomes formally invalid. In practice, it recovers from the situation, at the price of reduced quality control (i.e., worse product), so such situations should be avoided (Metzger & Polakow, 2011).

The scope of this present study is concerned with a part of the crude oil distillation called the platforming unit. It is made up of two subunits: the catalytic reforming or reaction unit and the distillation unit or train distillation. Most of the cited references are focused on optimizing the production rate of the distillation unit (Iranshahi, Bahmanpour, Paymooni, Rahimpour, & Shariati, 2011; Meidanshahi, Bahmanpour, Iranshahi, & Rahimpour, 2011), but if the focus is the heat recovery, 80% of the energy consumption (67% of the energy invoicing tasks) corresponds to the fuel consumption in the boilers of the previous task (the reaction unit).

* Corresponding author. Tel.: +34 954 552836; fax: +34 954 552833.
E-mail address: fbiscarri@us.es (F. Biscarri).

At present, research is not only focused in the rise of the production rate but also in making customized products (Frenkel, 2011) and in the improvement of product quality (Rahimpour, Vakili, Pourazadi, Iranshahi, & Paymooni, 2011). In this sense, classic applications of linear control theories on the distillation unit are widely available in the literature (Jabbar & Alatiqi, 1997). Also nonlinear state estimation research (Jana, Samanta, & Ganguly, 2009) and optimal planning strategy research (Kuo & Chang, 2008) are available. The main objective of these papers was to remove impurities in the distillate (i.e., C_5^+ in the debutanizer column) and maintain the minimum possible amount of product (butane) in the bottom residual fuel oil to maximize the yield of the product.

The energy management (de Lima & Schaeffer, 2011; Kansha, Kishimoto, & Tsutsumi, 2011) and the energy efficiency (Chiwewe & Hancke, 2011) become important problems. The objective is to perform a complete plant energy process optimization, including an adequate production rate with the required product quality while minimizing operating costs (fuel consumption in boilers) through a data mining approach. Several research endeavors have treated consumption analysis as a knowledge discovery problem using intelligence techniques (Li, Bowers, & Schnier, 2010). Both forms of learning, supervised and unsupervised, have been adopted in these studies (Hippert, Pedreira, & Souza, 2001; Metaxiotis, Kagiannas, Askounis, & Psarras, 2003). In Hippert et al. (2001), the unsupervised learning based on the SOM algorithm for the three tasks, namely classification, filtering and identification of customer load pattern, is proposed. The intelligent control algorithms applied to the control of combustion processes have produced satisfactory results and show a great potential for growth. Previous research has shown that boiler efficiency can be optimized with data-mining approaches (Miyayama et al., 1991; Ogilvie, Swidenbank, & Hogg, 1998). In Kusiak and Song (2006), the authors proposed an optimization with clustering-derived centroids. In Song and Kusiak (2007), the authors develop a data mining approach for optimizing the combustion efficiency of an electric-utility boiler subject to industrial operating constraints. The latest cited papers offer interesting researches about single boilers. These studies encourage the authors of the current paper to offer a mining approach to optimize the efficiency of a complete distillation plant, regarding the operating and economical constraints.

Since close monitoring of the process is, in practice, rarely available, only information collected in a historical database and the data mining software tools were used. The expert's performance is hidden in the collected dataset. This valuable knowledge feeds the proposed Decision Support System (DSS) framework. The global plant control model does not need to be reconfigured. The expert's information can simply be extracted.

The questions that emerge are: is it possible to extract expert information from the limited amount of data collected in the historical database, searching in past data optimal cost operating conditions? And, is it possible to improve energy efficiency result by the estimation of new operating condition with a DSS software tool? The feasibility and benefits of the proposed framework are demonstrated with a real case study reported. The proposed framework-based pilot commercial software is also presented.

The paper is organized as follows: in Section 2, the refinery platforming unit process is described. In Section 3, the data mining-based DSS framework is presented. It is divided into four subsections: the nature of the data set, the data preprocessing (cleaning and filtering), the data transformation and discretization and finally, the data reduction and prediction. In Section 4, a solution to increase the plant energy efficiency is proposed. Section 5 illustrates the quality of the framework by a case study considering a real database. In Section 6, a framework-based commercial software is presented. Section 7 outlines future directions and concluding remarks.

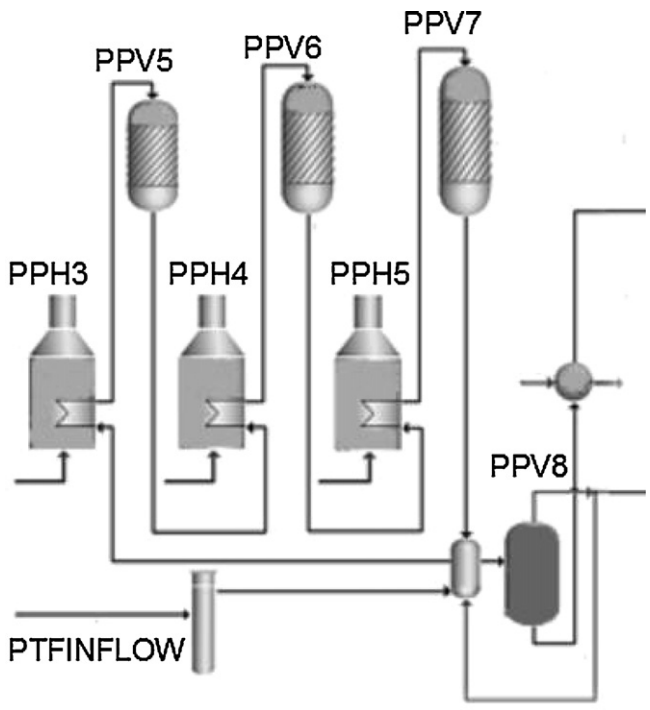


Fig. 1. Process flow diagram of the catalytic reforming plant.

2. The refinery platforming unit process

Refineries are composed of several operating units that are used to separate fractions, improve the quality of these fractions and increase the production of higher valued products like gasoline, jet fuel, diesel oil and home heating oil. The function of the refinery is to separate the crude oil into many kinds of petroleum products. This paper pays special attention to the Platforming Unit. This unit is constituted of two basic units: the catalytic reforming or reaction unit and the distillation unit or train distillation.

The conventional catalytic naphtha reforming process has been described in previous studies (Iranshahi, Rahimpour, & Asgari, 2010; Rahimpour, Iranshahi, & Bahmanpour, 2010). The process consists of three adiabatic reactors containing inter stage heaters to increase the reaction rates. The main idea of the process is to convert paraffins and naphthenes into aromatics (Fig. 1). The feed to the naphtha reformer is a crude oil fraction from the refinery crude unit with a boiling range between 100 °C and 180 °C. This process is adiabatically carried out at high temperatures, building up gasoline with a high octane number, LPG, hydrogen, fuel gas and coke, in three reformers. The coke deposits on the spent catalyst surface causing its deactivation. To recover its activation, the catalyst with coke is regenerated after a certain running time.

In the first reactor, the major reactions such as dehydrogenation of naphthenes are endothermic and very fast, causing a very sharp temperature drop. For this reason, this process is designed using a set of multiple reactors. Heaters between the reactors allow an adequate reaction temperature level to maintain the catalyst operation.

The effluent from the last reactor (PPV7, Fig. 1) is cooled partly by heat exchange with the reactor charge. The stream then enters the product separator and some of the light hydrocarbons are produced. The separator liquid product is pumped into the distillation unit. The function of the distillation unit is to separate the input product and to produce the aromatic fraction, i.e., benzene, toluene, $C8^-$ and $C9^-$ aromatics.

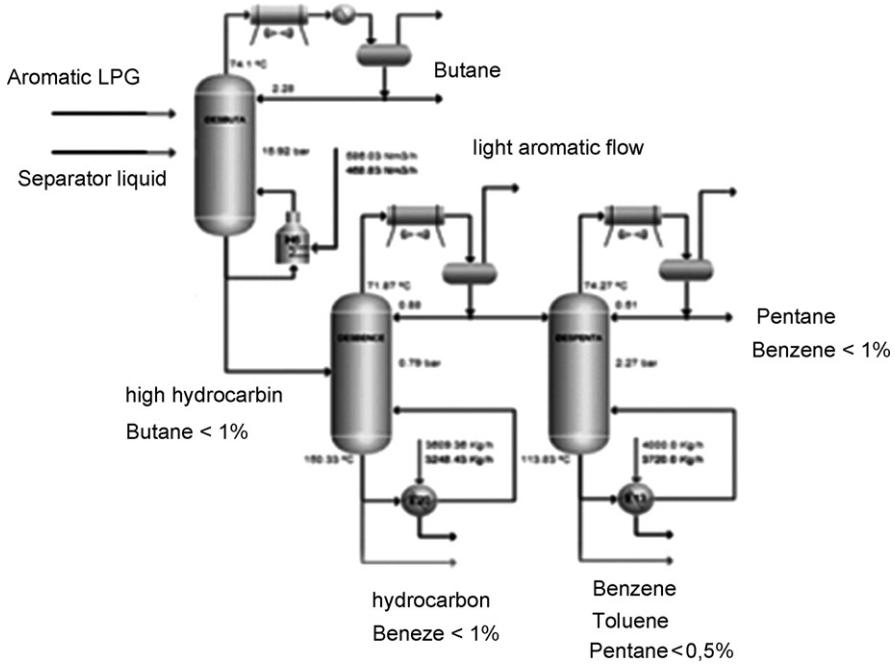


Fig. 2. Process flow diagram of the distillation unit.

This process is performed in three different distillation columns (Fig. 2). The separator liquid and a stream (called 'aromatic LPG') from the external platforming unit, feed off the first column, the debutanizer column. This column has coupled and strong nonlinear dynamics. To maintain the product specifications, a tightening process control is required, which is really a challenging task for control engineers (Jana, 2010). This column splits the input into two basic products: butane, to the top of the column and a high hydrocarbon flow, also called 'platformer', to the bottom of the column. Platformer feeds off the debenzenizer, the second distillation unit. Its goal is to obtain a light aromatic flow free of the high hydrocarbon. This stream is fed off the third distillation column that produces benzene and toluene. Benzene and toluene are the important products to the plant. The products are sent to the Morphylane Unit. The bottom product is sent to the second column and the top product to the third column, which are stored up or sent to the other units of the refinery.

As the platforming unit is one of the critical and important unit operations for the petroleum industry, the goal is to achieve a well-controlled and stable system, high production rate and product quality as well as low operating cost for the economic consideration. For this reason, attention has been paid to this unit to improve product rate, efficiency and quality assurance in petroleum industry in recent years.

3. The data mining-based DSS framework

Knowledge Discovery in Databases (KDD) is the process of identifying valid, new, useful and understandable patterns for large datasets. Data mining is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns-which are the essence of useful knowledge (Maimon & Rokach, 2010).

The data mining task can be specified in the form of a data mining query, which is the input to the data mining system. A data mining query is defined in terms of the following primitives (Han & Kamber, 2001).

- Task-relevant data. This is the database portion to be investigated in particular, the reaction zone. Attributes of interest (relevant attributes) to be considered in the mining process are defined in Appendix A.
- The kinds of knowledge to be mined. The data mining function to be performed is the classification of the operation features that belong to the EEI. The associations between the feature vector (operation plant features) and the predictor variable (the EEI) can be specified.
- Background knowledge. The knowledge about the domain to be mined is included in the historical database. This knowledge is useful for guiding the knowledge discovery process, in order to find the pattern of data. A new operating point, which is similar to a past operating point included in the historical database and can verify the quality of data constraint, would be found.
- Presentation and visualization of discovered patterns. Finally, the complete DSS is automatized through a practical software.

3.1. The nature of the data set

Appendix A describes the set of hourly available measurements: temperatures measured in the input and the output of the reactors, the fuel gas heaters' consumption, the platforming naphtha input flow, the product separator pressure that is controlled by the recycle gas density and the room temperature. Also, the temperature increase between the three reactors is available.

The frequency of the operation features is hourly, but the quality of the product is only analyzed once a day. The length of the sample of data is from January 2009 to May 2010; from January 2009 to January 2010 for training data, and the rest for validation subset. So, the full valid sample contains 12,149 records and the training sample 9402 records.

3.2. Data preprocessing: data cleaning and filtering

Real-world databases are highly susceptible to noisy, missing and inconsistent data. Data cleaning can be applied to correct inconsistencies (Buzzi-Ferraris, 2011; Buzzi-Ferraris & Manenti, 2011).

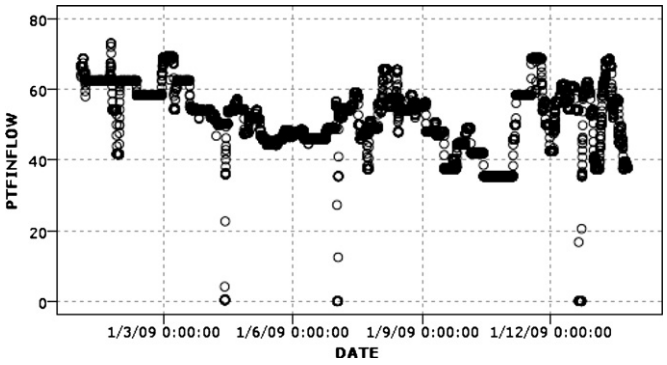


Fig. 3. Naphtha flow vs. date.

They are not clear outliers and spurious data, but the plant was partially stopped for five days during the period of study. This data would be removed from the sample. Fig. 3 shows clearly days whose input naphtha flow suffered a fall. These days (120 records) are classified as outliers and they are deleted from the training sample.

On the other hand, the quality of data is analyzed only once a day but not every day. The quality control of the plant limits the level of impurities in the distillate by two rules. The failure to comply with these rules set the "inconsistent days":

- In the bottom of the debutanizer column, the percentage of benzene (C_5^+ components) must be less than 1%.
- In the bottom of the debenzenizer column, the percentage of toluene must be less than 10%

Based on this first control objective, the sample is filtered. All data between the day before and the day after an "inconsistent day" are filtered. For example, a simple case is illustrated. Fig. 4 shows the percentage of toluene in the bottom of the debenzenizer column. The fourth, fifth and eighth July are outliers. The day before this range (third July) and the day after (ninth July) are also filtered. Note that no information is available of quality between fifth and eighth July. Then, all data between third July and ninth July (both included) are removed from the sample.

So, due to an excessive percentage of toluene, the "% toluene filter" deletes the 125 days from the training sample. At this point, there are valid 6482 records in the sample data.

The same procedure is applied in the case of the percentage of benzene in the bottom of the debutanizer column (Fig. 5). The added "% benzene filter" deleted 17 new days from the sample. The final training sample has reduced from 9402 to 6074 records due to the imposed quality restrictions. It is a strong but typical reduction in industrial process mining approach.

3.3. Data transformation and discretization

The data would be transformed or consolidated into a form appropriate for mining energy efficiency. Thus, new attributes are constructed and added for the given set of the operating plant features. The study of the influence of all process attributes in the Energy Efficient Indicator (EEI) allows one to identify the most influent attributes (attributes of interest) and the construction of a multivariate linear model to aid human expert to recover the energy losses. So, the EEI in the reaction zone on the plant measures the total reaction zone consumption with respect to the plant input flow. It is defined as follows:

$$EEI_{REACTION} = \frac{\sum_{i=3}^{5} CSMFGPPH_i}{PTFINFLOW} \quad (1)$$

Discretization techniques will be used to reduce the number of values for a given continuous attribute, particularly the EEI. Interval levels can then be used to replace actual data values in the classification mining process. Reducing the number of values for an attribute is especially beneficial if classification mining is to be applied to the preprocessed data. Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. In this sense, the analysis of the histogram is used. Partitioning rules defines the ranges of values. For instance, the standard deviation (σ) of the histogram splits the data distribution of $EEI_{REACTION}$ into five (using the $\pm 2\sigma$ range) disjoint subsets or buckets. Fig. 6 presents a histogram showing the data distribution of $EEI_{REACTION}$ and partitions.

From now on the new discrete variable, $EEI_{REACTION-SDBIN}$, replaces the continuous variable for classification purposes.

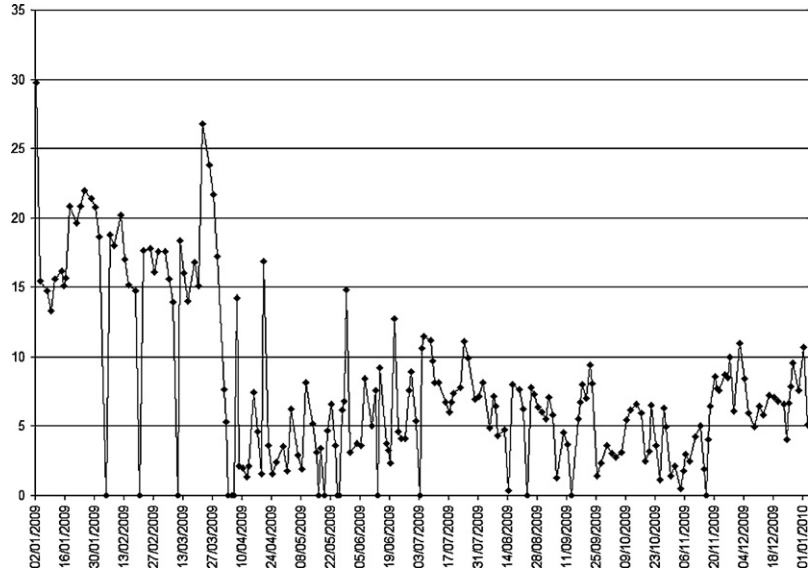


Fig. 4. Percentage of toluene after the debenzenizer column.

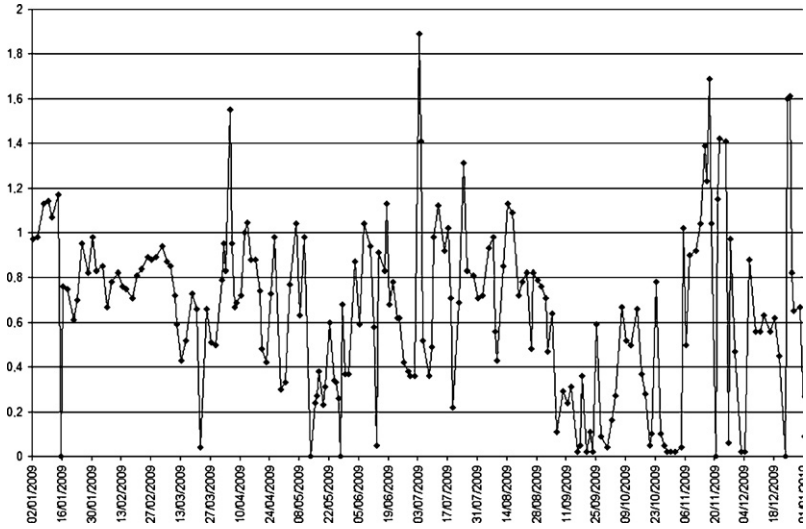


Fig. 5. Percentage of benzene after the debutanizer column.

Operation plant states are now classified according to their $EEI_{REACTION-SDBIN}$ value: very high and high consumption ($EEI_{REACTION-SDBIN} = \{2, 1\}$), normal consumption ($EEI_{REACTION-SDBIN} = \{0\}$) or low and very low consumption ($EEI_{REACTION-SDBIN} = \{-1, -2\}$). The mining objective becomes more specific after the data discretization step: the present operation point (state of the plant) is to be headed for an "optimal zone", defined by a minimum $EEI_{REACTION-SDBIN}$ zone.

3.4. Data reduction and prediction

Once the EEI is discretized, a discriminant analysis technique is used to predict this categorical response variable. Unlike generalized linear models, it assumes that the independent variables follow a multivariate normal distribution. The procedure attempts to determine several discriminant functions (linear combinations of the independent variables) that discriminate among the groups defined by the response variable.

Discriminant analysis is different from other techniques such as factor analysis, in that it is not an interdependence technique: an attribute relevance analysis that differentiates independent variables from dependent variables (also called criterion variables) must be made prior to use. Then, the first step is to measure the relevance of attributes. An automatic method should be introduced to perform attribute relevance analysis in order to filter statistically irrelevant or weak relevant attributes, and retain or even rank the most relevant attributes for the descriptive mining task at hand. There have been many studies in machine learning, statistics, fuzzy and rough set theories, and so on, on attribute relevance analysis (Gong, Huang, & Chen, 2008). The general idea behind attribute relevance analysis is to compute some measurements that are used to quantify the relevance of an attribute with respect to a given class or concept. Such measurements include information gain, uncertainty and correlation. In this sense, the p-value based Pearson chi-square tests for independence of the target and the predictor without indicating the strength or direction of any existing relationship is suitable for the framework purposes. From the list of predictors (Appendix A), eight attributes are selected (Table 1).

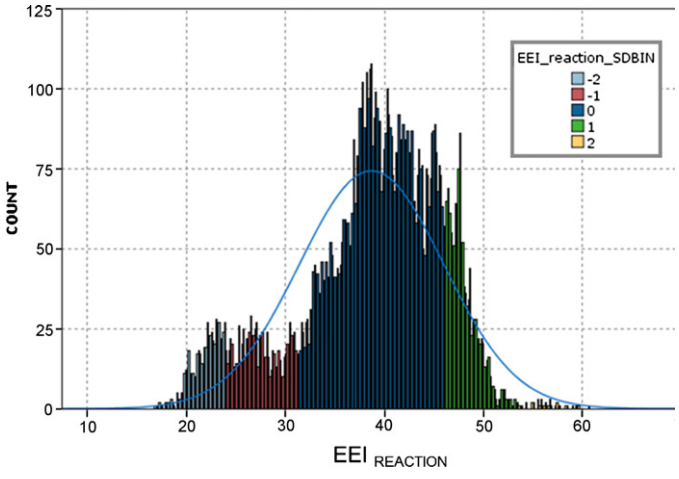


Fig. 6. Histogram of $EEI_{REACTION}$.

After inputs are selected, a canonical discriminant analysis separates the five classes of $EEI_{REACTION-SDBIN}$ through a linear combination of selected attributes. SPSS Modeler (originally, Statistical Package for the Social Sciences, since 1968) is used as the data mining tool for analysis. It is a data mining software tool by SPSS Inc., an IBM company. The data processing in SPSS Modeler is done through the use of nodes which are then connected together to form a stream frame. Designed to support CRISP-DM (Cross Industry Standard Process for Data Mining), which is *de facto* the standard for implementing data mining as a business process, Modeler's open architecture utilizes existing integrated investments to enable rapid predictive modeling and high Return On Investment (ROI) deployment. The processing time varies greatly depending on the size of the sample. This could take anywhere from a few minutes (hundreds of fields, hundreds of records) to several hours (a few million of data). In the case study that covers 74 fields and 25,400 records, the data mining processing time is reduced to a few seconds, using a standard PC.

Table 1
Attribute selection.

Rank	Attribute	Value
1	*CSMFGPPH3*	1.0
2	*CSMFGPPH5*	1.0
3	*PTFINFLOW*	1.0
4	*CSMFGPPH4*	1.0
5	*P_PPV88*	1.0
6	*TPROOM*	1.0
7	*DENRECGAS*	1.0
8	*TPRPPV567*	0.99

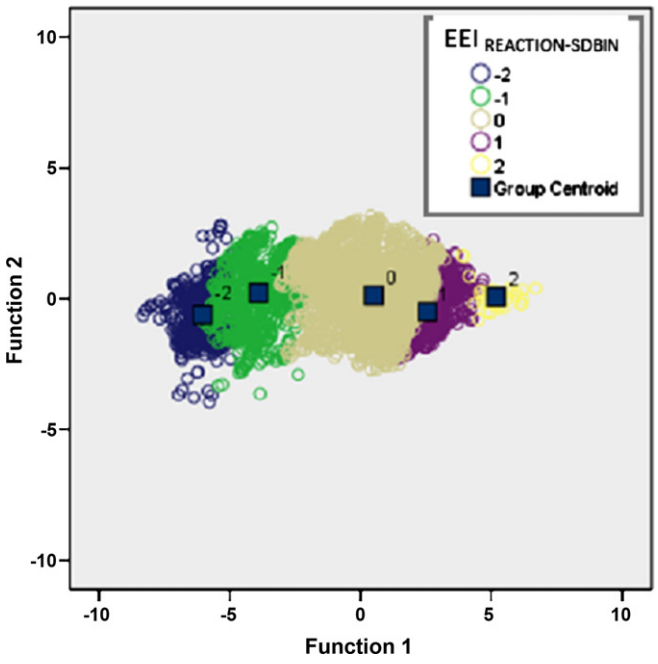


Fig. 7. Canonical discriminant functions.

The result offers a good classification, by means of which the first two canonical functions (named Function1 and Function2). The 92.9% of original grouped cases are classified correctly (92.8% of cross-validated grouped cases classified correctly). Function1 covers 96.8% of the variance. Function2 covers an additional 1.5%. Fig. 7 shows the canonical discriminant functions. Extreme groups, $\{-1, -2\}$ and $\{1, 2\}$, do not overlap. The model evaluation is first performed using ten-fold cross validation in the training sample. Later, a new validation by means of the testing sample is done. This kind of evaluation was selected to train the algorithms using the entire testing data set and obtaining a more precise model. This will increase the computational effort but improves the model's capacity for generating different data sets. The evaluation is performed by splitting the initial sample in 10 sub-samples in order to fill consumption range. The model is trained using 9/10 of the data set and tested with the 1/10 left. This is performed 10 times on different training sets and finally the ten estimated errors are averaged to yield an overall error estimate.

Using the normalized variables, the discriminant analysis also offers a structure matrix that allows building the discriminant functions from discriminating variables, without using the canonical form. From now on, the *N_* prefix indicates a normalized attribute. Table 2 shows the final form of Function1 and Function2. The variables are ordered by the absolute size of correlation within function.

Using the normalized variables, and by weighing up the high percentage of variance covered by Function1, the plant energy efficiency will be improved by means of the new attribute defined in (2).

Table 2
Discriminant functions with normalized attributes.

Attribute	Function1	Function2
N_CSMFGPPH5	0.280	−0.036
N_CSMFGPPH3	0.270	0.722
N_PTFINFLOW	−0.266	0.642
N_P_PPV8	−0.213	0.589
N_DENRECGAS	0.009	0.237
N_CSMFGPPH4	0.160	0.227
N_TPROOM	0.160	−0.262
N_TPRPPV567	0.013	−0.104

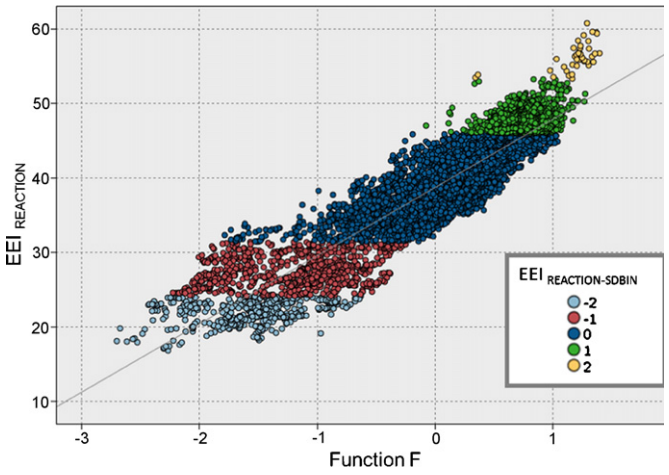


Fig. 8. Function *F* vs. EEI.

$$F = 0.28 * N\_CSMFGPPH5 + 0.27 * N\_CSMFGPPH3 - 0.27 * N\_PTFINFLOW - 0.213 * N\_P\_PPV8 + 0.009 * N\_DENRECGAS + 0.160 * N\_CSMFGPPH4 + 0.160 * N\_TPROOM + 0.013 * N\_TPRPPV567 \quad (2)$$

Fig. 8 shows the strong correlation between *F* and the energy efficiency indicator. The low negative values of *F* guarantee low consumption with respect to the platforming input flow. Thus (2), in order to improve the energy efficiency from a given operating point, the plant operator should increase *PTFINFLOW* or *P_PPV8*, or decrease the value of the other selected attributes. Is it possible, in practice, to move the operating point in this sense? It depends on the present plant operation constraints between attributes, but the historical database, with the help of the discriminant function *F*, suggests some possible ways, as one can see in the following section. In Fig. 8, also the regression line of the training sample is drawn as follows: $\$EEI_{REACTION} = 38.658 + 9.177 * F$. The regression model summary results in an $R^2 = 0.832$. The Pearson correlation between the dependent variable and the predictor is 0.912. The 95% confidence interval for coefficients are (38.597;38.719) and (0.093;9.261).

Data mining with large data set is a particularly hard problem to be solved for many observers. For example, it is not possible to use the arithmetic mean, since it is very efficient but at the same time fully nonrobust (masking and swamping problems arise). Moreover, it is not possible to use the median or the trimmed mean or other methods, since it is computationally hard to order very large sets of data like the industrial ones are considered in this work. The proposed model of $\$EEI_{REACTION}$ and the use of an efficient data mining tool overcomes difficulties and offers a practical approach.

4. A possible solution to increase the energy efficiency

Once the industrial process is characterized and the independent attributes are selected, the predictive model links $EEI_{REACTION}$ and attributes. There are several ways to derive the new information from the data. For example, a new predictive supervised classification model should be trained using the plant database; function *F* and the selected attributes as input. The supervised classification model output would be "a priori" estimated for training purposes: a set of "suggested" new optimal operation points would be established. Once the model is trained and validated, its output

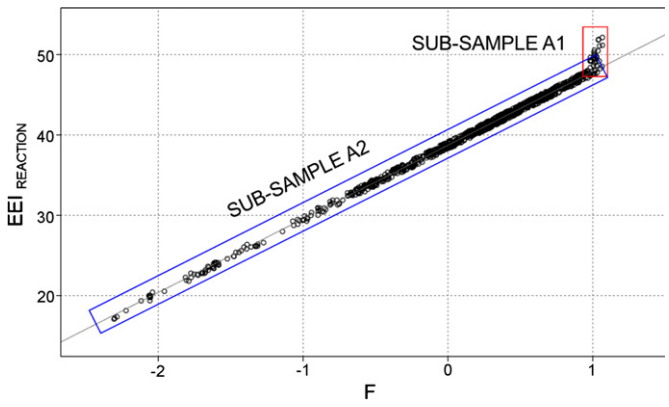


Fig. 9. Sub-sample A.

would offer a "suggested" new operation point in order to improve the energy efficiency. The new value of $EEI_{REACTION}$ will drop. The problem is the real knowledge of the "a priori" supervised classification model output. In the aromatic distillation process, there are some different possible solutions: the industrial process constraints are added to the economical constraints, usually "a priori" unknown (Robertson, Palazoglu, & Romagnoli, 2011). For example, sometimes the production rate of some subproduct (benzene, for example) would be fixed. Sometimes, the price of some subproduct varies and several operation changes are suggested. This particular procedure advises a human expert validation phase. At that point in the mining process, the semiautomated mining framework offers a clear decision aid system, but it is risky to adopt a fully automated system. At present, the suggested approach would be tested, in this sense, after the human expert validation phase. Section 5 illustrates a practical application of the method.

The authors propose a simple but visual, practical and effective method to aid the human expert to improve efficiency. It is based on the distance between the regression line and the plant operation point. It takes the plant operation constraint between attributes into account. The following steps describe the method.

- The source data are the training sample of data, the classification function F and the regression line (named $\$EEI_{REACTION}$) shown in Fig. 8. The collection of the selected attributes, plus calculated F and $\$EEI_{REACTION}$, is used to define the learning environment: $A|A = \{a_0, \ldots, a_n, F, \$EEI_{REACTION}\}$. Instances of the learning environment, e_m, are defined using values taken by vector $E = \{e_0, \ldots, e_m\}$.
- A function that measures the minimum absolute value of the distance between each $EEI_{REACTION}$ value and each point of the estimated regression line would be established as shown below:

$$Distance_e = min|EEI_{REACTION_e} - \$EEI_{REACTION_{j \in E}}| \quad (3)$$

where $Distance_e$ is the distance value of each individual state e.
- From a given original plant operation point (state e), with a certain value of variables ($F = F_0$; $Distance = Distance_0$), two sub-samples of data are selected. The sub-sample A1 should verify the following conditions: $Distance_{A1} \leq Distance_0$ and F_{A1} between $F_0 * (1 - F_{percent}/100)$ and F_0. $F_{percent}$ is the allowed percentage of change in F. The sub-sample A2 should verify: $F_{A2} < F_0$ and $Distance_{A2} < Distance_{threshold}$, with $Distance_{threshold}$ is a fixed limit of variable Distance. Joining A1 and A2 sub-samples, a possible theoretical way (sub-sample A) to "move" the original operation point to a new point of less $EEI_{REACTION}$ value is obtained. The objective is to move step by step the operation point toward the correct direction but with "small" (possible in practice) variations of the attributes. Fig. 9 shows an example of this process, with $F_0 = 1.069$, $Distance_0 = 3.97$. The threshold values are fixed as $F_{percent} = 5\%$ and $Distance_{threshold} = 0.5$.
- The last item is the most important task. The real industrial process imposes clear constraints to the attribute values. Let's suppose that the initial room temperature is $N\_TPROOM = 24.6\,^{\circ}C$. If this attribute decreases, the energy efficiency rises (2), but it is not very realistic to suggest much variation in N_TPROOM. In the same sense, the input naphtha flow ($N\_PTFINFLOW$) variation range is very limited in practice. Once the industrial process constraints are known, the operator must filter the sub-sample A. The reason for including constraint in the last step is simply because constraints are very changeable and it is difficult to establish its "a priori". It would be better if, first, the model is established and, then, the constraints filter the model. In this sense, the human expert must choose between several different options, knowing the present operation point and the present economic and the industrial constraints.

Table 3
Case study.

$CSMFGPPH5_0 = 328.77$	$CSMFGPPH3_0 = 1051.72$
$PTFINFLOW_0 = 41.39$	$P\_PPV8_0 = 14.23$
$DENRECGAS_0 = 0.38$	$CSMFGPPH4_0 = 790.39$
$TPROOM_0 = 24.65$	$TPRPPV567_0 = 102.79$

The correspondence in the past for the current conditions and the analogy between the current situation and certain previous conditions are two fundamental pillars of the proposed methodology.

In our battery test, we have found that given the dispersion of the set sample, it is possible to reach a significant improvement in the energy efficiency in all the operation points registered in the plant. It is because the environment conditions of the plant have not changed from the past and, therefore, the analogy between a previous and a current operation point is possible. The current condition (exactly values of all operation variables) is not probably found in the past, but the evolution of operation variables to a minimum EEI (the F vs. EEI, Fig. 8) is the same. Thus, an average rise in the energy efficiency is reached after the simulation of the full testing sample.

5. A case study

For a simple case study, a real operation state characterized by the attributes is shown in Table 3. This state is situated in the right-hand upper corner of a Fig. 9.

The calculated attributes are: $EEI_{REACTION_0} = 52.45$, $F_0 = 1.07$, $Distance_0 = 3.97$. The thresholds are fixed as $F_{percent} = 5\%$ and $Distance_{threshold} = 0.5$.

The room temperature is one of the most relevant attributes for the $EEI_{REACTION}$ evolution from a given operation point. Low temperatures help the plant efficiency. But, what is the $EEI_{REACTION}$ limit, with regards to a given range of N_TPROOM variation? From the proposed framework, just filter Sub-sample A. Fig. 10 shows the Sub-sample A with $N\_TPROOM >= 22\,^{\circ}C$. Two degrees from the initial point ($N\_TPROOM_0 = 24.65\,^{\circ}C$) limit the theoretical optimum to $EEI_{REACTION} \approx 38$. Nevertheless, there is still much more room for improvement. From the point of the view of this range of temperature, from $EEI_{REACTION_0} = 52$, the efficiency can improve to a significant $(52 - 38) * 100/52 = 27\,\%$.

The influence in the optimization process of the input naphtha flow variation range is in practice quantitatively different from the room temperature influence. Once the sub-sample is filtered by the room temperature, a new input flow filter reduces the points in the sub-sample A (the suggested future states), but not too much the optimal $EEI_{REACTION}$ range. Suppose that the input naphtha flow only

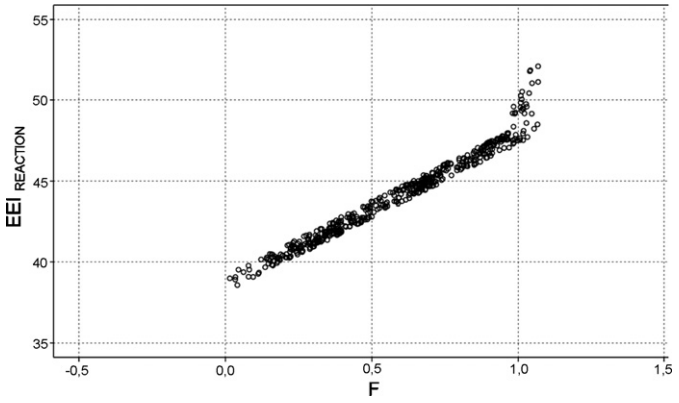


Fig. 10. Temperature influence.

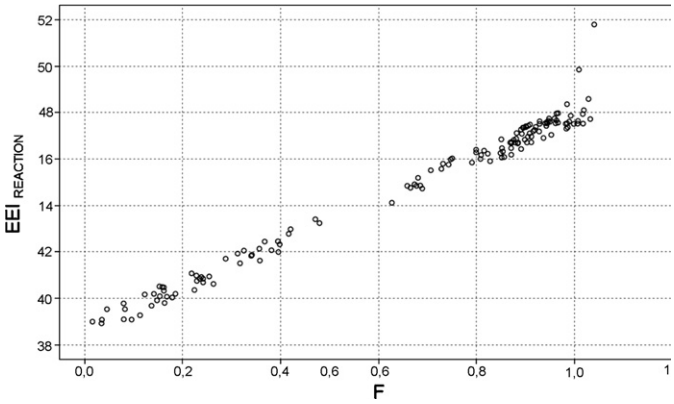
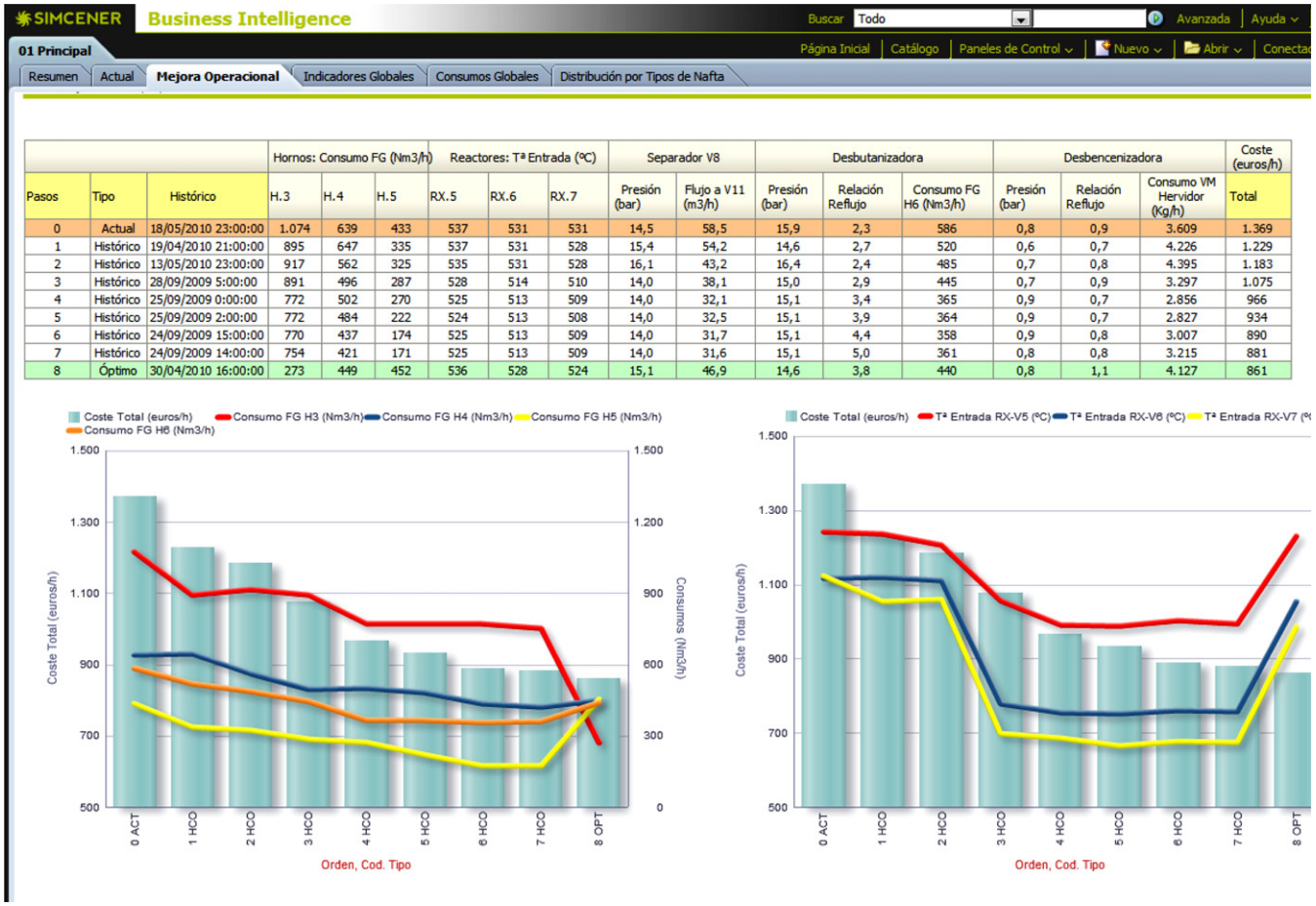


Fig. 11. Temperature and input flow influence.

could be increased from 41.39 just to let's say 45. It is a real possible variation range. Fig. 11 shows this new filter effect, added to the previous one. The problem is that successive filters highly reduce the number of suggested new operation points. In this case, the original training sample is not too large. In Fig. 11, for example, there are not any optimal future plant states in the range $F = (44.12 ; 43.24)$. Following the thread of this case, an automated procedure could present difficulties to move from one state to another. The only solutions are to increase the size of the training sample or estimate a new operation point not belonging to the real database. The expert engineering plant knowledge must be used to fill this gap. In future research, after the systematization of the human expert knowledge about economical constraint (i.e., as a rule set or an expert system) an automated procedure would lead the entire consumption optimization process.

6. Commercial application

The implementation of the approach proposed in this paper is the core of a pilot software developed by an engineering software company (ALIATIS). By means of an Oracle BI interactive Dashboard, the mining framework is updated from the enterprise database, helping experts to search for a new optimal



			Hornos: Consumo FG (Nm3/h)			Reactores: Tª Entrada (ºC)			Separador V8		Desbutanizadora			Desbencenizadora			Coste (euros/h)
Pasos	Tipo	Histórico	H.3	H.4	H.5	RX.5	RX.6	RX.7	Presión (bar)	Flujo a V11 (m3/h)	Presión (bar)	Relación Reflujo	Consumo FG H6 (Nm3/h)	Presión (bar)	Relación Reflujo	Consumo VM Hervidor (Kg/h)	Total
0	Actual	18/05/2010 23:00:00	1.074	639	433	537	531	531	14,5	58,5	15,9	2,3	586	0,8	0,9	3.609	1.369
1	Histórico	19/04/2010 21:00:00	895	647	335	537	531	528	15,4	54,2	14,6	2,7	520	0,6	0,7	4.226	1.229
2	Histórico	13/05/2010 23:00:00	917	562	325	535	531	528	16,1	43,2	16,4	2,4	485	0,7	0,8	4.395	1.183
3	Histórico	28/09/2009 5:00:00	891	496	287	528	514	510	14,0	38,1	15,0	2,9	445	0,7	0,9	3.297	1.075
4	Histórico	25/09/2009 0:00:00	772	502	270	525	513	509	14,0	32,1	15,1	3,4	365	0,9	0,7	2.856	966
5	Histórico	25/09/2009 2:00:00	772	484	222	524	513	508	14,0	32,5	15,1	3,9	364	0,9	0,7	2.827	934
6	Histórico	24/09/2009 15:00:00	770	437	174	525	513	509	14,0	31,7	15,1	4,4	358	0,9	0,8	3.007	890
7	Histórico	24/09/2009 14:00:00	754	421	171	525	513	509	14,0	31,6	15,1	5,0	361	0,8	0,8	3.215	881
8	Óptimo	30/04/2010 16:00:00	273	449	452	536	528	524	15,1	46,9	14,6	3,8	440	0,8	1,1	4.127	861

Fig. 12. Screen of costs from the commercial mining-based DSS.

Courtesy of ALIATIS.

operation point from the present one. This application uses a gateway in the sense of central access points to field-level data. It has become apparent in recent years that a direct connection to field-level networks causes security problems. Several authors recommend gateways as an appropriate means to cope with such security problems and to implement access control (Cheminod, Pironti, & Sisto, 2011).

The first screen of the commercial application shows the present state of the plant and the optimum efficiency point calculated from the constraint-based operator's knowledge. It also includes the present proposed saving cost and the DSS save report, through an energy efficiency operation screen (Fig. 12). Differences between the present and the optimal point allows one to calculate the energy savings easily (daily, weekly and yearly savings could be presented). This cost information is shown for the entire plant and also for every part (heaters) of the plant. The evolution of attributes and the savings from the heaters are displayed by graphics and tables. In the example that presents Fig. 12, a daily savings of 1369€/h to 861€/h as well as the step by step sequence to follow are shown. A simulation of the full testing sample indicates a rise in the energy efficiency between 35% and 45%.

7. Conclusion

This paper discusses the main attributes of an aromatic distillation unit in petrochemical industries to suggest a practical way for consumption optimization. A biographical review has been made, and a new framework is presented to find relevant knowledge about the particular characteristics of the distillation process and to describe the main features available.

The authors present an innovative data mining framework that uses the historical plant database to make the most of the expert knowledge. Every operation points used are points recorded in a historical database that presents an optimization program of the plant features, but not in the consumption optimization sense.

The mining process creates a characterization of the plant operation points based on the most relevant attributes previously selected. This classification can be used in two ways: to assign new points to existing classes (very high, high, low or very low $EEI_{REACTION-SDBIN}$ value) and to move the present operation point.

The main contribution of this paper is to suggest how to move from the present operation point to a close historical point that could improve energy efficiency. Both points, present and proposed points, carry out the industrial objective: high production rate with required product quality. The second one searches an optimal consumption as well. The human expert would choose between several optimal points once the industrial and economical constraints are fixed. This mining process optimization could be repeated just to achieve the minimum value of the energy efficiency indicator (EEI). In this sense, as a new contribution of this paper, the real industrial process constraints to the attribute values are considered "a posteriori". The model highlights the special importance of the use of real constraints.

The quality of this framework is illustrated by a case study that uses a real database. The framework presented in this paper is now in the testing phase. The previous results obtained were satisfactory considering the limitation of the available database. A rise in the energy efficiency from 35% to 45% is significantly improved from the previous company process. Also, the required product quality is maintained. As previously stated, as future research the human expert knowledge will be collected into an artificial intelligence-based expert system in order to fully automatize the mining framework.

Acknowledgment

The authors thank the *Corporación Tecnológica de Andalucía* (CTA) for providing the funds for this project.

Appendix A. Main attributes in the catalytic reforming

WAIPTF (°C): The variable that measures the catalyst deterioration.
TPRIN_PPPV7 (°C): The input PPV7 reactor's temperature.
TPROUT_PPV6 (°C): The output PPV6 reactor's temperature.
TPROUT_PPV5 (°C): The output PPV5 reactor's temperature.
TPRIN_PPV7 (°C): The input PPV7 reactor's temperature.
TPRIN_PPV6 (°C): The input PPV6 reactor's temperature.
TPRIN_PPV5 (°C): The input PPV5 reactor's temperature.
CSMFGPPH3 (m^3/h): The fuel gas PPH3 heater's consumption.
CSMFGPPH4 (m^3/h): The fuel gas PPH4 heater's consumption.
CSMFGPPH5 (m^3/h): The fuel gas PPH5 heater's consumption.
PTFINFLOW (m^3/h): The platforming input flow.
P_PPV8 (bar): The PPV8 product separator pressure.
TPROOM (°C): The room temperature.
DENRECGAS (kg/(Nm^3)): The recycle gas density. This variable maintains *P_PPV8* brought under control.
TPRPPV567 (°C): The temperature increase between the three reactors.

References

Alhajree, I., Zahedi, G., Manan, Z., & Zadeh, S. M. (2011). Modeling and optimization of an industrial hydrocracker plant. *Journal of Petroleum Science and Engineering*, *78*(3–4), 627–636.

Buzzi-Ferraris, G. (2011). New trends in building numerical programs. *Computers and Chemical Engineering*, *35*(7), 1215–1225.

Buzzi-Ferraris, G., & Manenti, F. (2011). Outlier detection in large data sets. *Computers and Chemical Engineering*, *35*(2), 388–390.

Cheminod, M., Pironti, A., & Sisto, R. (2011, February). Formal vulnerability analysis of a security system for remote fieldbus access. *IEEE Transactions on Industrial Informatics*, *7*(1), 30–40.

Chiwewe, T., & Hancke, G. (2011). A distributed topology control technique for low interference and energy efficiency in wireless sensor networks. *IEEE Transactions on Industrial Informatics*, *8*(1), 11–19.

Dave, D. J., Dabhiya, M. Z., Satyadev, S. V. K., Ganguly, S., & Saraf, D. N. (2003). Online tuning of a steady state crude distillation unit model for real time applications. *Journal of Process Control*, *13*(3), 267–282.

de Lima, R. S., & Schaeffer, R. (2011). The energy efficiency of crude oil refining in Brazil: A Brazilian refinery plant case. *Energy*, *36*(5), 3101–3112.

Frenkel, M. (2011). Thermophysical and thermochemical properties on-demand for chemical process and product design. *Computers and Chemical Engineering*, *35*(3), 393–402.

Ghashghaee, M., & Karimzadeh, R. (2011). Multivariable optimization of thermal cracking severity. *Chemical Engineering Research and Design*, *89*(7), 1067–1077.

Gong, R., Huang, S., & Chen, T. (2008, August). Robust and efficient rule extraction through data summarization and its application in welding fault diagnosis. *IEEE Transactions on Industrial Informatics*, *4*(3), 198–206.

Gueddar, T., & Dua, V. (2012). Novel model reduction techniques for refinery-wide energy optimisation. *Applied Energy*, *89*(1), 117–126. Special issue on Thermal Energy Management in the Process Industries.

Han, J., & Kamber, M. (Eds.). (2001). *Data mining. Concepts and techniques*. Morgan Kaufmann.

Hippert, H., Pedreira, C., & Souza, R. (2001, February). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, *16*(1), 44–55.

Iranshahi, D., Bahmanpour, A., Paymooni, K., Rahimpour, M., & Shariati, A. (2011). Simultaneous hydrogen and aromatics enhancement by obtaining optimum temperature profile and hydrogen removal in naphtha reforming process; a novel theoretical study. *International Journal of Hydrogen Energy*, *36*(14), 8316–8326, 4th Asian Bio-Hydrogen Symposium.

Iranshahi, D., Rahimpour, M., & Asgari, A. (2010). A novel dynamic radial-flow, spherical-bed reactor concept for naphtha reforming in the presence of catalyst deactivation. *International Journal of Hydrogen Energy*, *35*(12), 6261–6275.

Jabbar, N. A., & Alatiqi, I. (1997). Inferential-feedforward control of petroleum fractionators: A PNA approach. *Computational Chemical Engineering*, *21*, 255–262.

Jana, A. (2010, February). A hybrid flc-ekf scheme for temperature control of a refinery debutanizer column. *IEEE Transactions on Industrial Informatics*, *6*(1), 25–35.

Jana, A. K., Samanta, A. N., & Ganguly, S. (2009). Nonlinear state estimation and control of a refinery debutanizer column. *Computers and Chemical Engineering*, *33*(9), 1484–1490.

Julka, N., Karimi, I., & Srinivasan, R. (2002). Agent-based supply chain management: A refinery application. *Computers and Chemical Engineering*, *26*(12), 1771–1781.

Julka, N., Srinivasan, R., & Karimi, I. (2002). Agent-based supply chain management framework. *Computers and Chemical Engineering*, *26*(12), 1755–1769.
Kansha, Y., Kishimoto, A., & Tsutsumi, A. Application of the self-heat recuperation technology to crude oil distillation. Applied Thermal Engineering, in press.
Kuo, T.-H., & Chang, C.-T. (2008). Optimal planning strategy for the supply chains of light aromatic compounds in petrochemical industries. *Computers and Chemical Engineering*, *32*(6), 1147–1166.
Kusiak, A., & Song, Z. (2006, August). Combustion efficiency optimization and virtual testing: A data-mining approach. *IEEE Transactions on Industrial Informatics*, *2*(3), 176–184.
Li, X., Bowers, C., & Schnier, T. (2010, November). Classification of energy consumption in buildings with outlier detection. *IEEE Transactions on Industrial Electronics*, *57*(11), 3639–3644.
Liao, S.-H. (2005). Expert system methodologies and applications – A decade review from 1995 to 2004. *Expert Systems with Applications*, *28*(1), 93–103.
Liau, L. C.-K., Yang, T. C.-K., & Tsai, M.-T. (2004). Expert system of a crude oil distillation unit for process optimization using neural networks. *Expert Systems with Applications*, *26*(2), 247–255.
Maimon, O., & Rokach, L. (Eds.). (2010). *Data mining and knowledge discovery handbook*. Springer US.
Meidanshahi, V., Bahmanpour, A. M., Iranshahi, D., & Rahimpour, M. R. (2011). Theoretical investigation of aromatics production enhancement in thermal coupling of naphtha reforming and hydrodealkylation of toluene. *Chemical Engineering and Processing: Process Intensification*, *50*(9), 893–903.
Metaxiotis, K., Kagiannas, A., Askounis, D., & Psarras, J. (2003). Artificial intelligence in short term electric load forecasting: A state-of-the-art survey for the researcher. *Energy Conversion and Management*, *44*(9), 1525–1534.
Metzger, M., & Polakow, G. (2011, November). A survey on applications of agent technology in industrial process control. *IEEE Transactions on Industrial Informatics*, *7*(4), 570–581.
Miyayama, T., Tanaka, S., Miyatake, T., Umeki, T., Miyamoto, Y., Nishino, K., et al. (1991). A combustion control support expert system for a coal-fired boiler. In *Proceedings of the 1991 international conference on industrial electronics, control and instrumentation, IECON'91, October–November, 1991, vol. 2* (2nd ed., pp. 1513–1516).
Motlaghi, S., Jalali, F., & Ahmadabadi, M. N. (2008). An expert system design for a crude oil distillation column with the neural networks model and the process optimization using genetic algorithm framework. *Expert Systems with Applications*, *35*(4), 1540–1545.
Ogilvie, T., Swidenbank, E., & Hogg, B. (1998). Use of data mining techniques in the performance monitoring and optimisation of a thermal power plant. In *IEE colloquium on knowledge discovery and data mining (1998/434), May* , pp. 7/1–7/4
Ouattara, A., Pibouleau, L., Azzaro-Pantel, C., Domenech, S., Baudet, P., & Yao, B. (2012). Economic and environmental strategies for process design. *Computers and Chemical Engineering*, *36*, 174–188.
Rahimpour, M., Vakili, R., Pourazadi, E., Iranshahi, D., & Paymooni, K. (2011). A novel integrated, thermally coupled fluidized bed configuration for catalytic naphtha reforming to enhance aromatic and hydrogen productions in refineries. *International Journal of Hydrogen Energy*, *36*(4), 2979–2991.
Rahimpour, M. R., Iranshahi, D., & Bahmanpour, A. M. (2010). Dynamic optimization of a multi-stage spherical, radial flow reactor for the naphtha reforming process in the presence of catalyst deactivation using differential evolution (de) method. *International Journal of Hydrogen Energy*, *35*(14), 7498–7511.
Robertson, G., Palazoglu, A., & Romagnoli, J. (2011). A multi-level simulation approach for the crude oil loading/unloading scheduling problem. *Computers and Chemical Engineering*, *35*(5), 817–827. Selected Papers from ESCAPE-20 (European Symposium of Computer Aided Process Engineering – 20), 6–9 June 2010, Ischia, Italy.
Seo, J. W., Oh, M., & Lee, T. H. (2000). Design optimization of a crude oil distillation process. *Chemical Engineering and Technology*, *23*(2), 157–164.
Shiau, W.-L. (2011). A profile of information systems research published in expert systems with applications from 1995 to 2008. *Expert Systems with Applications*, *38*(4), 3999–4005.
Song, Z., & Kusiak, A. (2007, February). Constraint-based control of boiler efficiency: A data-mining approach. *IEEE Transactions on Industrial Informatics*, *3*(1), 73–83.