Cutpoint Temperature Modeling using a Coefficient Setup MIQP Technique for Distillation Yields and Properties

Robert E. Franzoi[a], *Brenno C. Menezes*[b], *Jeffrey D. Kelly*[c], *Jorge A. W. Gut*[a,*], *Ignacio E. Grossmann*[d]

[a] Department of Chemical Engineering, University of São Paulo, São Paulo, Brazil.

[b] Division of Engineering Management and Decision Sciences, College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

[c] Industrial Algorithms Ltd., 15 St. Andrews Road, Toronto, Canada.

[d] Chemical Engineering Department, Carnegie Mellon University, Pittsburgh, United States.

*Corresponding Author. E-mail address: jorgewgut@usp.br (J. A. W. Gut)

ABSTRACT

For high-performance operations in crude oil refinery processing, it is important to properly determine output stream yields and properties from distillation units. To address such complex representation, we propose a cutpoint temperature modeling framework using a coefficient setup MIQP (mixed-integer quadratic programming) technique to determine optimizable surrogate models to correlate independent X variables (crude oil compositions, temperatures) to dependent Y variables (yields and properties of distillates). The X inputs are generated by Latin Hypercube Sampling (LHS) and the experiments to obtain the synthetic Y outputs are simulated using the well-known conventional and improved swing-cut methods. Using the optimizable surrogate model determined in the MIQP, which is suitable to handle continuous data from the process and to use measurement feedback for adjustments and improvements, online outputs can be continuously updated. Furthermore, such updates of X and Y datasets can be used to re-identify the X-Y correlation-causation basis of the optimizable surrogate model. This MIQP surrogate identification technique may also be applied to other types of downstream process optimization problems such as reacting and blending unit-operations, as well as other separating processes.

Keywords: Crude oil assays; Distillation; Machine learning; Coefficient setup; Temperature cut-points; Nonlinear programming.

1. Introduction

Crude oil distillation units (CDUs) are complex sets of towers operated to separate liquid hydrocarbon feedstocks or crude oil raw materials into intermediate fractions or distillates according to boiling range temperatures.[1] As these distilled streams are processed in downstream unit-operations and blended into final products, for an overall high-performance operation of the refinery, it is important to precisely calculate the distillation unit's product yields and properties as a function of feed quality and operating conditions.[2] Both rigorous and surrogate models can be used to predict product amounts and properties of distillation processes for planning, scheduling, multi-unit coordinating, and real-time optimization (RTO) environments as well expectation monitoring. The rigorous, mechanistic, physics-based, first principles, white-box or engineering-based modeling typically consider molar, mass, energy, separation and equilibrium balances in the distillation columns. Compositions, flows and processing conditions may be accurately determined, but at a high computational effort, imposing convergence issues for their application in large-scale integrated problems. On the other hand, non-rigorous, black-box or empirical-based modeling can use surrogate or simplified shortcut correlations based on measured and/or synthetic data and data regression techniques. Due to their simplicity, effectiveness and acceptable accuracy within a localized region, surrogate modeling is commonly used for process optimization in oil refineries.[3]

In order to calculate CDU yields and cold-flow properties using non-rigorous models, one can use the temperature distribution from the crude oil true boiling point (TBP) curve, which represents how the crude oil yields and properties (such as specific gravity, sulfur content, etc.) vary with distillation temperature.[2,4] The TBP curve is related to the crude oil assay, which provides data on the quantities and qualities of each discretized temperature cut or micro-cut range through their distillation temperature distribution.[5] Due to operational limitations and inefficiencies regarding reflux and re-circulation rates, number of stages, etc., there is a well-known overlap in the TBP boiling ranges of adjacent fractions or compounds in any physical distillation column.[3] Therefore, this non-sharp fractionation between adjacent distillates should be considered to properly formulate cutpoint optimization methods.

By observational evidence in the oil refinery, the bulk quality of raw material composition (assay) of crude oils to be processed typically determines around 80 to 90% of the amounts and properties of the distillates, whereas the remaining part is determined by its operational variables such as internal reflux rates, system pressure profile, steam flows of side-strippers, pump-around re-circulation rates, parallel split ratios of pre-heat exchanger trains, feed and location temperatures in the furnace and tower, tray and/or packing characteristics, etc. Moreover, lower and upper bounds on quality specifications (circa 30 distinct types of properties such as specific gravity, sulfur concentration, acidity, carbon residue content, etc.) can be considered in the cutpoint temperature model. Due to uncertainties in the transformation equations, feed compositions and other processing data, high-fidelity modeling becomes unrealistic to be included in problems such as the blend scheduling and processing optimization of crude oils.

In addition to the above considerations, we are currently moving towards a more complex process optimization age in Industry 4.0,[6] driven by advancements in decision-making modeling, computer-aided capabilities, connectivity and solving algorithms. In this direction, big data,[7,8,9,10] data-driven models,[11,12,13,14,15] and machine learning techniques[16,17,18] have been used in a wide variety of engineering problems: a) to handle large, complex, and unreliable data sets; b) as a better or more efficient alternative to solve particular problems; and c) to solve problems which are either intractable or that require faster solutions for specific applications. Production and process optimization for the crude oil refinery industry typically handle large, complex, nonlinear, and nonconvex models. Thus, modeling and optimizing a fully integrated petroleum refinery problem is still not yet attainable in terms of complexity and uncertainty. Data from the plant is not always accurate and contains various noise levels, and model-plant mismatches can often impose issues either in model infeasibilities and solution implementation. Employing online measurements in the entire plant can help to mitigate errors and provide better fidelity; however, that is typically not as effective as identifying and estimating better local models. Hence, error propagation becomes significant, especially because of the nonlinearities associated with crude oil refinery problems.

In such context, the framework proposed in this work establishes surrogate models based on process analytics and machine learning techniques to reduce uncertainties around the determination of cutpoints in distillation units when incrementally optimizing and controlling the stream flow, yield and properties drawn from the CDU at a certain section in the tower. The main contributions of the proposed model are: a) it is as representative and accurate as typical models used for planning and scheduling environments such as the swing-cut methods[3,5]; b) real data from the plant can be embedded into the model to give it self adjustability and self improvability over time, minimizing the impact of uncertainties and disturbances in the process by correlated parameter updating (with re-estimation of coefficients); c) it is small in size, with fewer equations and degrees of freedom than the swing-cut models, and can be properly integrated into planning, scheduling, coordinating and RTO environments with minimal increase in the simulation and optimization effort.

This paper is outlined as follows. In Section 2, an overview of the cutpoint optimization approaches reported in the literature is presented. The problem statement is described in Section 3. The proposed algorithm for the identification of surrogate process analytics formulae known as coefficient-setup technique is presented in Section 4, whereby the swing-cut modeling (conventional and improved) as well as the inputs-outputs or X-Y data blocks or sets are given in the Supporting Information. Examples using the proposed algorithm for linear and interaction terms of bilinear correlations of the surrogate model are compared in Section 5. The conclusions and future work are discussed in Section 6.

2. Previous Shortcut Distillation Methods

The simplest approach to model a distillation unit is to use fixed yield and property values for its outputs, specifying the increments of the discretized CDU fractions and using crude oil assay data to calculate these fractions.[19] As this approach uses sharp fractionation to calculate the CDU yields, it does not compute rigorously the molecular behavior considering the non-perfect or non-sharp separation. A variation of the fixed yield model, known as multiple fixed yields, allows multiple and usually hypothetical operational modes in the process, whereby each mode is related to a distinct crude oil assay and hence, different yields and properties for the final cuts. This method creates an additional degree of freedom, which is represented by binary variables. Such improvement allows only a small number of different solutions (equal to the number of operational modes), despite the feasibility of intermediate solutions that lie between the range of pre-defined modes. Furthermore, crude oil refinery optimization problems are typically highly nonlinear and nonconvex due to the blending of streams and inventories throughout the processing network of unit-operations and intermediate tanks. The introduction of binary variables would lead to a nonconvex mixed-integer nonlinear programming (MINLP) problem, which is hard to solve for medium to large-scale cases. Brooks et al.[19] optimized product yields for a crude oil distillation unit by introducing eight pre-defined modes of operation. Each mode had a distinct choice of cutpoints and their approach allowed to blend the outputs of distinct modes to achieve required yields and properties of the final distillates. To handle the complex nature of the problem, the authors employed tabulated values of yields and properties of intermediate products as linear model for the distillation unit.

To better predict crude oil distillation unit outputs, there are models that consider non-sharp fractionation by including volume and/or mass variations for the cutpoints. A traditional empirical approach is known as delta-based or shift-vector modeling and uses small increments for product deviations in the TBP curve representing only first-order or linear effects. As example, there are the swing-cut methods,[3,5,20] which require both the TBP range for each product and the estimation of the size of each swing-cut. This information can be combined with the crude oil quality in the respective TBP range to calculate the properties of each distillate.[2] In addition, the swing-cut method creates additional degrees of freedom by optimizing the amount of each swing-cut to the lighter and heavier final distillates. The swing-cut methods became commonly used for distillation unit modeling due to their simplicity[21] and improvements when compared to the fixed yields method.

Zhang et al.[20] proposed a swing-cut method that considers operational conditions and feed properties in the distillation unit as variables in the optimization. However, this method considers the properties of adjacent distilled products fixed regardless the amounts and properties of the swing-cut splits added to the distillates, failing to represent the high non-linearity of the distillation process. The weight transfer ratio method (WTR), proposed by Li et al.[3], considers crude oil characteristics and products' yields and qualities in simplified empirical nonlinear models. The authors used an empirical procedure to calculate the mass transfer rates of each product in the CDU and to determine the size of each swing-cut. In addition, the authors used regression models based on the properties of the feed load to consider the variation of properties in each swing-cut. However, due to the possibility of processing more than one type of crude oil simultaneously, additional procedures are required to calculate the TBP curve of the crude oil mixture.

To deal efficiently with the variation of properties within the swing-cut, Menezes et al.[5] improved the traditional or conventional swing-cut (ISW) method by dividing each swing-cut into light and heavy fractions using arbitrary 10 °C increments for cuts. This method adds property information for both light and heavy split or swing fractions, in addition to the flow variables and nonlinear balance constraints. The properties of each fraction are calculated individually using interpolated quality information regarding their respective split quantities. This method predicts more accurately the properties of distillation unit outputs.

Alattas et al.[22] proposed a simplified nonlinear model for the distillation unit to be used in a production planning environment. The distillation unit is represented as flash towers operating in series. Fractionation indices (FI) are introduced for each layer of flashes and are calculated using characteristics of the columns such as temperature distribution. The model uses the fractionation indices and molar balances to predict the distillation tower operations more accurately than the conventional swing-cut models. Alattas et al.[23] improved their previous methodology by formulating the disjunction of the fractionation index with mixed-integer constraints, leading to a faster and more robust model. The authors also extended the application of their model to a multiperiod refinery planning problem based on an MINLP formulation.

Mahalec and Sanchez[24] proposed a hybrid model based on first principles considering mass and energy balances to optimize distillation unit towers. In the model, operational variables are correlated to product distillation curves by using partial least-square models, and to estimate the deviation between initial and end sections of the curve. By relying partially on a statistical modeling, the method manages to reduce the prediction error of the fractions in the distillation unit. However, the increment used to set the cutpoints (between 14 and 67 °C) may not be small enough to give a sufficiently good accuracy.

Kelly et al.[4] used a monotonic interpolation method to avoid Runge's phenomenon (oscillation at the edges of an interval when constructing a polynomial interpolant of high degree). First, this procedure uses analytical expressions to convert experimental methods to TBP temperatures. Second, monotonic interpolation converts the TBP temperatures into cumulative evaporations, which are blended linearly by mass or volume, and converted back using another monotonic interpolation. Third, a cutpoint temperature optimization is performed to adjust the front and back end of each distillation curve component using measured field or laboratory ASTM distillation data. Four case studies were provided, in which good agreement is shown between predicted and real blending properties.

Fu et al.[25] proposed a hybrid model to optimize a three-tower distillation unit. Partial least squares from the feed TBP curve and operational conditions were used to predict product TBP curves. Combined with volumetric and energy balances, this enables predictions with small discrepancies when compared to rigorous simulation. In their model, operating variables are used to compute the distances from the middle line of a product TBP curve, instead of relying on the internal reflux on selected trays since tray temperatures are not required for monitoring or optimization. This hybrid model has small size and good convergence, being suitable for planning, scheduling, coordinating and RTO environments.

For better predictions of process-shop yields and properties, Franzoi et al.[26] use predictive analytics techniques by doing constrained and weighted least squares to fit better base plus delta or shift-vector sub-models using data reconciliation and regression techniques. The reconciliation enforces the consistency of yields and the regression fits base and delta coefficients simultaneously across all yields. The proposed hybrid cutpoint optimization approach can be applied to online optimization of crude oil blend scheduling operations in complex industrial-sized refineries to determine the composition-quality feed demands for the amounts and properties of distillates in towers in cascade (as a real process equipment design). According to Kelly and Zyngier,[27] a continuous cycle of improvements can use process measured feedback, which leads to higher accuracy and reliability, and aims to reduce the gap between the model predictions and the actual plant values.

Cutpoint temperature modeling has also been used for energy efficient operations of crude oil distillation units. Durrani et al.[28] proposed a hybrid artificial neural network model based on the

Taguchi method and genetic algorithm to handle uncertainties in the crude oil feed compositions in order to reduce energy costs. Atmospheric distillation units may represent more than 25% of the potential of energy savings in crude oil refineries. Therefore, an efficient cutpoint temperature modeling saves a substantial amount of energy typically lost during the process.

Most cutpoint temperature methods presented in the literature do not take into account the highly dynamic and uncertain real process environment as well as the typical plant versus model mismatches. In such context, the methodology proposed in this work focuses in building an accurate cutpoint temperature model, suitable for planning, scheduling, coordinating and real time applications, and that uses continuous and real data from any reliable sources such as the production plant or rigorous simulation to improve the predictions of crude oil distillation units.

3. Problem Statement

For the sake of simplicity in the presentation, and to motivate the ideas behind the proposed method, we consider a specific case in which there are four crude oils (CO1 to CO4) feeding a crude distillation unit (CDU), which produces seven final cuts: fuel gas (FG), liquefied petroleum gas (LPG), naphtha (N), kerosene (K), light diesel (LD), heavy diesel (HD), and atmospheric residue (ATR). The yields and properties of the distillates are calculated using different methods: a) fixed yield (FY); b) conventional swing-cut (CSW); and c) improved swing-cut (ISW). Figure 1a shows the process flowsheet within a UOPSS (unit-operation-port-state superstructure) representation[29] for the distillation example in which swing-cuts are not considered, while Figure1b represents a scenario with three swing-cuts between naphtha and kerosene, kerosene and light diesel, light diesel and heavy diesel for CSW and ISW. The capacities for the crude oil pools and final product pools are 100 Mbbl. The maximum flowrate for the distillation unit is 100 Mbbl/day. The crude oil assay data is embedded in the optimization problem as well.

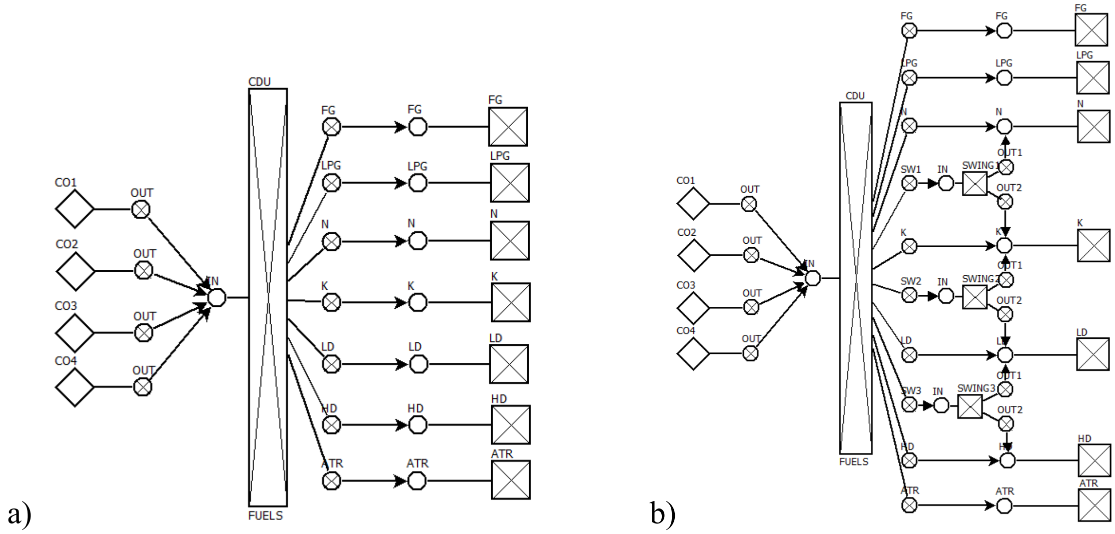


Figure 1: Crude oil distillation unit flowsheet a) without swing-cuts and b) with swing-cuts

The swing-cuts shown in Figure 1b represent hypothetical flows and they are used for modeling and optimization purposes only. In real distillation unit operations, variables such as flow, temperature and pressure are adjusted in the plant in order to control the production of fuels. For example, decreasing or increasing the trays' temperature profile changes the production of each distillate which has an important role on the refinery economics. In this work, we do not directly model these process variables, but the swing-cuts present the same purpose or degree-of-freedom as an outcome of the operational variations within the distillation column.

For instance, let us consider the hypothetical swing-cut between naphtha and kerosene. If this swing-cut splits equally between naphtha and kerosene in the optimization problem, the operational conditions in a real process should be chosen to meet this condition. In that case, the naphtha/kerosene swing-cut split fraction sw_1 is a proxy for the naphtha endpoint cutpoint temperature and relates the swing-cut naphtha (light-key) flow Q_N^{sw} to the swing-cut kerosene (heavy-key) flow Q_K^{sw}. Depending on the value of sw_1 chosen by the optimization, the swing naphtha/kerosene cutpoint temperature $T_{N/K}$ (to be adjusted in the actual unit in the tower) can be calculated as shown in Equations (1) and (2), in which T_N and T_K are the final cutpoint temperature of naphtha and the initial cutpoint temperature of kerosene, respectively. Because of the assumption of perfect- or sharp-separation, the final naphtha cutpoint equals the initial kerosene cutpoint.

$$sw_1 = \frac{Q_N^{sw}}{(Q_N^{sw} + Q_K^{sw})} = \frac{(T_{N/K} - T_N)}{(T_K - T_N)} \tag{1}$$

$$(1 - sw_1) = \frac{Q_K^{sw}}{(Q_N^{sw} + Q_K^{sw})} = \frac{(T_K - T_{N/K})}{(T_K - T_N)} \tag{2}$$

The final flows for naphtha and kerosene are their outlet CDU flows summed to their respective swing-cuts parts, as shown in Equations (3) and (4).

$$Q_N^{final} = Q_N^{cdu} + Q_N^{sw} \tag{3}$$

$$Q_K^{final} = Q_K^{cdu} + Q_K^{sw} \tag{4}$$

Substituting Equations (1) and (2) into Equations (3) and (4), the naphtha and kerosene final flows can be rewritten as a function of the naphtha/kerosene cutpoint temperature $T_{N/K}$:

$$Q_N^{final} = Q_N^{cdu} + \frac{(T_{N/K} - T_N)}{(T_K - T_N)}(Q_N^{sw} + Q_K^{sw}) \tag{5}$$

$$Q_K^{final} = Q_K^{cdu} + \frac{(T_K - T_{N/K})}{(T_K - T_N)}(Q_N^{sw} + Q_K^{sw}) \tag{6}$$

4. Proposed distillation cutpoint modeling

Three distillation unit models from the literature are reproduced in this work, which are described in the Supporting Information: a) fixed yield (FY); b) conventional swing-cut (CSW); and c) improved swing-cut (ISW). Moreover, a novel distillation model, based on data analytics identification and estimation using MIQP coefficient setup techniques to determine optimizable surrogate models, is proposed in this section. For the proposed distillation cutpoint model, the outputs (yields and properties) of the Y dataset are determined using the CSW and ISW models.

We propose a cutpoint temperature modeling framework using a data-driven coefficient setup MIQP technique to determine optimizable surrogate models to correlate variations of independent or X variables to dependent or Y variables inferred from X-Y datasets. This machine learning linear regression methodology focuses on establishing simple yet reliable correlations to estimate the yields and properties outputs from actual distillation units. These surrogate models are built from experiments, process simulations or any other reliable source of data, and can accurately predict the outputs of processes in which there is missing/uncertain data and that are typically very complex and require high effort to be simulated or obtained. Moreover, they can reduce the impact of variations in the crude oil assay and uncertainties in the process since they are predicted from a data-driven methodology (using crude-oil composition as inputs only from the crude-oils) rather than based on crude-oil assay (that can be out-of-date).

The proposed approach uses simulated or synthetic data from the conventional and improved swing-cut methods. In real processes, data from the plant or generated by calibrated rigorous process simulation or by any simpler yet accurate modeling procedure can be used. If using experimental or actual data from the field, the match of outputs (yields and properties) and inputs (crude oil composition, operational variables) using the proposed data-driven machine learning approach would eliminate the necessity of the distillation curves or distribution of yields and properties of the crude oil assays. Then, once identified the X-Y or input-output correlations, the estimation of the parameters or coefficients may be updated using active or passive historical data.

The proposed methodology is implemented and tested using the crude oil distillation or fractionation example presented in Figure 1b in which four crude oils feed a distillation unit to produce seven final products. The framework for the proposed model, shown in Figure 2, is implemented in Python 3 using the environment Microsoft Visual Studio 2015, and it is integrated to the Microsoft Excel to provide a more user-friendly approach for data manipulation and for better visualization of results. The modeling platform used is IMPL (Industrial Modeling & Programming Language), and the MIQP optimizations are carried out through the commercial solvers GUROBI 8.1.0 and CPLEX 12.8.0 connected to IMPL. The machine used was an Intel Core i7 with 2.90 GHz and 16 GB RAM.

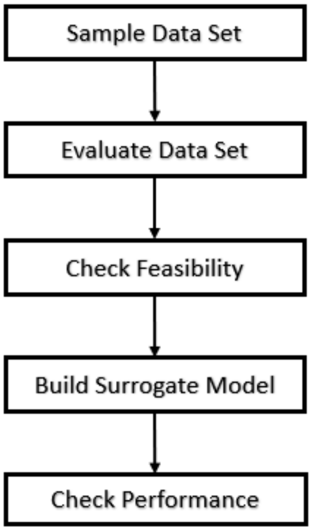


Figure 2: Framework for the proposed strategy

In summary, this methodology: a) builds a data set using randomly generated data; b) calculates or simulates the process variables of interest, which in this case are the outputs of the distillation unit, using this data set; c) checks the feasibility status for the optimal solutions previously found; d) builds or identifies a surrogate model which properly fits the data; and e) checks the performance of the surrogate model found by calculating the average mean square error between the surrogate model and the data set. Each step of the framework shown in Figure 2 is explained in detail as follows.

Sample Data Set: The proposed framework uses the well-established Latin Hypercube Sampling (LHS) technique to randomly sample points for the independent variables in order to build a data set to be used to identify the desired surrogate model. In order to provide a better insight into the reliability and robustness of our methodology, two randomly generated data sets have been created. As example, the Data Set 1 generated for the independent variables is presented in the Supporting Information (Table A1). Each data set is used in both the CSW and the ISW models. Thus, there are four different cases:

- Case 1: Data Set 1 using the CSW method;
- Case 2: Data Set 1 using the ISW method;
- Case 3: Data Set 2 using the CSW method;
- Case 4: Data Set 2 using the ISW method.

There are seven independent variables, related to the compositions of the four crude oils (x_1 to x_4) and the three swing-cut splits (x_5 to x_7). The variables x_5, x_6 and x_7 represent the light fractions for each swing cut, so that information for the heavy fractions is not accounted as the light and heavy fractions of each swing-cut are complementary to the unity. One hundred samples are generated in this step using LHS. In each sample, these seven independent variables are randomly created, respecting the composition consistency represented by Equation (7), in which the sum of compositions of all crude oils found in the set CR must be equal to the unity.

$$\sum_{j \in CR} x_j = 1 \tag{7}$$

Evaluate Data Set: The initial data sets are evaluated to calculate the yields and properties of the final distillation cuts. For that, Equations (A1) to (A15) are employed in an optimization problem respecting quality constraints (product specifications). The randomly generated independent variables, and the crude oil assay data are known. Although we do not need to optimize such problem (instead, we could calculate the final variables directly using Equations (A1) to (A15)), the optimization is useful to detect and avoid poor selection of feedstocks that would eventually result in infeasible solutions, potentially with unspecified final products. As example, the values of the dependent variables calculated using the ISW method based on Data Set 1 are shown in the Supporting Information (Table A2 for yields, Table A3 for specific gravity, and Table A4 for sulfur content).

Check Feasibility: A feasibility check is performed over all optimal solutions from the previous step. Infeasible or inconsistent solutions and low-quality sub-optimal solutions with objective functions 30% worse than the best solution found, are removed.

Build Surrogate Model: The final pool of solutions is used to train or build a surrogate model for the yields and properties of each distillate stream. Each model is a function of the independent variables of crude oils and of its respective swing-cuts. For example, the yields and properties of naphtha stream vary with the crude oil composition and with the naphtha-kerosene swing-cut (i.e., swing-cut 1 in Figure 1b). These surrogate models are intended to be a simple yet accurate correlation to replace the swing-cut models (or any other distillation unit model). For that, X basis are introduced in the problem to account for the independent variables. For each distillation component $i \in DC$ (i.e., any yield or property of a distillation cut), there are four linear basis l_{ji} to account for each crude oil $j \in CR$ (x_1, x_2, x_3, x_4), and three linear basis l_{ki} to account for each swing cut $k \in SW$ (x_5, x_6, x_7). There are twelve bilinear basis b_{jki} for the relations between them, i.e., for the products of coefficients of each crude oil by each swing-cut: $x_1 x_5,\ x_1 x_6,\ x_1 x_7,\ x_2 x_5,\ x_2 x_6,\ x_2 x_7,\ x_3 x_5,\ x_3 x_6,\ x_3 x_7,\ x_4 x_5,\ x_4 x_6,\ x_4 x_7$, also known as interaction second-order effect terms. Bilinear basis in which both coefficients represent crude oils (or if both represent swing-cuts) are not considered since these have no physical or relatable meaning. An intercept coefficient (I_i) has also been used in the model to account for any possible behavior not related to crude oils and swing-cuts.

For each case, six distinct models are proposed, in which three types of coefficients are employed. There are the intercept coefficients, which are not associated to any basis; the linear coefficients, which are multiplied for a linear basis (related to either a crude oil or a swing-cut); and the bilinear or interaction coefficients, which are multiplied by two basis, becoming a second order term in the equation. For the bilinear coefficients, only products between a crude oil basis and a swing-cut basis are considered. Products between two crude oils or two swing-cuts basis have not been considered, although they may prove beneficial when regressing with other data sets. We believe these three types of coefficient-basis (intercept, linear and bilinear) are enough to represent accurately the interactions for the crude oil distillation process addressed in this paper.

Moreover, an intelligent pre-choice or pre-elimination of coefficients is performed so as to be representative of real operations. For example, for naphtha components, the linear/bilinear coefficients used are related to all the four crude oils, but only to the first swing-cut, as there is no relation between naphtha components and swing-cuts 2 and 3 (as shown in Figure 1b). Similarly, kerosene relates to swing-cuts 1 and 2, light diesel relates to swing-cuts 2 and 3, and heavy diesel relates to swing-cut 3. Fuel gas, liquefied petroleum gas and atmospheric residue are not related to any swing-cut. The six distinct models proposed are:

- Model 1: intercept coefficient + linear coefficients for the crude oils;
- Model 2: intercept coefficient + linear coefficients for the crude oils + linear coefficients for the swing-cuts;

- Model 3: linear coefficients for the crude oils + bilinear coefficients for the products between one crude oil and one swing-cut;
- Model 4: linear coefficients for the crude oils + linear coefficients for the swing-cuts + bilinear coefficients for the products between one crude oil and one swing-cut;
- Model 5: intercept coefficient + linear coefficients for the crude oils + bilinear coefficients for the products between one crude oil and one swing-cut.
- Model 6: intercept coefficient + linear coefficients for the crude oils + linear coefficients for the swing-cuts + bilinear coefficients for the products between one crude oil and one swing-cut;

As an example, the surrogate Model 6 for a dependent variable (Y_{ip}) can be mathematically written as shown in Equation (8), in which CR and SW are the sets for crude oils and swing-cuts, respectively.

$$\begin{aligned} Y_{ip} = I_i + \sum_{j\,\in\,CR} l_{ji}\ X_{jp} + \sum_{j\,\in\,SW} l_{ki}\ X_{kp} & \qquad \forall\ i\ \in DC, \\ + \sum_{j\,\in\,CR}\sum_{k\,\in\,SW} b_{jki}\, X_{jp} X_{kp} & \qquad \forall\ p \in P \end{aligned} \tag{8}$$

For each point p in the data set P of independent variables (X), and for each dependent variable i (yields and properties of each output distillate from the distillation unit), Y_{ip} are the estimated outputs or dependent variable values, I_i are the intercept coefficients, l_{ji} and l_{ki} are the linear basis and b_{jki} are the bilinear basis with respect to the crude oils j and swing-cuts k, respectively, X_{jp} are the crude oil compositions, and X_{kp} are the swing-cut yields or split fractions.

The surrogate models are built through MIQP optimizations for the stream yields of each final cut (except fuel gas), for the specific gravity of each final cut (except fuel gas), and for the sulfur content of each final cut (except fuel gas and LPG). Therefore, for each scenario there are 17 optimization problems to be formulated and optimized, each one with 20 coefficients (one intercept, seven linear and twelve bilinear basis). Each optimal solution contains the active basis (binaries equal to one) and the respective coefficients. Each optimization leads to a surrogate model related to a specific process variable (i.e., yields or properties of distillation cuts), which are the dependent variables in our model. As an example of process variables of interest in our problem, we have: yield of naphtha (YLD_N), specific gravity of light diesel (SG_{LD}), sulfur content of atmospheric residue (S_{ATR}), etc. The optimization problems are formulated to minimize the least square error in Equation (9) subject to Equation (8) and Equations (10) to (14) that limit or bound the values of the coefficients and impose a maximum specified number of basis:

$$\text{Minimize } E_i = \frac{1}{n}\sum_{p=1}^{n}(y_{ip} - Y_{ip})^2 \tag{9}$$

$$-Mz_0 \leq I_i \leq Mz_0 \qquad \forall\ i\ \in DC \tag{10}$$

$$-Mz_j \leq l_{ji} \leq Mz_j \qquad \forall\ j \in CR, i\ \in DC \tag{11}$$

$$-Mz_k \leq l_{ki} \leq Mz_k \qquad \forall\ k \in SW, i\ \in DC \tag{12}$$

$$-Mz_{jk} \leq b_{jki} \leq Mz_{jk} \qquad \forall\ j\ \in CR, k\ \in SW, i\ \in DC \tag{13}$$

$$\sum_{j\,\in\,CR} z_j + \sum_{k\,\in\,SW} z_k + \sum_{j\,\in\,CR, k\,\in\,SW} z_{jk} + z_0 \leq B \qquad z_j, z_{jk}, z_{jk}, z_0 \in \{0,1\} \tag{14}$$

In Equations (9) to (14), the number of points in the data set is $n = 100$, M is a large enough number ($M = 1000$), z_j and z_k are binary variables that correspond to the linear basis, z_{jk} are binary variables for the bilinear basis, z_0 is the binary variable for the intercept, and B is the maximum number of basis. The real, physical or actual values for the dependent variables (y_{ip}) are calculated using either the conventional or the improved swing-cut method in order to provide acurate approximations. In order to build a surrogate model we need to identify which basis l_{ji}, l_{ji}, l_{ki}, and b_{jki} should be used and which are their respective coefficients. For that, the parameter B must be given in the optimization problem. In this work, the value of B is not limited as the problem to be optimized is small in size.

Check Performance: The surrogate models built for each dependent variable are compared to the original data set to calculate the error in the predictions (difference between the real values and the estimated values, i.e., $y_{ip} - Y_{ip}$). In order to allow an easier comparison between distinct models, Equation (9) takes the average least square error among all 100 points from the data set, and

Equation (15) takes the average error among each surrogate model, in which dv is the number of dependent variables within the set DV.

$$Error = \frac{1}{dv} \sum_{i \in DV} E_i \tag{15}$$

5. Results and Discussion

In this section, the results are presented for each case by using both the CSW and the ISW methods. Table 1 presents the least square errors for Yields, Specific Gravity and Sulfur Content, as well as the total error (average among yield, specific gravity, and sulfur content errors), from each surrogate model for each proposed case. Each of them is the average least square error among all data points for all final distillates, obtained from Equation (15). Detailed results for each distillate can be found in the Supporting Information (Tables B1 to B4). Also, we should note that the solution of the MIQPs require small computational times (around 3 seconds per MIQP).

Table 1: Yield, Specific Gravity and Sulfur Content least square errors for each model and case

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Data Set 1 CSW	Yield Error	8.10E-01	2.35E-02	5.40E-01	1.01E-01	7.13E-09	7.13E-09
	Specific Gravity Error	2.29E-03	3.22E-04	1.26E-03	8.21E-04	2.83E-04	2.83E-04
	Sulfur Content Error	7.76E-03	1.13E-03	4.02E-03	5.48E-03	9.48E-04	9.48E-04
	Total Error	2.89E-01	8.73E-03	1.92E-01	3.75E-02	3.79E-04	3.79E-04
Data Set 1 ISW	Yield Error	8.10E-01	2.35E-02	5.40E-01	1.24E-08	1.24E-08	1.24E-08
	Specific Gravity Error	2.29E-03	1.49E-04	1.29E-03	1.02E-04	1.02E-04	1.02E-04
	Sulfur Content Error	7.77E-03	5.83E-04	4.50E-03	3.82E-04	3.82E-04	3.82E-04
	Total Error	2.89E-01	8.51E-03	1.92E-01	1.49E-04	1.49E-04	1.49E-04
Data Set 2 CSW	Yield Error	8.77E-01	2.19E-02	5.92E-01	9.90E-02	8.59E-09	8.59E-09
	Specific Gravity Error	2.48E-03	3.46E-04	1.34E-03	7.95E-04	3.44E-04	3.44E-04
	Sulfur Content Error	7.99E-03	1.19E-03	4.34E-03	5.87E-03	1.17E-03	1.17E-03

	Total Error	3.13E-01	8.19E-03	2.11E-01	3.69E-02	4.65E-04	4.65E-04
Data Set 2 ISW	Yield Error	8.44E-01	2.09E-02	5.92E-01	1.25E-08	1.25E-08	1.25E-08
	Specific Gravity Error	2.38E-03	1.40E-04	1.24E-03	1.07E-04	1.07E-04	1.07E-04
	Sulfur Content Error	7.84E-03	6.07E-04	4.03E-03	3.23E-04	3.23E-04	3.23E-04
	Total Error	3.01E-01	7.60E-03	2.11E-01	1.33E-04	1.33E-04	1.33E-04

The lower the average error between the swing-cut approach and the surrogate model, the better is the performance of the model. The poor performance achieved by Model 1 (intercept + linear crude oil compositions) and Model 3 (linear crude oils and bilinear crude oil compositions multiplied by swing-cut yields), with errors in the order of 10^{-1} or 10^{-2}, indicates the need of swing-cut information in Model 1 to account for the fractionation and over the intercept in Model 3 for any other effects. Including the linear basis of the swing-cuts improves the performance of the predictions, especially for the yields, as shown by Models 2 and 4. However, the role of bilinear terms in Model 4 is shown to be much more significant than the intercept terms in Model 2. Although the intercept coefficient is not used in Model 4, it provides accurate results for the ISW method (errors under 10^{-4}) but not as good for the CSW method. Finally, from the results in Table 1 it can be noticed that both the intercept and the bilinear terms, in addition to linear terms of crude oils, are required to provide good accuracy as seen with Models 5 and 6, which have the best (and similar) performance. All basis (intercept, linear for crude oils, and bilinear) improve the accuracy of the model, although in that specific case in which all three types of basis are employed, the linear terms for the swing-cuts in Model 6 did not provide further improvements. As Models 5 and 6 present the same results, to simplify the surrogate model eliminating unnecessary terms, Model 5 has been chosen as the best fit, and therefore, the following results are presented based on this model.

For the case study considered, the yields from the proposed model are awfully close to the yields from both the conventional and improved swing-cut model, so that our model is as representative as the swing-cut model. As there is only one equation needed to calculate each yield/property of each final cut, our data-driven machine learning coefficient-setup model requires fewer equations and fewer degrees of freedom than other methods. It can be easily integrated into any planning/scheduling environment without significantly increasing the simulation/optimization effort. As example, we take the results from Model 5 using Data Set 1 and the ISW model to show the active basis and their respective coefficients being used to establish correlations to calculate the yields and properties of the final cuts:

$$YLD_{LPG} = 1.1823 - 0.2795x_{CO1} - 0.2061x_{CO2} + 0.0543x_{CO3} - 0.1143x_{CO4}$$

$$\begin{aligned} YLD_{N} = {} & -6.8787 + 14.0722x_{CO1} + 13.3600x_{CO2} + 13.8974x_{CO3} + 11.6549x_{CO4} \\ & + 3.4999x_{CO1}x_{SW1} + 3.4137x_{CO2}x_{SW1} + 2.7201x_{CO3}x_{SW1} + 2.1239x_{CO4}x_{SW1} \end{aligned}$$

$$\begin{aligned} YLD_{K} = {} & -11.3973 + 21.1936x_{CO1} + 21.2691x_{CO2} + 19.5103x_{CO3} + 18.1561x_{CO4} \\ & + 3.4999x_{CO1}x_{SW1} - 3.4137x_{CO2}x_{SW1} - 2.7201x_{CO3}x_{SW1} - 2.1239x_{CO4}x_{SW1} \\ & + 5.5787x_{CO1}x_{SW2} + 6.3726x_{CO2}x_{SW2} + 5.1732x_{CO3}x_{SW2} + 4.7251x_{CO4}x_{SW2} \end{aligned}$$

$$\begin{aligned} YLD_{LD} = {} & -11.6501 + 24.5691x_{CO1} + 26.2438x_{CO2} + 24.1220x_{CO3} + 23.5514x_{CO4} \\ & - 5.5787x_{CO1}x_{SW2} - 6.3725x_{CO2}x_{SW2} - 5.1732x_{CO3}x_{SW2} - 4.7251x_{CO4}x_{SW2} \\ & + 4.7096x_{CO1}x_{SW3} + 4.7427x_{CO2}x_{SW3} + 4.5682x_{CO3}x_{SW3} + 4.8972x_{CO4}x_{SW3} \end{aligned}$$

$$\begin{aligned} YLD_{HD} = {} & -3.7153 + 13.2021x_{CO1} + 13.1504x_{CO2} + 13.0235x_{CO3} + 13.7990x_{CO4} \\ & - 4.7096x_{CO1}x_{SW3} - 4.7427x_{CO2}x_{SW3} - 4.5683x_{CO3}x_{SW3} - 4.8972x_{CO4}x_{SW3} \end{aligned}$$

$$YLD_{ATR} = 63.2307 - 3.6515x_{CO1} - 4.7216x_{CO2} - 1.4779x_{CO3} - 2.0949x_{CO4}$$

$$SG_{LPG} = -0.1911 + 0.73400x_{CO1} + 0.7404x_{CO2} + 0.7379x_{CO3} + 0.7416x_{CO4}$$

$$\begin{aligned} SG_{N} = {} & 0.7453 - 0.0497x_{CO1} - 0.0297x_{CO2} - 0.0059x_{CO3} - 0.0089x_{CO4} + 0.213x_{CO1}x_{SW1} \\ & + 0.0151x_{CO2}x_{SW1} + 0.0151x_{CO3}x_{SW1} + 0.0143x_{CO4}x_{SW1} \end{aligned}$$

$$\begin{aligned} SG_{K} = {} & -0.2939 + 1.0778x_{CO1} + 1.0763x_{CO2} + 1.1121x_{CO3} + 1.1052x_{CO4} \\ & + 0.0123x_{CO1}x_{SW1} + 0.0137x_{CO2}x_{SW1} + 0.0096x_{CO3}x_{SW1} + 0.0092x_{CO4}x_{SW1} \\ & + 0.0142x_{CO1}x_{SW2} + 0.0222x_{CO2}x_{SW2} + 0.0126x_{CO3}x_{SW2} + 0.0142x_{CO4}x_{SW2} \end{aligned}$$

$$\begin{aligned} SG_{LD} = {} & -0.3113 + 1.1512x_{CO1} + 1.1624x_{CO2} + 1.1759x_{CO3} + 1.1705x_{CO4} \\ & + 0.0087x_{CO1}x_{SW2} + 0.0113x_{CO2}x_{SW3} + 0.0091x_{CO3}x_{SW3} + 0.0072x_{CO4}x_{SW3} \\ & + 0.0061x_{CO1}x_{SW3} + 0.0077x_{CO2}x_{SW3} + 0.0093x_{CO3}x_{SW3} + 0.0079x_{CO4}x_{SW3} \end{aligned}$$

$$SG_{HD} = -0.3195 + 1.1921x_{CO1} + 1.2118x_{CO2} + 1.2300x_{CO3} + 1.2168x_{CO4} + 0.0091x_{CO1}x_{SW3} + 0.0093x_{CO2}x_{SW3} + 0.0113x_{CO3}x_{SW3} + 0.0098x_{CO4}x_{SW3}$$

$$SG_{ATR} = -0.3377 + 1.3051x_{CO1} + 1.3143x_{CO2} + 1.3297x_{CO3} + 1.3255x_{CO4}$$

$$S_N = 0.0107 - 0.0101x_{CO1} - 0.0087x_{CO2} - 0.0062x_{CO3} - 0.0080x_{CO4} + 0.0025x_{CO1}x_{SW1} + 0.0046x_{CO2}x_{SW1} + 0.0112x_{CO3}x_{SW1} + 0.0068x_{CO4}x_{SW1}$$

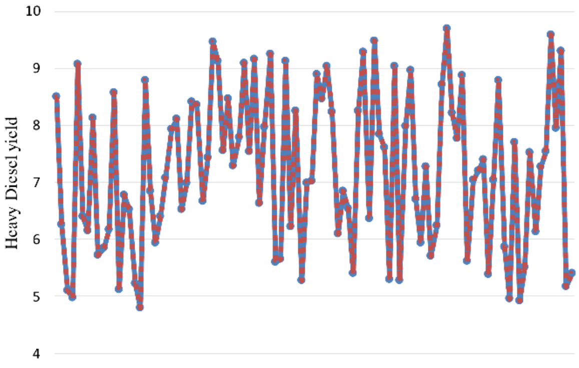
$$S_K = -0.0578 + 0.0879x_{CO1} + 0.0990x_{CO2} + 0.1370x_{CO3} + 0.1267x_{CO4} + 0.0139x_{CO1}x_{SW1} + 0.0141x_{CO2}x_{SW1} + 0.0160x_{CO3}x_{SW1} + 0.0182x_{CO4}x_{SW1} + 0.0242x_{CO1}x_{SW2} + 0.0311x_{CO2}x_{SW2} + 0.0320x_{CO3}x_{SW2} + 0.0352x_{CO4}x_{SW2}$$

$$S_{LD} = -0.1797 + 0.3235x_{CO1} + 0.3491x_{CO2} + 0.4161x_{CO3} + 0.4103x_{CO4} + 0.0444x_{CO1}x_{SW2} + 0.0454x_{CO2}x_{SW3} + 0.0549x_{CO3}x_{SW3} + 0.0529x_{CO4}x_{SW3} + 0.0537x_{CO1}x_{SW3} + 0.0448x_{CO2}x_{SW3} + 0.0062x_{CO3}x_{SW3} + 0.0657x_{CO4}x_{SW3}$$

$$S_{HD} = -0.2037 + 0.5566x_{CO1} + 0.6083x_{CO2} + 0.6949x_{CO3} + 0.7142x_{CO4} + 0.0228x_{CO1}x_{SW3} + 0.0575x_{CO2}x_{SW3} + 0.0349x_{CO3}x_{SW3} + 0.0589x_{CO4}x_{SW3}$$

$$S_{ATR} = -0.2800 + 1.0080x_{CO1} + 0.9456x_{CO2} + 1.0519x_{CO3} + 1.0944x_{CO4}$$

Figure 3 presents some plots to illustrate the accuracy of the Surrogate Model 5 for yields, specific gravity, and sulfur content of heavy diesel. The blue dots are the data samples calculated through the ISW model (y_{ip}), and the red dots are the surrogate model calculations (Y_{ip}) for the $n = 100$ data points. In the Supporting Information, all plots for the Surrogate Model 5 versus the ISW method using the data set 1 are presented (Figures D1 to D3).



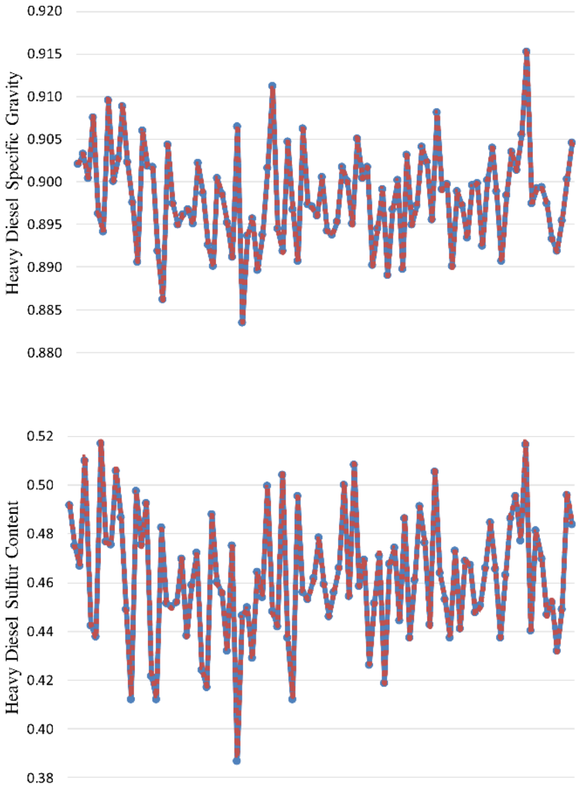


Figure 3: Plots for Surrogate Model 5 vs. ISW method of heavy diesel

Using all three types of coefficients (intercept, linear and bilinear) improves the accuracy and performance of the proposed methodology. The surrogate Model 5 has a nearly perfect fit to the data, as shown in the plots from Figure 3. However, if the bilinear coefficients are not used, there are significant mismatches between the model and the original data. As example, Figure 4 shows the difference of the 100 samples between the improved swing-cut approach and the surrogate Model 1 (which considers only the intercept and the linear coefficients of crude oils) for the sulfur content of heavy diesel.

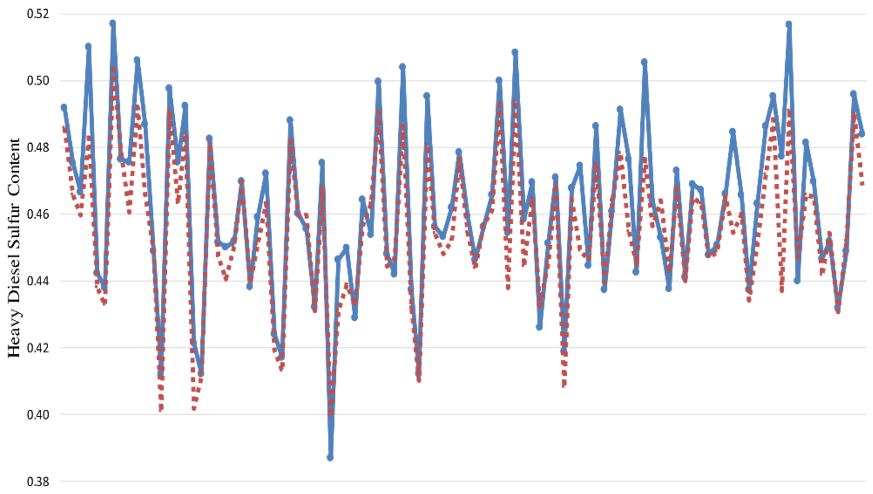


Figure 4: Plot for Surrogate Model 1 vs. ISW method for sulfur content of heavy diesel

Discrepancies can be observed between the ISW method and the correlation using only intercept and crude oil linear coefficients (Figure 4), which might have a significant impact depending on the application of this method. Poor predictions in the sulfur content of distillates may underspecify the product, what can potentially imply in costs for the refinery, to reprocess or blend the product with overspecified streams. As an alternative to mitigate off-spec products, the final fuels can be sold at reduced prices and waivers can be requested from regulatory agencies, which is not a good practice or sustainable procedure. On the other hand, overspecifications by bad predictions can produce distillates with product giveaways, which typically imply in losses at selling higher quality products at regular prices. To reduce both under and overspecification in the production, a better prediction of the inputs and outputs of the distillation process is required, whereby the surrogate model methodology proposed in this work is an alternative to improve the accuracy of predictions under simulation/optimization environments. For the predictions on yields and properties of distillation processes using the addressed coefficient setup MIQP technique, the use of bilinear terms, in addition to the intercept and linear terms (without the need of the linear term for the swing-cuts), is an efficient fashion to mathematically represent this type of process.

6. Conclusions

We are moving towards a more complex and detailed process optimization age, mainly because the advancements in decision-making modeling, computer-aided resources, and solution algorithms. Big data is becoming a reality and leads to opportunities of cost reduction in industrial processes. The surrogate modeling proposed in this work to estimate compositions and properties of distillates may be considered as machine learning and predictive analytics techniques, and can use either real (and uncertain) data from the plant or rigorous simulated data to improve the predictions of distillation units by using measurement feedback. In other words, the proposed surrogate model can be self-adjustable and self-improvable, although requiring engineering supervision. The results show that our model provides accurate predictions when compared to both conventional and improved swing-cut methods. Due to the small number of equations required, our shortcut sub-models can be easily integrated into any planning, scheduling, and coordinating environment with minimal increase in the simulation and optimization effort and data requirement.

ACKNOWLEDGMENTS

This work was supported by São Paulo Research Foundation – FAPESP [grants 2017/03310-1 and 2018/04942-4], Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq [grant 311868/2018-5], and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES [finance code 001].

References

(1) Riazi, M. R. Characterization and properties of petroleum fractions. *ASTM international*. **2005**.

(2) Fu, G.; Mahalec, V. Comparison of Methods for Computing Crude Distillation Product Properties in Production Planning and Scheduling. *Industrial & Engineering Chemistry Research*. **2015**, 54(45), 11371-11382.

(3) Li, W.; Hui, C.; Li, A. Integrating CDU, FCC and product blending models into refinery planning. *Computers & chemical engineering*. **2005**, 29 (9), 2010-2028.

(4) Kelly, J. D.; Menezes, B. C.; Grossmann, I. E. Distillation blending and cutpoint temperature optimization using monotonic interpolation. *Industrial & Engineering Chemistry Research*. **2014**, 53 (39), 15146-15156.

(5) Menezes, B. C.; Kelly, J. D.; Grossmann, I. E. Improved swing-cut modeling for planning and scheduling of oil-refinery distillation units. *Industrial & Engineering Chemistry Research*. **2013**, 52 (51), 18324-18333.

(6) Joly, M.; Odloak, D.; Miyake, M.; Menezes, B. C.; Kelly, J. D. "Refinery production scheduling toward Industry 4.0." *Frontiers of Engineering Management* 5. **2018**, 2, 202-213.

(7) Chen, M.; Mao, S.; Liu, Y. "Big data: A survey." *Mobile networks and applications* 19. **2014**, 2, 171-209.

(8) Chiang, L.; Lu, B.; Castillo, I. "Big Data Analytics in Chemical Engineering." *Annual Review of chemical and biomolecular engineering* 8. **2017**, 63-85.

(9) Maktoubian, J.; Ghasempour-Mouziraji, M.; Noori, M. Oil and Gas supply chain optimization using Agent-based modelling (ABM) integration with Big Data technology. *EAI Endorsed Transactions on Smart Cities*. **2020**, 4 (9).

(10) Wu, X.; Zhu, X.; Wu, G.; Ding, W. "Data mining with big data." *IEEE transactions on knowledge and data engineering* 26. **2013**, 1, 97-107.

(11) Boukouvala, F.; Li, J.; Xiao, X.; Floudas, C. A. "Data-Driven Modeling and Global optimization of industrial-scale petrochemical planning operations." *In 2016 American Control Conference (ACC)*. **2016**, 3340-3345.

(12) Li, J.; Xiao, X.; Boukouvala, F.; Floudas, C. A.; Zhao, B.; Du, G.; Su, X.; Liu, H. Data-driven mathematical modeling and global optimization framework for entire petrochemical planning operations. *AIChE Journal* 62. **2016**, 9, 3020-3040.

(13) Yu, W.; Morales, A. Data driven fast real-time optimization with application to crude oil blending. *In 2019 1st International Conference on Industrial Artificial Intelligence (IAI)*. **2019**, 1-6.

(14) Ahmad, I.; Ahsan A.; Manabu K.; Izzat I.C. Gray-box Soft Sensors in Process Industry: Current Practice, and Future Prospects in Era of Big Data. *Processes.* **2020**, 8(2), 243.

(15) McBride, K., Sanchez Medina, E.I. and Sundmacher, K. Hybrid Semi-parametric Modeling in Separation Processes: A Review. Chemie Ingenieur Technik. **2020**.

(16) Anderson, R. N. "'Petroleum Analytics Learning Machine'for optimizing the Internet of Things of today's digital oil field-to-refinery petroleum system." In 2017 IEEE International Conference on Big Data (Big Data). **2017**, 4542-4545.

(17) Beck, D. A. C.; Carothers, J. M.; Subramanian, V. R.; Pfaendtner, J. "Data science: Accelerating innovation and discovery in chemical engineering." AIChE Journal 62. **2016**, 5, 1402-1416.

(18) Wilson, Z. T.; Sahinidis, N. V. The ALAMO approach to machine learning. Computers & Chemical Engineering. **2017**, 106, 785-795.

(19) Brooks, R. W.; Van Walsem, F. D.; Drury, J. Choosing cutpoints to optimize product yields. Hydrocarbon Processing. **1999**, 78(11), 53-60.

(20) Zhang, J.; Zhu, X. X.; Towler, G. P. A level-by-level debottlenecking approach in refinery operation. Industrial & engineering chemistry research. **2001**, 40 (6), 1528-1540.

(21) Guerra, O. J.; Le Roux, G. A. C. Improvements in petroleum refinery planning: 1. Formulation of process models. Industrial & Engineering Chemistry Research. **2011**, 50 (23), 13403-13418.

(22) Alattas, A. M.; Grossmann, I. E.; Palou-Rivera, I. Integration of nonlinear crude distillation unit models in refinery planning optimization. Industrial & Engineering Chemistry Research. **2011**, 50 (11), 6860-6870.

(23) Alattas, A. M.; Grossmann, I. E.; Palou-Rivera, I. Refinery production planning: multiperiod MINLP with nonlinear CDU model. Industrial & engineering chemistry research. **2012**, 51(39), 12852-12861.

(24) Mahalec, V.; Sanchez, Y. Inferential monitoring and optimization of crude separation units via hybrid models. Computers & Chemical Engineering. **2012**, 45, 15-26.

(25) Fu, G.; Sanchez, Y.; Mahalec, V. Hybrid model for optimization of crude oil distillation units. AIChE Journal. **2015**, 62 (4), 1065-1078.

(26) Franzoi R. E.; Kelly J. D.; Menezes B. C.; Gut J. W. Advanced Data Analytics for Process-Shop Base+Delta Sub-Model Estimation in Planning and Scheduling Decision-Making. In: AICHE Annual Meeting, Pittsburgh, PA, United States. **2018**.

(27) Kelly J. D; Zyngier D. Continuously improve the performance of planning and scheduling models with parameter feedback. In: Proceedings of the foundations of computer-aided process operations. **2008**.

(28) Durrani, M.A.; Ahmad, I.; Kano, M.; Hasebe, S. An Artificial Intelligence Method for Energy Efficient Operation of Crude Distillation Units under Uncertain Feed Composition. *Energies*. **2018**, 11(11), 2993.

(29) Kelly, J. D. The Unit-Operation-Stock Superstructure (UOSS) and the Quantity-Logic-Quality Paradigm (QLQP) for Production Scheduling in The Process Industries. In Multidisciplinary International Scheduling Conference Proceedings: New York, United States. **2005**, 327-333.

Nomenclature

Continuous Variables

b_{jki}: bilinear basis of crude oil j and swing-cut k for dependent variable i

dv: number of dependent variables

E: objective function to be minimized

I_i: intercept basis for dependent variable i

B: maximum number allowed of basis

l_{ji}: linear basis of crude oil j for dependent variable i

l_{ki}: linear basis of swing-cut k for dependent variable i

n: total number of points in the data set

Q_K: final kerosene flow

Q_N: final naphtha flow

$T_{N/K}$: naphtha/kerosene cutpoint temperature

T_K: end cutpoint temperature of kerosene

T_N: initial cutpoint temperature of naphtha

x_j: generic independent variable for crude oil j

x_1 to x_4: independent variables for the four crude oils

x_5 to x_7: independent variables for the three swing-cuts

X_{jp}: value of independent decision variable of crude oil j at point p

X_{kp}: value of independent decision variable of swing-cut k at point p

Y_{ip}: value of dependent variable i at point p calculated by the surrogate model

y_{ip}: value of dependent variable i at point p calculated by the swing-cut methods

Y_{mc}: yields of micro cut mc

Binary Variables

z_0: binary variable for intercept coefficient

z_j: binary variable for linear coefficient of crude oil j

z_{jk}: binary variable for bilinear coefficient of crude oil j and swing-cut k

z_k: binary variable for linear coefficient of swing-cut k

Parameters

M: big M parameter

Sets

CR: crude oils

DC: distillation component

P: points in the data set

SW: swing-cuts

Subscripts

ATR: atmospheric residue

c: crude oil

HD: heavy diesel

i: dependent variable in the identification model

j: crude oil in the identification model

k: swing-cut in the identification model

K: kerosene

LD: light diesel

LPG: liquefied petroleum gas

N: naphtha

N/K: naphtha/kerosene

p: point of data set

sw: swing-cut

$sw1$: swing-cut 1, between naphtha and kerosene

$sw2$: swing-cut 2, between kerosene and light diesel

$sw3$: swing-cut 3, between light diesel and heavy diesel

Superscripts

sw: swing-cut

$final$: final distillate

cdu: crude distillation unit

Abbreviations

ATR: atmospheric residue

CDU: crude distillation unit

CO1 to CO4: crude oils

CSW: conventional swing-cut

FG: fuel gas

FI: fractionation indices

FY: fixed yields

HD: heavy diesel

IMPL: Industrial modeling and programming language

ISW: improved swing-cut

K: kerosene

LD: light diesel

LHS: Latin hypercube sampling

LPG: liquefied petroleum gas

MINLP: mixed integer nonlinear programming

MIQP: mixed-integer quadratic programming

N: naphtha

RTO: real time optimization

S: sulfur content

SG: specific gravity

SW1 to SW3: swing-cuts

TBP: true boiling point

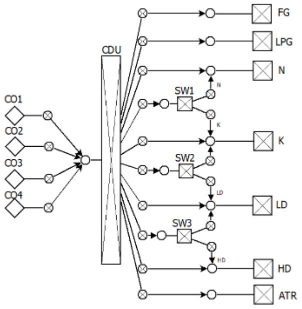
UOPSS: unit-operation-port-state superstructure

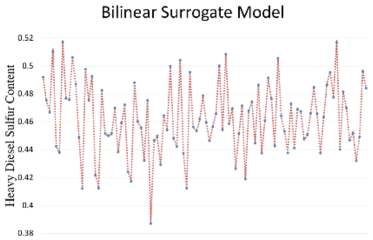
WTR: weight transfer ratio

YLD: yield

Contents

Graphical Abstract


CO1
CO2
CO3
CO4
CDU
SW1
SW2
SW3
N
K
LD
HD
FG
LPG
N
K
LD
HD
ATR


Bilinear Surrogate Model
Heavy Diesel Sulfur Content