Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes

Luca Mencarelli, Alexandre Pagot, Pascal Duchêne

Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes

Luca Mencarelli*, Alexandre Pagot, Pascal Duchêne

IFP Energies nouvelles, Solaize, France

Q1

A B S T R A C T

In this paper, we first briefly survey the main surrogate model building approaches discussed in the literature considering also design of experiments strategies and dimensionality reduction procedures: we mainly focus on sub-set approaches and sampling strategies for constrained regression problems. We delineate a systematic methodology for surrogate modelling in presence of model constraints, such as non-negativity of the model responses. The main contribution of this paper is twofold: from one side we extend the principal component analysis framework to the case of constrained regression problem, from the other we propose a novel methodology which integrates the subset selection and the previous principal component regression procedure. Finally, we apply the two novel algorithms to two fundamental chemical processes in petroleum refinery, namely catalytic reforming and light naphtha isomerization. The numerical results show the comparisons between the two algorithms in terms of computational and accuracy trade-offs.

Keywords:
Surrogate modeling
Principal component analysis
Subset selection
Catalytic reforming
Light naphtha isomerization

1. Introduction

The aim of chemical process synthesis engineering consists in modeling, designing and optimizing complex chemical processes. The possible high computational cost of the process estimation and optimization can be circumvented by means of surrogate models (or meta-models) that represent a systematic approximation of the mathematical relationships between the degrees of freedom (input data) and the variables of interest (output data). A systematic methodology to identify dependent and independent variables of a given chemical process unit is given by Henao and Maravelias (2010, 2011). Instead of obtaining the output data via experimental measurements, numerical simulators are often available, but in several cases obtaining output values from a given input configuration is rather time consuming.

Surrogate models can be useful either if the first-principle model is too complex or time consuming to optimize or if the first-principle model does not exist at all. In the first case it is possible to collect initial simulated data by choosing a sampling strategy and sampling more points later; in the latter case instead the data are obtained by physical experiments. Moreover, in both cases the data could be affected by noise or characterized by incomplete information.

The objective of the present paper consists in defining a methodology to derive a surrogate model starting from noiseless simulated data in presence of model constraints: in our case, in fact, a numerical simulator is available for the chemical processes we consider, and data can be obtained easily and rather quickly from the simulations.

From one side, the surrogate model should be sufficiently complex to catch the relationships of the given process, hence it should be characterized by high accuracy, from the other one, instead, it should be sufficiently simple to speed up the computational times, so that its complexity is low. High accuracy and low complexity are obviously conflicting targets and determine a trade-off between the quality of the approximation and the quantity of computational effort. Generally, in fact, surrogate models are preferred to other approaches, such as rigorous models or simplified physical approximation models, when the computational time is expected to be a crucial aspect Psaltis et al. (2016): rigorous or physical approximation models could be rather time consuming since they have been usually obtained by discretizing complex dynamic equations, such as systems of partial differential equations.

In our study we introduce a novel methodology to find sufficiently accurate surrogate models and to simultaneously perform dimensionality reduction with regards to the number of model parameters. Hence, procedures for reducing the total number of experiments are out of scope of the present paper.

Many different approaches have been discussed in the literature in order to build an effective surrogate model from a set of

* Corresponding author.
E-mail addresses: luca.mencarelli@ifpen.fr (L. Mencarelli), alexandre.pagot@ifpen.fr (A. Pagot), pascal.duchene@ifpen.fr (P. Duchêne).

https://doi.org/10.1016/j.compchemeng.2020.106772
0098-1354/

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

input and output couples of experimental or simulated data (see for instance the surveys Queipo et al., 2005; Forrester and Keane, 2009; Vu et al., 2017; Bhosekar and Ierapetritou, 2018; McBride and Sundmacher, 2019), but all the methodologies usually structure the surrogate modeling into roughly four main steps: (i) design of experiments (DOE), (ii) numerical simulations or experimental measurements, (iii) surrogate model selection and identification, and (iv) model testing. At step (i) the design space is conveniently sampled in order to define a set of input data configurations. At step (ii) several experiments are performed or a simulator is used to obtain the output data corresponding to the input configurations. Therefore, at step (iii) a specific surrogate model is selected and trained with respect to the so-called training set, which is composed of input and output data couples, i.e., the parameters of the surrogate model are estimated. Finally, at step (iv) the performances of the model are analyzed with respect to the so-called test set. If the performances of the surrogate model in terms of complexity and accuracy are not satisfactory, then the procedure restarts from step (i).

In our applications, we adopt a one-shot approach, i.e., we consider all the sampling points at once, so that our approach is composed of three main steps: (i) we sample the design space in order to obtain a training set of input/output values which opportunely cover the entire design space, (ii) we build the surrogate model by using all the sampling data, and (iii) we test the performance of our model on the test set.

Moreover, surrogate models are also used in a posterior optimization step to retrieve the optimal operating conditions for the chemical process for instance in a superstructure framework (we refer the interested reader to the survey Mencarelli et al., 2019a). In this approach a superstructure is defined by the set of all the possible alternative structures of a given chemical process: in surrogate driven superstructure approach the model units (reactors, distillation columns, or even entire sub-processes) are replaced by their surrogate models (Henao and Maravelias, 2010; 2011) and the superstructure is described by a general disjunctive problem (GDP) or a mixed integer (non)linear problem (MI(N)LP), which is then solved to determine the optimal alternative structure.

As aforementioned, we assume in our applications that a process simulator is available, but too time consuming to be used directly (Mencarelli et al., 2019b). Moreover, we would like to develop an appropriate surrogate model in order to exploit its analytic expression and derivatives during the posterior optimization phase, instead of directly applying a derivative-free approach. We have already shown that mesh-adaptive algorithms, such as NOMAD Audet et al. and Le Digabel (2011), or EGO-based methods (Jones et al., 1998) perform quite poorly in our applications (Mencarelli et al., 2019b).

Several recent papers deal with superstructure optimization by considering artificial neural networks (ANNs) as surrogate models (see, e.g., Altissimi et al., 1998; Nascimiento et al., 2000; Fahmi and Cremaschi, 2012): in particular, Fahmi and Cremaschi (2012) proposed a superstructure optimization methodology, by combining GDP and ANN, in which each process unit is replaced by an ANN which is trained by simulated data and embedded in a GDP formulation.

Generally, several outputs are considered at the same time so that we aim to build a surrogate model for each output. However, possible constraints should be enforced: these constraints can regard a single surrogate model independently from the other (in this case we have intra-model constraints) or contemporaneously a set of models (in this latter case we have inter-model constraints). Typical examples of intra-model constraints consist in non-negativity of the model responses (if the output represents physical measurements, or molar or mass fractions of several compounds): in this case the non-negativity constraint is referred to only one single model and each non-negativity constraint for a model is independent from the others. If the outputs represent compounds fractions, we should impose an equality inter-model constraints, enforcing the sum of the fractions is equal to 1: in this latter case the inter-model constraints depend contemporaneously from a set of models responses (for instance, in the previous example, the sum of the responses of all models should be 1, so that all the models responses appear in the corresponding inter-model constraint).

The presence of output constraints in derivative-free context has been investigated in recent papers. For instance, Conn and Le Digabel (2013) illustrate a hybrid methodology which combines quadratic models as surrogate models and mesh-adaptive direct search for constrained black-box problems, i.e., optimization problems in which the analytic expression of both the objective and the constraints is not available or it is too complex to evaluate. Boukoulava et al. (2017) and Boukoulava and Floudas (2017) introduce a data-driven methodology that combines surrogate modeling approach and deterministic global optimization algorithm, by extending the parallel AlgoRithms for Global Optimization of coNstrAined grey-box compUTational problems (p-ARGONAUT) algorithm (Beykal et al., 2018b). Moreover, p-ARGONAUT has been extended to multi-objective optimization problems in Beykal et al. (2018a). For a review in constrained derivative-free optimization we refer the interested reader to the survey (Boukoulava et al., 2016).

Therefore, the contribution of this paper is threefold: (i) we briefly summarize the main (recent) approaches to surrogate modeling together with sampling strategies and dimensionality reduction techniques focusing on constrained regression problems, (ii) we extend a well-known dimensionality reduction technique such as principal component analysis (PCA) to the case of regression problem with constraints for the response, and finally (iii) we propose a novel methodology by combining the previous version of PCA with subset selection (SS) in order to further reduce the dimensionality of the surrogate model, i.e., the number of parameters of the model. In fact, since the modelling step is usually followed by an optimization phase where the first-principle models are substituted by the surrogate model and the resulting optimization problem is solved, considering a model with a low number of parameters is crucial to solve the final optimization problem in a reasonable amount of time. Obviously, the dimensionality reduction procedure should maintain an acceptable quality of the surrogate model in terms of accuracy in data representation.

The rest of the paper is organized as follows. In Section 2 we briefly overview the sampling methods proposed in the literature. Section 3 is devoted to discuss the surrogate building procedures. In Sections 4.1 and 4.2 we deal with two kinds of possible additional constraints for the surrogate model, namely inter-model constraints and intra-model equality constraints, respectively. Then, we propose a two-phase PCA method (Section 5) and hybrid algorithm obtained by combining the two-phase PCA with SS (Section 6). The proposed methodologies are then applied to two relevant chemical processes in petroleum refinery framework, namely catalytic reforming in Section 7.1 and light naphtha isomerization in Section 7.2, respectively, which this report extensively discusses the numerical results of the computational experiments. Finally, conclusions and future work perspectives follow in Section 8.

In the rest of the paper we adopt the following notation. Given a positive integer scalar $N \in \mathbb{N}$, we indicate $[N] := \{1, \dots, N\}$. Moreover, given a set of N vectors $a_n \in \mathbb{R}^m$ ($n \in [N]$), a_n^m represents the m-th entry of the n-th vector in the set. In particular, we define a set of N data $x_n \in \mathbb{R}^K$ ($n \in [N]$) per the k-th input ($k \in [K]$) and a set $z_n \in \mathbb{R}^M$ ($n \in [N]$) of N data per the m-th output ($m \in [M]$).

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

2. Sampling step

In this section, we review the main techniques adopted for the sampling step. In particular, we report the papers which deal with sampling strategies for constrained problems: in our applications, in fact, the design space is described by a set (linear) constraints (see Section 4).

Two main sampling approaches have been exploited in the literature: (i) geometrical designs, and (ii) statistical designs. To the best of our knowledge, this distinction is introduced for the first time in Vu et al. (2017). In geometrical designs the DOE is defined by taking into account the geometrical shape of the design space; while in statistical designs the response of the surrogate model is assimilated to a realization of a random process.

Among geometrical designs the most adopted ones are: (i) full factorial design (FFD) (Box et al., 2005; Forrester et al., 2008), and (ii) latin hypercube design (LHD) (McKay et al., 1979). In both cases the design space is uniformly divided into regular cells with same dimensionality: in FFD the centre and the extreme points of each cell are selected, while in LHD we keep only a proper subset of the centres of the cells such that there is no couple of points sharing the same coordinate. It is worth to notice, in fact, that the FFD guarantees the design space is sampled in a uniform way; on the contrary, LHD does not guarantee the design space is sampled uniformly (see Fig. 2 in Viana (2016)). In order to avoid such a situation, several methods have been proposed in literature by choosing LHD according to space-filling criteria (Johnson et al., 1990; Pronzato, 2017), obtaining the so-called minimax LHD (van Dam, 2008), which minimize the covering radius, and maximin LHD (Morris and Mitchell, 1995; van Dam et al., 2007; Joseph and Hung, 2008; Husslage et al., 2011), which maximize the minimal pairwise distance between sampled points. Petelet et al. Petelet et al. (2010) introduce a sampling approach to deal with constrained LHD, i.e., LHD with inequality constraints, based on permutation technique applied to (unconstrained) LHD.

Among statistical designs the most common ones are (i) D-optimal, which aim to find the design which maximizes the determinant of the correlation matrix of the data, and (ii) I-optimal, whose designs minimize the average prediction variance (Goos et al., 2016) (for an insight about statistical designs see, e.g., Cornell (2002) and Smith (2005)). D-optimal and I-optimal designs have been extended to the case of linear constrained regression, in which (linear) additional constraints are presented, by Coetzer and Haines (2017).

Recently, an adaptive method for the sampling phase is introduced in Garud et al. (2017b): the idea consists in initially sampling the whole region according to a given sampling strategy described above and then iteratively placing the new sampling points in order to sample the original function as far as possible from the already placed points and in the region where the quality of the approximation is poor. Mixed adaptive sampling strategy for surrogate models represented by ANNs has been proposed by Eason and Cremaschi (2014). Straus and Skogestad (2019) have proposed another sampling algorithm relying on a termination criterion based on partial least squares regression (PLSR).

Sampling techniques and size choices are compared in Davis et al. (2018) with respect to different surrogate model building approaches. For an exhaustive discussion about the different strategies for sampling phase we refer the interested reader to survey (Garud et al., 2017a).

3. Surrogate model building

In this section, we survey the main techniques employed to develop surrogate models from input and output data. In statistical regression approaches, we consider a set $\mathcal{B}$ of basis functions $f_j(x) : \mathbb{R}^K \to \mathbb{R}$ ($j \in \mathcal{B}$) over the input variables, whose (linear) combinations give the responses of the surrogate model. The basis functions could be polynomials, transcendental and trigonometric functions, or even radial basis functions (RBFs).

RBFs arise from the seminal paper by Broomhead and Lowe (1988) and have been originally use to smoothly interpolate multivariable functions by Hardy (1971): they are universal approximators for functions over a finite number of real variables (Park and Sandberg, 1993). For a complete insight into RBF topics see the excellent survey by Buhmann (2000).

Other approaches include, for instance, Kriging methodology and support vector regression (SVR). Kriging dates back to the papers (Krige, 1952; Matheron, 1963) and have been applied to the design and analysis of computational experiments by Sacks et al. (1989). For a detailed analysis of the Kriging technique we refer the interested reader to the survey by Kleijnen (2009). Caballero and Grossmann (2008) propose a Kriging approach for the flowsheet optimization problem. Recently, Bouhlel et al. (2016) and Gaspar et al. (2017) have proposed two hybrid approaches by combining Kriging techniques with PLSR and with trust region method, respectively. On the contrary, SVR has been introduced by Vapnik (1995) (see also Vapnik et al., 1997) which extends the support vector machine technique to approximate nonlinear functions (for an introduction to SVR see the tutorial (Smola and Schölkopf, 2004)). Papers (Li et al., 2009; Ivanciuc, 2007) present several chemical applications for SVR.

Moreover, several papers are then devoted to compare the different approaches (see, e.g., Clarke et al., 2004; Amouzgar and Strömberg, 2017; Jin et al., 2001; Bhosekar and Ierapetritou, 2018 and references therein): however, there is no definitive understanding about the dominance relationships of one type of surrogate model with respect to the others in terms of accuracy in data representation. Müller and Shoemaker (2015) systematically address the influence of the surrogate model choice and the sampling method selection on the accuracy of the resulting model. In our study we restrict ourselves to polynomial surrogate models since we have already shown in Mencarelli et al. (2019b) that quadratic polynomial models perform sufficiently well with respect to the application we consider in the present paper; however the approaches we will describe can be extended to other types of basis functions.

Let $\mathcal{A} \subseteq \mathbb{R}^{|\mathcal{B}|}$ be the set of *a priori* constraints on the regression coefficients $\beta \in \mathbb{R}^{|\mathcal{B}|}$, such as, e.g., non-negativity constraints. In regression we aim to minimize an objective function $g(\beta) : \mathbb{R}^{|\mathcal{B}|} \to \mathbb{R}$ representing the distance between the simulated observations or measurements (output data) and the model responses. If the coefficients appear linearly (resp. nonlinearly) in the corresponding model, then we define the problem as linear regression (LR) (resp. nonlinear regression (NLR)). In our applications, we consider LR in which $f(x;\beta) := \Sigma_j \beta_j f_j(x)$, where $f_j(x) : \mathbb{R}^K \to \mathbb{R}$ may be general functions of x.

Typical objective functions for regression problems are the sum of the absolute distances:

$$g(\beta) := \sum_{n\in[N]} \left| z_n - \sum_{j\in\mathcal{B}} \beta_j f_j(x_n) \right|, \qquad \text{(LAD)}$$

which gives rise to the so-called least absolute deviation (LAD) criterion; or the sum of the squares of the residuals:

$$g(\beta) := \sum_{n\in[N]} \left(z_n - \sum_{j\in\mathcal{B}} \beta_j f_j(x_n) \right)^2, \qquad \text{(ORL)}$$

which defines the so-called ordinary least square regression (OLR) problem. For a detailed discussion about advantages and disadvantages of the two previous choices for the objective function, see

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

Chapter 1 in Huber (1981). Regression problems with (LAD) as objective function can be equivalently reformulated as constrained linear problems (LPs): it is sufficient to replace the absolute values with new variables t_n ($n \in |N|$) and consider the following inequalities for all $n \in [N]$:

$$t_n \geq z_n - \sum_{j \in \mathcal{B}} \beta_j f_j(x_n)$$
$$t_n \geq \sum_{j \in \mathcal{B}} \beta_j f_j(x_n) - z_n. \tag{1}$$

In most of the cases, however, the previous approach results into a highly dense surrogate model, i.e., with a large number of non-zero coefficients, so that the resulting surrogate model could be difficult to analyze and to optimize. In the surrogate-based superstructure approach, the curse of dimensionality is particularly critical, since the surrogate models are used as blocks in a more complex optimization framework.

Hence, in surrogate building process dimensionality reduction is therefore a key step to induce sparsity, i.e. to reduce the total number of non-zero coefficients: classical approaches include PCA (Cunningham, 2007; Fodor, 2002), random projections (RP) (Krahmer and Ward, 2016), and subset selection (SS) (Miller, 2002). Sparsity in the surrogate model can be induced also by adding a regularization parameter to the objective function such as in LASSO approach (Hastie et al., 2015): in this paper we consider only SS approaches since, in this latter case, we have a direct control on the number of non-zero coefficients in the surrogate models. More recently, Straus and Skogestad introduce two novel dimensionality reduction methods based on PLSR (Straus and Skogestad, 2017a; 2017b) and self-optimizing control (Straus and Skogestad, 2018), respectively.

Dimensionality reduction with application to water-flooding production optimization has been recently addressed by Sorek et al. (2017) by introducing the functional control method (FCM) and the interpolation control method (ICM) relying on polynomial approximation and piecewise polynomial interpolation controls, respectively (see also Beykal et al., 2018b).

In particular, in SS regression only a given subset of dimension $T \leq |\mathcal{B}|$ of basis functions is considered. Hence the following problem is defined:

$$\begin{aligned} \min_{y \in \mathbb{R}^{|\mathcal{B}|}, \beta \in \mathcal{A}} \quad & g(\beta) \\ \text{s.t.} \quad & \sum_{j \in \mathcal{B}} y_j = T \\ & \underline{\beta} y_j \leq \beta_j \leq \overline{\beta} y_j \quad \forall j \in \mathcal{B} \\ & y_j \in \{0,1\} \quad \forall j \in \mathcal{B}. \end{aligned} \tag{SS}$$

In (SS) $|\mathcal{B}|$ binary variables y_j are introduced to switch on/off the corresponding regression parameters β_j ($j \in \mathcal{B}$). Depending on the linearity (resp. nonlinearity) of the objective function, SS is a MILP (resp. MINLP). Coefficients $\underline{\beta}$ and $\overline{\beta}$ can be estimated in the preprocessing phase using any feasible value for β computed, e.g., by means of the (N)LR approach. In practical implementation we followed the procedure which is suggested in Cozad et al. (2014), i.e., we sum up the absolute values of β found for unconstrained regression and we set the obtained numerical value as the upper bound $\overline{\beta}$: then for the lower bound we simply set $\underline{\beta} := -\overline{\beta}$.

There exists several heuristic procedures to argue an opportune numerical value for T: the most adopted ones consist in forward- and backward-stepwise regressions. Forward-stepwise regression incrementally builds surrogate models by increasing T starting form $\mathcal{B} = \emptyset$ until a given information criterion, which includes the complexity and the accuracy of the model, is worsen. A possible information criterion is the correct Akaike criterion (AIC_c):

$$AIC_c(T, \beta) := N \log\left(\frac{1}{N} \sum_{n \in [N]} \left(z_n - \sum_{j \in \mathcal{B}} \beta_j f_j(x_n)\right)^2\right) + 2T + \frac{2T(T+1)}{N - T - 1}, \tag{AIC_c}$$

which is constituted by a weighted sum of the accuracy of the model, represented by the squares of the model residuals given by the distance between the output data and the surrogate model responses, and the relative complexity of the model, which takes into account the number of basis functions and the total number of observations. The flowchart of the SS algorithm is shown in Fig. 1, where (β^*, y^*) is the optimal solution for problem (SS). For other information criteria we refer the interested reader to the paper (Wilson and Sahinidis, 2017). The backward-stepwise regression approach, on the contrary, initially considers all the basis functions and progressively removes the less significant ones.

A comparison between different SS regression strategies is performed in Kim and Boukoulava (2019). Cozad et al. (2014) introduce a procedure, the automated learning of algebraic models for optimization (ALAMO), to solve (SS) with a forward-stepwise philosophy. A comprehensive description of ALAMO with applications to chemical problems is given by Wilson and Sahinidis (2017). Other software packages for surrogate building process are described in Bhosekar and Ierapetritou (2018).

4. Additional constraints

In practical applications several additional constraints on the responses of the surrogate models might be present. We divide them into two classes: (i) intra-model and (ii) inter-model constraints. Intra-model constraints regard the response of a single surrogate model, while inter-model constraints concern the responses of a subset of the models. The presence of inter-model constraints forces the procedure to address the corresponding subset of surrogate models at the same time.

4.1. Intra-model constraints

We consider a set of M outputs so that we have one problem of (SS) -type per output. In the notation the variables and the parameters of each output model are identified by the superscript $m \in [M]$.

We consider non-negativity constraints and we treat them by means of the approach introduced by Cozad et al. (2015). They propose a two-phase procedure: in the first step they build the surrogate model with respect to a given finite set of observations, while in the second one the points corresponding to the maximum violation with respect to the constraints for the resulting surrogate configuration are found and added to the set of the first step. The algorithm stops when no violated point is found.

In particular, in the first phase we solve a master problem with positivity constraints restricted to a (finite) subset $\mathcal{X}$ of the closed set $\mathcal{D} \subset \mathbb{R}^K$ describing the design space. Since in our case studies variables x represent mass fractions (see Section 7), we consider design spaces such that $\mathcal{D} := \{x \in \mathbb{R}^K : x \in [\underline{x}, \overline{x}] \wedge \sum_{k \in [K]} x^k = 1\}$.

The restricted master problem is:

$$\begin{aligned} \min_{y^m \in \mathbb{R}^{|\mathcal{B}|}, \beta^m \in \mathcal{A}} \quad & g(\beta^m) \\ \text{s.t.} \quad & \sum_{j \in \mathcal{B}^m} y_j^m = T^m \\ & \underline{\beta}^m y_j^m \leq \beta_j^m \leq \overline{\beta}^m y_j^m \forall j \in \mathcal{B}^m \\ & y_j^m \in \{0,1\} \forall j \in \mathcal{B}^m \end{aligned}$$

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

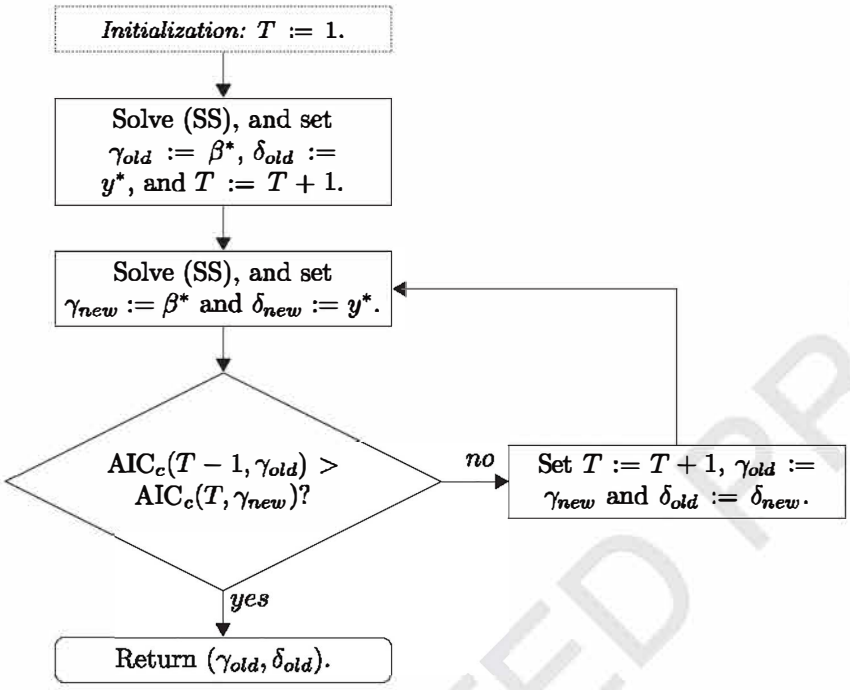


Fig. 1. Flowchart of the SS algorithm

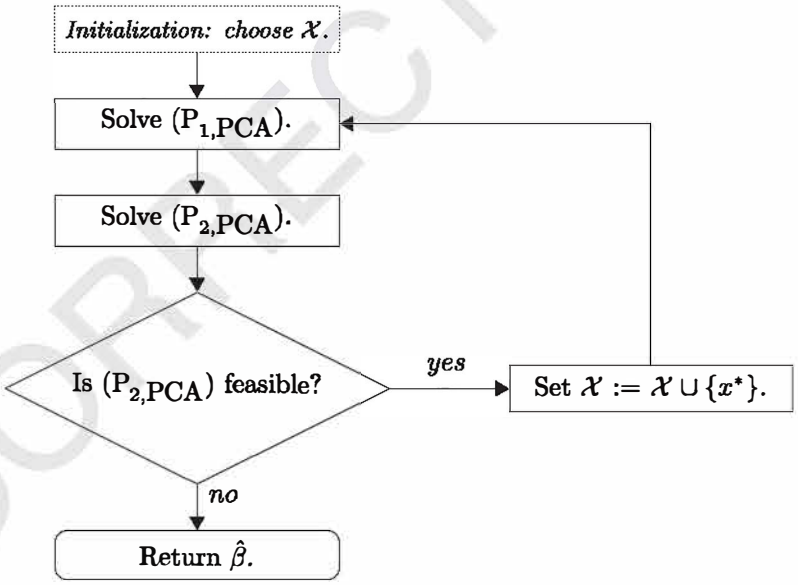


Fig. 2. Flowchart of the two-phase algorithm for intra-model constraints

$$\sum_{j \in \mathcal{B}^m} \beta_j^m f_j(x) \geq 0 \forall x \in \mathcal{X}. \qquad (P_1^m)$$

In the second phase, given is a feasible solution $(\mathcal{Y}^m, \hat{\beta}^m)$ of the problem (P_1^m), a non-negative scalar $\varepsilon_f \in \mathbb{R}_+$, representing the feasibility tolerance, and we solve the following optimization problem identifying the maximum violation:

$$\begin{aligned} \min_{x \in \mathcal{D}} \quad & \sum_{j \in \mathcal{B}^m} \hat{\beta}_j^m f_j(x) \\ \text{s.t.} \quad & \sum_{j \in \mathcal{B}^m} \hat{\beta}_j^m f_j(x) \leq -\varepsilon_f. \end{aligned} \qquad (P_2)$$

A positive small value for ε_f enforces strictly positivity for the violation (Cozad et al., 2015), or, in other words, we consider as feasible points the ones for which the corresponding violation is strictly less than ε_f.

The optimal solution x^* of problem (P_2) is then added to the set $\mathcal{X}$ and the first phase is performed again. As suggested in Cozad et al. (2015), in order to speed up the algorithm, instead of the optimal solution x^*, every set of (isolated) feasible solutions found for problem (P_2) by a state-of-the-art optimization solver, such as BARON (Tawarmalani and Sahinidis, 2005), can be added to $\mathcal{X}$. The procedure alternates the two phases until the problem (P_2) becomes infeasible. We note the previous procedure converges since at each iteration the amount of the violation of the current solution decreases because an increasing number of feasible points is taken into account in the first phase. Problems (P_1^m) can be solved separately for each output.

We initially set $\mathcal{X}$ equal to the set of all sampled points. As we said before, we use all the sampling points in one shot to build the surrogate model. Moreover, we observe that the objective function is always evaluated over the same set of initial points for which we know also the real outputs.

4.2. Equality inter-model constraints

Let $c \in \mathbb{R}^K$ and $d \in \mathbb{R}^M$ be L given vectors of opportune dimensions. We consider additional equality constraints linking the responses of several surrogate models, as follows:

$$\sum_{k\in[K]} c_\ell^k x^k = \sum_{m\in[M]} d_\ell^m \sum_{j\in\mathcal{B}^m} \beta_j^m f_j(x) \quad \forall x \in \mathcal{D} \wedge \forall \ell \in [L]. \tag{2}$$

In particular the previous relationship must hold for x_n ($n \in [N]$). The resulting problem is semi-infinite since it has an infinite number of constraints: we have one constraint for each design configuration $x \in \mathcal{D}$. In our computational experiments we practically consider problems with hundreds of constraints (see Section 7). In order to solve the resulting semi-infinite problem, in this case, we should consider all the M surrogate models at once by choosing the objective function $\Sigma_{m\in[M]}\omega^m g(\beta^m)$, which is the weighted sum of the objective functions of the models. In the implementation, we simply set $\omega^m := 1$ for all $m \in [M]$; however, the choice of the model weights ω^m constitutes a degree of freedom which can be further explored. In our approach we use the constraints (2) to express L variables as functions of the other ones for all the observations. In this way the equality constraint is automatically satisfied by definition.

To be more precise, we model only $K - L$ outputs and we derive the others from the equality constraints. We note that possible intra-model constraints, such as, e.g., non-negativity constraints, should be enforced for all the outputs in order to guarantee possible infeasible solutions are not generated. To the best of our knowledge, the previous technique for equality inter-model constraints is novel and can be applied to all the chemical balance constraints regarding, for instance, mass or energy balance.

5. Two-phase PCA

In order to reduce the number of basis functions in the previous procedure we propose an integration between the two-phase approach with PCA regression technique. In PCA regression approach a PCA step is performed before the regression procedure. We implement a PCA step over the basis functions, by decomposing in principal components the (Pearson) correlation matrix $C \in \mathbb{R}^{|\mathcal{B}|\times|\mathcal{B}|}$ of the basis functions evaluated over the initial input data. Then we consider only the eigenvectors corresponding to the first largest eigenvalues: hence, we derive new basis functions by projecting the original functions onto the subspace generated by the eigenvectors corresponding to the selected eigenvalues.

We note that performing the PCA over the correlation matrix can be seen as a standardization of the data in order to have the same variation data scale, since in order to define the (Pearson) correlation coefficients we subtract the means and we divide per the standard deviations. The (Pearson) correlation coefficients are defined as

$$C_{j_1,j_2} = \frac{\sum_{n\in[N]} \left(f_{j_1}(x_n) - \bar{f}_{j_1}\right)\left(f_{j_2}(x_n) - \bar{f}_{j_2}\right)}{\sqrt{\sum_{n\in[N]} \left(f_{j_1}(x_n) - \bar{f}_{j_1}\right)^2}\sqrt{\sum_{n\in[N]} \left(f_{j_2}(x_n) - \bar{f}_{j_2}\right)^2}}, \quad \forall (j_1, j_2) \in \mathcal{B}\times\mathcal{B}, \tag{3}$$

where $\bar{f}_{j_1} = (\sum_{n\in[N]} f_{j_1}(x_n))/N$ and $\bar{f}_{j_2} = (\sum_{n\in[N]} f_{j_2}(x_n))/N$.

In the two-phase approach we solve the first phase considering the new basis functions (reducing the dimensionality of the corresponding surrogate building problem) obtained by projecting the original basis function onto the space defined by the principal components of the correlation matrix. In the second phase we consider the original design space: the value of the function in the new point is then obtained by projecting the result of the second phase onto the new space.

The correlation matrix is decomposed as $C = \Lambda\Sigma\Lambda^T$, where Λ is the matrix whose columns are the orthonormal eigenvectors of the correlation data matrix and Σ is the diagonal matrix whose diagonal entries are the eigenvalues of matrix C sorted in non-increasing order. Let $\Lambda_{j'}$ be the matrix whose columns are the first $|\mathcal{B}'|$ eigenvectors of the correlation data matrix and $F(x)$ be the matrix whose rows are the basis function $f_j(x)$ ($j \in \mathcal{B}$), we define the new projected matrix of the basis functions as $F'(x) := \Lambda_{j'}^T F(x)$. The rows of the matrix $F'(x)$ give the new projected basis functions $f'_{j'}(x)$ ($j' \in \mathcal{B}'$). The problems solved in the first phase read as follows:

$$\begin{aligned} \min_{\beta^m\in\mathcal{A}'} \quad & g'(\beta^m) \\ \text{s.t.} \quad & \sum_{j'\in\mathcal{B}'} \beta_{j'}^m f'_{j'}(x) \geq 0 \qquad \forall x \in \mathcal{X}, \end{aligned} \tag{$P_{1,PCA}$}$$

where for instance we set $g'(\beta^m) := \sum_{n\in[N]} |z_n - \sum_{j'\in\mathcal{B}'} \beta_{j'} f'_{j'}(x_n)|$. The set $\mathcal{A}$ of *a priori* constraints over the surrogate parameters is replaced by its projected version $\mathcal{A}'$. Problem ($P_{1,PCA}$) is a (N)LP and depends on the number of principal components selected in the PCA step. We note that each principal component corresponds to a basis function. Therefore, in the two-phase PCA approach the selection of the basis functions is driven by the value of the eigenvalues of the correlation data matrix of the basis functions calculated over the input data.

Moreover, we observe that the PCA step is independent from the output data and can be performed in a preprocessing phase if different outputs should be considered at the same time. This is the case for example when different process alternative configurations should be evaluated: for instance, if a given chemical process can be realized with a different number of reactors (for a case study see Section 7.1), different surrogate models can be calculated for each possible number of reactors starting from the same input data opportunely generated according to a DOE strategy (see Section 2).

The problem addressed in the second phase is instead

$$\begin{aligned} \min_{x\in\mathcal{D}} \quad & \sum_{j'\in\mathcal{B}'} \hat{\beta}_{j'}^m f'_{j'}(x) \\ \text{s.t.} \quad & \sum_{j'\in\mathcal{B}'} \hat{\beta}_{j'}^m f'_{j'}(x) \leq -\varepsilon_f. \end{aligned} \tag{$P_{2,PCA}$}$$

In the objective function and in the inter-model constraints we project the original basis function onto the new space defined by the selected eigenvectors. The typology of the two problems and the stopping criterion follow the same philosophy as the two-phase approach sketched in the Section 4.1. The flowchart of the two-phase PCA algorithm is given in Fig. 2.

6. Hybrid approach

In this section we describe a hybrid approach obtained by combining SS philosophy and PCA regression. In particular, in the hybrid algorithm a SS step is performed once at the beginning to determine a lower number of representative basis functions and then the two-phase PCA procedure described in the previous section is applied to further reduce the dimensionality of the problem. The complete flowchart of the hybrid algorithm is given in Fig. 3.

To be more precise, we first implement the SS algorithm taking into account only the intra-model constraints: in this way we can solve a SS-type problem separately for each modelled output

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

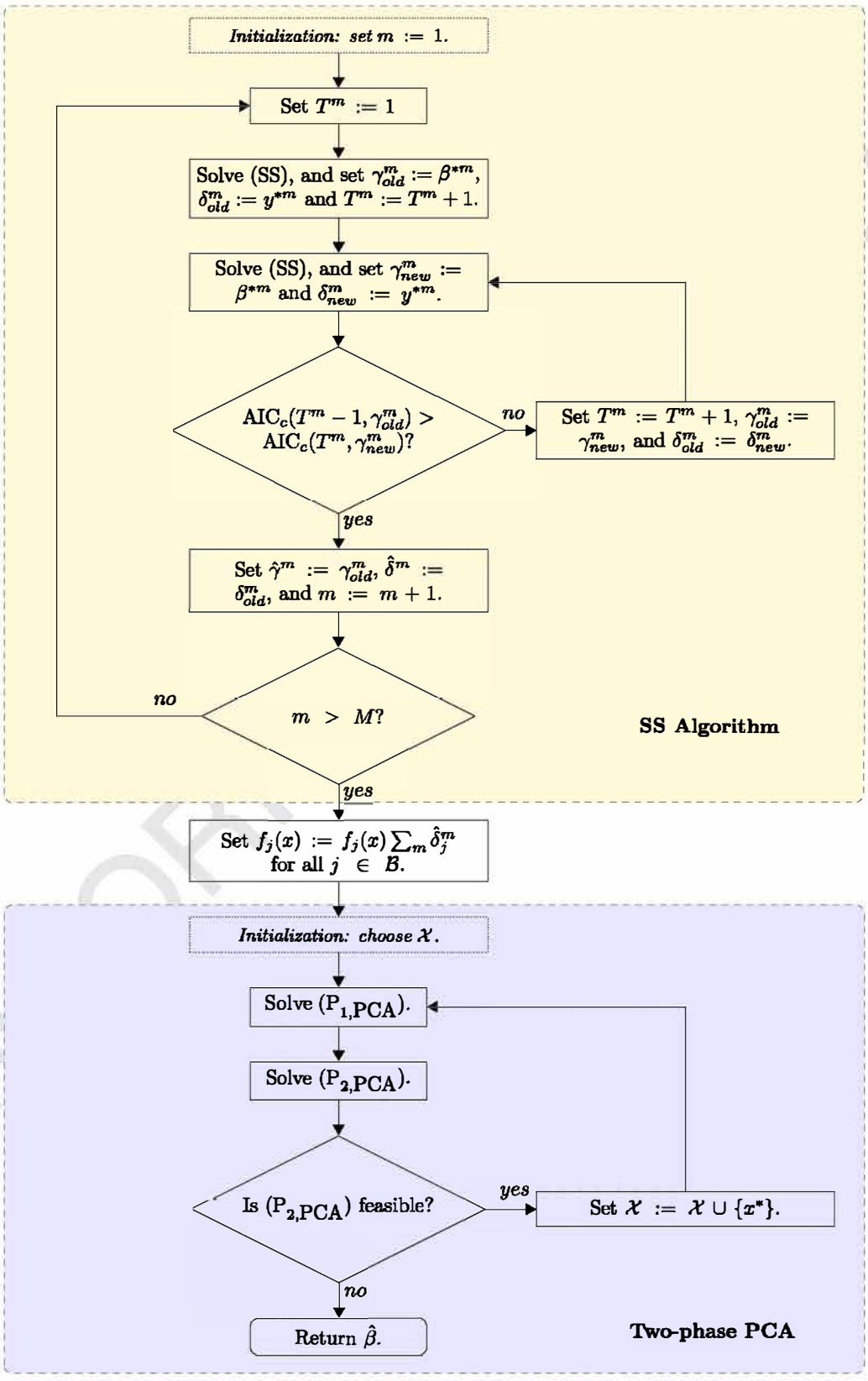


Fig. 3. Flowchart of the hybrid SS + two-phase PCA algorithm

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

by decomposing the original problem into M simpler independent subproblems with the same structure. We note a parallel implementation setting can be exploited in this context (in order to fairly compare the different approaches we consider only pure sequential implementations).

In particular, in presence of equality inter-model constraints, we can avoid to solve the SS-type problem for the L outputs obtained as functions of the other $K-L$ outputs through the equality constraints. Then, we sum up the values of the binary variables introducing to switch on/off the basis functions over the outputs and the values obtained per input are multiplied with the original functions in order to weight them. The functions selected for multiple outputs in the SS step will have a larger weight and will have, hence, a larger probability to be selected in the PCA step. We observe that the selected basis functions are employed to model only $K-L$ outputs, while L outputs are still obtained from the equality inter-model constraints.

This approach combines the efficacy of the SS strategy to find representative basis functions and the computational speed of the PCA regression. The numerical results show, in fact, that a lower number of principal components should be considered to obtain (in average) the same accuracy of the surrogate models in case a preliminary SS step is performed. Moreover, we note that in the case of pure SS approach a MI(N)LP should be solved at each iteration, while in the two-phase PCA only a (N)LP should be addressed: the (non)linearity of the problem depends on the (non)linearity of the objective function $g(\beta)$. Furthermore, performing a separate SS step per output allows to better capture the complexity of each surrogate model by dealing with different number T^m of basis function per the m-th output ($m \in [M]$).

7. Computational experiments

In our computational setting we divided the input variables into two classes, namely process variables p_j ($j \in [J]$), which represent the control variables in the posterior optimization phase, and composition variables c_i ($i \in [I]$), which coincide with the mass or molar composition of the compounds. We chose polynomial models because we are looking for simple surrogate models, since we would optimize them in a posterior phase in order to retrieve the best operational conditions for the analyzed process. We consider two types of surrogate models: (i) polynomial quadratic models and (ii) polynomial cubic models. In particular, in case (i) we consider models with the composition and the process variables occurring linearly and with bilinear interactions, i.e., bilinear mixed products, between process variables and composition variables: such models can be expressed in the form

$$\beta_0 + \sum_{i\in[I]} \beta_{1,i} c_i + \sum_{i\in[I]} \sum_{j\in[J]} \beta_{2,ij} c_i p_j + \sum_{i\in[I]} \beta_{3,i} c_i^2 \tag{4}$$

on the contrary in case (ii) we have models where the composition variables appear linearly and the process variables appear quadratically, i.e., such that can they be expressed in the form

$$\beta_0 + \sum_{i\in[I]} \beta_{1,i} c_i + \sum_{i\in[I]} \sum_{j\in[J]} \beta_{2,ij} c_i p_j + \sum_{i\in[I]} \beta_{3,i} c_i^2 + \sum_{i\in[I]} \sum_{j'\in[J]} \sum_{j''\in[J]} \beta_{4,ij'j''} c_i p_{j'} p_{j''}. \tag{5}$$

The number n_p of parameters in case (i) is given by $n_p = 1 + |I| + |I||J|$; in case (ii) we have instead $n_p = 1 + |I| + |I||J| + |I||J|^2$.

After preliminary computational tests with polynomial functions by considering the full cubic model with all the interactions, we decided to restrict ourselves to models resulting from the multiplication of the compositions variables appearing linearly and the process variables up to the quadratic terms, i.e. surrogate models (5) (considering other terms do not add much in terms of model accuracy for the case studies).

We choose function (LAD) as objective function for the regression step (first phase). Each MILP is solved by means of `IBM ILOG CPLEX 12.8` IBM ILOG CPLEX with option `numericalemphasis` and `scaind` activated and parallelization setting enable (up to 2 threads). The MINLPs are solved through `BARON 18.5.8` Tawarmalani and Sahinidis (2005). All the codes are implemented in `GAMS 25.1.2` McCarl et al. (2017) on a Dell machine with Intel(R) Xeon(R) CPU E5-1620 v3 at 3.50 GHz with 4 GB RAM.

We set a time limit of 100 CPU seconds for the first phase and 15 CPU seconds for the second phase. Preliminary computational experiments have shown that either the second phase finds a feasible solution relatively quickly or no solution is found within a larger CPU time, declaring hence the problem as infeasible. Moreover, in order to speed up the algorithm for the MILP solved in the SS step we stop at the feasible solution found at the root node in the Branch-and-Bound tree (we have observed that the feasible solution found at the root node is not so far from the optimum solution: however, a relatively large amount of time is needed to certify its optimality).

For the second phase, we keep the first feasible solution found by `BARON` and we add it to the restricted master problem. We set $\varepsilon_f := 0.2$. Generally, the choice of the numerical value for ε_f depends on the order of magnitude of the involved functions: in this case for the second phase we are considering as acceptable points the ones for which the corresponding response is greater than -0.2, which represents a reasonable value in our case studies. The computation of the eigenvalues and of the eigenvectors of the correlation data matrix is performed by means of the algebraic tools available in `GAMS` based on the `LAPACK DSYEV` routine Anderson et al. (1999).

7.1. Case study 1: Catalytic reforming

In this section we present a real-world application in petroleum refinery industry, namely the catalytic reforming process. Catalytic reforming (CR) is a chemical refinery process transforming raw naphtha into high octane gasoline called reformate containing aromatic hydrocarbons and iso-alcanes (Turaga and Ramanathan, 2003; Gjervan et al., 2004; Lapinski et al., 2014) (for an historical perspective on the studies about CR, see also Rahimpour et al. (2013)).

The catalytic reforming process was originally introduced in the 1940s by the Charles Stark Dreape laureate Vladimir Haensel[1], who proposed the so-called Platforming process, adopting a catalyst containing platinum. The CR unit is the most important process in refinery industry to produce lead-free automobile fuel and hydrogen. CR unit is composed of a sequence of reactors (usually from 3 to 5) characterized by operating conditions (temperature, pressure, molar hydrogen-to-hydrocarbon ratio, and feed composition) and equipped with catalyst (typically with platinum).

The main chemical reactions in CR are, in fact, dehydrogenation and hydroisomerization of naphtenes transforming them into aromatics, isomerization and dehydrocyclization of alcanes converting them into aromatics and iso-alcanes, hydrocraking of alcanes into smaller components, hydrogenolysis, and coke formation. Coke formation is a relatively slow process and it represents a damaging reaction since coke reduces the performances of the catalyst. Typical operating conditions are high temperature (450–500°C),

[1] see patents Alumina-platinum-halogen catalyst and preparation thereof. 1949, August 16. U.S. Patent No. 2,479,109 and Process of reforming a gasoline with an alumina-platinum-halogen catalyst. 1949, August 16. U.S. Patent No. 2,479,110.

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

Table 1
Acronyms.

Acronyms (alphabetical order)	
AIC_c	correct Akaike information criterion
ALAMO	Automated learning of algebraic models for optimization
ANN	Artificial neural network
ARGONAUT	Algorithms for global optimization of constrained grey-box computational problems
ATOUT	Advanced tools for optimization and uncertainty treatment
CR	Catalytic reforming
DOE	Design of experiments
FCM	Functional control method
FFD	Fully factorial design
GDP	General disjunctive problem
ICM	Interpolation control method
iP	iso-alcanes
IS	Isomerization
LAD	Least absolute deviation
LHD	Latin hypercube design
LP	Linear problem
LR	Linear regression
MILP	Mixed integer linear problem
MINLP	Mixed integer nonlinear problem
NLR	Nonlinear regression
nP	n-alcanes
OLR	Ordinary least square regression
PCA	Principal component analysis
PLSR	Partial least square regression
RBF	Radial basis function
RON	Research octane number
RP	Random projections
SIP	Semi-infinite problem
SS	Subset selection
SVR	Support vector regression
WHSV	Weight hourly space velocity

Table 2
Input compounds of the catalytic reforming process.

Input compounds	
nP_7	heptane
iP_7	isoheptane
N_7	naphthenes with 7 atoms of carbons
A_7	toluene
nP_8	octane
iP_8	isoctane
N_8	naphthenes with 8 atoms of carbons
A_8	ethyl-benzene + xylenes

Table 3
Output compounds of the catalytic reforming process.

Output compounds	
H_2	hydrogen
P_1	methane
P_2	ethane
P_3	propane
P_4	butane
P_5	pentane
$6P_6$	n-hexane
$5P_6$	simple branched alcanes with 6 atoms of carbons
$4P_6$	double branched alcanes with 6 atoms of carbons
N_6	naphthenes with 6 atoms of carbons
A_6	benzene
$7P_7$	n-heptane
$6P_7$	simple branched alcanes with 7 atoms of carbons
$5P_7$	double branched alcanes with 7 atoms of carbons
N_7	naphthenes with 7 atoms of carbons
A_7	toluene
$8P_8$	n-octane
$7P_8$	simple branched alcanes with 8 atoms of carbons
$6P_8$	double branched alcanes with 8 atoms of carbons
N_8	naphthenes with 8 atoms of carbons
A_8	ethyl-benzene + xylenes

medium level of pressure (3–35 atm) and molar hydrogen-to-hydrocarbon (H_2/HC) ratio between 3 and 8 Ancheyta-Juárez and Villafuerte-Macías (2001).

In particular, for illustration we consider the C7–C8 cut. The inputs of the model are the research octane number (RON) and the mass percentage of the hydrocarbon compounds occurring in the CR process (see Tables 2–3). The pressure and the temperature of the chemical reactions are treated as constants. Hence, we have one process variable (RON) and eight composition variables (mass percentages).

In this case, since the model outputs represent mass fractions, non-negativity and summing up to 1 constraint mush be enforced. Moreover, in CR we have one equality constraint (2) for each ℓ-th chemical element ($\ell \in [L]$) (hydrogen and carbon in hydrocarbon compounds) and n-th observation ($n \in [N]$), expressing the equivalence between the total number $\xi^{in}_{\ell,n}$ of moles of the ℓ-th element in the process inflow and the total number $\xi^{out}_{\ell,n}$ of the moles of the ℓ-th element in the process outflow in the n-th observation, i.e., $\xi^{in}_{\ell,n} = \xi^{out}_{\ell,n}$ for all $\ell \in [L]$ and $n \in [N]$. The number of moles for the ℓ-th element is given by the weighted sum of the molar percentage of all the chemical compounds in the stream, whose weights are the number of moles of hydrogen and carbon occurring in the corresponding compound, i.e.,

$$\xi^{in}_{\ell,n} = \sum_{k\in[K]} \frac{\mathrm{mol}^{in}_{\ell,k}}{\mathrm{mass}^{in}_{k}}\, c^{k}_{n}, \tag{6}$$

where $\mathrm{mol}^{in}_{\ell,k}$ and mass^{in}_{k} are the number of moles of the ℓ-th element in the k-th input compounds and the molar mass of the k-th input compounds, respectively, and

$$\xi^{out}_{\ell,n} = \sum_{m\in[M]} \frac{\mathrm{mol}^{out}_{\ell,m}}{\mathrm{mass}^{out}_{m}}\, z^{m}_{n}, \tag{7}$$

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

Table 4
R^2 and weighted R^2 (wR^2) for PCA and SS+PCA per number of principal components.

	PCA				SS+PCA			
	R^2		wR^2		R^2		wR^2	
No. comp.	Train	Test	Train	Test	Train	Test	Train	Test
30	0.85	0.78	0.94	0.90	0.87	0.81	0.95	0.92
31	0.85	0.78	0.94	0.90	0.88	0.87	0.96	0.95
32	0.88	0.87	0.96	0.95	0.88	0.87	0.96	0.95
33	0.88	0.87	0.96	0.95	0.88	0.87	0.96	0.95
34	0.88	0.87	0.96	0.96	0.88	0.87	0.96	0.95
35	0.88	0.87	0.96	0.95	0.89	0.87	0.96	0.96
36	0.89	0.88	0.96	0.96	0.89	0.86	0.96	0.96
37	0.89	0.88	0.96	0.96	0.88	0.88	0.96	0.96
38	0.89	0.89	0.96	0.96	0.89	0.88	0.96	0.96
39	0.89	0.89	0.96	0.96	0.90	0.89	0.97	0.95
40	0.90	0.91	0.97	0.97	–	–	–	–
avg	0.88	0.86	0.96	0.95	0.88	0.87	0.96	0.95

where $\text{mol}_{\ell,m}^{out}$ and mass_m^{out} are the number of moles of the ℓ-th element in the m-th output compounds and the molar mass of the m-th output compounds, respectively.

Note that satisfying the balance constraints on the number of moles implies all the mass fractions sum up to 1 and also the mass balance holds. Hence, these constraints have not been enforced in the first phase problem.

We have generated the plan of experiments by means of the software package `Design Expert 10` Design Expert. In particular, we adopt the D-optimal design for the training set and the I-optimal design for the test set for model (5). We chose statistical designs since for this kind of DOE Design Expert lets us to consider other additional constraints (such as that the sum of all the mass fractions for a given observation is equal to 1) when the plan of experiments is developed: D-optimal and I-optimal design are referred to the constraint case for the cubic model (5). In particular, we define a training set with 226 samples and a test set with 200 samples.

The software tool used to numerically simulate the catalytic reforming process is `OSCAR 1.1`, a numerical software developed at IFP Energies nouvelles.

We build individual surrogate models for each output component: as stated in the previous sections, we model only $m-2$ outputs since we have 2 equations for the molar balance of carbon and hydrogen, and we obtain the remaining 2 outputs from the equality inter-model constraints.

In order to evaluate the performance of a surrogate model, in addition to average coefficient of determination (R^2) over the outputs, we consider a weighted R^2 (wR^2) which consists in the average of the R^2 for the single models weighted with the average output mass fraction computed over the observations. In Table 4 we report the values of R^2 and wR^2 per number of principal components for the two-phase PCA approach and for the hybrid approach (SS+PCA).

The maximum number of principal components coincides with the number of considered basis functions: in the surrogate model for the catalytic reforming we are considering only the composition variables appearing linearly and the interaction between the process variable and the compositions variables up to the quadratic term, i.e., model (5) (see Section 7). Before dimensionality reduction the cubic model (5) has 40 parameters.

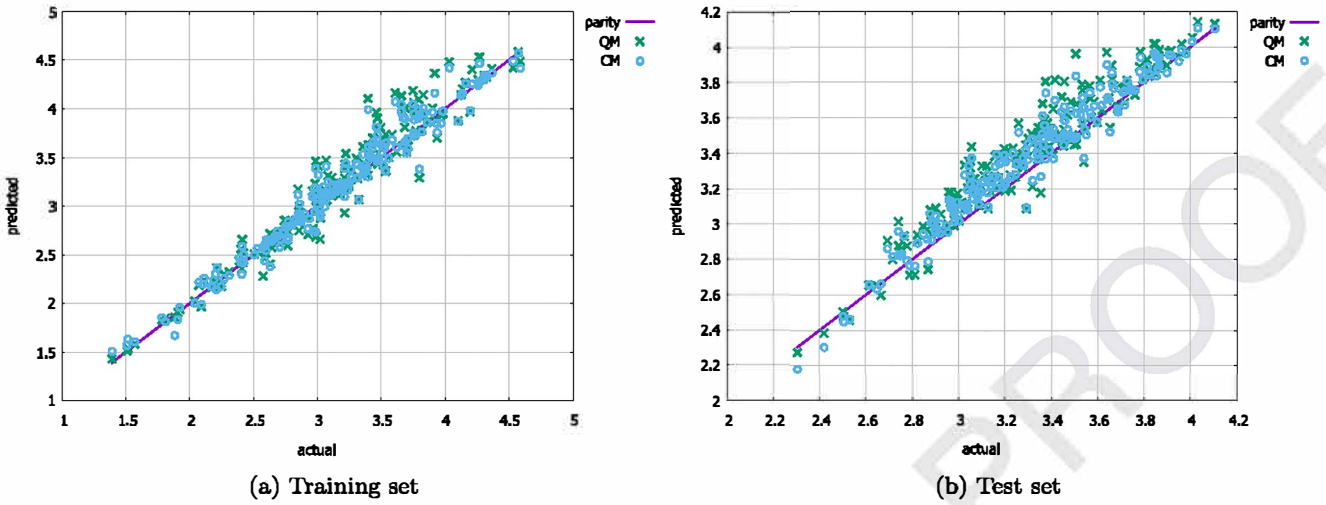
For the hybrid algorithm the maximum number of the principal components is given by the number of basis functions found in the SS step. It is worth observing that in the SS step the number of basis functions is already reduced from 40 to 39. The gain is relatively small: in this case we are not using the full cubic models with all the parameters, but we select *a priori* a subset of terms for the interactions between variables. Preliminary computational experiments with the full cubic model with all the possible mixed bilinear products have shown the SS step allows us to significantly reduce the number of parameters in the surrogate models. The total number of parameters for a polynomial model of degree d in n variables is $\binom{n+d}{d}$: a full cubic model has 220 parameters. In the case of full cubic model, the SS step let us to decrease the number of basis functions up to 18, obtaining a R^2 and a wR^2 indices for the training set equal to 0.87 and 0.96, respectively and a R^2 and a wR^2 indices for the test set equal to 0.89 and 0.96, respectively. In general, larger the number of basis functions the SS can chose among, better the performances of the SS step in terms of selection of the basis functions.

Table 5
CPU times (in seconds) for PCA and SS+PCA per number of principal components.

No. comp.	PCA	SS+PCA	Δ
30	457.89	543.34	15.73%
31	501.58	480.36	-4.42%
32	452.73	504.94	10.34%
33	476.43	504.58	5.58%
34	508.24	568.73	10.64%
35	497.81	566.78	12.17%
36	566.99	574.13	1.24%
37	587.06	591.95	0.83%
38	583.76	628.23	7.08%
39	603.74	625.85	3.53%
40	550.84	–	–
avg	526.10	558.89	6.27%

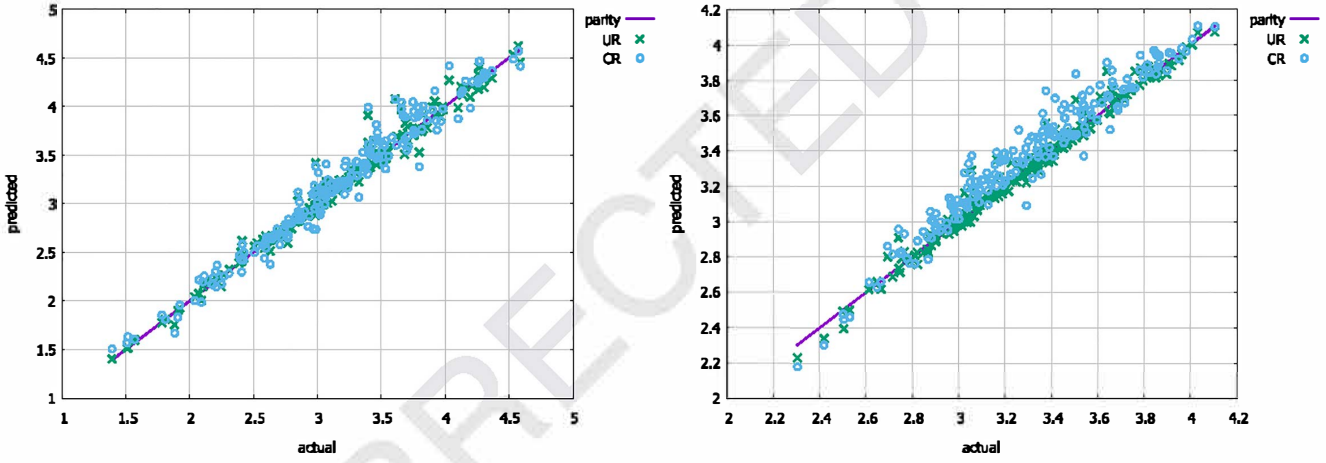
Table 4 shows the hybrid approach achieves the same performances as the two-phase PCA approach with a smaller number of principal components, both for the training and the test set. Moreover, we note the values of wR^2 is always larger than the ones of R^2, because the surrogate model in presence of equality constraints tends to better estimate the components characterized by a higher percentage concentration.

From Table 5, which reports the computational times for the two proposed algorithms and the percentage increase Δ between the computational time of SS+PCA and the computational time of PCA. It is clear that the SS step has an important impact on the CPU times: except for the cases of 31 principal components, in which the CPU time of the hybrid method is smaller than the one of the two-phase PCA, in all the other cases the CPU time of the hybrid method is the largest one. The average increase of the CPU time is 6.27% with a maximum of 15.73% for the case of 30 principal components. Higher the increase in the relative performance of the surrogate model, higher the increase of the CPU time. There-



(a) Training set (b) Test set

Fig. 4. H_2, predicted vs. actual plots. The values given by quadratic model (4) in green (QM); and the values given by the cubic model (5) in blue (CM).



(a) Training set (b) Test set

Fig. 5. H_2, predicted vs. actual plots, cubic model (5). The values given by the unconstrained regression in green (UR); and the values given by constrained regression in blue (CR).

fore, the results underline a trade-off between the relative reduction on the number of basis functions and the relative increase of computational effort.

Figs. 4a–7b show the scatter parity plots for the H_2 for the training set (figures with caption (a)) and the test set (figures with caption (b)). In the figures the x-axis reports the values of the simulated outputs, while the y-axis reports the values of the response of the surrogate models. In the limit case in which the response of the models coincides with the simulated output, the corresponding point lies on the bisector of the first quadrant. Smaller the distance between the value and the bisector of the first quadrant, better the performance of the surrogate model.

The quality of the surrogate models strongly depends on the grade of the mixed products between the process and the composition variables as shown in Fig. 4a and b which compare the quadratic model, i.e., Eq. (4), and the cubic model, i.e., Eq. (5), and on the fact that we are obliged to consider constrained regression (see Fig. 5a and b which report the parity plots obtained by constrained regression over all the basis functions and the unconstrained LAD for cubic model (5)).

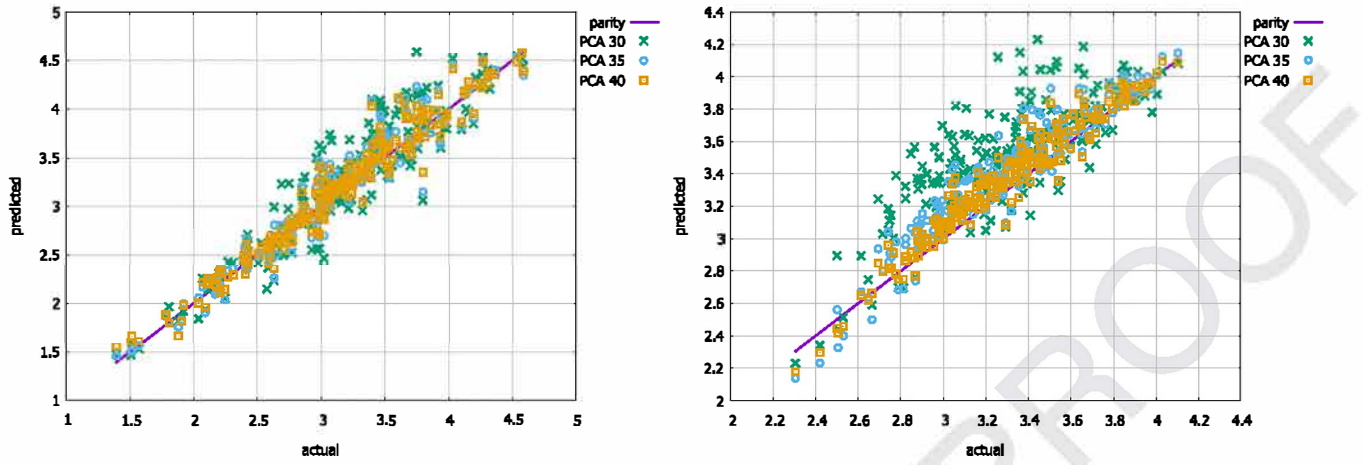
A key driver of the performance of the surrogate models is clearly the number of principal components and hence of basis functions considered in the estimation process: Fig. 6a and b show the predicted outputs for the two-phase PCA by varying the number of principal components. From the parity plot, it is possible to observe that the relative increase of estimation quality between 30 and 35 principal components is more accentuated than the one obtained by passing from 35 to 40 principal components (the relative R^2 and wR^2 increases for the 30/35 principal components and for the 35/40 principal components are reported in Table 6).

Fig. 7a and b report the comparison between the two-phase PCA and the hybrid approach for 35 principal components, showing the second method is slightly better than the first one in terms of distance between simulated and predicted outputs.

Table 6
Relative R^2 and wR^2 increases for the 30/35 and 35/40 principal components.

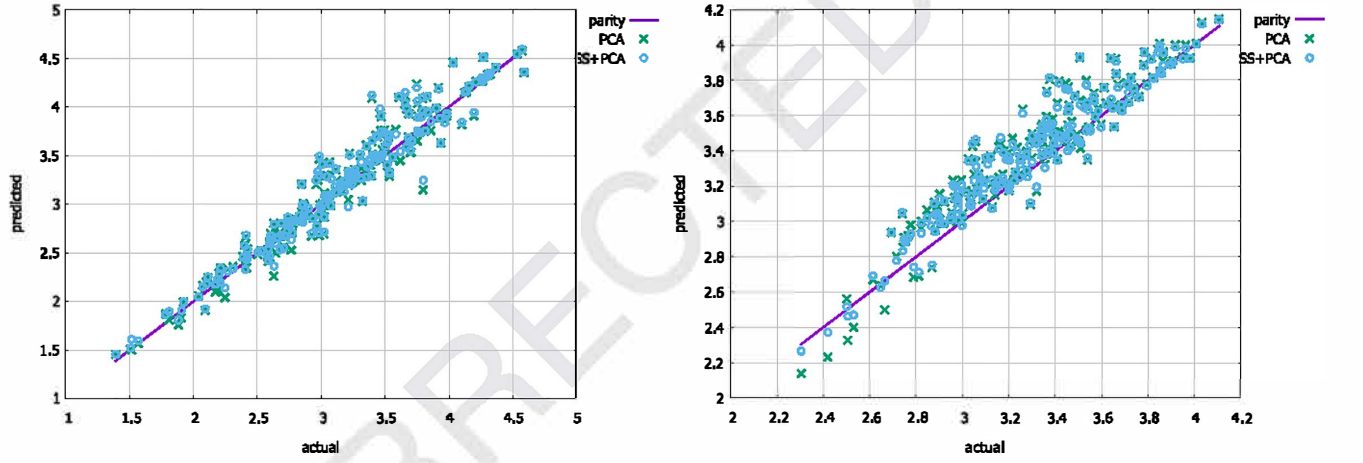
Index	30/35	35/40
R^2 train	3.41%	2.22%
R^2 test	10.34%	4.40%
wR^2 train	2.08%	1.03%
wR^2 test	5.26%	2.06%

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772



(a) Training set (b) Test set

Fig. 6. H_2, predicted vs. actual plots, cubic model (5). The values given by PCA with 30 principal components in green (PCA 30); the values given by PCA with 35 principal components in blue (PCA 35); and the values given by PCA with 40 principal components in orange (PCA 40).



(a) Training set (b) Test set

Fig. 7. H_2, predicted vs. actual plots, cubic model (5). The values given by the PCA with 35 principal components in green (PCA); and the values given by SS+PCA with 35 principal components in blue (SS+PCA).

Finally, Fig. 8a and b show the residual gap, calculated as

$$\frac{\sum_{j\in\mathcal{B}^m}\beta_j^m f_j(x) - z_n}{\sum_{j\in\mathcal{B}^m}\beta_j^m f_j(x)} \quad \forall n \in [N], \tag{8}$$

sorted according to the RON input value (in the figures we report only the values for H_2). It is worth noting that the surrogate models reproduced sufficiently the behaviour of the simulated values with regards to the process variable, which is indeed the control variable in the operating phase of the considered process: the values of the residual gaps for H_2 are in absolute value less than 0.25 for the training set and 0.15 for the test set.

In conclusion, in the CR case study the performances of the hybrid method are generally slightly better than the ones of the two-phase PCA.

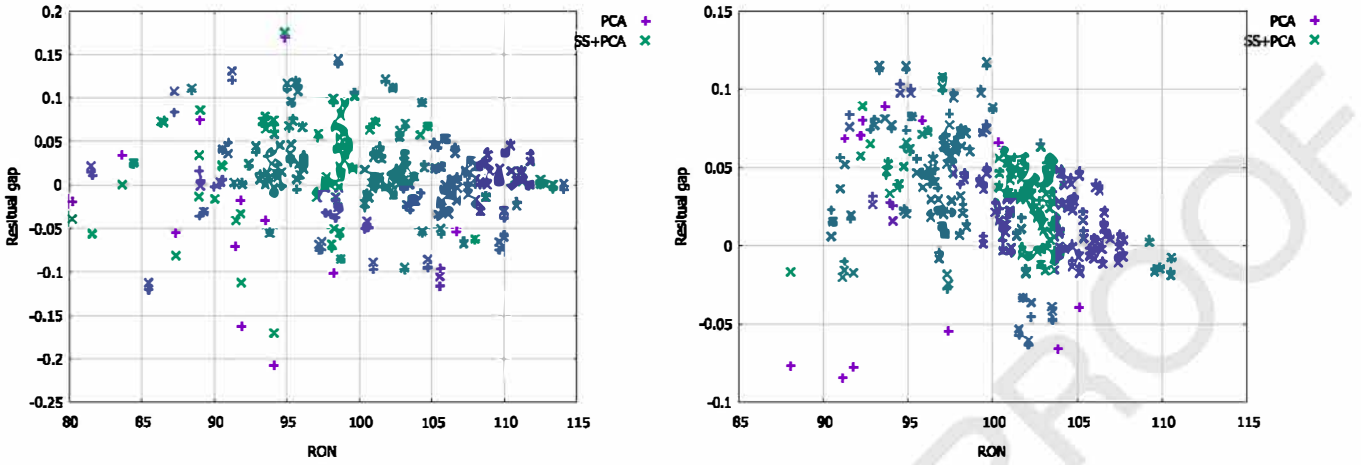
7.2. Case study 2: Isomerization

Isomerization (IS) is a chemical process which increases RON index of light hydrocarbon (C5–C6) by transforming n-alcanes into branched iso-alcanes with higher octane index (Valavarasu and Sairam, 2013; Sullivan et al., 2014). IS has been introduced in 1930s by Vladimir Ipatieff, who proposed a new chemical process to transform butane into isobutane, and used during the World War II to produce high octane aviation gasoline (for an historical analysis of IS process, see Sullivan et al. (2014)). Nowadays IS is fundamental to produce high octane fuel and reduce the level of benzene, aromatics and olefins in gasoline. IS unit is usually composed of a single reactors operating at relatively low temperatures (110–150°C) Valavarasu and Sairam (2013). In theory, in fact, hydrocarbons with C6 content could be treated via the catalytic reforming, but the constraint over the benzene content in the gasoline makes the process infeasible.

Low operating temperatures are necessary in order to minimize cracking of the hydrocarbons, and imply the chemical reactions are relatively slow: this effect is balanced, for instance, by means of highly active catalysts.

Table 7 reports the input and output compounds in the IS process. We consider the inverse of the weight hourly space velocity (WHSV), temperature, pressure, and mass fractions of the input compounds as input data and the mass fractions of the output compounds as output data.

As in the previous case study, we have one balance constraints for each chemical element (carbon and hydrogen) which should be satisfied by all the input configurations belonging to the design space. A latin hypercube design of experiments is gener-



(a) Training set (b) Test set

Fig. 8. H_2, RON vs. residual gaps, cubic model (5). The residual gap given by the PCA with 35 principal components in violet (PCA); and the residual gap given by SS+PCA with 35 principal components in green (SS+PCA).

Table 7
Input and output compounds of the isomerization process.

Input compounds	
nP_5	pentane
iP_5	isopentane
nP_6	hexane
$2iP_6$	2-methylpentane
$3iP_6$	3-methylpentane
$22iP_6$	2,2-dimethylhexane
$23iP_6$	2,3-dimethylhexane
Output compounds	
nP_4	butane
iP_4	isobutane
nP_5	pentane
iP_5	isopentane
nP_8	hexane
$2iP_6$	2-methylhexane
$3iP_6$	3-methylhexane
$22iP_6$	2,2-dimethylhexane
$23iP_6$	2,3-dimethylhexane

Table 8
R^2, weighted R^2 (wR^2), and time (in seconds) for PCA per number of principal components.

	R^2		wR^2		
No. comp.	Train	Test	Train	Test	Time
50	< 0	< 0	< 0	< 0	599.80
90	72.36	72.31	98.43	98.65	800.49
91	71.89	72.37	98.46	98.72	837.86
92	71.97	72.38	98.50	98.72	964.91
93	72.03	72.41	98.54	98.82	882.16
94	72.05	72.50	98.54	98.78	884.78
95	72.01	72.45	98.54	98.79	810.49
96	72.02	72.51	98.55	98.84	853.40
97	72.34	72.79	98.72	98.91	933.40
98	72.74	73.16	98.93	98.13	835.45
avg	72.16	72.54	98.58	98.71	866.99

Table 9
R^2, weighted R^2 (wR^2), and time (in seconds) for SS+PCA per number of principal components.

	R^2		wR^2		
No. comp.	Train	Test	Train	Test	Time
47	66.21	< 0	93.44	48.03	488.01
48	65.80	< 0	93.46	50.15	487.16
49	72.05	68.38	97.91	94.05	467.78
50	72.37	73.24	98.40	99.42	283.67
avg	72.21	70.81	98.16	96.74	375.73

ated by means of the software package Advanced tools for optimization and uncertainty treatment (ATOUT 1.1) developed and maintained by IFP Energies nouvelles. In particular, we define a training set with 200 samples and a test set with 5000 samples. We use the simulation software developed at IFP Energies nouvelles in order to simulate the industrial process performances. The total number of basis functions in the model before dimensionality reduction is 98 for the cubic model (5).

The same numerical trends observed for the CR process are valid also for the IS. Tables 8 and 9 show the R^2, wR^2 and the computational time (in seconds) for the two-phase PCA and the hybrid algorithm, respectively (the average referred to the principal components with positive value of R^2 indices). We note the two-phase PCA method is able to reduce the number of basis functions by maintaining however a comparable model accuracy. The hybrid method achieves the same grade of model accuracy with less basis functions: in particular, already with 49 basis function the hybrid algorithm is rather capable to capture the informations of the (simulated) data, while in order to achieve the same model quality the two-phase PCA method approach needs 90 basis functions.

For the computational time related to the surrogate model generation, the lower number of basis functions implies the computational time for the hybrid method to reach a satisfactory grade of accuracy is approximately half of the time needed for the two-phase PCA time to achieve the same quality level.

Fig. 9a and b show the scatter parity plots for the $2iP_6$ for the training set (figures with caption (a)) and the test set (figures with caption (b)). As for the CR process case study, the quality of the surrogate models clearly depends on the grade of the polynomial models (see Fig. 9a and b) and on the presence of the additional non-negativity and molar conservation constraints (see Fig. 10a and b). The previous plots refer to models (4) and (5) with the *a priori* selection of the interactions (bilinear terms) between process and composition variables.

Fig. 11a and b report the parity plot for the two-phase PCA with 90, 95, and 98 principal components, showing the light quality improvement of the surrogate models by considering an increasing number of principal components.

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

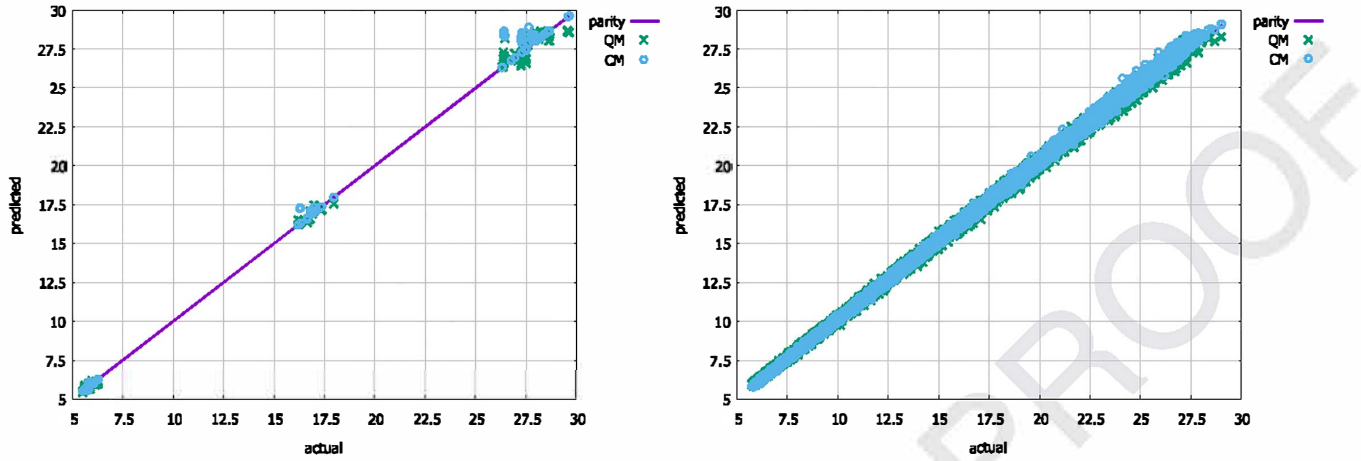


Fig. 9. $2iP_6$, predicted vs. actual plots. The values given by quadratic model (4) in green (QM); and the values given by the cubic model (5) in blue (CM).

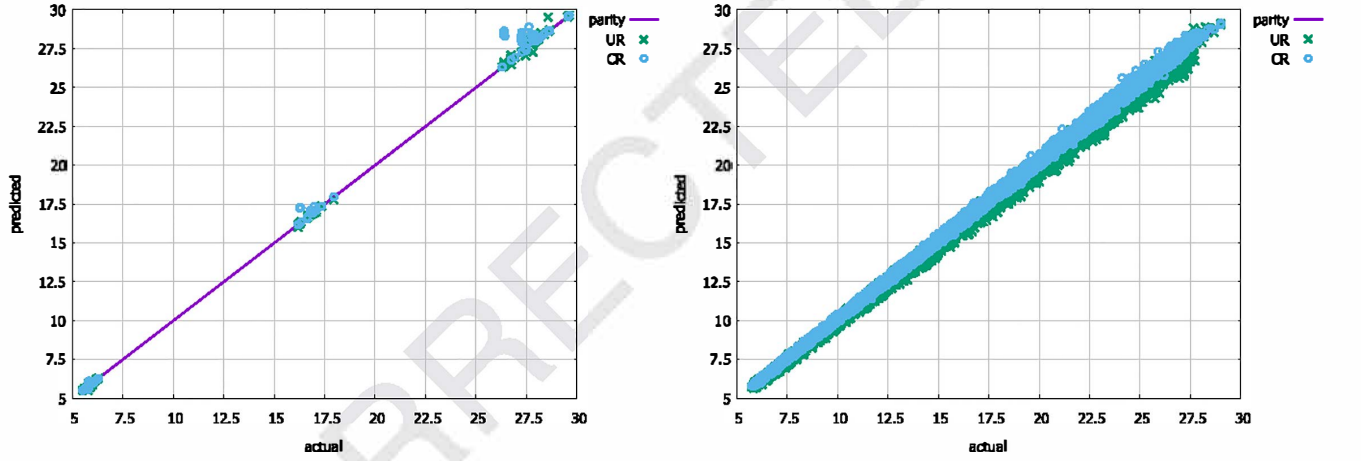


Fig. 10. $2iP_6$, predicted vs. actual plots, cubic model (5). The values given by the unconstrained regression in green (UR); and the values given by constrained regression in blue (CR).

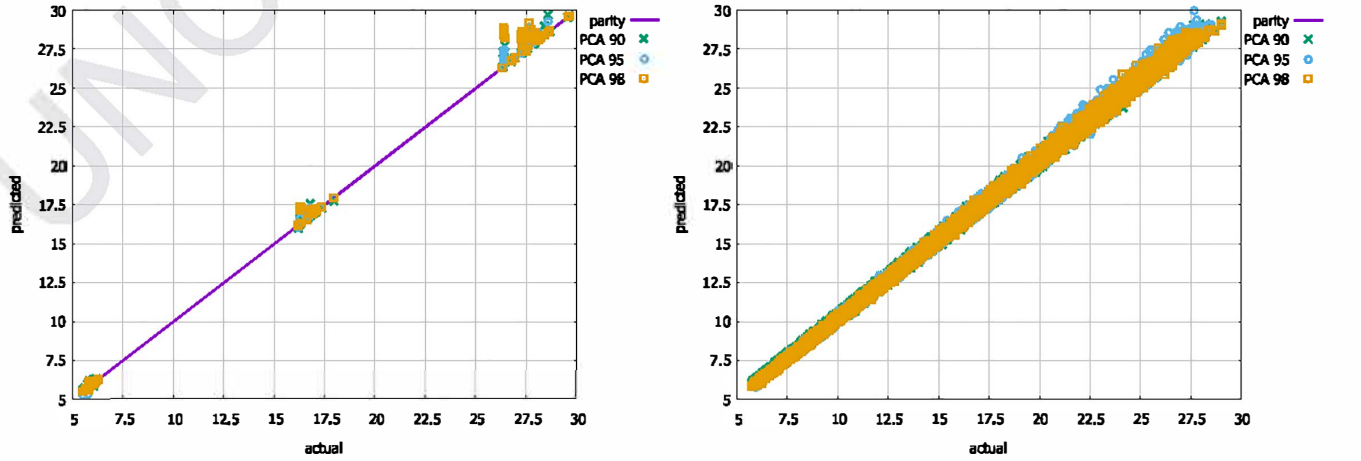


Fig. 11. $2iP_6$, predicted vs. actual plots, cubic model (5). The values given by PCA with 90 principal components in green (PCA 90); the values given by PCA with 95 principal components in blue (PCA 95); and the values given by PCA with 98 principal components in orange (PCA 98).

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

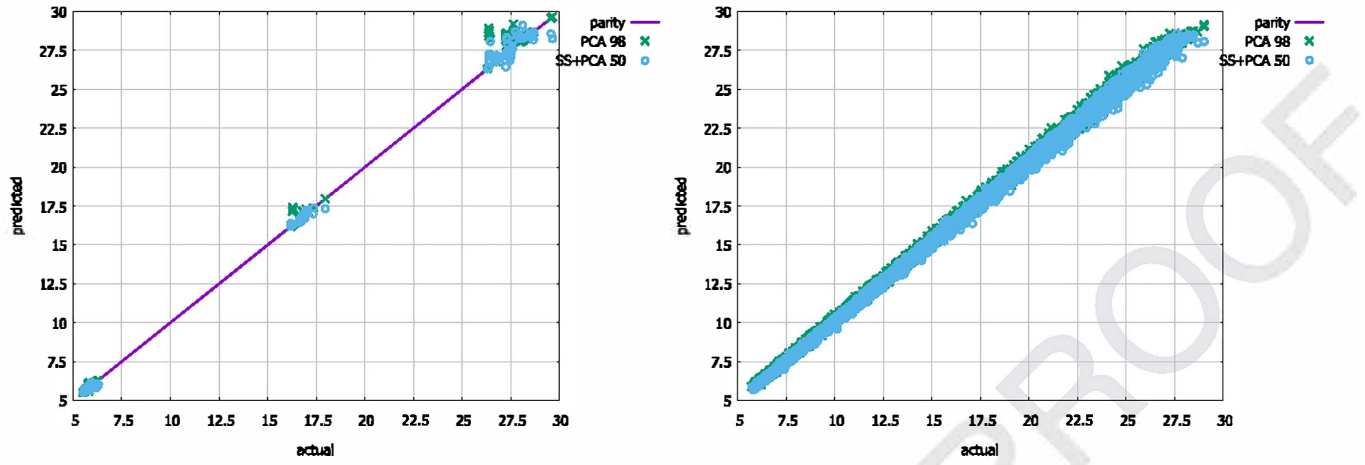


Fig. 12. $2iP_6$, predicted vs. actual plots, cubic model (5). The values given by the PCA with 98 principal components in green (PCA 98); and the values given by SS+PCA with 50 principal components in blue (SS+PCA 50).

Finally, Fig. 12a and b report the comparison between the two-phase PCA and the hybrid approach, graphically showing that the latter algorithm is characterized by a model accuracy comparable to the one of the two-phase PCA method, by considering, however, a lower number of basis functions.

In conclusion, the IS case study highlights the benefits of the hybrid approach, which is able to obtain the same model accuracy with about half the number of basis functions of the two-phase PCA approach.

8. Conclusions

In this paper, we have designed a systematic methodology to define and compute surrogate models for a given black-box process. In particular, we have discussed the DOE strategies and the main approaches for the identification of the surrogate model, focusing on the SS perspective. Moreover, we illustrate how to deal with possible intra-model and equality inter-model constraints. We have introduced a new two-phase PCA procedure for constrained regression problems by combining the two-phase approach in Cozad et al. (2015) with a PCA regression strategy. Therefore, we have discussed a possible hybrid strategy to combine the SS approach with the introduced two-phase PCA procedure.

The methods are evaluated and compared with respect to two case studies in petroleum refinery process, namely catalytic reforming and isomerization. For both case studies we have shown that the two-phase PCA method is able to reduce the number of basis functions required to obtain a satisfactory model accuracy, and the hybrid algorithm (the two-phase PCA preceded by a SS step) achieves a satisfactory model quality with a lower number of basis functions than the simple two-phase PCA methodology. The reduction in the number of basis function is significant in the isomerization case study, where the hybrid approach is able to obtain a satisfactory model accuracy with half the number of basis functions of the two-phase PCA.

In future work we would like to extend our methodology to consider also PLSR approaches. Moreover, our approaches can be easily extended to consider other functions than polynomials as basis functions: hence, analyzing the performances of these surrogate models in the context of PCA and SS+PCA could be another interesting future research axis. In our study, we have considered all the sampling points at once: consequently incrementally adding the sampling points could be an interesting strategy in order to obtain an accurate surrogate model with a lower number of sampling points. We have considered only noiseless data, we would like to test the two-phase procedure for noisy data relative to physical experiments in a further study.

Then, we are also interested in embedding the surrogate modeling approach into an optimization framework, where the surrogate model replaces the physical model to retrieve the optimal configuration and operating conditions of a given chemical process. The surrogate model is sufficiently accurate for this, and as its form is simple, it is fast to compute, and it allows the use of powerful global optimizers, that can fully exploit its analytic expression. Moreover, we are confident that the two-phase dimension reduction approaches described in the paper could be applied to other chemical processes. If needed, other basis functions than polynomial terms could be used as basis functions.

Acknowledgments

We are grateful to Master student Amina Bougueroua for preliminary computational experiments concerning the numerical characterization of the design of experiments for the isomerization process. We would thank also two anonymous referees whose suggestions significantly improve the quality of this paper.

References

Altissimi, R., Brambilla, A., Deidda, A., Semino, D., 1998. Optimal operation of a separation plant using artificial neural networks. Computers & Chemical Engineering 22 (Supplement 1), S939–S942.

Amouzgar, K., Strömberg, N., 2017. Radial basis functions as surrogate models with a priori bias in comparison with a posteriori bias. Structural and Multidisciplinary Optimization 55 (4), 1453–1469.

Ancheyta-Juárez, J., Villafuerte-Macías, E., 2001. Experimental validation of a kinetic model for naphtha reforming. Studies in Surface Science and Catalysis 133, 615–618.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J.J., Du Cruz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D., 1999. LAPACK User's Guide. Society for Industrial and Applied Mathematics (SIAM).

Audet, C., Le Digabel, S., Tribes, C., Rochon Montplaisir, V.,. The NOMAD project. Software available at https://www.gerad.ca/nomad/.

Beykal, B., Boukouvala, F., Floudas, C.A., Pistikopoulos, E.N., 2018. Optimal design of energy systems using constrained grey-box multi-objective optimization. Computers & Chemical Engineering 116, 488–502.

Beykal, B., Boukouvala, F., Floudas, C.A., Sorek, N., Zalavadia, H., Gildin, E., 2018. Global optimization of grey-box computational systems using surrogate functions and application to highly constrained oil-field operations. Computers & Chemical Engineering 114, 99–110.

Bhosekar, A., Ierapetritou, M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. Computers & Chemical Engineering 108, 250–267.

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

Bouhlel, M.A., Bartoli, N., Otsmane, A., Morlier, J., 2016. Improving kriging surrogates of high-dimensional design models by Partial Least Squares dimension reduction. Structural and Multidisciplinary Optimization 53 (5), 935–952.

Boukouvala, F., Floudas, C.A., 2017. ARGONAUT: AlgoRithms for Global Optimization of coNstrAined grey-box compUTational problems. Optimization Letters 11 (5), 895–913.

Boukouvala, F., Hasan, M.M.F., Floudas, C.A., 2017. Global optimization of general constrained grey-box models: new method and its application to constrained PDEs for pressure swing adsorption. Journal of Global Optimization 67 (1–2), 3–42.

Boukouvala, F., Misener, R., Floudas, C.A., 2016. Global optimization advances in mixed-integer nonlinear programming, MINLP, and constrained derivative-free optimization, CDFO. European Journal of Operational Research 255 (3), 701–727.

Box, G.E.P., Hunter, J.S., Hunter, W.G., 2005. Statistics for experimenters: Design, innovation, and discovery. John Wiley & Sons, Hoboken, NJ.

Broomhead, D.S., Lowe, D., 1988. Radial basis functions, multi-variable functional interpolation and adaptive networks. 4148. Royal Signal & Radar Establishment.

Buhmann, M.D., 2000. Radial basis functions. Acta Numerica 9, 1–38.

Caballero, J.A., Grossmann, I.E., 2008. An algorithm for the use of surrogate models in modular flowsheet optimization. AIChE Journal 54 (10), 2633–2650.

Clarke, S.M., Griebsch, J.H., Simpson, T.W., 2004. Analysis of support vector regression for approximation of complex engineering analyses. Journal of Mechanical Design 127 (6), 1077–1087.

Coetzer, R., Haines, L.M., 2017. The construction of D- and I-optimal designs for mixture experiments with linear constraints on the components. Chemometrics and Intelligent Laboratory Systems 171, 112–124.

Conn, A.R., Le Digabel, S., 2013. Use of quadratic models with mesh-adaptive direct search for constrained black box optimization. Optimization Methods and Software 28 (1), 139–158.

Cornell, J.A., 2002. Experiments with mixtures: Designs, models, and the analysis of mixture data. John Wiley & Sons, New York.

Cozad, A., Sahinidis, N.V., Miller, D.C., 2014. Learning surrogate models for simulation-based optimization. AIChE Journal 60 (6), 2211–2227.

Cozad, A., Sahinidis, N.V., Miller, D.C., 2015. A combined first-principles and data-driven approach to model building. Computers & Chemical Engineering 73, 116–127.

Cunningham, P., 2007. Dimension reduction. UCD-CSI-2007-7. University College Dublin.

Davis, S.E., Cremaschi, S., Eden, M.R., 2018. Efficient surrogate model development: Impact of sample size and underlying model dimensions. Computer Aided Chemical Engineering 44, 979–984.

van Dam, E.R., 2008. Two-dimensional minimax Latin hypercube designs. Discrete Applied Mathematics 158 (18), 3483–3493.

van Dam, E.R., Husslage, B.G.M., den Hertog, D., Melissen, H., 2007. Maximin Latin hypercube design in two dimensions. Operations Research 55 (1), 158–169.

Design Expert,. https://www.statease.com/software.html.

Eason, J., Cremaschi, S., 2014. Adaptive sequential sampling for surrogate model generation with artificial neural networks. Computers & Chemical Engineering 68 (4), 220–232.

Fahmi, I., Cremaschi, S., 2012. Process synthesis of biodiesel production plant using artificial neural networks as the surrogate models. Computers & Chemical Engineering 46, 105–123.

Fodor, I.K., 2002. A survey of dimension reduction techniques. Technical Report. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.

Forrester, A.I.J., Keane, A.J., 2009. Recent advances in surrogate-based optimization. Progress in Aerospace Sciences 45 (1–3), 50–79.

Forrester, A.I.J., Sobester, A., Keane, A.J., 2008. Engineering design via surrogate modelling: A practical guide. John Wiley & Sons, Hoboken, NJ.

Garud, S.S., Karimi, I.A., Kraft, M., 2017. Design of computer experiments: A review. Computers & Chemical Engineering 106, 71–95.

Garud, S.S., Karimi, I.A., Kraft, M., 2017. Smart sampling algorithm for surrogate model development. Computers & Chemical Engineering 96, 103–114.

Gaspar, B., Teixeira, A.P., Guedes Soares, C., 2017. Adaptive surrogate model with active refinement combining Kriging and a trust region method. Reliability Engineering & System Safety 165, 277–291.

Gjervan, T., Prestvik, R., Holmen, A., 2004. Catalytic reforming. In: Baerns, M. (Ed.), Basic Principles in Applied Catalysis. Springer, Berlin, Heidelberg, pp. 125–158.

Goos, P., Jones, B., Syafitri, U., 2016. I-optimal design of mixture experiments. Journal of the American Statistical Association 111 (514), 899–911.

Hardy, R.L., 1971. Multiquadratic equations of topography and other irregular surfaces. Journal of Geophysical Research 76 (8), 1905–1915.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity: The Lasso and generalizations. CRC Press.

Henao, C.A., Maravelias, C.T., 2010. Surrogate-based process synthesis. Computer Aided Chemical Engineering 28, 1129–1134.

Henao, C.A., Maravelias, C.T., 2011. Surrogate-based superstructure optimization framework. AIChE Journal 57 (5), 1216–1232.

Huber, P.J., 1981. Robust statistics. John Wiley & Sons, Hoboken, NJ.

Husslage, B.G.M., Rennen, G., van Dam, E.R., den Hertog, D., 2011. Space-filling Latin hypercube designs for computer experiments. Optimization and Engineering 12 (4), 611–630.

IBM ILOG CPLEX,. https://www.ibm.com/analytics/cplex-optimizer.

Ivanciuc, O., 2007. Application of support vector machine in chemistry. In: Lipkowitz, K., Cundari, T., Boyd, D. (Eds.), Reviews in Computational Chemistry, 23. John Wiley & Sons, Hoboken, NJ.

Jin, R., Chen, W., Simpson, T.W., 2001. Comparative studies of metamodelling techniques under multiple modelling criteria. Structural and Multidisciplinary Optimization 23 (1), 1–13.

Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. Journal of Statitical Planning and Inference 26 (2), 131–148.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13 (4), 455–492.

Joseph, V.R., Hung, Y., 2008. Orthogonal-maximin latin hypercube designs. Statistica Sinica 18 (1), 171–186.

Kim, S.H., Boukouvala, F., 2019. Machine learning-based surrogate modeling for data-driven optimization: A comparison of subset selection for regression technique. Optimization Letters 1–22.

Kleijnen, J.P.C., 2009. Kriging metamodeling in simulation: A review. European Journal of Operational Research 192 (3), 707–716.

Krahmer, F., Ward, R., 2016. A unified framework for linear dimensionality reduction in l1. Results in Mathematics 70 (1–2), 209–231.

Krige, D.G., 1952. A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy 52 (9), 201–203.

Lapinski, M.P., Metro, S., Pujadó, P.R., Moser, M., 2014. Catalytic reforming in petroleum processing. In: Treese, S., Jones, D., Pujadó, P. (Eds.), Handbook of Petroleum Processing. Springer, Cham, pp. 1–25.

Le Digabel, S., 2011. Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. ACM Transactions on Mathematical Software 37 (4), 44:1–44:15.

Li, H., Lianh, Y., Xu, Q., 2009. Support vector machines and its applications in chemistry. Chemometrics and Intelligent Laboratory Systems 95 (2), 188–198.

Matheron, G., 1963. Principles of Geostatistics. Economic Geology 58 (8), 1246–1266.

McBride, K., Sundmacher, K., 2019. Overview of surrogate modeling in chemical process engineering. Chemie Ingenieur Technik 91 (3), 228–239.

McCarl, B. A., Meeraus, A., van der Eijk, P., Bussieck, M., Dirkse, S., Nelissen, F., 2017. McCarl Expanded GAMS User Guide, GAMS Release 24.6.GAMS Development Corporation, Washington, DC, USA.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Mencarelli, L., Chen, Q., Pagot, A., Grossmann, I.E., 2019. A review on superstructure optimization approaches in process system engineering. Technical Report. IFP Energies nouvelles, Solaize, France, and Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh.

Mencarelli, L., Duchêne, P., Pagot, A., 2019. Optimization approaches to the integrated system of catalytic reforming and isomerization processes in petroleum refinery. Technical Report. IFP Energies nouvelles, Solaize, France.

Miller, A., 2002. Subset selection in regression. Chapman & Hall/CRC, Boca Raton, Florida.

Morris, M.D., Mitchell, T.J., 1995. Exploratory designs for computational experiments. Journal of Statitical Planning and Inference 43 (3), 381–402.

Müller, J., Shoemaker, C.A., 2015. Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems. Journal of Global Optimization 60 (2), 123–144.

Nascimiento, C.A.O., Giudici, R., Guardani, R., 2000. Neural network based approach for optimization of industrial chemical processes. Computers & Chemical Engineering 24 (9–10), 2303–2314.

Park, J., Sandberg, I.W., 1993. Approximation and radial-basis-function networks. Neural Computation 5 (2), 305–316.

Petelet, M., Iooss, B., Asserin, O., Loredo, A., 2010. Latin hypercube sampling with inequality constraints. AStA Advances in Statistical Analysis 9 (4), 325–339.

Pronzato, L., 2017. Minimax and maximin space-filling designs: Some properties and methods for construction. Journal de la Société Française de Statistique 158 (1), 7–36.

Psaltis, A., Sinoquet, D., Pagot, A., 2016. Systematic optimization methodology for heat exchanger network and simultaneous process design. Computers & Chemical Engineering 95, 146–160.

Queipo, N.V., Hafka, R.T., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P.K., 2005. Surrogate-based analysis and optimization. Progress in Aerospace Sciences 41 (1), 1–28.

Rahimpour, M.R., Jafari, M., Iranshavi, D., 2013. Progress in catalytic naphtha reforming process: A review. Applied Energy 109, 79–93.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Statistical Science 4 (4), 409–485.

Smith, W.F., 2005. Experimental design for formulation. ASA-SIAM Series on Statistics and Applied Probability, 15. SIAM, Philadelphia.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14 (3), 199–222.

Sorek, N., Gildin, E., Boukoulava, F., Beykal, B., Floudas, C.A., 2017. Dimensionality reduction for production optimization using polynomial approximations. Computational Geosciences 21 (2), 247–266.

Straus, J., Skogestad, S., 2017. Use of latent variables to reduce the dimension of surrogate models. Computer Aided Chemical Engineering 40, 445–450.

Straus, J., Skogestad, S., 2017. Variable reduction for surrogate modelling. In: Proceeding of Foundations of Computer-Aided Process Operations, Tucson, AZ, USA, 8-12 January 2017.

Straus, J., Skogestad, S., 2018. Surrogate model generation using self-optimizing variables. Computers & Chemical Engineering 119, 143–151.

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772

Straus, J., Skogestad, S., 2019. A new termination criterion for sampling for surrogate model generation using partial least squares regression. Computers & Chemical Engineering 121, 75–85.
Sullivan, D., Metro, S., Pujadó, P.R., 2014. Isomerization in Petroleum Processing. In: Treese, S., Jones, D., Pujadó, P. (Eds.), Handbook of Petroleum Processing. Springer, Cham, pp. 1–15.
Tawarmalani, M., Sahinidis, N.V., 2005. A polyhedral branch-and-cut approach to global optimization. Mathematical Programming 103 (2), 225–249.
Turaga, U.T., Ramanathan, R., 2003. Catalytic naphtha reforming: Revisiting its importance in the modern refinery. Journal of Scientific and Industrial Research 62 (10), 963–978.
Valavarasu, G., Sairam, B., 2013. Light naphtha isomerization process: A review. Petroleum Science and Technology 31 (6), 580–595.
Vapnik, V., 1995. The nature of statistical learning theory. Springer-Verlag, NY.
Vapnik, V., Golowich, S., Smola, A.J., 1997. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer, M., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA, pp. 281–287.
Viana, F.A.C., 2016. A tutorial on latin hypercube design of experiments. Quality and Reliability Engineering International 32 (5), 1975–1985.
Vu, K.K., D'Ambrosio, C., Hamadi, Y., Liberti, L., 2017. Surrogate-based methods for black-box optimization. International Transactions in Operational Research 24 (3), 393–424.
Wilson, Z.T., Sahinidis, N.V., 2017. The ALAMO approach to machine learning. Computers & Chemical Engineering 106, 785–795.

Please cite this article as: L. Mencarelli, A. Pagot and P. Duchêne, Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes, Computers and Chemical Engineering, https://doi.org/10.1016/j.compchemeng.2020.106772