Benefits of feasibility constrained sampling on unit operations surrogate model accuracy

Tesfayesus Zinare Mamo, Alessandro Di Pretoro, Valentina Chiari, Ludovic Montastruc, Stéphane Négny

▶ **To cite this version:**

Tesfayesus Zinare Mamo, Alessandro Di Pretoro, Valentina Chiari, Ludovic Montastruc, Stéphane Négny. Benefits of feasibility constrained sampling on unit operations surrogate model accuracy. Computers & Chemical Engineering, 2023, 173, pp.108210. 10.1016/j.compchemeng.2023.108210 . hal-04312517

HAL Id: hal-04312517
https://hal.science/hal-04312517

Submitted on 12 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Benefits of Feasibility Constrained Sampling on Unit Operations Surrogate Model Accuracy

Tesfayesus Zinare[a], Alessandro Di Pretoro[a,*], Valentina Chiari[b], Ludovic Montastruc[a], Stéphane Negny[a]

[a]*Laboratoire de Génie Chimique, Université de Toulouse, CNRS/INP/UPS, Allée "E. Monso", 4, Toulouse, 31432, France*
[b]*Politecnico di Milano, Piazza "L.Da Vinci", 32, Milano, 20133, Italy*

Abstract

Due to the increasing amount of data to be treated and, thus, to the need of more computational effective strategies, surrogate modeling has become a topic of major interest during the last decades. In the modeling procedure the data set generation plays a role of critical importance in terms of accuracy of the final solution. When dealing with chemical processes, some points generated during the sampling phase could fall into a physically unfeasible region of the domain. That is why, the design of experiment step needs to be integrated with constraints able to describe the feasibility limitations of the system. By means of a dedicated interface between ProSim® process simulator, ALAMO® modeling software and Matlab® for data processing, this research work assesses the benefits of constraints and subset-based approach on the accuracy of the model in terms of outlet streams and economic indicators for separation units.

Keywords: surrogate modeling, design of experiment, flash, distillation, chemical processes

1. Introduction

During the last decades, the digital transition is having an impact of non-negligible importance on all the aspects of the engineering domain. If,

*Corresponding author.
Email address: alessandro.dipretoro@ensiacet.fr (Alessandro Di Pretoro)

on the one hand, it allows to deal with systems with higher and higher complexity, on the other it implies a constantly increasing amount of data to be processed. As a consequence, the need of more computationally effective methodologies is of critical importance to make the promising innovation viable and feasible from a practical point of view. That is why, in recent years, surrogate modeling has seen a renewed interest in several scientific areas. The data-driven modeling approach can be exploited for three main purposes[1]:

- Compensate the absence of a phenomenological model
- Reduce the computational effort by replacing complex systems with more simple input-output relationships.
- Reduce computational effort and enhance convergence of optimization algorithms[2].

In particular, in chemical and process engineering considerable benefits were obtained so far by data-driven modeling since the phenomenological model equations are usually characterized by non-linear expressions related to equilibrium conditions and chemical kinetics. Examples of effective applications can be found for unit modeling[3, 4], energy systems[5], feasibility analysis and optimization[6, 7], optimal scheduling for Demand-Side Management[8], process control[9] or even integrated planning, scheduling and control[10].

In the majority of literature studies, a wide range of modeling procedures can be found. The most common among them are Response Surface Methodology (RSM)[11], Kriging[12], Artificial Neural Network (ANN)[13], Radial Basis Functions (RBF)[14] or again Support Vector Regression (SVR)[15]. However, beside the diversity of them, the common aspect of all these approaches is that the Design of Experiment step is the most crucial phase. In general, sampling strategies can be divided into two main categories, namely one-shot (or once-through, non-adaptive) and sequential (or adaptive, model-based). The former one has the purpose to generate all points at once by exploiting an optimal strategy for design space filling while the latter is a recursive method based on DoE adjustments at every modeling iteration loop. The first method implies a higher preliminary effort to generate a huge size dataset but lower amount of calculations during the modeling phase. On the contrary, the second approach distributes the computational effort

over the loops but is based on smaller samples whose quality increases at each iteration[1]. In fact, it was proved that the relationship between sample size and model dimension with the different sampling strategies has a non-negligible impact on the final outcome[16].

Furthermore, during recent years, many other strategies exist to improve sampling quality have been studied. Two of the most interesting are the subset partition of the global data set[17] and the inclusion of feasibility constraints[18]. In particular, the first one is able to increase the local accuracy of the model and the second one avoids the generation of output values that do not lie in the feasible region. As a consequence of these modeling procedure adjustments, computational times and convergence of the algorithm are affected by perturbations that still need more analysis.

A first relevant step towards the improvement of constrained modeling algorithm can be found in research works proposed by Beykal et al.[19] for DAE optimization algorithms and by Boukouvala et al.[20] that introduced the so called ARGONAUT optimization approach. For the second study, a global optimization algorithm for general constrained grey-box models has been proposed and tested on constrained PDE for a pressure swing adsorption case study[21]. However, in this case, the explicit function that correlates the independent variables involved in the feasibility constraint functions should be known.

Based on these latest advances then, in this research work, the impact of feasibility constraints and the use of sub-domains on data sampling is applied on unit operations. To be more precise, their benefits are assessed by means of two separations case study, namely a flash separation and a distillation column unit, and the resulting outcome is analyzed in detail. Further information about the case study and the methodology are provided in the next sections.

2. Case study

In this research work two case studies of different complexity are presented. They both refer to a toluene-biphenyl mixture separation. The choice of a mixture with an ideal thermodynamic behaviour has the purpose to highlight the benefits of surrogate modeling for equilibrium units even in absence of non-idealities and it represents the starting point for further analysis of more complex mixtures and units. The details for each unit along with

(a) *Flash separation unit* (b) *Distillation unit*

Figure 1: *Toluene – Biphenyl separation units*

the input and output parameters are better described in the corresponding sections here below.

2.1. Flash separation

Before the analysis of the distillation column, a simple flash separation is tested first (cf. Figure 1a). In particular, the selected unit is an adiabatic flash tank whose purpose is just to separate liquid and vapor fractions of the feed stream. The input variables for this study are toluene and biphenyl partial flowrates, feed stream temperature and pressure according to the variation ranges listed in Table 1. The output variables that are stored from simulation are toluene and biphenyl partial flowrates for the top and the bottom products.

For this unit, the separation feasibility constraint that will be employed in the next step concerns the enthalpic state of the feed, i.e. liquid phase below bubble point and vapor above dew point are not included in the model equations aimed at describing the equilibrium outlet streams.

2.2. Distillation column

The second case study is based on the same separation to be performed by means of a standard distillation column with 12 equilibrium stages. In this example, the output variables that are stored from simulation, beside toluene

Input variable	Symbol	Unit	min value	max value
Toluene flowrate	F_{tol}	$kmol/h$	30	50
Biphenyl flowrate	F_{biph}	$kmol/h$	1	6
Feed pressure	F_P	atm	1	4
Feed temperature	F_T	K	273.15	523.15

Table 1: Flash separation input variables

and biphenyl partial flowrates for distillate and bottom, are also OPerating (OPEX) and CAPital (CAPEX) EXpenses. In order to be uniquely defined, the distillation unit needs two degree of freedom to saturate. For this study distillate and reflux streams flowrates have been selected.

However, for this unit the feasibility constraints are more complex than for the flash. Since some of the input variables are correlated, the deviation range of part of them is related to that of the others. Therefore, the distillate flowrate cannot be higher than the feed one and the reflux ratio should have a definite value higher than zero. The DoE domain for distillation is then resumed in Table 2.

Input variable	Symbol	Unit	min value	max value
Toluene flowrate	F_{tol}	$kmol/h$	30	50
Biphenyl flowrate	F_{biph}	$kmol/h$	1	6
Feed pressure	F_P	atm	1	4
Feed temperature	F_T	K	273.15	523.15
Reflux flowrate	D_R	$kmol/h$	> 0	$< \infty$
Distillate flowrate	D_L	$kmol/h$	> 0	$< F_{in}$
Murphree efficiency	η	/	> 0	1

Table 2: Distillation column input variables

3. Methodology

The study was carried out by means of an interface between Matlab®, ProSimPlus® software for process simulation and ALAMO® for surrogate modeling. The data import and export is performed by means of .csv flowsheets that are updated at each optimization loop. Each of the modeling steps as well as the data analysis are discussed in detail in the following sections.

3.1. Design of experiment

Based on the selected domain for the input variables, the Design of Experiment (hereafter DoE) strategy is set up for the modeling phase. In this study two of the most common sampling strategies are adopted, namely Latin Hypercube Sampling (LHS)[22] and Quasi-Monte Carlo Halton's[23] sampling.

On the one hand, the LHS method was developed to reduce variance in the resulting distribution and increase the space-filling capacity. In fact, the Latin hypercube is a statistical sampling method that generalizes the concept of the Latin square: given a finite number of dimensions, each axis-aligned hyperplane only contains one sample point each; this way all portions of the continuous design space can be represented. On the other hand the Halton's method was conceived to have a faster rate of convergence with respect to the Monte-Carlo one. This is due to the fact that it is a low-discrepancy sequence, i.e. although it appears to be random, it is a set of deterministic samples with high space-filling efficiency and high domain uniformity. In fact, Halton's sampling is a general n-dimension sequence of samples in the unitary hypercube $[0,1]^n$; for each dimension, it uses several prime bases in order to fill the design space in a high-uniform manner.[24]. Graphical examples of these two sampling methods are given in Figure 2.

In this study the sampling step is performed by means of the Matlab® functions lhsdesign and haltonset respectively (the dedicated DoE toolbox could be used as well). Sets of 10000 elements are obtained and then saved on a .csv file that is later imported in ProSimPlus® for simulation as better explained in the following section.

3.2. Simulation

The process simulation is performed by means of ProSimPlus® simulator for samples generated with both the methodologies with conventional flash and distillation units modules. An embedded script is used for a quick data import and simulations are automatically run as shown in Figure 3.

At this stage, the first quality improvement concerning feasibility takes place. All points who generate unfeasible simulations according to the criteria explained in Section 2 are filtered and removed from the output dataset. The obtained results are then stored in a new .csv file thanks to an embedded script that respectively prints the conserved input and generated output parameters of interest that will be used for modeling as described in the fol-

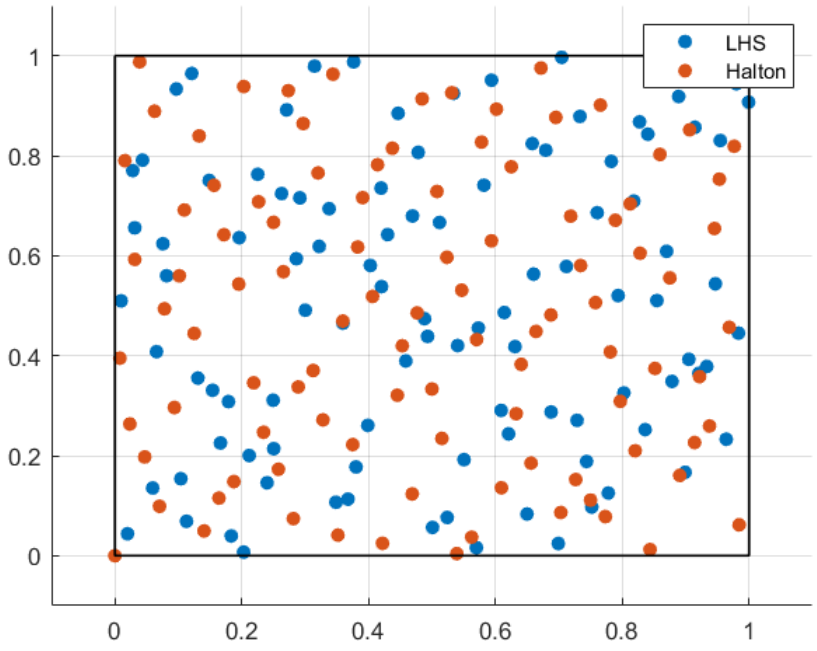


Figure 2: *ExampleofLHSandHaltonsampling*

lowing section. In terms of computational effort, one simulation of this kind takes about 1.2 s to be performed.

3.3. Surrogate modeling

The surrogate modeling step is then performed via ALAMO® ((Automatic Learning of Algebraic MOdels). ALAMO® is a software developed in 2014 by Cozad, Sahinidis and Miller[25]. Its purpose is to address the problem of derivative-free optimization and, thus, to generate the simplest and most accurate algebraic surrogate model of black-box systems, for which an experimental set-up or a simulator is currently available. It is based on a a three-step iterative process that can be resumed as follows:

- the N points from the initial DoE set are interrogated;
- In the second step, an initial algebraic model is created with integer optimization techniques, starting from the original training set. This

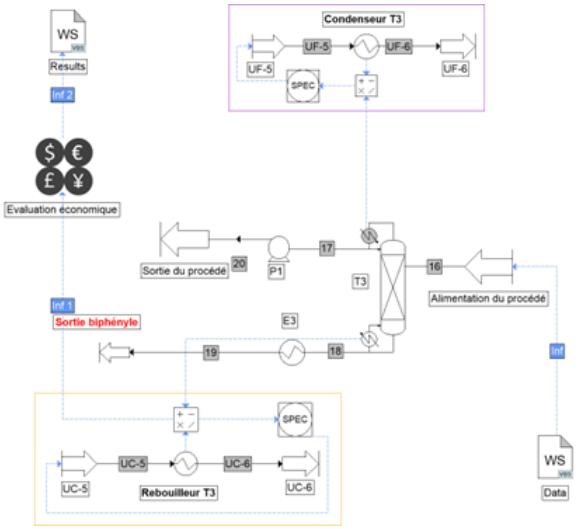


Figure 3: *Example of LHS and Halton sampling*

way, the best subset of simple basis functions is chosen from a set of potential functions, that can be used in construction of the model. The surrogate model is then obtained as a linear combination of non-linear basis functions. To reduce the coefficients regression complexity a "best-subset method" can be applied in order to list the possible combinations/subsets of base functions and to choose the best one via according to the best model fitness;

- Finally, the sample points where the model is lacking in accuracy are identified with an adaptive-sampling methodology based on derivative free optimization solvers. The initial training set is updated and the second step of the algorithm is repeated until the accuracy of the model is confirmed.

An example of the software interface is given in Figure 4.

In this research, the input variables are imported by the .cvs file obtained from Matlab® and the coefficients for the output analytical functions are shown in the results window.

In this phase, the second quality improvement strategy takes place. Two kind of data subsets are organised based on the inlet feed temperature and the inlet vaporization degree as discussed in the previous section.

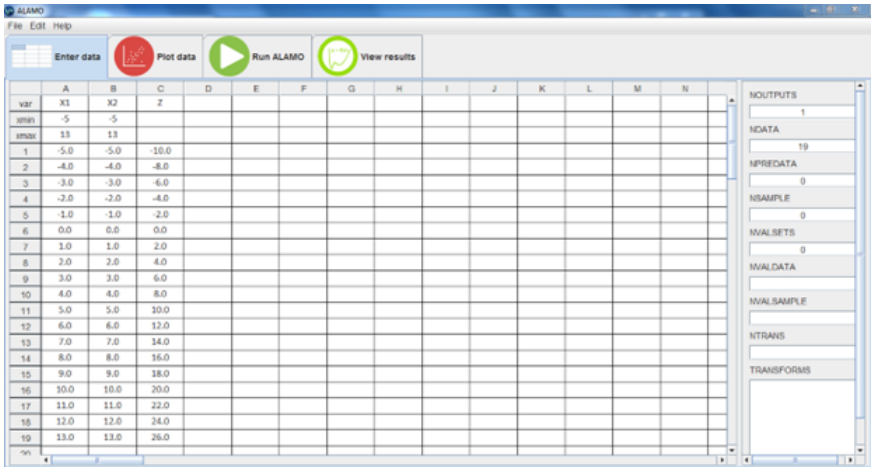


Figure 4: *ALAMO® interface*

Therefore, after the model equation is obtained for each of the output variables presented in Section 2, performance indicators are calculated to assess the quality of the result. To be more precise, the comparison has been performed with points that are generated on purpose and that do not belong to the set employed for the modeling phase, i.e. non-optimized points. Based on this methodology, the results of this study are presented in detail in the following section for each case study.

4. Results

According to the procedure explained in the previous section, the modeling algorithm was performed for the two case studies. The obtained results are discussed here below according to the specific unit. One run of the algorithm for a subset range took about 18 minutes for the flash and 30 minutes for the column simulations included. At the beginning all basis functions in ALAMO® have been selected for the functions analytical form and, afterwards, all the terms with negligible impact on the function values have been removed in order to further smooth the calculations. In particular, examples of the obtained model equations, quality fitting, errors and variance are provided to give a general overview of the research outcome.

4.1. Flash separation

The flash separation unit was modeled according to the procedure previously described. In a first phase, the modeling step with ALAMO® was

carried out on the global data set.

First of all, the so defined "unfeasible" points, according to the previously defined constraints, have been removed. On average, about the 23% of points have been ignored due to the combination of temperature and pressure falling outside the defined range. Then, the remaining samples have been imported in ALAMO® to model output variables with all basis functions. The obtained outcome was based on trigonometric functions such as:

$$F_{tol,bot} = -0.474 \cdot cos(F_{tol}) + 0.580 \cdot cos(F_T) \tag{1}$$

with a Maximum Absolute Error (MAE) in some points higher than 200%. This is due to the aliasing effect that depends on the amount of unknown variables to be regressed (in this case the model equation coefficients) and on the sample size according to the Nyquist theorem (undersampling or oversampling)[26]. Since there is no way to directly develop anti-aliasing filters[27] in the software, the action should be taken on sampling. That's is why the second adjustment related to subsets definition was considered strictly required.

Among the input parameters, temperature was selected as the first one to be used for this step. The interval delimited by the lowest and highest temperatures of the feasible points was divided into ten intervals of 12°C each. In this case, the obtained functions have a much more reliable form; for example the toluene bottom flowrate previously presented in Equation 1, is now given in the interval T=[470; 482] K by:

$$\begin{aligned} F_{tol,bot} = {} & 1.05 \cdot F_{tol} - 0.98 \cdot F_{biph} + 1.97 \cdot F_T - 877.68 \cdot ln(F_T) \\ & + 0.001 \cdot e^{F_P} + 1.53 \cdot cos(F_P) + 0.04 \cdot cos(F_T) + 4474.93 \end{aligned} \tag{2}$$

The analysis of the obtained results along with the MAE and the number of variables are summarized in Table 3. While the average error is always about a few percent, the maximum error for some points can also go up to 27%. In particular, it can be noticed that the MAE is usually higher when a lower amount of coefficients, i.e. basis functions, are present in the model equation. Although not discussed in this table, the average accuracy shows to be higher in central subsets than in boundary ones.

Furthermore, graphical examples for variables fitting with respect to temperature based subsets are shown in Figure 5.

	$F_{tol,top}$		$F_{biph,top}$		$F_{tol,bot}$		$F_{biph,bot}$	
F_T **range**	**MAE%**	n_{var}	**MAE%**	n_{var}	**MAE%**	n_{var}	**MAE%**	n_{var}
388-400	9.37	20	0.02	20	6.81	19	1.87	21
400-412	12.89	1	4.18	10	11.98	2	11.46	1
412-424	20.02	1	13.10	5	27.63	1	20.61	1
424-436	11.02	1	19.59	2	12.51	2	10.03	1
436-448	9.84	2	11.91	10	12.81	1	13.67	1
448-460	26.36	6	24.31	2	5.16	4	8.94	4
460-472	16.00	3	46.75	1	15.05	1	4.74	5
472-484	5.43	1	61.49	7	1.48	9	1.91	6
484-496	19.42	1	24.47	1	0.38	8	7.95	9
496-508	0.76	25	2.21	18	0.00	23	0.04	23

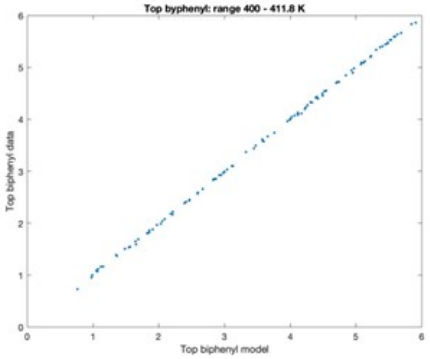
Table 3: MAE and number of variables vs. α

After that, the same approach was used accounting for the feed vaporization rate α. Subsets have been defined according to the amount of data. As already explained, the deviation range goes from the bubble point to the dew point. The size of each subset is equal to 0.1 except for the first and last ones whose size is 0.05 to have more accuracy in correspondence of the phase transition. For instance, the partial toluene flowrate in the bottom product in the interval α=[0.05; 0.15] is given by:

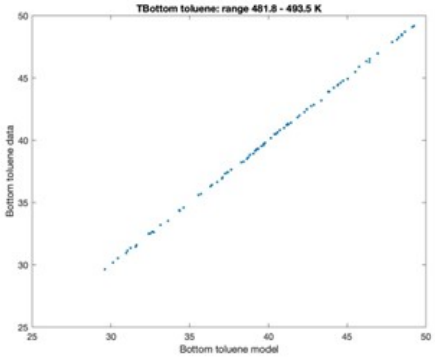
$$\begin{aligned} F_{tol,bot} = & -20.6 \cdot F_P - 0.82 \cdot F_T + 352.7 \cdot ln(F_T) + 0.003 \cdot e^{F_{biph}} \\ & + 0.019 \cdot sin(F_{tol}) - 0.04 \cdot F_{tol} \cdot F_P + 0.0025 \cdot F_{tol} \cdot F_T \\ & - 0.0027 \cdot F_{biph} \cdot F_T + 0.043 \cdot F_P \cdot F_T - 1779.3 \end{aligned} \tag{3}$$

Even for this case, the MAE and the number of variables are listed in Table 4 and examples for variables fitting with respect to vaporization fraction based subsets are plotted in Figure 6. Remarks similar to those of the temperature discretization apply for the vaporization fraction as well. In particular, for those points in proximity of highly vaporized feed, the error is almost equal to zero and the average error is lower than 1%. Thus, it can be concluded that the vaporization rate.

For the sake of completeness, although not reported in this publication, the analysis was carried out also with subsets based on the other input variables. The results were much less satisfying than those related to temperature

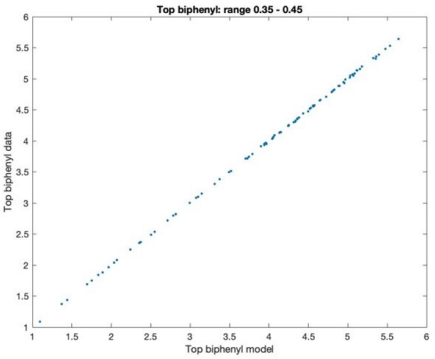


(a) $F_{biph,top}$ *in* $400 - 412$ K *interval*

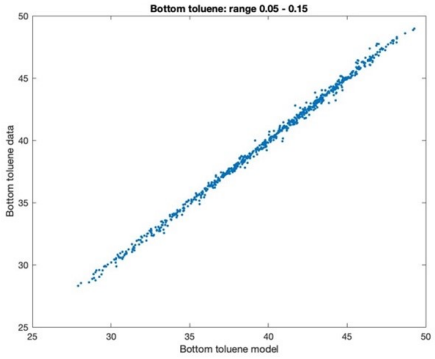


(b) $F_{tol,bot}$ *in* $482 - 494$ K *interval*

Figure 5: $Temperature - based\ subsets\ fitting$



(a) $F_{biph,top}$ *in* $0.35 - 0.45$ *interval*



(b) $F_{tol,bot}$ *in* $0.05 - 0.15$ *interval*

Figure 6: $Feed\ vapor\ fraction\ based\ subsets\ fitting$

	$F_{tol,top}$		$F_{biph,top}$		$F_{tol,bot}$		$F_{biph,bot}$	
α **range**	**MAE%**	n_{var}	**MAE%**	n_{var}	**MAE%**	n_{var}	**MAE%**	n_{var}
0-0.05	61.27	3	66.80	1	0.83	8	18.22	7
0.05-0.15	10.86	8	29.47	12	2.54	10	8.44	8
0.15-0.25	36.16	6	6.61	10	5.58	5	33.48	9
0.25-0.35	22.32	7	2.72	12	7.04	4	27.49	7
0.35-0.45	10.26	11	0.65	11	5.04	18	15.95	11
0.45-0.55	4.62	20	0.11	20	3.34	17	5.77	21
0.55-0.65	3.37	19	0.08	21	4.01	21	4.75	22
0.65-0.75	1.80	20	0.04	19	3.74	20	3.30	20
0.75-0.85	0.00	23	0.00	23	0.00	23	0.00	23
0.85-1	0.00	22	0.00	22	0.00	22	0.00	22

Table 4: MAE and number of variables vs. α

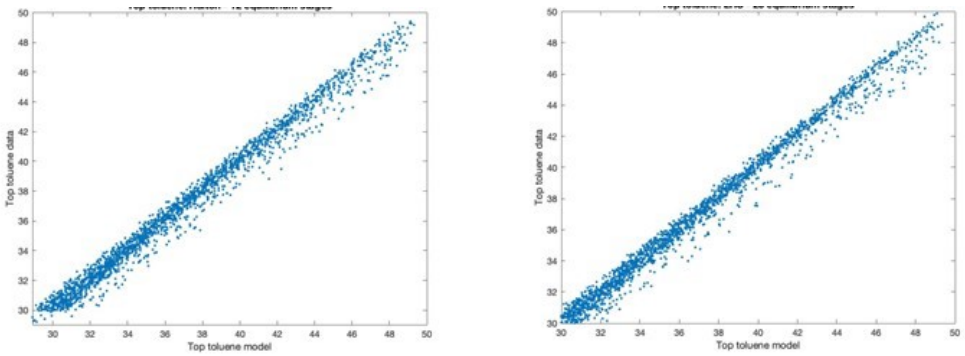
and vaporization rate. Therefore, it can be concluded that, in case of flash separation, since the liquid-vapor equilibrium is the aspect of main concern, feed vaporization fraction represents the key variable for sample subsets partition.

4.2. Distillation column

The same procedure was then performed with the 12 stages distillation column with seven input variables. In this part of the work the difference between Halton and LHS is one of the main focus along with the economic assessment. As usual, a first analysis was performed on the global dataset domain. The obtained equations for the distillate toluene are:

$$\begin{aligned} F_{tol,top}^{Halton} &= 0.94 \cdot D_L - 0.16 \cdot D_R - 1.45 \cdot F_P/3 - 0.42 \cdot F_T/232.55 \\ &+ 0.69 \cdot e^{F_P/3} + 2.05 \cdot sin(D_R/9) - 1.80 \cdot cos(F_{tol}/20) \\ &+ 1.33 \cdot cos(F_{biph}/4.46) \end{aligned} \tag{4}$$

$$\begin{aligned} F_{tol,top}^{LHS} &= 0.93 \cdot D_L - 0.069 \cdot D_R - 0.035 \cdot F_T + 0.33 \cdot ln(D_R) \\ &+ 3.04 \cdot ln(F_{tol}) + 0.097 \cdot ln(F_P) + 11.72 \cdot ln(F_T) \\ &+ 10^{-4} \cdot e^{D_R} - 1.4 \cdot 10^{-3} \cdot e^{F_{biph}} - 65.30 \end{aligned} \tag{5}$$



(a) $F_{tol,top}$ *Halton* (b) $F_{tol,top}$ *LHS*

Figure 7: *Fitting comparison between Halton and LHS sampling for distillate toluene partial flowrate*

In this case, fitting indicators are already quite accurate. The MAE are 6.6% and 6.8% respectively and the mean average errors are about 1.25% and 1.19%. The corresponding fitting plots are provided in Figure 7.

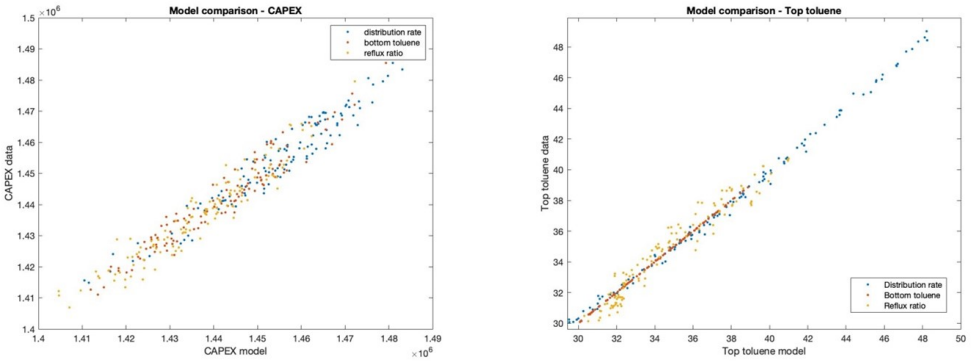
Therefore, the samples were ordered and distributed according to all possible parameters. The most promising results, listed in Table 5 for MAE and number of variables, have been obtained for the variables reflux ratio, bottom toluene flowrate and distribution rate defined as:

$$D_{tol/biph} = \frac{y_{tol}/x_{tol}}{y_{biph}/x_{biph}} \tag{6}$$

	Reflux ratio		$D_{tol/biph}$		$F_{tol,bot}$	
Output variable	**MAE%**	n_{var}	**MAE%**	n_{var}	**MAE%**	n_{var}
CAPEX	0.76	13	0.55	10	0.53	11
OPEX	18.25	7	9.13	2	22.12	13
$F_{tol,top}$	4.37	11	2.49	15	0.19	11
$F_{biph,top}$	10.93	1	32.29	2	35.39	6
$F_{tol,bot}$	12.90	6	12.71	1	0.62	11
$F_{biph,bot}$	9.64	11	3.57	5	1.62	10

Table 5: MAE and number of variables vs. α

In this case, the average error is lower than one percent for each variable and, except for some output parameters, the MAE as well is already quite



(a) $CAPEX$ (b) $F_{tol,top}$

Figure 8: *Comparison according to reflux ratio, bottom toluene flowrate and reflux ratio*

low. The best performing variable to order the data set was the reflux ratio. The comparison in terms of model fitting is shown in Figure 8 for CAPEX and $F_{tol,top}$.

Finally, the analysis is performed with the two sampling strategies by applying feasibility constraints on the subsets. The results for this part of the study are provided in Table for MAE, mean average error (mae) and number of variables.

	LHS			**Halton**		
Output variable	**MAE%**	**mae%**	n_{var}	**MAE%**	**mae%**	n_{var}
CAPEX	0.51	0.18	11	0.44	0.15	15
OPEX	2.16	3.18	8	2.66	5.08	14
$F_{tol,top}$	0.29	0.05	16	0.16	0.03	13
$F_{biph,top}$	9.99	4.99	4	14.71	6.59	3
$F_{tol,bot}$	0.89	0.16	16	0.45	0.08	14
$F_{biph,bot}$	9.98	4.99	3	14.71	6.59	3

Table 6: MAE, mae and number of variables for constrained Halton and LHS samplings

The first remark concerns, as usual, the lower quality of those parameters whose function is given by a lower number of variables. The best results are obtained for CAPEX, $F_{tol,top}$ and $F_{tol,bot}$ with an average error lower than 0.2%. On the other hand, when comparing the two sampling strategies, it

can be noticed that LHS performs slightly better for those variables who show worse quality indicators while both sampling approaches results almost in the same results quality.

In general, in all these studies concerning distillation column, it can be observed that the bottom biphenyl flowrate $F_{biph,bot}$ is the most difficult variable to regress with accuracy. If we take into account the flash unit as well, biphenyl shows in general more difficulties with respect to toluene in terms of behaviour prediction. This aspect could be related to the thermodynamic behaviour of the specific component in the bicomponent mixture. An analogous remark could be made for opertaing expenses. Since they are more affected by operating conditions than CAPEX, they are less simple to be regressed with an analytical fuction.

5. Conclusions

The proposed study proved to be effective to validate its purpose. In fact, the outcome of this research work allows to draw three main conclusions of interest as follows.

First, surrogate modeling for chemical processes performed by means of well established software packages showed good accuracy with respect to the phenomenological model with a considerable gain in terms of computational time.

Second, including feasibility constraints during sample points generation allows to have better models with the same modeling tools and approaches by considerably reducing the average and the maximum error with no increase in computational time.

Third, performing the modeling process with a local approach based on sub-domains permits to further improve the obtained accuracy with a non-substantial increase of the computational effort due to the fact that a set of more but simpler equations needs to be solved.

In the light of these outcome then, it could be stated that further studies concerning the inclusion of constraints and the domain subsets partition during the DoE step for more complex systems both in terms of units configuration and physical behaviour are worth to be performed in order to have a more reliable quantification of their impact on computational time and accuracy. This kind of studies could enable a more conscious decision about whether they are worth to be applied or not in terms of optimal costs-benefits compromise.

References

[1] K. McBride, K. Sundmacher, Overview of surrogate modeling in chemical process engineering, Chemie Ingenieur Technik 91 (2019).

[2] A. R. Conn, K. Scheinberg, L. N. Vicente, Introduction to Derivative-Free Optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009.

[3] F. N. Osuolale, J. Zhang, Energy efficiency optimisation for distillation column using artificial neural network models, Energy 106 (2016) 562–578. doi:https://doi.org/10.1016/j.energy.2016.03.051.

[4] K. Bi, B. Beykal, S. Avraamidou, I. Pappas, E. N. Pistikopoulos, T. Qiu, Integrated modeling of transfer learning and intelligent heuristic optimization for a steam cracking process, Industrial & Engineering Chemistry Research 59 (2020) 16357–16367. doi:10.1021/acs.iecr.0c02657.

[5] B. Beykal, F. Boukouvala, C. A. Floudas, E. N. Pistikopoulos, Optimal design of energy systems using constrained grey-box multi-objective optimization, Computers & Chemical Engineering 116 (2018) 488–502. doi:https://doi.org/10.1016/j.compchemeng.2018.02.017.

[6] F. Boukouvala, M. G. Ierapetritou, Feasibility analysis of black-box processes using an adaptive sampling kriging-based method, Computers & Chemical Engineering 36 (2012) 358–368. doi:https://doi.org/10.1016/j.compchemeng.2011.06.005.

[7] A. Bhosekar, M. Ierapetritou, Advances in surrogate based modeling, feasibility analysis, and optimization: A review, Computers & Chemical Engineering 108 (2018) 250–267. doi:https://doi.org/10.1016/j.compchemeng.2017.09.017.

[8] A. Di Pretoro, B. Bruns, S. Negny, M. Grünewald, J. Riese, Demand response scheduling using derivative-based dynamic surrogate models, Computers & Chemical Engineering 160 (2022) 107711. doi:https://doi.org/10.1016/j.compchemeng.2022.107711.

[9] A. Di Pretoro, A. Tomaselli, F. Manenti, L. Montastruc, Dynamic surrogate modeling for continuous processes control applications, in: L. Montastruc, S. Negny (Eds.), 32nd European Symposium on Computer

Aided Process Engineering, volume 51 of *Computer Aided Chemical Engineering*, Elsevier, 2022, pp. 91–96. doi:https://doi.org/10.1016/B978-0-323-95879-0.50016-3.

[10] L. S. Dias, M. G. Ierapetritou, Integration of planning, scheduling and control problems using data-driven feasibility analysis and surrogate models, Computers & Chemical Engineering 134 (2020) 106714. URL: https://www.sciencedirect.com/science/article/pii/S0098135419306982. doi:https://doi.org/10.1016/j.compchemeng.2019.106714.

[11] G. E. P. Box, K. B. Wilson, On the experimental attainment of optimum conditions, Journal of the Royal Statistical Society. Series B (Methodological) 13 (1951) 1–45.

[12] D. G. Krige, A statistical approach to some basic mine valuation problems on the witwatersrand, Journal of the South African Institute of Mining and Metallurgy 52 (1951).

[13] I. Mujtaba, N. Aziz, M. Hussain, Neural network based modelling and control in batch reactor, Chemical Engineering Research and Design 84 (2006) 635–644. doi:https://doi.org/10.1205/cherd.05096.

[14] H. Fang, M. F. Horstemeyer, Global response approximation with radial basis functions, Engineering Optimization 38 (2006) 407–424. doi:10.1080/03052150500422294.

[15] A. J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and Computing 14 (2004) 199 – 222. doi:https://doi.org/10.1023/B:STCO.0000035301.49549.88.

[16] S. E. Davis, S. Cremaschi, M. R. Eden, Efficient surrogate model development: Impact of sample size and underlying model dimensions, in: M. R. Eden, M. G. Ierapetritou, G. P. Towler (Eds.), 13th International Symposium on Process Systems Engineering (PSE 2018), volume 44 of *Computer Aided Chemical Engineering*, Elsevier, 2018, pp. 979–984. doi:https://doi.org/10.1016/B978-0-444-64241-7.50158-0.

[17] S. H. Kim, F. Boukouvala, Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques, Optimization Letters 14 (2020) 989 – 1010. doi:https://doi.org/10.1007/s11590-019-01428-7.

[18] C. Ding, M. Ierapetritou, A novel framework of surrogate-based feasibility analysis for establishing design space of twin-column continuous chromatography, International Journal of Pharmaceutics 609 (2021) 121161. doi:https://doi.org/10.1016/j.ijpharm.2021.121161.

[19] B. Beykal, M. Onel, O. Onel, E. N. Pistikopoulos, A data-driven optimization algorithm for differential algebraic equations with numerical infeasibilities, AIChE journal 66 (2020) 16657. doi:https://doi.org/10.1002/aic.16657.

[20] F. Boukouvala, C. A. Floudas, Argonaut: Algorithms for global optimization of constrained grey-box computational problems, Optimization Letters 11 (2017) 895 – 913. URL: https://doi.org/10.1007/s11590-016-1028-2. doi:10.1007/s11590-016-1028-2.

[21] F. Boukouvala, M. M. Hasan, C. A. Floudas, Global optimization of general constrained grey-box models: New method and its application to constrained pdes for pressure swing adsorption, J. of Global Optimization 67 (2017) 3–42. URL: https://doi.org/10.1007/s10898-015-0376-2. doi:10.1007/s10898-015-0376-2.

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

[23] J. H. Halton, Algorithm 247: Radical-inverse quasi-random point sequence, Commun. ACM 7 (1964) 701–702. URL: https://doi.org/10.1145/355588.365104. doi:10.1145/355588.365104.

[24] C. Wang, Q. Duan, W. Gong, A. Ye, Z. Di, C. Miao, An evaluation of adaptive surrogate modeling based optimization with two benchmark problems, Environmental Modelling & Software 60 (2014) 167–179. doi:https://doi.org/10.1016/j.envsoft.2014.05.026.

[25] A. Cozad, N. V. Sahinidis, D. C. Miller, Learning surrogate models for simulation-based optimization, AIChE Journal 60 (2014) 2211–2227. doi:https://doi.org/10.1002/aic.14418.

[26] H. Nyquist, Certain topics in telegraph transmission theory, Transactions of the American Institute of Electrical Engineers 47 (1928) 617–644. doi:10.1109/T-AIEE.1928.5055024.

[27] D. M. Boore, C. A. Goulet, The effect of sampling rate and anti-aliasing filters on high-frequency response spectra, Bulletin of Earthquake Engineering 12 (2014) 203–216. doi:DO - 10.1007/s10518-013-9574-9.