Comparative Study of Surrogate Modelling Techniques Applied to Three Different Chemical Processes

Ramón Mur, Ismael Díaz, Manuel Rodríguez
Departamento de Ingeniería Industrial y del Medio Ambiente, ETSI Industriales, Universidad Politécnica de Madrid, Calle de José Gutiérrez Abascal, 2, 28006 Madrid
r.mur@alumnos.upm.es

Abstract
In this paper, a comparative study of surrogate modelling techniques applied to chemical processes of different complexity is presented. The surrogate modelling techniques considered in this work are support vector regressions (SVR), Kriging and artificial neural networks (ANN). The surrogates were obtained by fitting to process data obtained from rigorous flowsheeting simulations previously developed in Aspen Plus v10. The processing schemes to be surrogated were (in order of decreasing complexity): 1) the separation of aromatics-aliphatics mixtures by liquid-liquid extraction using ionic liquids as novel and more sustainable solvents, 2) the toluene hydrodealkylation process and 3) a simple distillation of organic solvents. Besides, in the first process (aromatics-aliphatics separation), advanced predictive thermodynamic models based on quantum chemical calculations (COSMO-SAC) were considered, allowing for the prediction of fluid phase equilibria properties of mixtures containing ionic liquids. In addition, the robustness of the surrogate techniques was assessed by adding a random noise contribution to the variables sampled. Thus, the paper is organized as follows: in section 1 an introduction to the surrogate modelling techniques is presented along with a short description of the chemical processes modelled in Aspen Plus. Afterwards, section 2 includes the methodology detailing the computational approach presented. The main results obtained in the study are presented in section 3 and the final the conclusions summarized in section 4.

Keywords: Surrogate Modelling, Process Modelling

1. Introduction
1.1. Surrogate modeling techniques
There are many advanced surrogate modelling libraries available for the scientific/engineering community, ready to be used in the fields of engineering for optimization, parametric analysis, etc. Three of the most widely employed are support vector regression (SVR), Kriging interpolation and artificial neural networks (ANN); which have been considered in this work to test their performance for the modelling of chemical processes.

1.1.1. Support vector regression
SVR surrogates are represented as the weighted sum of basis functions added to a constant term. A general form of SVR surrogate is given in Eq. (1).
\[ \hat{f}(X) = \mu + \sum_{i=1}^{n} w^i \psi(X,X^i) \]  

(1)

Assuming a simple basis function \( \psi = X \), the surrogate can be written as per Eq. (2).

\[ \hat{f}(X) = \mu + w^T X \]  

(2)

The unknown parameters \( \mu \) and \( w \) in the model are obtained by solving an optimization problem (Cortes et al (1995)).

1.1.2. Kriging

Kriging is based on the idea that a surrogate can be represented as a realization of a stochastic process. Kriging is a two-step process. First, the regression function \( f(x) \) is constructed on the basis of Generalized Least Squares (GLS) regression. Then, a stochastic Gaussian process that represents the uncertainty about the mean of \( Y(x) \) with expected value zero is constructed based on the residuals \( (Z(x)) \).

\[ Y(x) = f(x) + Z(x) \]  

(3)

Where \( f(x) \) is a regression function and \( Z(x) \) is a Gaussian process with mean 0, variance \( \theta^2 \) and a correlation matrix \( \Psi \). Depending on the form of the regression function \( f(x) \), Kriging has different prefixes. Simple Kriging assumes the regression function to be a known constant. On the other hand, Ordinary Kriging assumes an unknown constant regression function \( (f(x) = a) \). In general, Universal Kriging assumes a multivariate polynomial regression function:

\[ f(x) = \sum_{i=1}^{p} \alpha_i b_i(x) \]  

(4)

Given \( p \) the number of candidate functions, \( b_i(x) \) the base functions and \( \alpha_i \) the coefficients determined by regression. Further details can be found at (Krige et al (1960)).

1.2. Artificial Neural Networks (ANN)

The structure of ANN typically consists of three distinct layers: the input layer, the hidden layer(s), and the output layer. The connections of neurons across layers represent the transmission of information between neurons. A typical ANN with three layers and one single output neuron has the following form:

\[ \hat{y} = \hat{f}(X) = \sum_{j=1}^{J} w_j \left( \sum_{i=1}^{k} v_{ij} f(x_i) + \alpha_j \right) + \beta \]  

(5)

Where \( X \) is an \( k \)-dimensional vector with \{\( x_1, x_2, \ldots, x_k \)\} as its elements, \( f \) is the transfer function, \( v_{ij} \) is the weight on the connection between the \( i^{th} \) input neuron and \( j^{th} \) hidden neuron, \( \alpha_j \) is the bias in \( j^{th} \) hidden neuron, \( w_j \) is the weight on connection between \( j^{th} \) hidden neuron and the output neuron, \( J \) is the total number of hidden neurons, and \( \beta \) is the bias of the output neuron. The weights and biases can be determined by the training procedure minimizing the training error (McCulloch et al (1943)).
1.3. Benchmark chemical processes

1.3.1. Aromatic-Aliphatic separation from low aromatic content naphta using ionic liquids

The liquid-liquid extraction process to separate aromatic components from a naphta stream using ionic liquids as solvents is studied in the present work as shown in Figure 1. For this purpose, an Aspen Plus flowsheet was developed based on the work of (de Riva et al (2016)). In this process, the naphtha entering the process (S-NAP-IN) contacts the extraction solvent stream (S7-IL) in a liquid-liquid extraction column (EXT). The raffinate (S-AL+N2) of this operation is the aliphatic product while the extract (S-AR+IL, rich in aromatic components) enters the intermediate stripper (STRIP) where the remaining aliphatic components are separated by an inert gas (N2) (S2) and exit the process in stream S-AL. After this gas stripping, the gas and the aliphatic hydrocarbons are separated in column SEP. The aromatic-rich stream (S1-IL) and the ionic liquid are separated in a vacuum distillation vessel (FLASH) after being conditioned (HEAT1). The aromatic components obtained in this vessel (S1-AR) are conditioned (HEAT2, PUMP1) before leaving the process as product (S3-AR). The regenerated ionic liquid (S3-IL) is recirculated as extraction solvent (S4-IL) to the extraction column (EXT-DIST), after being conditioned (HEAT1, HEAT3, PUMP2).

The following process variables have been sampled (input of the model):
- Theoretical number of stages of the liquid-liquid extraction column (EXT): 2-30.
- Molar solvent to feed ratio of the liquid-liquid extraction column (EXT): 1-4.
- Recovery temperature of the ionic liquid (FLASH): 130-230ºC.
- Nitrogen mass flow input into the stripper column (STRIP): 40-80 ton/h.

The final output variable to be regressed is the total mass fraction of aromatics (sum of all aromatics compound) in the product stream.

1.3.2. Toluene Hydrodealkylation

The purpose of this process is to produce benzene from toluene rich streams using H2 as reagent, see Figure 2. The toluene entering the process (S1-TO-IN) contacts the recovery toluene (S3-TOL) as well as the hydrogen entering the process (S-H2-IN) contacts the recovery hydrogen (S3-H2). These streams after being conditioned (HEAT1) enter the reactor (REACT) where the hydrodealkylation reaction occurs. Afterwards, methane and hydrogen are separated from the S2-TO-BE stream by flash separators (FLASH1...
and FLASH2) and then toluene and benzene are also separated using distillation columns (COL1 and COL2). Hydrogen and toluene are finally recycled to the process.

The following main chemical reaction has been assumed, and the kinetic equation and parameters adopted from (Dimian et al (2014)):

\[
H_2 + C_6H_5CH_3 \rightarrow C_6H_6 + CH_4
\]

The following process variables have been sampled (input of the model):
- Volume of the reactor (REACT): 0.8-200 \text{ m}^3.
- Temperature of the reactor (REACT): 550-850ºC.
- Pressure of the reactor (REACT): 3-240 bar.
- Fraction of recovered toluene after the first distillation column (COL1): 0.2-0.95.

The final output variable to be regressed is the mass fraction of benzene in the product.

### 1.3.3. Distillation of organic solvents

A solvent mixture of acetone, methanol and water enters the first distillation column (COL1) along with a water stream and the acetone is separated as depicted in figure 3. Then, the bottom stream (S-WA-MET) enters the second distillation column (COL2) where methanol and water are separated; the process scheme is shown in Fig 3.

The following process variables have been sampled (input of the model):
- Number of theoretical stages at the first column (COL1): 5-40.
- Number of theoretical stages at the second column (COL2): 5-40.
- Molar reflux ratio at the first column (COL1): 1-10.
- Molar reflux ratio at the second column (COL2): 1-10.

The final output variable to be regressed is the acetone mass fraction in the product.
2. Methodology
The methodology followed in this work can be split into three steps: the development of a communicating interface for sampling; the sampling and data preprocessing; and the fitting of the surrogate models. The first step is needed to call the variables values on the Aspen Plus models from MATLAB (COM interface), which is the main coding platform where the preprocessing and fitting steps are carried out. Both the input (before the execution of the process simulator) and the output variables (resulting from the Aspen Plus execution) were normalized to fit the [0,1] range. Data sampling was repeated 150 times using the LHS method adding a random noise value (low noise was considered to be $\eta \approx N(0,0.1)$ and large noise $\eta \approx N(0,0.2)$). The ranges of the input variables sampled are shortlisted for each process in sections 1.3.1 to 1.3.3. Then, data regression is carried out using different toolboxes available in MATLAB. Data fitting for Support Vector Regression and Artificial Neural Networks were performed using the built-in Statistics and Machine Learning and the Deep Learning toolboxes, whereas Kriging regressions were conducted using the ooDace toolbox developed by (Couckuyt et al (2014)) also available for MATLAB. The workstation used for the calculations was a computer with a 3.40GHz Intel Core i7-6700 CPU and 16 GB of RAM under Windows 10.

3. Results
In order to evaluate the performance of each surrogate modelling technique, the data sampling, noise addition (small or large) and regression process was repeated 25 times for each surrogate technique over each individual process in order to compute the average mean squared error (MSE) and the std deviation (the average and std deviation of the 25 surrogate models developed). Figures 4 to 6 show the average values and standard deviation for the three different processes studied in this work. The average and std deviation of the MSE values obtained have been normalized by dividing by the values obtained for SVR with large noise (deemed as the reference value).
Figure 4 shows that for the more complex process involving liquid-liquid extraction of aromatics-aliphatics mixtures with ionic liquids, the smaller MSE average value are obtained by ANN (SVR show similar performance), whereas the largest MSE standard deviation is also obtained by ANN (SVR show also similar values). In the other two processes (Figures 5 and 6, toluene hydrodealkylation and distillation of organic solvents), Kriging regression have shown to be the more accurate technique (lowest average deviations obtained) and the less sensitive to data changes (lowest standard deviation in the set of 25 regressions carried out). As it can be also seen in the previous figures, the imposed noise applied to the variables during the sampling step clearly influences the final deviation of the fitted models. It agrees with the fact that the more dispersed the sampled data to fit, more deviations are expected between the final surrogates fitted. However, the discussion presented herein is quite limited and is only valid in the restricted scope of this work. A broader study considering a larger pool of chemical processes, different variable nature (temperature, pressure, compositions, flows…) and different surrogate modelling parametrizations is under development at this moment which is expected to shed more light on the discussion presented herein.

4. Conclusions
In this work, a systematic study of surrogate modelling techniques (SVR, Kriging and ANN) applied to different chemical processes has been carried out. The performance of these techniques has been evaluated in the presence of imposed noise during variable sampling. Results show that for the more complex process model (aromatics-aliphatics separation), ANN and SVR have achieved the lowest MSE value. On the contrary, for the less complex processes (solvent distillation and toluene hydrodealkylation), Kriging regression was the most promising technique. However, a wider study is needed to obtain less case dependent results and deeper insights about the suitability of the different surrogate modelling techniques currently available.

Acknowledgements
The authors are grateful to the Com. de Madrid (project P2018/EMT-4348 SUSTEC) for financial support.

References