Surrogate modeling application for process system emissions assessment: improving computational performances for plantwide estimations

Giulio Carniob, Alessandro Di Pretoroa\*, Fabrizio Bezzob, Ludovic Montastruca

aLaboratoire de Génie Chimique, Université de Toulouse, CNRS/INP, Toulouse, France

bCAPE-Lab, Dipartimento di Principi e Impianti di Ingegneria Chimica, Università di Padova, via Marzolo 9, I-35131 Padova, Italy

alessandro.dipretoro@ensiacet.fr

Abstract

During the last decade, data driven modeling has gained a role of major interest all over the engineering fields mainly due to the need of higher computational power or, inversely, less computational demanding models for applications such as optimization, simulation, scheduling and control. Relevant contributions of process systems surrogate modeling as a support for operation optimization were already proved in literature with a reduction of the overall computational time by two orders of magnitude with respect to conventional simulations. In this research work a biogas-to-methanol plant case study is used to assess the total energy consumption and estimate the related emissions. Therefore, a modeling phase carried out via Response Surface Methodology is set up in order to obtain the analytical function that allows to estimate the equivalent CO2 emissions over an extended range of operating conditions representing a wide interval of biomass feed composition. The study has been performed over a wide independent variables domain as well as for different sample sizes in order to compare the computational performances and the accuracy of the obtained models accordingly. The computational time was reduced by two orders of magnitude with a mean relative error lower than 1%. Given the quality of the results, this approach could be further exploited for other system variables and processes including highly non-ideal behaviour of mixtures to be treated. Furthermore, more complex sampling and different surrogate modeling strategies could be tested in order to check if even higher computational effectiveness and model accuracy could be obtained in the process systems domain.

**Keywords**: Response Surface Methodology, sampling, uncertainty, computational efficiency, biogas to methanol.

* 1. Introduction

In recent years, the digitalization of the chemical industry on a massive scale has oriented the research efforts in process systems engineering domain towards more computational effective solutions. In this perspective, several modeling alternatives, applicable to other engineering fields as well, that were proposed during the last decades in order to ease the computational criticalities have seen a renewed interest as discussed in Alizadeh et al. (2020). In particular, when dealing with complex process systems, the conventional simulations based on phenomenological models result to be the most highly time demanding step within the optimization loops. A significant example regarding the gain in terms of effectiveness obtained when replacing simulations by data-driven models can be found in Di Pretoro et al. (2022). Their research proved indeed the capability of derivative-based surrogate models to reduce the required computational time for scheduling optimization by a 2,400 factor with 99.66% solution accuracy.

Along with the digital transition, sustainability can be defined as the second current target of major concern in the industrial domain. The most established indicator to assess the environmental impact of a chemical process is the Global Warming Potential whose unit of measure is the amount (kg) of equivalent CO2 emissions. Due to the uncertain nature of sustainable raw materials, a quick estimation of the expected emissions of the process under variable operating conditions could be a valuable decisional tool during the design and the scheduling phase of a process system. In particular, in a chemical process, the total energy consumption in terms of electricity and heat duties can be the identified as main emission item (Gadalla et al., 2006).

For this reason, a Response Surface Methodology approach, whose details are provided in the dedicated section, is exploited in this study to correlate the emissions of a biogas-to-methanol plant with the feed stream properties over a wide composition range. The obtained model was then tested to assess computational performance indicators such as computational time and model accuracy. The presentation of the case study, its process diagram as well as the operating parameters are provided in detail the next section followed by a thorough description of the modeling methodology.

* 1. Case study

The modeling strategy presented in this work was applied to a biogas-to-methanol case study. The choice of this system had two main purposes that are namely testing the performances of a modeling procedure on a plantwide scale, i.e. correlating the inlet and the outlet of a process involving several units in between, and assessing the accuracy of a model whose inlet parameters could vary over a wide range, since biogas properties depend on the biomass nature and the fermentation conditions. The process under analysis was modeled and simulated by means of Aspen HYSYS v12 process simulator and its simplified flow diagram is reported in Figure 1. Three main process sections can be identified: (*i*) steam reforming, (*ii*) methanol synthesis and (*iii*) product purification.



Figure 1 – Simplified Process Flow Diagram of the biogas-to-methanol plant

In the first section, the biogas fed to the process is preliminarily purified from the H2S, compressed and mixed with water. Then, the mixture is preheated thanks to the heat recovered from the reformer product stream. In the reforming furnace the biogas is then converted into bio-syngas by supplying heat to the system. The reactor is modeled as a fixed bed plug flow reactor with a given size and its products are cooled down by means of an air cooler and a series of heat exchangers in order to condense and separate the syngas from water. The latter is then recycled back mixed with a make up stream. In the second section, the syngas is recompressed to the reactor operating pressure and preheated as in the first one. In the reactor syngas is converted into methanol with a globally exothermic reaction scheme. This second reactor as well was modeled as a fixed bed plug flow reactor. After being purified from unconverted syngas, which is recycled back to the reactor, the methanol is laminated and sent to the third section consisting of a series of distillation steps. In the first column, light ends are removed from the methanol stream and the entrained part is washed with water in a second column in order to further enhance its recovery. Finally, the last column distills methanol from water to grade AA purity.

As previously introduced, the purpose of this study is to model the plantwide emissions for variable inlet feed composition. Therefore, Table 1 reports, along with the operating parameters and the process specifications required to fulfill the final product quality, the selected domain in terms of methane and carbon dioxide partial flowrates over which the Design of Experiment will be performed.

Table 1 - Operating parameters, specifications and modeling domain

|  |  |  |
| --- | --- | --- |
| **System parameters** | **Value** | **Unit** |
| Inlet methane flowrate | 10.95 ± 4.38 | kmol/h |
| Inlet carbon dioxide flowrate | 6.71 ± 2.68 | kmol/h |
| **Process specifications** |  |  |
| Water/methane ratio @ reformer   | 3.075 | mol/mol |
| Outlet methanol purity | 0.9985 | kg/kg |

The assessment of the environmental impact requires a deep understanding of the process. The selected indicator is the Global Warming Potential expressed as CO2 equivalent emissions. These emissions are primarily related to heating utilities and electricity usage while the impact due to the equipment production and dismission can be neglected as already proved by the study of Gadalla et al. (2006). To be more precise, the emissions related to electricity represents only the 7.9% of the entire process energy demand while the most energy-intensive units are the reforming furnace and the reboilers of the two distillation columns. As concerns the cooling duties, since they work at ambient temperature, cooling water can be selected as utility stream, and their contribution to the overall GWP can be thus considered negligible.

So, finally, the plantwide CO2 equivalent emissions of the biogas to methanol process were calculated according to the correlations provided in the same research article in order to generate a dataset of simulated points corresponding to different methane and carbon dioxide partial flowrate points. These values could then be employed in the following modeling step to build a surrogate model of the plant emissions according to the methodology whose details are discussed in the next section.

* 1. Methodology
		1. Data generation and sampling

All literature studies dealing with metamodels agree about the fact that data sampling is the most critical step of the entire modeling procedure (McBride & Sundmacher, 2019). The quality, the nature and the distribution of the employed dataset indeed strongly affects the accuracy of the obtained model. For this reason, the data generation should follow dedicated Design of Experiment (hereafter DoE) strategies. A preliminary DoE distinction can be made between stationary and sequential modes: in the former case the entire dataset is generated at once and used in a single modeling step while, in the latter, the DoE and the modeling steps are reiterated by updating the data sampling criteria according to the model results of the previous loop with a higher computational effort. In this research study, the once-through approach is used with the goal of having an optimal space-filling over the modeling domain. Several methods for the sampling generation exist such as Monte Carlo sampling, Quasi-Monte Carlo (QMC) Sobol’ design or again QMC Halton. For the biogas-to-methanol emissions modeling, the Latin Hypercube Sampling (LHS) (McKay et al., 1979) will be used since it provides good space filling and no particular criticalities in the resulting function are expected. An example of the space partition according to the LHS is provided in Figure 2 (left).

Finally, as concerns data acquisition, two main types of data are mainly exploited for surrogate modeling, namely experimental and generated ones. In this study, as already mentioned in the previous section, data were generated over the selected domain by means of phenomenological models implemented in the process simulation modules.

 

Figure 2 – left) Example of LHS sample distribution - right) Response Surface example

* + 1. Modeling

The modeling strategy exploited in this research study is the Response Surface Methodology introduced by Box and Wilson (1951). The principle behind this approach is the use of a Response Surface, usually polynomial, in a n-D space to describe the relationship between n-1 explanatory variables and one or more response variables. From a mathematical point of view, the objective function to minimize is the mismatch between the *F* function values corresponding to the explanatory variables $\tilde{x}$ and the dataset provided by the DoE step. On the other hand, the output variables of the optimization algorithm are the coefficients *Ki* and *ki* that respectively multiply each term and argument of the Response Surface according to the expression reported in Equation 1.

|  |  |
| --- | --- |
| $$F(\tilde{x})=\sum\_{i=1}^{n}K\_{i}⋅f\_{i}(\tilde{x},\tilde{k})$$ | (1) |

For a better understanding, a graphical example is provided in Figure 2 (right).

In this study, the optimization is performed on Matlab® by means of the dedicated Curve Fitting Toolbox® that allows to select among polynomial, logarithmic or statistical functions as basis functions of the RSM. As a result, the software provides the coefficients of each function term as well as their confidence interval. Moreover, the toolbox enables the comparison between the obtained surface and a validation dataset provided by the user in order to assess model quality parameters, e.g. the sum of squared errors, while the mean relative error is calculated separately and used as reference indicator for this study.

The detailed results for the biogas-to-methanol plant emissions over the selected domain are presented in the next section for different sample sizes.

* 1. Results

The procedure was tested for a variable sample size in order to assess the impact of the dataset dimension on the results quality. Figure 3 (left) shows the comparison between the obtained data-driven model and the simulation points for CO2 and methane partial flowrate variations in the feed. In particular, in order to allow the distinction between the Response Surface and the validation data (green surface), the plot reports the outcome that does not correspond to the highest obtained accuracy (i.e. only 10 samples). In fact, according to the curve fitter optimizer, the best option for the approximation of the biogas-to-methanol plant emissions for any sample with 50 points or more resulted to be a 5th grade polynomial function expressed as:

|  |  |
| --- | --- |
| $$RS\left(X\_{1},X\_{2}\right)=k\_{1}+k\_{2}⋅X\_{1}+k\_{3}⋅X\_{2}+k\_{4}⋅X\_{1}⋅X\_{2}+k\_{5}⋅X\_{1}^{2} + k\_{6}⋅X\_{2}^{2}+k\_{7}⋅X\_{1}^{3}+k\_{8}⋅X\_{2}^{3}+k\_{9}⋅X\_{1}⋅X\_{2}^{2}+k\_{10}⋅X\_{2}⋅X\_{1}^{2}+k\_{11}⋅X\_{1}^{4}+k\_{12}⋅X\_{1}^{3}⋅X\_{2}+k\_{13}⋅X\_{1}^{2}⋅X\_{2}^{2}+k\_{14}⋅X\_{1}⋅X\_{2}^{3}+k\_{15}⋅X\_{2}^{4}+k\_{16}⋅X\_{1}^{5}+k\_{17}⋅X\_{1}^{4}⋅X\_{2}+k\_{18}⋅X\_{1}^{3}⋅X\_{2}^{2}+k\_{19}⋅X\_{1}^{2}⋅X\_{2}^{3}+k\_{20}⋅X\_{1}⋅X\_{2}^{4}+k\_{21}⋅X\_{2}^{5}$$ | (2) |

where *X1* and *X2* are respectively the methane and carbon dioxide flowrates and *ki* are the model parameters provided by the optimizer. It is worth remarking that this model is just the result of an optimal data correlation and the obtained parameters have no particular meaning from a physical point of view.

  

Figure 3 – left) RS vs. validation dataset - right) Computational time and mean relative error

As concerns the dataset size related performances, results are provided in Figure 3 (right). As expected, the overall computational time, given by the sum of the DoE and the function generation over the entire domain, increases with the sample size as well as the model accuracy. However, beyond 40 experimental values, the mean relative error decrease is almost negligible while the computational time keeps increasing with an almost linear trend. Therefore, beside providing a polynomial function able to replace the plantwide simulation for process environmental impact prediction, this methodology allows also to define the order of magnitude of the sample size suitable for the model generation. In practice, in case one of the operating parameters should vary, e.g. feed inlet temperature or reactor conversion, without a considerable change of the process thermodynamics or components, a new model could be easily obtained by using a dataset of about 50 samples.

* 1. Conclusions

The proposed methodology, based on RSM for surrogate modeling generation, provided both positive and reliable outcome for the specific case study and promising results of general validity. In particular, the research showed that building a plantwide surrogate model that allows to correlate input variables with the overall process energy consumption by keeping good accuracy is a viable solution. In terms of computational effort, the obtained data-driven model allows to perform reliable process emissions estimations over a wide range of operating conditions with a computational time reduced by orders of magnitude with respect to the conventional process simulation approach. Moreover, an optimal sample size in terms of the best compromise between accuracy and computational effort can be easily identified.

In terms of future perspectives, the obtained data-driven model could be used for optimal demand response scheduling in order to test the performances in terms of optimal solution accuracy as well as to reduce the computational time when optimal process operating conditions need to be identified during the design phase. Furthermore, different sampling strategies could be compared to detect the most suitable one for the specific system under analysis.

References

R. Alizadeh, J.K. Allen, F. Mistree, 2020. Managing computational complexity using surrogate models: a critical review. Research in Engineering Design 31, 275–298.

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

A. Di Pretoro, B. Bruns, S. Negny, M. Grünewald, J. Riese, 2022. Demand Response Scheduling Using Derivative-Based Dynamic Surrogate Models. Computers & Chemical Engineering 160, 107711.

M. Gadalla, Ž. Olujić, M. Jobson, R. Smith, 2006. Estimation and reduction of CO2 emissions from crude oil distillation units, Energy 31(13), 2398-2408.

M. Fedeli, F., Manenti, 2022. Assessing process effectiveness with specific environmental and economic impact of heat, power & chemicals (HPC) option as future perspective in biogas. Cleaner Chemical Engineering 2, 100016.

K. McBride, K. Sundmacher, 2019. Overview of Surrogate Modeling in Chemical Process Engineering. Chemie Ingenieur Technik 91, 228–239.

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