Flexibility Analysis Using Surrogate Models Generated via Symbolic Regression

Tim Forster*a*, Daniel Vázquez*b*, Isabela Fons Moreno-Palancas*a*, Gonzalo Guillén-Gosálbez*a,\**

a Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, Vladimir Prelog Weg 1, 8093 Zurich, Switzerland

b IQS School of Engineering, Universitat Ramon Llull, Via Augusta 390, 08017 Barcelona, Spain

\* Corresponding author: gonzalo.guillen.gosalbez@chem.ethz.ch

Abstract

Computing the flexibility index to quantify the extent to which disturbances and uncertainties can affect a given process can be very challenging, especially if constraints are hard to describe algebraically or if they are unavailable as closed-form expressions. Here, we tackle the challenge of solving a flexibility index problem in the presence of such constraints by using symbolic regression. In essence, we replace those constraints in the flexibility index problem by an algebraic surrogate built using symbolic regression. This facilitates the solution process of the flexibility index problem by allowing the user to apply off-the-shelf deterministic solvers. We showcase the capabilities of our approach in a case study, discussing the pros and cons of the suggested approach relative to other existing approaches.

**Keywords**: Uncertainty, Flexibility Index, Symbolic Regression, Surrogate Modelling.

* 1. Introduction

To continuously satisfy safety, operational or cost constraints, a production process needs to be operable even in the presence of slight disturbances or uncertainties during the operational phase (Grossmann et al., 1983). In a pioneering work, Swaney and Grossmann (1985a, 1985b) introduced a method to compute an index that assesses the flexibility of a given process design under operation, i.e., the ability to continue operation under uncertainty. Due to the deterministic form of the flexibility index problem, it can only be computed if constraints are available as closed-form mathematical expressions. This makes the flexibility index computation challenging if very complex constraints are involved in the model or if the constraints are not available analytically (Floudas et al., 2001). To overcome this challenge, several approaches have been described to assess the feasible space and the flexibility of a model for cases where only input-output data is observed. Some of these works use surrogate models, such as Kriging (Boukouvala and Ierapetritou, 2012), neural networks (Metta et al., 2021), or high-dimensional model representations (Boukouvala et al., 2010). One limitation of such modelling approaches is that the user must assume an *a priori* model structure. Another approach was described by Sachio et al. (2023), where the authors created an available Python package to obtain representations of the design space boundary. Further, Zhao et al. (2021) used derivative free optimization to calculate the flexibility index. Alternatively, analytical equations could be derived using symbolic regression to replace only some of the constraints, which would allow using the originally described flexibility index problem (Grossmann et al., 1983) and state-of-the-art solvers. Building on this idea, our proposed approach relies on the application of a symbolic regression algorithm that finds analytical equations that can reproduce given data precisely. The identified models can subsequently be incorporated in the flexibility problem, thereby simplifying the calculations, and enabling the user to apply off-the-shelf deterministic solvers. This approach allows to decouple the surrogate model training from solving the flexibility index problem. Below, we illustrate the capabilities of our approach applied to a continuous stirred tank reactor, where we discuss the model building time, prediction accuracy of the model, and the solution approach for the resulting hybrid flexibility formulation.

* 1. Problem Statement

Some process parameters might be affected by uncertainties. Additionally, some process variables could be adjusted to counteract these uncertainties. In what follows, we shall consider a given number of process constraints . To result in a feasible operation, each of those constraints must be fulfilled, meaning that the inequalities should hold for all constraints. Usually, a nominal operating point is defined for a process, without considering variations in the uncertain parameters. At this point, the question is how far the values can deviate from such a point such that the process remains feasible. A first step to assess the process flexibility, according to Grossmann et al. (1983), could be the evaluation of the worst constraint violation , which can be minimized by adjusting the control variable. This leads to the so-called feasibility function of the process, defined in Eq. (1):

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

For the case when , the process remains feasible for the given realization of , and infeasible otherwise. A modeler might face difficulties to formulate closed-form constraints when they are very complex to be described, or even inaccessible in an algebraic form (i.e., if the constraints are expressed as a system of ordinary or partial differential equations). In this case, the solution of the system in Eq. (1) is not straightforward. To overcome this challenge, we split the set of constraints into the two subsets and . The first subset, , contains the constraints  that are already available as closed-form expressions. On the other hand, subset contains the constraints , which are unavailable as closed-form constraints or very difficult to describe. Subsequently, the constraints in are replaced with closed-form surrogate models that are constructed via symbolic regression.

* 1. Methodology
		1. Fundamentals of the Flexibility Index Problem

For the sake of brevity, we do not reproduce here the entire fundamentals of feasibility and flexibility theory. Instead, we refer readers to the original works by Grossmann et al. (1983), Halemane and Grossmann (1983), and Swaney and Grossmann (1985a, 1985b). A short summary of the derivation of the flexibility index problem is given below for the sake of completeness and readability. Originally, Grossmann et al. (1983) proposed an approach to quantify the largest possible uncertainty set , such that the process remains feasible over the entire range of . For this, the variable 𝛿 was introduced, which can be regarded as a scaled deviation from a nominal point , such that any realization of around that nominal point results in a feasible solution (i.e., the feasibility function remains non-positive). For this work, we assume a rectangular form for given by , while other options are documented and compared in literature (Pulsipher et al., 2019). In this rectangular form, the parameters and represent the minimum lower and maximum upper deviations (Grossmann et al., 2014). This deviation is optimized to maximize the area around the nominal operating point , represented by . This results in a bi-level optimization problem, given in Eq. (2), which determines the flexibility index .

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

After a reformulation of the problem given by Swaney and Grossmann (1985a, 1985b), the application of the Karush-Kuhn-Tucker conditions, and the introduction of an active set strategy by Grossmann and Floudas (1987), the flexibility index problem can be posed as a single-level optimization problem, given in Eq. (3). A summary of these reformulations is comprehensively covered by Pulsipher et al. (2019).

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

In this formulation, , , , and refer, respectively, to the maximum constraint violation, the slack variables, the Lagrange multipliers, and binary variables of constraint . Furthermore, represents the dimensionality of adjustable process variables, and a large-enough parameter that acts as upper bound for the slack variables .

* + 1. Replacement of Constraints by Surrogate Models

It might be challenging to formulate closed-form constraints to solve the problem given in Eq. (3). In such a case, we define the two sets of constraints, and , respectively, as described in section 2. Subsequently, we replace the constraints  (or parts of them) by algebraic surrogates built with a symbolic regression algorithm, where the surrogate functions replacing the constraints will be denoted by . The flexibility index problem in Eq. (3) is therefore reformulated into the hybrid expression in Eq. (4), which combines the backbone of the original flexibility index problem with a data-driven model component for the replaced constraints. As models for , we build algebraic expressions using a symbolic regression algorithm that identifies closed-form equations that map specific inputs to observed target variables. This leads to the following model with analytical surrogates embedded.

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

* 1. Case Study
		1. Software Implementation

The calculations discussed in this work were carried out on an AMD Ryzen 7 Pro CPU and 32 GB of RAM. We used Python v3.10.13 with NumPy v1.26.1, SciPy v1.11.3, and pyDOE v0.3.8 to construct the sampling dataset. The symbolic regression algorithm provided by Guimerà et al. (2020), the Bayesian Machine Scientist (BMS), was used to train the surrogate models. The resulting flexibility index problem was modeled in Pyomo (Hart et al., 2011) v6.6.2 and solved using BARON (Sahinidis, 1996) v22.9.30.

* + 1. Continuous Stirred Tank Reactor

We apply the hybrid flexibility approach discussed above to a continuous stirred tank reactor (CSTR). In the reactor, a first order reaction takes place, where a raw material A reacts irreversibly to form a product B. The reaction rate () is calculated through an Arrhenius expression, from the temperature (), activation energy (), universal gas constant (), and a pre-exponential factor (), i.e., , where () is the concentration of the raw material in the reactor. The temperature and volume in the vessel are constant and set to , and , respectively. The feed concentration of material A is denoted by (). The feed rate () represents the control variable and is bounded between . Additionally, the inlet concentration and the pre-exponential factor are regarded as uncertain parameters within the bounds and , respectively. The nominal operating point is . The goal of the flexibility analysis is to quantify the largest possible rectangular set , such that the process is still feasible. Besides imposing bounds on the variables, we seek to keep the concentration of B at the outlet of the reactor above a given minimum required concentration . The mass balances are given in Eq. (5).

|  |  |  |
| --- | --- | --- |
|  | and  | (5) |

To identify the surrogate model – which should map the feed rate, the inlet concentration of A and the pre-exponential factor to the concentration of product B at the outlet of the reactor – the ODE system given in Eq. (5) is solved for different input vectors with samples, each with of A and no B as starting conditions in the reactor. A Latin Hypercube Sampling approach was used to generate the samples, of which 80% were used for model training and 20% for model testing. After simulating the reactor for each sample , the outlet concentration of B was obtained. To train the BMS model, a variety of unary () and binary () operators were allowed to be selected by the algorithm. 1000 Markov chain Monte Carlo steps were used for model training. The model was allowed to consider up to eight differentiated parameters.

* 1. Results and Discussion

First, the model building results are discussed, followed by an assessment of the flexibility index calculations. Figure 1 (a) shows the training and testing performance of the identified algebraic model, while Figure 1 (b) summarizes the results of the flexibility index problem.



Figure 1 (a)Observed values are plotted against the model predictions. The circles represent the training data, whereas the diamonds correspond the test data. (b) Graphical representation of the solution for the flexibility index problem. The bright shaded area represents the feasible region. The dashed lines represent the bounds of the uncertain parameters . The bold solid nonlinear line represents the surrogate constraint which can be influenced by the control variable 𝑧. The considered nominal operating point (triangle) is located in the set 𝑇(𝛿) (dark-shaded solid-lined box). The critical realization of (pentagon) is visualized where 𝑇(𝛿) touches the closest constraint, which happens to be the surrogate constraint.

For training the BMS model, 2 min were required. The resulting training and testing root mean squared errors were , and , respectively, and the coefficients of determination were for training and testing, respectively. The model could therefore map the inputs remarkably well to the outputs.

The flexibility index problem could be solved with the identified algebraic model incorporated. Less than one second was required to reach global optimality (zero relative optimality gap). In the optimal solution, the control variable was at its bound of . The solution provides the maximum allowed deviation of the uncertain parameters such that the process remains feasible. The flexible region is therefore spanned by the parameter intervals and , corresponding to a flexibility index of .

* 1. Conclusion

We introduced an approach to compute the flexibility index of a process in cases where constraints are present that are very complex to be described, or even inaccessible in an algebraic form. The original deterministic flexibility index problem is combined with a surrogate model built by symbolic regression. An advantage of the symbolic regression algorithm used, the Bayesian Machine Scientist, is that no aprioristic model structure for the approximated constraints is required. Further, the surrogate model training can be decoupled from the flexibility index problem, which reduces the complexity of the entire solution procedure. The approach was applied to a continuous stirred tank reactor example implementing a chemical reaction. A model that precisely maps the uncertain parameters and control variables to a desired output could be identified, where the closed-form expression then allowed the use of a global solver to solve the resulting hybrid flexibility index problem.

References

F. Boukouvala, M.G. Ierapetritou, 2012, Feasibility analysis of black-box processes using an adaptive sampling Kriging-based method, Computers & Chemical Engineering, 36, 358–368.

F. Boukouvala, F.J. Muzzio, M.G. Ierapetritou, 2010, Design Space of Pharmaceutical Processes Using Data-Driven-Based Methods, Journal of Pharmaceutical Innovation, 5, 119–137.

C.A. Floudas, Z.H. Gümüş, Marianthi G. Ierapetritou, 2001, Global Optimization in Design under Uncertainty: Feasibility Test and Flexibility Index Problems, Industrial & Engineering Chemistry Research, 40, 4267–4282.

I.E. Grossmann, B.A. Calfa, P. Garcia-Herreros, 2014, Evolution of concepts and models for quantifying resiliency and flexibility of chemical processes, Computers & Chemical Engineering, Manfred Morari Special Issue, 70, 22–34.

I.E. Grossmann, C.A. Floudas, 1987, Active constraint strategy for flexibility analysis in chemical processes, Computers & Chemical Engineering, 11, 675–693.

I.E. Grossmann, K.P. Halemane, R.E. Swaney, 1983, Optimization strategies for flexible chemical processes, Computers & Chemical Engineering, 7, 439–462.

R. Guimerà, I. Reichardt, A. Aguilar-Mogas, F.A. Massucci, M. Miranda, J. Pallarès, M. Sales-Pardo, 2020, A Bayesian machine scientist to aid in the solution of challenging scientific problems, Science Advances, 6.

K.P. Halemane, I.E. Grossmann, 1983, Optimal process design under uncertainty, AIChE Journal, 29, 425–433.

W.E. Hart, J.-P. Watson, D.L. Woodruff, 2011, Pyomo: modeling and solving mathematical programs in Python, Mathematical Programming Computation, 3, 219–260.

N. Metta, R. Ramachandran, M. Ierapetritou, 2021, A novel adaptive sampling based methodology for feasible region identification of compute intensive models using artificial neural network, AIChE Journal, 67, e17095.

J.L. Pulsipher, D. Rios, V.M. Zavala, 2019, A computational framework for quantifying and analyzing system flexibility, Computers & Chemical Engineering, 126, 342–355.

S. Sachio, C. Kontoravdi, M.M. Papathanasiou, 2023, A model-based approach towards accelerated process development: A case study on chromatography, Chemical Engineering Research and Design, 197, 800–820.

N.V. Sahinidis, 1996, BARON: A general purpose global optimization software package, Journal of Global Optimization, 8, 201–205.

R.E. Swaney, I.E. Grossmann, 1985a, An index for operational flexibility in chemical process design. Part I: Formulation and theory, AIChE Journal, 31, 621–630.

R.E. Swaney, I.E. Grossmann, 1985b, An index for operational flexibility in chemical process design. Part II: Computational algorithms, AIChE Journal, 31, 631–641.

F. Zhao, I.E. Grossmann, S. García-Muñoz, S.D. Stamatis, 2021, Flexibility index of black-box models with parameter uncertainty through derivative-free optimization, AIChE Journal, 67, e17189.