Evaluating solution performance under uncertainty in superstructure optimization

Julia Granacher[a,1], Rafael Castro-Amoedo[a], Ivan Daniel Kantor[a,b] and François Maréchal[a]

[a] Industrial Process and Energy Systems Engineering, Ecole Polytechnique Fédérale de Lausanne, Sion, Switzerland,
[b] Department of Chemical and Materials Engineering, Concordia University, Montreal, Canada

Abstract

Superstructure optimization problems, as they often arise in the domain of Process Systems Engineering, are highly subjected to uncertainty inherent in different aspects of the optimization formulation, such as process performance, pricing parameters of resources and products, environmental impact factors, availability and service demands. While the matter of uncertainty in such problems has been addressed with a variety of methods and research contributions, we are proposing a framework to explore the effects of uncertainty in parameters on the obtained solution space and the decision-making process. Our approaches address the identification of important process unit decisions that drive the objective functions, the identification process unit sizes that are preferable for a variety of parameterised scenarios, and the parameter domains for which certain process unit sizes might be preferred compared to others. Furthermore, we elaborate on choosing from a set of solutions obtained for a superstructure optimization problem, considering parameter uncertainty in both the generation stage and the exploration of the solution space. The suggested approaches are applied to the design of an integrated pulp biorefinery, focusing the uncertainty analysis on economic parameters.

Introduction

Superstructure optimization has been established as an important method to derive sustainable and robust designs of process and energy systems in many domains of Process Systems Engineering (PSE). However, such optimization problems are subject to uncertainty - an inherent feature of any superstructure, affecting process performances, pricing parameters of resources and products, availabilities and service demands. Uncertainty in superstructure optimization has been addressed by a variety of research contributions (Sahinidis, 2004; Ning and You, 2019; Moret, 2017). Li and Grossmann (2021) provide a comprehensive overview of available contributions in the domain of PSE and uncertainty analysis, with a focus on stochastic optimization. Stochastic optimization is a major research area in the domain of optimization under uncertainty, relying on probability distributions of input parameters in the optimization problem. Classic textbooks address the main developments in the field (Birge and Louveaux, 2011; Shapiro, 2008; Pistikopoulos and Ierapetritou, 1995); a comprehensive review of existing methods is made available including mathematical formulations and illustrative examples by Li et al. (2022). Robust opti-

mization is another approach, based on a set of uncertainties in parameters that ensures the best benefits in the worst case of possible scenarios (Ben-Tal et al., 2004; Yanıkoğlu et al., 2019). As another powerful paradigm for optimization under uncertainty, chance constrained programming aims to optimize an objective while ensuring constraints to be satisfied with a specified probability in the uncertainty domain (Prékopa, 2013). Similarly to stochastic programming, uncertainty is tackled using probability distributions. More recently, the application of big data and machine learning for dealing with uncertainty in optimization has gained interest. Ning and You (2019) provide - besides a review of concepts and applications of stochastic programming, robust optimization and chance-constrained programming - insights on data-driven approaches towards optimization under uncertainty.

In general, many methods and approaches exist to tackle uncertainty in PSE-related design and optimization problems. However, only few of them exist that include uncertainty for both, solution generation and solution ranking, and that provide guidance throughout the decision-making process based on the uncertainty inherent in the model. The aim of the research herein is to develop approaches on how to identify well-performing and robust solutions of a superstructure optimization problem by considering the existence of uncertainty in data during solution generation and the selection of solutions from a set of unique configurations. In

---

[1] Corresponding author: J. Granacher (E-mail: julia.granacher@epfl.ch).

our context, robustness describes the ability of a configuration to perform well in a wide range of parameter uncertainty. Depending on the knowledge of the decision maker on the degree of uncertainty inherent in the problem, different approaches are suggested. Hence, different aspects the decision maker faces during the decision process are addressed with visually-assisted, data-driven approaches.

Methodology

In this contribution, we present a framework to address the impact of parameter uncertainty on superstructure optimization problems. Multiple aspects and considerations are discussed, each of them being applied to a PSE problem afterwards. The applications are centered around uncertainties in economic parameters for the design of integrated industrial biorefineries.

In the first part of this methodology section, we give a short overview of the overall method on how we generate solutions for a superstructure optimization problem. In the second part, we propose an approach for identifying the configuration among a set of obtained configurations that performs best under a wide range of scenarios. Finally, information available from correlations between optimization objectives, parameters and decisions is discussed.

Superstructure formulation and result generation

A process superstructure aims at describing the system's units and the way they interact with each other. By activating certain units and their connections, different system configurations can be achieved. In this work, the methodology for superstructure modelling and optimization is adapted from (Gassner and Maréchal, 2010; Kantor et al., 2020). For each unit in the system, energy and mass flow balances are formulated describing demand and supply properties. Binary decision variables describe whether a unit is installed, and whether it is used in a respective time-step. Continuous decision variables describe the size of the installed unit and the level of usage at which it is operated in each period. Both continuous and binary variables are constrained by parameterized bounds. Operating and investment costs are calculated as a function of equipment size; environmental impacts of the system are estimated using the LCI (Life Cycle Inventory) Ecoinvent database (Wernet et al., 2016). Pinch analysis is applied to model heat recovery opportunities and investigating the integration of the utility system by introducing heat cascade constraints as elaborated by Marechal and Kalitventzeff (1998). For solution generation, the decision variables are determined by solving a Mixed-Integer Linear Programming (MILP) problem formulated in the AMPL optimization language (Fourer et al., 2002), using the CPLEX (IBM, 2017) branch-and-bound algorithm.

The defined superstructure model, from now on referred to as lower-level framework, is integrated in an upper-level framework, in which optimization problems are formulated for exploring the impact of uncertainty in the parameter domain.
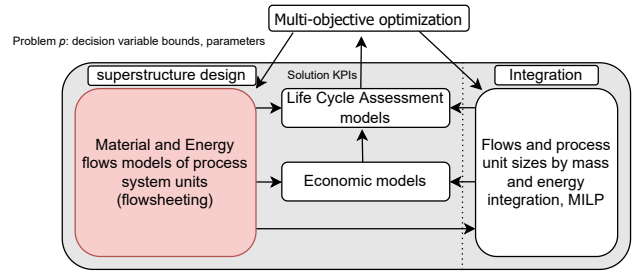

Figure 1: Optimization framework of the solution generation strategy, adapted from Granacher et al. (2022)

For generating solutions, the design space, consisting of parameters that are subjected to uncertainty, is defined and a design of experiment is conducted in the upper level to obtain the parameter distribution D_1. Obtaining distributions of environmental and economic parameters is challenging; often, they are not available. To avoid a bias in results induced by the assumption of inaccurate uncertainty distributions of parameters, parameters are sampled from an equally distributed parameter space, applying Latin Hypercube Sampling (LHS) (Iman, 2014). For each sample in the distribution, multi-objective optimization is performed applying the ε-constraint method, and the respective constraints as well as the parameter samples are communicated from the upper to the lower level, where the optimization problem is solved (Figure 1). The procedure yields a Pareto front for the objectives of interest for each scenario in D_1. In a next step, the set of unique S_{unique} process configurations S obtained over all samples in D_1, characterized by the activation status and installed size of the units present in the superstructure is identified.

Validating robustness of obtained configurations under parameter uncertainty

For identifying solutions among the obtained configurations that perform well under parameter uncertainty, the performance of all unique configurations S_{unique} that are present on the Pareto fronts obtained for D_1 is recalculated *a posteriori* to optimization. For this recalculation, a set of scenarios D_2 describing the uncertainty in parameters is considered. Similarly to D_1, D_2 is sampled using LHS. This procedure leads to $|D_2|x|S|$ datapoints describing the configuration's performance regarding objective functions throughout different scenarios.
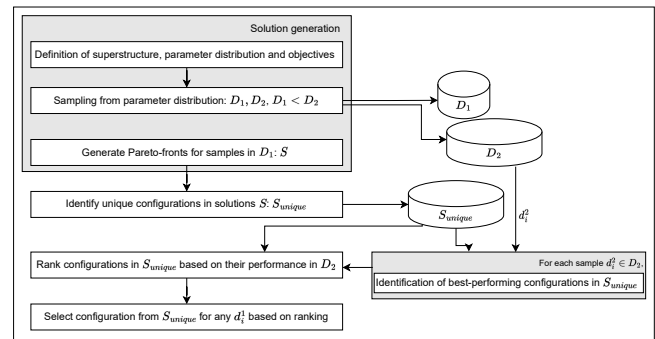

Figure 2: Proposed algorithm to rank configurations.

The recalculated objectives are used to identify the best-performing configurations for each scenario in D_2. 'Best' in

this case is defined as being among the configurations that are closest to Pareto optimality from all available configurations in a given scenario. In other words, configurations that are not dominated by others regarding the relevant objectives are identified. This classification allows to rank the configurations that are Pareto optimal for a scenario in D_1 regarding their presence in the near-Pareto optimal domain for a large variety of scenarios (D_2), and include this information in the decision-making process when choosing configurations. The described procedure is visualized in Figure 2 and further illustrated in Figure 3; for each scenario in D_2, configurations closest to Pareto optimality are identified. Unique configurations are then ranked based on their occurrence in domains close to Pareto optimality in D_2 (Figure 3, top). For an scenario in D_1, the computed rank of the configurations can then be used to prioritize Pareto optimal configurations (Figure 3, bottom).
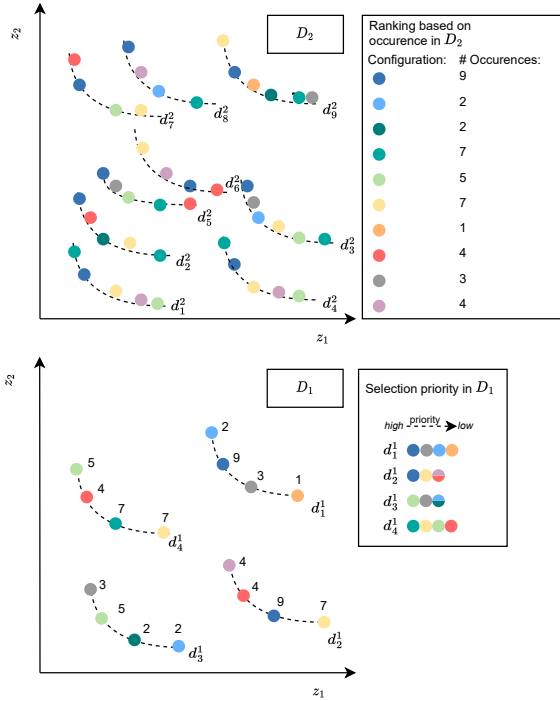

Figure 3: Ranking solutions based on close-to-Pareto optimality in D_2 (top) and prioritizing them based on the obtained ranking in D_1 (bottom) for selected objectives z_1 and z_2.

Interpreting the solution space

Besides the above-discussed approaches to validate and prioritize obtained Pareto optimal solutions based on their performance in the domain of parameter uncertainty, we suggest approaches and techniques to identify relevant unit decisions, their impact on changes in the objectives, as well as the impact of parameter variation on unit selection, to better understand the solution space of a superstructure optimization problem. We aim to address a set of questions: (i) What are the most important process units that influence the objectives of a superstructure optimization problem, and at which size are they typically installed? (ii) How do changes in the parameter space affect the optimizer unit choices? (iii) Under which conditions is a solution preferable compared to others?

For the first question, the unit choices that affect the objectives most are determined using the Spearman Rank correlation coefficient, relating the changes in objectives of the optimization problem to the changes in unit configurations and sizes (Dodge, 2008). The most correlated process units are identified based on the sum of the correlation coefficients determined for both objectives, and the distribution of the obtained process unit sizes are visually presented. For this visual representation, the configurations closest to Pareto optimality in 90% of the scenarios in D_2 are considered, meaning that we are showing the process unit sizes that are present in the robust solutions obtained from the solution generation. Robust in this case is meant to define the competitiveness of a certain configuration compared to others for a multitude of scenarios.

The next question addresses the conditions under which a process unit is typically installed in robust configurations, and in this way, identifying the conditions under which a certain unit is most likely to be competitive. Furthermore, typical sizes at which process units are emerging in configurations closest to Pareto optimality for certain scenarios are identified. This information might be relevant in high-uncertainty situations, where the decision maker is keen on having an overview on a wide range of possibilities.

Correlations between the parameter distribution and the objectives are derived for each unique configuration in the solution space, as well as the correlations between the unit choices and the objectives for each scenario. These correlations are used to identify the process units and parameters influencing the objectives. For each relevant process unit, its characteristics in robust configurations across the range of considered scenarios is displayed, showing both unit installation and size of the process unit.

For identifying typical solutions and their validity ranges in the scenario space, a similar approach to the one in Section 2.2 is elaborated. The difference is that previously, the decision maker had a strong guess about the scenario to consider in D_1 and uses the distributions in D_2 to evaluate the performance of the obtained Pareto optimal configurations for this chosen scenario, whereas in the approach suggested hereafter, no strong guess is available. Therefore, validity regions for obtained configurations, i.e. scenarios in which a configuration is preferable compared to the others, are identified. Typical configurations are obtained by clustering the unique solutions on unit decisions; the number of clusters is identified using the k-elbow approach (Satopaa et al., 2011). The data is reduced to two dimensions by applying Principal Component Analysis (PCA) with high explained variance, followed by a t-Distributed Stochastic Neighbor Embedding that allows for adequate visualization of the clusters. For clustering, the k-Medoid partitioning technique is applied, yielding cluster representatives from the respective cluster members rather than average data (Park and Jun, 2009). For each cluster, the representative is selected based on its occurrence in Pareto optimal domain of D_1 and near Pareto optimal domains in D_2. For visualizing the ranges of the scenarios where each cluster is performing well, we identify the sce-

narios in D_2 and D_1 where the cluster members belong to the configurations that are closest to Pareto optimality (D_2) or Pareto optimal ($D1$). The obtained parameter space is displayed in parallel coordinates, showing the parameter distribution and the resulting objectives in Pareto optiomal or near Pareto optimal domains for each cluster. Furthermore, we show the distribution of the clusters for binned objective functions. For each bin, we determine the number of times a cluster member is close to Pareto optimality or Pareto optimal and leads to a objective that is in the respective bin; allowing to identify typical solutions that are preferable if a certain overall performance regarding the objectives is desired. The described procedure is visualized in Figure 4.
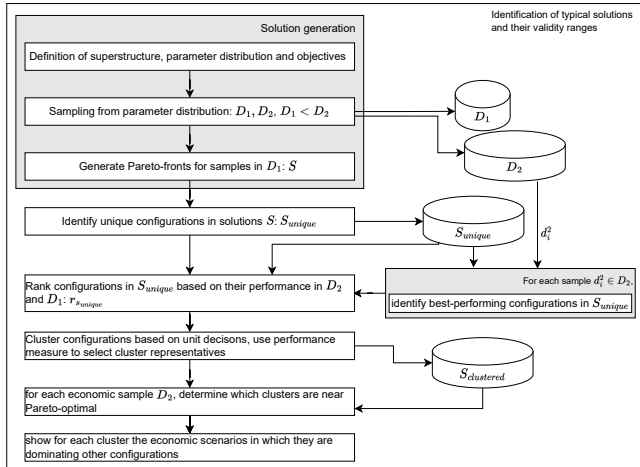

Figure 4: Identification of validity ranges of clusters.

Application and results

Solution generation

We apply the suggested approaches to the solutions generated for the efficient design of an integrated industrial biorefinery, where a Kraft pulp mill is enriched with biomass-based process units that convert excess electricity and biogenic residual streams such as bark and black liquor to storable energy in the form of fuel. The biomass-based processes can be divided into two main pathways, dry gasification of bark and hydrothermal gasification of black liquor, both followed by fuel synthesis. More details on the considered superstructure are available in Granacher et al. (2022). Optimizer decisions include system configurations which contain the installation of units as well as their sizes. The initial results were generated for competing objectives total expenditure (TOTEX) and environmental impact measured in Global Warming Potential (GWP), applying the method described in Section 2.1. In the current example, we are considering uncertainties inherent in economic parameters. Our sample set D_1 has a size of 50, containing 25 economic parameters to be varied with equal probability in ranges that are defined by $\pm$ 20%-50% of the mean of historic market observations, depending on the observed profiles. For each of the samples in D_1, a total of 10 Pareto optimal configurations are generated with the methods described in 2.1, using the ε-constraint method for GWP and TOTEX as objectives, which yields a solution space of 500 configurations. Out of

the 500 configurations, 487 are identified as unique regarding unit choice and size. For generating the set D_2, 1000 samples of economic conditions are created, following the same sampling approaches as the ones used for D_1. For each unique configuration, the objectives are recalculated for all samples in D_2, yielding a total of 487000 datapoints.

Validation of solution robustness

We use the proposed algorithm (Fig 2) to identify preferable solutions for a certain economic scenario in D_1. For each economic sample in D_2, the solutions closest to Pareto optimality from the 487 unique configurations are identified, which allows to derive a performance measure for each configuration. Going back to the economic sample in D_1 and the corresponding Pareto front, each configuration on the Pareto front can now be associated with this performance measure. For the economic scenario displayed in Figure 5, a clear preference on which configurations to choose based on this ranking can be observed: the configuration with minimum impact and the one one next to it are the most competitive ones in terms of occurrence on the near-Pareto optimal domain for different economic scenarios, the occurrence being visualized by the size of the points.
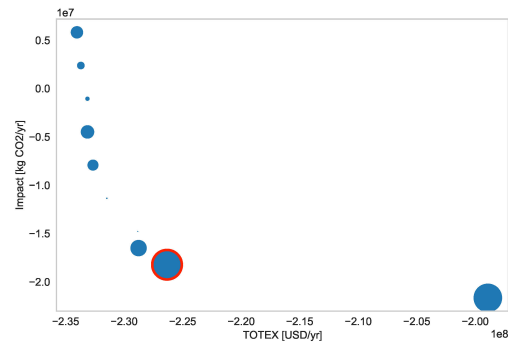

Figure 5: Resulting Pareto front for economic sample d_i in D_1, the size of the points indicating their occurrence in the near-optimal domains in D_1 and D_2.

We investigate the second configuration (marked in red in Figure 5) further by looking at the characteristics of the process unit installations. Figure 6 displays the sizes of the process units active in this configuration (marked as white), compared to the observed distribution of the respective process unit sizes over all obtained unique configurations. For readability, only the nine active process units with the highest correlation to the objectives are displayed.

Even though not all the units considered in the superstructure are displayed, some general conclusions regarding the characteristics of the analyzed system configuration can be drawn. It can be observed that for the selected configuration, the largest size of the dry gasification line in all of the obtained solution space is present, indicated by the size of the water gas shift reactor, the gasifier and the gas cleaning unit, which affect also the flowrate of offgases in the lime kiln and the amount of natural gas needed, which is at the lower end of the observed distribution. Regarding CO_2 capture and the Selexol unit, the configuration is positioned on the average over the solution space, thereby, the mass flow

into the Selexol unit is and indicator for the amount of black liquor going through hydrothermal gasification and consecutive cleaning.
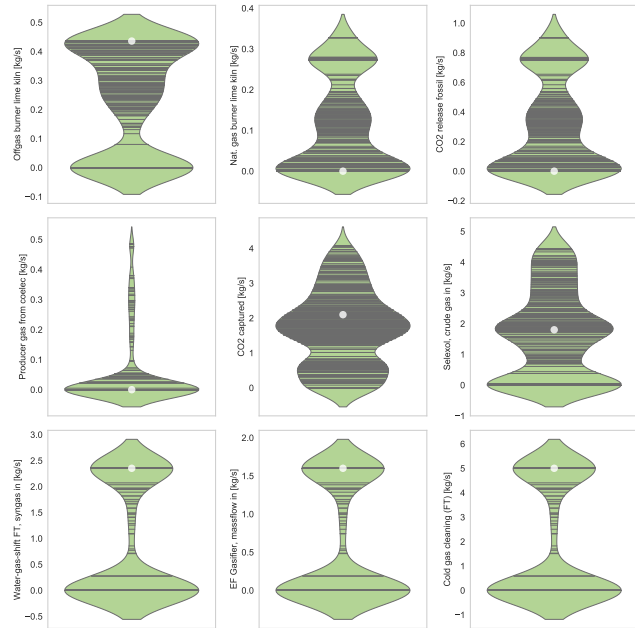

Figure 6: Position of identified configuration regarding its process unit sizes, compared to the process unit sizes in all of the unique configurations.

Generally, it can be observed that the chosen configuration consists of process unit sizes that are common among a wide range of obtained unique configurations (indicated by the length of the horizontal lines in Figure 6); this observation is strengthened by the fact that the process unit sizes of the selected configuration also lead to preferable performance in most of the scenarios in D_2. Therefore, we conclude that for obtaining robust configurations, installing the displayed process units in the respective sizes seem favourable.

Interpretation of the solution space

As suggested in 2.3, visually assisted approaches are followed to understand the correlations between scenarios, optimizer decisions and effects on the objectives. To identify the changes in process unit size that drive the objective functions, the Spearman Correlation method is applied as previously described. The typical sizes of the highly correlated process units and their frequency throughout all obtained unique configurations are displayed in Figure 7. Furthermore, the unit sizes found in the configurations that perform closest to Pareto optimality in 90% of the considered scenarios are marked as dots; from now on they are referred to as dominating sizes. The process units showing the highest correlations to the objectives over all unique configurations are the ones describing the wet gasification line and its effects on the recausticiser in the pulp mill. For most process units, between five and seven dominating sizes can be identified. It needs to be mentioned that, since these are the process units that influence the objective functions significantly, only few dominating sizes are detected. Nevertheless, the obtained Figures help to identify critical process units that influence the ob-

jectives under parameter variation, as well as sizes for the respective units in which the system configuration's performance is likely to be preferable in the majority of economic scenarios.
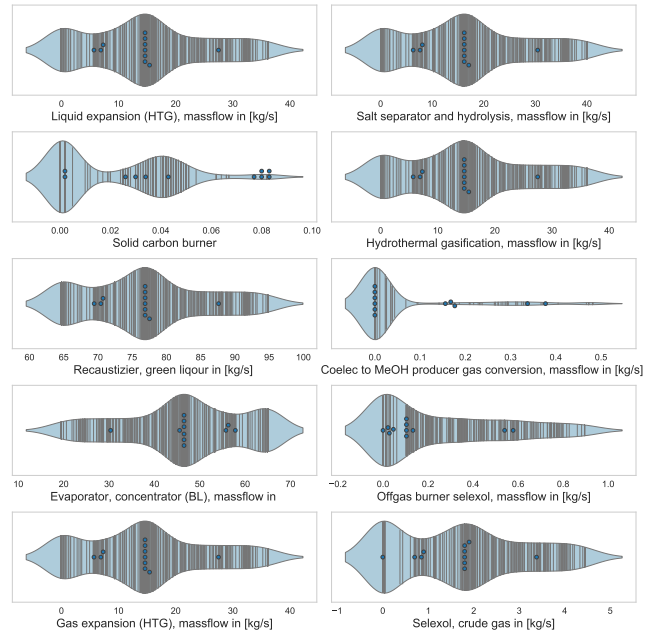

Figure 7: Typical process unit sizes of units with high correlation to objectives in the near-Pareto optimal domains (dots) vs. the observed sizes over all unique configurations.

After analyzing the relation between process unit sizes and obtained performance of related configurations in the uncertainty domain, the process units and their sizes are furthermore related to the economic conditions in which their performance is preferable compared to other unit installations.
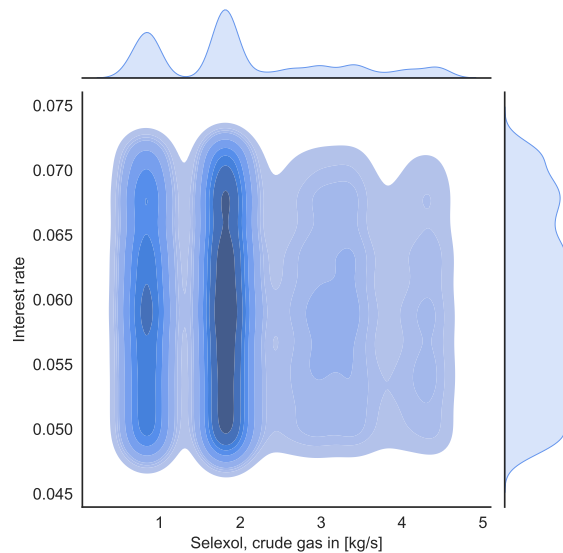

Figure 8: Heat map of Selexol process unit sizes in near-Pareto optimal domains for varying interest rate.

For each identified relevant process unit, the size distribution is visualized in relation to the economic parameters at which the respective size is present in a configuration selected as closest to Pareto optimal in D_2 or Pareto optimal in

D_1.

For the example given in Figure 8, it can be noted that one process unit size seems to be equally competitive or not competitive for the range of economic parameters analyzed in D_2, meaning that a process unit size is performing strongly or less strongly over all of the economic parameter space, and no relevant variation over the economic space can be observed. Therefore, for the given example, no conclusions regarding economic conditions are preferably for the installation of certain process units sizes can be drawn; however, it can be noted that a certain size, in the case of the visualized Selexol process unit, a gas flow of 2 kg/s, is preferable compared to other sizes.

Following the identification of units that characterize a configuration and influence the objectives and their typical sizes and economic ranges at which they are installed, we take one step back and again look at the ensemble unit decisions that define a obtained configuration. In 3.2, we discussed the selection of configurations from a set of Pareto optimal solutions for one economic scenario in D_1, considering an assumed uncertainty in economic parameters. If however, the decision maker is not confident to make a guess about which economic scenario is the most adequate one to choose from, they can select configuration on an ensemble of observations rather than a strong guess. More specifically, the whole ensemble of economic scenarios is equally weighted when ranking configurations, instead of a-priori choosing one most likely one from D_1 with the corresponding Pareto front.

In our example, the approach described in Section 2.3 yields five cluster representatives describing the solution space. For each of the cluster representatives, the economic scenarios for Pareto optimal or close to Pareto optimal performance are identified. These economic conditions are then displayed in parallel coordinates, where we differentiate between conditions where a cluster performs close to, or strictly Pareto optimal (Figure 9).
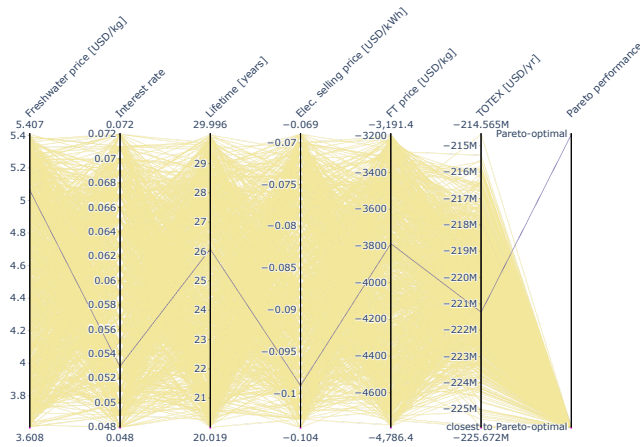
Figure 9: Economic conditions for near-Pareto optimality (yellow) and Pareto optimality (black) of cluster 4.

This visual representation aims to assist in identifying the economic domains in which a configuration is preferable compared to others, however, a significant shortcoming hereby is currently the lack of representation of actual regions in the economic domain rather than ranges of parameters as they are currently displayed in Figure 9; this issue will be addressed in future work.

For refining the selection between the cluster representatives, their individual performance regarding the objectives is analyzed. For the observed range of economic performance (TOTEX), ten bins are created, and for each cluster, the number of times it performs closest to Pareto optimality with a TOTEX in the respective bin is derived. In the visual representation, the size of the circles represents the relative number of times the cluster is selected for the binned objective, the color represents the cluster. The number of times selected is displayed in % of times it could theoretically be selected at closest to Pareto optimal.
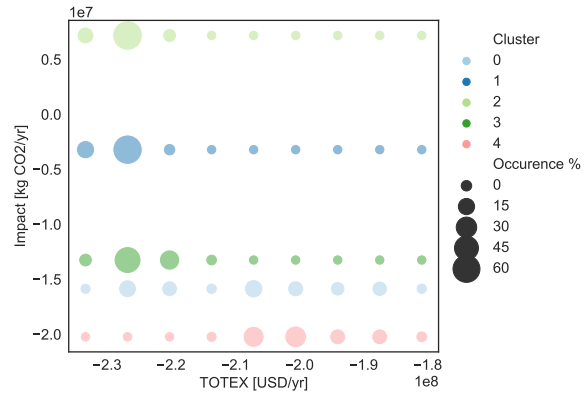
Figure 10: Performance of clusters on binned Pareto front. Size of the dots represents how often a cluster is occurring among the configurations closest to Pareto optimality in the scenarios of D_2 while resulting in a TOTEX within the respective bin.

For the analyzed superstructure optimization problem, the obtained cluster performance is presented in Figure 10. For two clusters, namely 0 and 4, the environmental impact is rather low compared to the average over the cluster representatives, with the occurrence on near-Pareto optimal domains being relatively constant over all considered bins of TOTEX for cluster 0, while cluster 4 is more likely to result in a higher TOTEX. The other three clusters show a similar distribution regarding TOTEX, in 45% of the analyzed economic scenarios, they appear in the domains closest to Pareto optimality with a resulting TOTEX between -235 and -228 MUSD, while differing significantly in environmental performance.

Conclusion

In this paper, approaches on decision-making under uncertainty for superstructure optimization problems have been presented. While the analysis can surely not lead to one "optimal" solution to pick, we hope that the suggested visual representations help to inspire and facilitate decision-making under uncertainty, by providing insights on correlations between unit choices, objectives and economic scenarios, and thereby deepening the understanding of solutions obtained from superstructure optimization problem. In future work, we aim to generalize the suggested approaches further, so

that uncertainty in other parameters, such as environmental assumptions can be considered. In addition, the consideration of more than two objectives will be taken into account in future versions of this work. Furthermore, we are aiming at making the developed tool available as an open source platform for interactive visualization of superstructure optimization solution characteristics considering parameter uncertainty.

Acknowledgements

References

Ben-Tal, A., A. Goryashko, E. Guslitzer, and A. Nemirovski (2004, March). Adjustable robust solutions of uncertain linear programs. *Mathematical Programming 99*(2), 351–376.

Birge, J. R. and F. Louveaux (2011, June). *Introduction to Stochastic Programming*. Springer Science & Business Media. Google-Books-ID: Vp0Bp8kjPxUC.

Dodge, Y. (2008). Spearman Rank Correlation Coefficient. In *The Concise Encyclopedia of Statistics*, pp. 502–505. New York, NY: Springer.

Fourer, R., D. Gay, and B. Kernighan (2002, January). *AMPL: A Modeling Language for Mathematical Programming* (2 ed.). Duxbury-Thomson. Journal Abbreviation: Management Science - MANAGE SCI Publication Title: Management Science - MANAGE SCI.

Gassner, M. and F. Maréchal (2010). Increasing Conversion Efficiency in Fuel Ethanol Production from Lignocellulosic Biomass by Polygeneration - and a Paradoxon between Energy and Exergy in Process Integration. In *Infoscience*. Number: CONF.

Granacher, J., T.-V. Nguyen, R. Castro-Amoedo, and F. Maréchal (2022, January). Overcoming decision paralysis—A digital twin for decision making in energy system design. *Applied Energy 306*, 117954.

IBM (2017). CPLEX User's Manual.

Iman, R. L. (2014). Latin Hypercube Sampling. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd.

Kantor, I., J.-L. Robineau, H. Bütün, and F. Maréchal (2020). A Mixed-Integer Linear Programming Formulation for Optimizing Multi-Scale Material and Energy Integration. *Frontiers in Energy Research 8*, 49.

Li, C. and I. E. Grossmann (2021). A Review of Stochastic Programming Methods for Optimization of Process Systems Under Uncertainty. *Frontiers in Chemical Engineering 2*.

Li, Y., J. Wei, Z. Yuan, B. Chen, and R. Gani (2022, January). Sustainable synthesis of integrated process, water treatment, energy supply, and CCUS networks under uncertainty. *Computers & Chemical Engineering 157*, 107636.

Marechal, F. and B. Kalitventzeff (1998, March). Process integration: Selection of the optimal utility system. *Computers & Chemical Engineering 22*, S149–S156.

Moret, S. (2017). *Strategic energy planning under uncertainty*. Ph. D. thesis, EPFL.

Ning, C. and F. You (2019, June). Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. *Computers & Chemical Engineering 125*, 434–448.

Park, H.-S. and C.-H. Jun (2009, March). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications 36*(2, Part 2), 3336–3341.

Pistikopoulos, E. N. and M. G. Ierapetritou (1995, October). Novel approach for optimal process design under uncertainty. *Computers & Chemical Engineering 19*(10), 1089–1110.

Prékopa, A. (2013, March). *Stochastic Programming*. Springer Science & Business Media. Google-Books-ID: SH7vCAAAQBAJ.

Sahinidis, N. V. (2004, June). Optimization under uncertainty: state-of-the-art and opportunities. *Computers & Chemical Engineering 28*(6), 971–983.

Satopaa, V., J. Albrecht, D. Irwin, and B. Raghavan (2011, June). Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, Minneapolis, MN, USA, pp. 166–171. IEEE.

Shapiro, A. (2008, March). Stochastic programming approach to optimization under uncertainty. *Mathematical Programming 112*(1), 183–220.

Wernet, G., C. Bauer, B. Steubing, J. Reinhard, E. Moreno-Ruiz, and B. Weidema (2016, September). The ecoinvent database version 3 (part I): overview and methodology. *The International Journal of Life Cycle Assessment 21*(9), 1218–1230.

Yanıkoğlu, , B. L. Gorissen, and D. den Hertog (2019, September). A survey of adjustable robust optimization. *European Journal of Operational Research 277*(3), 799–813.