CHEMICAL ENGINEERING TRANSACTIONS
VOL. 105, 2023
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: David Bogle, Flavio Manenti, Piero Salatino

ISBN 979-12-81206-04-5; **ISSN** 2283-9216
DOI: 10.3303/CET23105051

Environmental Assessment of Glucose Production using Neuronal Networks

Nancy Prioux*, Rachid Ouaret, Jean-Pierre Belaud

Laboratoire de Génie Chimique, Université de Toulouse, CNRS, Toulouse, France
nancy.prioux@toulouse-inp.fr

A previously conducted study has developed a comprehensive framework that combines Machine Learning techniques and environmental assessment to facilitate decision-making in Process System Engineering (Prioux et al., 2023). This framework consists of five distinct steps: goal and scope definition, data architecture establishment, sustainability assessment, data visualization and result analysis, and the final decision-making process. The data architecture itself is directly influenced by the construction of big data architecture and comprises five sub-steps: (i) data collection and extraction from Knowledge Engineering, (ii) data enrichment and storage facilitated by Process Engineering and Machine Learning (ML), (iii) data cleaning, (iv) analysis of (raw) data, and (v) visualization of (raw) data. In the third step, environmental impacts are calculated, while the fourth step involves the utilization of ML tools for analyzing and visualizing the obtained results. To evaluate the efficacy of this approach, a comparison of various biomass pretreatment processes for glucose production was conducted. The primary objective of this study is to examine the steps of data enrichment and data cleaning in detail. To obtain relevant scientific articles, a targeted search using specific keywords was conducted on platforms such as Science Direct and Web of Science. However, it was observed that the data quality among these articles is inconsistent, with the presence of potential inconsistencies. Leveraging Machine Learning (ML) techniques enables us to enhance the quality of the data. This research paper seeks to emphasize the significance of ML, particularly neural network tools, in the context of environmental assessments within the realm of Big Data. The key focal points of our approach encompass enriching data from literature sources to simulate processes and identifying outliers within the extracted data.

1. Introduction

Recent years have seen the emergence of action plans for climate, sustainable development or the environment such as the European "Green Deal" plan of January 2020. One of the main action plans, entitled "Action Plan for the Circular Economy," proposes to use the circular economy (CE) model to simulate the use of sustainable models. To achieve sustainable models, "life cycle thinking" can help improve environmental performance while maximizing economic and social benefits (Belaud et al., 2017). Several comprehensive methods have emerged to design waste recycling processes that fit into the circular economy (Grimaud et al., 2017). Agriculture is a particular area where CE and "life cycle thinking" have developed over the past decade. A large portion of agricultural waste is lignocellulosic by products that can be converted into bioenergy, biomolecules, or biomaterials.

In the past three decades, a considerable amount of research has been conducted and published on various pretreatment processes (Davis et al., 2017). However, a notable gap exists in terms of standardized criteria that can aid in selecting the most suitable technology pathway among the available options. Incorporating environmental, economic, and social assessments within a Circular Economy (CE) framework offers a promising approach to address this challenge. These assessments necessitate a substantial volume of data, particularly process-related data that can potentially be sourced from scientific or technical literature or generated through simulation techniques (Morales-Mendoza et al., 2012). Scientific articles outlining the recovery processes for agricultural by-products represent a significant yet underutilized data source.

Paper Received: 5 February 2023; Revised: 12 July 2023; Accepted: 19 July 2023
Please cite this article as: Prioux N., Ouaret R., Belaud J.-P., 2023, Environmental Assessment of Glucose Production Using Neuronal Networks, Chemical Engineering Transactions, 105, 301-306 DOI:10.3303/CET23105051

Handling such a large quantity of data can pose challenges; however, numerous tools and methods now exist for this purpose, including those derived from "Big Data" or artificial intelligence, such as machine learning (ML). By leveraging these technologies, environmental assessment tools like Life Cycle Assessment (LCA) can effectively analyze process data and assessment outcomes. The proposed approach establishes a connection between machine learning and sustainability assessment, providing valuable support to researchers and engineers in the analysis and comparison of diverse lignocellulosic biomass pretreatment processes and biomass types within the circular economy context.
Previous work has developed a complete framework for decision-making in Process System Engineering by coupling Machine Learning techniques and environmental assessment (Prioux, Ouaret, & Belaud, 2022; Prioux, Ouaret, Hetreux, et al., 2022). Five steps characterize this framework: goal and scope, data architecture, sustainability assessment, data visualization and analysis of results and decision. The data architecture is directly inspired by the construction of big data architecture and consists of five sub-steps: (i) data collection and extraction from Knowledge Engineering (ii) data enrichment and storage thanks to Process Engineering, and thanks to Machine Learning (ML) (iii) data cleaning (iv) (raw) data analysis, and (v) (raw) data visualization. The third step, environmental impacts are calculated. In the fourth step, the use of ML tools for analysis and visualization of results. The approach is tested on the comparison of biomass pretreatment processes for glucose production. This paper focuses on the data enrichment and data cleaning steps.

2. General approach

In this section, the steps of our approach are briefly explained and the global methodology is presented in Figure 1. In the first step, the purpose and boundaries of the study must be clearly defined. "Life cycle thinking" is recommended. This thinking encourages a "cradle-to-grave" or "cradle-to-gate" approach if the logistics of a value chain is difficult to obtain or even "cradle-to-cradle" in circular engineering. The system boundaries and functional unit significantly influence the evaluations. For example, it needs to be clarified whether the upstream biomass supply chain is taken into account. Once the objective and scope are properly defined, the supply chain, technologies and transformation processes must also be described.

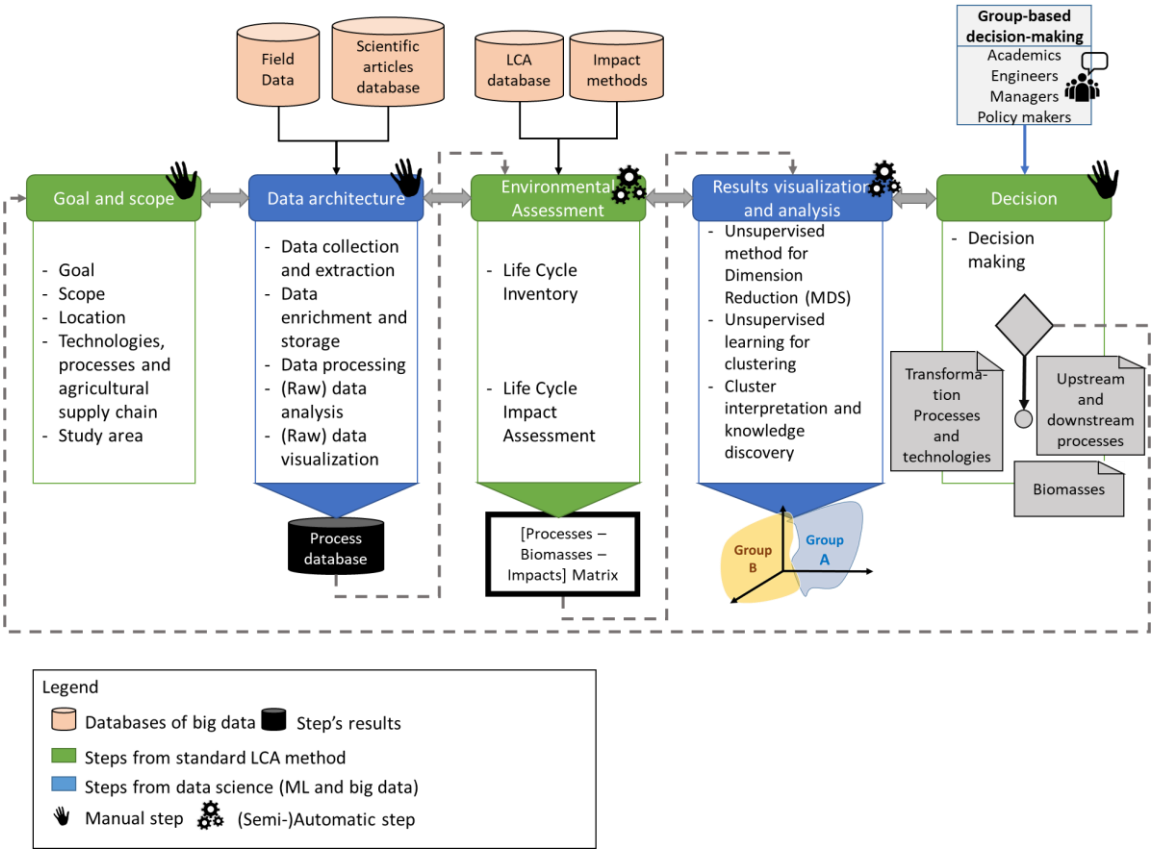

Figure 1. Five main steps of the approach from (Prioux, Ouaret, Hetreux, et al., 2022)

The data architecture is directly inspired by the construction of the Big Data architecture and consists of five sub-steps: (i) data collection and extraction (ii) data enrichment and storage (iii) data processing (iv) (raw) data analysis and (v) (raw) data visualization. This step can be automatic, semi-automatic or manual and uses methods from knowledge engineering or ML. The prediction of glucose yield has been accomplished using an optimized artificial neural network model. Subsequently, the interaction between inputs and outputs is visualized through the utilization of partial dependence plots (PDP) and individual conditional expectation plots (ICE).
The third step consists of completing the life cycle inventory using the process data from the previous step. The inventory is completed with free or commercial databases such as the EcoInvent database containing background data. Next, one or more impact calculation methods must be determined in accordance with the first step, which incorporates the nature of the study and the system. The sustainability impacts or even the damage spheres are calculated. At the end of this step, the main result is a structure [processes: biomasses: impacts] that is difficult to analyze.
The step of visualization and analysis of the results includes ML-derived methods to assist in the analysis of this matrix. Starting from the statistical literature, traditional Multi-Dimensional Scaling and unsupervised clustering techniques (k-means) are combined to extract relevant information. Finally, the user analyzes the points grouped in clusters to link them to significant processes/impacts. The visualization of the clusters will help the researcher or R&D engineers in the final decision according to the objective of their studies.
The following paragraphs will focus on the data enrichment and data processing steps. After the data collection and extraction step, all data needed to pursue our methodological framework are stored in structured, easy-to-use databases. It is from this base that we will carry out the enrichment and the processing, analysis and visualization of the data.

3. Data enrichment and processing

Data enrichment is the process of adding data from experts or data from experts or models in the studied domain. Depending on the domain studied, the models can be empirical numerical simulations, thermodynamic or energetic simulations or energy simulations, material transfer or transformation functions or even chemical reactions. This enrichment depends on the stakeholders of the study and the models available in the field of study. It is also possible to use neural networks to enrich the data. In our case study, for example, processes can be simulated. Indeed, the data extracted from the scientific literature will be used as a learning base and a test base for our model and once our model is set, processes can be simulated.
The data processing sub-step involves cleaning, adding and deleting data for volume and value management of the database under study. After this processing, a final, more accurate and accessible database can be generated. The processing saves time in the following steps and avoids incorrect interpretation in the during the raw data analysis sub-step and the sustainability analysis step. This step is done by the expert in a manual or semi-automatic way. An use in this case of neural network interpretation methods to understand the influence of the inputs on the result allows analyzing the processes one by one and to notice aberrant processes. In our case study, we will use the methods of partial dependence plots (PDP) and individual conditional expectation plots (ICE) (Friedman, 2001). Partial dependence helps to understand the marginal effect of a predictor (or a subset of it) on the predicted outcome. Essentially, it allows us to understand how the variable of interest changes when we change the value of one predictor while accounting for the average effect of all other predictors in the model. The ICE method, instead of averaging the predicted values for all observations, plots the predictions of each copy for each observation. Thus, an ICE plot visualizes the dependence of the predicted response to a predictor for each observation separately. This results in several lines, one for each observation, compared to one line for the PDP plot. A PDP plot is the average of all the plots in an ICE curve.

4. Case study

The approach is tested by comparing biomass pretreatment processes for glucose production. Only the environmental area is considered.The goal of the study is to help a researcher select a process for glucose production. The boundaries range from biomass to the enzymatic hydrolysis step i.e., a “cradle-to-gate” approach. Biomass is considered as a waste – the impacts of agricultural phases are attributed to the end product. If the biomass is considered as a co-product, the impacts of the production of the final product will be split between it and the biomass. The biomass transport phase impact is minor – the biorefinery is close to the field. The functional unit is “1 g of glucose” and all results are expressed based on this unit.
Thanks to specific keywords in Science Direct and Web of Science, twenty articles have been selected. Relevant data from these articles are extracted semi-automatically using an ontology (Lousteau-Cazalet et al., 2016). This represents more than 23.000 data (numeric or text). Each scientific article is entered in the ontology with its meta-information (source type, reputation, citation data).

The ontology structures the process data and ensures an export in CSV files supplying internal software. This software developed on Microsoft Excel conducts a first “cleaning” of the data by simulating the processes to calculate and check the mass balance. After this sub-step, we remove the data of three articles because they contained inconsistencies or many missing data points that are not amenable to be verified by the simulation. In the following we study 78 processes that deal with four different biomasses (bagasse, corn stover, rice straw, wheat straw) and two types of processes (purely mechanical processes with or without ultra-fine milling).
For the regression, Orange allows making the succession of the steps in a visual and automatic way. After the data are extracted in the software and extracted in the software and then normalized, the variable of interest is specified as well as the input variables. Normalization is the transformation of the values into centered and reduced values. The input variable is the glucose yield (glucose_yield). Only the numerical input variables are considered: Theoretical possible glucose yield, enzymatic activity, minimum size of the solid output constituent of the milling steps, temperature (maximum), quantity of biomass input, quantity of buffer liquid and the duration of the enzymatic hydrolysis step (Figure 2).

The variables are separated into two groups:

- the training group: these are the data that will train the ML model.
- The test group: these are the data that will allow evaluating the reliability of the model by comparing the value predicted by the model to the value corresponding to this group.

The training group represents 80% of the data, i.e., 63 experiments and the test group 20% or 15 experiments. Five models are tested on Orange and it is the RN model which is the best with R2 = 0.974. The RN is composed of 4 layers of 100 neurons each. The visualization of the predicted data according to the real data is represented on the Figure 3.

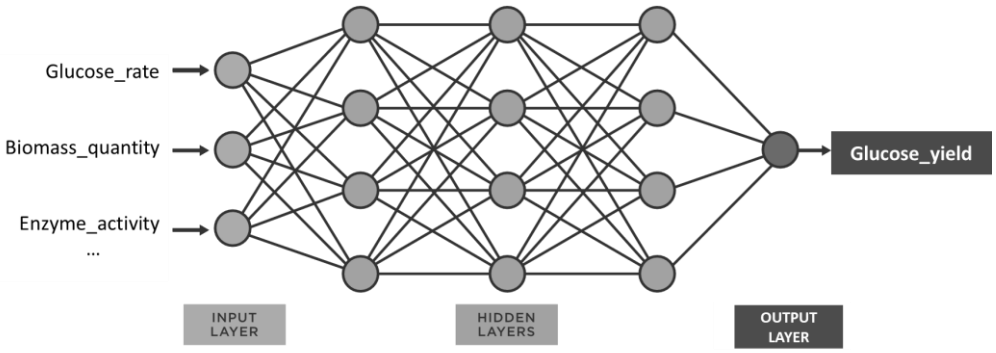

Figure 2. Prediction of glucose yield by neural networks

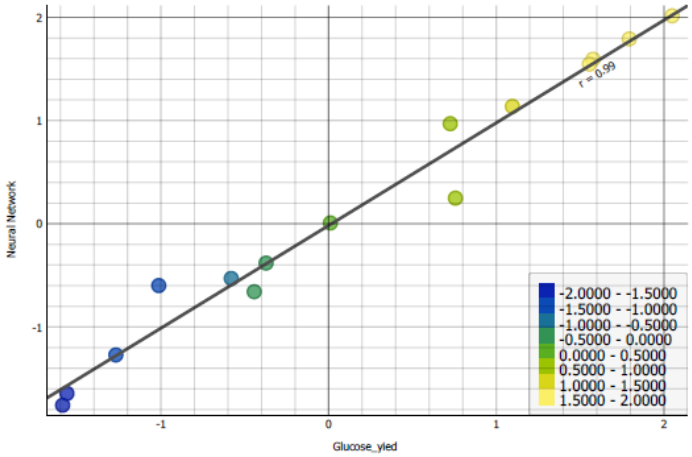

Figure 3. Visualization of the glucose yields predicted by the neural network model as a function of the actual real glucose yields (normalized values)

Once the regression is done, it is possible to use the model to "simulate" experiments without having to do real experiments. It is recommended to stay within the ranges of the input variables. To visualize the distribution of the different input variables, the PDP and ICE graphs are. An example of the impact of glucose rate buffer liquide quantity are presented in the Figure 4. We notice that the ICE curves are globally similar to each graph, that is to say that the experiments evolve in the same way if we vary the variables 1 to 1. The PDP and ICE illustrate the relationship between a selected input variable and the output prediction of the ANN while holding other input variables at fixed values. It provides insights into the impact and influence of a specific input variable on the prediction outcome. By examining the shape and trend of the PDP curve, one can infer the nature and magnitude of the relationship between the input variable and the prediction. A horizontal PDP curve indicates that the input variable has no significant influence on the prediction, while non-linear or varying curves suggest a more complex relationship.

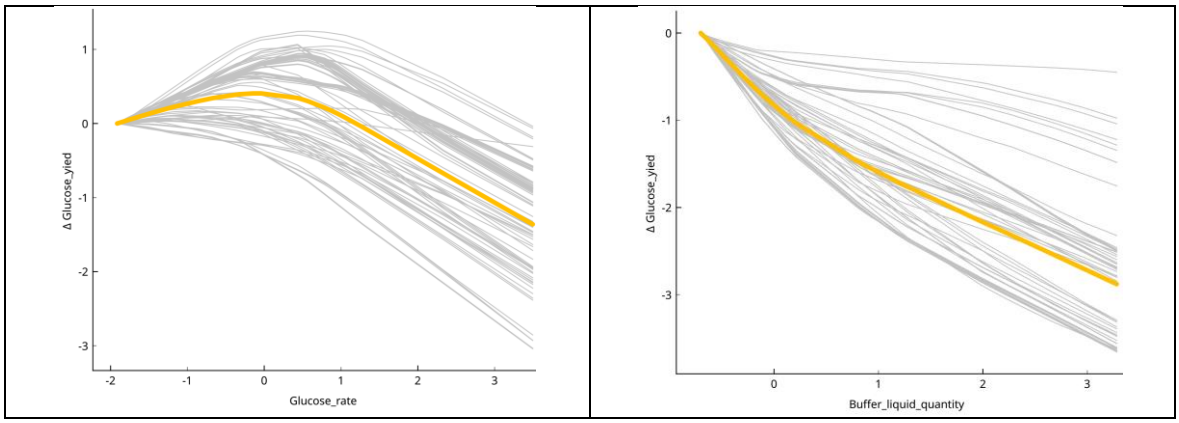

Figure 4. Centered partial dependence plot (in yellow) and individual conditional expectation (in black)

5. Conclusions

To conclude, a five-step approach coupling data science and environmental assessment for researchers or R&D engineers in a preliminary study has been presented in this paper. The enrichment and processing steps have been the most detailed with the addition of AI methods such as neural networks and two methods of interpretation of these networks: PDP and ICE graphs. We have seen that neural networks can simulate new processes thanks to the data extracted from the literature. However, these processes will have to remain within the limits of the processes that were used to train the model.

Several limitations have been identified:

- The data from the scientific literature are by nature data from a series of batch experiments in the laboratory. The life cycle analysis (LCA) is therefore performed for a low level of technology readiness level (TRL) or maturity (TRL 1/2).
- The neural networks can simulate processes which are not in the scope of literature data.
- The abundance and the quality of the data are not sufficient for these new technological processes.
- Some knowledge of the data and models is needed to find outliers in the PDP and ICE graphs.

Several perspectives on the method and the case study can be identified:

- Integration of the agricultural upstream phase and consideration of the overall supply chain according to a dynamic analysis.
- The policy of impacts related to effluents can be modified by taking into account the future of these effluents and their valorization in the framework of the circular economy.
- The implementation of another ML method to find the outlier.

Acknowledgments

This work has been sponsored by the French government research program “Investissements d’Avenir” through the Research National Agency (ANR-18-EURE-0021).

References

Belaud JP, Adoue C, Sablayrolles C, Vialle C, Chorro A, 2017, Decision making approach for industrial ecology: layout and commercialization of an industrial park, Chemical Engineering Transactions, 57, 1561- 1566

Davis, C. B., Aid, G., & Zhu, B. (2017). Secondary Resources in the Bio-Based Economy : A Computer Assisted Survey of Value Pathways in Academic Literature. *Waste and Biomass Valorization*, *8*(7), 2229-2246. https://doi.org/10.1007/s12649-017-9975-0

Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *Annals of statistics*, 1189-1232.

Grimaud, G., Perry, N., & Laratte, B. (2017). Decision Support Methodology for Designing Sustainable Recycling Process Based on ETV Standards. *Procedia Manufacturing*, *7*, 72-78. https://doi.org/10.1016/j.promfg.2016.12.020

Morales-Mendoza, L. F., Azzaro-Pantel, C., Belaud, J.-Pierre., Pibouleau, L., & Domenech, S. (2012). An integrated approach combining process simulation and life cycle assessment for eco-efficient process design. In I. D. L. Bogle & M. Fairweather (Éds.), *Computer Aided Chemical Engineering* (Vol. 30, p. 142-146). Elsevier. https://doi.org/10.1016/B978-0-444-59519-5.50029-0

Prioux, N., Ouaret, R., & Belaud, J.-P. (2022). Machine Learning Based Framework for Biorefinery Environmental Assessment. *Chemical Engineering Transactions*, *96*, 517-522. https://doi.org/10.3303/CET2296087

Prioux, N., Ouaret, R., Hetreux, G., & Belaud, J.-P. (2023). Environmental assessment coupled with machine learning for circular economy. *Clean Technologies and Environmental Policy*. https://doi.org/10.1007/s10098-022-02275-4