CHEMICAL ENGINEERING TRANSACTIONS
VOL. 96, 2022
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: David Bogle, Flavio Manenti, Piero Salatino

ISBN 978-88-95608-95-2; **ISSN** 2283-9216
DOI: 10.3303/CET2296087

Machine Learning Based Framework for Biorefinery Environmental Assessment

Nancy Prioux*, Rachid Ouaret, Jean-Pierre Belaud

Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France
nancy.prioux@ensiacet.fr

The transformation of actual processes into sustainable processes is a major study subject over recent years, particularly through the circular economy. However, the environmental assessments require a huge quantity of data and many of these data are heterogeneous. Environmental evaluation tools would clearly benefit from Data Science approaches in the Big Data context. This paper focuses on developing a framework for decision-making in Process System Engineering by coupling Machine Learning techniques and environmental assessment. Five-steps framework have been deployed in a framework and tested on the comparison of biomass pretreatment processes for glucose production. Some scientific articles have been selected thanks to specific keywords in Science Direct and Web of Science. The data architecture and in particular the data analysis allows us to bring data to higher quality such as a material balance check. The approach gives access to a process-impact matrix which is analyzed through Dimensional Reduction methods in order to highlight similar impacts and/or processes.

1. Introduction

According to the French Agency for Environment and Energy Management (ADEME), the Circular Economy (CE) takes into account three areas of action: (1) consumption through consumer demand and behavior, (2) supply and economic actors for whom industrial ecology is an accepted and promising path from the initial design of a territorial area and (3) waste management (Belaud et al., 2019a). These three areas describe as the entire life cycle of a process or a product. Life cycle thinking is used in sustainable models to improve environmental performance while maximizing economic and social benefits. In recent years, several global methods have emerged to design biorefineries in CE models (Grimaud et al., 2017). One of the challenges in biorefineries is to design processes that are as sustainable as possible. The supply chain includes several operational steps, from biomass selection to waste disposal, and goes through various processing steps. Each step in this chain can be described with its material flows (inputs and outputs), operating parameters, energy and economic data. All these data are required so that the environmental assessment could be carried out considering the data diversity and its heterogeneous sources. However, without proper data science tools, it can be difficult to valorize the collected data and better decision-making. Environmental evaluation tools such as Life Cycle Assessment (LCA) would clearly benefit from Data Science approaches in the Big Data context (Belaud et al., 2019b).

The present paper examines the use of Machine Learning (ML) approaches to LCA for the evaluation of biorefinery processes. The main topics are concentrated on the decision-oriented problems of sustainability and eco-design. It structures around the valorization and representativeness of data with the help of data Dimensional Reduction (DR) (Cox and Cox, 2001) and clustering. Despite the importance of ML, a review of the literature revealed that sustainability and environmental assessment are largely not yet part of the popular lexicon of Data Science in action. This study addresses this gap. This paper focuses on developing a framework for decision-making in Process System Engineering by coupling Machine Learning techniques and environmental assessment. It also aims to identify the driving factors of the bio-process that have a major impact on the outputs of LCA.

Paper Received: 18 January 2022; Revised: 5 August 2022; Accepted: 29 July 2022
Please cite this article as: Prioux N., Ouaret R., Belaud J.-P., 2022, Machine Learning Based Framework for Biorefinery Environmental Assessment, Chemical Engineering Transactions, 96, 517-522 DOI:10.3303/CET2296087

2. General approach

2.1 Materials and methods

Big data can be used at various levels of sustainability management. One of the challenges in sustainability management is designing the “best” process in the three areas of sustainability – environmental, economic and social (Santoyo-Castelazo and Azapagic, 2014). A supply chain includes several operational stages, from input choice to waste disposal, and it passes through various transformation stages and upstream/downstream processes. Indicators of sustainability impacts can describe each stage. The main goal of this approach is to analyze the different systems and provide support for group-based decision-making. The presented approach was adapted to any industry by making use of concepts from industry 4.0 and sustainability management. In particular, we retained the Big Data pillar from Industry 4.0 and sustainability assessment from sustainability management. Figure 1 illustrates one path of digital transformation based on the integration of big data into an industry.

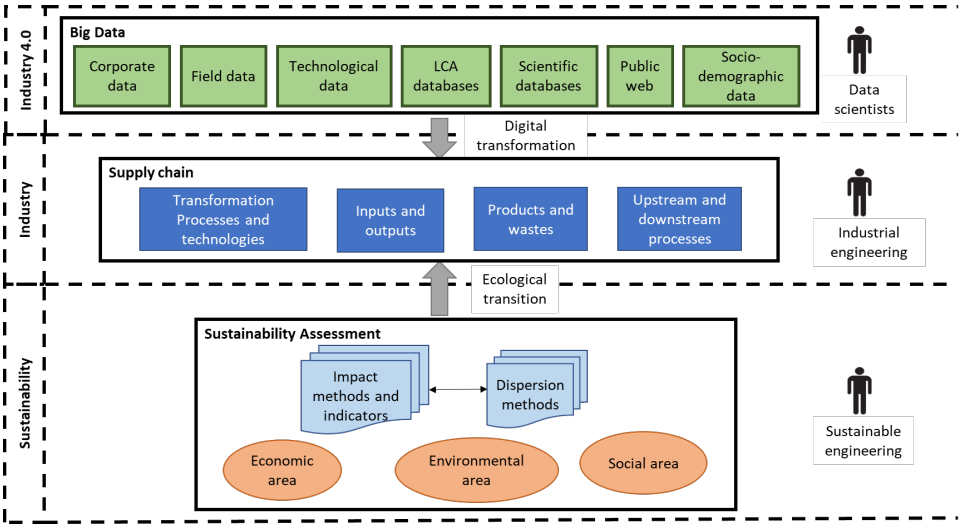

Figure 1: Industry 4.0 and sustainability for industries.

Each supply chain is described by several categories of data that are heterogeneous and can influence the other categories. For example, the type of input can influence the type of transformation processes. All data are importance and influence the social, economic and environmental indicators. Obtaining such data is time-consuming and requires expensive experiments. Alternatively, data can be obtained from scientific publications and other sources, with automatic or manual use of data analysis. It is indeed possible to use these data to obtain foreground data for sustainability analysis, whereas background data are usually available from the LCA database.

2.2 Detailed Framework

Our approach’s goal is the development of a methodological framework centered on intensive data and knowledge extractionfor an economically viable and ecologically responsible design of industrial processes or systems. The framework is divided into five major steps: (i) goal and scope (ii) data architecture (iii) sustainability assessment (iv) results visualization and analysis and (v) decision.

In the first step, the *goal and scope of the study* must be clearly defined. The life cycle thinking being the foundation of the approach, it is recommended to follow the LCA ISO norm (ISO 14040:2006, 2006). A “cradle-to-grave” approach is preferred or a “cradle-to-gate” if the logistics of a value chain are difficult to obtain. After the goal and the scope – the system boundaries - are well defined, the functional unit, the study function, the supply chain, technologies, and transformation processes should also be described.

The *data architecture* step consists of the treatment of the processing of data from scientific papers or private databases. It is directly inspired by the construction of big data architecture and consists of five sub-steps: data collection and extraction, data enrichment and storage, data processing, (raw) data analysis, and (raw) data visualization.

This step can be automatic, semi-automatic, or manual and it uses data technic e.g., machine-learning methods for the (raw) data analysis. These substeps are detailed in Belaud et al. (2019). The last two sub-steps – analysis and visualization – can take benefit from ML methods like we describe in the fourth step.
The third step consists on *sustainability assessment,* which is divided into two parts. First, life cycle inventory lists and quantifies every input and output required for the sustainability assessment. There are two categories, the foreground data which is the process data from the previous step and the background data are available into specific LCA database. Then, sustainability assessment involves choosing the impact methods, the indicators, and the dispersion methods in accordance with each area of sustainability management. A [Process-Impacts] Matrix is the result of this step but this matrix is difficult to analyze.
The fourth step, *results visualization and analysis*, summarizes the analysis of the impact-process matrix. The step provides the methods derived from ML to help this analysis. Based on statistical literature, a combination of traditional DR and unsupervised clustering techniques was chosen to extract information from the impacts. More precisely, this hybrid approach is based on the Multi-Dimensional Scaling (MDS) using the Canberra distance and k-means (Lance and Williams, 1967). The objective is to search for "hidden" structures in multidimensional data and to help interpret the area of clustered midpoints in the assessment matrix. The advantage of this approach is that data-based methods require very little knowledge of processes to perform. Figure 2 summarizes the treatment of the [Process-Impacts] matrix.

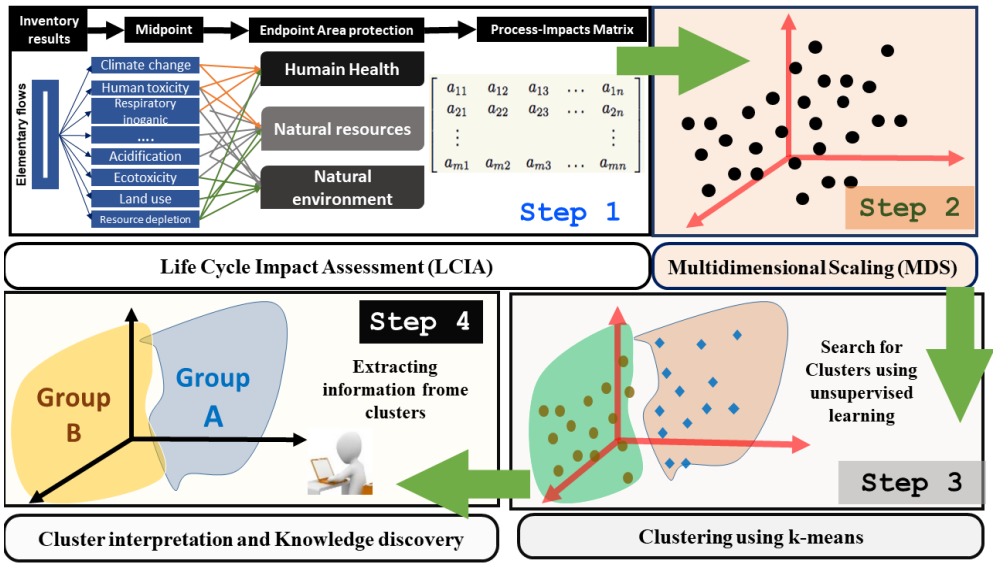

Figure 2: Treatment of the [Process-Impacts] Matrix (Step 4)

For the final step, the researcher can use the previous step – the analysis and the visualization of clusters - as a basis for his decision. This decision can be made by the researcher himself or by a group composed of different engineers/researchers from different fields

3. Case study

The approach is tested on the comparison of biomass pretreatment processes for glucose production. Only the environmental area is considered.

3.1 Goal and scope step

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.

3.2 Data architecture step

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. For this paper, no raw data analysis and simulation are carried out.

3.3 Environmental assessment

For the life cycle inventory, the process data comes from the end of the previous step – the cleaned process data - and the background data from EcoInvent v2.2. Then an attributional LCA method is applied: the ReCiPe 2016 method. The calculation comes from SimaPro® a LCA software. The environmental assessment evaluates 17 "midpoint" impacts. The result of this step is a [Process-Impacts] matrix of dimensions 17x17.

3.4 Results visualization and analysis

The previous matrix is then analyzed by MDS. The interpretation of an MDS result is simple the closer the objects are in the scatter plot, the more similar they are. That is the projected points are arranged in such a way that the grouped ones (small geometrical dissimilarity between them) will reflect original closeness in the data. After that, a clustering algorithm has been applied to the MDS projection to highlight the most similar objects (Impacts - Process) as illustrated in Figure 2. There are two types of k-means clustering possible: one based on impacts distance matrix and one based on the process distance matrix. The two-dimensional of MDS results of projected impacts (17 impacts) is shown in Figure 3. The abbreviation of the impacts is introduced to facilitate the visualization (Table 1). It presents 4 sub-figures of the first 4 dimensions with the most significant combinations. For example, in the first figure (top left), we have represented the projection of the 17 impacts on the first two dimensions, which represent a total variance of 45%. The percentage of explained variance for the first four components is 70%. The visualization of the four dimensions shows the same three groups and we can clearly distinguish three clusters using k-means:

- Group 1: Almost all impacts that concern chemical pollution of soil and water are found in this group. The exemption is the marine ecotoxicity found in group 2.
- Group 3: This group mainly includes impacts related to land use and land transformation.
- Group 2: This group forms three sub-clusters with superposed points (from the 2-D perspective). This suggests that these points are highly similar based on the Canberra distance. Here, we find a group quite heterogeneous where impacts not presented in groups 1 and 3 are found. The marine ecotoxicity expected rather in group 1 is found in this group.

Table 1: Abbreviations of impacts

Impacts	Abbreviations	Impacts	Abbreviations
Climate change Human Health	CCHH	Terrestrial ecotoxicity	TecoX
Ozone depletion	OD	Freshwater ecotoxicity	FrEco
Human toxicity	HT	Marine ecotoxicity	MaEco
Photochemical oxidant formation	Pohto_ChOx	Agricultural land occupation	AgLOcc
Particulate matter formation	PM	Urban land occupation	UrbLOcc
Ionising radiation	IR	Natural land transformation	NLTran
Climate change Ecosystems	CCE	Metal depletion	MeDe
Terrestrial acidification	TA	Fossil depletion	FossDe
Freshwater eutrophication	FrEu		

The two-dimensional of MDS results of process projected reveal results quite similar to those obtained by using MDS on the impacts matrix. In this case, the percentage of explained variance for the first four components is 98%, which is an excellent representation in lower-dimensional spaces. Three distinct groups of processes can be identified. Very tight and separate clusters appear in the process data, which may suggest that each cluster is a domain or subdomain that needs to be analyzed individually. A group referring to processes whose pre-treatments are purely mechanical. Going back to the impacts, we find that these two pre-treatments have a very significant impact on the depletion of fossils compared to the others.

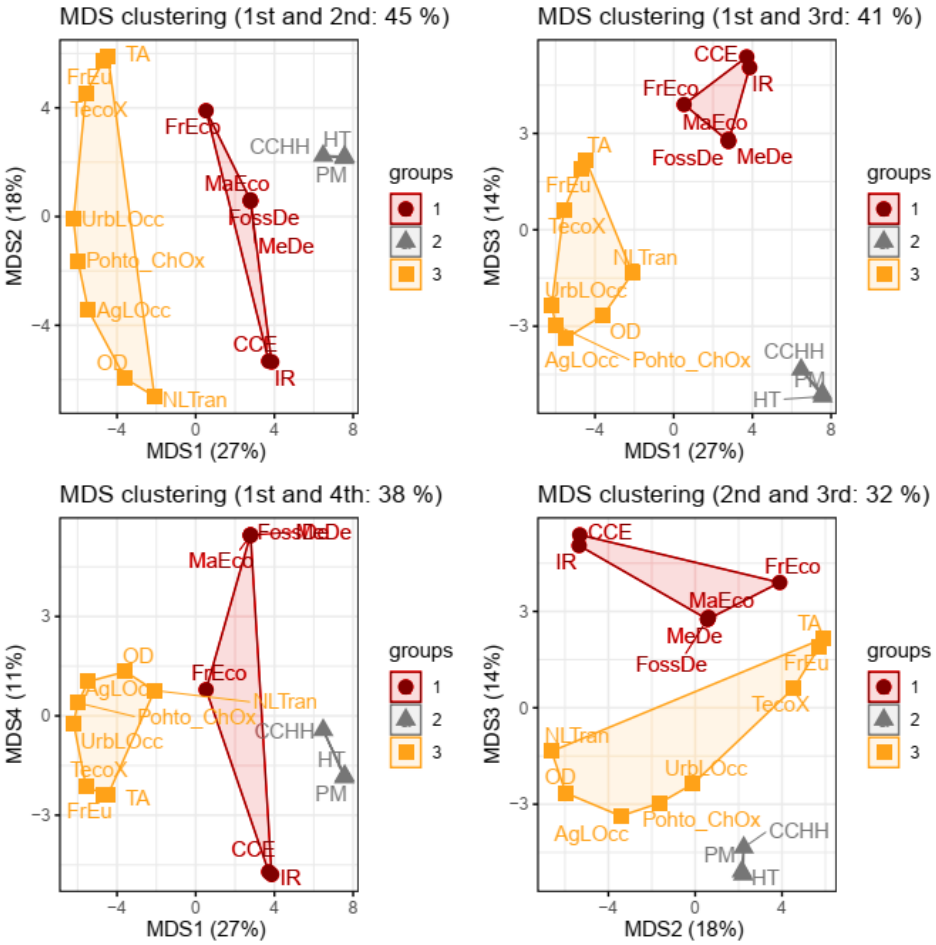

Figure 3: Scatter plot of MDS projections (two dimensions) and k-means clustering based on impacts distance matrix. Percentage of explained information for the first four components is 70%.

3.5 Decision

The hybrid ML techniques, here MDS and k-means clustering allows the researcher, engineer or decision maker to make rapid clusters on processes without loss of information. Following the goal of the study, the choice could be made on a process of a specific group. The specialists can also make a loop of the framework for a specific group to detail the result.

4. Conclusions

This paper proposes a generic and practical approach for the researcher or research and development engineer in the use of Machine Learning in the field of sustainability within the Big Data context**.** The improvement of the traditional LCA method by coupling the tools of (big) data science and artificial intelligence algorithms allows a different discussion of environmental impacts. **I**ndeed, on the one hand, data science tools allow extracting and collecting data directly from scientific articles. On the other hand, the MDS can simplify the discussion of environmental impacts.

Composed of five steps, the approach is shown as a decision help in a pre-study. It is designed to save time and money by including no experiments and using public scientific data as a database. After structuring and processing process data from scientific literature, the LCA step give an environmental impact-process matrix which can be analyzed by MDS method. In the case study, the comparison of biomass pretreatment processes for glucose production, this MDS clustering methods highlight major findings: (i) a group includes impacts related to land use, and land transformation is detected, and (ii), a cluster of all impacts related to chemical pollution of soils and water.

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 approach does not integrate the change of scale required to implement a semi-industrial pilot, especially if the process becomes semi-continuous
- The abundance and the quality of the data are not sufficient for these new technological processes.

The most ambitious perspective is the automation of the database enrichment phase. A further research objective will include the comparison of several ML clustering tools. Other points for progress are to reconsider the functional unit, the global environmental assessment strategy by integrating the upstream agricultural phase (consequential LCA, system allocation and system extension policy) and considering the global supply chain according to a dynamic analysis, spatial, or even temporal.

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, J.-P., Adoue, C., Vialle, C., Chorro, A., Sablayrolles, C., 2019a. A circular economy and industrial ecology toolbox for developing an eco-industrial park: perspectives from French policy. Clean Technologies and Environmental Policy 21, 967–985. https://doi.org/10.1007/s10098-019-01677-1

Belaud, J.-P., Prioux, N., Vialle, C., Sablayrolles, C., 2019b. Big data for agri-food 4.0: Application to sustainability management for by-products supply chain. Computers in Industry 111, 41–50. https://doi.org/10.1016/j.compind.2019.06.006

Cox, T.F., Cox, M.A.A., 2001. Multidimensional scaling. Chapman & Hall/CRC, Boca Raton.

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

ISO 14040:2006, 2006. Environmental management - Life cycle Assessment - Principles and Framework. International Organization for Standardization, Geneva, Switzerland.

Lance, G.N., Williams, W.T., 1967. Mixed-data classificatory programs i - agglomerative systems. Australian Computer Journal 1, 15–20.

Lousteau-Cazalet, C., Barakat, A., Belaud, J.-P., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C., Vialle, C., 2016. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. Computers and Electronics in Agriculture 127, 351–367. https://doi.org/10.1016/j.compag.2016.06.020

Santoyo-Castelazo, E., Azapagic, A., 2014. Sustainability assessment of energy systems: integrating environmental, economic and social aspects. Journal of Cleaner Production 80, 119–138. https://doi.org/10.1016/j.jclepro.2014.05.061