Environmental assessment coupled with machine learning for circular economy

Nancy Prioux, Rachid Ouaret, Gilles Hetreux, Jean-Pierre Belaud

ENVIRONMENTAL ASSESSMENT COUPLED WITH MACHINE LEARNING FOR CIRCULAR ECONOMY

N. PRIOUX, R. OUARET, G. HETREUX and J.-P. BELAUD
Laboratoire de Génie Chimique, Université de Toulouse, CNRS, Toulouse, France
nancy.prioux@ensiacet.fr

ABSTRACT: The circular economy and its various recirculation loops have become a major study subject over recent years, particularly in the field of agriculture, which is a significant source of waste production. There have been several studies focused on transforming agricultural lignocellulosic waste with "sustainable" processes: economically viable, socially accepted, and environmentally friendly. Thanks to "*life cycle thinking*", it is possible to assess these potential environmental impacts. However, these environmental analyses generally require a massive volume of specific data, the collection of which can be time-consuming and tedious, or impossible to practice. On the other hand, scientific articles describing the processes for the valorization of agricultural by-products are intriguing but rarely exploited sources of data. In this paper, a hybridization of data science techniques and environmental analysis proposed to improve Life Cycle Analysis (LCA) thanks to Machine Learning (ML). ML part of the proposed approach is based on unsupervised learning, which is composed by two methods: dimension reduction using the Multidimensional Scaling and clustering technique using k-means. Composed by five steps and dedicated to researchers or R&D engineers, the approach is oriented towards offering decision on technologies and processes for waste to energy in early eco-design step. The case study in the domain of pre-treatment processes for corn stover and rice straw is detailed. The results show that all impacts that concern chemical pollution of soil and water are found in the same cluster. Other impacts are detected in the same cluster which are related to the land use and the land transformation. In the same vein, two purely mechanical pre-treatments has been identified and grouped by Multidimensional Scaling and k-means.

Keywords: Agricultural Waste, Life Cycle Assessment, Artificial Intelligence, Data Science, Clustering, Biorefinery

1. INTRODUCTION AND SCIENTIFIC CONTEXT

The idea of sustainable development is a complex problem applying in almost all areas of life, from social and health to economic development and environmental assessment (WCED 1987). In recent years, policymaking increasingly involves action plans for climate, sustainable development or environment like Kyoto Protocol or Paris climate agreement. In December 2019 at the COP25 Climate Summit in Madrid Spain (COP25), the European Green Deal is presented by the European Commission and is accepted in January 2020 (European Commission 2019). The European Green Deal plan includes several actions aimed at stopping climate change and reducing the level of pollution emitted into the atmosphere. One of the main action plan named "*Circular Economy Action Plan*" proposes to use the Circular Economy (CE) model to simulate the use of sustainable models. This plan completes the 2015's first Circular Economy Action Plan whose all 54 actions under the plan have been delivered or are being implemented. Ellen Macarthur Foundation (2015) defines the circular economy as "*be one that is restorative and regenerative by design and aims to keep products, components, and materials at their highest utility and value at all times, distinguishing between technical and biological cycles*". This new concept of sustainable development could reconcile some environmental, economic, and social aspects. A review paper proposed by Ghisellini et al. (2016) describes the origins, principles and the limitations of CE models. Following the first plan in 2015, the French government recommended the SNTEDD (National Ecological Transition Strategy for Sustainable Development) which consists of nine areas, including the CE. According to the French Agency for Environment and Energy Management (ADEME), the CE takes into account three areas of action: (1) consumption through consumer demand and behaviour, (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 policy (Belaud et al. 2019a). These three areas represent the entire life cycle of a product, service, or a process. To achieve sustainable models, life cycle thinking can help improve environmental performance while maximizing economic and social benefits. Several global methods have emerged to design waste recycling processes that fit into the circular economy (Grimaud et al. 2017). Agriculture is a particular area in which CE and life-cycle thinking have developed over the last decades. Singh et al. (2021) highlights the connections between CE and various major themes and notably the links between CE and LCA, biomass, biorefinery, bioenergy and waste management.

In France, the agri-food industry produces about 2.6 million tons of organic waste per year, a figure that is constantly increasing (Barry 2020). This trend is accompanied by the projected increase in the world's population. At the same time, human activities are reducing the amount of land available for agriculture, which inevitably has an impact on farming systems. For some people, new agricultural technologies that facilitate sustainable intensified agriculture seem to be the best solution for the future (Garnett et al. 2013). However, such intensification in agriculture could result in more waste of products and resources (West et al. 2014). According to Horton et al. (2016), a major challenge in achieving sustainability

in agriculture is categorizing wastes. They distinguished them into two categories: wastes generated by inputs, such as fertilizers remains or used water, and wastes generated from following treatment processes. The latter, composed of huge amounts of lignocellulosic by-products, comes mainly from incomplete conversion of biomass or raw materials processing in the supply chain. Lignocellulosic biomass is one of the most abundant and least expensive renewable resources on Earth. The production of biomaterials, biomolecules and bioenergy often relies on the bioconversion of it, where it is enzymatically hydrolysed to produce glucose. It is also possible to produce different biomaterials to replace plastic material, such as composite beads based on olive pomace (Lissaneddine et al. 2021). Lignocellulosic biomass consists of four main elements: lignin, cellulose, hemicellulose, and phenolic acids. Only cellulose and hemicellulose can be hydrolysed to generate glucose. Although lignocellulosic biomass is a renewable resource, transformation processes must be sustainable to participate in sustainable development. For this reason, more and more agri-food processes have been incorporated with various sustainability assessments (Food SCP Round Table European Commission 2012; Raymond 2012). In order to generate a good yield of glucose, it is important to pre-treat the biomass before hydrolyzing it. Over the past 30 years, numerous pre-treatment processes have been studied and published (Davis et al. 2017). Various factors were applied to evaluate and compare the performance, efficiency, or environmental influences of these processes (Joglekar et al. 2019). Environmental factors, energy consumption and energy efficiency were considered as 3 classical factors (Zhu et al. 2010; Barakat et al. 2013). However, there is a lack of criteria to guide the choice between all these processes. The use of environmental, economic, and social assessment in a CE context is a good way to guide this choice. In this paper, only the environmental dimension is studied, even though it is entirely possible to complete our approach through economic and social manners. Our choice for the environmental analysis is based on the Life Cycle Assessment (LCA) method published by the International Organization for Standardization (ISO 14040:2006 2006).

From the literature on LCA, it appears that the most commonly used approaches rely on knowledge-based approaches and multi-criteria decision-making. The Life Cycle Impact Assessment (LCIA) knowledge acquisition is based on several experts from different backgrounds (ecologist, chemist, lab technician, epidemiologist, ...). Most previous LCA studies and methods for assessing sustainability are based on environmental, social, and economic dimensions (Svensson et al., 2018; Zhao et al., 2019). For measuring country sustainability performance, Tan et al. (2017) used an adaptive neuro-fuzzy inference system (ANFIS) approach. The proposed approach is effective to measure the countries' sustainability performance, however, the results of the ANFIS method are strongly dependent on appropriate training data. In this regard, we have recently seen an increase in the fuzzy-set theory in the sustainability assessment field. Many researchers have applied this approach for country sustainability assessment. The advantage of the fuzzy-set and fuzzy inference system is it can emulate the behaviour of skilled humans and handle the multidimensional complexity of vague situations. Several studies have been conducted using other Machine Learning (ML) algorithms. For example, Kouchaki-Penchah et al. (2017) and Zhao et al. (2019) have combined Data Envelope Analysis (DEA) with LCA. Such hybrid systems have been tested for determining the energy efficiency, for measuring the efficiency of a sustainable development system by tracking economic, environmental, and social dimensions.

Some studies have examined the spatiotemporal patterns of Life-Cycle Environmental (LCE) based on multi-year assessment over a large geographic region, and improving the spatial resolution of LCE (Lee et al. 2020). In the same context, Romeiko et al. (2020) utilized boosted regression tree models to identify the top influential factors among soil, climate, and farming practices, which drive the spatial and temporal heterogeneity of life cycle environmental impacts. The results of this study showed that soil organic content and nitrogen application rate were the top influencing factors for life cycle eutrophication (EU) and acidification (AD). The top influencing factors for life cycle global warming (GWP) impacts were soil texture, nitrogen application rates, and temperature in March. By using Boosted Regression Tree (BRT) model with Gaussian distribution, the same team researchers have conducted another study based on the prediction of LCE impacts of corn production under climate variability and scenarios. To the best of our knowledge, many methods for LCIA or for general sustainability assessment are "supervised", meaning that learning algorithms are performed from input-output data (Abdella et al. 2020). However, in many cases, one needs to extract latent data structures based on the observations and the mixture nature of the endpoint area protection. For this reason, our research relies on the use of Dimension Reduction (DR) techniques for the assessment of sustainability through a set of impacts (Climate change, Human Health, Ozone depletion, Human toxicity, ...) and underlying process. Several strategies have been explored to achieve this goal using "Unsupervised" learning which includes dimension reduction and clustering.

The purpose of this paper is to present an approach to aid in the analysis and comparison of different pre-treatment processes for lignocellulosic biomass and different types of biomass in the context of the circular economy. The main contribution of this study is the development of an approach that suggests the enhancement of the traditional LCA method by coupling methods from ML. This enhancement can be found in two main items: (i) the use of experimental opendata extracted from scientific web engines and (ii) the inclusion of unsupervised ML for the interpretation and analysis of environmental impacts. To date, few research studies have been conducted using big data technologies for LCA. In particular, there are studies to complete the background data, necessary for LCA, or to adapt LCA to technological developments (Cooper et al. 2013; Bhinge et al. 2015). However, in our knowledge, no study uses experimental data from public web, such as data from scientific article or from scientific open database, to complete the foreground database

required in a LCA. This point is developed in the step 2 of our approach, and to this end, the proposed approach constitutes a contribution to the classical literature (section 2.2). The use of unsupervised learning for LCA is motivated to detect the environmental fingerprint of lignocellulosic biomass processes. More precisely, this study lies in the characterization of different environmental impacts into clusters to help analysis, interpretation and decision. This point corresponds to the step 4 of our approach (section 2.2 and result section).

This approach is intended for researchers or research and development engineers who would like to make the first screening in pre-treatment processes or biomasses using data contained in the scientific literature. It can be deployed during a preliminary study phase and be used to assist an initial decision for the development of a laboratory or semi-industrial pilot. After an introduction to the description of data availability, section 2 outlines this general approach. Section 3 will use this approach to deal with the study of different pre-treatments processes of rice straw and corn stover and then provides the results and interpretations.

2. METHODOLOGICAL APPROACH

2.1. LCA and ML methods

Data science can be used at different levels of sustainability management. One of the challenges in valorising by-products in the agricultural supply chain is to design a process that is 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 the chain can be described with its inputs and outputs, as well as its energy and economic data. All these data are required so that the LCA could be carried out taking into account the data diversity and its heterogeneous sources. Various types of data from heterogeneous sources are required for environmental analysis. ML would provide appropriate methods and technologies for this analysis. The main objective of this approach is to analyse diverse technological processes and to provide decision supports by analysing the results.

2.1.1. LCA method

LCA is a tool for assessing the potential environmental impact of a product or service through its entire life cycle. The life cycle of a product/service can be broken down into several stages which starts from the designing of the product to various stages of processing and utilization, and ends with the disposal or recycling stage of the product. LCA involves four steps: defining the objective and scope of the study, conducting a life cycle inventory, conducting a LCIA, and, at last, analysing and interpreting the results.

The first step is to define the objective, system boundaries, and the functional unit of the life cycle. The functional unit must be well defined so that it will not disturb the outcome of the LCA (Burgess and Brennan 2001).

The Life Cycle Inventory (LCI), which is the second step, aims at collecting data about the quantities of pollutants emitted and resources extracted throughout the life cycle. This inventory concerns two kinds of data: the background data and the foreground data. The foreground data represents data on processes of interest to decision-makers, while the background data includes all the other processes data that is related to processes of interest (Clift et al. 1998; Elghali et al. 2007). Data for these inventories can be obtained directly, through on-site measurements (primary data), or indirectly, from published scientific papers, models, and databases (secondary data). The foreground system is generally based on primary data, while the background system is based on secondary data sources (Guinée 2002).

The third step is the LCIA, during which numerous data of pollutants and resources collected at the LCI are calculated through specific methods to evaluate their environmental impacts (Suh and Huppes 2005).

The last step is the interpretation of the results, which involves identifying important issues based on the results of the LCI and LCIA, assessing their sensitivities, checking for their consistency and completeness, and formulating a report with conclusions, recommendations, as well as limitations. This last stage of LCA is "delicate" for novices and, sometimes, even experts. It may be, therefore, beneficial to enrich the original LCA method with other techniques to assist researchers or engineers during their analysis and interpretation.

In our approach, we use only secondary data. Foreground data are taken from the public web through scientific articles (Web of Science and Science Direct) to constitute an unstructured database, whereas background data are available in structured databases (public or private) such as EcoInvent (Frischknecht et al. 2005). To analyse this data and structured them, engineering and ML techniques for clustering were used.

2.1.2. Multidimensional Scaling method

In this article, we use Multidimensional Scaling (MDS), also called Principal Coordinates Analysis (PCoA) which is a method for visualizing (de)similarity between objects in a reduced dimensional space. It is designed to understand the proximity and opposition structures. Starting from information about the mutual similarities of n objects, often with the

similarity matrix where $\Delta = \delta(i,j)_{0\le i\le n;0\le j\le n}$, we look for a configuration of n points (R^2) which would be like that if the object i resembles to j more than the object l to k, we have $\delta(i,j) < \delta(l,k)$.

For any distance matrix of size $n \times n$, the MDS allows us to find a set of n points marked by their coordinates whose similarity matrix is equal or very close according to the data.

Let x_r $(r = 1,\dots,n)$ be the coordinates of n points in a p dimensional Euclidean space where $x_r = (x_{r1}, x_{r2},\dots,x_{rp})^{\top}$ and $[B]_{rs} = b_{rs} = x_r^{\top}x_s$. From the Euclidean distance $\Delta = \delta_{rs}$, and a matrix A of elements $[A]_{rs} = a_{rs} = -\frac{1}{2}\delta_{rs}^2$, which is deducted from the decomposition $x_r^{\top}x_s$, the matrix B is obtained using the following relation :

$$B = HAH \quad (1)$$

where H is the centring matrix: $H = I - n^{-1}I \cdot I^{\top}$, where I , a $n \times n$ identity matrix and$I = (1,1,\dots,1)^{\top}$, a vector of n ones. The elements of $b_{rs} = a_{rs} - a_{r\bullet} - a_{\bullet s} + a_{\bullet\bullet}$, where ::

$$a_{r\bullet} = n^{-1}\sum_{s} a_{rs}\,,$$

$$a_{\bullet s} = n^{-1}\sum_{r} a_{rs}\,, and$$

$$a_{\bullet\bullet} = n^{-2}\sum_{r}\sum_{s} a_{rs}.$$

The algorithmic procedure of MDS can be summarized, as illustrated in (Cox and Cox 2001) by the following steps:

1. Obtain the matrix $\Delta = \delta_{rs}$ of dissimilarities.
2. Find the matrix $A = [-\frac{1}{2}\delta_{rs}^2]$.
3. Find the matrix $B = [a_{rs} - a_{r\bullet} - a_{\bullet s} + a_{\bullet\bullet}]$.
4. Find the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$ and the eigenvectors $v_1, v_2, \dots, v_{n-1}$. If B is semi-defined positive (some eigenvalues are negative), either (i) ignore the negative values and continue, or (ii) add an appropriate constant c to the (dis)similarities
5. Choose an appropriate size number l, possibly using $\frac{\sum_1^l \lambda_i}{\sum (positive\ eigenvalues)}$.
6. The coordinates of the *n* points in Euclidean dimension space l are given by
   $x_{ri} = v_{ir}$ $(r = 1,\dots,n;\ i = 1,\dots,l)$.

2.2. Approach

Using a general architecture based on sustainability principles (Belaud et al. 2019b), our approach is composed of five steps (Figure 1): (1) goal and boundaries, (2) data architecture, (3) the environmental assessment, (4) results visualization and analysis, and (5) decision. Each step has sub-steps and the transition from one to the other can be done in either direction through feedback loops. It is recommended to iterate to consolidate the results and, subsequently, the choices resulted from the interpretations.

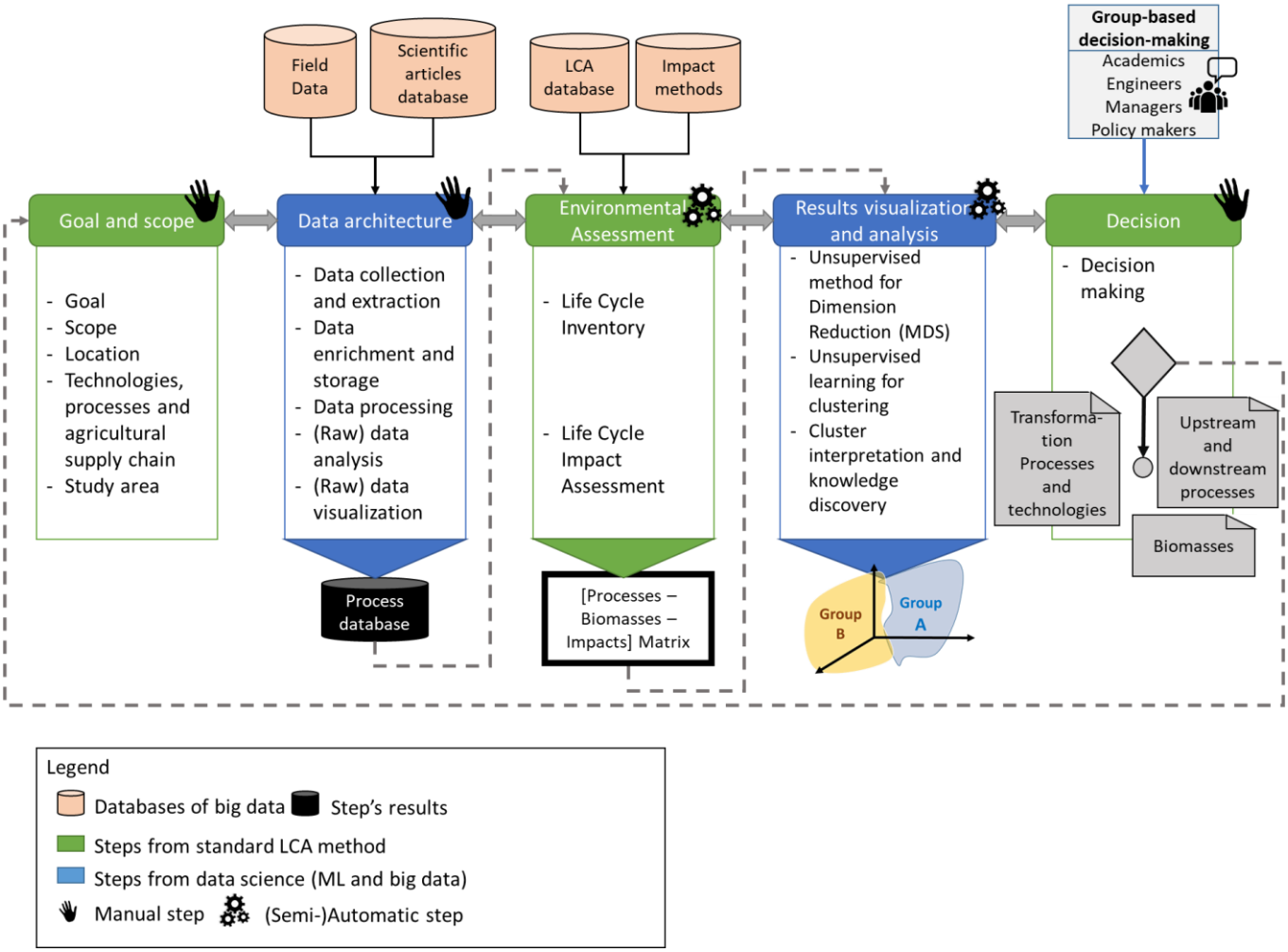

Figure 1: Five main steps of the approach

2.2.1. Goal and scope

In this step, the goal of the study case and the boundaries of the system must be clearly defined. It is possible to come back to this step if a blur or a problem is noticed in the following steps. Life cycle thinking is recommended. This thinking encourages a "cradle to grave" or "cradle to gate" approach. In the CE model, the part of the life cycle in which the product is used is a key element in moving towards the ecological transition. “Cradle-to-Gate” approaches are often preferred because integrating downstream elements into sustainability analyses can be tedious and difficult. Especially, scientists and engineers often find it impossible to consider end-user or consumer behaviour in their models. The system boundary has a significant effect on subsequent evaluations. For example, it needs to be clarified whether the upstream biomass supply chain is considered. Once the objective and scope have been properly defined, the supply chain, technologies and transformation processes should also be described. This description should be as complete as possible concerning the operations of the process, the location of the study, the various inputs and outputs, and the type of energy used. These details ensure the relevancy of the collected data.

2.2.2. Data architecture

The data architecture is directly inspired by the construction of massive 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.

Collecting and extracting data from unstructured databases requires special tools, such as data queries (SQL queries) or Online Analytical Processing (OLAP). This sub-step is more complicated for unstructured data collected from the web. The meta-data associated with these web pages can be used for their classification and to provide access to their content. For example, the definition of the MARC (MAchine-Readable Cataloging) format in the early 1960s standardized metadata in documentary resources. With the development of Linked Web, especially the Resource Description Framework (RDF,

a W3C standard), it is possible to use SPARQL Protocol and RDF Query Language to request the RDF. The extraction process generates a structured table which differs according to the format of the web pages: API, HTML, or pdf. The extraction process can be automatic, semi-automatic or manual. Data can be extracted automatically from web pages with an API format. However, extracting relevant data from scientific articles requires a reading guide, which is generated from ontologies and pre-processing analysis of experts. This type of extraction is therefore semi-automatic.

For data enrichment and storage, the extracted data is stored in Relational Database Management Systems (RDBMS). Enrichment is the process of adding data to the DBMS. This data comes from experts or pre-processing models. The models are empiric numerical simulations of unit operation type, thermodynamics, and energy for the control of flows, transformations, and transfers.

Data processing involves cleaning, adding, and deleting data for volume and value management. After this step, a second, more accurate and more accessible database can be generated. Keeping this second database saves time for the following sub-steps and avoids misinterpretations during the data analysis sub-step and the environmental analysis step. The data analysis sub-step depends on the objective, data domains and decision-makers. Different types of analysis are possible: descriptive analysis (what happened?), diagnostic analysis (why did it happen?), predictive analysis (what is going to happen?) and prescriptive analysis (how can we make it happen?). Some of these methods are visual, so data visualization can be included in the data analysis sub-step. Data visualization can also be achieved by plotting the raw data in the form of a simple or interactive graph.

2.2.3. Environmental assessment

In this step, the LCI and the LCIA is performed. It is very important to follow the steps recommended by ISO (ISO 14040:2006 2006) and to use the data from the previous step for the process data. The background data is taken directly from the EcoInvent database. Once the inventory (LCI) is completed, one (or more) impact calculation method(s) must be chosen under Step 1 (The purpose and the boundary of the study). It is possible to choose several impact calculation methods to compare the results, but this may hinder the final analysis and the choice of process or biomass for the researcher in his or her first analysis.

2.2.4. Results visualization and analysis

It is in this stage where we bring the ML method into use to help analyse environmental impacts which is the focus of this paper. At the end of the previous step, the result obtained is a matrix with the processes, biomasses, and environmental impacts, which is difficult to analyse for a non-expert in LCA. From the statistical literature, this step combines traditional techniques for dimension reduction and unsupervised clustering to extract knowledge about life cycle impact assessment (LCIA). More specifically, the hybrid approach is based on Multidimensional Scaling (MDS) using Canberra distance (Lance and Williams 1966) and $k-$means. MDS is an algorithm that transforms a distance matrix into a set of coordinates such that the distances derived from these coordinates approximate as well as possible the original distances. In this work, we used Canberra distance (a weighted version of the Manhattan distance) for constructing the proximity matrix and then the data is mapped on a lower-dimensional (two or dimensions) spatial representation. These methods reduce the dimensional space (the variable set) while preserving the maximum amount of information. The goal is to seek "*hidden*" structures in the multidimensional data and to help to interpret the grouped endpoint area of the LCIA assessment matrix. The benefit of such approach is that the data-driven methods require minimal process knowledge to perform this task.

Figure 2 outlines the data-driven processing for LCIA. In the proposed method, first, the process-impacts matrix is used as input (similarity matrix) for learning from data (Part A). Second, DR techniques project the raw process data to a lower-dimensional space (2 or 3) (Part B). There are several statistical methods to achieve this mapping such that uninformative variance in the data is discarded (Burges 2010). After projection raw data by a DR technique, the clustering approach is then applied to consider similar life cycle impacts as well as in the process within the lower-dimensional space (Part C). Finally, the user (expert) analyses the grouped data in clusters to relate them to meaningful process/impacts (Part D). This last cluster assignment and extracting information step is called knowledge discovery. An advantage of the workflow is that it is relatively simple to use because each DR and clustering combination requires the specification of only one or two parameters and techniques. The idea behind the proposed workflow is motivated by the fact that sustainability assessment requires both qualitative and quantitative criteria.

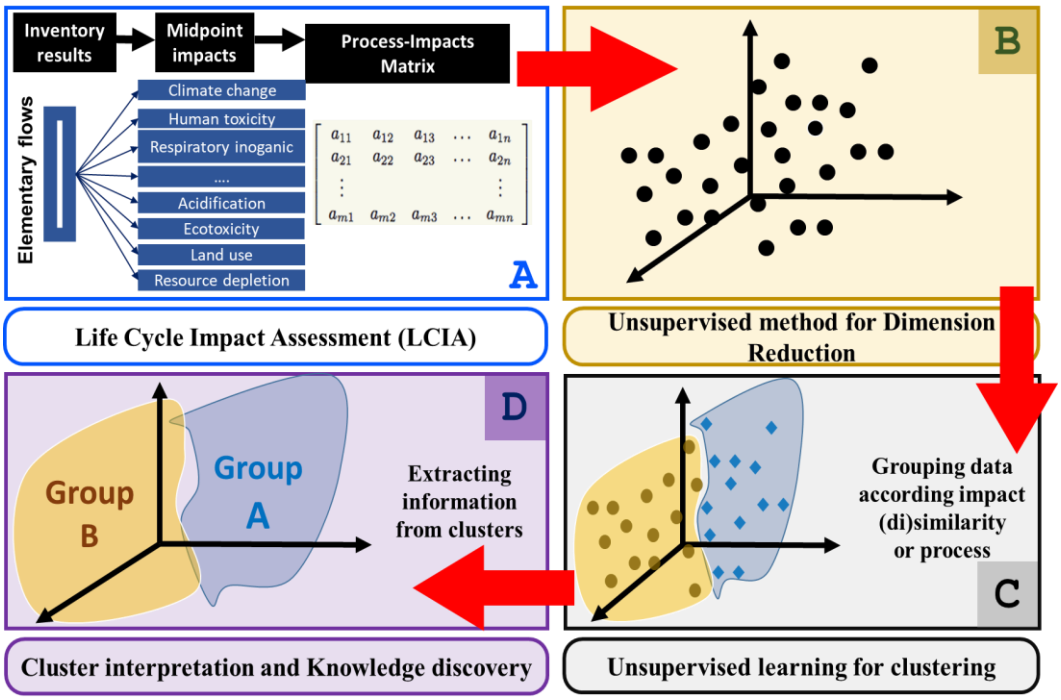

Figure 2: Schematic of data driven processing for LCIA

2.2.5. Decision making

The visualization of data clusters from the previous step can help the researcher with his decision-making. This decision can be made by the researcher himself or by a group composed of different engineers/researchers from different fields.

3. RESULTS AND DISCUSSION

In this section, the first results of our approach are presented with a case study comparing different pretreatment processes of two biomasses: corn stalk and rice straw. A software programmed on Excel supports the data architecture and the environmental steps as well as on R for results analysis and visualization step. The software part is shown in Figure 3. The internal software has been verified with:

- well-establisted ProSim for the process simulation part. The process simulation is used to raw data processing (step 2 iii), especially for the verification of mass balances. An article whose material balance could not be verified is deleted from the database.
- SimaPro for the environmental impact simulation part (step 3)

For the origins of data, foreground data come from data collection of our approach - they are called process data - and background data come from. EcoInvent Database. This study concerns mainly the foreground data, the background data are not modified or analysed.

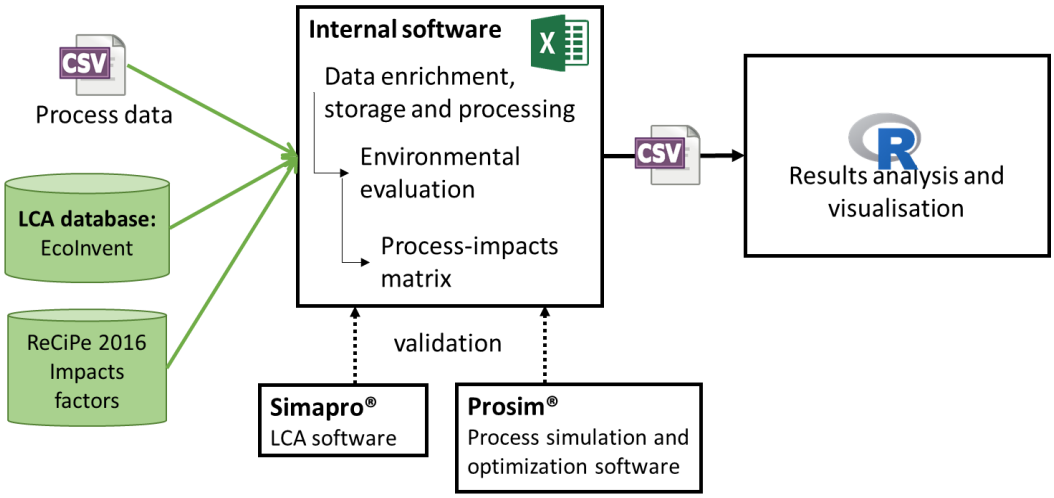

Figure 3: Software tools scheme

3.1. Study case first results

The purpose of the study is to help the researcher to select a process and/or a biomass for the production of glucose. The study has a cradle-to-grave approach i.e. the boundaries of the study range from biomass to the enzymatic hydrolysis. Biomass is here considered as a waste from agriculture that has no impact - the impacts are attributed to the final product of agriculture (corn and rice). Besides, the biorefinery is considered to be relatively close to the site, and therefore the transport stage is negligible. The function of the system is, therefore “producing glucose” and the functional unit is the "production of 1 kg of glucose". All results can therefore be expressed in terms of the functional unit. Figure 4 shows a instance pre-treatment process from a Liu et al. (2013)’s paper.

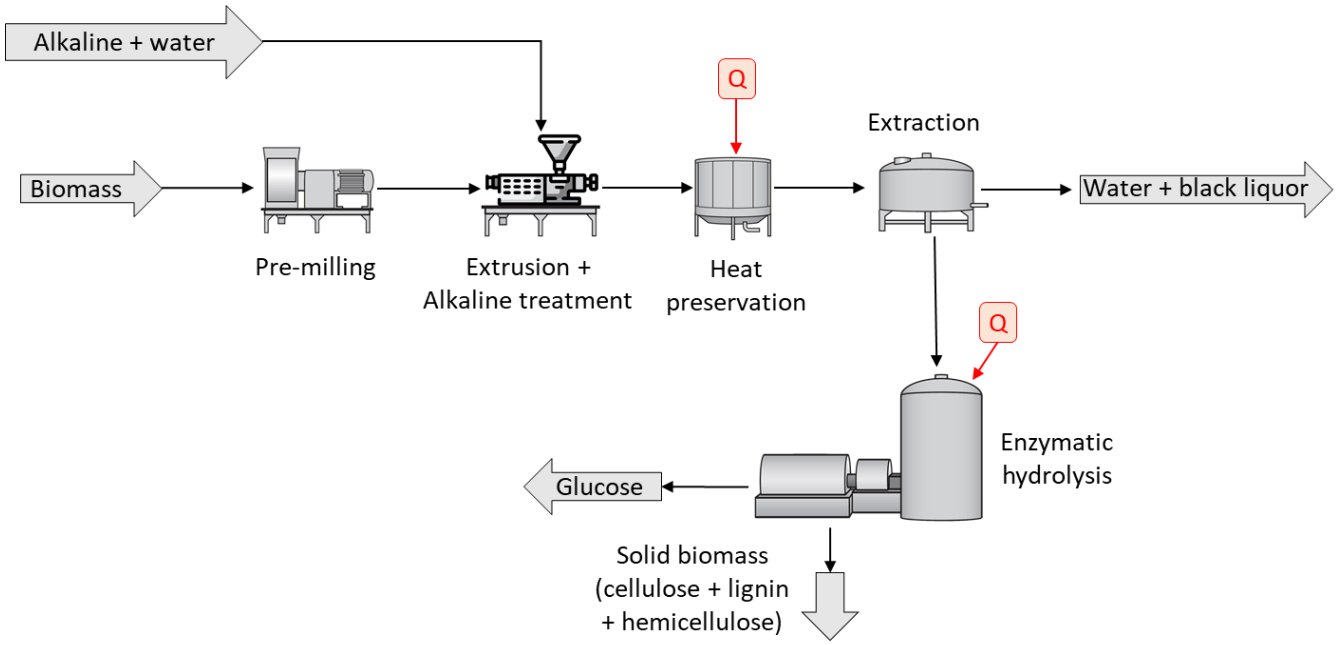

Figure 4: Pretreatment process example adapted from (Liu et al. 2013)

Now, the step 2 of the approach - data architecture - was developed. A more detailed step can be found in Belaud et al. (2021). The selection of articles describing the processes was made by process experts on corn stove and rice straw thanks to keywords: “rice straw”, “corn stover”, “treatment”,“hydrolyze” and “milling” in scientific databases, such as Web of Science or Science Direct. Eighteen papers were initially selected and stored by topics (different type of process). The data from the papers were then extracted and stored in CSV files before being passed into our software (cf. Figure 3). This software developed on Excel did a first cleaning of the data using process simulations to calculate and verify the mass balance. Thirteen articles were selected because they had either too much missing data for verification or too many inconsistencies. The raw data analysis and visualisation are not used in this case study and they will be the subject of future research.

Then the environmental assessment is performed on the remaining process data. The LCA method used here is ReCiPe 2016 (Huijbregts et al. 2017) , the background database is EcoInvent and the processed data comes from the previous steps. The result of this step is a matrix of 18 mid-point impacts and 13 processes. This matrix is then analyzed by multidimensional scaling.

The interpretation of an MDS result is the same as for any other dimension reduction; objects that are closer together on the scatter plot are more alike than those further apart. That is the projected points are arranged in such a way that the grouped ones (small geometrical distance between them) will reflect original relationships in the data. However, additional information to make the projection more informative. As illustrated in Figure-workflow 1, a clustering algorithm has been applied to the MDS projection to highlight the most similar objects (Impacts - Process). The abbreviation of the impacts is introduced to facilitate the visualization (Table 1, Annex 1). The two-dimensional of MDS results of projected impacts (17 impacts) is shown in Figure 5. It presents 4 sub-figures of the first 4 dimensions with the most significant combinations (dimension 1 vs. dimension 2, ..., dimension 2 vs. dimension 3). 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 visualisation of the four dimensions shows the same three groups and we can clearly distinguish three clusters using k-means:

- Group 1: Terrestrial Acidification (TA), Freshwater Eutrophication (FrEu), Terrestrial Ecotoxicity (TecoX) and Freshwater Ecotoxicity (FrEco). TA is most closely associated with FrEu, with respondents considering them almost identical. Other points of Impacts in this cluster are considered similar based on their proximity such TecoX with TA. The FrEco item is furthest in 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 2: Human Toxicity (HT), Particulate Matter formation (PM), Climate Change to Human Health (CCHH), Marine Ecotoxicity (MaEco), Metal Depletion (MeDe), Fossil Depletion (FossDe), Climate Change to Ecosystems (CCE) and Ionizing Radiation (IR). 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. For example, HT and PM, CCE and IR and,

MaEco, MeDe and FossDe. 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.

- Group 3: Urban Land Occupation (UrbLOcc), Photo-Chemical Oxidant formation (Pohto_ChOx), Agricultural Land Occupation (AgLOcc), Ozone Depletion (OD) and Natural Land Transformation (NLTran). This group mainly includes impacts related to land use and land transformation.

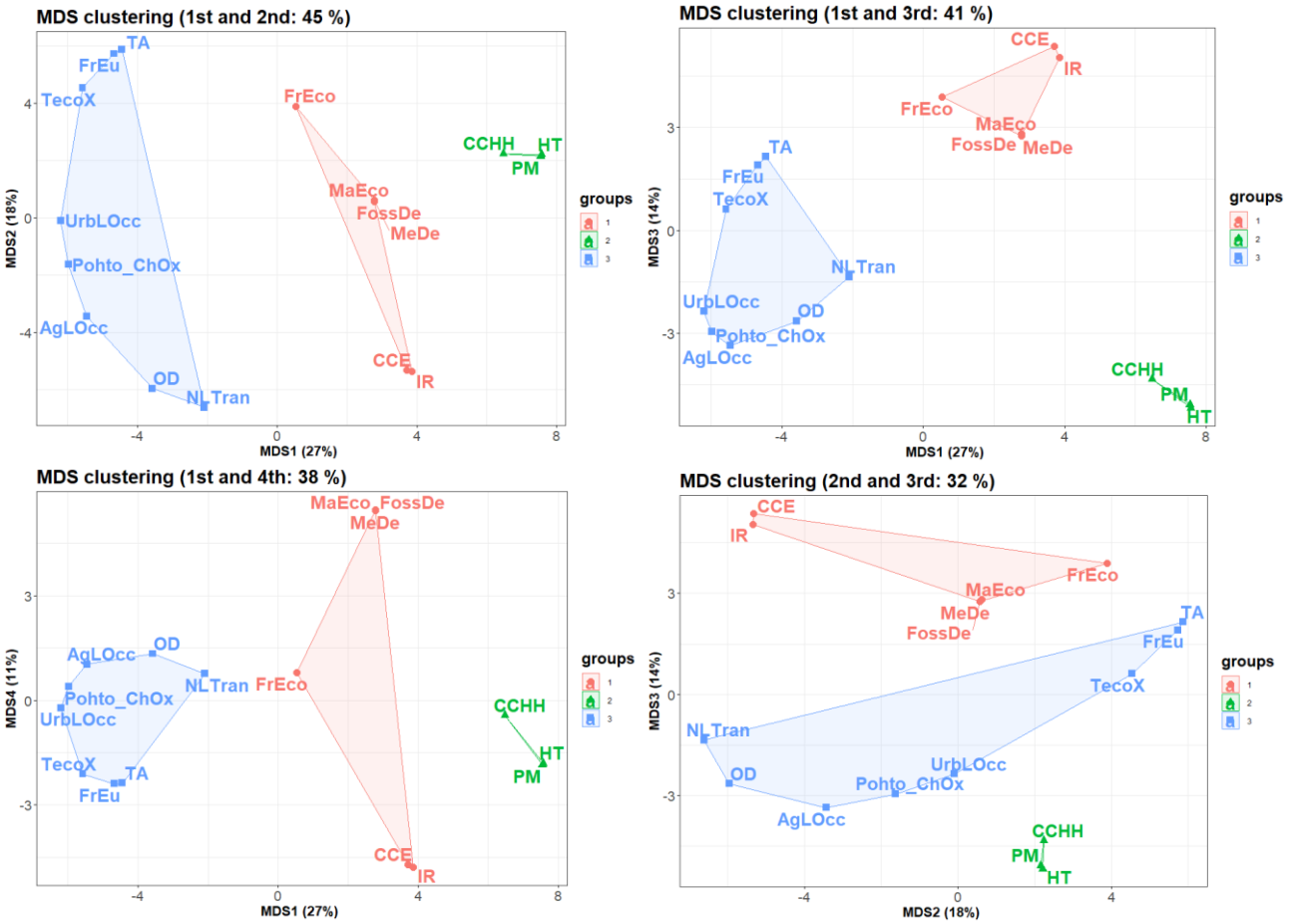

Figure 5: Scatter plot of MDS projection (two dimensions) and k-means clustering based on Impacts distance matrix. Percentage of explained information for the first four components is 70%.

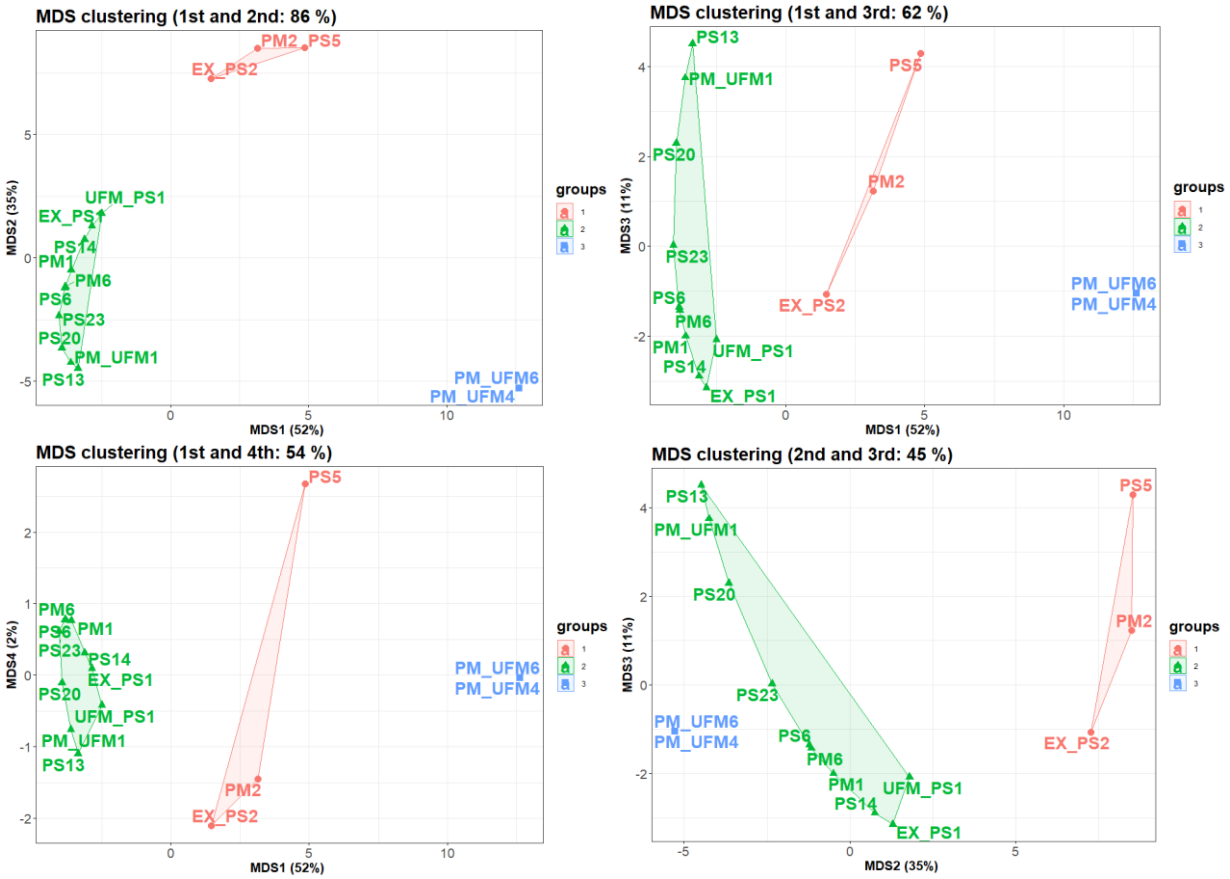

Figure 6: Scatter plot of MDS projection (two dimensions) and k-means clustering based on Process distance matrix. Percentage of explained information for the first four components is 98%.

The results of MDS using process distance matrix is presented in the Figure 6. In this case, the percentage of explained variance for the first four components is 98%, which is an excellent representation in lower-dimensional spaces. That is from a matrix with dimension 15, we lose only 2% to represent the first 4 dimensions. The results are quite similar to those obtained by using MDS on the impacts matrix (Figure 5): three distinguished groups have been identified. The acronyms captioning the dots represent each type of process. These will not be explained here. 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 sub-domain that needs to be analysed individually. In group 3, for example, there are two purely mechanical pre-treatments (PM-UFM for "pre-milling and ultra-fine milling"). 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. For Group 1, the three pre-treatments have relatively similar impacts. For all impacts and fossil depletion, the impact costs around $10, whereas group 2 pre-treatments have an impact costs around $1. Through these groups, we can find an analysis that could be done by an LCA expert, whereas the visualization was done using tools from data science field.

3.2. Discussion and perspectives

The approach proposed in this article is, to our knowledge, unique. It is intended to be generic and practical for the researcher or the R&D engineer. For instance, before starting a project involving the recovery of energy and valuables from wastes, the researcher will be able to compare different processes existing in the literature. The enhancement of the traditional LCA method by coupling tools from (big) data science and algorithms from artificial intelligence allows to have different discussion on environmental impacts. Indeed, on the one side, tools from data science allow to extract and collect data directly from scientific papers. And on the other, the MDS allows to simplify the discussion on environmental impacts. In this paper and this case study, the MDS method was studied because it was the one that gave the best percentage of explained variance.

3.2.1. Limits of this approach

This case study has shown various limitations of the data, and especially data from scientific databases generated by laboratories. The life cycle analysis is therefore carried out at a research TRL scale (TRL scale 1/2) which can lead to a change of scale if we want to switch to an industrial pilot (Bianco et al. 2021). Moreover, as the processes are not continuous in the labs, the data had to be processed so that the inputs of a process step correspond to the outputs of the previous step. The data processing can be improved, for example, adding scale changes made in the process engineering. Another limitation in the data concerning the abundance and quality, which can be very low for some innovative and new processes such as biomass pre-treatment processes. It would be interesting for future studies to change the functional unit and broaden the boundaries of the study to take into account the overall logistical scale or to consider effluent recycling. Indeed, all impacts have been related to the final product (glucose) and no recycling of effluents is taken into account. Recycling could reduce the impacts of some processes more than others. In fact, this approach serves more like aid in deciding on which process or biomass should not be used or which routes could be taken, instead of deciding which one is the best process. A final limitation comes from the main input, which is biomass, a waste product from agriculture which can be of variable quality depending on time and storage. Furthermore, it is necessary to evaluate its durability and supply, which could also vary over time. These three criteria might be a problem for both the economic and environmental process.

3.2.2. Perspectives of the work

Further improvements to the proposed approach could be obtained with any of the following:

(i) including additional specific data sources, methods and visualisations for the economic and social areas. These aspects could enhance the sustainability data inventories and assessment methods. We can cite Social LCA, Life Cycle Cost (LCC), consequential LCA (Curran et al. 2005) or dynamic LCA (Collinge et al. 2013).

(ii) addition of new circular indicators like service-oriented optimization model (Al-Aomar and Alshraideh, 2019)

(iii) progress towards the automation in the data extraction step (step 2 in our approach). This would make it possible to save time and to add new sources of data more easily.

(iv) test different ML techniques to analyse and visualise the raw data such as:

a) Natural Language Processing (NLP) to analyse the text of the scientific literature
b) Other unsupervised learning methods, in particular Dimensionality Reduction Techniques.

(v) from our feedback with the Excel-VBA research tool, the development of a complete ergonomic computing framework remains to be achieved. This would encourage stakeholders to adopt this approach and would facilitate decision-making through the implementation of collaborative decision-making techniques, such as Delphi-SWOT;

(vi) the design of models for calculating energies, for assessing the impacts on the environment of activities linked to new energy systems;

(vii) the generalisation of this principle and the development of a library of business and domain-specific models from agri-food process engineering. These models could be used to check and validate the data in the data architecture step. Controls could include, for example, an advanced material balance or energy analysis;

(viii) the development of data dispersion propagation and automatic qualitative explanation systems for stakeholders.

4. CONCLUSION

A general approach coupling Multidimensional Scaling, k-means algorithm and environmental analysis was proposed. Composed of five steps, this approach is presented as a decision aid for the researcher in a pre-study. It is designed to save time and money by including no experiments and using public scientific data as a database. This approach has been tested in the example of the valorisation of lignocellulosic biomass into glucose through the comparison of pre-treatment processes and two biomasses: corn stalk and rice straw.

After structuring the data and life cycle analysis steps, the environmental impact-process matrix is analysed using an MDS method. A major result of this article is that it highlights the importance of using a hybridization of LCA and ML methods. In light of the results presented above, it can be concluded that:

- This study demonstrates the importance of data science methods to shed new light on LCIA;
- These results contribute significantly to the very small data set available in the literature on using unsupervised learning for LCA purposes.

DECLARATION

- **Funding**

No funds, grants, or other support was received.

- **Conflicts of interest/Competing interests**

The authors have no conflicts of interest to declare that are relevant to the content of this article.

- **Availability of data and code**

The datasets and code generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

- **Authors' contributions**

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Nancy Prioux, Rachid Ouaret and Jean-Pierre Belaud. The first draft of the manuscript was written by N. Prioux and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. The work is supervised by Gilles Hetreux and Jean-Pierre Belaud.

REFERENCES

Abdella GM, Kucukvar M, Onat NC, et al (2020) Sustainability assessment and modeling based on supervised machine learning techniques: The case for food consumption. J Clean Prod 251:119661. https://doi.org/10.1016/j.jclepro.2019.119661

Al-Aomar R and Alshraideh H (2019) A service-oriented material management model with green options. J Clean Prod 236:117557. https://doi.org/10.1016/j.jclepro.2019.07.032.

Barakat A, de Vries H, Rouau X (2013) Dry fractionation process as an important step in current and future lignocellulose biorefineries: A review. Bioresour Technol 134:362–373. https://doi.org/10.1016/j.biortech.2013.01.169

Barry C (2020) La production de déchets non dangereux dans les industries agroalimentaires. Agreste

Belaud J-P, Adoue C, Vialle C, et al (2019a) A circular economy and industrial ecology toolbox for developing an eco-industrial park: perspectives from French policy. Clean Technol Environ 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. Comput Ind 111:41–50. https://doi.org/10.1016/j.compind.2019.06.006

Belaud, J. P., Prioux, N., Vialle, C., Buche, P., Destercke, S., Barakat, A., & Sablayrolles, C. (2021). Intensive Data and Knowledge-Driven Approach for Sustainability Analysis: Application to Lignocellulosic Waste Valorization Processes. Waste and Biomass Valorization, 1-16. https://doi.org/10.1007/s12649-021-01509-8

Bhinge R, Srinivasan A, Robinson S, Dornfeld D (2015) Data-intensive Life Cycle Assessment (DILCA) for Deteriorating Products. Procedia CIRP 29:396–401. https://doi.org/10.1016/j.procir.2015.02.192

Bianco F, Race M, Forino V, et al (2021) Chapter 4 - Bioreactors for wastewater to energy conversion: from pilot to full scale experiences. In: Bhaskar T, Varjani S, Pandey A, Rene ER (eds) Waste Biorefinery. Elsevier, pp 103–124

Burges CJC (2010) Dimension Reduction: A Guided Tour. Found Trends® Mach Learn 2:275–365. https://doi.org/10.1561/2200000002

Burgess AA, Brennan DJ (2001) Application of life cycle assessment to chemical processes. Chem Eng Sci 56:2589–2604. https://doi.org/10.1016/S0009-2509(00)00511-X

Clift R, Frischnecht R, Huppes G, et al (1998) Toward a Coherent Approach to Life Cycle Inventory Analysis. Report of the Working Group on Inventory Enhancement, Brussels

Collinge, W.O., Landis, A.E., Jones, A.K. et al. (2013) Dynamic life cycle assessment: framework and application to an institutional building. Int J Life Cycle Assess 18:538–552. https://doi.org/10.1007/s11367-012-0528-2

Cooper J, Noon M, Jones C, et al (2013) Big Data in Life Cycle Assessment: Big Data in Life Cycle Assessment. J Ind Ecol 17:796–799. https://doi.org/10.1111/jiec.12069

Cox TF, Cox MAA (2001) Multidimensional scaling. Chapman & Hall/CRC, Boca Raton

Curran, M. A., Mann, M., & Norris, G. (2005). The international workshop on electricity data for life cycle inventories. Journal of cleaner production, 13(8): 853-862.

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

Elghali L, Clift R, Sinclair P, et al (2007) Developing a sustainability framework for the assessment of bioenergy systems. Energy Policy 35:6075–6083. https://doi.org/10.1016/j.enpol.2007.08.036

Ellen Macarthur Foundation (2015) Towards a circular economy: business rationale for an accelerated transition. Ellen Macarthur Foundation

European Commission (2019) The European Green Deal

Food SCP Round Table European Commission (2012) Continous Environmental Improvement - Final Report

Frischknecht R, Jungbluth N, Althaus H-J, et al (2005) The ecoinvent Database: Overview and Methodological Framework (7 pp). Int J Life Cycle Assess 10:3–9. https://doi.org/10.1065/lca2004.10.181.1

Garnett T, Appleby MC, Balmford A, et al (2013) Sustainable Intensification in Agriculture: Premises and Policies. Science 341:33–34. https://doi.org/10.1126/science.1234485

Ghisellini P, Cialani C, Ulgiati S (2016) A review on circular economy: the expected transition to a balanced interplay of environmental and economic systems. J Clean Prod 114:11–32. https://doi.org/10.1016/j.jclepro.2015.09.007

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

Guinée JB (ed) (2002) Handbook on life cycle assessment: operational guide to the ISO standards. Kluwer Academic Publishers, Dordrecht ; Boston

Horton P, Koh L, Guang VS (2016) An integrated theoretical framework to enhance resource efficiency, sustainability and human health in agri-food systems. J Clean Prod 120:164–169. https://doi.org/10.1016/j.jclepro.2015.08.092

Huijbregts MAJ, Steinmann ZJN, Elshout PMF, et al (2017) ReCiPe2016: a harmonised life cycle impact assessment method at midpoint and endpoint level. Int J Life Cycle Assess 22:138–147. https://doi.org/10.1007/s11367-016-1246-y

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

Joglekar SN, Tandulje AP, Mandavgane SA, Kulkarni BD (2019) Environmental Impact Study of Bagasse Valorization Routes. Waste Biomass Valorization 10:2067–2078. https://doi.org/10.1007/s12649-018-0198-9

Kouchaki-Penchah H, Nabavi-Pelesaraei A, O'Dwyer J, Sharifi M (2017) Environmental management of tea production using joint of life cycle assessment and data envelopment analysis approaches. Environ Prog Sustain Energy 36:1116–1122. https://doi.org/10.1002/ep.12550

Lance GN, Williams WT (1966) Computer programs for hierarchical polythetic classification ("similarity analyses"). Comput J 9:60–64

Lee EK, Zhang X, Adler PR, et al (2020) Spatially and temporally explicit life cycle global warming, eutrophication, and acidification impacts from corn production in the U.S. Midwest. J Clean Prod 242:118465. https://doi.org/10.1016/j.jclepro.2019.118465

Lissaneddine A, Mandi L, El Achaby M, et al (2021) Performance and dynamic modeling of a continuously operated pomace olive packed bed for olive mill wastewater treatment and phenol recovery. Chemosphere 280:130797. https://doi.org/10.1016/j.chemosphere.2021.130797

Liu C, van der Heide E, Wang H, et al (2013) Alkaline twin-screw extrusion pretreatment for fermentable sugar production. Biotechnol Biofuels 6:97. https://doi.org/10.1186/1754-6834-6-97

Raymond R (2012) Improving food systems for sustainable diets in a green economy. 2012 United Nations Conference on Sustainable Development: Governance for Greening the Economy with Agriculture

Romeiko XX, Lee EK, Sorunmu Y, Zhang X (2020) Spatially and Temporally Explicit Life Cycle Environmental Impacts of Soybean Production in the U.S. Midwest. Environ Sci Technol 54:4758–4768. https://doi.org/10.1021/acs.est.9b06874

Singh S, Babbitt C, Gaustad G, et al (2021) Thematic exploration of sectoral and cross-cutting challenges to circular economy implementation. Clean Technol Environ Policy 23:915–936. https://doi.org/10.1007/s10098-020-02016-5

Suh S, Huppes G (2005) Methods for Life Cycle Inventory of a product. J Clean Prod 13:687–697. https://doi.org/10.1016/j.jclepro.2003.04.001

Tan Y, Shuai C, Jiao L, Shen L (2017) An adaptive neuro-fuzzy inference system (ANFIS) approach for measuring country sustainability performance. Environ Impact Assess Rev 65:29–40. https://doi.org/10.1016/j.eiar.2017.04.004

WCED (1987) World commission on environment and development. Our common future. 17:1–91

West PC, Gerber JS, Engstrom PM, et al (2014) Leverage points for improving global food security and the environment. Science 345:325–328. https://doi.org/10.1126/science.1246067

Zhao L, Zha Y, Zhuang Y, Liang L (2019) Data envelopment analysis for sustainability evaluation in China: Tackling the economic, environmental, and social dimensions. Eur J Oper Res 275:1083–1095. https://doi.org/10.1016/j.ejor.2018.12.004

Zhu JY, Pan X, Zalesny RS (2010) Pretreatment of woody biomass for biofuel production: energy efficiency, technologies, and recalcitrance. Appl Microbiol Biotechnol 87:847–857. https://doi.org/10.1007/s00253-010-2654-8

ANNEX 1: NOMENCLATURE OF THE ABBREVIATION

Life Cycle	
Terms	**Abbreviation**
Life Cycle Impact Assessment	LCIA
Life Cycle Assessment	LCA
Life-Cycle Environmental	LCE
Life Cycle Inventory	LCI
Circular Economy	CE
Machine Learning	
Terms	**Abbreviation**
Adaptive Neuro-Fuzzy Inference System	ANFIS
Data Envelope Analysis	DEA
Multidimensional Scaling	MDS
Dimension Reduction	DR
Boosted Regression Tree	BRT
Principal Coordinates Analysis	PCoA
Database management	
Terms	**Abbreviation**
MAchine-Readable Cataloging	MARC
Relational Database Management Systems	RDBMS
Machine Learning	
Terms	**Abbreviation**
Reaserch and developpement	R&D
National Ecological Transition Strategy for Sustainable Development	SNTEDD
French Agency for Environment and Energy Management	ADEME