DOI: 10.1111/jiec.13305

RESEARCH ARTICLE

JOURNAL OF INDUSTRIAL ECOLOGY WILEY

A review of inventory modeling methods for missing data in life cycle assessment

Shiva Zargar[1] | Yuan Yao[2] | Qingshi Tu[1]

[1]Sustainable Bioeconomy Research Group, Department of Wood Science, The University of British Columbia, Vancouver, Canada

[2]Center for Industrial Ecology, Yale School of the Environment, Yale University, New Haven, Connecticut, USA

Correspondence
Qingshi Tu, Sustainable Bioeconomy Research Group, Department of Wood Science, The University of British Columbia, Vancouver, Canada.
Email: qingshi.tu@ubc.ca

Editor Managing Review: Mark Huijbregts.

Funding information
Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number RGPIN-2021-02841].

Abstract

Missing data is the key challenge facing life cycle inventory (LCI) modeling. The collection of missing data can be cost-prohibitive and infeasible in many circumstances. Major strategies to address this issue include proxy selection (i.e., selecting a surrogate dataset to represent the missing data) and data creation (e.g., through empirical equations or mechanistic models). Within these two strategies, we identified three approaches that are widely used for LCI modeling: Data-driven, mechanistic, and future (e.g., 2050) inventory modeling. We critically reviewed the 12 common methods of these three approaches by focusing on their features, scope of application, underlying assumptions, and limitations. These methods were characterized based on the following criteria: "domain knowledge requirement" (both as a method developer and a user), "post-treatment requirement," "challenge in assessing data quality uncertainty," "challenge in generalizability," and "challenge in automation." These criteria can be used by LCA practitioners to select the suitable method(s) to bridge the data gap in LCI modeling, based on the goal and scope of the intended study. We also identified several aspects for future improvement for these reviewed methods.

KEYWORDS
future inventory, industrial ecology, life cycle assessment, life cycle inventory modeling, machine learning, proxy selection

1 | INTRODUCTION

Life cycle inventory (LCI) modeling is the foundation of life cycle assessment, which transforms the knowledge of a product system into quantifiable unit processes and relevant input/output flows for environmental impact assessment. In general, LCI modeling involves four steps: (1) Identifying the reference product flow (e.g., 1 MJ of biodiesel), elementary flows (e.g., air pollutant emissions from combustion), and other intermediate flows (e.g., methanol, electricity); (2) constructing the product system by identifying relevant unit processes that generate these flows, which can be accomplished by either selecting existing ones from the database (also known as "proxy selection") (Meron et al., 2020) or creating new ones based on domain knowledge (Huijbregts et al., 2001); (3) adjusting each unit process for the context of interest (e.g., update the electricity input of a manufacturing process to match the technical specification of the study); and (4) assessing the uncertainty of flows or unit processes (e.g., pedigree matrix for "data quality uncertainty"). There are two types of uncertainty associated with LCI modeling. Data quality uncertainty is "contextual," (Henriksen et al., 2020), referring to the uncertainty originating from the differences in scope (e.g., geographic representativeness) between the flow/unit process of interest and the corresponding LCI data created from steps 1–3. Model uncertainty, also known as basic uncertainty, refers to the uncertainty stemming from the inherent variations of the model/method that generates the data (e.g., fluctuation of process yield over time, measurement errors) (Ecoinvent, 2018).

A common practice for LCI modeling is summarized in Figure 1. A query is made to the database (e.g., ecoinvent) to determine if there is an exact match for the technical description of the unit process of interest (e.g., "flat glass production, uncoated"). When such a match is found, the next step

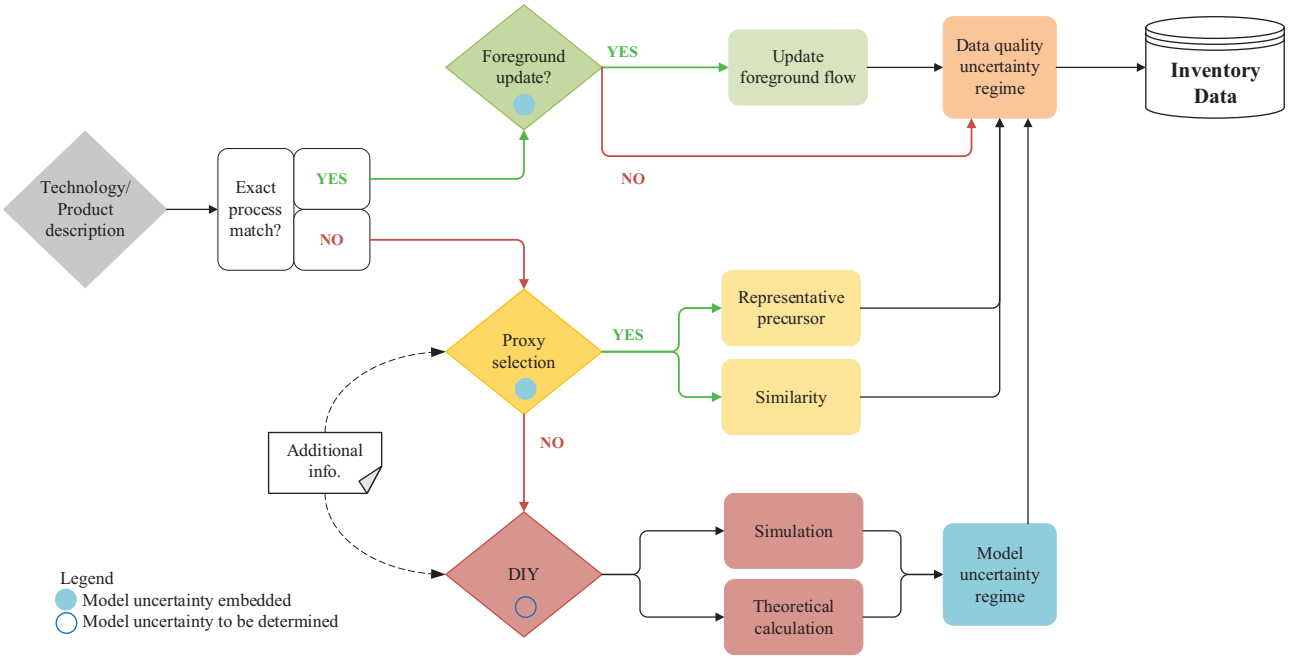

FIGURE 1 Illustration of current practice for inventory modeling and associated uncertainty assessment (DIY = generate the inventory or compiling from existing inventory)

is to determine whether any update is needed for the foreground exchanges (e.g., select a region-specific electricity provider) of the selected unit process. Switching, adding, or removing a foreground exchange automatically incorporates the concomitant model uncertainty that is embedded within the new exchange (e.g., the uncertainty associated with modeling the electricity generation technology of a new electricity provider). On the other hand, if an exact match of technical description is not found, two available options are either "proxy selection" or "compiling from existing inventory." For the former, two approaches are typically used: (1) "representative precursor approach" refers to using the unit process of producing the immediate precursor (e.g., formic acid, the precursor chemical of oxalic acid production) as the surrogate when the unit process for producing the actual product of interest (e.g., oxalic acid) is not available (Kendall et al., 2010). (2) "Similarity-based approach" refers to the selection of a surrogate unit process based on the general similarity in technology description. For example, "heat, from steam, in the chemical industry" could be used to represent the source of heating at an electrocatalysis-based chemical plant when a plant-specific unit process for heating is not available (Tu et al., 2021). As both approaches map the actual unit process of interest to existing unit processes available in the database, they rely on the embedded assumptions associated with existing unit processes to assess the corresponding model uncertainty. On the other hand, the option of "compiling from existing inventory," instead of selecting a single unit process as the surrogate from the database, refers to the situation where one can build a new unit process by compiling the relevant unit processes from the database. For example, the LCI of the synthesis process of choline chloride could be represented by the combination of the unit processes and flows describing the synthesis of ethylene oxide, hydrochloric acid, and trimethylamine, as well as electricity and heat generations. This requires detailed information (e.g., reaction equation for the synthesis of choline chloride from these chemicals and corresponding energy consumption) in order to quantify the exchange of these relevant flows (and to identify the unit processes that generate those flows) throughout the synthesis process of choline chloride, which is often achieved by theoretical and/or empirical calculations (e.g., enthalpy of reaction, sensible heat) (Zargar et al., 2022) or process simulation (e.g., using AspenPlus®) (De et al., 2019; Han et al., 2019; Parvatker & Eckelman, 2019;). As the new unit process is created by explicitly quantifying the exchanges, the corresponding model uncertainty could be described by assigning uncertainty information (e.g., a probability distribution) to flow exchanges. After a proper inventory is determined using one of the earlier-mentioned options, the corresponding data quality (e.g., the temporal correlation between the unit process of choice and actual inventory data of interest) can be assessed using standard methods such as a pedigree matrix (Heijungs & Huijbregts, 2004; Huijbregts et al., 2001; Weidema & Wesnæs, 1996). Pedigree matrix is the most commonly used evaluation method for data quality uncertainty, which converts the qualitative assessment of the appropriateness of using a unit process (with respect to a specific context) into quantitative scores used to construct an underlying distribution (typically lognormal) for uncertainty analysis (Bamber et al., 2020; Ciroth et al., 2016).

Although ISO 14044 (2006) provides general guidance, there lacks a standard approach for LCI modeling. Subjectivity in many aspects, from proxy selection (Canals et al., 2011) to data quality scoring (Cooper & Kahn, 2012), results in debates about the efficacy of LCA results for decision-making (Säynäjoki et al., 2012). This study critically reviewed the existing methods for proxy selection and data creation in LCI modeling with a focus on modeling missing inventory data. In particular, we focused on the methods belonging to data-driven and mechanistic approaches, two major

categories of LCI modeling approaches, as well as those for future (e.g., 2050) inventory modeling. We aimed to answer the following questions, which can collectively guide LCA practitioners to select a suitable method for filling LCI data gaps :

1. What are the commonly used methods to fill LCI data gaps?
2. What are the features, scope of application, and underlying assumptions of these methods?
3. What are the limitations of these methods?

We acknowledge the concomitant uncertainties associated with these inventory modeling methods, yet a thorough understanding of how to quantify these uncertainties entails a separate critical review and hence is not included in the scope of this study.

2 | METHODS

To identify relevant studies, we used multiple search engines, including Science Direct, Web of Science Core Collection, Engineering Village, World Wide Science, and Google Scholar. We used the following keywords: Prospective life cycle inventory approach*, prospective inventory creation, prospective inventory modeling, proxy selection in LCA, LCI AND machine learning, LCI AND data mining, LCI AND emerging technology*, LCA AND inventory AND chemical*, LCA AND inventory AND emerging technology*. This process was iterative and included canvassing the reference lists of identified literature sources. To this end, 12 inventory modeling methods were found and classified into 3 categories based on their modeling approaches: Data-driven approach (9), mechanistic approach (4), and future inventory data modeling approach (3).

These methods were characterized based on six criteria: (1) "domain knowledge requirement (developer)" needed to create a method. The developers refer to the original authors of the method articles reviewed. (2) "Domain knowledge requirement (user)" needed to apply an existing method to a specific case. The users refer to the LCA practitioners who intend to apply the method developed by the authors of the original article. The domain knowledge requirement criteria (both developer and user) further differentiate "subject-specific knowledge" (e.g., operation conditions for synthesizing a particular chemical) from "method-specific knowledge" (e.g., how to train a particular machine learning model). (3) "Post-treatment requirement" describes the further adjustments (e.g., change of foreground electricity provider) needed after the inventory data is created (either from proxy selection or new inventory creation). (4) "Challenge in assessing data quality uncertainty" refers to the additional information required to assess the data quality uncertainty of the inventory data created by a particular method. (5) "Challenge in generalizability" indicates the challenges facing a method when applied outside the domain where the method was initially developed. (6) "Challenge in automation" refers to the requirement of manual inputs when applying a method. For each criterion (except "domain knowledge requirement" [developer]), three levels of characteristic descriptions were conceived (Table S1 in the Supporting Information) in order to explain the challenge of developing and/or applying each method. Level 1 descriptions generally correspond to a low degree of challenge (e.g., application of a method is highly automated), whereas level 3 descriptions typically indicate a high degree of challenge (e.g., both subject-specific and method-specific knowledge are required when applying a method in a domain where the method was not originally developed from). The degree of challenge (e.g., domain knowledge requirement, automation) increases as the descriptions move from level 1 to level 3.

3 | RESULTS

The following sections provide a detailed description of the features and scope of application and underlying assumptions of each method reviewed. Table 1 summarizes the commonly used methods with their key characteristics against the six criteria defined in the Methods section.

3.1 | Data-driven approach

Machine learning (ML) models have become an important part of the data-driven approach in recent years. For example, Meron et al. (2020) proposed a similarity-based method for proxy selection using a group of characteristics (e.g., regional population density, average waste composition index for a region-specific municipal solid waste [MSW] management process) to describe individual unit processes or inventory data (e.g., kg CO_2 emissions) contained in a data pool. These characteristics define a multidimensional space where each unit process is situated based on the values of its characteristics. Domain knowledge is required to identify the appropriate set of characteristics and their value ranges, such that the Euclidean distance (calculated by the values of characteristics) between two more similar unit processes is shorter than that of two less similar ones. When the values of characteristics are provided for the unknown process, the most relevant (i.e., shortest distance) unit process or inventory data could be automatically selected as a proxy. The efficacy of this proxy selection approach is sensitive to the choice of the characteristics and their value ranges. Another similarity-based method is developed for automatically creating the missing (intermediate and elementary) flow data for a unit

TABLE 1 Selected methods for bridging inventory data gap and their characteristics against the criteria matrix

Approach	Domain knowledge requirement (Developer)	Domain knowledge requirement (User)	Post-treatment requirement	Challenge in assessing data quality uncertainty	Challenge in generalizability[a]	Challenge in automation	Study reviewed
Data-driven							
Similarity-based model for proxy selection	– [subject-specific knowledge] identify a set of attributes for encoding a specific database – [method-specific knowledge] KNN	For other applications: – Need to encode (based on subject knowledge) a database other than the default one used by the authors – Need to train a KNN model from scratch	The selected proxy can be used directly (uncertainty needs to be assessed separately)	Pedigree matrix values should be adjusted based on the relationship between selected proxy and the "ground-truth" (e.g., geographical correlation)	"The proposed methodology and algorithm are general, and the scope of the methodology's applicability is broad." —Meron et al., 2020	A meta-analysis of relevant LCA data (e.g., literature, database) is needed to produce the set of characteristics for encoding (can be both database- and subject-specific)	(Meron et al. 2020)
Similarity-based model for data creation[a]	– [subject-specific knowledge] need to prepare the database to be ready for ML model (manual) – [method-specific knowledge] KNN	For other applications: – Need to prepare the database to be ready for ML model (manual) – Need to train a KNN model from scratch	The resulting unit process can be used directly (uncertainty needs to be assessed separately)	The resulting process (and its flows) is aggregated from other processes, uncertainty information is lost	"Although we envision broad applications of our method, …our method is more applicable in the situation that most data are known and only a few data points are missing." —Hou et al., 2018	– Calculating similarity scores based on flows of a unit process is automated once the model is built – Need to prepare the database to be ready for ML model (manual)	(Hou et al. 2018)
eXtreme gradient boosting (XGBoost)	– [subject-specific knowledge] need to prepare the database to be ready for ML model (manual) – [method-specific knowledge] classification model to identify the relevance (i.e., existence of linkage) of the flows; regression model to predict value of relevant flows	For other applications: – Need to prepare the database to be ready for ML model (manual) – Need to train one or two ML models from scratch	The resulting unit process can be used directly (uncertainty needs to be assessed separately)	The value of missing flows of the resulting process are generated using those of the existing flows (of the same process) as predictors, uncertainty cannot be assessed	"Although we envision broad applications of our method, …The application domain of our model is the unit processes with no less than 10 nonzero flows excluding those flows with extremely small values." —Zhao et al., 2021	– Need to prepare the database to be ready for ML model (manual) – Once the model is created, training is automated	(Zhao et al. 2021)

(Continues)

TABLE 1 (Continued)

Approach	Domain knowledge requirement (Developer)	Domain knowledge requirement (User)	Post-treatment requirement	Challenge in assessing data quality uncertainty	Challenge in generalizability[a]	Challenge in automation	Study reviewed
Artificial neural network	– [subject-specific knowledge] input array preparation (specific to the subject of interest) – [method-specific knowledge] ANN	For other applications: – Need to prepare the input array to be ready for ML model (subject-specific) – May need to retrain the ANN or create a new architecture	The resulting unit process can be used directly (uncertainty needs to be assessed separately)	Uncertainty assessment not mentioned	ANN is a generic ML method, so the generaliability is very high; the only challenge will the need for large training dataset	– Need to prepare the input array to be ready for ML model (subject-specific) – Once the model is created, training is automated	(Liao et al., 2020)
Probability density functions	– [subject-specific knowledge] variable of interest (case specific) – [method-specific knowledge] maximum likelihood estimation method	Need to manually select variable of interest (subject-specific) to create PDF (method-specific); unless directly use PDF generated by others	The resulting data can be used directly (uncertainty needs to be assessed separately)	– Data quality uncertainty assessment may not be possible, as each datapoint for PDF creation may come with very different contexts (e.g., data collected from different times, locations) – Uncertainty assessment on model output can be assessed using Monte Carlo simulation (based on the PDF)	– Creating PDF for variable of interest is commonly used in many different fields of modeling – Applicable only to "flows" (e.g., steam consumption of a chemical manufacturing process)	Need to manually select variable of interest to create PDF	(Pereira et al., 2018)
Classification trees	– [subject-specific knowledge] defining classes (case-specific) – [method-specific knowledge] classification tree	Requirement is the same as those for the developer (unless the results generated by developers are directly used by others)	The resulting data can be used directly (uncertainty needs to be assessed separately)	Data quality uncertainty assessment may not be possible, as each data point contributing to the interval may come with very different contexts (e.g., data collected from different times, locations)	Applicable only to "flows" (e.g., steam consumption of a chemical manufacturing process)	Need to manually define classes to crate the classification tree	(Pereira et al. 2018)

(Continues)

TABLE 1 (Continued)

Approach	Domain knowledge requirement (Developer)	Domain knowledge requirement (User)	Post-treatment requirement	Challenge in assessing data quality uncertainty	Challenge in generalizability[a]	Challenge in automation	Study reviewed
Data mining[b]	– [subject-specific knowledge] defining filtering rules (case-specific) – [method-specific knowledge] building data mining workflow (e.g., in Excel)	Requirement is the same as those for the developer (unless the results generated by developers are directly used by others)	The resulting data can be used directly (uncertainty needs to be assessed separately)	DQI using pedigree matrix (indicator values based on expert adjustment)	Applicable mostly to chemical manufacturing	– Need manual analysis of the database, defining filter rules	(Meyer et al., 2020)
Mechanistic							
Process simulation	– [subject-specific knowledge] process information, e.g., reaction conditions, purification requirement (case-specific) – [method-specific knowledge] simulation software	Requirement is the same as those for the developer (unless directly use the results generated by others)	The resulting data can be used directly (uncertainty needs to be assessed separately)	DQI using pedigree matrix (indicator values based on expert adjustment)	Mostly applicable to chemical and biological processes	Mostly rely on commercial software (manual operations)	Meyer et al., 2019; Montazeri & Eckelman, 2016; Morales-Mendoza & Azzaro-Pantel, 2017; (Parvatker & Eckelman, 2019; Tu et al., 2021)
Theoretical method	– [subject-specific knowledge] e.g., reaction stoichiometry	Requirement is the same as those for the developer (unless the results generated by developers are directly used by others)	The resulting data can be used directly (uncertainty needs to be assessed separately)	Not mentioned by the authors, but likely DQI using pedigree matrix (indicator values based on expert adjustment)	Mostly applicable to chemical and biological processes	Calculation can be automated, although subject-specific knowledge is required to provide information for calculation	(Cuéllar-Franca et al., 2016; Righi et al., 2020; Tsoy et al., 2020)
Lineage and process ontologies	– [subject-specific knowledge] e.g., components involved in the synthesis pathways of a certain chemical – [method-specific knowledge] ontology modeling	Requirement is the same as those for the developer (unless the results generated by developers are directly used by others)	This method guides inventory modeling (i.e., what to model), rather than providing inventory data	Not mentioned by the authors; as the output of the method is an intermediate step toward actual inventory modeling, it is not clear how data quality uncertainty can be assessed	Ontology modeling is applicable to a wide range of domains	Ontology modeling can be automated (after architecture is set up manually), however, inventory modeling is still required separately	Barrett et al., 2019); (Mittal et al., 2018

(Continues)

TABLE 1 (Continued)

Approach	Domain knowledge requirement (Developer)	Domain knowledge requirement (User)	Post-treatment requirement	Challenge in assessing data quality uncertainty	Challenge in generalizability[a]	Challenge in automation	Study reviewed
Future inventory data							
Learning curve	– [subject-specific knowledge] e.g., advancement trajectory of similar technologies, basic laws of physics, chemistry and thermodynamics (to set boundary conditions)	Requirement is the same as those for the developer (unless the results generated by developers are directly used by others)	The resulting data can be used directly (uncertainty needs to be assessed separately)	Not mentioned by the authors, but likely DQI using pedigree matrix (indicator values based on expert adjustment)	Applicable to a wide range of domains	Heavy reliance on subject-specific knowledge (mostly manual knowledge synthesis process)	Tsoy et al. (2020)
Alteration based on scenario model results	– [subject-specific knowledge] structure of a specific database – [method-specific knowledge] working knowledge of a specific IAM; programing language	Requirement is the same as those for the developer (unless the results generated by developers are directly used by others)	The resulting data can be used directly (uncertainty needs to be assessed separately)	Use different future scenarios to generate inventory data (not directly analyzing data quality uncertainty)	Applicable to energy systems	Automation through soft linking IAM results with database of interest using tools such as "Wurst"	(Mendoza Beltran et al., 2020)

[a]Examples of database preparation: Database cleaning to retain parent processes only, normalization.
[b]Automation is possible when combining text mining techniques with subject-specific terminology.

process based on the assumption that similar processes (e.g., organic chemical production processes) in a network of unit processes are likely to have similar inputs and outputs of intermediate and elemental flows (Hou et al., 2018). The similarity is based on Minkowski distance, a generic formula that can represent distance calculations (e.g., Manhattan and Euclidean) by changing the values of its parameter. While Meron et al. (2020) selected the most similar unit process from existing candidates, Hou et al. (2018) created a unit process by aggregating the inventory data (e.g., the weighted average of electricity uses) of the most similar unit processes in the existing pool. The data quality uncertainty of the resulting inventory data, however, cannot be assessed due to the aggregation from multiple unit processes whose underlying data quality indicators may vary significantly. Besides, since the approach is highly sensitive to the structure and completeness of the network, its efficacy is limited even when the fraction of missing flow data is small (e.g., 5%) in a unit process. Zhao et al. (2021) developed an eXtreme Gradient Boosting (XGBoost)-based framework to estimate the values of missing flows. XGBoost is a tree-based ensemble ML algorithm that combines the decisions (classification or regression values) of individual trees to arrive at the final prediction (Chen & Benesty, 2016; Romeiko et al., 2020). The two-step modeling framework started with first an XGBoost classification model to distinguish between flows with zero and nonzero values of LCI (e.g., whether a certain resource flow is used as the input to a given unit process). In the second step, an XGBoost regression model was constructed to estimate the value of those nonzero flows. This method showed an improvement in tolerance for missing flows, compared to the 5% limit from Hou et al. (2018). For example, when 10% of flow data was missing, the misclassification rate was very low (0.79%). When less than 20% of the flow data was missing, the regression model could still generate a reasonably accurate estimation ($R^2 > 0.7$). This improvement is likely due to the general observation in ML field that ensemble decision tree-based models (e.g., XGBoost in this case) are generally more robust against outliers compared to similarity-based models. Although the issue of missing flow data is considerably mitigated, the XGBoost-based modeling framework by Zhao et al. (2021) still faces several limitations. For instance, the predictions were poor when the unit process of interest either had very few nonzero flows or the value of the nonzero flow was extremely small (e.g., $< 10^{-7}$). The prediction performance was also weak for the flows that were only used by a small number of unit processes in the training database, which is a common issue with the imbalanced training data. As a rule of thumb, this modeling framework should be applied when the unit process of interest had no less than 10 nonzero flows whose values were not extremely small.

Another major ML algorithm is ANN, a computational model structure based on the brain's neuron interaction principles (Swingler, 2001; Zhu et al., 2020). ANN has been applied to create molecular structure-based models to estimate the environmental impacts of chemicals (Parvatker & Eckelman, 2019; Wernet et al., 2012).

Liao et al. (2020) coupled ANN with kinetic modeling (for the pyrolysis step) and process simulation (for the entire production facility) to generate a comprehensive inventory dataset for activated carbon (AC) preparation from a wide range of woody biomass and operational conditions. An ANN model was trained to predict the yield of AC prepared from steam activation of pyrolyzed woody biomass, using input parameters such as biomass characteristics, pyrolysis conditions, and activation conditions. The yield is a crucial input to the process simulation that was used to estimate the LCI of AC production. This example demonstrates the advantage of combining ANN and mechanistic, knowledge-based models to create inventory data when large training datasets are not available. Nevertheless, mechanistic models are domain-specific, and data availability is always a major challenge that may limit the wide application of ANN (and ML models in general) for inventory data generation in different domains. In addition to ANN, many other ML approaches can be used to estimate the key process parameters of biomass conversion to be combined with mechanistic models. Those approaches have been reviewed in a previous study (Liao & Yao, 2021).

Statistical analysis is another data-driven approach. A few types of statistical methods have been applied in LCA studies. For example, probability density functions (PDF) and classification trees were used to estimate the steam consumption for chemical and pharmaceutical productions by Pereira et al. (2018). For both methods, the foundational hypothesis was that steam consumption correlated mostly with reaction classes rather than specific reactants and products involved; therefore, the estimation of steam consumption can be simplified (with reasonable estimation errors) by focusing only on the information relevant to reaction classes. For a given group of reaction classes (e.g., alkylations and arylations), the type of distribution and associated parameter values of a PDF were determined by the maximum likelihood estimation method, using primary data collected from industries and/or secondary data from literature. Accordingly, the interquartile range of steam consumption for a given reaction group was generated through Monte Carlo simulation using the corresponding PDF. On the other hand, the classification tree approach utilizes additional attributes such as operational conditions in tandem with reaction classes to categorize steam consumptions into several classes (e.g., "low/medium/high steam consumption classes"). Unlike the PDF approach, the interval of the steam consumption of each class (e.g., "high steam consumption") was deterministic (e.g., aggregated from respective data of that class) in the classification tree approach. In order to have sufficient data for constructing a PDF, aggregation of similar reaction classes (e.g., based on the type of molecules involved in the reaction) into one reaction group is needed, which entails the application of domain knowledge. For instance, alkylations and arylations are included in the same reaction group as alkyl and aryl groups are both hydrocarbons. The downside is the potential increase in uncertainty of predicted results. Also, as indicated by Pereira et al. (2018), it is important to validate the underlying assumption that steam consumption is similar between reaction classes within the same reaction group, when new reaction classes are to be investigated. For example, a univariate analysis of variance (ANOVA) can be applied to comparing the observed steam consumption values (e.g., from facility operations) between reaction classes of interest and those within the same general reaction group.

Data mining has been used to consistently collect, harmonize, process, and convert data into LCIs to gain valuable insights and support the conversion of data into machine-readable queries for automation (Cashman et al., 2016; Ortmeier et al., 2021). A key benefit of the data mining method

is its high-degree automation for information extraction and update, whereas a major challenge is the representativeness of the mined data (Cashman et al. (2016). (Meyer et al. (2020) proposed a procedure to create context-based filtering rules, using metadata (e.g., "Source Classification Code" and "Emission Unit Description" from National Emission Inventory (EPA, 2014)), to improve technological correlation and completeness of the resulting inventory data from data mining. The authors proposed a set of filtering rules to delineate species hierarchies (e.g., ethylbenzene as a speciated VOC under "total VOCs") and intersource overlap (e.g., air pollutant emissions are reported in both National Emission Inventory and Toxic Release Inventory) (EPA, 2014), in order to minimize double-counting in reported emission inventory data. The authors also presented a sanitization method for data quality improvement by reducing the exclusion of confidential information from the industry. Despite setting filtering rules based on manual data source analysis in this study, the authors expected that an automated creation of context-based filtering rules could be achieved by combining the text mining techniques from cheminformatics with manufacturing process terminology, which may enable a more accurate application of secondary data for chemical manufacturing.

3.2 | Mechanistic approach

Methods of mechanistic approach rely largely on the physical and/chemical relationship within a product system for inventory data modeling (e.g., the heat consumption for an endothermic reaction for chemical synthesis). LCI databases (e.g., ecoinvent, GaBi) contain inventory data for about 500 chemicals, mainly bulk chemicals or intermediates; however, the number of chemicals in commerce has reached 85,000 (National Academies Press, 2014). This substantial gap in data availability causes challenges in obtaining representative LCA results for systems involving those chemicals that do not have inventory data representing their production and associated emissions. Accordingly, multiple methods have been developed to estimate LCI data for chemicals. Process simulations, using tools such as AspenPlus, CHEMCAD, DWSIM, HYSYS, and PROSIM, have been an effective method for estimating LCI data for chemicals and fuel productions where commercial-scale operations do not already exist or when primary data is unavailable (Montazeri & Eckelman, 2016; Morales-Mendoza & Azzaro-Pantel, 2017; Parvatker & Eckelman, 2019). We also acknowledge the enormous LCA publications of biomass-derived fuels and materials that use process simulations to estimate LCI data. These publications are not included in this review as they have been extensively reviewed in prior review articles (Fröhling & Hiete, 2020; Lan et al., 2019). Process simulation transforms the configuration of a production facility into a connected process model of unit operations (e.g., reactor and distillation column) and auxiliary equipment (e.g., pump and valve). Particularly for technologies at a low technology readiness level (TRL), process simulation enables a comprehensive estimation of utility demand (e.g., for cooling and heating), material and waste flow, as well as potential energy savings through plant-wide heat integration (e.g., through pinch analysis). For example, a novel chemical synthesis technology at TRL 4 "Component and/or validation in a laboratory environment") (Innovation Canada, 2018) may lack a proper laboratory setup to estimate the energy and material consumptions of downstream treatment at the commercial scale (e.g., coproduct purification, solvent recovery). With chemical engineering expertise and proper domain knowledge, a process simulation can be set up to investigate the operational conditions of distillation columns and hence, determine the purity of main and coproducts as well as associated utility demand (Tu et al., 2021). Despite these advantages, the requirement for chemical engineering expertise and knowledge of plant-specific design parameters constrains the application of process simulations in LCI modeling (Meyer et al., 2019; Parvatker & Eckelman, 2019). When a resource (e.g., software access) or expertise (e.g., a proper choice of the equation of state for the thermodynamics system of interest) is not available, obtaining inventory data for core steps of the manufacturing processes, such as estimating heat duty for reaction and distillation, may still be feasible by a combination of theoretical (e.g., enthalpy of reaction based on stoichiometry) and shortcut methods. The theoretical method requires chemical and physical properties data of components, such as molecular weight, standard enthalpy of formation, boiling point, and so on. (Cuéllar-Franca et al., 2016; Righi et al., 2020). The sources of relevant information include chemical process encyclopedia, design equations, physical and thermodynamic data for chemical handbooks and literature, such as Perry's Chemical Engineers' Handbook and NIST database (Kim & Overcash, 2003; Parvatker & Eckelman, 2019; C. I. Watson, 1992; Perry, 1950).

Mittal et al. (2018) developed a lineage ontology-based method to map out the chemical production supply chains by identifying the synthesis steps contained in the production processes. In order to manage the data in various unit processes, a process ontology-based method was developed to connect the process data with lineage data. To generate a lineage for a chemical from a set of reaction data, a series of SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language) queries were applied. When a lineage is established, process ontology will be used to assist modeling the inventory based on two approaches: Top-down (e.g., data mining) and bottom-up (e.g., simulation) (Mittal et al., 2018).

3.3 | Future inventory data modeling approach

Modeling the changes in the foreground and background inventory data for a product system, with respect to future technological (e.g., improvement in yield of a process) and market scenarios (e.g., a shift in electricity grid mix for energy supply), is a unique challenge facing prospective LCA (Tsoy et al., 2020). For the technological changes in foreground LCI data, one method is to apply learning curves that reflect the change of foreground inventory data (e.g., yield and energy use) in response to the progress of technology development and implementation (Tsoy et al., 2020).

Traditionally, learning curves are used to describe the technology advancement in terms of "decreasing costs as a function of accumulating experience with that technology" or "as a function of cumulative production," which does not seem to be directly relevant for prospective LCA. However, the learning curve for the future advancement of emerging technology (e.g., improved material efficiency in foreground system) can be characterized by referencing the trajectory of advancement of the similar existing technologies, in combination with expert knowledge and basic laws of physics, chemistry, and thermodynamics (Tsoy et al., 2020).

Transforming the existing background LCI databases toward future contexts (e.g., inventory data for electricity grid mix in the next 20 years) via the scenario assumptions is a common approach to avoid the temporal mismatch between background and foreground systems (Tsoy et al., 2020). In order to create a set of consistent and transparent scenario assumptions, comprehensive data collection, harmonization, and interpretation are crucial, representing a challenge for the scenario approach (Fishman et al., 2021).The Shared Socioeconomic Pathways (SSPs) (Dellink et al., 2017; Mendoza Beltran et al., 2020; O'Neill et al., 2014, 2017; Riahi et al., 2017) developed by the climate change research community are one of the most commonly consulted scenario descriptions. SSPs are used to generate socioeconomic and technological development assumptions for a given future scenario employed by an integrated assessment model (IAM) (O'Neill et al., 2017). These assumptions represent the linkage among socioeconomic drivers (e.g., urbanization) and natural environment and technological changes (e.g., decarbonization of electricity grid), based on which IAMs calculate the future GHG emissions due to changes in energy systems, resource uses, and so on. IAMs are considered a creditable source of future inventory data for prospective LCA research, because they provide the opportunity for exploring plausible future scenarios and the mechanism to endogenously generate relevant technological and resource use data make.

Besides the capacity of transforming the scenario descriptions into the information for LCI modeling through IAMs, tools have also been developed to facilitate the modification of a given database based on this information. For example, an open-source tool "Wurst" was created to import, filter, and modify the ecoinvent database according to the IMAGE-based scenarios (Beltran et al., 2020). Futura is another open-source tool for generating customized background databases (Joyce & Björklund, 2021). LCA practitioners can import a default database (e.g., ecoinvent) in Futura and then apply a variety of modifications to represent the technological changes under a given future scenario. The modifications include creating unit processes for new technologies, regionalizing new or existing unit processes, and altering the market composition of unit processes of different technologies (e.g., changing production volumes of different producers, etc.). The creation of a customized unit process in Futura (e.g., electricity generation with carbon capture, utilization and storage in a specific geographic location) is accomplished by linking the relevant existing unit processes from the default database.

4 | DISCUSSIONS

We developed a criteria matrix (Table 1) to characterize the key requirements of the reviewed methods to answer the first and second research questions defined in the earlier section.

We observed that all methods except "lineage and process ontologies" show a low degree of effort for post-treatment requirement, once the inventory data is selected or created. In addition, we found that the development of data-driven methods generally entails a high requirement of both subject-specific and method-specific knowledge. The method-specific knowledge can be minimal when the methods are applied to the same domain, for example, predict the yield of the same technical process under different operational conditions using a pretrained ANN model. Nonetheless, if the method is applied to a different domain, both subject-specific knowledge (e.g., understanding of the new technical process) and method-specific knowledge (e.g., train an ANN model from scratch) are typically required of the users. Another common observation of data-driven methods is the challenge of assessing the data quality uncertainty associated with the resulting inventory data. This is mostly due to the lack of a mechanism to reconcile the different data quality descriptions (e.g., geographical location, temporal difference) during the aggregation of source data. On the other hand, ML models are more amendable to automation and tend to have a higher generalizability, compared to other data-driven methods and methods from other approaches. This may be attributed to the fact that ML methods primarily aim to create a correlation between input variables and model output, rather than developing a causal relationship that is context-specific and hence often difficult to generalize.

Process simulation and theoretical calculations are mostly used for estimating the energy consumption, material use, and waste generation of the chemical and biological manufacturing processes. Both methods require a strong subject-specific knowledge (e.g., chemical engineering expertise, reaction stoichiometry, and standard enthalpy of formation for reactants). Method-specific knowledge is often a barrier hampering the use of process simulation software. The resulting inventory data from these two methods can be used directly (e.g., for quantifying the electricity consumption) and the data quality uncertainty can be accessed using the pedigree matrix (e.g., for technological difference between the simulation and real-world process).

Ontology-based methods are widely used in many domains and the applications can often be automated after the architecture is created manually. The knowledge required for building and querying the data graph of interest can be a barrier for both the method developers and users. Ontology-based methods can guide inventory data modeling (e.g., to identify a specific step of a chemical synthesis process), whereas additional modeling is required for creating or selecting proxy data.

Applying a learning curve requires subject-specific knowledge. The uncertainty due to the technological difference between the product system of interest and the reference may be evaluated by the pedigree matrix. Updating background data using future scenario modeling results requires both subject-specific knowledge (e.g., understanding the structure of a given LCA database) and method-specific knowledge (e.g., working knowledge of an IAM). Method-specific knowledge is required for the method developer to automate such complex alternations on an LCA database, yet understanding the procedure to use the method may still present a knowledge barrier for some users.

To answer the third research question (mentioned in Section 1), we identified limitations and several aspects of existing methods for future improvement. For example, the efficacy of the similarity-based prediction method for inventory data creation is highly volatile and the contributing factors are not systematically studied. The domain knowledge requirement is a barrier for nonexperts in using the method. Besides, the limited availability of training data for ML-based methods is a bottleneck constraining the efficacy and applicability of these methods, as a majority of the studies used the ecoinvent database which has a highly imbalanced coverage (e.g., number of products, variety of technological pathways) of industrial sectors. In addition, some methods have limited case studies, for instance, IAM can be powerful for prospective LCA, but the current use cases are predominantly focusing on energy sectors. Likewise, although effective, mechanistic approach has found limited applications outside the chemical industry.

For future improvement, the collaboration between LCA community and domain researchers should be enhanced, and it also will be helpful to clearly document the domain knowledge requirement to ease the use of the method by nonexpert users. Besides, improving the degree of automation will be beneficial, as a method with a higher degree of automation is more likely to "scale up" (i.e., being applied to a more extensive database or a different product system). To improve the degree of automation, leveraging the fast development of computer science and engineering is necessary. Diversifying the data sources for data-driven methods, ML, in particular, is crucial for improving the efficacy and applicability of these methods to different domains. For instance, a systematic review of the data sources available for applying ML algorithms to LCI data creation is necessary. Accordingly, the advantages and limitations of these data sources need to be assessed, which may entail the development of the LCA-specific evaluation metrics that complement the metrics already existing in the ML field. Besides improving the availability and quality of data, more ML algorithms should also be explored. For example, using matrix-based LCA, the flows of a foreground LCI model can be represented in a vector format (Heijungs, 2010). By training a generative adversarial network (GAN) model to learn the vector representations of different foreground LCI models in a database, it may be possible to infer the missing flow(s) of a new foreground LCI model by generating a vector based on certain conditions (e.g., industrial category or reaction type). Another potentially viable algorithm is the graph neural network (GNN) where the nodes can represent the activities (e.g., electricity generation) and the edges can be the flows (e.g., quantify of biomass input to the generation) of an LCI model. By training a GNN model to learn the relationship between nodes and edges represented in different LCI models, it may be possible to infer the missing flow(s) of a new LCI model (i.e., an incomplete graph) via the trained GNN model. Another important aspect for future research is developing uncertainty assessment methods for the inventory data created from models (e.g., ML-based methods), when existing methods (e.g., Pedigree matrix) are not applicable (e.g., geographic correlation cannot be assessed when the inventory data is created using the average value of the data collected from different geographic locations).To assist LCA practitioners in choosing suitable method(s) for bridging the data gap in LCI modeling, we recommend the consideration of the criteria in the following aspects. Firstly, the method selection should be aligned with the goal and scope of the intended LCA study. For example, if it is crucial to report the data quality, then the LCA practitioner should prioritize the use of those methods that have a low level of challenge in "handling data quality uncertainty." If the goal of the LCA study is to produce a generalized, easy-to-use model that minimizes the method-specific knowledge requirement for potential users, then methods with a low level of challenge in "automation" and "generalizability" should be considered. Another aspect is to align the "domain knowledge requirement" of the method with the available knowledge of the users. For instance, to train and adequately use an ANN model for predicting the electricity consumption of a chemical manufacturing process, both subject-specific knowledge (e.g., type of reaction, operational conditions) and method-specific knowledge (e.g., how to create an ANN architecture using Keras, an open-source python library for creating ANN models) are necessary.

5 | CONCLUSION

Three commonly used approaches to address the challenge of missing data in LCI modeling were identified in this study: Data-driven, mechanistic, and future inventory modeling. A total of 12 methods belonging to these 3 approaches were characterized according to a criteria matrix: Domain knowledge requirement (developer) and (user), post-treatment requirement, challenge in assessing data quality uncertainty, challenge in generalizability, and challenge in automation. We found that domain knowledge requirement was generally high for data-driven methods, ML-based methods, in particular. Many of the reviewed methods could be applied to a wide range of subjects and have the potential to be automated. The detailed characterization results of each method (Table 1) provided guidance for the use of the criteria matrix for future research. Our study shows that the efficacy of bridging the data gap in LCI modeling can be enhanced by leveraging the diverse expertise within the industrial ecology community, such as ML, statistics, and process simulation. Equally important is that the increased availability of complete and more high-fidelity LCI data can benefit the other industrial ecology methods, for instance, through providing additional technological information for material flow analysis (MFA) and physical input–output analysis.

ACKNOWLEDGMENTS

The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number RGPIN-2021-02841]. Support from the Faculty of Forestry at the University of British Columbia (Faculty of Forestry Doctoral Fellowship—FFDF) is gratefully acknowledged.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Shiva Zargar https://orcid.org/0000-0003-4374-8628
Yuan Yao https://orcid.org/0000-0001-9359-2030
Qingshi Tu https://orcid.org/0000-0001-7113-0564

REFERENCES

Bamber, N., Turner, I., Arulnathan, V., Li, Y., Zargar Ershadi, S., Smart, A., & Pelletier, N. (2020). Comparing sources and analysis of uncertainty in consequential and attributional life cycle assessment: Review of current practice and recommendations. *International Journal of Life Cycle Assessment*, *25*(1), 168–180. http://link.springer.com/10.1007/s11367-019-01663-1

Barrett, W. M., Takkellapati, S., Tadele, K., Martin, T. M., & Gonzalez, M. A. (2019). Linking molecular structure via functional group to chemical literature for establishing a reaction lineage for application to alternatives assessment. *ACS Sustainable Chemistry and Engineering*, *7*(8), 7630–7641. https://pubs.acs.org/doi/full/10.1021/acssuschemeng.8b05983

Canals, L. M. I., Azapagic, A., Doka, G., Jefferies, D., King, H., Mutel, C., Nemecek, T., Roches, A., Sim, S., Stichnothe, H., Thoma, G., & Williams, A. (2011). Approaches for addressing life cycle assessment data gaps for bio-based products. *Journal of Industrial Ecology*, *15*(5), 707–725. http://doi.org/10.1111/j.1530-9290.2011.00369.x

Cashman, S. A., Meyer, D. E., Edelen, A. N., Ingwersen, W. W., Abraham, J. P., Barrett, W. M., Gonzalez, M. A., Randall, P. M., Ruiz-Mercado, G., & Smith, R. L. (2016). Mining available data from the United States environmental protection agency to support rapid life cycle inventory modeling of chemical manufacturing. *Environmental Science and Technology*, *50*(17), 9013–9025. https://pubs.acs.org/sharingguidelines

Chen, T., & Benesty, M. (2016). *XGBoost: eXtreme gradient boosting*. R package version 0.4-3. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Xgboost%3A+Extreme+Gradient+Boosting%2C+R+package+version+0.4-2&btnG=

Ciroth, A., Muller, S., Weidema, B., & Lesage, P. (2016). Empirically based uncertainty factors for the pedigree matrix in ecoinvent. *International Journal of Life Cycle Assessment*, *21*(9), 1338–1348. https://link.springer.com/article/10.1007/s11367-013-0670-5

Cooper, J. S., & Kahn, E. (2012). Commentary on issues in data quality analysis in life cycle assessment. *International Journal of Life Cycle Assessment*, *17*(4), 499–503.

Cuéllar-Franca, R. M., García-Gutiérrez, P., Taylor, S. F. R., Hardacre, C., & Azapagic, A. (2016). A novel methodology for assessing the environmental sustainability of ionic liquids used for CO_2 capture. *Faraday Discussions*, *192*(0), 283–301. https://pubs.rsc.org/en/content/articlehtml/2016/fd/c6fd00054a

De, R., Bhartiya, S., & Shastri, Y. (2019). Multi-objective optimization of integrated biodiesel production and separation system. *Fuel*, *24*, 519–532.

Dellink, R., Chateau, J., Lanzi, E., & Magné, B. (2017). Long-term economic growth projections in the Shared Socioeconomic Pathways. *Global Environmental Change*, *42*, 200–214.

Ecoinvent. (2018). *How to interpret the uncertainty fields in ecoinvent. FAQs*. https://www.ecoinvent.org/support/faqs/methodology-of-ecoinvent-3/how-to-interpret-the-uncertainty-fields-in-ecoinvent.html

EPA. (2014). National emissions inventory. In 2014 National Emissions Inventory Data Facilities (May): Data obtained September 2019. https://www.epa.gov/air-emissions-inventories/national-emissions-inventory-nei

Fishman, T., Heeren, N., Pauliuk, S., Berrill, P., Tu, Q., Wolfram, P., & Hertwich, E. G. (2021). A comprehensive set of global scenarios of housing, mobility, and material efficiency for material cycles and energy systems modeling. *Journal of Industrial Ecology*, *25*(2), 305–320. https://onlinelibrary.wiley.com/doi/full/10.1111/jiec.13122

Fröhling, M., & Hiete, M. (2020). Sustainability and life cycle assessment in industrial biotechnology: A review of current approaches and future needs. In *Advances in biochemical engineering/biotechnology* (Vol. 173, pp. 143–203). Springer. https://link.springer.com/chapter/10.1007/10_2020_122

Han, D., Yang, X., Li, R., & Wu, Y. (2019). Environmental impact comparison of typical and resource-efficient biomass fast pyrolysis systems based on LCA and Aspen Plus simulation. *Journal of Cleaner Production*, *231*, 254–267.

Heijungs, R. (2010). Sensitivity coefficients for matrix-based LCA. *International Journal of Life Cycle Assessment*, . https://link.springer.com/article/10.1007/s11367-010-0158-5

Heijungs, R., & Huijbregts, M. A. J. (2004). A review of approaches to treat uncertainty in LCA. *IEMSs 2004 International Congress*: *8*. https://scholarsarchive.byu.edu/iemssconference/2004/all/197

Henriksen, T., Astrup, T. F., & Damgaard, A. (2020). Data representativeness in LCA: A framework for the systematic assessment of data quality relative to technology characteristics. *Journal of Industrial Ecology*, . https://onlinelibrary.wiley.com/doi/abs/10.1111/jiec.13048

Hou, P., Cai, J., Qu, S., & Xu, M. (2018). Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environmental Science and Technology*, *52*(9), 5259–5267. https://pubs.acs.org/doi/abs/10.1021/acs.est.7b05366

Huijbregts, M. A. J., Norris, G., Bretz, R., Ciroth, A., Maurice, B., Von Bahr, B., Weidema, B., & De Beaufort, A. S. H. (2001). Framework for modelling data uncertainty in life cycle inventories. *International Journal of Life Cycle Assessment*, *6*(3), 127–132. https://link.springer.com/article/10.1007/BF02978728

Innovation Canada. (2018). *Technology readiness levels*. https://www.ic.gc.ca/eic/site/080.nsf/eng/00002.html

ISO. (2006). *ISO 14044: Environmental management—Life cycle assessment—Requirements and guidelines*. The International Organization for Standardization (ISO).

Joyce, P. J., & Björklund, A. (2021). Futura: A new tool for transparent and shareable scenario analysis in prospective life cycle assessment. *Journal of Industrial Ecology*, . https://github.com/pjamesjoyce/futura

Kendall, A., Yuan, J., Brodt, S., & Jan, K. (2010). Carbon footprint of U.S. *honey production and packing report to the National Honey Board. Scenario: 1–23*.

Kim, S., & Overcash, M. (2003). Energy in chemical manufacturing processes: Gate-to-gate information for life cycle assessment. *Journal of Chemical Technology and Biotechnology*, *78*(9), 995–1005. http://doi.org/10.1002/jctb.821

Lan, K., Park, S., & Yao, Y. (2019). Key issue, challenges, and status quo of models for biofuel supply chain design. In *Biofuels for a more sustainable future: Life cycle sustainability assessment and multi-criteria decision making* (pp. 273–315). Elsevier.

Liao, M., Kelley, S., & Yao, Y. (2020). Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. *ACS Sustainable Chemistry and Engineering*, *8*(2), 1252–1261. https://pubs-acs-org.ezproxy.library.ubc.ca/doi/full/10.1021/acssuschemeng.9b06522

Liao, M., & Yao, Y. (2021). Applications of artificial intelligence-based modeling for bioenergy systems: A review. *GCB Bioenergy*. https://onlinelibrary.wiley.com/doi/full/10.1111/gcbb.12816

Mendoza Beltran, A., Cox, B., Mutel, C., van Vuuren, D. P., Font Vivanco, D., Deetman, S., Edelenbosch, O. Y., Guinée, J., & Tukker, A. (2020). When the background matters: Using scenarios from integrated assessment models in prospective life cycle assessment. *Journal of Industrial Ecology*, *24*(1), 64–79. www.wileyonlinelibrary.com/journal/jie

Meron, N., Blass, V., & Thoma, G. (2020). Selection of the most appropriate life cycle inventory dataset: New selection proxy methodology and case study application. *International Journal of Life Cycle Assessment*, *25*(4), 771–783. https://link.springer.com/article/10.1007/s11367-019-01721-8

Meyer, D. E., Cashman, S., & Gaglione, A. (2020). Improving the reliability of chemical manufacturing life cycle inventory constructed using secondary data. *Journal of Industrial Ecology*, . https://onlinelibrary.wiley.com/doi/abs/10.1111/jiec.13044

Meyer, D. E., Mittal, V. K., Ingwersen, W. W., Ruiz-Mercado, G. J., Barrett, W. M., Gonzalez, M. A., Abraham, J. P., & Smith, R. L. (2019). Purpose-driven reconciliation of approaches to estimate chemical releases. *ACS Sustainable Chemistry and Engineering*, *7*(1), 1260–1270. https://pubs.acs.org/doi/abs/10.1021/acssuschemeng.8b04923

Mittal, V. K., Bailin, S. C., Gonzalez, M. A., Meyer, D. E., Barrett, W. M., & Smith, R. L. (2018). Toward automated inventory modeling in life cycle assessment: The utility of semantic data modeling to predict real-world chemical production. *ACS Sustainable Chemistry and Engineering*, *6*(2), 1961–1976. https://pubs.acs.org/sharingguidelines

Montazeri, M., & Eckelman, M. J. (2016). Life cycle assessment of catechols from lignin depolymerization. *ACS Sustainable Chemistry and Engineering*, *4*(3), 708–718.

Morales-Mendoza, L. F., & Azzaro-Pantel, C. (2017). Bridging LCA data gaps by use of process simulation for energy generation. *Clean Technologies and Environmental Policy*, *19*(5), 1535–1546.

National Academies Press. (2014). *Identifying and reducing environmental health risks of chemicals in our society*, National Academies Press.

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., van Ruijven, B. J., van Vuuren, D. P., Birkmann, J., Kok, K., Levy, M., & Solecki, W. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change*, *42*, 169–180.

O'Neill, B. C., Kriegler, E., Riahi, K., Ebi, K. L., Hallegatte, S., Carter, T. R., Mathur, R., & van Vuuren, D. P. (2014). A new scenario framework for climate change research: The concept of shared socioeconomic pathways. *Climatic Change*, *122*(3), 387–400. https://link.springer.com/article/10.1007/s10584-013-0905-2

Ortmeier, C., Henningsen, N., Langer, A., Reiswich, A., Karl, A., & Herrmann, C. (2021). Framework for the integration of process mining into life cycle assessment. In *Procedia CIRP*, *98*, 163–168. https://www.sciencedirect.com/science/article/pii/S2212827121000470

Parvatker, A. G., & Eckelman, M. J. (2019). Comparative evaluation of chemical life cycle inventory generation methods and implications for life cycle assessment results. *ACS Sustainable Chemistry and Engineering*, *7*(1), 350–367. https://pubs.acs.org/sharingguidelines

Pereira, C., Hauner, I., Hungerbühler, K., & Papadokonstantakis, S. (2018). Gate-to-gate energy consumption in chemical batch plants: Statistical models based on reaction synthesis type. *ACS Sustainable Chemistry and Engineering*, *6*(5), 5784–5796. https://pubs.acs.org/doi/abs/10.1021/acssuschemeng.7b03769

Perry, J. H. (1950). Chemical engineers' handbook. *Journal of Chemical Education*, *27*(9), 533.

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., Bauer, N., Calvin, L., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J. C., KC, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., …, Tavoni, M. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, *42*, 153–168.

Righi, S., Dal Pozzo, A., Tugnoli, A., Raggi, A., Salieri, B., & Hischier, R. (2020). The availability of suitable datasets for the LCA analysis of chemical substances. In *Life cycle assessment in the chemical product chain: Challenges, methodological approaches and applications* (pp. 3–32). Springer. https://link.springer.com/chapter/10.1007/978-3-030-34424-5_1

Romeiko, X. X., Guo, Z., Pang, Y., Lee, E. K., & Zhang, X. (2020). Comparing machine learning approaches for predicting spatially explicit life cycle global warming and eutrophication impacts from corn production. *Sustainability (Switzerland)*, *12*(4), 1481. https://www.mdpi.com/2071-1050/12/4/1481/htm

Säynäjoki, A., Heinonen, J., & Junnila, S. (2012). A scenario analysis of the life cycle greenhouse gas emissions of a new residential area. *Environmental Research Letters*, *7*(3), 034037. https://iopscience.iop.org/article/10.1088/1748-9326/7/3/034037

Swingler, K. (2001). Applying neural network: A practical guide. Morgan Kauffman. https://books.google.com/books?hl=en&lr=&id=bq0YnP4BNKsC&oi=fnd&pg=PP7&ots=BXEtNR7LNk&sig=wotB9-NPjGIVU21_Zmnq-dc1LxQ

Tsoy, N., Steubing, B., van der Giesen, C., & Guinée, J. (2020). Upscaling methods used in ex ante life cycle assessment of emerging technologies: A review. *International Journal of Life Cycle Assessment*, *25*, 1680–1692.

Tu, Q., Parvatker, A., Garedew, M., Harris, C., Eckelman, M., Zimmerman, J. B., Anastas, P. T., & Lam, C. H. (2021). Electrocatalysis for chemical and fuel production: Investigating climate change mitigation potential and economic feasibility. *Environmental Science and Technology*, *55*(5), 3240–3249. https://pubs.acs.org/doi/abs/10.1021/acs.est.0c07309

Watson, C. I. (1992). *NIST Special Database 14*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.3485

Weidema, B. P., & Wesnæs, M. S. (1996). Data quality management for life cycle inventories—An example of using data quality indicators. *Journal of Cleaner Production*, *4*(3–4), 167–174.

Wernet, G., Hellweg, S., & Hungerbühler, K. (2012). A tiered approach to estimate inventory data and impacts of chemical products and mixtures. *International Journal of Life Cycle Assessment*, *17*(6), 720–728. http://www.sust-chem.ethz.ch/tools/finechem

Zargar, S., Jiang, J., Jiang, F., & Tu, Q. (2022). Isolation of lignin-containing cellulose nanocrystals: Life-cycle environmental impacts and opportunities for improvement. *Biofuels, Bioproducts and Biorefining*, *16*(1), 68–80.

Zhao, B., Shuai, C., Hou, P., Qu, S., & Xu, M. (2021). Estimation of unit process data for life cycle assessment using a decision tree-based approach. *Environmental Science & Technology*, *55*(12), 8439–8446. https://pubs.acs.org/doi/abs/10.1021/acs.est.0c07484

Zhu, X., Ho, C. H., & Wang, X. (2020). Application of life cycle assessment and machine learning for high-throughput screening of green chemical substitutes. *ACS Sustainable Chemistry and Engineering*, *8*(30), 11141–11151. https://pubs.acs.org/doi/full/10.1021/acssuschemeng.0c02211

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zargar, S., Yao, Y., & Tu, Q. (2022). A review of inventory modeling methods for missing data in life cycle assessment. *Journal of Industrial Ecology*, 1–14. https://doi.org/10.1111/jiec.13305