Combined use of MILP and multi-linear regression to simplify LCA studies

Janire Pascual-González[a], Carlos Pozo[a], Gonzalo Guillén-Gosálbez[a,b,*], Laureano Jiménez-Esteller[a]

[a] Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Campus Sescelades, Avinguda Països Catalans, 26, 43007 Tarragona, Spain
[b] Centre for Process Integration, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, UK

ARTICLE INFO

ABSTRACT

Life cycle assessment (LCA) has become the prevalent approach for quantifying the environmental impact of products over their entire life cycle. Unfortunately, LCA studies require large amounts of data that are difficult to collect in practice, which makes them expensive and time consuming. This work introduces a method that simplifies standard LCA studies by using proxy metrics that are identified following a systematic approach. Our method, which combines multi-linear regression and mixed-integer linear programming, builds in an automatic manner simplified multi-linear regression models of impact that predict (with high accuracy) the damage in different environmental categories from a reduced number of proxy metrics. Our approach was applied to data retrieved from ecoinvent. Numerical results show that few indicators suffice to describe the environmental performance of a process with high accuracy. Our findings will help develop general guidelines for simplified LCA studies that will focus on quantifying a reduced number of key indicators.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Life cycle assessment (LCA) has recently become the prevalent approach for quantifying the environmental impact of products and processes over their entire life cycle. LCA has expanded rapidly in both academia and industry, finding applications in a wide variety of fields. For instance, Finnveden et al. (2009) and Hellweg and Mila i Canals (2014) provided a review of recent developments in all LCA phases, including existing and emerging applications, whereas Jeswani et al. (2010) explored the options for broadening the LCA methodology beyond the current ISO framework for improved sustainability analysis.

One of the main drawbacks of LCA is that it requires large amounts of data of processes that are operated in disperse facilities across the product supply chain. In practice, gathering full information of the operations of complex, interrelated industrial systems including all emissions and activities for each of them is often prohibitive. First, since data collection tends to be highly time consuming and expensive, companies typically store information of only a subset of regulated compounds for which records are mandatory. Second, a full LCA may require data from external companies that might consider them too confidential to be released for external use. This situation creates data gaps that might affect critically the outcome of the LCA analysis, thereby leading to spurious conclusions and wrong advice. Data availability is therefore a major issue in sustainability assessment that can hinder the widespread adoption of sustainability principles in industry.

Streamlined LCA (SLCA) techniques aim to simplify the LCA analysis by reducing the amount of data required in the calculations (Marwah et al., 2011; Sundaravaradan et al., 2011). According to the Society of Environmental Toxicology and Chemistry (SETAC, 1999), SLCA methods can be roughly classified into 3 main groups that differ in the type of simplification underlying them: (1) those based on a contraction of the system boundary by which some upstream and/or downstream components are removed; (2) those based on the use of qualitative and/or less accurate data; and (3) those based on a reduction in the number of impact categories or inventory data.

* Corresponding author at: Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Campus Sescelades, Avinguda Països Catalans, 26, 43007 Tarragona, Spain. Tel.: +34 977558618; fax: +34 977559621.

E-mail addresses: janire.pascual@urv.cat (J. Pascual-González), carlos.pozo@urv.cat (C. Pozo), gonzalo.guillen@urv.cat (G. Guillén-Gosálbez), laureano.jimenez@urv.cat (L. Jiménez-Esteller).

Nomenclature

Acronyms

CED	cumulative energy demand
ICT	information and communications technology
LCA	life cycle assessment
LCI	life cycle inventory
LCIA	life cycle impact assessment
MILP	mixed-integer linear programming
MTrain	regression model obtained by minimizing the error in the training set
MVal	regression model obtained by minimizing the error in the validation set
SLCA	streamlined LCA

Index

i	LCIA metric
j	product
r	iteration of the algorithm

Sets

A	set of LCIA metrics whose values will be estimated from those of proxy LCIA metrics. Note that in our approach, this set is not defined beforehand.
B	set of proxy LCIA metrics whose values will be used to estimate those of other LCIA metrics. Note that in our approach, this set is not defined beforehand.
I	set of LCIA metrics
J	set of products
$ONE(r)$	set of binary variables whose value is 1 in the iteration r of the algorithm
R	set of iterations
$ZERO(r)$	set of binary variables whose value is 0 in iteration r of the algorithm solution in the iteration r of the algorithm

Parameters

$\underline{b}$	lower bound on the regression coefficient
$\overline{b}$	upper bound on the regression coefficient
n	number of LCIA metrics to be included as proxy indicators in the regression model
$y_{ob}(j, i)$	"true" value of metric i in observation j, which is obtained from a detailed LCA analysis

Variables

ARE	average relative error of the multi-linear regression model
$b(i, i')$	regression coefficient of proxy metric i in the predictive multi-linear equation of metric i'
$bin(i)$	binary variable that equals 1 if metric i is used as proxy indicator to estimate the value of other metrics and 0 otherwise
$bin(i, r)^*$	value of the ith component of the vector of binary variables in the optimal solution of iteration r
$y_{pr}(j, i)$	predicted value of LCIA metric i in observation j

Group 3 methods, which constitute so far the most widespread approach, restrict the analysis to a specific subset of life cycle inventory (LCI) entries and/or life cycle impact assessment (LCIA) categories. This approach has found applications in many sectors, including vehicle development (Arena et al., 2013; Moriarty and Honnery, 2008), oil refineries and industrial facilities (Weston et al., 2011), global warming potential evaluation (Bala et al., 2010), coal-fired electricity plants (Steinmann et al., 2014), pharmaceuticals (Jiménez-González et al., 2013), and food processing (Sanjuán et al., 2014), among others.

The simplification of the LCA study might come at the cost of excluding important environmental factors, thereby leading to uncertainties as well as potentially wrong conclusions (Hunt et al., 1998). To avoid this, SLCA studies are constructed using detailed knowledge of the process. This makes standard SLCA studies very specific, that is, they are only valid for particular industrial sectors and cannot be readily applied to other areas.

In this work the focus is on simplifying LCA studies by reducing the number of impacts to be assessed. A large amount of LCIA metrics are presently available, but no consensus has been reached yet on which one should be universally adopted. Quantifying all of them is highly data intensive, because it requires detailed information on many primary feedstocks, emissions and waste.

In a pioneering work, Huijbregts et al. (2006, 2010) proposed to use the cumulative energy demand to predict other LCIA metrics through linear regression. Hanes et al. (2013) showed that this approach has some limitations stemming from the use of a single log transformed metric. The debate on which indicator to use continues, but so far no systematic method has been developed to provide insight into this problem.

This work introduces a rigorous approach for selecting proxy LCIA metrics in streamlined LCA analysis. Our systematic method relies on a novel mixed-integer linear programming (MILP) model that identifies a reduced subset of key impacts that are used to estimate the others through multi-linear regression. The main advantages of this methodology are two-fold. First, no significant environmental data will be lost, since all the LCIA metrics are either measured or estimated. Second, it requires no aprioristic knowledge on the system, because the selection of metrics is performed using a systematic approach based on discrete-continuous optimization.

Our approach has been applied to data retrieved from ecoinvent as a first step towards its future application to a more complete dataset constructed from several LCA databases (i.e., GaBi, Simapro, ecoinvent, ELCD, NREL) (LBP, 2015; National Renewable Energy Laboratory, 2012; Simapro manual PRe Consultants, 2013; Swiss Centre For Life Cycle Inventories, 2013; Wolf et al., 2008). Numerical results show that few indicators suffice to describe the environmental performance of a process with high accuracy and that several LCIA metrics tend to be highly correlated (Huijbregts et al., 2006, 2010). The application of our algorithm provides deep insight into the relationships between impacts. This fundamental knowledge might be used to develop other streamlined LCA methods as well as more efficient environmental regulations.

The paper is organized as follow. Section 2 provides the problem statement, while Section 3 describes the mathematical formulation. Section 4 describes the ecoinvent database and presents the numerical results. In Section 5, the conclusions of the work are drawn.

2. Problem statement

The problem under study can be formally stated as follows. We are given environmental data expressed in the form of a matrix containing $|I|$ LCIA metrics i and $|J|$ observations j (each one corresponding to a different product). The goal of the analysis is to first identify, among the whole range of LCIA metrics available, a given number of them that will be taken as a basis for building regression models of impact. These regression models will then be used for estimating other impacts with the maximum accuracy possible (which are not quantified using LCI data, but rather predicted from the proxy impact values).