Computers and Chemical Engineering 108 (2018) 179–193

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Predicting the cradle-to-gate environmental impact of chemicals from molecular descriptors and thermodynamic properties via mixed-integer programming

Raul Calvo-Serrano[a], María González-Miquel[b], Stavros Papadokonstantakis[c], Gonzalo Guillén-Gosálbez[a,d,*]

[a] Centre for Process Systems Engineering (CPSE), Imperial College London, SW7 2AZ, United Kingdom
[b] School of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom
[c] Chalmers University of Technology, Maskingränd 2, 412 58, Göteborg, Sweden
[d] Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007, Tarragona, Spain

ARTICLE INFO

Article history:
Received 10 July 2017
Received in revised form 5 September 2017
Accepted 10 September 2017
Available online 12 September 2017

ABSTRACT

Life Cycle Assessment (LCA) has recently gained wide acceptance in the environmental impact evaluation of chemicals. Unfortunately, LCA studies require large amounts of data that are hard to gather in practice, a critical limitation when assessing the processes and value chains present in the chemical industry. We here develop an approach that predicts the cradle-to-gate life cycle production impact of organic chemicals from attributes related to their molecular structure and thermodynamic properties. This method is based on a mixed-integer programming (MIP) optimisation framework that systematically constructs short-cut predictive models of life cycle impact. On applying our approach to a data set containing 88 chemicals, 17 molecular descriptors and 15 thermodynamic properties, we estimate with enough accuracy (for the purposes of a standard LCA) several impact categories widely applied in LCA studies, including the cumulative energy demand, global warming potential and Eco-indicator 99. Our framework ultimately leads to linear models that can be easily integrated into existing modelling and optimisation software, thereby facilitating the design of more sustainable processes.

1. Introduction

The chemical industry is at present striving to decrease its environmental footprint in the transition towards a more sustainable manufacturing sector. In this context, systematic methods are required to assist in the environmental assessment and optimisation of chemicals, identification of critical hotspots across products' supply chains and definition of guidelines to effectively retrofit processes so that they adhere to sustainability principles. The amount of new chemicals produced annually has exponentially increased in the last decades. The database CAS registry (http://www.cas.org/, 2016) from Chemical Abstracts Service, a division of the American Chemical Society devoted to authoritatively collect disclosed chemical substance information, contains at present more than 120 million organic and inorganic substances and 66 million sequences, with approximately 15,000 new chemicals being added each day. Hence, an enormous amount of chemical species exists and for many of them the corresponding chemical hazards and potential environmental impacts are not fully understood or characterised. This information, however, is critical for the proper evaluation of their wider sustainability impacts and future market feasibility.

Under the above scenario, the recent trend towards the development of more sustainable products has led to a plethora of environmental assessment tools(Constable et al., 2002; Tobiszewsk, 2016). Among them, Life Cycle Assessment (LCA) (Guinée et al., 1993a,b; Sheldon, 2015) has become (arguably) the prevalent approach to quantify the environmental burdens of products from cradle to grave, finding applications in many areas, including the holistic assessment of the environmental footprint of chemicals considering all the stages in their life cycle (Anastas and Lankey, 2000; Azapagic and Clift, 1999; Bojarski et al., 2009; Burgess and Brennan, 2001; Cespi et al., 2015; Gerber et al., 2011; Jimenez-Gonzalez and Overcash, 2014; Khoo et al., 2016; Kralisch et al., 2014; Wen and Shonnard, 2003). Furthermore, the combined use of LCA and optimization has gained increasing popularity in process systems engineering and is now being used to tackle a wide variety of prob-

* Corresponding author at: Centre for Process Systems Engineering (CPSE), Imperial College London, SW7 2AZ, United Kingdom.
E-mail address: g.guillen05@ic.ac.uk (G. Guillén-Gosálbez).

http://dx.doi.org/10.1016/j.compchemeng.2017.09.010
0098-1354/

Nomenclature

Acronyms

LCA	Life cycle assessment
MIP	Mixed-integer programming
SLCA	Streamlined life cycle assessment
MLR	Multi-linear regression
ANN	Artificial neural network
CED	Cumulative energy demand
GWP	Global warming potential
COD	Chemical oxygen demand
BOD5	Biological oxygen demand
TOC	Total organic carbon
EI99 Total	Eco-indicator 99 total
EI99 HH	Eco-indicator 99 human health
EI99 EQ	Eco-indicator 99 ecosystem quality
EI99 Res	Eco-indicator 99 resources
LOOCV	Leave one out cross validation
MIQCP	Mixed-integer quadratically constrained programming
MINLP	Mixed-integer non-linear programming
ARE	Average relative error
MD	Molecular descriptors
TP	Thermodynamic properties

Indices

i	Chemical
j	Chemical attribute

Sets

I	Set of chemicals
J	Set of chemical attributes

Parameters

x_{ij}	Normalised value of attribute j in chemical i
$y_o\ i$	Normalised value of the impact category being predicted in chemical i
$\overline{b_j}$	Upper bound on the regression coefficient b_j
$\underline{b_j}$	Lower bound on the regression coefficient b_j
$\overline{k}$	Number of attributes to consider in the regression

Variables

a	Constant regression coefficient
b_j	Regression coefficient for attribute j
z_j	Binary variable that equals one if attribute j is used in the regression and zero otherwise
RSS	Sum of squared residuals
AIC	Akaike information criterion

lems (Azapagic, 1999; Galán-Martín et al., 2016; Guillén-Gosálbez and Grossmann, 2009; Mele et al., 2011; Pieragostini et al., 2012; You and Wang, 2011).

Unfortunately, LCA requires large amounts of data from several echelons in the product's supply chain, which are often owned by different companies that might be reluctant to share this environmental information and/or might even lack the necessary measurements. This limitation is particularly critical in the chemical industry, where complex process networks operate exchanging energy, mass and water and consuming a wide variety of intermediate products and feedstocks.

LCA calculations can be simplified via Streamlined LCA (SLCA) (Shatkin and Larsen, 2011; Todd and Curran, 1999), whose goal is to reduce the upstream and downstream information required in a standard LCA by using proxy data, qualitative models and/or regression equations (Hunt et al., 1998). To increase their accuracy, these methods must be tailored to a given sector, ensuring that the simplifications and assumptions made hold in that particular context. Specific SLCA methods have been developed for a wide range of systems, such as buildings (Malmqvist et al., 2011; Yeo et al., 2016; Zabalza Bribián et al., 2009), water treatment plants (Quirante and Caballero, 2016; Schulz et al., 2012), power plants (Moreau et al., 2012), oil refineries (Weston et al., 2011), general chemical processes (Eckelman, 2016; Guillen-Gosalbez et al., 2007; Hugo et al., 2004; Karka et al., 2014; Marvuglia et al., 2015; Tula et al., 2017; Wernet et al., 2009, 2008) and other manufacturing facilities (Kaebemick et al., 2003).

In a seminal work, Wernet et al. (2008) proposed a method to estimate the life cycle impact of several organic chemicals from specific molecular descriptors including the molecular weight and the number of several functional groups (e.g. hydroxyl groups, chlorine atoms, etc.). Following this approach, the authors applied Multi-Linear Regression (MLR) and Artificial Neural Networks (ANN) to a data set containing 103 chemicals in order to predict several impact categories (i.e. LCA metrics), including the Cumulative Energy Demand (CED), a widely used LCA metric previously found to correlate with several life cycle impacts (Huijbregts et al., 2010). In a follow-up work, Wernet et al. (2009) applied the ANN model to a larger data set of chemicals. These works designate the state-of-the-art with respect to publicly available model predictive life cycle impact assessment using molecular information. Life cycle inventory estimation methods have also been presented by GlaxoSmithKline (Curzons et al., 2007; Jiménez-González et al., 2004), however the associated data and model transparency is very low due to confidentiality reasons.

Here we develop an SLCA approach for simplifying the environmental impact assessment of chemicals that makes use of mathematical techniques to predict their life cycle environmental footprint from information readily available in practice. Our computational framework shows three main novelties compared to existing approaches: (i) it includes thermodynamic properties in the predictive models, which greatly improve the quality of the estimate when compared to predictions based exclusively on molecular descriptors (as was the case in Wernet et al., 2008); (ii) it identifies in an automatic manner chemical attributes that lead to better predictions and low risk of over-fitting; and (iii) it relies on simple linear models that can be easily integrated into software packages for process and molecular simulation. Our approach is based on MIP techniques that construct in an automatic manner linear regression models of environmental impact using binary variables to select chemical attributes and continuous variables to represent regression coefficients. This MIP approach, which can be easily extended to construct Quantitative Structure-Activity Relationships (QSAR) (Birkved and Heijungs, 2011; Eklund et al., 2012; Pirhadi et al., 2015; Yang et al., 2011) models for other environmental, health and safety properties, opens up new research avenues for the incorporation of sustainability principles in the design of greener chemical products and processes (e.g. Bio-chemicals and Bio-fuels (Karka et al., 2014)).

The paper is organized as follows: First we specify the problem to solve together with the underlying assumptions and the methodology proposed to solve it. Then, we present in detail the modelling approach and solution procedure. Finally, we demonstrate the capabilities of the proposed methodology using the same dataset presented by Wernet et al. (2008).

2. Problem statement

Our hypothesis is that thermodynamic properties contain a plethora of information concerning the energy, water and feedstock