UNIVERSITAT ROVIRA I VIRGILI

DEVELOPMENT OF SYSTEMATIC METHODS FOR THE ASSESSMENT AND OPTIMIZATION OF LIFE CYCLE ENVIRONMENTAL IMPACTS.

Janire Pascual González

Dipòsit Legal: T 1570-2015

DOCTORAL THESIS

DEVELOPMENT OF SYSTEMATIC METHODS FOR THE ASSESSMENT AND OPTIMIZATION OF LIFE CYCLE ENVIRONMENTAL IMPACTS

Janire Pascual González

Janire Pascual González

DEVELOPMENT OF SYSTEMATIC METHODS FOR THE ASSESSMENT AND OPTIMIZATION OF LIFE CYCLE ENVIRONMENTAL IMPACTS

DOCTORAL THESIS

Supervised by: Dr. Gonzalo Guillén Gosálbez
Dr. Laureano Jiménez Esteller

Department of Chemical Engineering

SUSCAPE research group

Universitat Rovira i Virgili

Tarragona, 2015

Universitat Rovira i Virgili

Department of Chemical Engineering

Av. Països Catalans, 26

47007-Tarragona, Spain

Phone +34 977 55 86 43

Dr. Gonzalo Guillén-Gosálbez and Dr. Laureano Jiménez Esteller

CERTIFY:

That the present study entitled "Development of systematic methods for the assessment and optimization of life cycle environmental impacts", presented by Janire Pascual González for the award of the degree of Doctor, has been carried out under our supervision at the Chemical Engineering Department of the University Rovira i Virgili.

Tarragona, 22th July 2015,

Dr. Gonzalo Guillén Gosálbez Dr. Laureano Jiménez Esteller

AGRADECIMIENTOS

Esta tesis no habría sido posible sin la constante ayuda y dedicación de mis supervisores Dr. Gonzalo Guillén y Dr. Laureano Jiménez. Gracias a vuestros consejos y apoyo he podido crecer como investigadora y como persona. Agradezco también a la Universidad Rovira i Virgili, al departamento de Ingeniería Química y al grupo de investigación SUSCAPE la oportunidad que me brindaron para incorporarme al programa de doctorado. También quisiera agradecer su colaboración en esta tesis doctoral al Dr. Josep Maria Mateo, Dr. Ignacio Grossmann y Dr. Jeffrey Siirola.

A mis compañeros del SUSCAPE, gracias por compartir conmigo estos tres años, por presarme vuestra ayuda siempre que la he necesitado y por todos los momentos divertidos que hemos vivido. No podría olvidarme de mis amigos del SEES:lab, gracias por todos esos momentos de descanso, por las risas y por los debates, habéis hecho que todo sea mucho más fácil.

Gracias también a mi familia por estar siempre a mi lado. Gracias por vuestro apoyo y confianza en todo momento. En especial, quiero darle las gracias a mi madre, sin tu esfuerzo yo no podría estar escribiendo estas líneas. Gracias a ti he llegado hasta aquí y soy la persona que quería ser. Os quiero mucho.

A Toni, gracias por creer en mí incluso en los momentos en que yo misma no creía. Gracias por tu incondicional apoyo y por transmitirme tu seguridad. Gracias por hacerme sonreír cada día. T’estimo molt.

Por último, gracias a todos los que han estado a mi lado todo este tiempo y no he nombrado. De corazón: muchas gracias.

SUMMARY

Engineers are nowadays encouraged to incorporate the principles of sustainability in new and existing facilities. In this scenario, the question that still remains is how to assess the environmental sustainability of a product in a fast, reliable, accurate and economic manner. As an example, there are several millions of products officially recognized by the United Nations, but so far we have only fully assessed the global environmental impact of a few thousand. This is so because in today's globalized markets environmental studies have become data intensive and time consuming, while at the same time new technologies and products are being developed at a very fast pace. In this challenging context, the widespread adoption of sustainability principles in industry will hardly take place without the proper decision-making support tools.

It is important to follow a sustainable approach not only at the engineering scale, but also at a wider global scale. Nowadays, governments are looking for policies promoting a more sustainable development. However, the design of these polices is challenging, particularly when several countries are involved in the life cycle of products and detailed knowledge of the international channels through which goods and services are traded is required.

This thesis proposes systematic methods for the assessment and optimization of life cycle impact assessment (LCIA) metrics that measure the environmental performance of a system, either at the smaller engineering scale or at the bigger macroeconomic level. The methods investigated include multivariate statistical analysis, life cycle assessment, environmentally extended multi-regional input-output models, mixed

integer-linear programming and multi-objective optimization. We apply these methods to international databases to shed light on how the environmental impacts are generated and how to design environmental policies in a simpler and effective way, at both engineering and global scales.

The work compiled in this PhD dissertation comprises four papers for their publication in international peer reviewed journals (three of them have been published, and the last one is ready to be submitted). The first paper describes the application of multivariate statistical analysis to assess the relationship between the different LCIA metrics at an engineering scale [1]. Then, we apply optimization tools to identify proxy LCIA metrics that can simplify the LCA analysis by reducing the amount of input data required [2]. In the other two papers macroeconomic data are analyzed. First, we use multivariate statistical analysis to identify environmental impact patterns at a global scale [3], and finally, we develop a multi-objective optimization model to identify economic sectors whose regulation will lead to a reduction of the environmental impact at a global scale with minimum changes in the economic flows [4].

A comprehensive multivariate statistical analysis of LCIA data of 4087 products classified into 18 categories is performed to study the level of correlation between impacts [1]. A total of 32 LCIA metrics, calculated using three different methodologies (cumulative energy demand (CED), impact-oriented characterization (CML 2001) and eco-indicator 99 (Eco-99)), are investigated. In this approach, the starting point is the identification and elimination of outliers using the robust correlation matrix. Then, a correlation analysis based on the Pearson's correlation coefficient is performed considering a level of significance of 0.001. The dispersion of the data is assessed using the coefficient of variation. Finally, least-squares linear regression models are constructed in order to evaluate whether there is a single metric which suffices to predict the others. Results show a high level of correlation between most impact metrics and that the most correlated categories are those with a low coefficient of variation. Despite these findings, it is not possible to make accurate predictions using a single proxy LCIA metric.

The assessment of environmental metrics explained above shows that many LCIA metrics are highly correlated. In view of this, it might be possible to develop streamlined LCA methods capable of predicting environmental metrics with accuracy from a reduced set of proxy indicators. Bearing this goal in mind, in the second paper we propose a rigorous approach [2] based on the combined use of multi-linear regression and mixed-integer linear programming for predicting in a fast, reliable and accurate manner the damage caused by a product over its entire life cycle from a reduced number of proxy environmental metrics. This methodology is applied to data retrieved from the ecoinvent database considering 17 LCIA metrics associated with products in the electricity and oil categories. This approach requires no aprioristic knowledge on the system. Results show that few indicators suffice to describe the environmental performance with accuracy. This simplification can lead to significant savings in time and resources during data collection in LCA analysis.

We next perform the multivariate statistical analysis explained below to macroeconomic data to identify global environmental impact patterns [3]. The aim of this section is to study the impact patterns of the wealthiest nations using environmentally extended multi-regional input-output tables. The statistical analysis is performed on data retrieved from the World Input-

Output Database. This database covers more than 30 million economic transactions taking place between 35 economic sectors of 40 countries (27 EU countries and 13 other major countries that represent 85% of the world's gross domestic product) and 69 environmental metrics classified into air emissions, land occupation, and the consumption of energy, water and natural resources. In this case study, before performing the statistical analysis, we first calculate the consumption-based impact of nations using multi-regional input-output models. Assessing the consumption-based impact instead of the production-based impact avoids the potential masking of the environmental impact that occurs when displacing the manufacturing tasks to countries with soft environmental regulations. The outcome of the statistical analysis applied to these data shows that most of the nations have similar environmental impact patterns despite polluting with different intensity. This information can be very useful during the development of unified environmental regulations.

The identification of environmental impact patterns provides us with the possibility of designing effective environmental policies, which can be applied to countries displaying similar pollution patterns. In paper four [4] we apply a systematic multi-objective optimization approach that provides decision-support for environmental policy makers. This method simultaneously minimizes the global CO_2 emissions (assessed via LCA) and maximizes the demand satisfaction of a nation. This approach relies on a bi-objective linear programming model that contains the basic equations of a multi-regional environmentally extended input-output table. The calculations are performed using data retrieved from the World Input-Output database for the year 2009. Numerical results produced for the case of US show that it is possible to reduce the global CO_2 emissions with little impact on the US economy by controlling key economic sectors. Furthermore, we observe that the global CO_2 emissions can drop by approximately 2% (without any perturbation on the economy) if we increase the share of cleaner energy sources (*i.e.*, shale gas) in the electricity grid of US.

TABLE OF CONTENTS

1 INTRODUCTION

In the recent past, there has been a growing interest on the development of more sustainable products. The question of how to assess the environmental performance of a product in a fast, reliable, accurate and economic manner remains still open. What has become clear is that the three main pillars of sustainability (economic performance and social and environmental impact) should be assessed over the entire life cycle of a product. Life cycle assessment (LCA) is a well-established methodology that quantifies the life-cycle environmental impact of a product following well defined and documented principles and guidelines [5, 6]. A wide variety of impact assessment methods based on LCA currently exist [7]. However, their calculation requires large amounts of data that are difficult to collect in practice.

At the macroeconomic level, globalization of markets and the consequent international trade have accelerated the socioeconomic development of nations, but have in turn led to undesirable effects like the externalization of environmental impacts. The design of sustainable policies becomes therefore challenging when more than one country is involved in the life cycle of a product. The environmental impact should be assessed on a life cycle basis and across nations in order to avoid outsourcing. However, performing an LCA study at the macroeconomic level is often hindered by the lack of information on the life cycle of products being internationally traded.

This thesis is devoted to overcoming such limitations by developing a set of systematic methods for assessing and optimizing life cycle impact assessment (LCIA) metrics from a sustainable perspective at both the macroeconomic and engineering scales. The final goal is to facilitate the development of simpler and effective environmental policies.

Fig. 1. illustrates the work developed in this thesis. We first conduct a multivariate statistical analysis using environmental data at the engineering scale in order to assess relationships between LCIA metrics. Then, we apply optimization tools to identify proxy LCIA metrics that will reduce the amount of data required for the execution of an LCA analysis. After that, we perform a multivariate statistical analysis to identify environmental impact patterns at a global scale using environmentally extended multi-regional input-output models (EEMRIO). Finally, we develop a multi-objective optimization (MOO) model to identify which economic sectors must be regulated first in order to reduce the environmental impact at a global scale with minimum changes in an economy.

The document is organized as follows. Section 1 introduces the challenging problems addressed in this thesis and provides a general background on mathematical programming and other techniques used in this thesis. The assessment of environmental impact metrics at an engineering scale is presented in section 2. The following section (section 3) applies optimization tools to identify proxy LCIA metrics. Then, in section 4, macroeconomic data are first assessed using multivariate statistical analysis tools. Finally, in section 5 a MOO problem is proposed to design effective environmental policies.

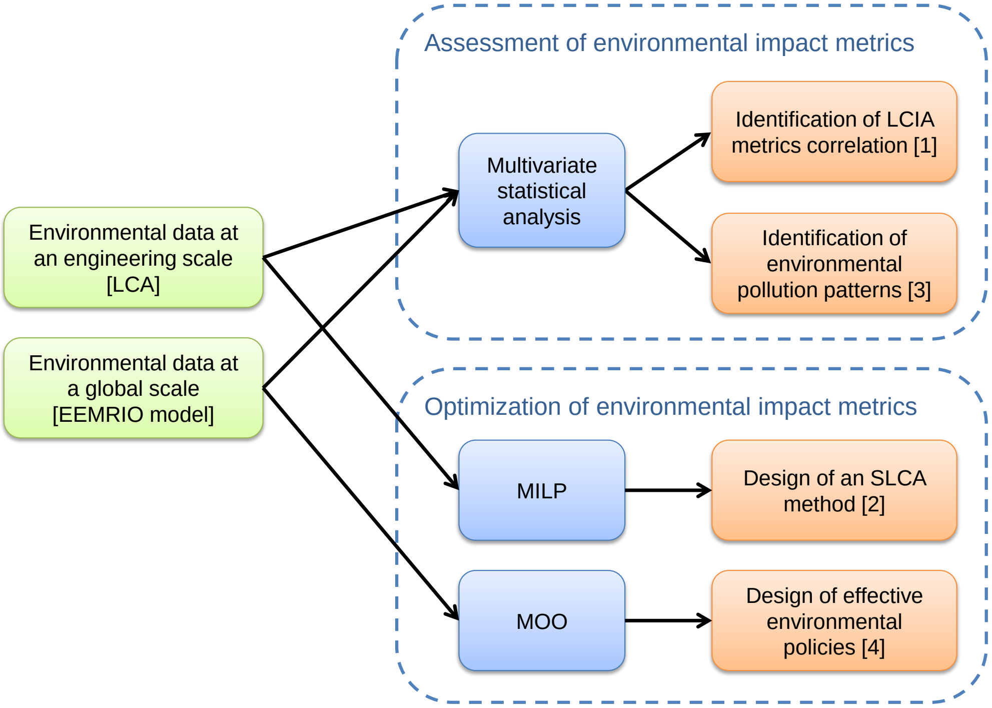

Assessment of environmental impact metrics
Environmental data at an engineering scale [LCA]
Environmental data at a global scale [EEMRIO model]
Multivariate statistical analysis
Identification of LCIA metrics correlation [1]
Identification of environmental pollution patterns [3]
Optimization of environmental impact metrics
MILP
Design of an SLCA method [2]
MOO
Design of effective environmental policies [4]

Fig. 1. Roadmap of the thesis. Green squares are the input data used at both the engineering and macroeconomic levels, blue squares are the methodologies presented in this thesis and orange squares represent the objectives of our work. The references to papers in which the objectives are accomplished are included in brackets.

1.1 Objectives

The particular objectives of this doctoral thesis are:

- To study the relationship between LCIA metrics at an engineering scale by applying a multivariate statistical analysis in order to determine if impacts within a given category are correlated and the intensity of such correlation.

- To develop a rigorous and systematic approach for identifying and selecting a reduced subset of proxy LCIA metrics to be used in simplified multi-linear regression models that predict other impacts with high accuracy.

- To analyze environmental impact patterns at a global scale by applying a statistical analysis on data retrieved from an environmentally extended multi-regional input-output (EEMRIO) table covering a wide variety of nations and impacts.

- To develop a systematic tool for identifying economic activities that need to be modified in order to reduce the environmental impact to the maximum extent possible while minimizing, at the same time, the changes to be performed in an economy.

1.2 Assessment of environmental impacts: Life Cycle Assessment

There is no consensus yet on how to assess the environmental sustainability of a product in a reliable and accurate manner. As a result, a wide range of impact assessment methodologies have been developed for quantifying the life cycle environmental impact of a product or a process. Among them, life cycle assessment (LCA) has become the prevalent approach [5, 8, 9].

LCA is a well-established methodology that has recently expanded rapidly in both academia and industry, finding applications in a wide variety of fields [8]. The main merit of LCA lies in the holistic view adopted, which avoids shifting environmental burdens between echelons of the product supply chain. A standard LCA comprises four main phases [6]:

- goal and scope definition, where the boundaries of the analysis are defined;

- inventory phase, which calculates the life cycle inventory of inputs and outputs associated with the product (*i.e.*, life cycle emissions to air, soil and water, amount of waste generated and feedstock requirements);

- impact assessment phase, which determines the life cycle impact in several damage categories from the life cycle inventory; and

- interpretation of results (results are analyzed and recommendations are made to improve the life cycle environmental performance of the product).

In today's globalized markets, the calculation of life cycle impact assessment (LCIA) metrics has become highly data intensive and time consuming. This thesis seeks to simplify these calculations and give a clear answer to the open challenges in this area, which are discussed in detail in the ensuing sections.

1.3 LCA at an engineering scale: challenges and proposed solutions

LCA studies need to collect large amounts of product data in the inventory phase, which are required to fully characterize the whole range of upstream and downstream processes associated with a given product [10]. In practice, gathering full information of the operations of complex, interrelated industrial systems including all emissions and activities for each of them is often a prohibitive task for several reasons. First, data collection tends to be highly time consuming and expensive, and for this reason 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 (in many cases, instead of having data of each industry, data available are usually an average of the data of all industries within a sector or band-specific). Data availability is therefore a major issue in sustainability assessment that can hinder the widespread adoption of sustainability principles in industry.

Streamlined LCA methods (SLCA) have been devised to this end [11, 12]. The concept of SLCA appeared originally to reduce the amount of data required by a standard full LCA [13]. The goal of a SLCA is to approximate the results of a full LCA (*i.e.*, the one that could be developed with full information of all the industrial processes related to the main product) but using less data.

According to the Society of Environmental Toxicology and Chemistry [13], SLCA methods typically follow nine approaches, which can be roughly classified into three main groups: (1) to contract the system boundary and remove some upstream and/or downstream components; (2) to use qualitative and/or less accurate data; and (3) to calculate the impact from selected inventory entries.

Most of the SLCA methods developed so far belong to the third group of approaches. Part of the research efforts here have focused on defining a universal proxy indicator which could be used to predict a wide range of life cycle impacts [14–17]. In addition, customized streamlined methods have been developed for many industrial sectors, including vehicle development [18, 19], oil refineries and industrial facilities [10], coal-fired electricity plants [20], pharmaceuticals [21], food processing [22] and plastic bags and recycled materials [23]. However, these studies require a detailed knowledge of the process in order to select the appropriate proxy indicator and thus avoid potentially wrong conclusions. This makes standard SLCA case-dependant and for this reason these methods cannot be readily applied to other areas. With new technology and products being developed at a very fast pace, the use of proxy impact indicators would simplify LCA studies to a large extent, since only the data related to the quantification of the proxy would be required. Furthermore, comparisons between alternative products would become easier, as they could be performed on the basis of an analysis of a single category. Therefore, developing fast, reliable and accurate SLCA methods capable of predicting the life cycle impact of a product from limited data readily quantified in practice is a priority to ensure a more sustainable development.

1.3.1 Towards a systematic streamlined life cycle analysis

1.3.1.1 Identification of correlated metrics

The extent to which the data required by a standard LCA can be simplified depends on the decisions to be made along with the final goal of the analysis. Most SLCA strategies use specific data to represent impacts or life cycle inventory entries. These methods attempt to identify a specific subset of LCIA categories from which to predict the outcome of a full space LCA with the maximum possible accuracy. In general, SLCA studies exclude factors that are relevant for the analysis, thereby leading to uncertainties as well as potentially wrong conclusions [24]. To minimize this effect, it is crucial to make the right simplifications (*e.g.* remove the proper upstream/downstream components, use appropriate surrogate data, identify key inventory entries to be used as proxy, etc.).

A first step towards this goal would involve the identification of correlated metrics which could be omitted from the study. Despite its importance, the study of the relationships between environmental metrics has received little attention to day.

Multivariate statistical analysis becomes a powerful tool to shed light on how to simplify the calculation of LCIA metrics, as it is the area of statistics that deals with observations made on many variables (*i.e.,* environmental impacts) [25]. These tools allow us to analyze and quantify the extent to which environmental impacts are correlated, an information that can be used to simplify LCA studies.

In this thesis, we use the correlation matrix, based on Pearson's correlation coefficient [26, 27], to study the relationships between impact metrics (see section *2.5.Multivariate statistical analysis* in [1]). In addition to the correlation analysis, the dispersion of the data is assessed using the coefficient of variation. This analysis requires data on LCIA metrics that can be retrieved from several LCA databases (*i.e.,* GaBi, Simapro, ecoinvent, ELCD, NREL) [28–32]. Without loss of generality, in this thesis the environmental data at the engineering scale have been retrieved from the ecoinvent database (see section *2.1 Ecoinvent Database* in [1] and section *4.1 Ecoinvent Database* in [2] for further details).

The statistical analysis performed in this thesis is applied considering data split into 18 categories. We consider that a data set is heterogeneous when the products/technologies within the group display dissimilar features (*e.g.,* chemicals). On the opposite case, when the processes within a category are similar, the set is homogeneous (*e.g.,* oil). The aim of this analysis is therefore to assess the level of correlation and dispersion of the data.

The identification of impacts with similar behavior has a two benefit: (i) It provides valuable insight for developing SLCA methods in which redundant metrics are omitted with little loss of information; (ii) it assists in the development of simpler environmental policies focusing on regulating a reduced number of proxy impacts.

1.3.1.2 Combination of optimization with multi-linear regression

Once the relationships between LCIA metrics are understood, the next step involves the development of customized SLCA methods based on this information. In this thesis, we propose to combine multi-linear regression models with mathematical programming (see section 1.5.2.1) for systematically selecting proxy LCIA metrics in streamlined LCA analysis (see [2]). In particular, we propose to split the LCIA metrics of interest into two groups. The first group will contain a subset of proxy LCIA metrics whose values (which will be measured) will be used as the input of a set of multi-linear regression models that will predict the value of the second group of LCIA metrics. The selection of the metrics which will be used as proxy as well as the parameters for the multi-linear regression models will be obtained automatically by formulating and solving a mixed-integer linear programming problem (MILP, see section 1.5.2.1). In addition, we will test the validation of multi-linear regression models using the *k*-fold cross-validation (see section 1.5.3). The main advantages of this methodology are two:

- All significant environmental data will be used, since all the LCIA metrics are either measured or estimated.
- It requires no aprioristic knowledge on the system and, thus, can be easily extended to new processes.

1.4 LCA at a global scale: challenges and proposed solutions

Social pressure towards the adoption of sustainability principles at a wider macroeconomic level has encouraged governments to incorporate environmental concerns in public policies. Consequently, countries must face the challenge of improving their environmental performance while still remain economically competitive. It seems clear that in a globalized international market, the impact should be assessed on a life cycle basis and across nations (*i.e.,* on a consumption-based basis) in order to avoid outsourcing. However, LCA studies at a macroeconomic level require information on mass and energy flows embodied in the life cycle of products being internationally traded. This information is seldom available.

Environmentally-extended input-output (EEIO) models (see section 1.5.1) aggregate LCA data into economic sectors, which simplifies the environmental analysis. These models assess the environmental and economic performance of a system by establishing a link between the total economic output and the associated environmental impact of each economic sector of a region [33] (see sections *3.2. Environmental extension of the IO Model* and *3.3 Multi-regional IO Model* in [4] for further details in the procedure).

Isolated, EEIO models can only be used to assess the effect that different scenarios have on the economic and environmental performance of a region. However, these models can be combined with multi-objective optimization (see section 1.4.1.2) in order to automatically generate optimal alternatives for the current situation. This combined approach has been applied by several authors. Cho [34] combined multi-objective programming and input-output (IO) models to maximize the economic growth and simultaneously minimize the environmental pollution and the energy consumption of a region in Korea. Oliveira and Antunes [35] developed multi-objective optimization model for Portugal using IO models. Hondo et al. [36] applied IO models to technology selection for housing policy toward the long-term reduction of CO_2 emissions in Japan. San Cristóbal [37] proposed an EEIO linear programming problem combining two types of restrictions: environmental restrictions establishing GHG emission targets, and economic restrictions. These works focused on optimizing single economies without considering international trade, thereby neglecting the impact that changes in the economy of a region may have on other overseas economies.

Environmentally extended multi-regional input-output (EEMRIO) models (see section 1.5.1) attribute pollution or resources depletion to the final demand of a product or service following a consistent holistic approach [38] and considering economic transaction between countries, which makes them very useful for policy making. These models have never been combined with multi-objective optimization. In this thesis, the environmental assessment and the optimization at a global scale of an economy is conducted on the basis of these models (see [3, 4]).

1.4.1 Towards an effective environmental legislation

In addition to the challenges stated in the previous section, there is one consideration which should not be overlooked when conducting environmental assessment studies at the macroeconomic level. It is well known that nations displace the manufacturing tasks to countries with softer environmental regulations in order to mitigate their own environmental impact [39–44]. To avoid having inaccurate results, environmental policies should be based on consumption-based emissions rather than on production-based emissions.

The production-based emissions are those associated to activities of facilities operating within the limits of a country regardless of whether their products are consumed locally or externally exported. Therefore, policies based on production penalize the producer rather than the consumer. This weakness allows the trade of emissions between countries, which mask the impact of the developed countries.

On the contrary, consumption-based emissions refer to those emissions caused by all the facilities located anywhere in the world that cover the demand of a region. The definition of policies based on consumption, rather than production, ensures that final consumers are penalized for the emissions associated with the consumed goods, thereby preventing the masking of impact via displacement of production facilities.

Fig. 2 illustrates the differences in the quantification of impacts between the production based and the consumption based perspective. In this example we consider 4 countries. From a production based approach, A and D are slightly polluting countries, B is highly polluting and C is totally clean. On the contrary, from the consumption based approach, A and C become the most polluting countries, while country B changes from the most polluting to a totally clean country.

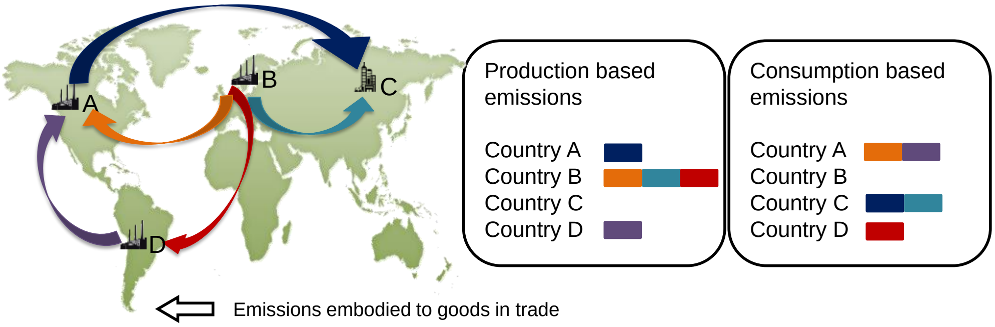

A
B
C
D
Production based emissions
Country A
Country B
Country C
Country D
Consumption based emissions
Country A
Country B
Country C
Country D
Emissions embodied to goods in trade

Fig. 2. Illustrative example of the differences in the quantification of impacts between the production based and the consumption based perspective. The arrows represent the emissions embodied to goods in trade between countries.

1.4.1.1 Identification of similar pollution patterns

The implementation of sustainability principles at a global scale can only be achieved by imposing effective environmental regulations targeting the appropriate drivers of environmental impact. Different tools can assist policy makers during the development of these regulations. One analysis that is valuable is to identify countries showing similar environmental impact patterns, which enables the definition of effective unified environmental policies for similar nations [45]. Such study has yet to be conducted on the context of EEMRIO models. In this thesis, we use the correlation matrix, based on Pearson's correlation coefficient [26, 27], to study the relationships between impact metrics at the macroeconomic scale [3] (see section 2.4 *Multivariate statistical analysis in* [3] for further details of the procedure). For this, we use the World Input-Output Database (WIOD), which is an IO database that covers 35 manufacturing sectors and 41 major countries in the world for the period 1995 to 2009 (see section *2.1 WIOD database* and *2.3 Pressure indicators* in [3] and section *4.1 Data source* in [4] for further details).

1.4.1.2 Combination of multi-objective optimization with environmentally extended input-output models

EEMRIO models can be combined with multi-objective optimization to identify key economic activities that need to be modified in order to minimize the environmental impact at a global macroeconomic scale. This tool could assist policy makers in the development of public policies targeting key sectors to be regulated.

In particular, we combine EEMRIO models with MOO to maximize the demand satisfaction while minimizing the environmental impact at a global scale (see section *3.4. Multi objective optimization problem based on linear programming* in [4] for further details in the procedure).

The main novelty of our approach is that it makes use of an EEMRIO model that identifies optimal environmental strategies in a single region (in this case US) taking into account the impact that those strategies will have globally, thereby leading to a decrease of the global impact rather than the local impact. The goal of the analysis is to maximize the demand satisfaction of the US economy and simultaneously minimize the CO_2 emissions at the global macroeconomic scale, and therefore, identify the sectors to be regulated first.

The approach presented relies on a multi-objective linear programming model (see section 1.5.2.2), since the environmental impact must be minimized while at the same time maximizing the economic output. As a consequence, the solution of the MOO problem is given by a set of Pareto points rather than a single optimal solution. There are several methods available for solving MOO problems. Without loss of generality, the epsilon constraint (EC) method is applied in this thesis (see section 1.5.2.3).

Our analysis identifies economic sectors whose regulation leads to major CO_2 emissions savings at marginal decreases in economic performance. The combination of EEMRIO models with MOO problems proves to be a powerful tool in the development of more effective environmental policies.

In addition, this approach allows us to identify the effect that introducing greener energy sources will have on the economy. Specifically, we analyze the effect of that increasing the share of shale gas in the electricity grid of US will have on its overall environmental performance (see [4]).

1.5 Methodology

This section provides an overview of the main techniques used in this thesis.

1.5.1 Environmentally extended multi-regional input-output models

Input-output (IO) models [46, 47] were conceived as an economic tool to analyze the interdependence of industries/sectors in an economy. They allow predicting how changes in the final demand of services could affect the whole economic system (see section *3.1 Input-Output (IO)* model in [4] for further details in the procedure). Table 1 shows a generic IO table, in which the rows represent the intermediate sales form one sector to the others and the columns the purchases from one sector to the others.

Table 1. Illustrative example of an IO table for the case of 1 region and 3 industrial sectors.

		Sales				
		Sector 1	**Sector 2**	**Sector 3**	**Final demand**	**Total output**
Purchases	**Sector 1**	$x_{(1,1)}$	$x_{(1,2)}$	$x_{(1,3)}$	$y_{(1)}$	$X_{(1)}$
	Sector 2	$x_{(2,1)}$	$x_{(2,2)}$	$x_{(2,3)}$	$y_{(2)}$	$X_{(2)}$
	Sector 3	$x_{(3,1)}$	$x_{(1,2)}$	$x_{(3,3)}$	$y_{(3)}$	$X_{(3)}$

The equations of an IO model can be expressed in compact form as follows:

$$X(i) = \sum_{j} a(i,j)X(j)+y(i) \qquad \forall i \tag{1}$$

$$a(i,j) = \frac{x(i,j)}{X(j)} \qquad \forall i,j \tag{2}$$

where *X(i), X(j)* are the total output in currency units of sector *i* and *j*, y*(i)* is the final demand (end user) of sector *i*, *a(i,j)* are technological coefficients and *x(i,j)* is the output of sector *i* acting like an input for sector *j*.

Environmental aspects can be integrated into IO models giving rise to EEIO models [33]. To this end, additional rows denoting the pollution intensity of each sector (*i.e.,* impact per unit of money traded) are added to the original table, obtaining the following equation:

$$TImp = \sum_i Imp(i) = \sum_i X(i)e(i) \tag{3}$$

where *Imp(i)* is the environmental impact associated with sector i, while *e(i)* is the environmental pollution intensity for sector *i* (*i.e.*, impact per monetary unit traded). Finally, *TImp* is the total environmental impact generated by all sectors of the economy.

When more than one country is considered in the analysis, the IO model becomes a multi-regional IO model. Then, IO equations should be rewritten as follows.

$$X(i,r) = \sum_j \sum_{r\prime} X(j,r')a(i,j,r,r') + y(i,r) \qquad \forall i,r \tag{4}$$

$$a(i,j,r,r') = \frac{x(i,j,r,r')}{X(j,r')} \qquad \forall i,j,r,r' \tag{5}$$

where *X(i,r), X(j,r')* are the total output in currency units (*e.g.* US$) of sector *i* and *j* in region *r/r'*, *a(i,j,r,r')* are technological coefficients, *y(i,r)* is the final demand (end user) of sector *i* of region *r* and *x(i,j,r,r')* is the output of sector *i* of region *r* acting like an input for sector *j* of region *r'*.

Taking this into account, the environmental equations can be rewritten as follows:

$$TImp = \sum_{i}\sum_{r} Imp(i,r) = \sum_{i}\sum_{r} X(i,r)e(i,r) \tag{6}$$

where *e(i,r)* is the environmental pollution intensity for sector i of region r (*i.e.*, impact per monetary unit traded). Finally, *TImp* is the total environmental impact generated by all of the sectors of the economy.

EEMRIO models are typically used for predicting changes in an economy according to changes in the demand of a single or several sectors. However, this methodology can be used in turn for the evaluation of the effect of introducing greening energy sources in an economy (see section *4.4. Impact of Shale Gas* in [4] for further details).

1.5.2 Mathematical programming: optimization

Optimization problems usually consist in maximizing or minimizing an objective function in the presence of constraints, which define the search space or solution space. In mathematical programming, optimization problems are usually expressed as minimizations:

$$\begin{array}{lll} SOO & \min & f(x,y) \\ & s.t. & h(x,y) = 0 \\ & & g(x,y) \leq 0 \\ & & x \in \Re; y \in \mathbb{Z} \end{array}$$

Optimization problems are composed of an objective function (*f(x,y)*), a set of constraints that can be either inequalities (*g(x,y)*) or equalities (*h(x,y)*) and the decision variables that can be either continuous (denoted by *x*) or integer (denoted by *y*). Note that widely-used binary variables are a particular case of the more general integer ones.

The structure of the objective function and constrains determines the type of optimization problem addressed. Linear programming problems (LP) have continuous variables and linear equations. Non-linear programming problems (NLP) have continuous variables and at least one non-linear equation, either in the objective function or in the constraints. Mixed integer linear programming problems (MILP) have continuous and integer variables and linear equations. Mixed integer non-linear programming problems (MINLP) contain continuous and integer variables and one or more non-linear equations. The models presented in this thesis are MILP (see [2]) and LP (expressed as MOO) (see [4]).

1.5.2.1 Mixed integer-linear programming based on multi-linear regression models

The SLCA methodology proposed in this thesis performs an analysis of LCA data of thousands of products in order to construct simplified predictive regression models that will estimate the life cycle impact of a product from key limited data (see section *2 Problem statement* in [2] for further details). The task of building these simplified models can be expressed in mathematical terms as an MILP model that seeks to find the parameters of the predictive model that minimizes the error of the approximation (the difference between the values of the metrics obtained from a detailed LCA analysis and those predicted by the model) subject to some equality and inequality constraints given by the type of regression approach followed. The MILP for multiple data regression shows the general following form (see section *3 Mathematical formulation* in [2] for further details of the procedure).

$$
\begin{aligned}
MILP \quad \min \quad & f(x,y) \\
s.t. \quad & h(x,y) = 0 \\
& g(x,y) \leq 0 \\
& x \in \Re; y \in \{0,1\}
\end{aligned}
$$

where *y* are binary variables that indicate if whether a metric is measured, and therefore included in the regression model as predictor, or not; and *x* are continuous variables that represent the parameters of the regression model. In addition, the model contains the following equations:

- Equality and inequality constraints (*h(x,y)* and *g(x,y)*, respectively) are used to model multi-linear regression equations and logic constraints:
  - ➢ Multi-linear regression equations are based on a given canonical formalism that express the impact values as a multi-linear function of some LCIA metrics.
  - ➢ Logic constraints impose conditions on the number and nature of metrics selected. In the context of our problem, binary variables model the decision of whether an LCIA metric is included in the regression model (and used as predictor) or not. Limits on the total number (and type) of metrics to be used in the regression models are imposed using algebraic constraints containing binary variables.
- The value of the objective function *f(x,y)* is the number to minimize. In this case it corresponds to the approximation error, which is the difference between the values of the metrics obtained from a detailed LCA analysis and those predicted by the model.

1.5.2.2 Multi-objective optimization

The optimization problem exposed in section 1.4.1.2 aims to evaluate alternative policies considering more than one criterion (*i.e.,* environmental and economic). To this end, a multi-objective optimization (MOO) model is developed.

$$\begin{aligned} MOO \quad \min \quad & F = \{f_1, \dots, f_N\} \\ s.t. \quad & h(x) = 0 \\ & g(x) \leq 0 \\ & x \in \Re \end{aligned}$$

The solution to a MOO problem is not a single point, but rather a set of Pareto solutions that represents the optimal trade-off between the conflicting objectives considered in the analysis. A solution is said to be Pareto optimal when it cannot be improved simultaneously in all the objectives without necessarily worsening at least one of them. Therefore, all the Pareto solutions are considered to be equally optimal (see [48] for further information).

Fig. 3 illustrates de concept of Pareto optimality for the optimization of two objectives (*i.e.,* OF1 and OF2). In this case, OF1 is minimized while OF2 is maximized. The grey curve is the Pareto front and the points in this curve (*i.e.,* green dots) are optimal solutions. The region above the curve is infeasible, since no real alternative can improve one objective, either OF1 or OF2, without worsening the other objective. Points below the curve (*i.e.,* orange dots) are sub-optimal, since they can be improved in both criteria by the points lying in the Pareto front. Several methods exist for obtaining Pareto optimal solutions in MOO problems.The epsilon constraint (EC, see section 1.5.2.3) method has been used in this thesis to solve the MOO problem.

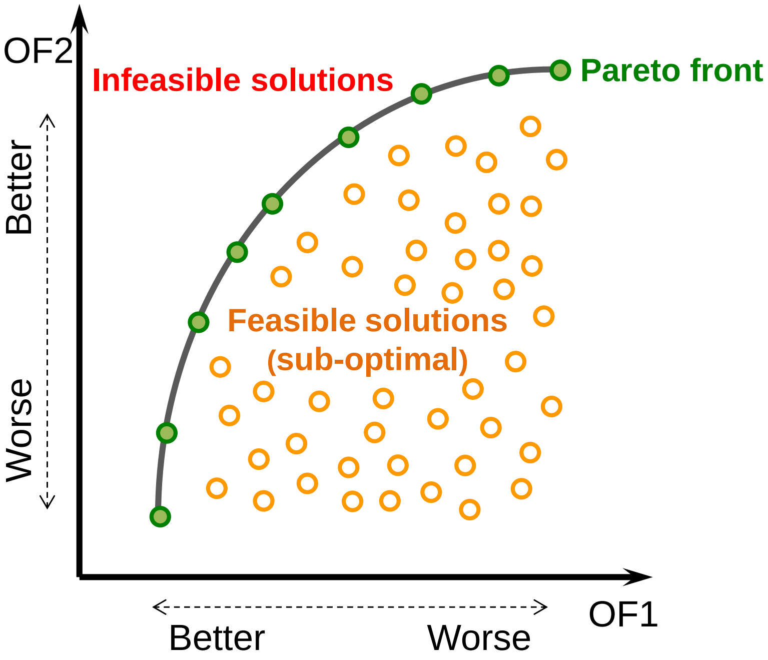


Fig. 3. Example of a bi-criteria Pareto optimal frontier for two conflictive objectives.

1.5.2.3 Epsilon constraint

The epsilon constraint method (EC) is an algorithm widely used to solve MOO problems. This method tackles MOO problems by solving a series of single objective sub-problems where all the objective but one are transferred to auxiliary constraints that impose bounds on them [49].

$$
\begin{array}{llll}
EC \quad \min & f_1 & & \\
s.t. & f_n \leq \varepsilon_n^m & n = 2, \dots, N & m = 1, \dots, M \\
 & h(x) = 0 & & \\
 & g(x) \leq 0 & & \\
 & x \in \Re & &
\end{array}
$$

The epsilon parameters (ε_n^m) are obtained by optimizing every single objective individually, storing the best and worst values of each objective in their optimization and then splitting this interval into a set of subintervals (see section *3.5 Solution method* in [4] for further details of the procedure).

1.5.3 Cross validation

Cross-validation is a model evaluation method for assessing how accurate a predictive model is. Fig. 4 illustrates de cross-validation procedure. In a prediction model, the data are split into two subsets: (1) training set, which are the known data with which the prediction model is build and (2) validation set, which are the unknown data on which the model is tested. This method tests the model obtained from the training set using the validation set and gives insight on the model performance in an independent dataset [50].

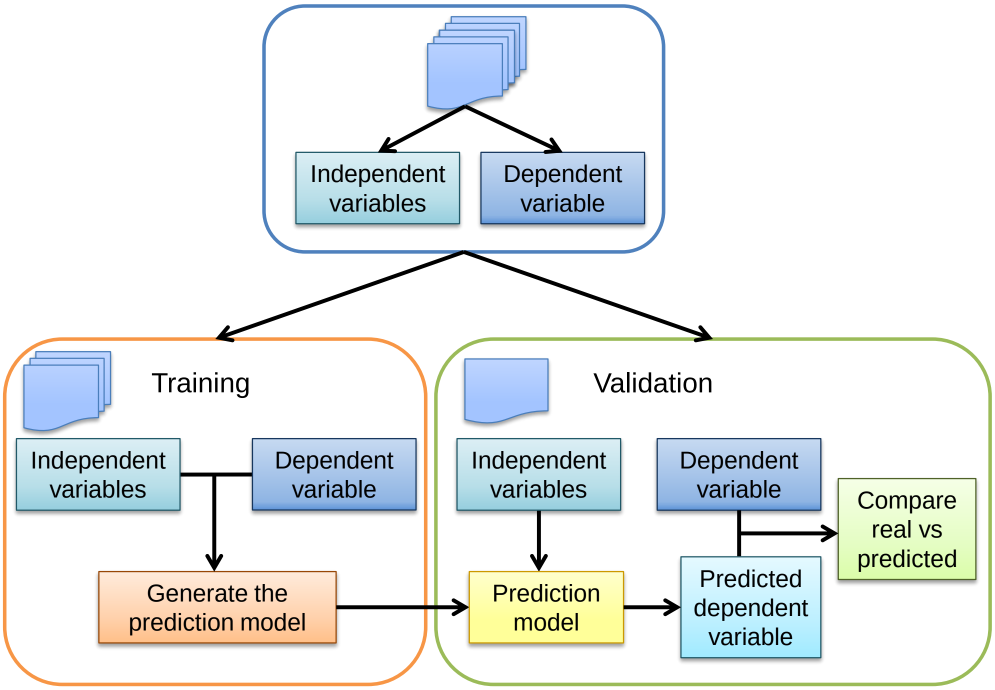


Fig. 4. Cross-validation procedure. In this example, the independent variables are those features used for make the predictions and the dependent variable is the feature to predict.

There are three types of cross validation [51]:

- The holdout method, which is the simplest type of cross-validation, is based on splitting the data into two exclusive subsets. The first one (*i.e.*, training set) is used to build the model and the second one (*i.e.,* validation set) is used to test the model. This kind of cross-validation requires very low computational time, but it highly depends on the choice of training and validation sets.

- *k*-fold cross-validation is a type of validation that improves the holdout method. In this case the data are split into *k* subsets and the holdout method is repeated *k* times. Each of these times, *k*-1 subsets are used as training set and the remaining subset is used as validation set. The advantage of this method over the holdout method is that all observations are used in both training and validation sets and each observation is used exactly once in the validation.

- Leave-one-out cross-validation is the extreme form of *k*-fold cross-validation being *k* equal to *r*, where *r* is the number of observations. This type of validation builds the model *r* different times using all the data but one point, and tests it with that point. This method has the lowest variance in the evaluation, yet the computational time associated is very high.

In this thesis, the *k*-fold cross-validation is used because it guarantees that each data point is used exactly once as validation data and the required computational time is significantly lower than that of the leave-one-out cross-validation.

1.6 General conclusions

This doctoral thesis focuses on the development of systematic methods for the assessment and optimization of life cycle environmental impacts from a sustainable perspective at an engineering and macroeconomic scale. The results obtained from the accomplishment of this work have provided a set of conclusions that are listed below:

- A systematic method based on multivariate statistical analysis for the assessment of the relationships between LCIA metrics has been presented using data from 4087 processes related to human activities (see section *4. Conclusions* [1] in for further details).

- Numerical results show that LCIA metrics are highly correlated and that the level of this correlation is larger in homogeneous data sets. Understanding how impacts are generated help us to develop SLCA methods. (see section *3. Results* in [1] for further details).

- A systematic approach to simplify LCA studies based on multi-linear regression models and mixed-integer linear programming has been devised. This methodology automatically builds multi-linear regression models that predict, with high accuracy, the impact value in different categories from a set of key proxy impact metrics (see section *5. Conclusions* in [2] for further details).

- The streamlined LCA method developed in this thesis proves that few LCIA metrics suffice to describe the environmental performance with accuracy. Our approach could lead to significant saving in time and resources in data collection (see section *4.2. Numerical results* in [2] for further details).

- The combination of EEMRIO models with statistical analysis reveals that environmental indicators are highly correlated and that most of the wealthiest nations display similar environmental impact patterns. These findings can be used to develop simpler environmental policies (see section *3. Results and discussion* in [3] for further details).

- A systematic method that combines multi-objective optimization and EEMRIO models within a single unified framework has been developed for optimizing global economies. This methodology identifies economic changes leading to significant environmental improvements with little impact on the economy (see section *4.3. Multi-objective optimization* in [4] for further details).

- This approach takes into account the life cycle impact of the products, so it leads to solutions yielding true environmental savings at a global scale (see section *4.2. Data analysis* in [4] for further details).

- This approach shows that improving the environmental efficiency of an economic sector (*i.e.* electricity production) by including cleaner energy sources (*i.e.* shale gas) leads to significant environmental savings without modifying the economic structure of a region (see section *4.4 Impact of shale gas* in [4] for further details).

1.7 Future work

We present a set of potential research lines to be addressed in future work on this domain:

- A sustainable approach might include social concerns in the assessment. Therefore, social indicators should be considered in the identification of more sustainable impact patterns.

- The SLCA method proposed in this thesis uses key LCIA metrics to predict others. Future work could consider the use of elementary mass and energy flow data to make those predictions.

- The approximation error was used to evaluate the quality of the SLCA models. Information theory metrics such as the Akaike and Bayesian information criteria could be used instead to minimize simultaneously the error of the approximation and the level of complexity of the model.

- Nonlinear models for data regression could be developed to predict LCIA metrics in order to obtain SLCA methods with higher accuracy.

- In this thesis EEMRIO models were combined with MOO to decrease the CO_2 emissions at a global scale by performing changes in the US economy. This approach could be easily extended to deal with other economic regions and environmental impacts.

1.8 Nomenclature

1.8.1 *Acronyms*

CED	Cumulative energy demand
CML 2001	Impact-oriented characterization
EC	Epsilon constraint
EEIO	Environmentally extended input-output
EEMRIO	Environmentally extended multi-regional input-output
Eco-99	Eco-indicator 99
GHG	Greenhouse gas
IO	Input-output
LCA	Life cycle assessment
LCIA	Life cycle impact assessment
LP	Linear programming
MILP	Mixed-integer linear programming
MINLP	Mixed integer non-linear programming
MOO	Multi-objective optimization
NLP	Non-linear programming
OF1	Objective function 1
OF2	Objective function 2
SLCA	Streamlined life cycle assessment
SOO	Single objective optimization
US	United States
WIOD	World input-output database

1.8.2 *Index*

n	Objective function
m	Epsilon constraint subinterval

1.8.3 Parameters

ε_n^m	Epsilon parameter for subinterval m on objective n
F	Vector of objective functions
k	Number of substets of the cross validation
M	Total number of epsilon subintervals
N	Total number of objectives
r	Number of observations

1.8.4 Variables

f_n	Individual objective function
x	Generic continuous variable
y	Generic integer variable

1.9 References

1. Pascual-González J, Guillén-Gosálbez G, Mateo-Sanz JM, Jiménez-Esteller L: **Statistical analysis of the EcoInvent database to uncover relationships between life cycle impact assessment metrics**. *J Clean Prod* 2015. doi:10.1016/j.jclepro.2015.05.129 .

2. Pascual-González J, Pozo C, Guillén-Gosálbez G, Jiménez-Esteller L: **Combined use of MILP and multi-linear regression to simplify LCA studies**. *Comput Chem Eng* 2015, **82**:34–43.

3. Pascual-González J, Guillén-Gosálbez G, Mateo-Sanz JM, Jiménez-Esteller L: **Statistical analysis of global environmental impact patterns using a world multi-regional input–output database**. *J Clean Prod* 2015, **90**:360–369.

4. Pascual-González J, Guillén-Gosálbez G, Jiménez-Esteller L, Siirola JJ, Grossmann IE: **Multi-objective multi-regional input-output model for minimizing CO2 emissions at a macro-economic scale:Application to the US economy**.(Ready to be submitted to *AIChE Journal).*

5. Hellweg S, Mila i Canals L: **Emerging approaches, challenges and opportunities in life cycle assessment**. *Science (80-)* 2014, **34**:1109–1113.

6. Guinée JB, Heijungs R, Huppes G, Kleijn R, de Koning A, van Oers L, Wegener Sleeswijk A, Suh S, Udo de Haes HA, de Bruijn H, van Duin R, Huijbregts MAJ, Gorrée M: *Life Cycle Assessment: An Operational Guide to the ISO Standards*. 2002.

7. Hermann BG, Kroeze C, Jawjit W: **Assessing environmental performance by combining life cycle assessment, multi-criteria analysis and environmental performance indicators**. *Journal of Cleaner Production* 2007:1787–1796.

8. Finnveden G, Hauschild MZ, Ekvall T, Guinée J, Heijungs R, Hellweg S, Koehler A, Pennington D, Suh S: **Recent developments in Life Cycle Assessment.** *J Environ Manage* 2009, **91**:1–21.

9. Jeswani HK, Azapagic A, Schepelmann P, Ritthoff M: **Options for broadening and deepening the LCA approaches**. *J Clean Prod* 2010, **18**:120–127.

10. Weston N, Clift R, Holmes P, Basson L, White N: **Streamlined Life Cycle Approaches for Use at Oil Refineries and Other Large Industrial Facilities**. *Ind Eng Chem Res* 2011, **50**:1624–1636.

11. Sundaravaradan N, Marwah M, Shah A, Ramakrishnan N: **Data mining approaches for life cycle assessment**. *Proc 2011 IEEE Int Symp Sustain Syst Technol* 2011:1–6.

12. Marwah M, Shah A, Bash C, Patel C, Ramakrishnan N: **Using data mining to help design sustainable products**. *Computer (Long Beach Calif)* 2011, **44**:103–106.

13. SETAC: **Streamlined Life-Cycle Assessment : A Final Report from the SETAC North America Streamlined LCA Workgroup**. 1999(July).

14. Hanes R, Bakshi BR, Goel PK: **The Use of Regression in Streamlined Life Cycle Assessment**. *Proc ISST* 2013, **v1**.

15. Ong SK, Koh TH, Nee AYC: **Development of a semi-quantitative pre-LCA tool**. *J Mater Process Technol* 1999, **89-90**:574–582.

16. Park, Ji-hyung K-KS: **Approximate life cycle assessment of product concepts using multiple regression analysis and artificial neural networks**. *KSME Int J* 2003, **17**:1969–1976.

17. Sousa I, Eisenhard JL, Wallace D: **Approximate Life-Cycle Assessment of Product Concepts Using Learning Systems**. *J Ind Ecol* 2000, **4**:61–81.

18. Arena M, Azzone G, Conte A: **A streamlined LCA framework to support early decision making in vehicle development**. *J Clean Prod* 2013, **41**:105–113.

19. Moriarty P, Honnery D: **The prospects for global green car mobility**. *J Clean Prod* 2008, **16**:1717–1726.

20. Steinmann ZJN, Venkatesh A, Hauck M, Schipper AM, Karuppiah R, Laurenzi IJ, Huijbregts MAJ: **How to address data gaps in life cycle inventories: A case study on estimating CO2 emissions from coal-fired electricity plants on a global scale**. *Environ Sci Technol* 2014, **48**:5282–5289.

21. Jiménez-González C, Ollech C, Pyrz W, Hughes D, Broxterman QB, Bhathela N: **Expanding the Boundaries : Developing a Streamlined Tool for Eco-Footprinting of Pharmaceuticals**. *Org Process Res Dev* 2013, **17**:239–246.

22. Sanjuán N, Stoessel F, Hellweg S: **Closing data gaps for LCA of food products: Estimating the energy demand of food processing**. *Environ Sci Technol* 2014, **48**:1132–1140.

23. Bala A, Raugei M, Benveniste G, Gazulla C, Fullana-I-Palmer P: **Simplified tools for global warming potential evaluation: When "good enough" is best**. *Int J Life Cycle Assess* 2010, **15**:489–498.

24. Hunt RG, Boguski TK, Weitz K, Sharma A: **Case studies examining LCA streamlining techniques**. *The International Journal of Life Cycle Assessment* 1998:36–42.

25. Johnson RA, Wichern DW: *Applied Multivariate Statistical Analysis*. *Volume 47*; 2007.

26. Montgomery DC, Runger GC: *Applied Statistics and Probability for Engineers*. 2003.

27. Walpole R, Myers RH: *Probability and Statistics for Engineers and Scientists*. *Volume 3rd*; 2012.

28. **GaBi 6 Software-System and Databases for Life Cycle Engineering** [http://www.gabi-software.com/]

29. Simapro manual PRe Consultants: **Introduction to LCA with SimaPro 8**. *PRé Consult Netherlands Version* 2013:1–77.

30. Swiss Centre For Life Cycle Inventories: **Ecoinvent Database 3.0**. *Ecoinvent Centre* 2013.

31. Wolf M, Pennington D, Pant R, Chomkhamsri K, Pretato U, Commission E: **European Reference Life Cycle Database (ELCD) European Reference Life Cycle Database (ELCD)**. *Database* 2008:1–30.

32. National Renewable Energy Laboratory: **U.S. Life Cycle Inventory Database**. 2012.

33. Leontief W: **Environmental Repercussions and the Economic Structure: An Input-Output**. *Rev Econ Stat* 1970, **52**:262–271.

34. Cho C-J: **The economic-energy-environmental policy problem: An application of the interactive multiobjective decision method for Chungbuk Province**. *J Environ Manage* 1999, **56**:119–131.

35. Oliveira C, Antunes CH: **A multiple objective model to deal with economy-energy-environment interactions**. In *European Journal of Operational Research*. *Volume 153*; 2004:370–385.

36. Hondo H, Moriizumi Y, Sakao T: **A method for technology selection considering environmental and socio-economic impacts input-output optimization model and its application to housing policy**. *Int J Life Cycle Assess* 2006, **11**:383–393.

37. San Cristóbal JR: **An environmental/input–output linear programming model to reach the targets for greenhouse gas emissions set by the kyoto protocol**. *Econ Syst Res* 2010, **22**:223–236.

38. Wiedmann T: **A review of recent multi-region input-output models used for consumption-based emission and resource accounting**. *Ecological Economics* 2009:211–222.

39. Davis SJ, Caldeira K: **Consumption-based accounting of CO2 emissions.** *Proc Natl Acad Sci U S A* 2010, **107**:5687–5692.

40. Davis SJ, Peters GP, Caldeira K: **The supply chain of CO2 emissions.** *Proc Natl Acad Sci U S A* 2011, **108**:18554–9.

41. Hertwich EG, Peters GP: **Carbon footprint of nations: A global, trade-linked analysis**. *Environ Sci Technol* 2009, **43**:6414–6420.

42. Peters GP: **From production-based to consumption-based national emission inventories**. *Ecol Econ* 2008, **65**:13–23.

43. Wiebe KS, Bruckner M, Giljum S, Lutz C, Polzin C: **Carbon and materials embodied in the international trade of emerging economies: A multiregional input-output assessment of trends between 1995 and 2005**. *J Ind Ecol* 2012, **16**:636–646.

44. Peters GP, Hertwich EG: **CO2 embodied in international trade with implications for global climate policy.** *Environ Sci Technol* 2008, **42**:1401–1407.

45. Baumann H, Cowell SJ: **An Evaluative Framework for Conceptual and Analytical Approaches Used in Environmental**. *Greener Manag Int* 1999:109.

46. Leontief WW: **Quantitative Input and Output Relations in the Economic Systems of the United States**. *Rev Econ Stat* 1936, **18**:105–125.

47. Wiedmann T, Lenzen M, Turner K, Barrett J: **Examining the global environmental impact of regional consumption activities - Part 2: Review of input-output models for the assessment of environmental impacts embodied in trade**. *Ecological Economics* 2007:15–26.

48. Ehrgott M: **Multicriteria Optimization**. In *Multicriteria Optimization*. Edited by Springer. Berlin, Germany; 2005.

49. Bérubé J-F, Gendreau M, Potvin J-Y: **An exact -constraint method for bi-objective combinatorial optimization problems:**

Application to the Traveling Salesman Problem with Profits. *European Journal of Operational Research* 2009:39–50.

50. Kohavi R: **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In *International Joint Conference on Artificial Intelligence*. *Volume 14*; 1995:1137–1143.

51. Browne M: **Cross-Validation Methods.** *J Math Psychol* 2000, **44**:108–132.

2 Paper 1: Statistical analysis of the ecoinvent database to uncover relationships between life cycle impact assessment metrics

Janire Pascual-González[1], Gonzalo Guillén-Gosálbez[1,2], Josep M. Mateo-Sanz[1], Laureano Jiménez-Esteller[1]*

[1]Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Campus Sescelades, Avinguda Països Catalans, 26, 43007 Tarragona, Spain

[2]Centre for Process Integration, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, UK

Abstract

A wide range of impact assessment methodologies are available for quantifying the life cycle environmental impact of anthropogenic activities. The calculation of these metrics requires typically large amounts of data that are hard to collect in practice. To shed light on the extent to which these input data can be reduced (while yet obtaining accurate impact assessment values), this work applies a multivariate statistical analysis to the ecoinvent database. Numerical results show that many life cycle impact assessment (LCIA) metrics are highly correlated, but despite this high level of correlation no single indicator is capable of predicting the others with accuracy via univariate linear regression. Our findings open new avenues for the development of advanced streamlined LCIA methods based on multiple data regression that could exploit this high level of correlation and potentially lead to significant savings in time and resources associated with LCA studies.

Keywords: Multivariate statistical analysis; Environmental metrics relationship; Life Cycle Assessment; Streamlined LCA.

2.1 Introduction

The calculation of life cycle impact assessment (LCIA) metrics requires large amounts of data that are hard to collect in practice. This represents a major obstacle towards the widespread adoption of LCA in industry and academia.

Streamlined LCIA methods (Hunt et al., 1998; Marwah et al., 2011; Sundaravaradan et al., 2011) were proposed originally to simplify the LCA calculations by following 9 different approaches that can be roughly classified into three main groups: (i) those based on removing upstream and/or downstream components from the analysis; (ii) those based on using qualitative or less accurate data, surrogate data or “showstoppers”; and (iii) those based on using specific information to represent impacts or life cycle inventory entries. In essence, the first group restricts the scope of the analysis by making the boundaries for the LCA calculations tighter, which reduces the LCI data to collect (Graedel, 1998; Todd and Curran, 1999). The second uses mainly qualitative information from the process that does not rely on true mass balances (as they do not include quantities at each step). The last group focuses on determining only a reduced set of proxy indicators (which should be easy to quantify) that could ultimately replace a full-scale LCIA analysis (Huijbregts et al., 2006).

Since the concept of "streamlined LCA" appeared, there have been numerous attempts to incorporate it in different industrial sectors, including vehicle development (Arena et al., 2013; Moriarty and Honnery, 2008; Sundaravaradan et al., 2011), oil refineries and industrial facilities (Weston et al., 2011), coal-fired electricity plants (Steinmann et al., 2014), pharmaceuticals (Jiménez-González et al., 2013), food processing (Sanjuán et al., 2014) and plastic bags and recycled materials (Bala et al., 2010).

Most of the SLCA methods developed so far belong to the third group of approaches. Particularly, a topic that has attracted great attention in this field is the definition of a universal proxy indicator that could be used to predict a wide range of impacts in other categories (Hanes et al., 2013; Ong et al., 1999; Park, Ji-hyung, 2003; Sousa et al., 2000). Using a unique proxy impact indicator would simplify LCA studies to a large extent, since only the life cycle entries affecting the calculation of the proxy would be required. In addition, comparisons between alternative products would become easier, as they could be performed on the basis of an analysis of a single category. As will be later discussed in detail, this work investigates this topic in great detail. In a seminar work, Huijbregts, M.A et al (2006, 2010) correlated linearly the cumulative energy demand CED with a set of LCIA metrics, working under the assumption that energy consumption is, in general, what ultimately drives the impact in many categories. Hanes et al. (2013), however, showed that this approach has some limitations.

A literature review on the topic of LCIA methods reveals that research efforts have been devoted primarily towards the definition of novel indicators and/or a unique proxy metric. On the contrary, the study of the relationships between impact metrics, which are still poorly understood, has received much less attention. It is well known in the LCA community that some impacts are highly correlated, mainly because they are ultimately caused by similar (or the same) substances. Despite this observation, to the best of our knowledge, no rigorous study has been carried out on this topic in the open literature.

With the aim to fill this research gap, this work presents a comprehensive statistical study of the extent to which impacts are correlated using LCIA data of thousands of products retrieved from the ecoinvent database. Numerical results show that most LCIA metrics are highlycorrelated (with 60% of thembeing correlatedwith more than 50% of the others), with the level of correlation and the most correlated metrics varying from one product category to another. It was also found that it is not possible to predict the whole range of impacts with accuracy using a single LCIA metric. The high level of correlation between metrics, however, opens new avenues for the development of advanced multiple regression models (not necessarily linear) for simplifying LCIA studies. Our findings enhance our understanding on howimpacts are generated, illustrating clearly the need to develop general guidelines for streamlined LCIA studies leading to significant savings in time and resources.

2.2 Methods

2.2.1 Ecoinvent Database

We use in our calculations environmental data of thousands of products that have been retrieved from the ecoinvent database. Note that the quality and validity of the data is a key issue in any LCA analysis. In fact, different LCA data sources and tools might lead to different results for the same LCA analysis (Herrmann and Moltesen, 2015). We are aware of the fact that there might be discrepancies between databases, but we do think that these might not have a significant impact on the outcome of our statistical analysis, as systematic errors in data collection may cancel out.

The choice of ecoinvent (Frischknecht and Rebitzer, 2005) as source of LCA data for the calculations is motivated by the fact that it is one of the most comprehensive international Life Cycle Inventory (LCI) databases. It provides relevant, reliable, transparent and accessible information of several thousands of LCI datasets in the area of agriculture, energy supply, transport, biofuels and biomaterials, bulk and specialty chemicals, construction materials, packaging materials, basic and precious metals, metals processing, ICT and electronics as well as waste treatment. Ecoinvent covers 4087 processes related with human activities, which are classified by region, economic sector and product type.

The quality and reliability of all the data present in the ecoinvent database are both guaranteed by a peer review process by which data are revised by an internal LCA expert before being fed into the database. This revision, which affects both, calculated and measured data, involves also the assessment of the uncertainty of the data (Swiss Centre For Life Cycle Inventories, 2007).

Our study analyzes the relationships between 32 LCIA metrics considering impact data of 4087 products divided into the same 18 categories covered by ecoinvent (see Table 1). The 32 LCIA metrics are calculated using three different methodologies (CED, CML and Eco-indicator 99) (see Table 2), which are next described in detail. Let us clarify that the product categories used in the calculations are the ones defined in ecoinvent. Hence, no clustering method has been applied to create a new taxonomy of products to perform the statistical analysis. Details on each impact assessment methodology covered in the study are provided next.

Table 1. List of categories in ecoinvent version 2.2.

Chemicals	Agricultural means of production	Nuclear
Waste Management	Electronics	Agricultural production
Metals	Oil	Photovoltaic
Natural gas	Biomass	Construction materials
Electricity	Wood energy	Wooden materials
Transport	Hard coal	All data

Table 2. Life cycle impact assessment methods (and its subcategories) covered in this work.

Methodology	Subcategories	Unit	Code
Cumulative energy demand (CED)	renewable energy resources, biomass	MJ-eq	CED1
	non-renewable energy resources, fossil	MJ-eq	CED2
	non-renewable energy resources, nuclear	MJ-eq	CED3
	non-renewable energy resources, primary forest	MJ-eq	CED4
	renewable energy resources, solar converted	MJ-eq	CED5
	renewable energy resources, potential (in barrage water), converted	MJ-eq	CED6
	renewable energy resources, kinetic (in wind), converted	MJ-eq	CED7

Methodology	Subcategories	Unit	Code
Impact-oriented characterization (CML 2001)	acidification potential	kg SO_2-Eq	CML1
	climate change	kg CO_2-Eq	CML2
	eutrophication potential	kg NO_x-Eq	CML3
	freshwater aquatic ecotoxicity	kg 1,4-DCB-Eq	CML4
	freshwater sediment ecotoxicity	kg 1,4-DCB-Eq	CML5
	human toxicity	kg 1,4-DCB-Eq	CML6
	ionizing radiation	DALYs	CML7
	land use	m^2a	CML8
	malodours air	m^3 air	CML9
	marine aquatic ecotoxicity	kg 1,4-DCB-Eq	CML10
	marine sediment ecotoxicity	kg 1,4-DCB-Eq	CML11
	photochemical oxidation (summer smog)	kg formed ozone	CML12
	resources	kg antimony-Eq	CML13
	stratospheric ozone depletion	kg CFC-11-Eq	CML14
	terrestrial ecotoxicity	kg 1,4-DCB-Eq	CML15
Eco-indicator 99 (Eco-99)	ecosystem quality, acidification & eutrophication	ecopoints	ECO1
	ecosystem quality, ecotoxicity	ecopoints	ECO2
	ecosystem quality, land occupation	ecopoints	ECO3
	human health, carcinogenics	ecopoints	ECO4
	human health, climate change	ecopoints	ECO5
	human health, ionizing radiation	ecopoints	ECO6
	human health, ozone layer depletion	ecopoints	ECO7
	human health, respiratory effects	ecopoints	ECO8
	resources, fossil fuels	ecopoints	ECO9
	resources, mineral extraction	ecopoints	ECO10

2.2.2 Cumulative Energy Demand

The cumulative energy demand (CED) methodology quantifies the total energy use throughout the life cycle of a product/good or a service, including the direct as well as indirect uses of energy during the extraction of raw materials, manufacturing phase and waste disposal (Boustead and Hancock, 1979; Dones et al., 2007; Faist-Emmenegger et al., 2007; Frischknecht et al., 1998; Jungbluth, 2007). Huijbregts, M.A et al (2006) claimed that this metric could be used as proxy indicator to predict a wide variety of impacts via univariate linear regression. As will be shown later in the article, our numerical results show that linear regression based solely on the CED indicator may lead to large approximation errors when predicting some LCIA metrics.

2.2.3 Impact-oriented Characterization

The impact-oriented characterization (CML 2001) methodology is an impact assessment method that restricts the quantitative modelling to the early stages of the cause-effect chain. This approach attempts to avoid the main uncertainty sources associated with the use of damage assessment methods (Guinée et al., 2002).

2.2.4 Eco-indicator

The eco-indicator 99 (EI) methodology is an endpoint, top-down approach (damage oriented) that evaluates the environmental damages in three different categories: human health, ecosystem quality and resources depletion (Goedkoop and Spriensma, 2000; Goedkoop et al., 1998).

2.2.5 Multivariate statistical analysis

A multivariate statistical analysis is conducted to study the level of correlation between the LCIA metrics presented above. All the statistical calculations were performed using the XLSTAT software (version 2013.3.02), a statistical add-in available in Microsoft Excel (Addinsoft, 2013), along with the R package (version 3.0.1), a widely used software for statistical computing (R Core Team, 2013).

A preliminary linear correlation analysis is first performed to quantify the strength of the correlation between any two variables (each one representing a different environmental metric) considering a set of samples (recall that each sample/observation corresponds to a different product). The calculations are performed for every individual subcategory separately, as well as for all of them simultaneously. Outliers are identified and eliminated using the robust correlation matrix, which is calculated with the package MASS of the R software. Further details on this method can be found in Rousseeuw et al. and Venables and Ripley (1999; 2002).

Pearson correlation coefficients (r) are calculated after eliminating the outliers following Eq. 1:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

where x_i and y_i are the observed value of the independent and dependent variable, respectively, $\bar{x}$ and $\bar{y}$ are the mean of the independent and dependent variable, respectively, and n is the total number of samples left after applying the outliers' methodology. In our case, the dependent and independent variables correspond to the LCIA metrics.

These coefficients range from totally correlated (−1 or 1), to randomly distributed (0). The sign of the correlation coefficient (positive or negative), defines the direction of the relationship, while the absolute value indicates the strength of the correlation (Montgomery and Runger, 2003; Walpole and Myers, 2012).

The Pearson correlation coefficient follows a t-Student distribution with n-2 degrees of freedom, where n is the total number of samples left after applying the outliers' methodology. Hence, the significance of the correlation might be assessed by calculating the test statistic t:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{2}$$

The test statistic t is usually converted into t-Student p-values for the analysis. We considered that a given correlation is significant if its p-value is lower than the level of significance alpha (i.e., alpha equal to 0.001) (Montgomery and Runger, 2003; Walpole and Myers, 2012).

Hence, the outcome of the t-Student statistic test is used to calculate a correlation index that quantifies the extent to which a metric correlates with the others. This index is defined as the percentage of LCIA metrics that are statistically correlated with the one being assessed (i.e., p-value lower than level of significance). Hence, the index is mathematically calculated as follows:

$$I_k = \frac{m_k'}{m} \qquad \forall k = 1, \dots, m \tag{3}$$

where I_k is the correlation index for LCIA metric k, m_k' is the number of LCIA metrics correlated with the metric k (considering a level of significance alpha equal to 0.001), and m is the total number of metrics.

In addition to calculating the Pearson coefficient, we also determine the coefficient of variation (CV) for every category and LCIA methodology. This metric, which quantifies the dispersion of the data, is defined as the ratio between the standard deviation (σ) to the mean (μ) (see Eq. 4). It is useful because the standard deviation must always be understood in the context of the mean of the data (Lapin, 1998). Higher CV values (expressed as a percentage) indicate a stronger dispersion of the data.

$$CV = \frac{\sigma}{\mu} \cdot 100 \tag{4}$$

2.2.6 Linear Regression

For every pair of impacts, an ordinary least-squares linear regression model is constructed to predict one impact from the other one. This analysis aims to answer the question of whether a single impact suffice to predict the others, if not globally, at least in a particular set of products. These linear equations are forced to pass through the origin in order to avoid negative impact values. Hence, the equation is as follows:

$$y = a \cdot x \tag{5}$$

where *y* is the predicted LCIA metric and *x* is the measured LCIA metric (which should be calculated from full LCI data). The least-squares calculations provide as output the slope (*a*) of the regression, the correlation coefficient (*r*) and the relative error *RE* (Eq. 6).

$$RE = \frac{100}{n} \cdot \sum_{p=1}^{n} \frac{\left|\hat{y}_p - y_p\right|}{\hat{y}_p} \tag{6}$$

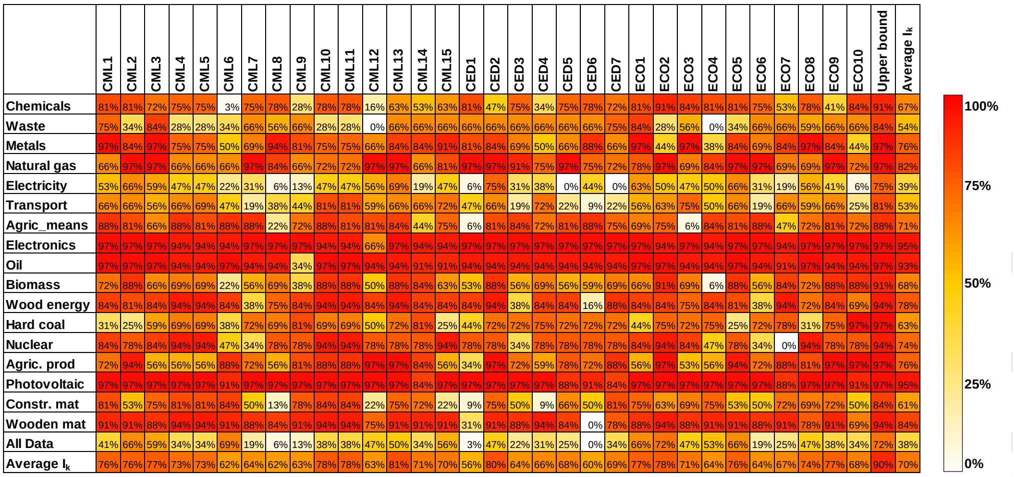
where $\hat{y}_p$ denotes the predicted value (generated via univariate linear regression) of the LCIA metric for product *p* and y_p represents the real value of the metric for the same product (Montgomery and Runger, 2003; Walpole and Myers, 2012).

Regression models of higher quality will yield lower relative errors (a perfect regression model would lead to a zero relative error). The linear regression models were constructed with 80% of the observations (i.e., using 80% of the points as training-set), and performing a cross-validation with the remaining 20% (in order to test the robustness of the results). The error referred to throughout the paper applies to the error in the cross-validation set (do not confuse with the error of the linear regressions in the training set). The split of the data was made following a random procedure and avoiding concentration of points in one single region (in order to avoid extrapolation, the lower and upper bounds of the impact values in the validation set should fall within the lower and upper bounds in the training set).

2.3 Results

Fig. 1 shows a heat map of the correlation index, where red squares denote totally correlated metrics, while white squares represents randomly distributed metrics. Every entry of the matrix shows the correlation index of an impact metric within a product category. To facilitate the analysis, the matrix displays as well the average correlation index of each impact over all the categories and of each category over all the impacts. As observed, the distribution of correlation indexes over the 18 ecoinvent categories is irregular, since many metrics are highly correlated in some categories and poorly in others (e.g. human toxicity (CML6) shows correlations indexes above 97% in oil and below 3% in chemicals). An in-depth analysis of the results reveals that the most correlated methodology is the impact-oriented characterization (CML), whose average correlation index considering all its impact categories is 71% (highest value of 93% in electronics and lowest of 42% in electricity, note that these average values are not shown in the figure). On the other hand, cumulative energy demand shows the lowest correlation index (average correlation index of 66% and values far from 100% in most categories).

The most correlated CED metric is non-renewable energy resources, fossil (average correlation index of 80%). The most correlated CML metric is resources (average correlation index of 81%), while for the eco-indicator 99, it is ecotoxicity (average correlation index of 78%). Note, however, that the fact that a metric will exhibit a high correlation index and strength of the correlation does not guarantee that the error obtained when using the metric to estimate the remaining impacts in all the categories via univariate linear regression will be low.



	CML1	CML2	CML3	CML4	CML5	CML6	CML7	CML8	CML9	CML10	CML11	CML12	CML13	CML14	CML15	CED1	CED2	CED3	CED4	CED5	CED6	CED7	ECO1	ECO2	ECO3	ECO4	ECO5	ECO6	ECO7	ECO8	ECO9	ECO10	Upper bound	Average I_k
Chemicals	81%	81%	72%	75%	75%	3%	75%	78%	28%	78%	78%	16%	63%	53%	63%	81%	47%	75%	34%	75%	78%	72%	81%	91%	84%	81%	81%	75%	53%	78%	41%	84%	91%	67%
Waste	75%	34%	84%	28%	28%	34%	66%	56%	66%	28%	28%	0%	66%	66%	66%	66%	66%	66%	66%	66%	66%	75%	84%	28%	56%	0%	34%	66%	66%	59%	66%	66%	84%	54%
Metals	97%	84%	97%	75%	75%	50%	69%	94%	81%	75%	75%	66%	84%	84%	91%	81%	84%	69%	50%	66%	88%	66%	97%	44%	97%	38%	84%	69%	84%	97%	84%	44%	97%	76%
Natural gas	66%	97%	97%	66%	66%	66%	97%	84%	66%	72%	72%	97%	97%	66%	81%	97%	97%	91%	75%	97%	75%	72%	78%	97%	69%	84%	97%	97%	69%	69%	97%	72%	97%	82%
Electricity	53%	66%	59%	47%	47%	22%	31%	6%	13%	47%	47%	56%	69%	19%	47%	6%	75%	31%	38%	0%	44%	0%	63%	50%	47%	50%	66%	31%	19%	56%	41%	6%	75%	39%
Transport	66%	66%	56%	66%	69%	47%	19%	38%	44%	81%	81%	59%	66%	66%	72%	47%	66%	19%	72%	22%	9%	22%	56%	63%	75%	50%	66%	19%	66%	59%	66%	25%	81%	53%
Agric_means	88%	81%	66%	88%	81%	88%	88%	22%	72%	88%	81%	81%	84%	44%	75%	6%	81%	84%	72%	81%	88%	75%	69%	75%	6%	84%	81%	88%	47%	72%	81%	72%	88%	71%
Electronics	97%	97%	97%	94%	94%	94%	97%	97%	97%	94%	94%	66%	97%	94%	94%	97%	97%	97%	97%	97%	97%	97%	97%	94%	97%	94%	97%	97%	94%	97%	97%	97%	97%	95%
Oil	97%	97%	97%	94%	94%	97%	94%	94%	34%	97%	97%	94%	94%	91%	91%	94%	94%	94%	94%	94%	94%	94%	97%	97%	94%	94%	97%	94%	91%	97%	94%	94%	97%	93%
Biomass	72%	88%	66%	69%	69%	22%	56%	69%	38%	88%	88%	50%	88%	84%	63%	53%	88%	56%	69%	56%	59%	69%	66%	91%	69%	6%	88%	56%	84%	72%	88%	88%	91%	68%
Wood energy	84%	81%	84%	94%	94%	84%	38%	75%	84%	94%	94%	84%	94%	84%	84%	84%	94%	38%	84%	84%	16%	88%	84%	84%	75%	84%	81%	38%	94%	72%	84%	69%	94%	78%
Hard coal	31%	25%	59%	69%	69%	38%	72%	69%	81%	69%	69%	50%	72%	81%	25%	44%	72%	72%	75%	72%	72%	72%	44%	75%	72%	75%	25%	72%	78%	31%	75%	97%	97%	63%
Nuclear	84%	78%	84%	94%	94%	47%	34%	78%	78%	94%	94%	78%	78%	78%	94%	78%	78%	34%	78%	78%	78%	78%	84%	94%	84%	47%	78%	34%	0%	94%	78%	78%	94%	74%
Agric. prod	72%	94%	56%	56%	56%	88%	72%	56%	81%	88%	88%	97%	97%	84%	56%	34%	97%	72%	59%	78%	72%	88%	56%	97%	53%	56%	94%	72%	88%	81%	97%	97%	97%	76%
Photovoltaic	97%	97%	97%	97%	97%	91%	97%	97%	97%	97%	97%	97%	97%	84%	97%	97%	97%	97%	97%	88%	91%	84%	97%	97%	97%	97%	97%	97%	88%	97%	97%	91%	97%	95%
Constr. mat	81%	53%	75%	81%	81%	84%	50%	13%	78%	84%	84%	22%	75%	72%	22%	9%	75%	50%	9%	66%	50%	81%	75%	63%	69%	75%	53%	50%	72%	69%	72%	50%	84%	61%
Wooden mat	91%	91%	88%	94%	94%	91%	88%	84%	91%	94%	94%	75%	91%	91%	91%	31%	91%	88%	94%	84%	0%	78%	88%	94%	88%	91%	91%	88%	91%	78%	91%	69%	94%	84%
All Data	41%	66%	59%	34%	34%	69%	19%	6%	13%	38%	38%	47%	50%	34%	56%	3%	47%	22%	31%	25%	0%	34%	66%	72%	47%	53%	66%	19%	25%	47%	38%	34%	72%	38%
Average I_k	76%	76%	77%	73%	73%	62%	64%	62%	63%	78%	78%	63%	81%	71%	70%	56%	80%	64%	66%	68%	60%	69%	77%	78%	71%	64%	76%	64%	67%	74%	77%	68%	90%	70%

Fig. 1. Heat map of the correlation index of the LCIA metrics over the 18 categories. The column “upper bound” shows the maximum correlation index in each product category, the column “average I_k” displays the average of the correlation index over all the LCIA metrics, while the row “average I_k” shows the average over all the product categories.

The correlation results can be displayed in the form of cumulative probability curves that provide the degree of correlation between LCIA metrics for every product category. These curves display in the *x* axis the correlation index and in the *y* axis the percentage of metrics with a correlation index lower or equal than the one shown in the *x* axis (i.e., percentile of impacts with an index lower or equal to that shown in the *x* axis). Hence, a vertical line in the rightmost side of the figure would represent a category fully correlated (in which all the impacts would be linearly correlated with the rest). A vertical line in the leftmost side of the figure would represent the opposite situation, that is, a category containing no single pair of correlated impacts (see Fig. A1). Following this reasoning, curves closer to the right hand side of the figure are more correlated than those lying on the left hand side.

Fig. 2 shows the cumulative probability curve of each category. Recall that each category contains a given set of products that have been grouped together by the developers of ecoinvent (following some general standards). An upper bound curve is provided as well for comparative purposes (this curve corresponds to the column "upper bound" in the correlation index matrix). This upper bound curve, which is ideal (i.e., does not reflect any product category), is constructed by taking, for each cumulative percentage, the largest correlation index of each impact among all the categories. For clarity, a split version of Fig. 2 is provided in the appendix (see Figs. A2-4).

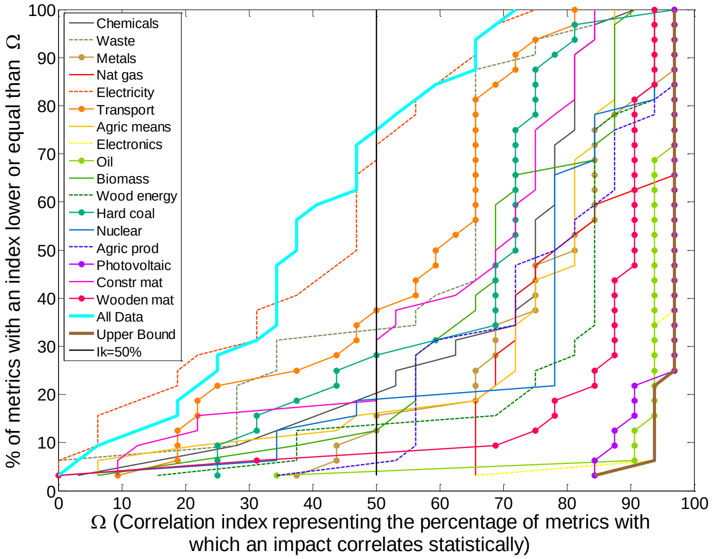


Fig. 2. Cumulative curves of the correlation index. Each line represents one category. The intersecting points between the vertical line and the cumulative curves indicate the percentage of impacts with a correlation index below 50%.

As observed, the LCIA metrics are highly correlated, since in almost all of the categories, more than 60% of the impacts are correlated with more than 50% of the other metrics. The intersection between the vertical line and the cumulative curves represents the percentage of metrics with a correlation index lower or equal than 50% (black line in Fig. 2). As an example, 3% of the impacts in the subcategory agriculture production (dark blue line in Fig. 2), show a correlation index lower or equal to 50%. Hence, 97% of the impacts show correlations indexes above 50%.

As observed in Fig. 2, higher correlation indexes are obtained by analyzing the data by categories instead of all together, since all the categories show a better cumulative curve than the *All Data* curve (blue line) (except for the category electricity).

The most correlated categories are electronics and photovoltaics (32 metrics correlated with more than 50% of the impacts), followed closely by oil (31 metrics correlated with more than 50% of the impacts). By contrast, the category electricity is the less correlated one, yielding a curve very close to the one associated to all the data together (with 28% of the LCIA metrics correlating with more than 50% of the remaining impacts). In general, it is observed that the most correlated categories contain more homogeneous data, as it is the case for industrial technologies available to provide oil, electronics components and photovoltaic energy. The electricity category, on the contrary, is more heterogeneous due to the existence of a wide range of industrial technologies for electricity generation.

Fig. 3 shows in the *x* axis the ecoinvent categories, and in the *y* axis the average correlation index and the coefficient of variation. As observed, categories that are more correlated show lower coefficients of variation (i.e., are more homogeneous). For instance, the category electronics (with an average correlation index of 95%) has a coefficient of variation of 6%. In contrast, the category electricity, which is the less correlated category (average correlation index of 39%), has a coefficient of variation of 55%. Note that in homogeneous samples, it might be easier to identify proxy LCIA metrics yielding good approximation errors (Hanes et al., 2013).

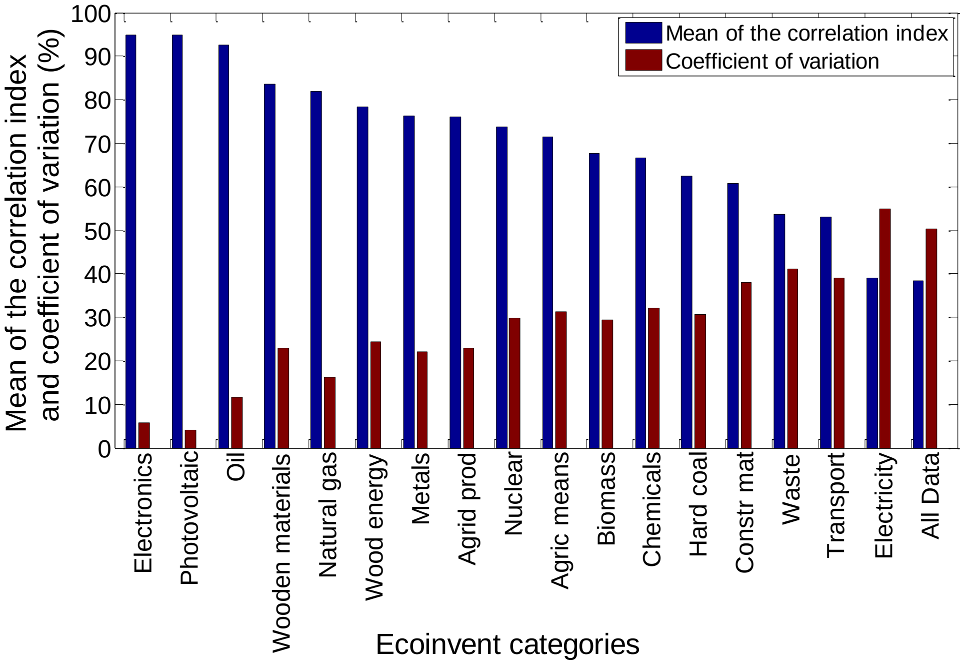
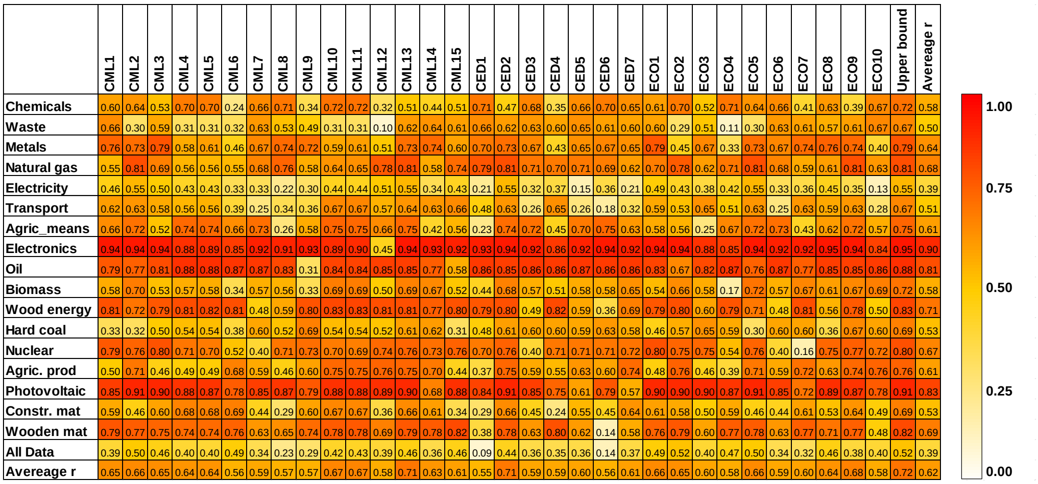


Fig. 3. Correlation index and coefficient of variation for the ecoinvent categories.

Fig. 4 shows a heat map of the Pearson correlation coefficient, where red squares denote totally correlated metrics, while white squares represent randomly distributed metrics. Every entry of the matrix shows the average Pearson coefficient of an impact metric within a product category. Similarly, as with the previous case, we add one column and one row representing the average Pearson in the categories and impacts, respectively. As observed, the strength of the correlation increases when the impacts are analyzed by categories instead of all together. If we analyze the relationship between the correlation index and the Pearson coefficient for each impact and category, results show that the strength of the correlation grows with the correlation index .In other words, the impacts that are more correlated are also the ones showing larger correlation indexes.



	CML1	CML2	CML3	CML4	CML5	CML6	CML7	CML8	CML9	CML10	CML11	CML12	CML13	CML14	CML15	CED1	CED2	CED3	CED4	CED5	CED6	CED7	ECO1	ECO2	ECO3	ECO4	ECO5	ECO6	ECO7	ECO8	ECO9	ECO10	Upper bound	Avereage r
Chemicals	0.60	0.64	0.53	0.70	0.70	0.24	0.66	0.71	0.34	0.72	0.72	0.32	0.51	0.44	0.51	0.71	0.47	0.68	0.35	0.66	0.70	0.65	0.61	0.70	0.52	0.71	0.64	0.66	0.41	0.63	0.39	0.67	0.72	0.58
Waste	0.66	0.30	0.59	0.31	0.31	0.32	0.63	0.53	0.49	0.31	0.31	0.10	0.62	0.64	0.61	0.66	0.62	0.63	0.60	0.65	0.61	0.60	0.60	0.29	0.51	0.11	0.30	0.63	0.61	0.57	0.61	0.67	0.67	0.50
Metals	0.76	0.73	0.79	0.58	0.61	0.46	0.67	0.74	0.72	0.59	0.61	0.51	0.73	0.74	0.60	0.70	0.73	0.67	0.43	0.65	0.67	0.65	0.79	0.45	0.67	0.33	0.73	0.67	0.74	0.76	0.74	0.40	0.79	0.64
Natural gas	0.55	0.81	0.69	0.56	0.56	0.55	0.68	0.76	0.58	0.64	0.65	0.78	0.81	0.58	0.74	0.79	0.81	0.71	0.70	0.71	0.69	0.62	0.70	0.78	0.62	0.71	0.81	0.68	0.59	0.61	0.81	0.63	0.81	0.68
Electricity	0.46	0.55	0.50	0.43	0.43	0.33	0.33	0.22	0.30	0.44	0.44	0.51	0.55	0.34	0.43	0.21	0.55	0.32	0.37	0.15	0.36	0.21	0.49	0.43	0.38	0.42	0.55	0.33	0.36	0.45	0.35	0.13	0.55	0.39
Transport	0.62	0.63	0.58	0.56	0.56	0.39	0.25	0.34	0.36	0.67	0.67	0.57	0.64	0.63	0.66	0.48	0.63	0.26	0.65	0.26	0.18	0.32	0.59	0.53	0.65	0.51	0.63	0.25	0.63	0.59	0.63	0.28	0.67	0.51
Agric_means	0.66	0.72	0.52	0.74	0.74	0.66	0.73	0.26	0.58	0.75	0.75	0.66	0.75	0.42	0.56	0.23	0.74	0.72	0.45	0.70	0.75	0.63	0.58	0.56	0.25	0.67	0.72	0.73	0.43	0.62	0.72	0.57	0.75	0.61
Electronics	0.94	0.94	0.94	0.88	0.89	0.85	0.92	0.91	0.93	0.89	0.90	0.45	0.94	0.93	0.92	0.93	0.94	0.92	0.86	0.92	0.94	0.92	0.94	0.94	0.88	0.85	0.94	0.92	0.92	0.95	0.94	0.84	0.95	0.90
Oil	0.79	0.77	0.81	0.88	0.88	0.87	0.87	0.83	0.31	0.84	0.84	0.85	0.85	0.77	0.58	0.86	0.85	0.86	0.86	0.87	0.86	0.86	0.83	0.67	0.82	0.87	0.76	0.87	0.77	0.85	0.85	0.86	0.88	0.81
Biomass	0.58	0.70	0.53	0.57	0.58	0.34	0.57	0.56	0.33	0.69	0.69	0.50	0.69	0.67	0.52	0.44	0.68	0.57	0.51	0.58	0.58	0.65	0.54	0.66	0.58	0.17	0.72	0.57	0.67	0.61	0.67	0.69	0.72	0.58
Wood energy	0.81	0.72	0.79	0.81	0.82	0.81	0.48	0.59	0.80	0.83	0.83	0.81	0.81	0.77	0.80	0.79	0.80	0.49	0.82	0.59	0.36	0.69	0.79	0.80	0.60	0.79	0.71	0.48	0.81	0.56	0.78	0.50	0.83	0.71
Hard coal	0.33	0.32	0.50	0.54	0.54	0.38	0.60	0.52	0.69	0.54	0.54	0.52	0.61	0.62	0.31	0.48	0.61	0.60	0.60	0.59	0.63	0.58	0.46	0.57	0.65	0.59	0.30	0.60	0.60	0.36	0.67	0.60	0.69	0.53
Nuclear	0.79	0.76	0.80	0.71	0.70	0.52	0.40	0.71	0.73	0.70	0.69	0.74	0.76	0.73	0.76	0.70	0.76	0.40	0.71	0.71	0.71	0.72	0.80	0.75	0.75	0.54	0.76	0.40	0.16	0.75	0.77	0.72	0.80	0.67
Agric. prod	0.50	0.71	0.46	0.49	0.49	0.68	0.59	0.46	0.60	0.75	0.75	0.76	0.75	0.70	0.44	0.37	0.75	0.59	0.55	0.63	0.60	0.74	0.48	0.76	0.46	0.39	0.71	0.59	0.72	0.63	0.74	0.76	0.76	0.61
Photovoltaic	0.85	0.91	0.90	0.88	0.87	0.78	0.85	0.87	0.79	0.88	0.88	0.89	0.90	0.68	0.88	0.84	0.91	0.85	0.75	0.61	0.79	0.57	0.90	0.90	0.90	0.87	0.91	0.85	0.72	0.89	0.87	0.78	0.91	0.83
Constr. mat	0.59	0.46	0.60	0.68	0.68	0.69	0.44	0.29	0.60	0.67	0.67	0.36	0.66	0.61	0.34	0.29	0.66	0.45	0.24	0.55	0.45	0.64	0.61	0.58	0.50	0.59	0.46	0.44	0.61	0.53	0.64	0.49	0.69	0.53
Wooden mat	0.79	0.77	0.75	0.74	0.74	0.76	0.63	0.65	0.74	0.78	0.78	0.69	0.79	0.78	0.82	0.38	0.78	0.63	0.80	0.62	0.14	0.58	0.76	0.79	0.60	0.77	0.78	0.63	0.77	0.71	0.77	0.48	0.82	0.69
All Data	0.39	0.50	0.46	0.40	0.40	0.49	0.34	0.23	0.29	0.42	0.43	0.39	0.46	0.36	0.46	0.09	0.44	0.36	0.35	0.36	0.14	0.37	0.49	0.52	0.40	0.47	0.50	0.34	0.32	0.46	0.38	0.40	0.52	0.39
Avereage r	0.65	0.66	0.65	0.64	0.64	0.56	0.59	0.57	0.57	0.67	0.67	0.58	0.71	0.63	0.61	0.55	0.71	0.59	0.59	0.60	0.56	0.61	0.66	0.65	0.60	0.58	0.66	0.59	0.60	0.64	0.68	0.58	0.72	0.62

Fig. 4. Heat map of the Pearson correlation coefficient of the LCIA metrics over the 18 categories. The column "average *r*" displays the average of the correlation index over all the LCIA metrics, while the row "average *r*" shows the average over all the product categories.

Table 3 displays the most correlated metric within each category and the average Pearson correlation coefficient for each of them. As observed, there is no single proxy metric that prevails over the rest (i.e., no single metric behaving better than the others in all the categories simultaneously), on the contrary, the most correlated metric in each category changes. For instance, the ecotoxicity metric of the eco-indicator 99 is the most correlated metric in four categories (chemicals, biomass, agricultural production and all data). However, in other categories like waste management or metals is one of the less correlated metrics. This might be attributed to the different nature and features of the products under study (i.e., different sources of impact, emissions, etc.).

As observed, LCIA metrics related to climate change are the most correlated in products categories related to energy (i.e., gas natural, electricity and photovoltaic). On the other hand, in products in which the life cycle stage of raw materials extraction (i.e., chemicals, nuclear, metals and hard coal) and/or alter the soil during their normal activity (i.e., agricultural, waste management and biomass), the more correlated metrics are those related to ecosystems quality, like ecotoxicity or eutrophication.

Note that Table 3 shows only the most correlated LCIA metric in each category, yet in some categories other metrics may show very similar correlation indexes (see Fig. 1).

Some previous studies suggested that CED may be used as a predictor for the environmental burden of commodity production (Huijbregts et al., 2006, 2010). Remarkably, this indicator does not appear among the most correlated ones in Table 3. In fact, it is the less correlated indicator (average correlation index among all the categories of 66%). Furthermore, the CED metrics show the largest dispersion coefficients (average coefficient of dispersion of the CED metrics of 42%). This implies that the same metric might be highly correlated in one category and poorly in another one.

Table 3. Most correlated metric and correlation index and average Pearson correlation coefficient of the most correlated metric within each category.

Category	Most correlated LCIA metric	Correlation index of the most correlated metric in the category	Average Pearson correlation coefficient of the most correlated metric in the category
Chemicals	Eco99: Ecotoxicity	91%	0.703
Waste Management	Eco99: Acidification & eutrophication	84%	0.600
Metals	CML:Eutrophication potential	97%	0.794
Natural gas	Eco99:Climate change	97%	0.810
Electricity	CED:Non-renewable energy resources, fossil	97%	0.551
Transport	CML:Marine sediment ecotoxicity	81%	0.672
Agricultural means of production	CML:Marine aquatic ecotoxicity	88%	0.753
Electronics	CML:Resources	97%	0.945
Oil	CML:Human toxicity	97%	0.872
Biomass	Eco99:Ecotoxicity	91%	0.658
Wood energy	CML:Marine sediment ecotoxicity	94%	0.832
Hard coal	Eco99:Mineral extraction	97%	0.600
Nuclear	CML:Terrestrialecotoxicity	94%	0.762
Agricultural production	Eco99:Ecotoxicity	97%	0.760
Photovoltaic	CML:Climate change	97%	0.908
Construction materials	CML:Human toxicity	91%	0.691
Wooden materials	CED:Non-renewable energy resources, primary forest	94%	0.803
All data	Eco99:Ecotoxicity	72%	0.523

We next address the issue of whether it is possible to use a single proxy LCIA indicator for predicting the remaining metrics precisely using linear models. To this end, we construct, for each LCIA metric, a set of linear models (one regression model for correlating the metric under study with each of the remaining metrics) that predict a given impact from the former one. The median relative error of the predictions (see Table 4) is calculated assuming a linear relationship between the two LCIA metrics. We used the median error instead of the mean error to deal with the skewness of the relative error distribution. Recall that the data are split into two sets: the training set, which contains 80% of the points, and the validation set, which contains the remaining 20%.

Results show that there are only two categories (photovoltaic and oil) in which accurate estimations (i.e., with a median error of the cross-validation set of 5% and 23%, respectively) can be obtained using linear regressions of one single metric. In the best predictions that could be made in the remaining categories (using the metric yielding the lowest error) the median relative error of the cross-validation set lies between 33 and 82%. This is a very high value that makes univariate linear regression inappropriate for streamlined LCIA studies in those categories. Note that the metric leading to the minimum approximation error via linear regression might not be the one showing the largest correlation index.

Table 4. Median, minimum relative error and average coefficient of variation for each category.

Category	Median RE (%)	Minimum RE (%)	Metric with the minimum RE (%)	Average coefficient of variation (%)
Chemicals	46	30	CML: marine sediment ecotoxicity	32
Waste Management	85	54	CED: biomass	41
Metals	65	46	ECO: acidification & eutrophication	22
Natural gas	59	42	CML: resources	16
Electricity	52	37	CED: primary forest	55
Transport	59	37	CED: primary forest	39
Agricultural means of production	51	28	CML: marine aquatic ecotoxicity	31
Electronics	50	34	CML: acidification potential	6
Oil	23	17	CED: kinetic (in wind), converted	12
Biomass	67	55	CED: kinetic (in wind), converted	29
Wood energy	33	21	CED: biomass	24
Hard coal	42	29	CML: malodours air	31
Nuclear	33	20	CML: eutrophication potential	30
Agricultural production	46	27	ECO: ozone layer depletion	23
Photovoltaic	5	3	ECO: respiratory effects	4
Construction materials	62	54	CML: resources	38
Wooden materials	68	48	ECO: mineral extraction	23
All data	82	68	CML: acidification potential	50

Table 5. Median and minimum relative error and average coefficient of variation for each LCIA metric.

LCIA metric	Median RE (%)	Minimum RE (%)[a]	Average coefficient of variation (%)
CML1	53	4	25
CML2	51	4	28
CML3	58	3	21
CML4	60	4	29
CML5	59	4	29
CML6	67	9	48
CML7	64	13	42
CML8	62	5	51
CML9	61	4	44
CML10	56	4	27
CML11	54	4	26
CML12	54	5	46
CML13	45	4	18
CML14	54	5	30
CML15	60	5	32
CED1	64	4	62
CED2	46	4	20
CED3	60	12	41
CED4	63	9	38
CED5	61	15	40
CED6	61	6	55
CED7	55	11	37
ECO1	56	4	21
ECO2	53	6	27
ECO3	53	5	33
ECO4	60	5	45
ECO5	51	4	28
ECO6	62	13	42
ECO7	55	4	42
ECO8	52	3	25
ECO9	49	5	26
ECO10	64	8	39

[a]The minimum relative error belongs to the photovoltaic category in all the LCIA metrics.

Analyzing the LCIA metrics instead of the categories (see Table 5), we find that all the metrics display a median relative error over 45% over all the product categories. This relative error is too high for any LCIA metric to be used as universal proxy indicator. These results are consistent with the work by Hanes et al (2013), which found that LCIA metrics are poor predictors when the data analyzed is not homogeneous. Regarding the methodologies, the three of them show similar median relative errors (CML-2001 = 58%, CED=61% and ECO-99= 54%), being cumulative energy demand the worst.

Note that the big differences in the orders of magnitude of the impacts lead to large relative errors in the predictions (over 40%), even when the correlation coefficients are high. For instance, in the electronics category, which shows the largest average Pearson correlation coefficient value, the relative errors for every single impact are quite large (over 34%). As an example, Fig. 5 shows the scatter plot of the prediction of ozone layer depletion as a function of fossil fuel. The regression coefficient is close to one (R^2=0.9708), but the median relative error is above 50%. If we enlarge the linear regression plot (Fig. 6), focussing on the points close to the origin (0,0), we understand why this happens. There, we observe that points close to the origin have the same tendency as the others, yet their relative errors are high because their magnitudes are small in comparison with the points lying on the right hand side of the plot. This can be understood (arguably) as a limitation of the linear regression, which is unable to predict accurately over the entire domain.

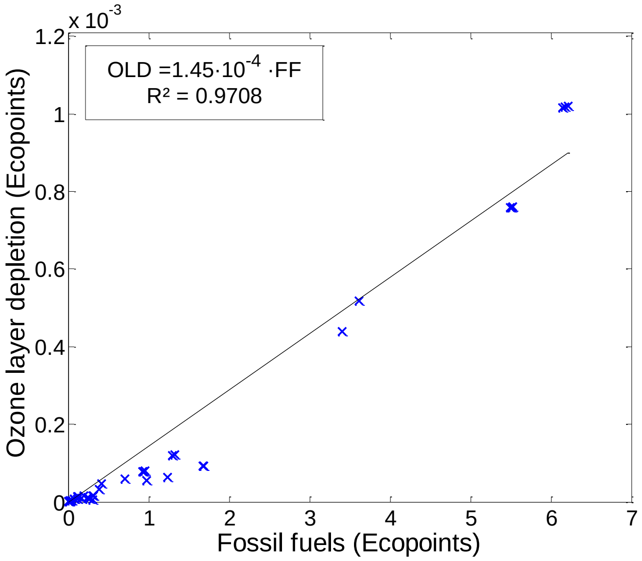


Fig. 5. Linear regression plots, based on 99industrial processes, for impacts fossil fuel and ozone layer depletion.

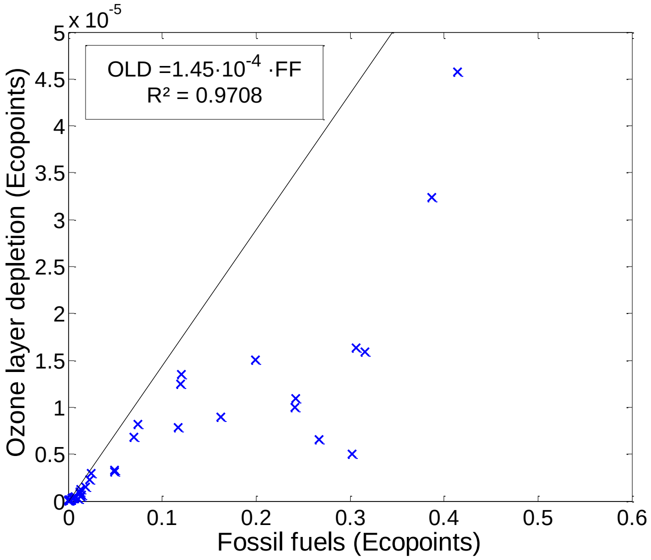


Fig. 6. Linear regression plots of the observations close to the origin, for impacts fossil fuel and ozone layer depletion.

2.4 Conclusions

This paper studied the relationships between LCIA metrics (as applied to the environmental assessment of products) using data retrieved from the ecoinvent database. We analyzed 4087 processes related to human activities that are grouped into 18 categories and whose environmental performance was quantified according to 32 different LCIA metrics.

Our results show that there is a strong correlation between the analyzed LCIA metrics, with more than 60% of them being correlated with more than 50% of the rest within each product category. The intensity of the correlation increases with the correlation index, so in general those metrics that correlate with a higher number of impacts show larger Pearson coefficients in the correlations.

Furthermore, higher correlation indexes and intensities are obtained when the data are analyzed in each isolated subcategory rather than as a whole. The most correlated categories show lower coefficients of variation, indicating that LCIA metrics tend to be more correlated in more homogeneous products datasets. In addition, the most correlated metric differs from one category to another and it happens that some LCIA metrics correlate with a large number of impacts in one category and with very few in others.

The analysis of the errors obtained through the application of univariate linear predictions shows that it is not possible to make accurate predictions of impact using a single LCIA indicator. Our results thus suggest the need to use more sophisticated regression models for making predictions, either based on nonlinear relationships between metrics or on multivariate approaches accounting for more than one LCIA metric in the calculations.

We do not claim that we should analyze only a reduced set of impacts when assessing the environmental performance of a process. Each damage category provides valuable information that covers a wide spectrum of environmental aspects. Our results, however, show that the high level of correlation between impact metrics makes it possible to develop streamlined LCIA methods that will focus on quantifying a reduced number of damage categories and estimating the rest from them. Future work will therefore focus on devising advanced multivariate regression models of this type. A comparison between different databases (i.e., gabi, simapro) would also be a potential area of improvement, including an uncertainty analysis of the data using stochastic modelling. With regard to this last point, it should be mentioned that there is little information available in ecoinvent on the characterization of the uncertain parameters affecting the LCA calculations (i.e., only uncertainties affecting a few products are fully described, typically through lognormal distributions). With this information at hand, it would be possible to apply a sampling method and generate representative samples of the impact values of each product, and then conduct the statistical analysis for all these samples rather than for the nominal values.

2.5 Acknowledgements

The authors wish to acknowledge the financial support received from the University Rovira i Virgili and from the Spanish Ministry of Education and Science (CTQ2012-37039 and DPI2012-37154).

2.6 Nomenclature

Acronyms

CED	Cumulative energy demand
CML 2001	Impact-oriented characterization
EI	Eco-indicator 99
LCI	Life Cycle Inventory
LCIA	Life Cycle Impact Assessment

Index

k	LCIA metric

Parameters

m	Total number of metrics
$m_{k'}$	Number of LCIA metrics correlated with the metric k
n	Number of samples left after applying the outliers' methodology
x	Measured LCIA metric
y_p	Real value of the metric for process p

Variables

a	Slope of the linear regression
CV	Coefficient of variation
I_k	Correlation index
r	Correlation coefficient of the linear regression
RE	Relative error of the prediction
y	Predicted LCIA metric
$\hat{y}_p$	Predicted LCIA metric for process p
μ	Mean of the data set
σ	Standard deviation of the data set

2.7 References

Addinsoft, S., 2013. 2013: XLSTAT software.

Arena, M., Azzone, G., Conte, A., 2013. A streamlined LCA framework to support early decision making in vehicle development. J. Clean. Prod. 41, 105–113.

Bala, A., Raugei, M., Benveniste, G., Gazulla, C., Fullana-I-Palmer, P., 2010. Simplified tools for global warming potential evaluation: When "good enough" is best. Int. J. Life Cycle Assess. 15, 489–498.

Boustead, I., Hancock, G.F., 1979. Handbook of industrial energy analysis. Ellis Horwood Limited.

Dones, R., Bauer, C., Röder, A., 2007. Teil VI Kohle. Ecoinvent Rep. 6-VI, v2.0.

Faist-Emmenegger, M., Heck, T., Jungbluth, N., 2007. Teil V Erdgas. Ecoinvent Rep. 6-V, v2.0.

Frischknecht, R., Heijungs, R., Hofstetter, P., 1998. Einstein's lesons for energy accounting in LCA. Int. J. Life Cycle Assess. 3(5), 266-272.

Frischknecht, R., Rebitzer, G., 2005. The ecoinvent database system: a comprehensive web-based LCA database. J. Clean. Prod. 13, 1337–1343.

Goedkoop, M., Hofstetter, P., Müller-Wenk, R., Spriemsma, R., 1998. The ECO-indicator 98 explained. Int. J. Life Cycle Assess. 3(6), 352-360.

Goedkoop, M., Spriensma, R., 2000. The Eco-indicator 99: a damage oriented method for life cycle impact assessment. Methodology report. Netherlands:Amersfoort.

Graedel, T.E., 1998. Streamlined life-cycle assessment. Prentice Hall Up. Saddle River, NJ.

Guinée, J.B., Heijungs, R., Huppes, G., Kleijn, R., de Koning, A., van Oers, L., Wegener Sleeswijk, A., Suh, S., Udo de Haes, H.A., de Bruijn, H., van Duin, R., Huijbregts, M.A.J., Gorrée, M., 2002. Life Cycle Assessment: An Operational Guide to the ISO Standards, The Netherlands.

Hanes, R., Bakshi, B.R., Goel, P.K., 2013. The Use of Regression in Streamlined Life Cycle Assessment.In: Proc. ISST v1.

Herrmann, I.T., Moltesen, A., 2015. Does it matter which Life Cycle Assessment (LCA) tool you choose? – a comparative assessment of SimaPro and GaBi. J. Clean. Prod. 86, 163–169.

Huijbregts, M.A.J., Hellweg, S., Frischknecht, R., Hendriks, H.W.M., Hungerbühler, K., Hendriks, A.J., 2010. Cumulative energy demand as predictor for the environmental burden of commodity production. Environ. Sci. Technol. 44, 2189–2196.

Huijbregts, M.A.J., Rombouts, L.J.A., Hellweg, S., Frischknecht, R., Hendriks, A.J., Van De Meent, D., Ragas, A.M.J., Reijnders, L., Struijs, J., 2006. Is cumulative fossil energy demand a useful indicator for the environmental performance of products? Environ. Sci. Technol. 40, 641–648.

Hunt, R.G., Boguski, T.K., Weitz, K., Sharma, A., 1998. Case studies examining LCA streamlining techniques. Int. J. Life Cycle Assess. 3(1), 36-42.

Jiménez-González, C., Ollech, C., Pyrz, W., Hughes, D., Broxterman, Q.B., Bhathela, N., 2013. Expanding the Boundaries : Developing a

Streamlined Tool for Eco-Footprinting of Pharmaceuticals. Org. Process Res. Dev. 17, 239–246.

Jungbluth, N., 2007. Teil IV Erdöl. Ecoinvent Rep. 6-IV, v2.0.

Lapin, L.L., 1998. Probability and Statistics for Modern Engineering, 2nd editio. ed. Waveland Pr Inc.

Marwah, M., Shah, A., Bash, C., Patel, C., Ramakrishnan, N., 2011. Using data mining to help design sustainable products. Computer (Long. Beach. Calif). 44, 103–106.

Montgomery, D.C., Runger, G.C., 2003. Applied Statistics and Probability for Engineers, Phoenix USA.

Moriarty, P., Honnery, D., 2008. The prospects for global green car mobility. J. Clean. Prod. 16, 1717–1726.

Ong, S.K., Koh, T.H., Nee, A.Y.C., 1999. Development of a semi-quantitative pre-LCA tool. J. Mater. Process. Technol. 89-90, 574–582.

Park, Ji-hyung, K.-K.S., 2003. Approximate life cycle assessment of product concepts using multiple regression analysis and artificial neural networks. KSME Int. J. 17, 1969–1976.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rousseeuw, P.J., Driessen, K. Van, 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics 41, 212–223.

Sanjuán, N., Stoessel, F., Hellweg, S., 2014. Closing data gaps for LCA of food products: Estimating the energy demand of food processing. Environ. Sci. Technol. 48, 1132–1140.

Sousa, I., Eisenhard, J.L., Wallace, D., 2000. Approximate Life-Cycle Assessment of Product Concepts Using Learning Systems. J. Ind. Ecol. 4, 61–81.

Steinmann, Z.J.N., Venkatesh, A., Hauck, M., Schipper, A.M., Karuppiah, R., Laurenzi, I.J., Huijbregts, M.A.J., 2014. How to address data gaps in life cycle inventories: A case study on estimating CO2 emissions from coal-fired electricity plants on a global scale. Environ. Sci. Technol. 48, 5282–5289.

Sundaravaradan, N., Marwah, M., Shah, A., Ramakrishnan, N., 2011. Data mining approaches for life cycle assessment. Proc. 2011 IEEE Int. Symp. Sustain. Syst. Technol. 1–6.

Swiss Centre For Life Cycle Inventories, 2007. Ecoinvent data V2.0. Ecoinvent Cent.

Todd, J., Curran, M., 1999. Streamlined Life-cycle Assessment: A Final Report From The SEATC North America Streamlined LCA Workgroup. Tech. report, Soc. Environ. Toxicol. Chem.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S Fourth Edition, World. Springer, New York.

Walpole, R., Myers, R.H., 2012. Probability and Statistics for Engineers and Scientists, Power.

Weston, N., Clift, R., Holmes, P., Basson, L., White, N., 2011. Streamlined Life Cycle Approaches for Use at Oil Refineries and Other Large Industrial Facilities. Ind. Eng. Chem. Res. 50, 1624–1636.

2.8 Appendix

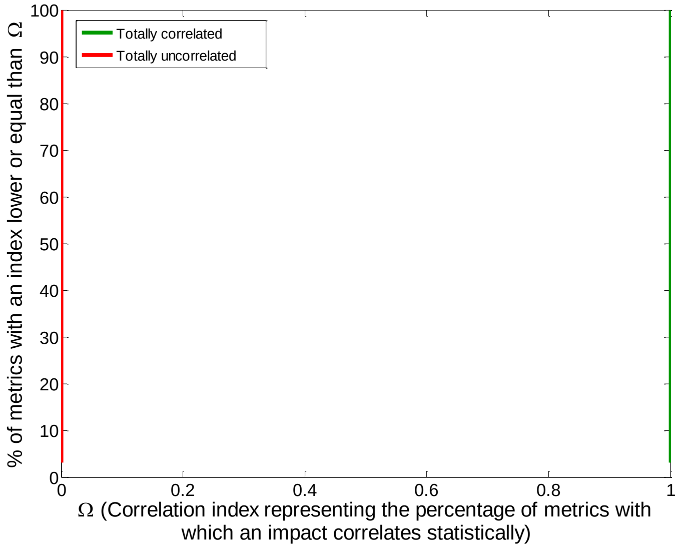


Fig. A.1 Totally correlated (green line) and totally uncorrelated (red line) correlation index's cumulative curves.

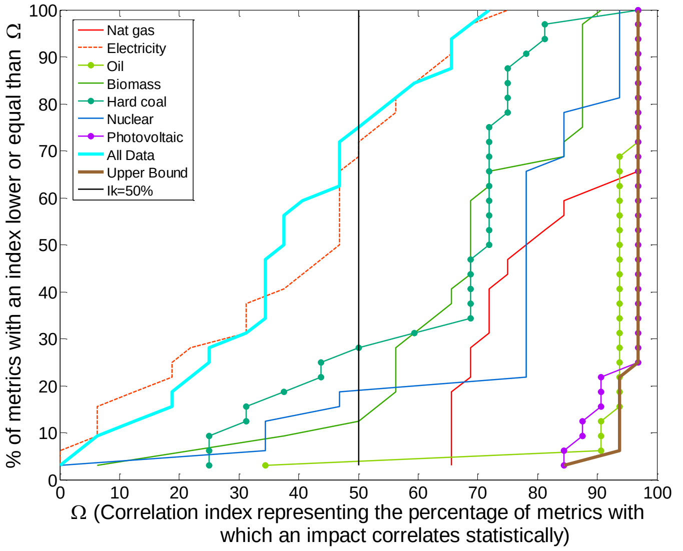


Fig. A.2. Cumulative curves of the correlation index for categories related to energy. Each line represents one category. The intersecting points between the vertical line and the cumulative curves indicate the percentage of impacts with a correlation index below 50%.

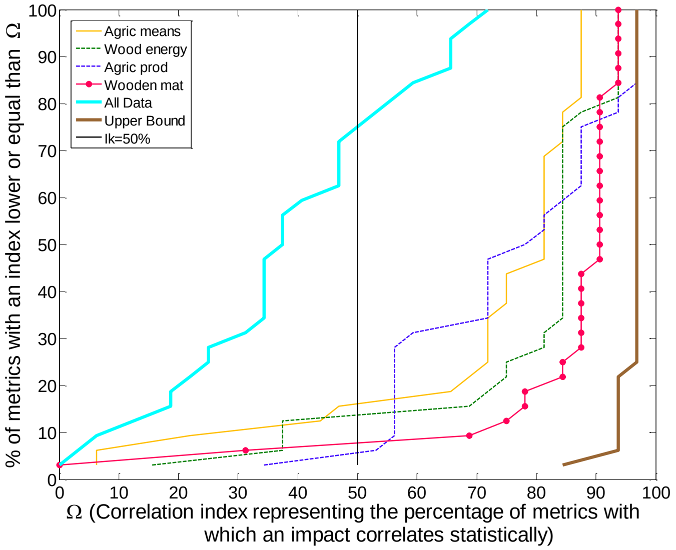


Fig. A.3. Cumulative curves of the correlation index for categories related to wood and agriculture. Each line represents one category. The intersecting points between the vertical line and the cumulative curves indicate the percentage of impacts with a correlation index below 50%.

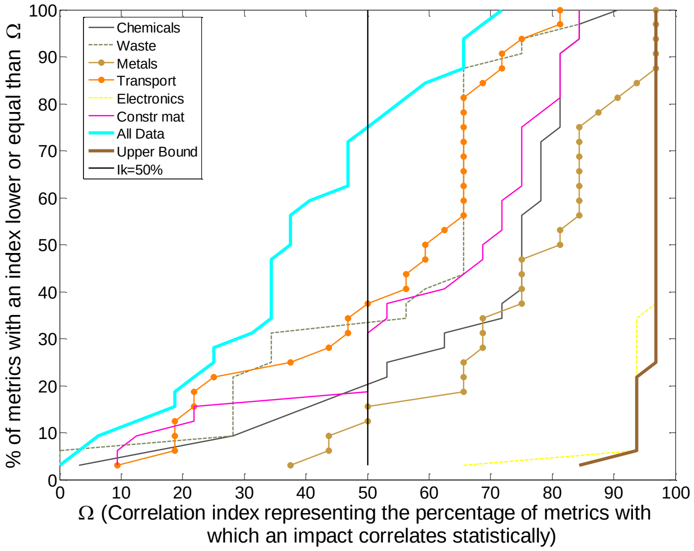


Fig. A.4. Cumulative curves of the correlation index for the other categories. Each line represents one category. The intersecting points between the vertical line and the cumulative curves indicate the percentage of impacts with a correlation index below 50%.

3 Paper 2: Combined use of MILP and multi-linear regression to simplify LCA studies

Janire Pascual-González[1], *Carlos Pozo*[1], *Gonzalo Guillén-Gosálbez*[1,2*], *Laureano Jiménez-Esteller*[1]

[1]Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Campus Sescelades, Avinguda Països Catalans, 26, 43007 Tarragona, Spain

[2]Centre for Process Integration, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, UK

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 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.

Keywords: Multi-linear regression; Streamlined LCA analysis; Environmental impact prediction; Mixed-integer linear programming, Life Cycle Assessment.

3.1 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, Finneveden et al. (2009) and Hellweg & 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, Shah, Bash, Patel, & Ramakrishnan, 2011; Sundaravaradan, Marwah, Shah, & Ramakrishnan, 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.

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, Azzone, & Conte, 2013; Moriarty & Honnery, 2008), oil refineries and industrial facilities (Weston, Clift, Holmes, Basson, & White, 2011), global warming potential evaluation (Bala, Raugei, Benveniste, Gazulla, & Fullana-I-Palmer, 2010), coal-fired electricity plants (Steinmann et al., 2014), pharmaceuticals (Jiménez-González et al., 2013), and food processing (Sanjuán, Stoessel, & Hellweg, 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, Boguski, Weitz, & Sharma, 1998). To avoid this, SLCA studies are constructed from 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.

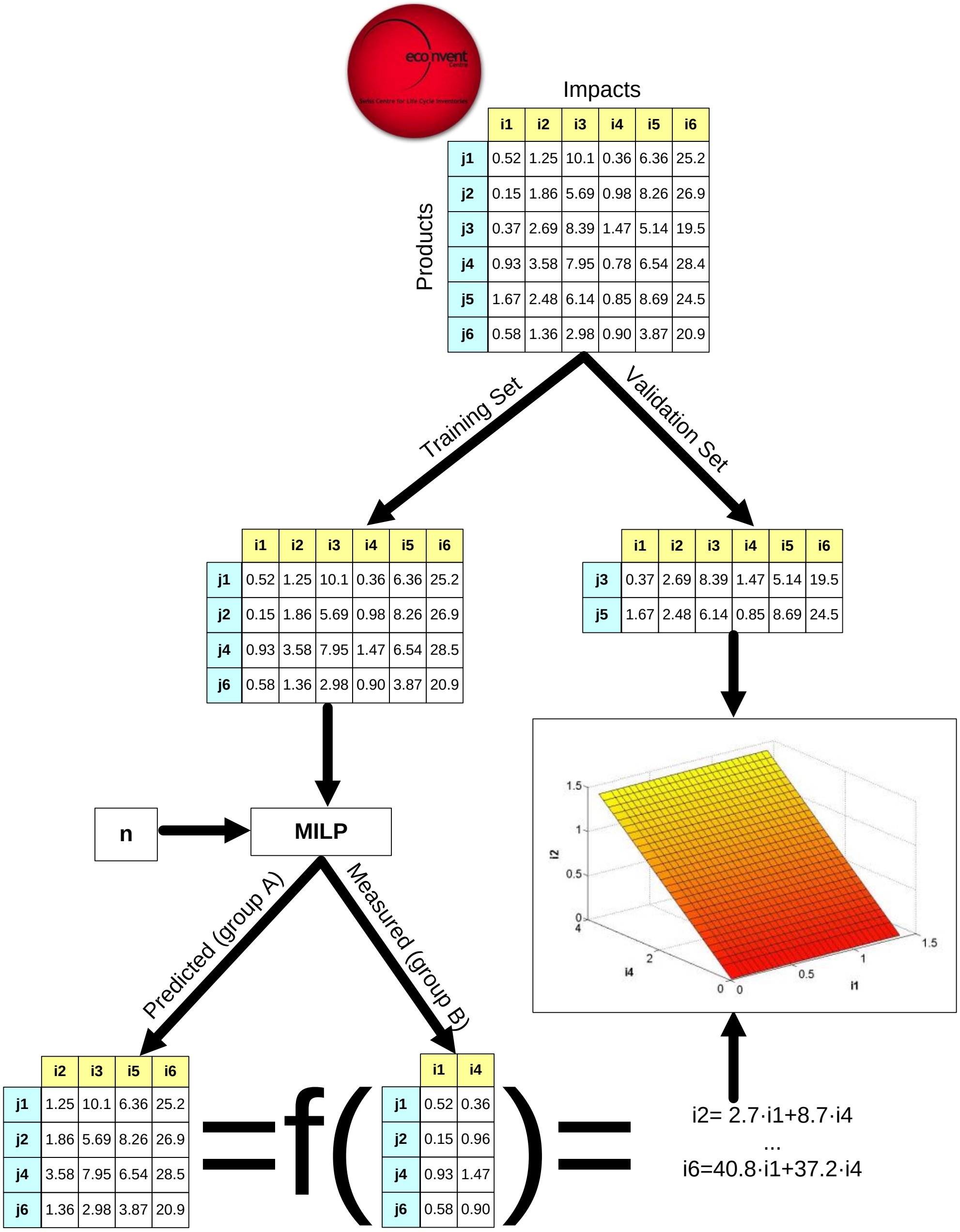
The paper is organized as follow. Section 2 provides the problem statement, while section 3 describes the mathematical formulation. Section 4 describes the ecoinvent database and presents the numerical results. In section 5, the conclusions of the work are drawn.

3.2 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). The selection of metrics to be measured must be optimal, in the sense that no other combination of such number of metrics exists that will yield a better approximation error (when predicting the remaining metrics that are not selected as proxy). The quality of this approximation is quantified by the error, which is given by the difference between the values of the metrics obtained from a detailed LCA analysis and those predicted by the model. Hence, the error is originated from the use of a regression model that estimates some LCIA metrics from a reduced set of key proxy impact indicators. Note that the set of LCIA metrics selected to predict the others is part of the outcome provided by the optimization model. Hence, the proxy indicators are identified by the optimization model rather than fixed according to some knowledge of the system.

3.3 Mathematical formulation

An MILP formulation is developed to tackle the problem defined above. Binary variables are used to denote the inclusion of a specific LCIA metric in the regression model, while continuous ones denote the coefficients of the regression models. The MILP identifies the LCIA metrics that should be measured in order to minimize the overall approximation error (note that this error is given by the remaining LCIA metrics, that is, by the metrics whose values are predicted by the model instead of measured). The MILP model is built using a training set that only contains part of the original data. The regression model provided as output of the MILP is then evaluated using the remaining points in the original data, that is, the so-called validation set. Note that the points in this second split of data fall within the limits of those in the training set. Hence, no extrapolation is made, which could potentially lead to large predictive errors. Fig. 1 summarizes the overall approach. Note that the example shown in the figure is for illustrative purposes only, as the ratio between the number of processes and impacts is too low to ensure reliable results. Some authors suggest that this ratio should be at least equal to 5 (Hair, Black, Babin, & Anderson, 2009).



	i1	i2	i3	i4	i5	i6
j1	0.52	1.25	10.1	0.36	6.36	25.2
j2	0.15	1.86	5.69	0.98	8.26	26.9
j3	0.37	2.69	8.39	1.47	5.14	19.5
j4	0.93	3.58	7.95	0.78	6.54	28.4
j5	1.67	2.48	6.14	0.85	8.69	24.5
j6	0.58	1.36	2.98	0.90	3.87	20.9

	i1	i2	i3	i4	i5	i6
j1	0.52	1.25	10.1	0.36	6.36	25.2
j2	0.15	1.86	5.69	0.98	8.26	26.9
j4	0.93	3.58	7.95	1.47	6.54	28.5
j6	0.58	1.36	2.98	0.90	3.87	20.9

	i1	i2	i3	i4	i5	i6
j3	0.37	2.69	8.39	1.47	5.14	19.5
j5	1.67	2.48	6.14	0.85	8.69	24.5

	i2	i3	i5	i6
j1	1.25	10.1	6.36	25.2
j2	1.86	5.69	8.26	26.9
j4	3.58	7.95	6.54	28.5
j6	1.36	2.98	3.87	20.9

	i1	i4
j1	0.52	0.36
j2	0.15	0.96
j4	0.93	1.47
j6	0.58	0.90

Fig. 1. Outline of the approach. In this example, 6 products *j* and 6 impacts *i* are considered. Part of the data is used as training set and part as validation set. For a given number of metrics to be measured, say *n*, the MILP is run to identify the optimal metrics used to predict the others

through multi-linear regression. The multi-linear models are then assessed using the points in the validation set.

Let us consider LCA data expressed in terms of a matrix with $|I|$ LCIA metrics i and $|J|$ observations j (each one corresponding to a different product). We propose to use this data to build and train a multi-linear regression model that predicts the values of a subset of LCIA metrics, say group A metrics, from a set of proxy indicators, say group B metrics. In practice, we will quantify with the maximum accuracy possible metrics in group B, and use their values to predict metrics in group A. Groups A and B are disjoint, since their intersection is the empty set, that is, either a metric falls in A or B, but not in both at the same time.

Let $A \subset I$ be the set of LCIA metrics whose values are predicted (group A), and $B = I \backslash A$ the set of LCIA metrics calculated from complete LCI data and used as proxy indicators (group B). The estimated value of LCIA metric i' in observation j (recall that an observation corresponds to a particular product) can be predicted from the values of the measured LCIA metrics $i \neq i'$ in that observation.

$$y_{pr}(j,i') = \sum_{i \in B} b(i,i') \cdot y_{ob}(j,i) \qquad\qquad i' \in A, j \in J \qquad (1)$$

where $y_{pr}(j,i')$ is the predicted value of metric i' in observation j (i.e., in product j), $b(i,i')$ is the regression coefficient of the proxy metric i used in the regression model that estimates metric i' and $y_{ob}(j,i)$ is the “true value” of metric i in observation j. By “true value” we mean the value of the impact obtained by performing a detailed LCA analysis on product j. Note that all the impacts are forced to pass through the origin in order to avoid negative estimations. The regression coefficient *b(i,i')* can take any value between a lower ($\underline{b}$) and an upper bound ($\overline{b}$) following Eq.2:

$$\underline{b} \leq b(i,i') \leq \overline{b} \qquad\qquad i \in I, i' \in I, i \neq i' \tag{2}$$

Eq. 1 can only be used when the set B is defined beforehand. In the proposed approach, however, the selection of the metrics used as proxy indicators in the regression models is not made in advance, but it is rather an outcome of the optimization model. To model this logic decision (whether a metric is used as proxy and therefore obtained from a detailed LCA analysis), we define the following disjunction:

$$\begin{bmatrix} Y(i) \\ b(i',i) = 0 \quad i' \in I, i' \neq i \end{bmatrix} \vee \begin{bmatrix} \neg Y(i) \\ b(i,i') = 0 \quad i' \in I, i' \neq i \end{bmatrix} \quad \mathrm{i} \in \mathrm{I} \tag{3}$$

That is, if metric *i* is selected as proxy, the Boolean variable *Y(i)* will be true, and the regression coefficients for predicting this metric from any other metric *i'* will be set to zero. This is because the values of metrics selected as proxy will be measured rather than predicted. On the contrary, if metric *i* is not selected as proxy, the Boolean variable will be false, and the regression coefficients for predicting any other metric *i'* from this metric will be set to zero. That is, metrics not selected as proxy cannot be used to predict the values of other metrics.

The disjunctive term in Eq. 3 can be reformulated into standard algebraic equations using the big-M reformulation (Vecchietti, Lee, & Grossmann, 2003), which yields the following equations after appropriate simplifications.

$$-M \cdot (1 - bin(i)) \leq b(i', i) \leq M \cdot (1 - bin(i)) \quad i \in I, i' \in I, i \neq i' \tag{4}$$

$$-M \cdot bin(i) \leq b(i, i') \leq M \cdot bin(i) \qquad i \in I, i' \in I, i \neq i' \tag{5}$$

In Eqs. 4-5, the binary variable *bin(i)* denotes whether an impact is selected or not. Hence, *bin(i)* will take a value of one if impact i is selected as proxy and used to predict the others (note that this implies that such an impact i will be calculated from a detailed LCA analysis). The binary variable will be zero otherwise (which implies that impact i will not be calculated using LCI data, but rather estimated from the proxy LCIA metrics through multi-linear regression). Hence, the values of *bin(i)* define the two sets of LCIA metrics: those calculated with LCI data and used as proxy indicators to build regression models (group *B*) (for which *bin(i)* is one); and those estimated from the former metrics (group *A*) (for which *bin(i)* is zero). Note that the value of the binary variable is not defined beforehand, but instead obtained after running the optimization model.

Additionally, if the value of the big-M parameter *M* is selected so that $M=|\overline{b}|=|\underline{b}|$, Eq. 2 is no longer required, as it becomes redundant once Eqs. 4-5 are included in the model.

Note that limits on *b(i,i')* allow the coefficient to take a zero value even for a metric used as proxy. This situation should be distinguished from the one in which the regression coefficient $b(i,i')$ is forced to take a zero value because the corresponding metric i is predicted rather than measured. The definition of binary variable *bin(i)* allows us to rewrite Eq. 1 as in Eq. 6, so that it no longer requires the *a priori* definition of the group *B*:

$$y_{pr}(j,i') = \sum_{i \in I, i \neq i'} b(i,i') \cdot y_{ob}(j,i) + y_{ob}(j,i') \cdot bin(i') \qquad i' \in I, j \in J \tag{6}$$

Eq. 6 works as follows: if metric i' is not selected as proxy but instead it is predicted, then *bin(i')* will be zero and will make the second term of the right-hand side of the equation zero as well. Consequently, the predicted value of the LCIA metric $y_{pr}(j,i')$ will be calculated from the corresponding terms of the multi-linear regression model. On the contrary, if metric i' is used as proxy to predict other metrics, Eq. 4 will force the first term in the right-hand side of Eq. 6 to be zero and the predicted value $y_{pr}(j,i')$ will equal the value observed $y_{ob}(j,i')$. Hence, in this latter case the error of predicting a proxy LCIA metric will be zero, because this indicator is indeed calculated from LCI data rather than estimated.

An important clarification should be made at this point. Our approach assumes that the proxy indicators are calculated from LCI data with full accuracy, so the term "error" refers to the error of predicting LCIA metrics (not quantified from LCI data) from proxy LCIA indicators. In practice, the calculation of the proxy LCIA metrics themselves (which are used in the regression models) may be affected in turn by several uncertainty sources. For simplicity, we neglect these uncertainties in our calculations.

A maximum limit *n* on the number of proxy LCIA metrics is imposed, because otherwise the model would define all the metrics as proxy indicators so as to make the error zero:

$$\sum_{i \in I} bin(i) \leq n \tag{7}$$

The model seeks to minimize the average relative error of the multi-linear regression models, which is denoted by *ARE* and determined as follows:

$$ARE = \frac{100}{|I| \cdot |J|} \cdot \sum_{i \in I} \sum_{j \in J} \frac{error(j,i)}{y_{ob}(j,i)} \tag{8}$$

Here, *error(j,i)* is the absolute value of the difference between the values of the metrics that would be obtained from a detailed LCA analysis and those predicted by the model. This value can be obtained from Eqs. 9-10:

$$error(j,i) \geq y_{pr}(j,i) - y_{ob}(j,i) \qquad i \in I, j \in J \tag{9}$$

$$error(j,i) \geq y_{ob}(j,i) - y_{pr}(j,i) \qquad i \in I, j \in J \tag{10}$$

The overall model can be expressed in compact form as follows:

$$\min ARE \tag{11}$$

$$\begin{aligned} s.t. \quad & \text{Eqs. 4-10} \\ & b(i,i') \in \mathbb{R}, y_{pr}(j,i) \in \mathbb{R}^{+} \\ & bin(i) \in \{1,0\} \end{aligned}$$

This MILP can be solved by standard branch and cut methods implemented in powerful software packages, like CPLEX (IBM ILOG, 2012).

3.4 Results and discussion

The methodology proposed is applied to 2 categories (electricity and oil) of the ecoinvent database, which is arguably the most extensive database of life cycle inventory data. These categories include 141 and 90 products, respectively. Data in the electricity category is heterogeneous, since there are a wide range of different industrial technologies available to provide electricity. The oil category, on the contrary, is more homogeneous. Both categories satisfy the rule of thumb of 5:1 regarding the ratio between the number of products and impacts that ensures reliable results. In this section, the fundamentals of ecoivent are first described before presenting in detail the numerical results generated with our approach.

3.4.1 Ecoinvent database

The ecoinvent database v.2.2 (Frischknecht & Rebitzer, 2005) is used in the calculations. This database contains LCA data of 4087 products related with human activities classified by region, economic sector and product type.

Ecoinvent provides relevant, reliable, transparent and accessible information of several thousands of life cycle inventory datasets in the area of agriculture, energy supply, transport, biofuels and biomaterials, bulk and specialty chemicals, construction materials, packaging materials, basic and precious metals, metals processing, ICT (information and communications technology) and electronics as well as waste treatment.

Our calculations consider 17 LCIA metrics associated with technologies belonging to the categories electricity and oil. These 17 LCIA metrics, which are shown in Table 1, were calculated following two methodologies: the cumulative energy demand approach (CED) (Boustead &

Hancock, 1979; Dones, Bauer, & Röder, 2007; Faist-Emmenegger, Heck, & Jungbluth, 2007; Frischknecht, Heijungs, & Hofstetter, 1998; Jungbluth, 2007) and the Eco-indictor 99 methodology (ECO) (M Goedkoop & Spriensma, 2000; Mark Goedkoop, Hofstetter, Müller-Wenk, & Spriemsma, 1998).

Table 1. Life cycle impact assessment methods (and its metrics) covered in this work.

Methodology	LCIA metrics	Unit	Code
Cumulative Energy Demand (CED)	renewable energy resources, biomass	MJ-eq	CED1
	non-renewable energy resources, fossil	MJ-eq	CED2
	non-renewable energy resources, nuclear	MJ-eq	CED3
	non-renewable energy resources, primary forest	MJ-eq	CED4
	renewable energy resources, solar converted	MJ-eq	CED5
	renewable energy resources, potential (in barrage water), converted	MJ-eq	CED6
	renewable energy resources, kinetic (in wind), converted	MJ-eq	CED7
Eco-indicator 99 (ECO)	ecosystem quality, acidification & eutrophication	ecopoints	ECO1
	ecosystem quality, ecotoxicity	ecopoints	ECO2
	ecosystem quality, land occupation	ecopoints	ECO3
	human health, carcinogenics	ecopoints	ECO4
	human health, climate change	ecopoints	ECO5
	human health, ionizing radiation	ecopoints	ECO6
	human health, ozone layer depletion	ecopoints	ECO7
	human health, respiratory effects	ecopoints	ECO8
	resources, fossil fuels	ecopoints	ECO9
	resources, mineral extraction	ecopoints	ECO10

3.4.2 Numerical results

Multi-linear regression models were constructed using the *k*-fold cross validation for a *k* equal to 5, that is, using 80% of the observations as training set and the remaining 20% as validation set (Kohavi, 1995). The algorithm is first run for the training set, providing as output the proxy LCIA indicators and the multi-linear regression models (MTrain models from here on). MTrain models are then assessed using the validation set. The split of the data was made following a random procedure and avoiding concentration of points in one single region. The split is also done in a manner such that the points in the validation set are guaranteed to fall within the limits of the training set. The aim here is to avoid extrapolating impact values, which would eventually lead to larger approximation errors.

The model is first run for an increasing number of allowable proxy indicators. The MILP was implemented in GAMS 23.7 and solved with CPLEX 12.3.0.0 on an Intel Core i5-3470 3.20GHz computer. The model features 4149 continuous variables, 6853 constraints and 17 binary variables, and it takes around 200 CPU seconds to find the optimal solution with an optimality gap of 0%.

Figs. 2 and 3 show the errors in the training and validation sets as a function of the number of proxy LCIA metrics used in the regression models.

Note that MTrain models provide the minimum error in the training set, but not necessarily in the validation set. This is because these models are built with data of the training set only. To get deeper insight into the results, the MILP is run for the validation set data in order to construct another set of regression models called MVal. The performance of these models (in the validation set) corresponds to the best performance that could be attained by any model in the validation set. Hence, this analysis sheds light into how well the regression model constructed with the training set data (MTrain) performs in the validation set data.

As expected, the error decreases in both the training and the validation set with the number of proxy impacts included in the regression, first sharply and then marginally after a given number of impacts. When the model includes just one LCIA metric in the regression, the error is too high to guarantee accurate predictions in both categories. This result is consistent with the work of Hanes et al. (2013), which found that the use of a single proxy indicator as predictor results in poor estimates. The error, however, decreases significantly with the number of proxy LCIA metrics. For instance, 5 LCIA metrics suffice to predict the remaining 11 metrics with errors below 20% in both the training and the validation set for the electricity category (Figs. 2A and 2B). This number is even lower in the oil category, in which errors below 15% are obtained in both the training and the validation set using 3 proxy metrics (Figs. 3A and 3B).

Comparing the results by categories, we found that the MILP works better with the homogeneous set (oil category) and worse with the heterogeneous set (electricity category). The error in the homogeneous set is always below 16% in the validation set regardless of the number of proxy metrics, while the heterogeneous set needs 6 LCIA metrics to show similar error values. These results are consistent with the work of Hanes et al. (2013), which found that regression models have better predictive capability in homogeneous sets than in heterogeneous sets. Overall, the proposed methodology is able to make accurate predictions in both cases for a sufficient number of proxy indicators.

Figs. 2B and 3B compare the performance, in the validation set, of the best models obtained using points in the training set (MTrain models) and in the validation set (Mval models). Recall that to obtain the latter, the algorithm was run for the points in the validation set (assuming that these points could be used for building the multi-linear regression models and not only for validation purposes). Note that in both categories, the errors in the validation set of the MTrain models (the models constructed using 80% of the data and tested with the remaining 20%) is slightly higher than the best possible errors that could be attained (those associated with MVal models, which were constructed and tested with the points of the validation set). Differences between both approaches of around 9% are obtained, which demonstrates that our methodology provides predictions rather close to the best possible prediction that could be made with the data in the validation set (even in heterogeneous datasets).

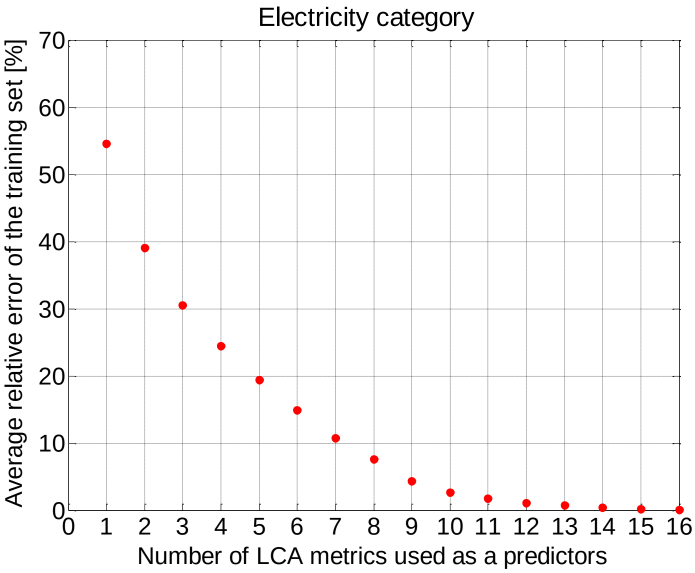


Fig. 2A. Average relative error for the electricity category in the training set. The red dots represent the results of models MTrain (when the models are constructed using 80% of the data and then tested using the remaining 20%).

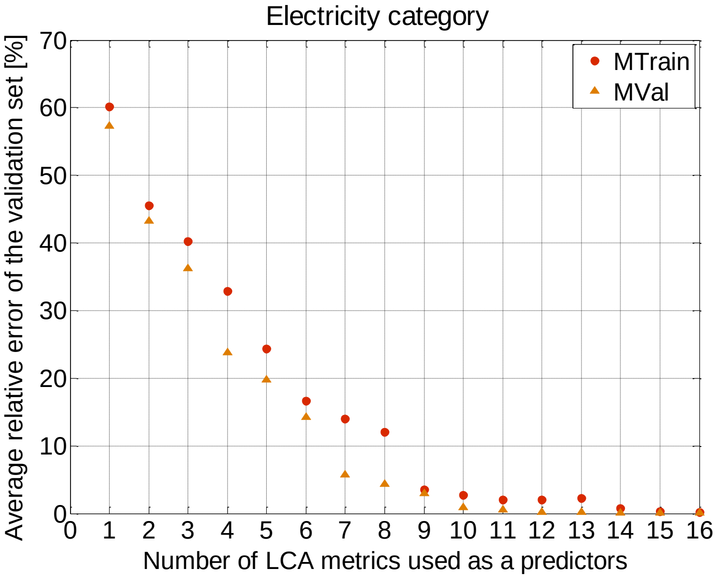


Fig. 2B. Average relative error for the electricity category in the validation set. The red dots represent the results of models MTrain (when the models are constructed using 80% of the data and then tested using the remaining 20%), while the orange triangles represent the minimum errors for the validation set obtained with models MVal (when the models are both built and tested with the points belonging to the validation set).

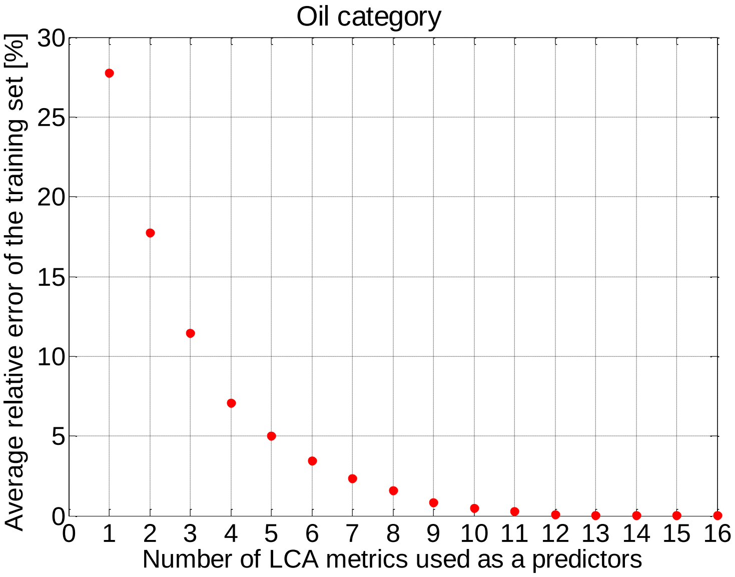


Fig. 3A. Average relative error for the oil category in the training set. The red dots represent the results of models MTrain (when the models are constructed using 80% of the data and then tested using the remaining 20%).

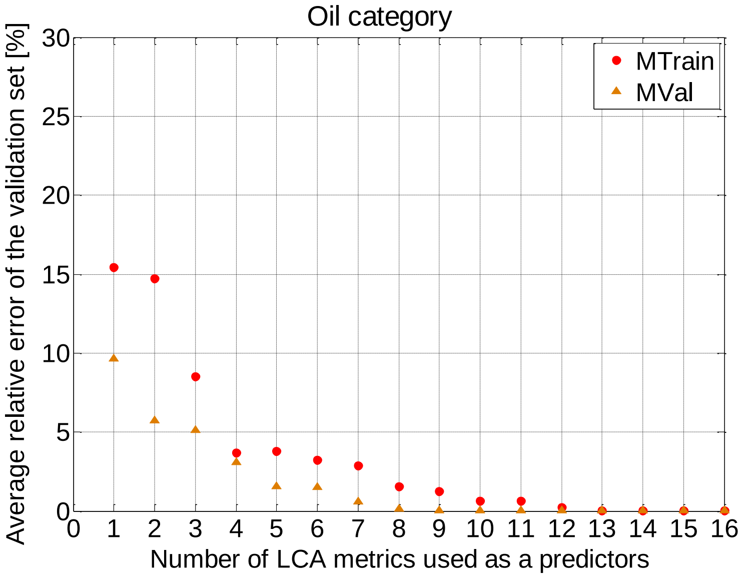


Fig. 3B. Average relative error for the oil category in the validation set. The red dots represent the results of models MTrain (when the models are constructed using 80% of the data and then tested using the remaining 20%), while the orange triangles represent the minimum errors for the validation set obtained with models MVal (when the models are both built and tested with the points belonging to the validation set).

Each of the red dot points in Figs. 2 and 3 corresponds to a given regression model involving a specific set of proxy indicators. Note however, that not all the metrics may be equally easy to quantify in practice. For instance, the impact on climate change tends to be readily available because it reflects a major social concern. Hence, it may be desirable to include it as proxy in the regression model. Thus, an interesting analysis is to study whether the MILP shows a similar performance even when it is forced to include specific metrics. To this end, the MILP is run again forcing it to select the following specific metrics as proxy: (1) a metric widely used (climate change, ECO5); and (2) a metric seldom selected as proxy (ionizing radiation, ECO6).

Figs. 4 and 5 show the errors obtained in the following three cases: (1) MTrain free, that is, the regression models generated when the MILP can select freely the proxy metrics; (2) MTrain climate change fixed, that is, the regression models obtained when the MILP is forced to include the metric climate change as proxy indicator; and (3) MTrain ionizing radiation fixed, that is, the regression models resulting from forcing the MILP to include ionizing radiation as proxy indicator.

The results in the electricity category (Fig. 4) show that MTrain with climate change fixed models lead to very similar results as those in the base case (models MTrain free). On the other hand, when a unique proxy indicator is allowed (i.e., $n = 1$), model MTrain with ionizing radiation fixed shows a larger error than the base case (model MTrain free) in both the training set (where the error increases from 54.6% to 72.1%) and the validation set (where the error increases from 60.2% to 86.0%).

Note that in both, the training and the validation sets, the errors diminish significantly as we increase the number of proxy indicators, even reaching values close to the MTrain free case. Note also that in some cases the error in the validation set of the regression models with fixed impacts is slightly lower than that of the MTrain free case (i.e., model MTrain with all the binary variables free). This is because the optimal MILP solution in the training set might not be optimal in the validation set.

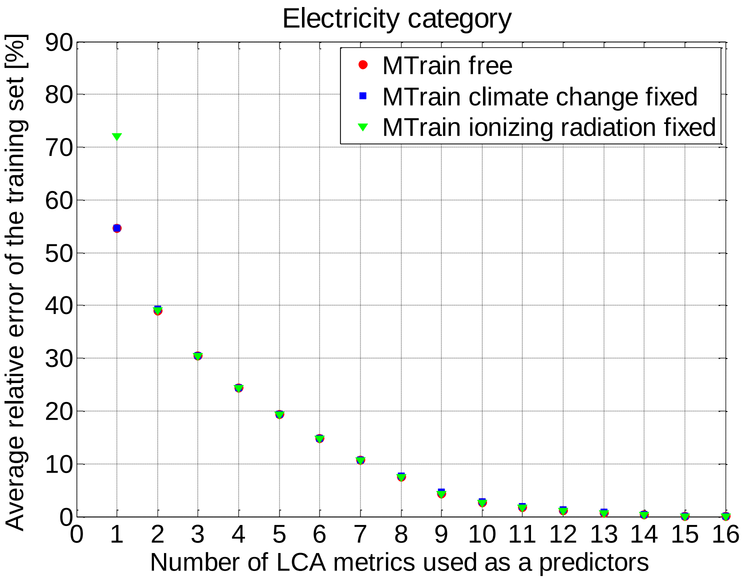


Fig. 4A. Comparison of the average relative error curve of the electricity category for the training set. The red dots represent the MTrain free case, the blue squares the MTrain climate change fixed case and the green triangles the MTrain ionizing radiation fixed case.

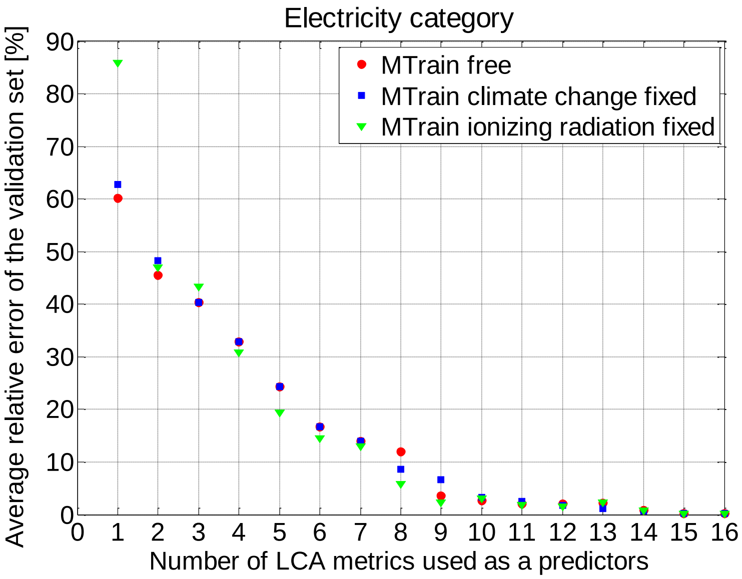


Fig. 4B. Comparison of the average relative error curve of the electricity category for the validation set. The red dots represent the MTrain free case, the blue squares the MTrain climate change fixed case and the green triangles the MTrain ionizing radiation fixed case.

Repeating the same calculations for the oil category (Fig. 5), it is found that the relative error increases considerably when a specific metric (i.e., climate change and ionizing radiation) is forced to be selected as the unique proxy indicator (i.e., *n* = 1). In particular, the error in the training set increases from 27.7% to 39.8% when we use the MTrain climate change fixed model instead of the MTrain free model, and to 36.4% when we employ the MTrain ionizing radiation fixed model. Regarding the error in the validations set, it increases from 15.4% to 77.3% when using the MTrain climate change fixed model instead of the MTrain free model and to 18.3% when resorting to the MTrain ionizing radiation fixed model. However, the relative error decreases significantly as we include more proxy indicators, leading to results really close to those produced by the MTrain free case.

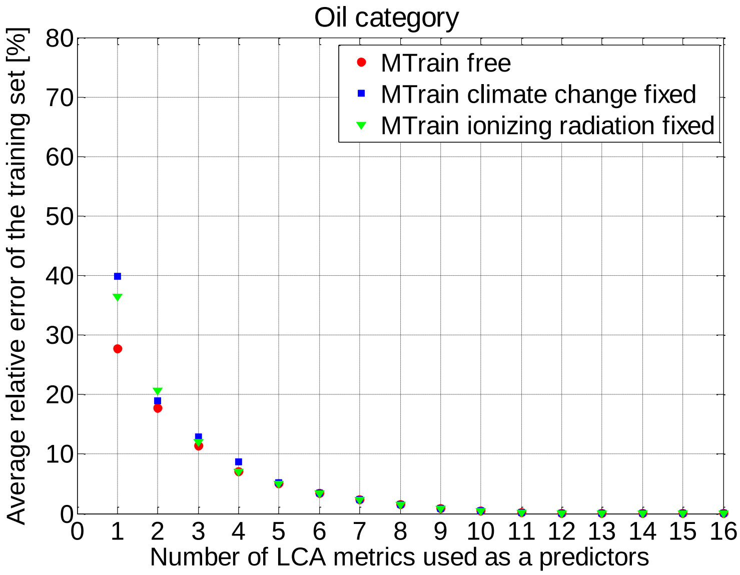


Fig. 5A. Comparison of the average relative error curve of the oil category for the training set. The red dots represent the MTrain free case, the blue squares the MTrain climate change fixed case and the green triangles the MTrain ionizing radiation fixed case.

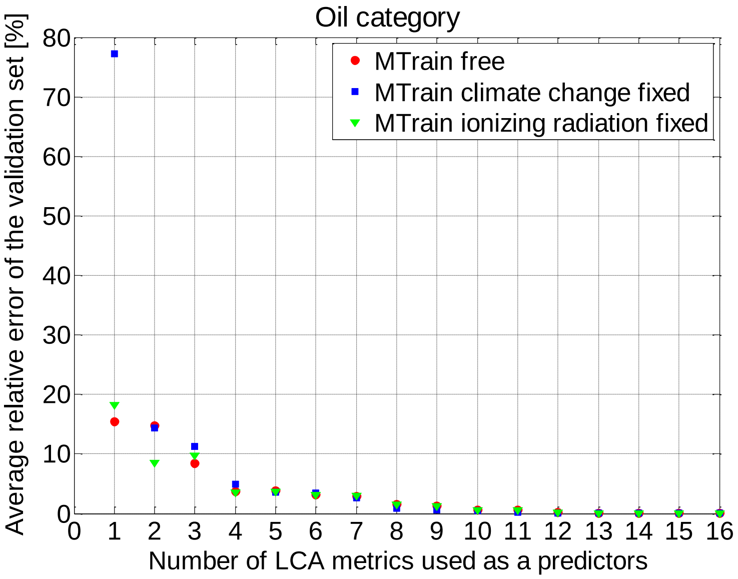


Fig. 5B. Comparison of the average relative error curve of the oil category for the validation set. The red dots represent the MTrain free case, the blue squares the MTrain climate change fixed case and the green triangles the MTrain ionizing radiation fixed case.

Figs. 6 and 7 show the LCIA metrics selected in each run of the algorithm for the MTrain free case, the MTrain model climate change fixed and the MTrain model ionizing radiation fixed. Blue squares denote that a given metric is selected as proxy, while white squares represent the opposite (that the metric is estimated from the proxy indicators). Columns represent the number of proxy LCIA metrics identified by the MILP, while rows denote the LCIA metrics selected (according to the notation given in Table 1).

As observed, the algorithm changes some proxy metrics as we increase their number. In other words, the best combination of *n* proxy metrics does not necessarily belong to the best combination of *n+1* proxy indicators and so on. In fact, as the number of proxy metrics increases, the MILP tends to replace two or more LCIA metrics selected in previous iterations by new ones.

Furthermore, comparing the MTrain free case with the case in which climate change and ionizing radiation are fixed, it is observed that the combination of metrics is kept as constant as possible (in most cases, the algorithm simply replaces one LCIA metric by the one forced to be selected).

	MTD free															
	Number of metrics selected															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CED1						■	■	■	■	■	■	■	■	■	■	■
CED2		■						■	■	■	■	■	■			
CED3		■	■	■	■	■	■								■	■
CED4											■	■	■	■	■	■
CED5							■	■	■	■	■	■	■	■	■	■
CED6					■	■	■	■	■	■	■	■	■	■	■	■
CED7			■	■	■	■	■		■	■	■	■	■	■	■	■
ECO1	■									■	■	■	■	■	■	■
ECO2												■	■	■	■	■
ECO3														■	■	■
ECO4								■	■	■	■	■	■	■	■	■
ECO5			■	■	■	■	■								■	■
ECO6								■	■	■	■	■	■	■		■
ECO7								■	■	■				■	■	■
ECO8													■	■	■	■
ECO9				■	■	■	■				■	■	■	■	■	■
ECO10								■	■	■	■	■	■	■	■	■

	MTD climate change fixed															
	Number of metrics selected															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CED1						■	■	■	■	■	■	■	■	■	■	■
CED2																
CED3		■	■	■	■	■	■	■					■		■	■
CED4									■	■	■	■	■	■	■	■
CED5							■	■	■	■	■	■	■	■	■	■
CED6					■	■	■	■	■	■	■	■	■	■	■	■
CED7			■	■	■	■	■	■	■	■	■	■	■	■	■	■
ECO1										■	■	■	■	■	■	■
ECO2												■	■	■	■	■
ECO3														■	■	■
ECO4									■	■	■	■	■	■	■	■
ECO5	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
ECO6									■	■	■	■		■		■
ECO7													■	■	■	■
ECO8															■	■
ECO9				■	■	■	■	■			■	■	■	■	■	■
ECO10								■	■	■	■	■	■	■	■	■

	MTD ionizing radiation fixed															
	Number of metrics selected															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CED1						■	■	■	■	■	■	■	■	■	■	■
CED2		■						■	■	■	■	■	■			
CED3																■
CED4											■	■	■	■	■	■
CED5							■	■	■	■	■	■	■	■	■	■
CED6					■	■	■	■	■	■	■	■	■	■	■	■
CED7			■	■	■	■	■		■	■	■	■	■	■	■	■
ECO1										■	■	■	■	■	■	■
ECO2												■	■	■	■	■
ECO3														■	■	■
ECO4								■	■	■	■	■	■	■	■	■
ECO5			■	■	■	■	■								■	■
ECO6	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
ECO7								■	■	■				■	■	■
ECO8													■	■	■	■
ECO9				■	■	■	■				■	■	■	■	■	■
ECO10								■	■	■	■	■	■	■	■	■

Fig. 6. Comparison of the LCIA metrics selected in the electricity category for every number of proxy indicators. The first matrix represents the MTrain free case (all the variables are free), the second and third matrixes correspond to the MTrain climate change fixed case and to the MTrain ionizing radiation fixed case, in which the MILP is forced to select climate change and ionizing radiation, respectively, as proxy indicators.

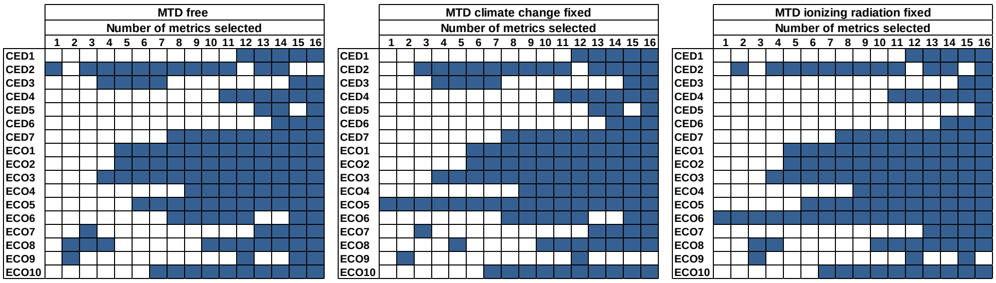


Fig. 7. Comparison of the LCIA metrics selected in the oil category for every number of proxy indicators. The first matrix represents the MTrain free case (all the variables are free), the second and third matrixes correspond to the MTrain climate change fixed case and to the MTrain ionizing radiation fixed case in which the MILP is forced to select climate change and ionizing radiation, respectively, as proxy indicators.

To study redundancies in the impact categories, the model is next solved in an iterative manner. That is, for a given number of proxy indicators, the MILP is first run to identify a set of impact metrics to be used by the multi-linear regression. An integer cut (Balas & Jeroslow, 1972) is then added to exclude this combination of proxy metrics in further iterations. The MILP is then solved again for the same number of proxy indicators but with the following integer cut added:

$$\sum_{i \in ONE(r)} bin(i) - \sum_{i \in ZERO(r)} bin(i) \leq |ONE(r)| - 1 \qquad r \in R \qquad (12)$$

Here $ONE(r) = \{i | bin(i,r)^* = 1\}$ and $ZERO(r) = \{i | bin(i,r)^* = 0\}$, with $bin(i,r)^*$ being the value of the i^{th} component of the vector of binary variables in the optimal solution in iteration *r* of the algorithm. Note that *ONE(r)* and *ZERO(r)* are both obtained from the optimal MILP solution in iteration *r*. This procedure is repeated iteratively in order to generate 10 models (all of them with the same total number of metrics selected, but with different combinations of indicators and approximation errors) for each number of proxy indicators.

Figs. 8 and 9 compare the errors of the best (first) solution identified by the models MTrain and the solutions obtained in the tenth iteration of the iterative procedure applied to the same models. As observed, both solutions behave similarly in the training set (differences in error between 0% and 12.1%), but significantly different in the validation sets (differences in error between 2.0% and 44.4%). In particular, in the training set, the biggest difference between the base case (i.e., the model generated without any integer cut) and the solution obtained with the model with integer cuts added corresponds to the case of selecting one metric, in both, electricity and oil category, with an increase of the error of 11.6% and 12.1% respectively . For more than one LCIA metric, the differences between the base case and the integer cut solution decrease considerably.

Regarding the validation set, in the electricity category the differences between the base case and the model obtained after the tenth iteration are significant. This happens not only for the case of one proxy, but also for up to 8 metrics, point at which the differences between both solutions decrease significantly. The error differences vary from 1% to 15% in the 8 first iterations and are close to 0% in the remaining ones.

On the other hand, in the validation set of the oil category, the models with integer cuts show worse performance than the base case model only for the case of selecting one single proxy indicator (i.e., *n*=1, with an increase in the error of around 45%), while they behave better in the remaining cases.

In addition, the numerical results show that the error is higher in the electricity category than in the oil category. This is because the former contains heterogeneous data, as it covers a wide range of different industrial technologies available to provide electricity. The environmental impact patterns of the products within this category differ substantially, and for this reason the correlation between impacts is low. Hence, eliminating combinations of proxy indicators from the MILP will have a strong negative effect on the predictive capabilities of the multi-linear regression, as the remaining candidates to which the model will have to resort to will likely show lower correlations with the predicted metrics.

Note that in some iterations the integer cut solution behaves slightly better in the validation set than the solution of the base case. This is because the optimal solution of the training set is not guaranteed to be optimal in the validation set.

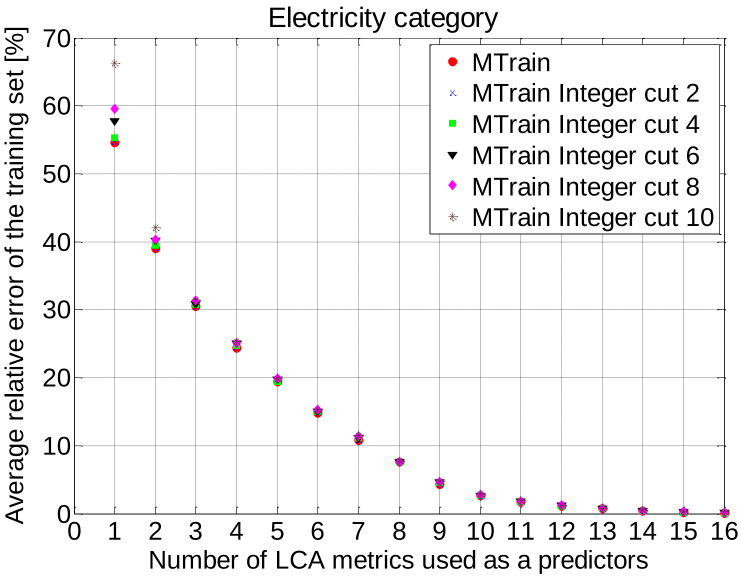


Fig. 8A. Comparison of the average relative error curve of the electricity category for the training set. Each series of points corresponds to a different MTrain model generated after a given number of integer cuts have been included into the MILP formulation.

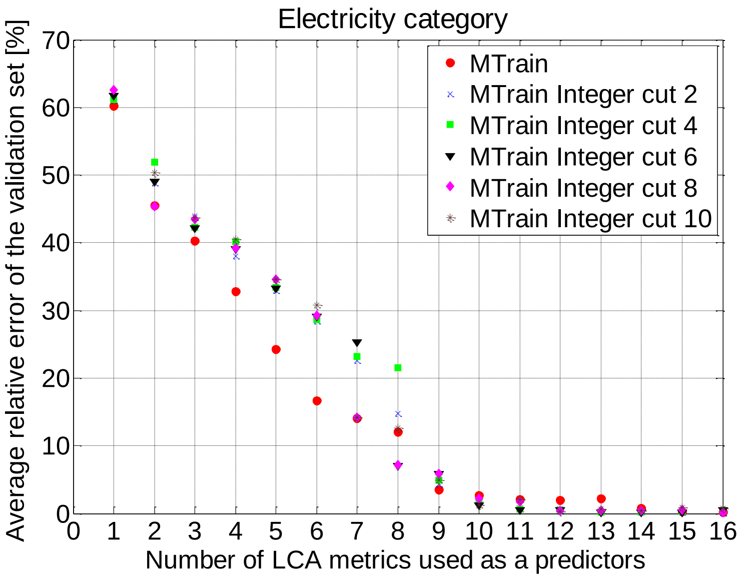


Fig. 8B. Comparison of the average relative error curve of the electricity category for the validation set. Each series of points corresponds to a different MTrain model generated after a given number of integer cuts have been included into the MILP formulation.

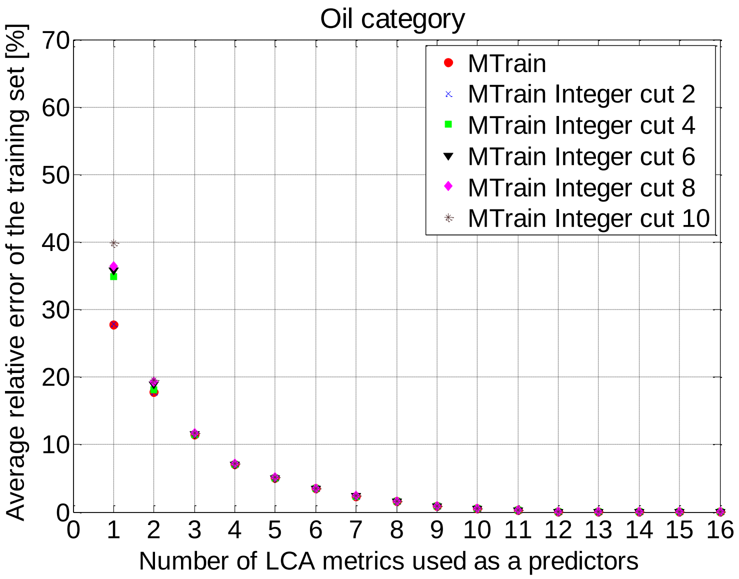


Fig. 9A. Comparison of the average relative error curve of the oil category for the training set. Each series of points corresponds to a different MTrain model generated after a given number of integer cuts have been included into the MILP formulation.

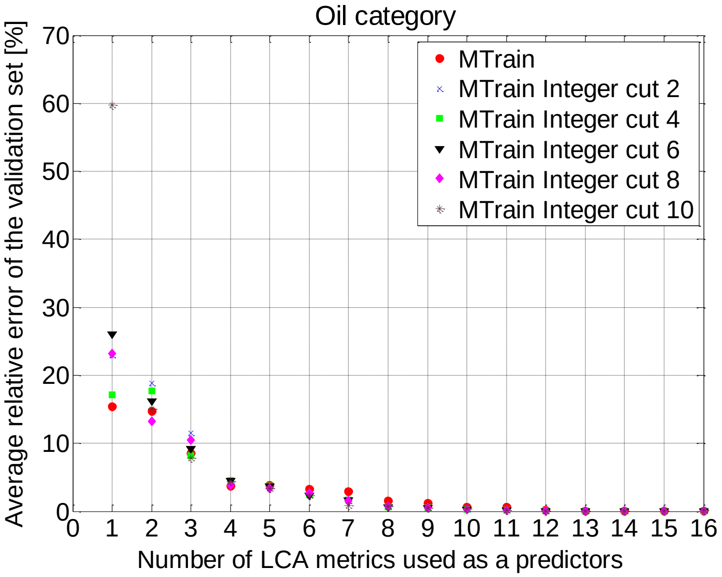


Fig. 9B. Comparison of the average relative error curve of the oil category for the validation set. Each series of points corresponds to a different MTrain model generated after a given number of integer cuts have been included into the MILP formulation.

3.5 Conclusions

This paper presented a systematic approach to simplify LCA studies that combines multi-linear regression and mixed-integer linear programming (MILP). Our algorithm automates the construction of multi-linear regression models that make use of a reduced set of key proxy impact metrics for predicting impact values in other categories with high accuracy (i.e., at minimum relative error).

This approach was applied to 2 categories of the ecoinvent database, electricity and oil, which include 141 and 90 products, respectively. The numerical analysis, which covered 17 LCIA metrics, aimed to identify the best impacts for predicting the others via multi-linear regression. Numerical results show that few indicators suffice to describe the environmental performance with accuracy, with error values below 20% using 5 proxy LCIA metrics in the electricity category, and below 15% for 3 proxy indicators in the oil category. Furthermore, it was found that the model constructed using 80% of the products (training set) behaves in the remaining 20% of the products (validation set) almost as well as the best model that could be built for the validation set (errors below 9%). This demonstrates that the proposed approach is able to make accurate environmental predictions for products for which they were not trained. In addition, it was found that different combinations of LCIA metrics lead to similar approximation errors, suggesting the existence of redundant LCIA metrics.

Our findings will help to develop streamlined LCA studies that will focus on predicting the environmental impact of a product from a reduced set of key proxy indicators. Our approach could lead to significant savings in time and resources associated with data collection. In addition, this work opens new avenues for developing more effective environmental regulations that will focus on controlling a reduce number of key impacts (as their minimization will very likely result in the minimization of other damage categories). Future work will apply a similar approach to the analysis of LCA data retrieved from other databases.

3.6 Acknowledgements

The authors wish to acknowledge the financial support received from the University Rovira i Virgili and from the Spanish Ministry of Education and Science (CTQ2012-37039 and DPI2012-37154).

3.7 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

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

$yob(j,i)$ “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 i^{th} component of the vector of binary variables in the optimal solution of iteration r

$ypr(j,i)$ Predicted value of LCIA metric i in observation j

3.8 References

Arena M, Azzone G, Conte A. A streamlined LCA framework to support early decision making in vehicle development. J Clean Prod 2013;41:105–13.

Bala A, Raugei M, Benveniste G, Gazulla C, Fullana-I-Palmer P. Simplified tools for global warming potential evaluation: when good enough is best. Int J Life Cycle Assess 2010;15:489–98.

Balas E, Jeroslow R. Canonical cuts on the unit hypercube. SIAM J Appl Math 1972;23:61–79.

Boustead I, Hancock GF. Handbook of industrial energy analysis. Ellis Horwood Limited; 1979.

Dones R, Bauer C, Röder A. Teil VI Kohle. Ecoinvent report, 6–VI(6), v2.0; 2007.

Faist-Emmenegger M, Heck T, Jungbluth N. Teil V Erdgas. Ecoinvent report, 6–V(6), v2.0; 2007.

Finnveden G, Hauschild MZ, Ekvall T, Guinée J, Heijungs R, Hellweg S, et al. Recent developments in life cycle assessment. J Environ Manag 2009;91:1–21, Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/19716647.

Frischknecht R, Heijungs R, Hofstetter P. Einstein's lessons for energy accounting in LCA. Int J Life Cycle Assess 1998;3(5):266–72.

Frischknecht R, Rebitzer G. The ecoinvent database system: a comprehensive webbased LCA database. J Clean Prod 2005;13(13–14):1337–43.

Goedkoop M, Hofstetter P, Müller-Wenk R, Spriemsma R. The ECO-indicator 98 explained. Int J Life Cycle Assess 1998;3(6):352–60.

Goedkoop M, Spriensma R. The eco-indicator 99: a damage oriented method for life cycle impact assessment. Methodology report. Netherlands: Amersfoort; 2000.

Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 7th ed. Upper Saddle River: Prentice Hall; 2009.

Hanes R, Bakshi BR, Goel PK. The use of regression in streamlined life cycle assessment. In: Proc. ISST; 2013. p. v1.

Hellweg S, Mila i Canals L. Emerging approaches, challenges and opportunities in life cycle assessment. Science 2014;34:1109–13.

Huijbregts MAJ, Hellweg S, Frischknecht R, Hendriks HWM, Hungerbühler K, Hendriks AJ. Cumulative energy demand as predictor for the environmental burden of commodity production. Environ Sci Technol 2010;44:2189–96.

Huijbregts MAJ, Rombouts LJA, Hellweg S, Frischknecht R, Hendriks AJ, Van De Meent D, et al. Is cumulative fossil energy demand a useful indicator for the environmental performance of products? Environ Sci Technol 2006;40:641–8.

Hunt RG, Boguski TK, Weitz K, Sharma A. Case studies examining LCA streamlining techniques. Int J Life Cycle Assess 1998;3(1):36–42.

IBM ILOG, C. CPLEX 12; 2012.

Jeswani HK, Azapagic A, Schepelmann P, Ritthoff M. Options for broadening and deepening the LCA approaches. J Clean Prod 2010;18:120–7.

Jiménez-González C, Ollech C, Pyrz W, Hughes D, Broxterman QB, Bhathela N. Expanding the boundaries: developing a streamlined tool for eco-footprinting of pharmaceuticals. Org Process Res Dev 2013;17:239–46.

Jungbluth N. Teil IV Erdöl. Ecoinvent report, 6–IV(6), v2.0; 2007.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence; 1995. p. 1137–43.

LBP P. GaBi 6 software-system and databases for life cycle engineering; 2015, Retrieved from: http://www.gabi-software.com/

Marwah M, Shah A, Bash C, Patel C, Ramakrishnan N. Using data mining to help design sustainable products. Computer 2011;44:103–6.

Moriarty P, Honnery D. The prospects for global green car mobility. J Clean Prod 2008;16:1717–26.

National Renewable Energy Laboratory. U.S. Life Cycle Inventory Database; 2012, Retrieved from: https://www.lcacommons.gov/nrel/search

Sanjuán N, Stoessel F, Hellweg S. Closing data gaps for LCA of food products: estimating the energy demand of food processing. Environ Sci Technol 2014;48:1132–40.

SETAC. Streamlined life-cycle assessment: afinal reportfromthe SETAC NorthAmerica Streamlined LCA Workgroup, July; 1999.

Simapromanual PRe Consultants.Introduction to LCA with SimaPro 8.Version, 1–77. The Netherlands: PRe Consultants; 2013.

Steinmann ZJN, Venkatesh A, Hauck M, Schipper AM, Karuppiah R, Laurenzi IJ, et al. How to address data gaps in life cycle inventories: a case study on estimating CO_2 emissions from coal-fired electricity plants on a global scale. Environ Sci Technol 2014;48:5282–9.

Sundaravaradan N, Marwah M, ShahA, Ramakrishnan N. Data mining approaches for life cycle assessment. In: Proceedings of the 2011 IEEE international symposium on sustainable systems and technology; 2011. p. 1–6.

Swiss Centre For Life Cycle Inventories. Ecoinvent Database 3.0. Ecoinvent Centre; 2013, Retrieved from: http://www.ecoinvent.org/database/

Vecchietti A, Lee S, Grossmann IE. Modeling of discrete/continuous optimization problems: characterization and formulation of disjunctions and their relaxations. Comput Chem Eng 2003;27:433–48.

Weston N, Clift R, Holmes P, Basson L, White N. Streamlined life cycle approaches for use at oil refineries and other large industrial facilities. Ind Eng Chem Res 2011;50(3):1624–36.

Wolf M, Pennington D, Pant R, Chomkhamsri K, Pretato U, European Commission. European Reference Life Cycle Database (ELCD). Database 2008:1–30.

4 PAPER 3: STATISTICAL ANALYSIS OF GLOBAL ENVIRONMENTAL IMPACT PATTERNS USING A WORLD MULTI-REGIONAL INPUT–OUTPUT DATABASE

Janire Pascual-González[1], Gonzalo Guillén-Gosálbez[1,2], Josep M. Mateo-Sanz[1], Laureano Jiménez-Esteller[1]*

[1]Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Campus Sescelades, Avinguda Països Catalans, 26, 43007 Tarragona, Spain

[2]Centre for Process Integration, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, UK

Abstract

Understanding how anthropogenic impacts are generated at a global scale is a major challenge to face. This work studies the environmental impact patterns of the wealthiest nations using environmentally extended multi-regional input-output tables. A multivariate statistical analysis is performed on data covering 69 environmental indicators (classified into 5 main categories: energy, emissions, material, water and land), and 41 countries. This analysis shows that damages in different categories (and also within the same one) are highly correlated and that the wealthiest countries display very similar environmental impact patterns. These findings might help to develop more effective environmental regulations that will focus on controlling a reduced number of key indicators. In addition, the analysis of pollution patterns at a global scale will help to establish unified environmental regulations in countries with similar patterns.

Keywords: Environmentally extended multi-regional input-output model; Multivariate statistical analysis; Environmental impact pattern; Life Cycle Assessment.

4.1 Introduction

The study of the mechanisms by which environmental impacts are generated and embodied in international trade channels has recently gained wider interest. In this context, environmentally extended multi-regional input-output tables (EEMRIO) have emerged as a useful tool to assess the impact of economic activities on the environment (Tukker et al., 2013; Watson and Moll, 2008). These models attribute pollution or resources depletion to the final demand of a product or service following a consistent holistic approach (Wiedmann, 2009) that makes them very useful in the development of environmental policies.

A key point in their use and, more generally, in the area of environmental engineering, involves the way in which the environmental performance is assessed. A plethora of environmental indicators have been proposed for quantifying the anthropogenic damage in different categories (Arvidsson et al., 2012; Cerdan et al., 2009; Herva et al., 2011; Veleva and Ellenbecker, 2001). Among them, those based on Life Cycle Assessment (LCA) principles have recently become the prevalent approach (Finnveden et al., 2009; Hellweg and Mila i Canals, 2014; Jeswani et al., 2010). A wide variety of impact assessment methods based on LCA currently exist (Hermann et al., 2007; Lu et al., 2013). Some of these metrics have been used in the context of input-output (IO) models (Chang et al., 2014; Cicas et al., 2007; Hendrickson et al., 2006; Junnila, 2008), but their interactions at a global scale are still poorly understood.

Environmentally extended input-output models can be used as a standalone tool or combined with multi objective optimization. The latter approach has been used to identify key economic sectors in the economies of Korea (Cho, 1999), Taiwan (Hsu and Chou, 2000), Portugal (Oliveira and Antunes, 2004), Spain (San Cristóbal, 2010), Greece (Hristu-Varsakelis et al., 2010) and Japan (Lin, 2011), whose regulation reduces significantly the total impact without compromising to a large extent the economic output.

The aforementioned works focus on single economies. On the contrary, the study of environmental impact patterns of nations at a global scale has received much less attention to date. The analysis of the impact patterns of nations can play a major role in sustainability because it could assist in the development of more effective environmental policies in several ways (Baumann and Cowell, 1999). First, identifying impacts that behave similarly is important for developing simpler regulations that will focus on a reduced number of key pressure indicators (since their control will very likely keep other similar indicators within the desired limits). Second, nations with similar environmental impact patterns might implement unified regulations that are likely to be effective for all of them.

This work analyzes environmental impact patterns at a global scale by applying a statistical analysis on data retrieved from a world multi-regional input-output table covering a wide variety of nations and impacts. Numerical results show that the environmental pressure indicators as well as the environmental impact patterns of the wealthiest nations tend to be highly correlated. This insight might be used by public policy makers seeking to reduce the impact at a global scale.

4.2 Methods

In this section, the fundamentals of input-output tables are discussed in first place before describing the statistical analysis carried in this work.

4.2.1 WIOD database

The World Input-Output Database (WIOD) is used in the calculations. This database was developed to analyze the effects of globalization on trade patterns, environmental pressures, and socio-economic development across a wide set of countries (Timmer et al., 2012). WIOD describes the economic inputs and outputs (in monetary terms) of 35 manufacturing sectors covering 27 EU countries and 13 other major countries in the world for the period 1995-2009. The level of disaggregation of the database, which was originally selected based on initial data-availability, ensures a maximum level of detail without the need to specify additional information that are usually lacking in some national data records. The list of countries included in the database is given in Table 1. Note that the input-output data for the rest of the world is aggregated into a single hypothetical country called “Rest of World”.

Table 1. List of countries that appear in WIOD database

European Union		America	Asia and Pacific
Austria	Latvia	Brazil	Australia
Belgium	Lithuania	Canada	China
Bulgaria	Luxembourg	Mexico	India
Cyprus	Malta	United States	Indonesia
Czech Republic	Netherlands		Japan
Denmark	Poland		Russia
Estonia	Portugal		South Korea
Finland	Romania		Taiwan
France	Slovak Republic		Turkey
Germany	Slovenia		
Greece	Spain		
Hungary	Sweden		
Ireland	United Kingdom		
Italy			

4.2.2 Multi-regional IO model

The WIOD database contains a multi-regional input-output (IO) table that covers a wide range of transactions of goods and services between several economic regions (Leontief, 1936; Miller and Blair, 1985). The multi-regional IO model used in the calculations is based on the original formulation of Leontief (1970), which is adequately modified in order to account for several economic regions. Let us consider an economy with *R* regions *r* and *I* sectors *i* in each region. The equations of the IO model for this system can be expressed in compact form as follows:

$$X(i,r) = \sum_{j}\sum_{r'} X(j,r')a(i,j,r,r') + y(i,r) \qquad \forall i,r \qquad (1)$$

where *X(i,r)* is the total output in currency units (e.g., US$) of sector *i* in region *r*, *a(i,j,r,r')* denotes the technological coefficients calculated with Eq. (2) and *y(i,r)* is the final demand (end user) of sector *i* for region *r*.

$$a(i,j,r,r') = \frac{x(i,j,r,r')}{X(j,r')} \qquad \forall i,j,r,r' \tag{2}$$

In Eq. (2), the symbol *x(i,j,r,r')* represents the output of sector *i* of region *r* acting like an input for sector *j* of region *r'*. The coefficients *a(i,j,r,r')* represent the amount (in US$) of output of sector *i* of region *r* necessary to produce one dollar of output of sector *j* of region *r'*. Note that Eq. (1) defines a system of linear equations with I·R equations and unknowns. This system can be solved for a given fixed demand *y* and a set of technological coefficients *a*. The environmental impact of an economy is calculated using the so-called "pollution intensity" vector, which indicates the impact caused in a given environmental category per monetary unit traded:

$$IMPACT(k) = \sum_i \sum_r impact(i,k,r) = \sum_i \sum_r X(i,r)e(i,k,r) \qquad \forall k \tag{3}$$

where *impact(i,k,r)* is the environmental indicator in category *k* associated with sector *i* of region *r*, while *e (i,k,r)* is the environmental pollution intensity for sector *i* of region *r* (i.e., pressure indicator per monetary unit traded). Finally, *IMPACT(k)* is the total environmental impact in category *k* generated by all of the sectors of the economy.

A distinction is made between production-based and consumption-based impact of a nation (Peters, 2008). The former is caused by the economic activities taking place within the limits of the country (these include activities producing goods that are either exported or consumed internally). The consumption-based impact of a region is caused by all the economic activities (taking place anywhere in the world) that generate the amount of goods/services demanded by that region. Note that some of these economic activities will be located in the region of interest while others may operate abroad (and all together cover the whole life cycle of those goods and services consumed by the country under study). The study of consumption-based emissions (as opposed to production-based emissions), avoids the potential masking of the environmental impact of a nation that might occur when displacing the manufacturing tasks to countries with weaker environmental regulations.

The consumption-based emissions are therefore obtained as follows:

$$CBE(r) = \sum_{i} X^{C}(i,r)e\,(i,k,r) \qquad \forall r \qquad (4)$$

where X^C denotes the economic transactions required to fulfill the demand of region r. The value of X^C is obtained by solving the following system of linear equations with I·R equations and unknowns:

$$X^{C}(i,r) = \sum_{j}\sum_{r'} X(j,r)a(i,j,r,r') + y(i,r) \qquad \forall i,r \qquad (5)$$

where the demand y corresponds to the demand of the region.

4.2.3 Pressure indicators

The WIOD database includes a set of environmental satellite accounts that cover energy, air emission accounts, materials extraction, land use and water use indicators. The calculation of these indicators was carried out using information provided by the International Energy Agency (IEA) (Table 2) for the year 2009. Note that the approach followed takes into account the impact generated in all of the stages in the life cycle of the goods/services being analyzed, regardless of the location where the impact occurs.

Table 2. List of energy commodities, materials, land and water covered in the WIOD database

Flow	
Blue water	Electricity
Green water	Heat
Grey water	Nuclear
CO_2	Hydroelectric
CH_4	Geothermal
N_2O	Solar power
NOx	Wind power
SOx	Other sources
CO	Distribution losses
Non-methane volatile organic compounds	Animal biomass (used)
NH_3	Feed biomass (used)
Arable land	Food biomass (used)
Permanent crops	Forestry biomass (used)
Permanent meadows and pastures	Other biomass (used)
Productive forest area	Coal (used)
Hard coal and derivatives	Natural gas (used)
Lignite and derivatives	Crude oil (used)
Coke	Other fossil fuels (used)
Crude oil, NGL and feedstocks	Non-metallic minerals for construction (used)
Diesel oil for road transport	Other non-metallic minerals (used)
Motor gasoline	Metals (used)
Kerosene and gasoline	Animal biomass (unused)
Light fuel oil	Feed biomass (unused)
Heavy fuel oil	Food biomass (unused)
Naphta	Forestry biomass (unused)
Other petroleum products	Other biomass (unused)
Natural gas	Coal (unused)
Derived gas	Natural gas (unused)
Industrial and municipal waste	Crude oil (unused)
Biogasoline also including	Non-metallic minerals for construction (unused)
Biodiesel	Other non-metallic minerals (unused)
Biogas	Metals (unsed)
Other combustible renewables	

4.2.4 Multivariate statistical analysis

The consumption-based environmental impact associated with each economic region is calculated in first place (Eqs. (4) and (5)). Recall that this impact considers all the economic activities required to fulfill the demand of a region. To perform the calculations, the demand of the region is fixed while the others are set to zero. Let *impact(k,r)* be the environmental indicator in category *k* of region *r* (the pressure indicator associated with the fulfillment of the demand of region *r*). The generic consumption-based matrix shown in Table 3 is therefore obtained in first place:

Table 3. Illustrative example of the impact matrix for the case of *r* regions and *k* categories.

	Impact category 1	Impact category 2	…	Impact category *k*
Region 1	Impact (1,1)	Impact (2,1)	…	Impact (*k*,1)
Region 2	Impact (1,2)	Impact (2,2)	…	Impact (*k*,2)
…	…	…	…	…
Region *r*	Impact (1,*r*)	Impact (2,*r*)	…	Impact (*k*,*r*)

A multivariate statistical analysis is then performed on this matrix using the XLSTAT software (version 2013.3.02), a statistical add-in for Microsoft Excel (Addisonft, 2013).

A correlation analysis is carried out in first place to measure the strength of potential linear and nonlinear relationships between any two variables (i.e., environmental indicators). Each sample/observation used in the study covers the impacts associated with a given economic region. The Mahalanobis Distance methodology (MD) is applied (Mahalanobis, 1936) to discard outliers, thereby improving the robustness of the analysis. Hence, for each of the *R* observations (in this case 41 regions), in a *p*-dimensional multivariate sample (where *p* is the number of variables, which equals 69 environmental pressure indicators), a distance value D_r is calculated as follows:

$$D_r = \sqrt{(CBE_{(r)} - \overline{CBE})^T S^{-1} (CBE_{(r)} - \overline{CBE})} \qquad \forall r \qquad (6)$$

where T is the estimated multivariate location and S the sample covariance matrix.

Under the assumption of multivariate normally distributed data, the impact values follow a Chi-square distribution with p degrees of freedom (Rousseeuw and van Zomeren, 1990). Hence, the MD is usually converted into Chi-square p-values for the analysis. Multivariate outliers are defined as observations having a large (squared) MD. A level of significances of 0.00025 is considered in the analysis.

The Pearson correlation coefficients (PCC) between pressure indicators are calculated after removing the outliers. These coefficients range from totally correlated (−1 or 1), to randomly distributed (0). The sign of the correlation coefficient (positive or negative), defines the direction of the relationship, while the absolute value indicates the strength of the correlation. An index is next calculated for each variable (i.e., pressure indicator) to quantify the extent to which a given pressure indicator correlates with the others. This index is defined as the percentage of pressure indicators for which the correlation test yields a positive result (i.e., percentage of pressure indicators that are correlated with the indicator being analyzed). Hence, the index is calculated as follows:

$$C_k = \frac{p'}{p} \qquad \forall k \qquad (7)$$

where C_k is the correlation index of pressure indicator *k*, *p'* is the number of pressure indicators correlated with it (considering a level of significances of 0.001), and *p* is the total number of pressure indicators left after applying the outliers' methodology. Finally, once the most correlated pressure indicators within each group are identified, an analysis is carried out in order to assess the type of relationship between environmental indicators (i.e., linear, quadratic, logarithmic, exponential, etc.). The regression equations are then fitted using the least-squares approach.

4.3 Results and discussion

The methodology used to eliminate outliers identifies 5 atypical countries: China, India, Poland, United States and "Rest of world", whose pressure indicators' values, in absolute terms, are between 50 and 250 times higher than the average. These regions are removed in order to avoid distortions in the final results of the analysis. After removing them, the next step is to determine the number of indicators that are correlated with each pressure indicator. For this analysis, Pearson correlation coefficients (PCC) for a level of significance of 0.001 are considered (see Fig. 1). The pressure indicators are classified into 5 main categories: energy (orange bars), emissions (brown bars), material (grey bars), water (blue bars), and land (green bars). As an example, the indicator diesel is correlated with 86% of the remaining metrics, while the metric OTHSOURC, which accounts for electricity and heat sources not included as individual categories in the database (i.e., sources that do not belong to any single category), correlates with only 5% of them.

Pressure indicators are, in general terms, highly correlated, with 72% of them being correlated with more than 50% of the others. These results reveal that pressure indicators behave similarly, that is, when a pressure indicator increases so do the others and vice-versa. The most correlated pressure indicator is diesel (86%), followed closely by Non-methane volatile organic compounds (NMVOC) (85%), and NOx (83%). These findings are consistent with others studies performed on individual process technologies (Huijbregts et al., 2010, 2006).

Considering the different categories, it is observed that indicators based on emissions are the most correlated ones (all the pressure indicators of this category show a correlation index above 74%), while land indicators show lower correlation indexes, varying from 28% to 58%.

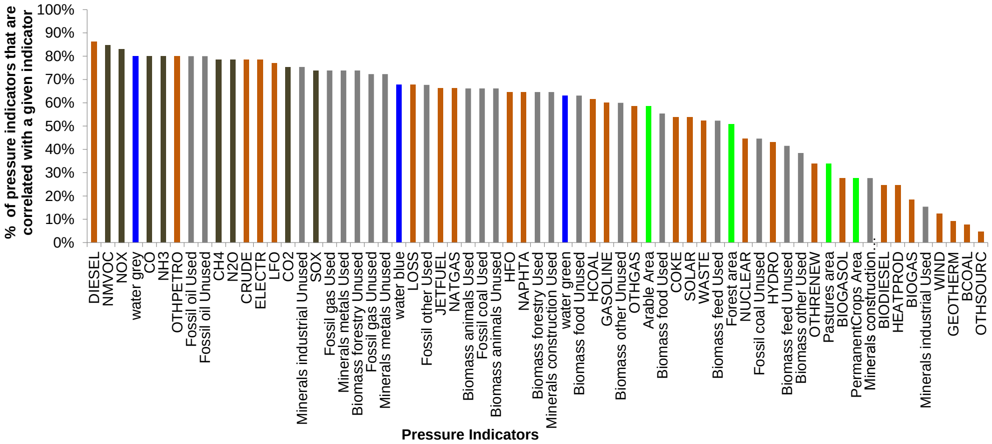


Fig. 1. Percentage of correlated pressure indicators. Pressure indicators are divided into five categories: energy (orange bars), emissions (brown bars), material (grey bars), water (blue bars) and land (green bars). The height of the bars indicate the percentage of pressure indicators that are correlated with a given indicator (i.e., the diesel bar shows the degree of correlation between diesel and the remaining indicators, that is, the percentage of indicators that are correlated with diesel).

A heat map is next constructed from the correlation matrix in order to get further insight into the strength of the correlation between indicators,. In the heat map, black squares denote totally correlated indicators, while white squares represent randomly distributed indicators (see Fig. 2). As noted, the strength of the correlation grows with the correlation index.

For convenience in the presentation of the results, pressure indicators are grouped into clusters according to the intensity of the correlation (i.e., from more correlated to less correlated). Hence, the indicators corresponding to the first 6 bars in Fig.1 are grouped into cluster C1, the next 6 indicators into cluster C2 and so on. The final clusters of pressure indicators are shown in Table 4. The full version of the heat map without grouping pressure indicators is given in Appendix. The heat map shows the average Pearson's correlation coefficient between the pressure indicators belonging to each cluster.

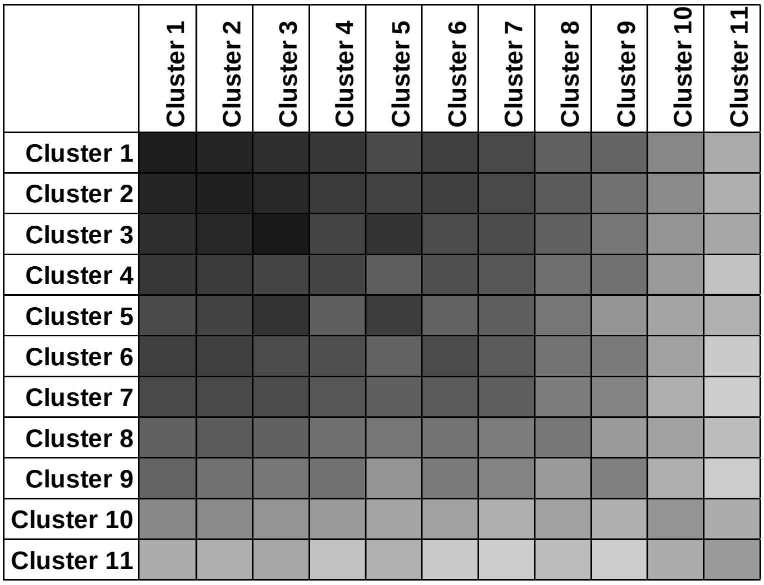


Fig. 2. Heat map of the pressure indicators matrix based on Pearson correlation coefficients.

Table 4. Pressure indicators belonging to each cluster

C1	Diesel oil for road transport	C2	Other petroleum products
	Non-methane volatile organic compounds		Crude oil (used)
	NO_x		Crude oil (unused)
	Grey water		CH_4
	CO		N_2O
	NH_3		Crude il, NGL and feedstocks
C3	Electricity	C4	Metals (used)
	Light fuel oil		Forestry biomass (unused)
	CO_2		Natural gas (unused)
	Other non-metallic minerals (unused)		Metals (unsed)
	SO_x		Blue water
	Natural gas (used)		Distribution losses
C5	Other fossil fuels (used)	C6	Heavy fuel oil
	Kerosene and gasoline		Naphta
	Natural gas		Forestry biomass (used)
	Animal biomass (used)		Non-metallic minerals for construction (used)
	Coal (used)		Green water
	Animal biomass (unused)		Food biomass (used)
C7	Food biomass (unused)	C8	Coke
	Hard coal and derivatives		Solar power
	Motor gasoline		Industrial and municipal waste
	Other biomass (unused)		Feed biomass (used)
	Derived gas		Productive forest area
	Arable land		Nuclear
C9	Coal (unused)	C10	Biogasoline also including
	Hydroelectric		Permanent crops
	Feed biomass (unused)		Non-metallic minerals for construction (unused)
	Other biomass (used)		Biodiesel
	Other combustible renewables		Heat
	Permanent meadows and pastures		Biogas
C11	Other non-metallic minerals (used)		
	Wind power		
	Geothermal		
	Lignite and derivatives		
	Other sources		

Figs. 3-7 depict scatter plots for the most correlated pressure indicators within each category (i.e., the indicators with the largest correlation index within each category). As observed, diesel consumption (which belongs to the "energy" category), depicts linear correlation with the remaining indicators (R^2=0.54-0.82; P< 0.001). On average, 70% of the data variability could be explained by a linear regression. A deeper analysis of the linear correlation reveals that diesel correlates better with pressure indicators belonging to the categories emissions (R^2=0.82), energy (R^2=0.72), and materials (R^2=0.73); and worse with pressure indicators belonging to the categories water (R^2=0.68), and land (R^2=0.47). A plausible explanation of why diesel consumption is highly correlated with some emissions (i.e., NMVOC) as well as with crude oil and minerals use is that they are all correlated with non-renewable resources depletion. On the contrary, diesel shows lower correlation indexes with land and water indicators, mainly because they are not directly related to the extraction and processing of fossil fuels (Hischier et al., 2005; Jungbluth et al., 2005).

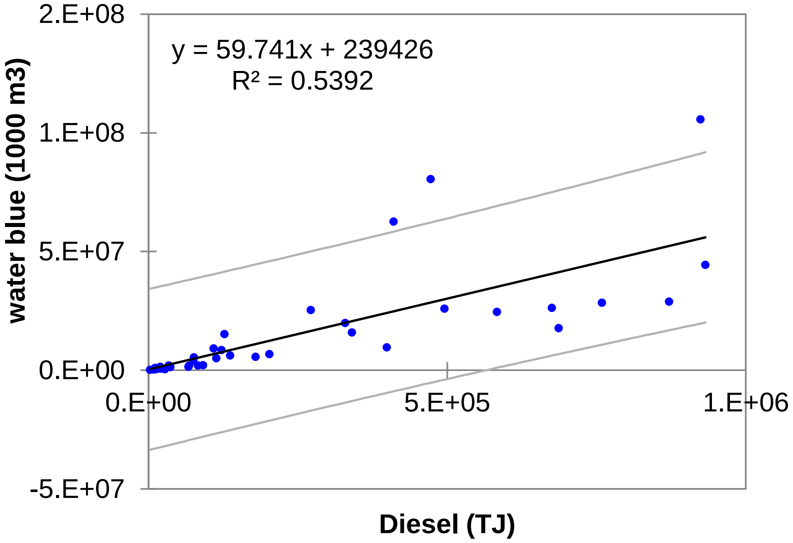


Fig. 3. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 41 countries, for indicators diesel and water blue.

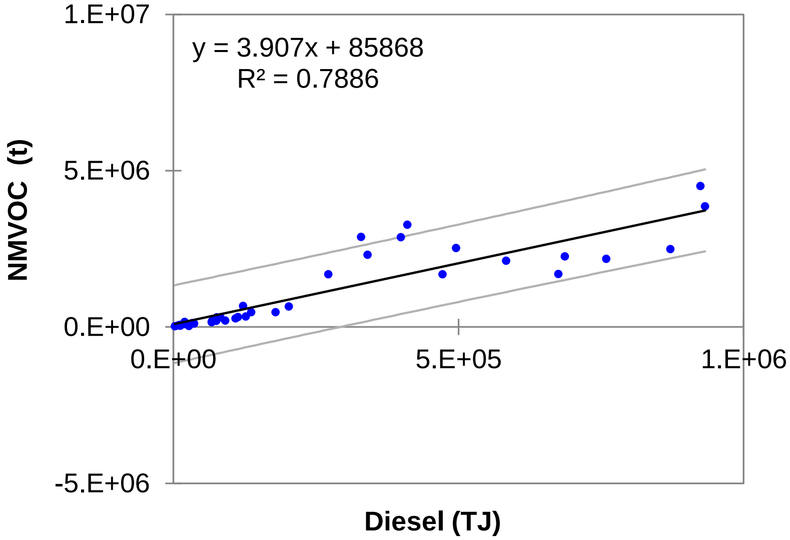


Fig. 4. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 41 countries, for indicators diesel and NMVOC.

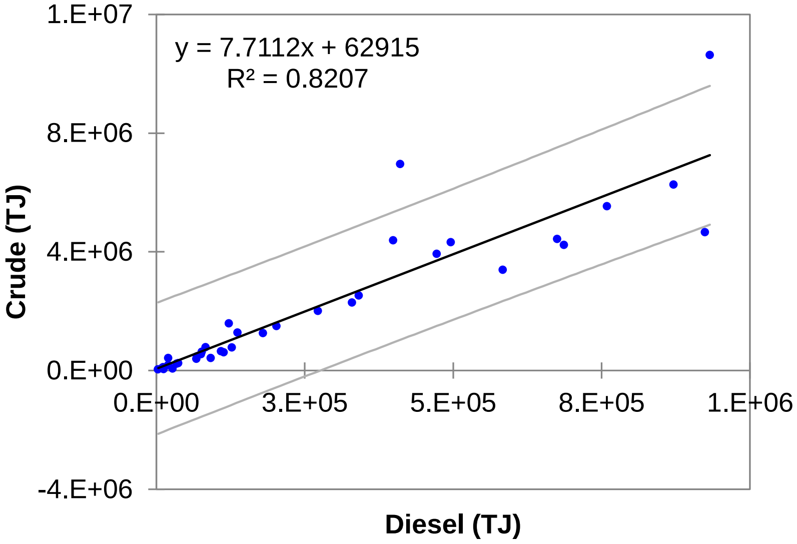


Fig. 5. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 41 countries, for indicators diesel and crude oil.

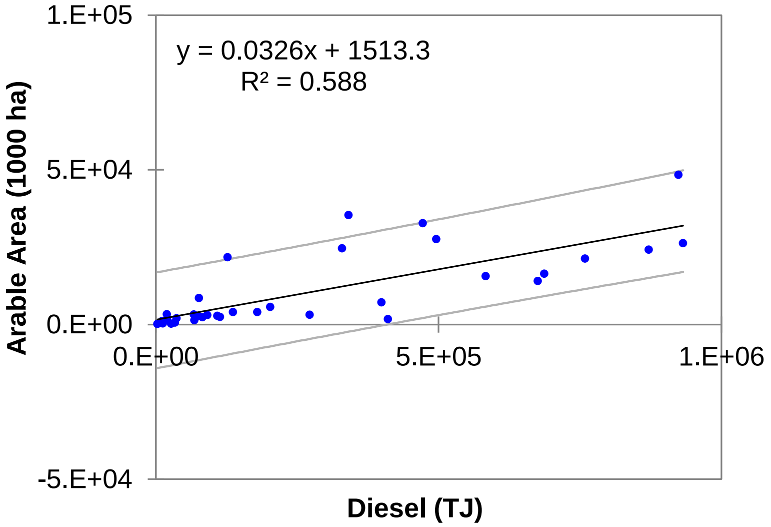


Fig. 6. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 41 countries, for indicators diesel and arable area.

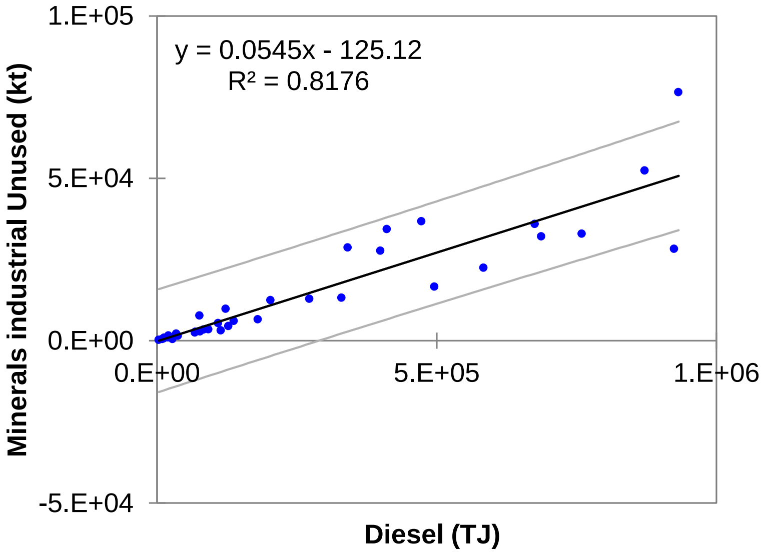


Fig. 7. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 41 countries, for indicators diesel and mineral industrial unused.

The degree of similarity between the environmental impact patterns of nations is finally investigated. The preprocessing step identifies 12 atypical pressure indicators (outliers): water blue, water green, water grey, CH_4, SOx, CO, HCOAL, Crude, Diesel, HFO, and NATGAS, whose values are up to 1000 times higher than the average.

As observed in Fig. 8, the correlation between regions is almost total, since 100% of the countries are correlated with more than 95% of the remaining nations (considering a level of significance of 0.001 based on the Pearson correlation coefficient, PCC). The most correlated country is United Kingdom, whose average PCC is 0.90, followed by Poland (PCC=0.90), and Luxemburg (PCC=0.89). On the other hand, there are only three countries that have a correlation index below 100%: Sweden (98%), France (98%) and Estonia (95%), with average PCC values of 0.73, 0.70 and 0.59, respectively.

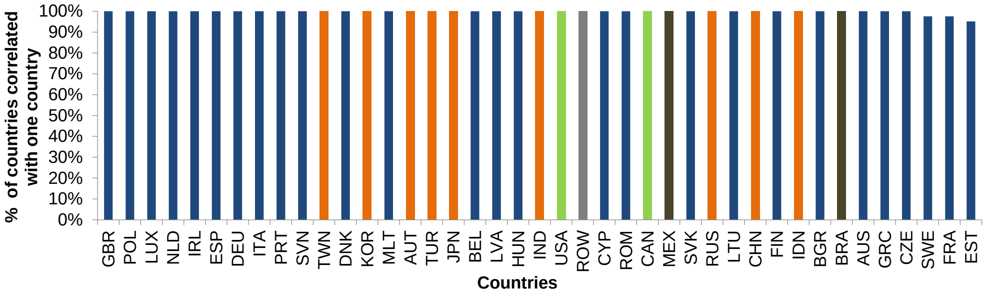


Fig. 8. Percentage of correlated countries. Countries are divided in five groups: European Union (blue bars), North America (green bars), Latin America (brown bars), Asia and Pacific (orange bars) and Rest of World (grey bars).

Fig. 9 shows the average values of pressure indicators in each category along with the lower and upper limits within which they fall in the different nations. As observed, countries pollute with different intensities despite showing similar pollution patterns. The ratios between pressure indicators are almost constant, while absolute values differ from one country to another.

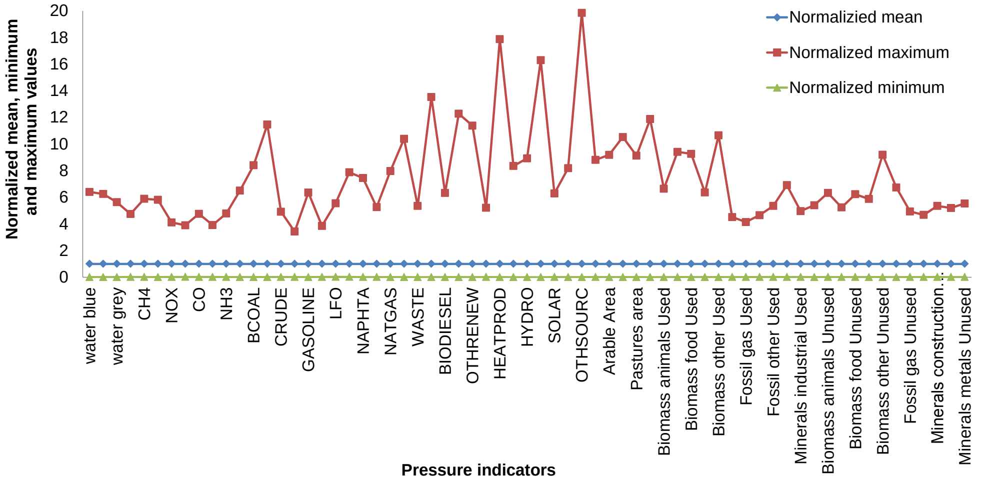


Fig. 9. Mean, minimum and maximum values of the pressure indicators in each category. The data provided is normalized by the average value of the pressure indicators in each category.

Fig. 10 shows, as an example, the pollution intensity map for two environmental indicators, CO_2 and gasoline. As seen, countries like United States or Canada show larger pressure indicator values (in those categories) than Brazil or India, despite having similar environmental impact patterns.

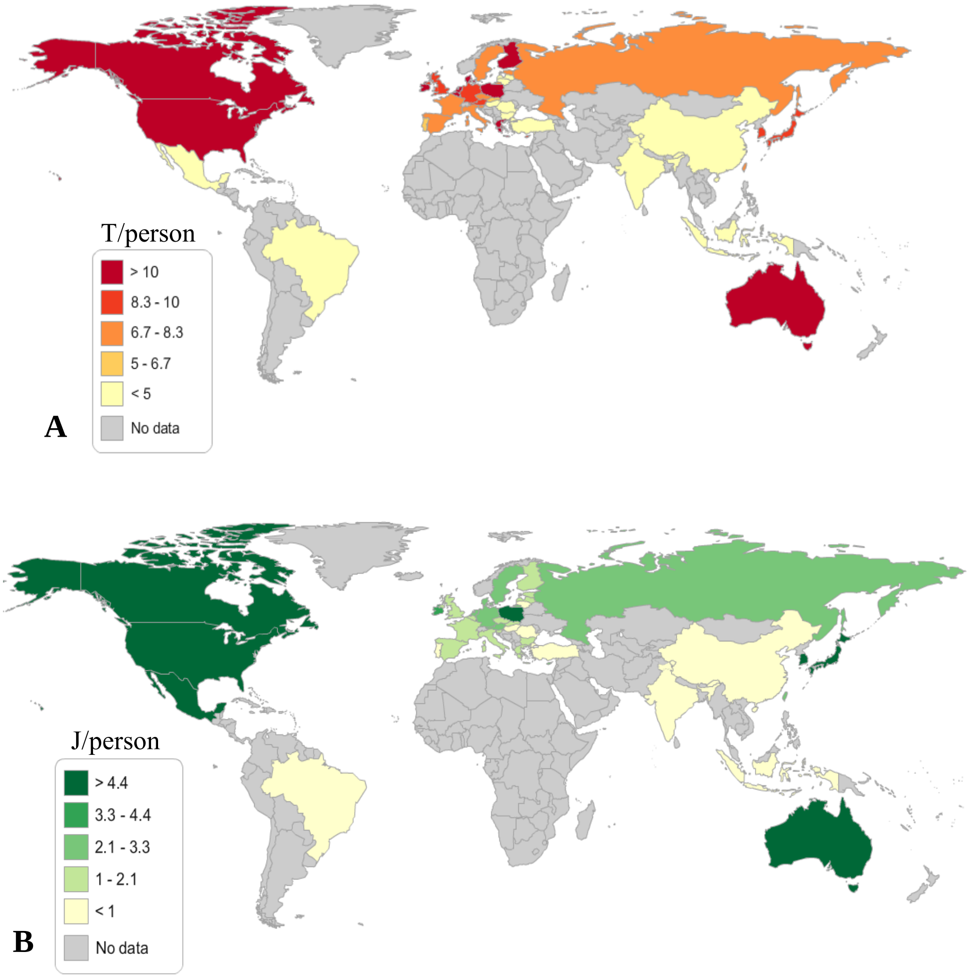


Fig. 10. Consumption-based CO_2 emissions (A) and gasoline use (B) per capita in 2009.

Figs. 11-15 show scatter plots for the most correlated countries within each region (i.e., European Union, North America, Latin America, Asia and Pacific). As observed, United Kingdom correlates linearly with the remaining countries (R^2=0.80-0.97; P< 0.001). In fact, on average, 81% of the data variability could be explained by a linear regression.

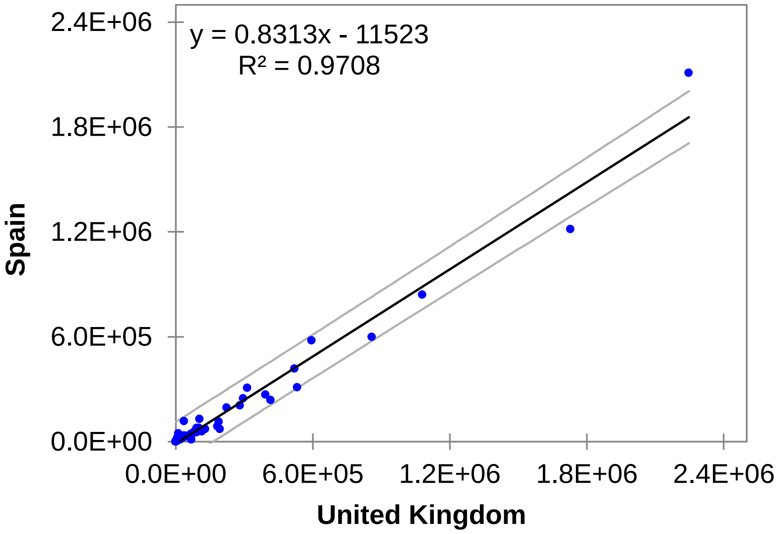


Fig. 11. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 65 pressure indicators for United Kingdom and Spain.

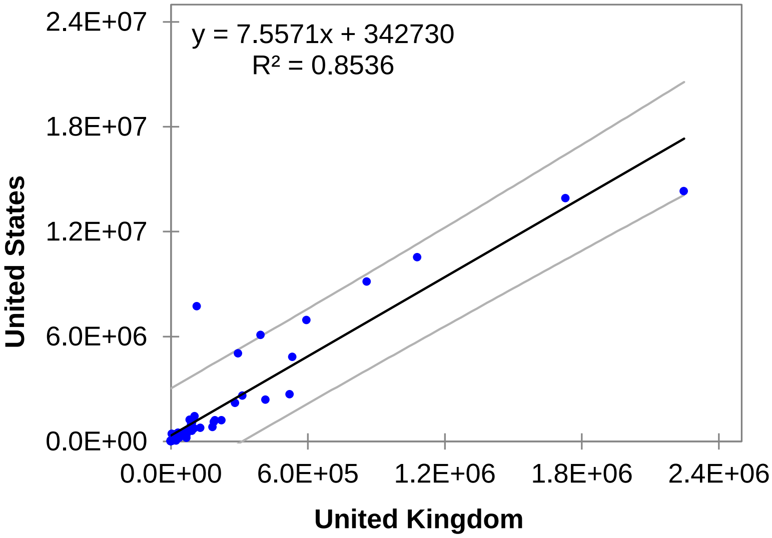


Fig. 12. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 65 pressure indicators for United Kingdom and United States.

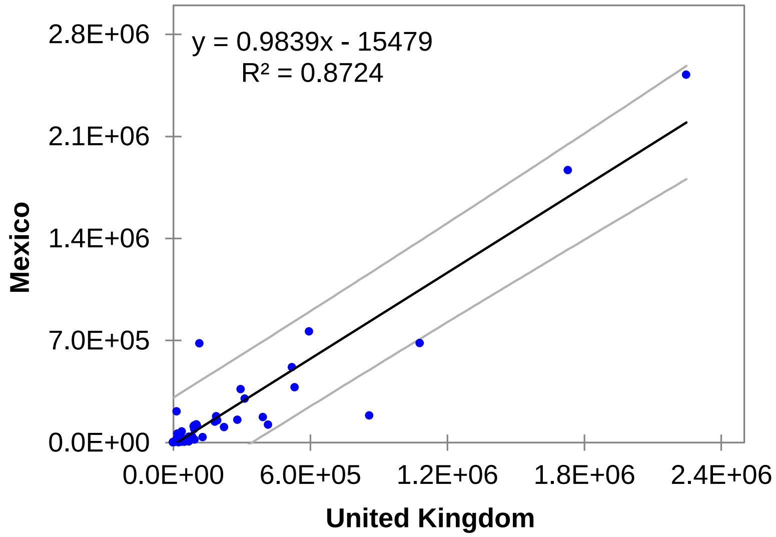


Fig. 13. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 65 pressure indicators for United Kingdom and Mexico.

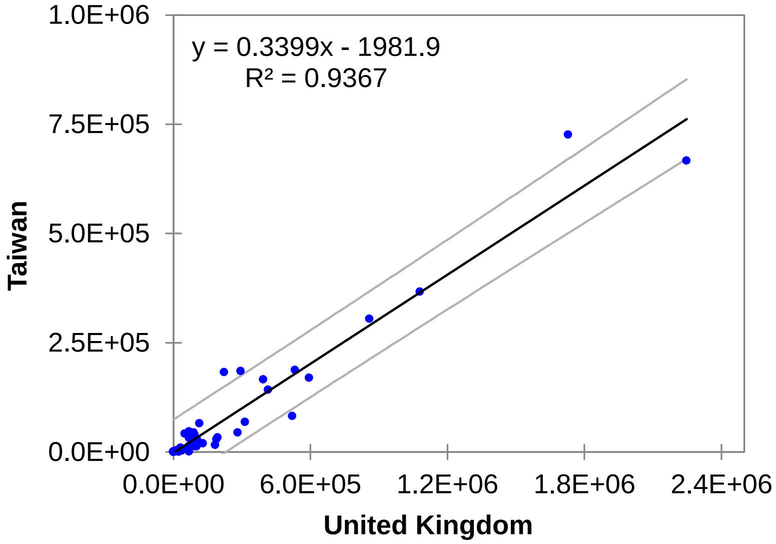


Fig. 14. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 65 pressure indicators for United Kingdom and Taiwan.

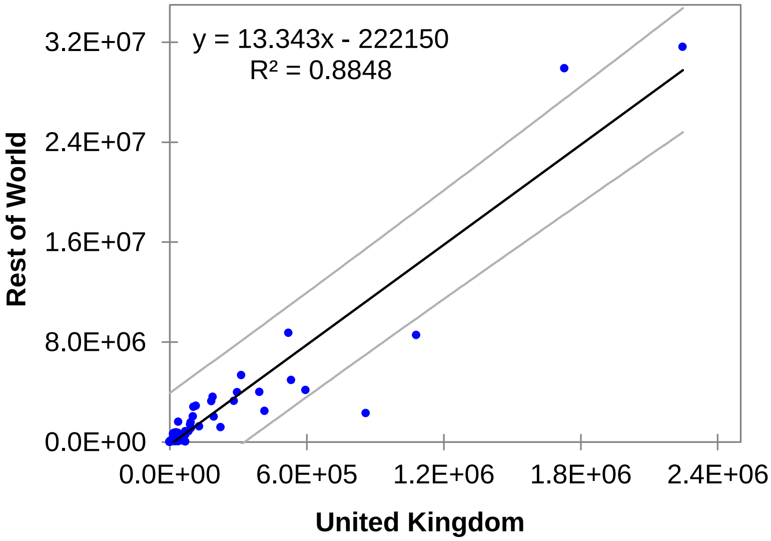


Fig. 15. Linear regression plots with 95-percentile confidence intervals (grey lines), based on 65 pressure indicators for United Kingdom and Rest of World.

4.4 Conclusions

This paper studied the environmental impact patterns of the wealthiest economies using environmentally extended input-output tables. The analysis covered 69 environmental indicators and 41 countries representing more than 85% of the world's GDP. The methodology followed combines multivariate statistical analysis and multi-regional input-output models within a single unified framework. This approach allows identifying relationships between pressure indicators as well as environmental impact patterns.

The results show that the environmental indicators are highly correlated, with 72% of them being correlated with more than 50% of the rest. Furthermore, it was found that indicators based on emissions show higher degrees of correlations (they are correlated, on average, with 79% of the remaining indicators), while land use metrics show lower correlations (they are correlated, on average, with 43% of the indicators).

In addition, it was found that nations show similar environmental impact patterns, despite polluting with different intensity.

These findings shed light on how the impact is generated at a global scale. This knowledge can be used to develop simpler environmental regulations that will focus on a reduced number of key indicators, thereby leading to significant savings in time and resources. Furthermore, a unified environmental legislation might be developed and effectively applied to countries displaying similar environmental impact patterns.

4.5 Acknowledgements

The authors wish to acknowledge the financial support received from the University Rovira i Virgili and from the Spanish Ministry of Education and Science (CTQ2012-37039 and DPI2012-37154).

4.6 Nomenclature

a(i,j,r,r')	Technological coefficients
CBE(r)	Consumption-based emissions
C_k	Correlation index of pressure indicator *k*
D_r	Mahalanobis Distance
e (i,k,r)	Environmental pollution intensity for sector *i* of region *r*
EEMRIO	Environmentally extended multi-regional input-output
GDP	Gross Domestic Product
IEA	International Energy Agency
impact(i,k,r)	Environmental indicator in category *k* associated with sector *i* of region *r*

IMPACT(k)	Total environmental indicator in category *k* generated by all the sectors of the economy
IO	Input-output
LCA	Life Cycle Assessment
MD	Mahalanobis Distance methodology
p	Total number of pressure indicators after applying the outliers' methodology
p'	Number of pressure indicators correlated with it
PCC	Pearson correlation coefficient
S	Sample covariance matrix
WIOD	World input-output database
X^C	Economic transactions required to fulfill the demand of region *r*
x(i,j,r,r')	Output of sector *i* of region *r* acting like an input for sector *j* of region *r'*
X(i,r)	Total output in currency units (e.g., US$) of sector *i* in region *r*
y(i,r)	Final demand (end user) of sector *i* for region *r*

4.7 References

Addinsoft, S., 2013. 2013: XLSTAT software.

Arvidsson, R., Fransson, K., Fröling, M., Svanström, M., Molander, S., 2012. Energy use indicators in energy and life cycle assessments of biofuels: Review and recommendations. J. Clean. Prod. 31, 54–61.

Baumann, H., Cowell, S.J., 1999. An Evaluative Framework for Conceptual and Analytical Approaches Used in Environmental. Greener Manag. Int. 109.

Cerdan, C., Gazulla, C., Raugei, M., Martinez, E., Fullana-i-Palmer, P., 2009. Proposal for new quantitative eco-design indicators: a first case study. J. Clean. Prod. 17, 1638–1643.

Chang, Y., Ries, R.J., Man, Q., Wang, Y., 2014. Disaggregated I-O LCA model for building product chain energy quantification: A case from China. Energy Build. 72, 212–221.

Cho, C.-J., 1999. The economic-energy-environmental policy problem: An application of the interactive multiobjective decision method for Chungbuk Province. J. Environ. Manage. 56, 119–131.

Cicas, G., Hendrickson, C.T., Horvath, A., Matthews, H.S., 2007. A regional version of a US economic input-output life-cycle assessment model. Int. J. Life Cycle Assess.

Finnveden, G., Hauschild, M.Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., Suh, S., 2009. Recent developments in Life Cycle Assessment. J. Environ. Manage. 91, 1–21.

Hellweg, S., Mila i Canals, L., 2014. Emerging approaches, challenges and opportunities in life cycle assessment. Science 344(6188), 1109–1113.

Hendrickson, C.T., Lave, L.B., Matthews, H.S., 2006. Environmental Life Cycle Assessment of Goods and Services: An Input-Output Approach, RFF Press. doi:10.2307/302397

Hermann, B.G., Kroeze, C., Jawjit, W., 2007. Assessing environmental performance by combining life cycle assessment, multi-criteria analysis and environmental performance indicators. J. Clean. Prod. 15(18), 1787-1796.

Herva, M., Franco, A., Carrasco, E.F., Roca, E., 2011. Review of corporate environmental indicators. J. Clean. Prod. 19, 1687–1699.

Hischier, R., Althaus, H.-J., Werner, F., 2005. Developments in Wood and Packaging Materials Life Cycle Inventories in ecoinvent. Int. J. Life Cycle Assess. 10, 50–58.

Hristu-Varsakelis, D., Karagianni, S., Pempetzoglou, M., Sfetsos, A., 2010. Optimizing production with energy and GHG emission constraints in Greece: An input-Output analysis. Energy Policy 38, 1566–1577.

Hsu, G.J.Y., Chou, F.Y., 2000. Integrated planning for mitigating CO2 emissions in Taiwan: A multi-objective programming approach. Energy Policy 28, 519–523.

Huijbregts, M.A.J., Hellweg, S., Frischknecht, R., Hendriks, H.W.M., Hungerbühler, K., Hendriks, A.J., 2010. Cumulative energy demand as predictor for the environmental burden of commodity production. Environ. Sci. Technol. 44, 2189–2196.

Huijbregts, M.A.J., Rombouts, L.J.A., Hellweg, S., Frischknecht, R., Hendriks, A.J., Van De Meent, D., Ragas, A.M.J., Reijnders, L., Struijs, J., 2006. Is cumulative fossil energy demand a useful indicator for the environmental performance of products? Environ. Sci. Technol. 40, 641–648.

Jeswani, H.K., Azapagic, A., Schepelmann, P., Ritthoff, M., 2010. Options for broadening and deepening the LCA approaches. J. Clean. Prod. 18, 120–127.

Jungbluth, N., Bauer, C., Dones, R., Frischknecht, R., 2005. Life cycle assessment for emerging technologies: Case studies for photovoltaic and wind power. Int. J. Life Cycle Assess. 10, 24–34.

Junnila, S., 2008. Life cycle management of energy-consuming products in companies using IO-LCA. Int. J. Life Cycle Assess. 13, 432–439.

Leontief, W., 1970. Environmental Repercussions and the Economic Structure: An Input-Output. Rev. Econ. Stat. 52, 262–271.

Leontief, W.W., 1936. Quantitative Input and Output Relations in the Economic Systems of the United States. Rev. Econ. Stat. 18, 105–125.

Lin, C., 2011. Identifying lowest-emission choices and environmental Pareto frontiers for wastewater treatment wastewater treatment input-output model based linear programming. J. Ind. Ecol. 15, 367–380.

Lu, S.-M., Lu, C., Chen, F., Chen, C.-L., Tseng, K.-T., Su, P.-T., 2013. Low Carbon Strategic Analysis of Taiwan. Low Carbon Econ. 4, 12–24.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Nat. Inst. Sci. India 2.

Miller, R.E., Blair, P.D., 1985. Input-output analysis: foundations and extensions. Prentice-Hall.

Oliveira, C., Antunes, C.H., 2004. A multiple objective model to deal with economy-energy-environment interactions, in: European Journal of Operational Research. 370–385.

Peters, G.P., 2008. From production-based to consumption-based national emission inventories. Ecol. Econ. 65, 13–23.

Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc. 85, 633–639.

San Cristóbal, J.R., 2010. An environmental/input–output linear programming model to reach the targets for greenhouse gas emissions set by the kyoto protocol. Econ. Syst. Res. 22, 223–236.

Timmer, M., Erumban, A.A., Gouma, R., Los, B., Temurshoev, U., Vries, G.J., Arto, I., Genty, V.A.A., Neuwahl, F., Rueda-Cantuche, J.M., Villanueva, A., Fracois, J., Pindyuk, O., Pöschl, J., Stehrer, R., 2012. The World Input-Output Database (WIOD): Contents, Sources and Methods, version 0.9.

Tukker, A., Bulavskaya, T., Giljum, S., de Koning, A., Lutter, S., Simas, M., Stadler, K., Wood, R., 2013. The Global Resource Footprint of nations.

Veleva, V., Ellenbecker, M., 2001. Indicators of sustainable production: Framework and methodology. J. Clean. Prod. 9, 519–549.

Watson, D., Moll, S., 2008. Environmental benefits and disadvantages of economic specialisation within global markets and implications for SCP monitoring, in: Paper for the SCORE! Conference. Brussels, Belgium.

Wiedmann, T., 2009. A review of recent multi-region input-output models used for consumption-based emission and resource accounting. Ecol. Econ. 69(2), 211-222.

4.8 Appendix

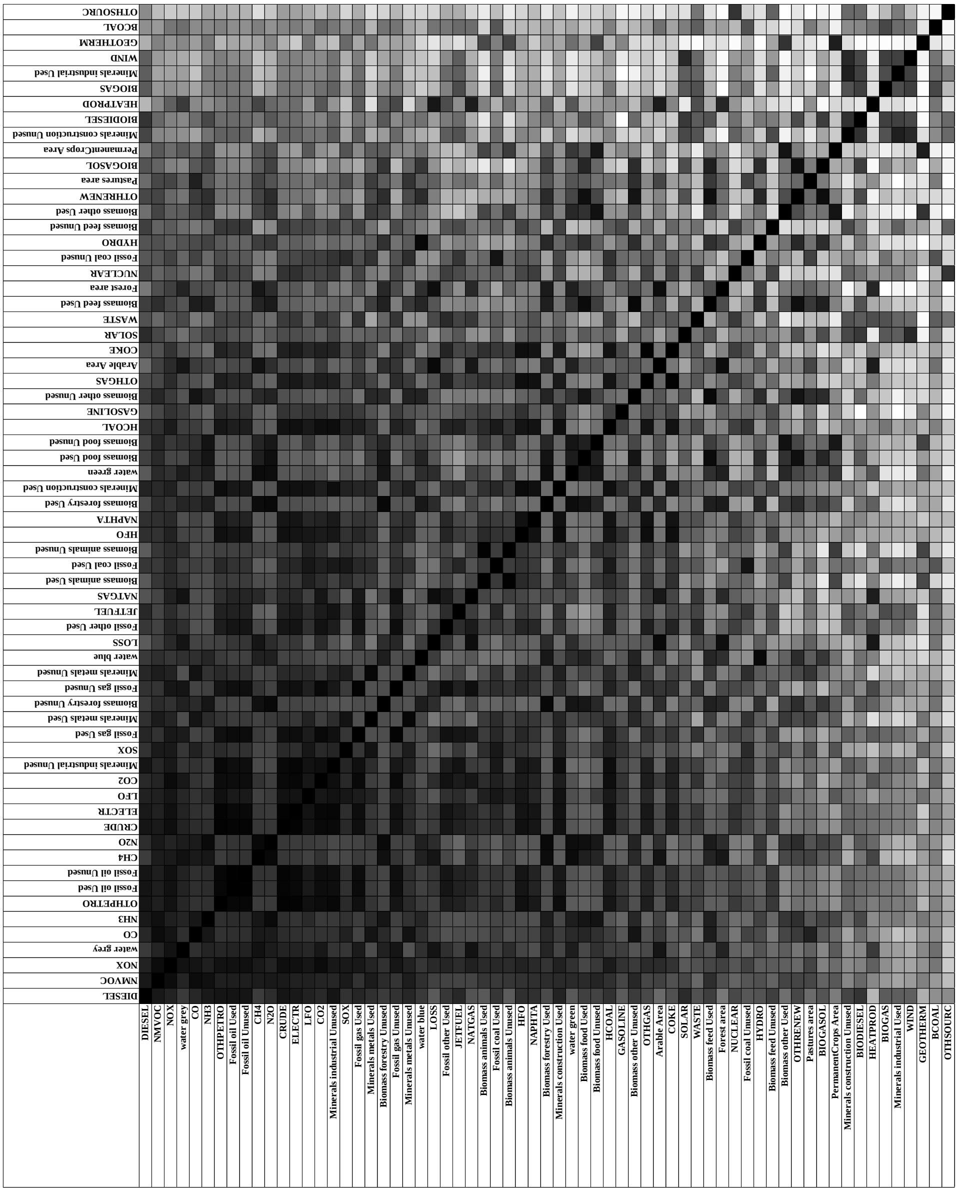


Fig. A.1. Heat map of the pressure indicators matrix based on Pearson correlation coefficients.

5 PAPER 4: MULTI-OBJECTIVE MULTI-REGIONAL INPUT-OUTPUT MODEL FOR MINIMIZING CO_2 EMISSIONS AT A MACRO-ECONOMIC SCALE: APPLICATION TO THE US ECONOMY

Janire Pascual-González[1], , Gonzalo Guillén-Gosálbez[1,2] , Laureano Jiménez-Esteller[1],Jeffrey J. Siirola[3], Ignacio E. Grossmann[3]*

[1]Departament d'Enginyeria Química, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Campus Sescelades, Avinguda Països Catalans, 26, 43007 Tarragona, Spain

[2]Centre for Process Integration, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, UK

[3]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States

Abstract

Designing effective environmental policies for mitigating global warming at a global scale is a very challenging task that requires detailed knowledge of the international channels through which goods and services are traded. Standard environmental regulations focus on reducing the impact in the place of origin regardless of the final destination of the goods produced. This narrow scope might lead to an unfair allocation of responsibilities among the parties involved. This work presents a decision-support tool that minimizes the impact at a global macroeconomic scale by performing changes in the economic sectors of an economy. Our tool combines multi-objective optimization, environmentally extended input-output tables and life cycle assessment within a unified framework. The capabilities of our approach are illustrated through its application to the US economy. Our results identify sectors that should be regulated first to reach a given environmental target while maximizing the demand satisfaction. From the analysis performed, it is concluded that the application of process systems engineering tools at a macroeconomic level can provide valuable insight for public policy makers during the development of more effective environmental regulations.

Keywords: Input-Output analysis, Multi-objective optimization, Linear programming, Global warming potential.

5.1 Introduction

In today's globalized market, countries must face the challenge of reducing their greenhouse gas (GHG) emissions while remaining economically competitive. Policies like the Kyoto Protocol have focused on reducing the direct emissions of nations in an attempt to mitigate global warming on time. It is well known, however, that countries can mask their environmental impact by displacing the manufacturing tasks to regions with softer environmental regulations[1–6]. To avoid this, environmental policies should distinguish between production-based and consumption-based impact. The production-based impact is caused by the facilities operating within the limits of a country. Some of these facilities might produce goods that are exported overseas, so the responsibility of their impact should be assigned to the final consumer rather than to the producer. Conversely, consumption-based impact refers to the impact caused by all the facilities (located anywhere in the world) that produce the goods demanded by a region. By defining environmental policies based on consumption, final customers are penalized for the impact associated with the goods they consume, thereby ensuring a fair scenario where a potential masking of impact via displacement of production facilities is prevented.

It seems clear that in a globalized international market the impact should be assessed on a life cycle basis and across nations (i.e., on a consumption-based basis). Unfortunately, the calculation of the consumption-based impact of a region at a global scale requires large amounts of data that are difficult to collect in practice. The theory behind consumption-based calculations, however, was developed in economics long time ago through the use of input-output models (IO)[7]. These models study economic flows between sectors of the same or different nations and allow for the prediction that changes in the demand of a region have on an entire economy. The original IO approach focused on a single economic region, but was later enlarged in scope in order to deal with several regions simultaneously by covering international transactions between sectors of different nations[8].

Furthermore, it is possible to integrate environmental aspects into IO models, thereby giving rise to environmentally extended input-output models (EEIO)[9]. These models are constructed from standard IO tables by incorporating an additional column that displays the impact associated with the monetary flows between economic sectors. Recent efforts have been undertaken to gather the necessary data to build environmentally extended multi-regional input-output tables (EEMRIO) at a global scale[10,11]. EEMRIO models attribute pollution or resources depletion to the final demand of a product or service following a consistent holistic approach[12], which makes them very useful for policy making.

EEIO models have been integrated recently with multi-objective optimization as a manner to automate the search for alternatives with improved performance at a global macroeconomic level. Some authors applied this approach to the minimization of the environmental impact in the economies of Korea[13], Taiwan[14], Portugal[15], Spain[16], Greece[17] and Japan[18]The aforementioned works have focused primarily on optimizing single economies (without considering international economic transactions). This narrow scope neglects the impact that changes in the economy of one region may have on other overseas economies.

This work introduces a systematic strategy that combines multi-objective optimization and multi-regional input-output models within a unified framework that enables the identification of key economic activities that are contributing marginally to the economy but significantly to the total impact. The main novelty of our approach is that it makes use of a multi-regional model that enables the assessment of the effects that the environmental strategies adopted in a region will have on other nations. This approach leads ultimately to solutions that decrease the impact globally rather than locally. The capabilities of our approach are illustrated through its application to the US economy using information retrieved from the World Input-Output Database (WIOD)[19]. Our final aim is to develop a tool to assist public policy makers in the development of more effective environmental regulations.

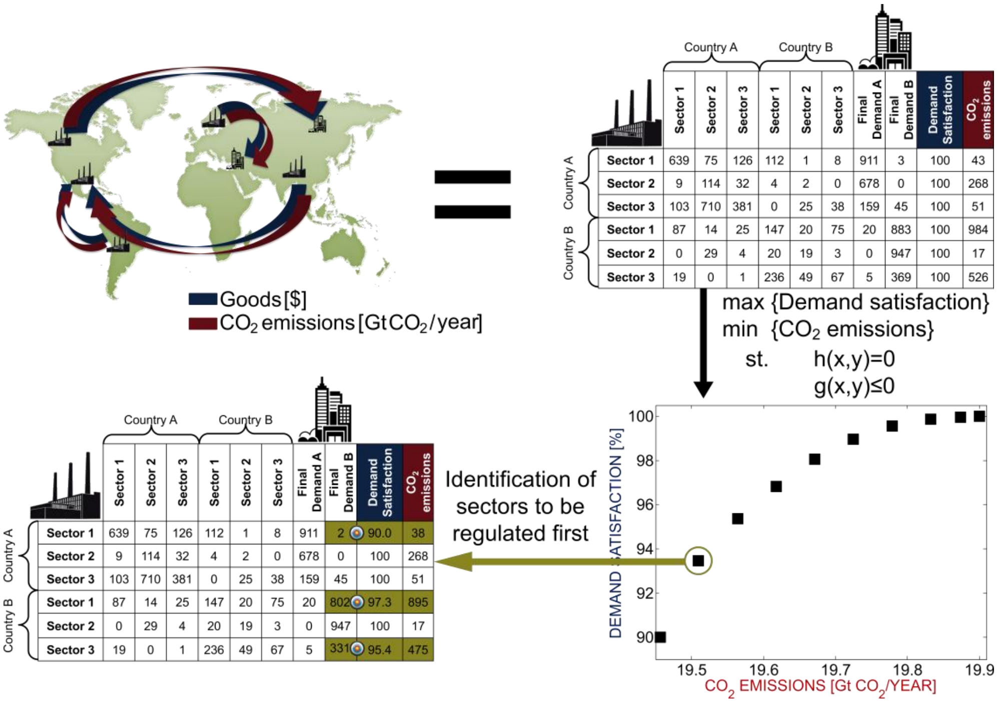
The paper is organized as follows. In section 2 the problem of interest is formally defined, while section 3 introduces the mathematical formulation and the solution method. Section 4 summarizes the main results, including a preliminary analysis of the IO data and a discussion of the optimization results produced by the model, which are generated also for the case of replacing coal by shale gas. Finally, section 5 summarizes the conclusions of our work.

5.2 Problem statement

The problem we aim to solve can be formally stated as follows. We are given macroeconomic information of a set of economic regions. This information covers the economic transactions (sales and purchases of goods and services) taking place between the economic sectors (located in different nations as well as within the same country) that produce the goods and services demanded by the global population. The impact associated with each economic transaction is expressed in the form of pollution intensity vectors that represent the impact caused per unit of money traded. The goal of the analysis is to find the sectors to be regulated in order to simultaneously minimize the CO_2 emissions at a global macroeconomic scale and the changes that need to be performed in the economy in order to achieve such reductions. As will be discussed in more detail later in the article, the second objective is represented through the maximization of the demand satisfaction of the economy.

Note that the outcome of this optimization provides valuable insight for public policy makers, which can use it in different ways. The most straightforward one is to define taxes on the most polluting sectors so as to reduce their demand and therefore the corresponding environmental impact. A decrease in the demand will result in turn in a reduction of the economic flows, and therefore of the gross domestic product of the country. Hence, a more appealing alternative to decrease the impact (without modifying the economy to a large extent) is to foster research on cleaner technologies that will improve the environmental efficiency of the target sectors. This positive environmental effect will eventually propagate to other industrial sectors via trade, thereby enhancing the level of sustainability of the global economy.

5.3 Mathematical formulation


Goods [$]
CO_2 emissions [GtCO_2/year]
=
max {Demand satisfaction}
min {CO_2 emissions}
st. h(x,y)=0
g(x,y)≤0
Identification of sectors to be regulated first
100
98
96
94
92
90
19.5
19.6
19.7
19.8
19.9
DEMAND SATISFACTION [%]
CO_2 EMISSIONS [Gt CO_2/YEAR]

		Country A			Country B						
		Sector 1	Sector 2	Sector 3	Sector 1	Sector 2	Sector 3	Final Demand A	Final Demand B	Demand Satisfaction	CO_2 emissions
Country A	Sector 1	639	75	126	112	1	8	911	3	100	43
	Sector 2	9	114	32	4	2	0	678	0	100	268
	Sector 3	103	710	381	0	25	38	159	45	100	51
Country B	Sector 1	87	14	25	147	20	75	20	883	100	984
	Sector 2	0	29	4	20	19	3	0	947	100	17
	Sector 3	19	0	1	236	49	67	5	369	100	526

		Country A			Country B						
		Sector 1	Sector 2	Sector 3	Sector 1	Sector 2	Sector 3	Final Demand A	Final Demand B	Demand Satisfaction	CO_2 emissions
Country A	Sector 1	639	75	126	112	1	8	911	2	90.0	38
	Sector 2	9	114	32	4	2	0	678	0	100	268
	Sector 3	103	710	381	0	25	38	159	45	100	51
Country B	Sector 1	87	14	25	147	20	75	20	802	97.3	895
	Sector 2	0	29	4	20	19	3	0	947	100	17
	Sector 3	19	0	1	236	49	67	5	331	95.4	475

Fig. 1. Outline of the approach. Environmental impacts are embodied in the flows of goods. Input-output tables describe the economic transactions taking place between sectors of an economy. The solution of a multi-objective model based on input-output tables identifies the sectors that need to be regulated first so as to attain significant improvements in environmental performance with little impact on the economy.

Fig. 1 summarizes the overall approach. In this example, 2 countries and 3 sectors per country are considered. An input-output table, discussed in more detail in the ensuing sections, is constructed in the first place with data on economic transactions between sectors. In this table, rows represent sales of goods/services from one sector to the others as well as to the final consumer, while columns denote purchases from one sector to the others. As an example, sector 1 of country A sells 75 monetary units of goods/services to sector 2 of country A, and purchases 87 monetary units of goods/services from sector 1 of country B.

The input-output tables allow us to quantify the impact from production based and from consumption based perspective, Fig. 2 illustrates the differences between thos two approaches. In this example we consider 4 countries. From a production based approach, A and D are slightly polluting countries, B is highly polluting and C is totally clean. On the contrary, from the consumption based approach, A and C become the most polluting countries, while country B changes from the most polluting to a totally clean country.

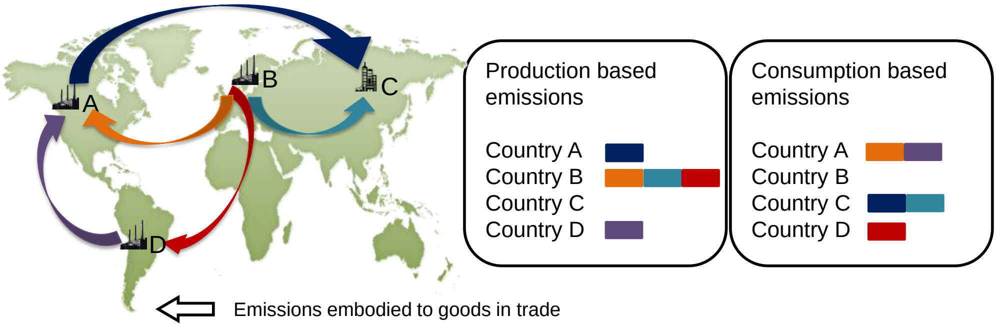

A
B
C
D
Production based emissions
Country A
Country B
Country C
Country D
Consumption based emissions
Country A
Country B
Country C
Country D
Emissions embodied to goods in trade

Fig. 2. Illustrative example of the differences in the quantification of impacts between the production based and the consumption based perspective. The arrows represent the emissions embodied to goods in trade between countries

Taking this IO table as starting point, an optimization model is formulated next and then efficiently solved via optimization methods. The outcome of the bi-objective model (minimization of CO_2 emissions and maximization of demand satisfaction) consists of a Pareto set of alternatives, each representing a different economic plan. The analysis of these Pareto points provides information on the sectors that should be regulated in the very first place to achieve a given environmental target while causing minimum disturbances in the economy (i.e., while maximizing the satisfaction of the current demand).

The approach presented here relies on a bi-objective linear programming model that contains the basic equations of an environmentally extended multi-regional input-output (EEMRIO) table. This section starts by describing IO models, a topic that is typically missing in the standard chemical engineering literature, before presenting the complete mathematical formulation.

5.3.1 Input-Output (IO) model

In its basic form, an input-output model is based on a system of linear equations that describe the distribution of the outcome of an economic sector throughout the economy. Table 1 shows a generic IO table, in which the rows represent the sales between sectors and the columns the purchases.

Table 1. Illustrative example of an IO table for the case of 1 region and 3 industrial sectors.

Sales →

		Sector 1 [$]	Sector 2 [$]	Sector 3 [$]	Final demand [$]	Total output [$]
Purchases ↓	Sector 1 [$]	$x_{(1,1)}$	$x_{(1,2)}$	$x_{(1,3)}$	$y_{(1)}$	$X_{(1)}$
	Sector 2 [$]	$x_{(2,1)}$	$x_{(2,2)}$	$x_{(2,3)}$	$y_{(2)}$	$X_{(2)}$
	Sector 3 [$]	$x_{(3,1)}$	$x_{(1,2)}$	$x_{(3,3)}$	$y_{(3)}$	$X_{(3)}$

For an economy with sectors i, the equations of an IO model can be expressed in compact form as follows:

$$X(i) = \sum_{j} a(i,j)X(j) + y(i) \qquad \forall i \tag{1}$$

where:

X(i), X(j) are variables denoting the total output in currency units (e.g. US$) of sector *i*/*j*.

y(i) is a parameter representing the final demand (end user) of sector *i*.

a(i,j) are parameters denoting the technological coefficients, which are calculated with Eq 2 (note that this equation contains only parameters, so it can be left out of the pure IO model).

$$a(i,j) = \frac{\bar{x}(i,j)}{\bar{X}(j)} \qquad \forall i,j \tag{2}$$

where, $\bar{x}(i,j)$ is the current output of sector *i* acting like an input for sector *j*, while $\bar{X}(j)$ is the current total output in currency units (e.g. US$) of sector *j*. The coefficients *a(i,j)* represent the amount (in US$) of output of sector *i* necessary to produce one dollar of output of sector *j*. The IO model assumes that there is a direct proportionality between the total output of sector *j* and the inputs that this sector acquires from its supplying sectors. Accepting this premise, the technological coefficients *a(i,j)* can be considered constant for a certain period, assuming that the technological conditions of the total production of an economy remain unchanged. IO tables are typically used for predicting changes in the sectors of an economy according to changes in the demand of a single (or several) sectors. This analysis is carried out by fixing the demand to the predicted value and then solving the resulting system of linear equations. This calculation provides the economic flows (corresponding to sectorial transactions) required to satisfy the new demand.

As will be explained in more detail later in this article, our IO model is based on the WIOD database, which covers a wide range of transactions of goods and services between several world economic regions[7,20].

5.3.2 Environmental extension of the IO Model

The purely economic IO table can be modified so as to include environmental aspects, which gives rise to an environmentally extended input-output table (EEIO). To this end, additional rows denoting the pollution intensity of each sector (i.e., impact per unit of money traded) are added to the original table. These new rows contain environmental coefficients for each sector and impact. For an economy with *i* sectors, the following equation is used:

$$Imp(i) = X(i)e(i) \qquad \forall i \tag{3}$$

$$TImp = \sum_i Imp(i) = \sum_i X(i)e(i) \tag{4}$$

where *Imp(i)* is the environmental impact (i.e., global warming potential) associated with sector *i*, while *e(i)* is the environmental pollution intensity for sector *i* (i.e., impact per monetary unit traded). Finally, *TImp* is the total environmental impact generated by all of the sectors of the economy.

5.3.3 *Multi-regional IO Model*

Multi-regional IO tables cover transactions of goods and services between economic sectors of different countries. For an economy with regions *r* and sectors *i* in each region, Eq. 1 should be rewritten as follows:

$$X(i,r) = \sum_{j}\sum_{r\prime} X(j,r')a(i,j,r,r') + y(i,r) \qquad \forall i,r \qquad (5)$$

The following notation is used here:

X(i,r), X(j,r') are variables denoting the total output in currency units (e.g. US$) of sector *i/j* in region *r/r'*.

a(i,j,r,r') are parameters representing the technological coefficients, which are calculated via Eq. 6.

y(i,r) is a parameter denoting the final demand (end user) of sector i of region r.

Note that, similarly to the previous case, for a given demand and technical coefficients, the model takes the form of a system of linear equations with the same number of equations and unknowns. The values of the technical coefficients are obtained from the current values of the economic flows as follows (again note that this equation contains parameters only, so it can be left out of the pure IO model):

$$a(i,j,r,r') = \frac{\bar{x}(i,j,r,r')}{\bar{X}(j,r')} \qquad \forall i,j,r,r' \qquad (6)$$

In Eq. 6, $\bar{x}(i,j,r,r^{'})$ is a parameter denoting the current output of sector i of region r acting like an input for sector j of region r', while $\bar{X}(j,r')$ is another parameter that represents the total current output in currency units (e.g. US$) of sector *j* in region *r'*. Note again that we assume here that the relationship between the amount purchased from a sector to its neighboring sectors and the total output of the sector is constant in a given time period. Hence, the current values of the economic flows are used to calculate the values of the technical coefficients, and these technical coefficients are then employed in the calculation of the economic flows that would be required to satisfy another given demand. Hence, the reader should not confuse the current economic flows (i.e., parameters $\bar{x}(i,j,r,r')$ and $\bar{X}(j,r')$) corresponding to the current demand, with those calculated for a different demand (i.e., variables $x(i,j,r,r')$ and $X(j,r')$). The technical coefficients *a(i,j,r,r')* represent the amount (in US$) of output of sector *i* in region *r* necessary to produce one dollar of output of sector *j* in region *r'*. Taking this into account, the environmental equations can be rewritten as follows:

$$Imp(i,r) = X(i,r)e(i,r) \qquad \forall i,r \tag{7}$$

$$TImp = \sum_{i}\sum_{r} Imp(i,r) = \sum_{i}\sum_{r} X(i,r)e(i,r) \tag{8}$$

where *e(i,r)* is the environmental pollution intensity for sector *i* of region *r* (i.e., impact per monetary unit traded). Finally, *TImp* is the total environmental impact generated by all of the sectors of the economy.

5.3.4 Multi objective optimization problem based on linear programming.

As already mentioned, an IO table leads to a system of linear equations in which the total output of each sector is the unknown variable, while its demand is a fixed parameter. The system of linear equations is typically solved for different demand values (*y(i,r)*), which provides valuable insight into the effect that demand changes have on the economic and environmental performance of the overall economy.

Bearing all this in mind, we use the basic EEMRIO table to develop a multi-objective LP model. On the one hand, we would like to minimize the environmental impact. Since it is assumed that the technologies (and therefore the corresponding pollution intensities) are given, the only option to accomplish this goal is to reduce the economic flows (*x(i,r)*), that is, the economic activity of each sector. This action will reduce in turn the demand satisfaction level attained by the economy. Hence, the goal of the optimization is twofold: to minimize the environmental impact and to minimize the extent to which the economy needs to be modified in order to reduce the impact to the level sought. The latter objective is here modeled through the maximization of the demand satisfaction (i.e., maximization of demand flows, *y(i,r)*). In our case, the environmental impact is quantified via the total CO_2 emissions (note however that any other impact indicator could be used instead). Finally, our approach leads to the following bi-criterion optimization problem:

$$\min\left\{-\sum_{i}\sum_{\mathrm{r}} y(i,r),\ TImp\right\} \tag{9}$$

$$s.t.\quad X(i,r) = \sum_{j}\sum_{r'} X(j,r)a(i,j,r,r') + y(i,r) \qquad \forall i,r$$

$$TImp = \sum_{i}\sum_{r} Imp(i,r) = \sum_{i}\sum_{r} X(i,r)e(i,r)$$

$$\underline{y_0(i,r)} \leq y(i,r) \leq \overline{y_0(i,r)} \qquad \forall i,r$$

$$X(i,r),\ y(i,r),\ TImp,\ Imp(i,r) \in \mathbb{R}^{+}$$

where *Imp(i,r)* denotes the environmental impact (i.e., the CO_2 emissions) produced by sector *i* of region *r*, while *e(i,r)* is the environmental coefficient for sector *i* of region *r*. Finally, *TImp* is the total impact generated by the sectors of the economy.

This LP model seeks to optimize simultaneously the demand satisfaction and the associated CO_2 emissions (*Timp*) at a global scale (i.e., across the world), subject to the standard equations of the input output tables, the environmental equation that quantifies the CO_2 emissions, and a flexible demand constraint. Thus, the model minimizes the total CO_2 emissions regardless of the place where the emissions are released. This approach avoids solutions in which the emissions of a country are minimized by displacing the manufacturing tasks to other regions.

In this formulation, the demand is represented by a continuous variable which is constrained within realistic lower and upper bounds. Hence, as opposed to standard IO tables where *y(i,r)* is a parameter, here it is defined as a variable. With this modeling approach, the model is flexible enough to leave part of the demand unsatisfied, which reflects the situation that would arise when regulating the demand of the sector. The LP identifies in a systematic manner those sectors whose demand needs to be modified in first place so as to achieve a given environmental target while maximizing the demand satisfaction. This information provides valuable insight for public policy makers on how to improve the environmental performance of the global economy. Specifically, the solution calculated by the optimization algorithm can be implemented in practice by: (i) imposing taxes on these key sectors; (ii) improving the environmental efficiency of their technologies; (iii) combining both strategies simultaneously.

5.3.5 Solution method

The solution of the bi-criterion optimization problem described above is given by a set of Pareto solutions representing the optimal trade-off between the conflicting objectives. These Pareto points show the property that it is impossible to improve them simultaneously in all of the objectives without necessarily worsening at least one of the others. There are several methods available for solving multi-objective optimization problems. Without loss of generality, this work applies the epsilon constraint method, which solves a series of single objective sub-problems where one objective is selected as main criterion while the others are transferred to auxiliary constraints that impose bounds on them[21].

5.4 Results

The approach presented was applied to the US economy in order to minimize the CO_2 emissions at a global scale by regulating its economic sectors. This part of the paper is organized as follows. Section 4.1 describes the database used in this work. Section 4.2 provides a preliminary analysis that assesses the CO_2 emissions embodied in the trade of goods and services within US sectors, and between US sectors and other foreign sectors. Section 4.3 summarizes the results obtained with the bi-objective model. Section 4.4 analyzes the effect that replacing coal by shale gas, an emerging trend in the US economy, will have on the outcome of the optimization.

5.4.1 Data source

The World Input-Output Database (WIOD) was used in our calculations. This database was originally developed to analyze the effects of globalization on trade patterns, environmental pressures and socio-economic development across a wide set of countries[19]. The WIOD describes the economic inputs and outputs (in monetary terms) of 35 manufacturing sectors, covering 27 EU countries and 13 other major countries in the world for the period 1995 to 2009. The level of disaggregation, which was chosen on the basis of initial data-availability exploration, ensures a maximum level of detail without the need for additional information that is typically lacking in the system of national accounts. The 35-industry list is identical to the list used in the EUKLEMS database[22], but shows an additional breakdown of the transport sector. The list of manufacturing sectors is given in Table 2, while the list of countries covered by the database is given in Table 3. The preliminary analysis is simplified by grouping the 35 manufacturing sectors into 6 main sectors according to the type of activity (see Table 2).

Table 2. List of manufacturing sectors that appear in WIOD-database

Business	**Services**
Financial Intermediation	Hotels and Restaurants
Renting of M&Eq and Other Business Activities	Education
Construction	Health and Social Work
Retail Trade, Except of Motor Vehicles ; Repair of Household Goods	Other Community, Social and Personal Services
Sale, Maintenance and Repair of Motor Vehicles Retail Sale of Fuel	Public Admin and Defense; Compulsory Social Security
Wholesale Trade and Commission Trade, Except of Motor Vehicles	Private Households with Employed Persons
Industry	Real Estate Activities
Coke, Refined Petroleum and Nuclear Fuel	**Technology**
Chemicals and Chemical Products	Electrical and Optical Equipment
Rubber and Plastics	Post and Telecommunications
Other Non-Metallic Mineral	Machinery, Nec
Electricity, Gas and Water Supply	Manufacturing, Nec; Recycling
Food, Beverages and Tobacco	**Transport**
Textiles and Textile Products	Transport Equipment
Leather, Leather and Footwear	Inland Transport
Pulp, Paper, Paper , Printing and Publishing	Water Transport
Primary sector	Air Transport
Agriculture, Hunting, Forestry and Fishing	Other Supporting and Auxiliary Transport Activities; Activities of Travel Agencies
Mining and Quarrying	
Wood and Products of Wood and Cork	
Basic Metals and Fabricated Metal	

Table 3. List of countries that appear in WIOD database

European Union		America	Asia and Pacific
Austria	Latvia	Brazil	Australia
Belgium	Lithuania	Canada	China
Bulgaria	Luxembourg	Mexico	India
Cyprus	Malta	United States	Indonesia
Czech Republic	Netherlands		Japan
Denmark	Poland		Russia
Estonia	Portugal		South Korea
Finland	Romania		Taiwan
France	Slovak Republic		Turkey
Germany	Slovenia		
Greece	Spain		
Hungary	Sweden		

5.4.2 Data analysis

Production-based emissions of US industrial sectors

We first studied the extent to which every sector of the economy contributes to the overall CO_2 emissions. Fig. 3 shows a breakdown of the US production-based CO_2 emissions according to the sector of origin. Every bar in the figure represents the total emissions of each economic sector, which was quantified following a production-based approach; that is, the figure shows the emissions released within the limits of US (and regardless of the final destination of the goods produced). The production-based CO_2 emissions of sector i of country r (denoted by $Imp^P(i,r)$) are calculated from the sales of the sector and the associated pollution intensity, as follows:

$$Imp^P(i,r) = X^P(i,r)e(i,r) \qquad \forall i, r = US \qquad (10)$$

where $X^P(i,r)$ represents the sales of sector i of region r, and $e(i,r)$ is the pollution intensity (environmental coefficient for sector i of region r expressed in Gt CO_2 per US$).

Note that the CO_2 emissions are originated from economic transactions that produce goods consumed by either national (dark blue bars in Fig. 3) or international (light blue bars in Fig. 3) customers.

The total production-based US emissions were 4.2 Gt in 2009, while the total exported emissions were 0.3 Gt. More than half of the emissions generated within US belong to the sector industry. A more disaggregated analysis (see Fig. A.1. in the appendix) shows that activities related to chemical engineering (sectors: *coke, refined petroleum and nuclear fuel, chemicals and chemical products and rubber and plastics*) represent 9% of the total emissions, while the production of utilities (sector *electricity, gas and water supply*) represents 48% of the total emissions.

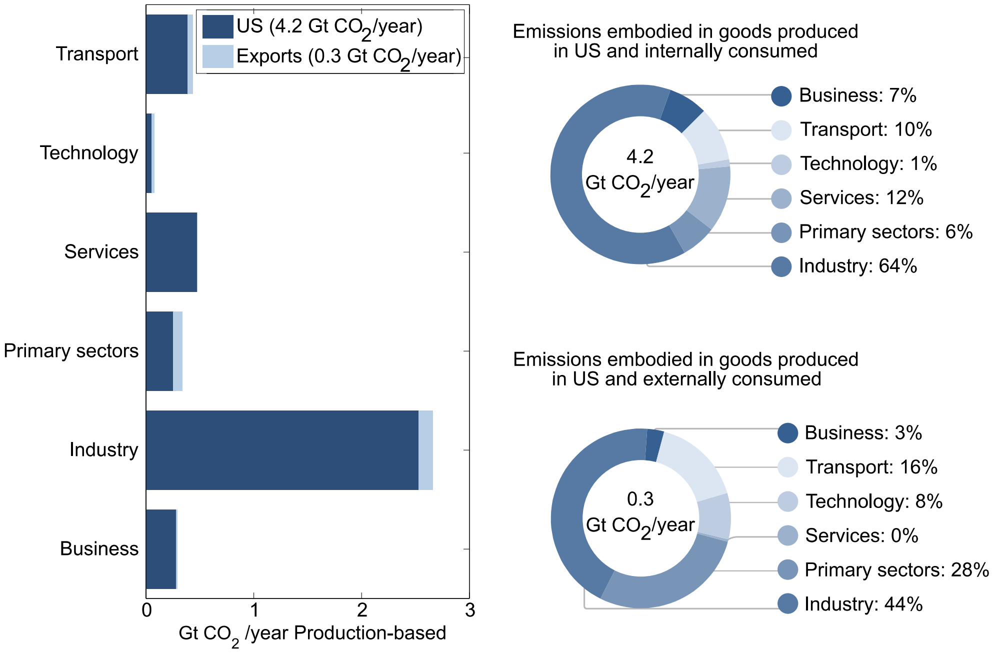


Fig.3. Dark blue bars represent the breakdown of total production-based CO_2 emissions generated within the limits of US (total emissions equal 4.2 Gt CO_2/year). Light blue bars are the breakdown of CO_2 emissions exported via trade (total exported emissions equal 0.3 Gt CO_2/year.

Consumption-based emissions of US industrial sectors

The consumption-based emissions of US consider the CO_2 emissions associated with all the facilities located anywhere in the world that cover the demand of every single sector of US, either directly (i.e., sectors that send goods that cover the demand of the US sector) or indirectly (sectors whose output is used as intermediate input by other sectors that ultimately cover the demand of the US sector). The consumption-based CO_2 emissions (denoted by $Imp^C(i,r)$) are therefore obtained as follows:

$$Imp^C(i,r) = \sum_{r} \sum_{i\prime} X^C(i',r)e\,(i',r) \qquad \forall i, r = US \quad (11)$$

where X^C denotes the economic transactions required to fulfill the demand of sector i of region r. Note that, as opposed to the production-based emissions of sector i, the consumption-based ones might be associated with sectors different from i that produce goods used as intermediate products to ultimately cover the demand of i. The value of X^C is obtained by solving the following system of linear equations with |I|·|R| equations and unknowns:

$$X^C(i,r) = \sum_{j} \sum_{r'} X^C(j,r')a(i,j,r,r') + y(i,r) \qquad \forall i, r = US \quad (12)$$

where demand $y(i,r)$ corresponds to the demand of sector i in region r (i.e., US). Note that this equation considers all the economic transactions required to satisfy the demand of every sector of the US economy regardless of the place where they take place.

The total US consumption-based emissions are 4.8 Gt (versus 4.2 Gt of production-based emissions), while the total imported emissions are 1.1 Gt (versus 0.3 Gt of CO_2 emissions exported). Hence, almost 90% of the total CO_2 emissions (4.2 out of 4.8 Gt) attributed to the US economy are generated by internal activities, while the remaining 10% are imported from abroad via trade. This 10% mismatch between production-based and consumption-based emissions shows that the US is masking part of its impact by importing goods and services from abroad.

Fig.4 shows the results of this analysis, where each bar denotes the total emissions associated with the manufacturing tasks (taking place in any sector of any country) required to fulfill the demand of every US sector (regardless of the region and sector where they occur). As an example, to fulfill the demand of the sector industry, US needs to emit 1.5 Gt of CO_2 emissions within its boundaries, while other countries need to emit 0.31 Gt that are "imported" by the US economy via trade. On the other hand, this sector produces 2.5 Gt of CO_2, 0.14 Gt of which are exported (see Fig.3). Note that these 2.5 Gt of CO_2 are associated with the facilities of this sector that aim to fulfill either the intermediate demand of other sectors or the final demand of the sector itself.

As observed, the economic activities associated with the sector industry are responsible for a large amount of emissions (2.53 Gt CO_2, which represents 64% of the total US production-based emissions, as shown in Fig. 3), while the emissions released for satisfying the demand of the sector are significantly lower (1.51 Gt CO_2, which represents 38% of the total US consumption-based emissions in Fig. 4). This means that most of the emissions generated by the sector are ultimately associated with other sectors that purchase goods/services from it and use them as intermediate products. Hence, the sector industry is indeed the largest ultimate source of impact, but in practice its outputs are used by other sectors that should share the corresponding responsibility.

Within the sector industry (see Fig. A.2 in the appendix), 22% of the direct consumption-based emissions are associated with the subsector electricity, gas and water supply. Chemical engineering sectors represent 9% of the production-based emissions, and 7% of the consumption-based ones.

The mismatch between production-based and consumption-based emissions is further explored in Fig.5, which shows a breakdown of the emissions of the industry sector according to the ultimate destination of the goods. As observed, the main sectors that have transactions with the sector industry are the same sector itself (54%), followed by services (23%) and business (11%).

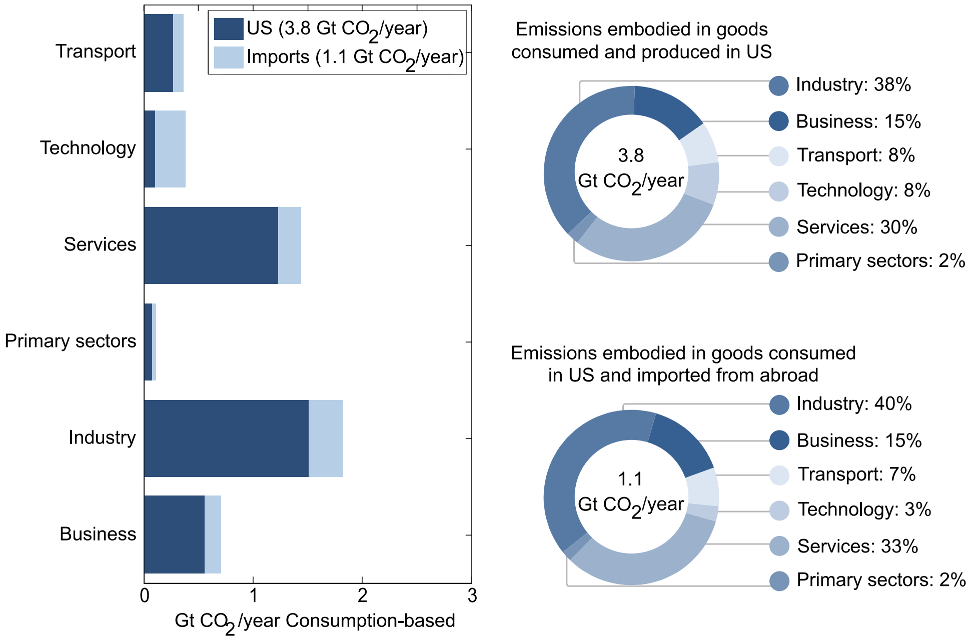


Fig. 4. Dark blue bars represent the breakdown of total consumption-based CO_2 emissions generated to satisfy the demand of each US sector (total emissions equal 3.8 Gt CO_2/year). Light blue bars are the sectorial breakdown of CO_2 emissions imported via trade (total imported emissions equal 1.1 Gt CO_2/year.

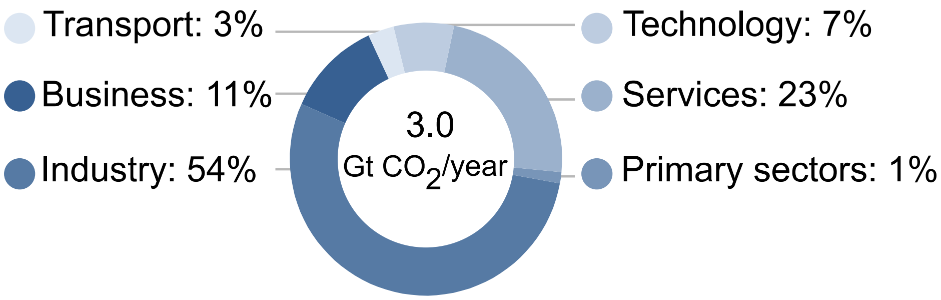


Fig. 5. Breakdown of the emissions of the sector industry in 2009 according to the final demand of the sectors. Each portion represents the percentage of production-based CO_2 emissions generated by the sector industry that are attributed to the intermediate demand of each US sector.

Fig. 6 shows a more detailed comparison between consumption-based and production-based emissions for each of the sectors of the US economy. Those sectors close to the line have a lower mismatch between production-based and consumption-based emissions (e.g., sector transport). In sectors below the line, the production-based emissions exceed the consumption-based ones (e.g., sector industry), while in the sectors above the line, the opposite situation occurs (e.g., sector technology). As already discussed, the overall mismatch between production-based and consumption-based emissions is around 10%. However, this mismatch can be significantly larger on a sectorial basis. More precisely, consumption-based emissions are significantly higher than production-based emissions in the sectors business (ratio of 143%), services (202%) and technology (401%), while they are lower in sectors industry (32%) and primary sectors (67%). This was expected, as part of the output of industrial and primary sectors is used to provide services, develop technology and run businesses. A more detailed analysis of this issue covering the subsectors within each sector is provided in Fig. A.4. of the appendix. Regarding the chemical engineering activities, we found that sector *coke, refined petroleum and nuclear fuel* is a net producer sector (its consumption-based emissions are 34% lower than its production-based emissions); while sectors *chemicals and chemical products and rubber and plastics* are net consumer sectors (consumption-based emissions are 4% and 52% higher than production-based emissions, respectively.

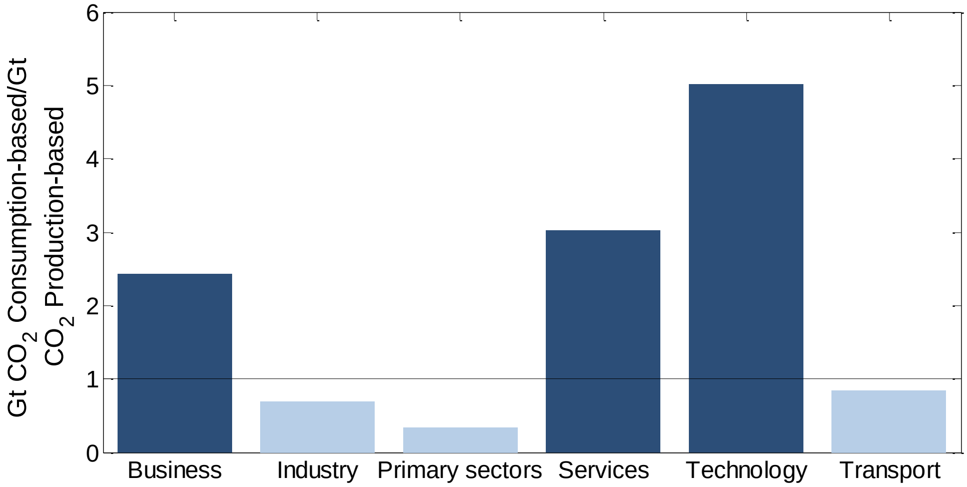


Fig. 6. Comparison between the consumption (dark blue bars) and production-based (light blue bars) accounting approaches in 2009. Each bar represents one industrial sector.

Fig. 7 shows a more detailed spatial analysis of the geographical distribution of the emissions traded that covers the top countries (and their industrial sectors) with which US exchange goods and services. Note that "Rest of World" (ROW) accounts for the joint emissions of several countries.

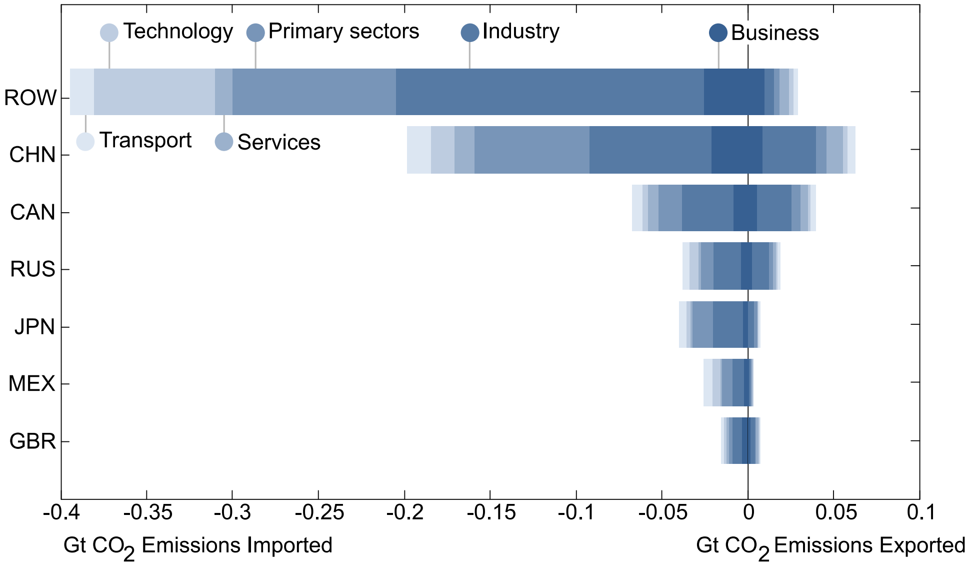


Fig. 7. Countries with higher trade of CO_2 with US in 2009. ROW = Rest of World; CHN = China; CAN = Canada; RUS = Russia; JPN = Japan; MEX = Mexico; GBR = United Kingdom.

As observed, trade is larger between countries like China, Canada, Russia, Japan, Mexico, Great Britain and the nations accounted for in "Rest of the World". Regarding the breakdown of emissions by sectors, we found that industry and primary sectors cover 68% and 55% of the USA imported/exported emissions, respectively. These results are consistent with the work by David and Caldeira (2010)[1].

5.4.3 *Multi-objective optimization*

The multi-objective IO model described previously was applied to minimize the impact of the US economy at a global scale (considering all the emissions required to satisfy the US demand). For convenience in the presentation of the results, the demand satisfaction level is expressed as the percentage of the total demand that is effectively covered (note however that the objective that is maximized is the summation of the demand flows rather than the percentage of demand satisfied). This percentage is obtained as follows:

$$DSat = 100 \sum_{i} \frac{y(i,r)}{y_0(i,r)} \qquad \forall r = US \qquad (13)$$

where demand *y(i,r)* corresponds to the optimized demand of sector *i* in region *r* (i.e., US) and *y₀(i,r)* is the current demand of sector *i* in region *r* (i.e., US). In the calculations, we assume that the optimized demand must fall within 90% to 100% of the actual demand.

The resulting LP model features 5,742 variables and 4,308 constraints. It was implemented in the General Algebraic Modeling Software (GAMS v 24.4.1) and solved with CPLEX v12.6.1.0. The CPU time varied between 15.77 and 44.35 CPU seconds depending on the instance being solved.

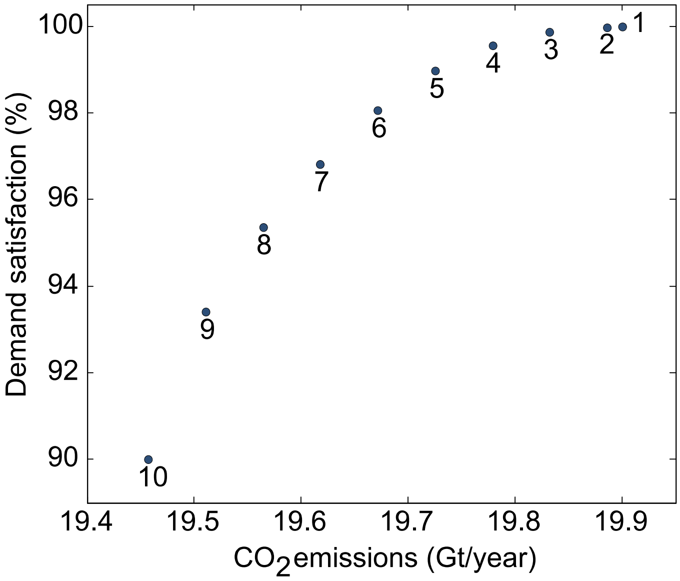


Fig. 8. Pareto optimal frontier for global CO_2 production-based emissions (Gt/year) vs demand satisfaction (%) in 2009.

Fig. 8 shows the 10 Pareto points obtained using the epsilon constraint method. The Pareto frontier, as expected from the LP nature of the model, is concave with the slope increasing as we move to the left. Hence, as we go from the maximum demand satisfaction solution (solution 1) to the minimum impact one (solution 10), greater reductions of demand satisfaction are required for a given reduction of CO_2 emissions.

Each point of the curve corresponds to a different macroeconomic alternative in which sectors are classified into 3 main groups: Those with a demand hitting its lower bound, those with a demand hitting its upper bound, and only one sector with a demand lying between the lower and upper bound. Hence, an important outcome of the optimization is the number of sectors whose final demand is modified to reach a given environmental target. The number of sectors regulated increases as we move from the maximum demand satisfaction solution (all sectors fully cover the final demand) to the minimum impact one (all the demands hit the lower bound of 90%).

Table 4 displays the ratio between the optimal reduction in CO_2 emissions and the corresponding drop in demand satisfaction for every point of the Pareto frontier:

$$Ratio = \frac{\mathrm{CO_2\ emissions\ reduction\ \ (\%)}}{\mathrm{demand\ unsatisfaction\ \ \ (\%)}} \tag{14}$$

Note that the values of this *Ratio* are consistent with the concave nature of the Pareto set. In the same table, the *Cut sectors* row indicates the number of productive sectors whose final demand must be modified to reach the corresponding environmental target (note that there are in total 1435 sectors, that is, 35 sectors and 41 countries).

Table 4. Optimal solutions found for the CO_2 emissions minimization for 2009. The number of sectors refers to the disaggregated sectors provided in the Appendix.

Pareto Points	1	2	3	4	5	6	7	8	9	10
CO_2 emissions reduction (%)	0.0	0.3	0.5	0.8	1.1	1.4	1.6	1.9	2.2	2.4
Demand satisfaction (%)	100	99.9	99.9	99.6	99.0	98.1	96.8	95.4	93.4	90.0
Ratio	-	3.9	3.9	1.8	1.1	0.7	0.5	0.4	0.3	0.2
Cut sectors	0	14	14	261	449	734	885	885	1075	1435

In the maximum demand solution, all of the sectors fulfill the maximum demand. The minimum impact solution (i.e., solution 10) shows the lowest ratio (0.2), but allows for the largest reduction in CO_2 emissions (2.4%) at the expense of reducing the demand by 10%, and cutting 1,435 sectors. In contrast, the intermediate Pareto point 6 shows a ratio close to 0.7 with a reduction of 1.35% in CO_2 emissions and a demand satisfaction of 98.1%.

Fig. 9 shows the reduction in production-based CO_2 emissions of each country compared to the base case (current situation) in the minimum impact solution, in an intermediate solution (i.e., solution 6) and in the solution with the highest improvement ratio (i.e., solution 2).

As seen, the largest reduction in emissions occurs in United States, followed by Canada and Mexico. These last two countries exchange a large amount of goods/services with US via trade, and for this reason their CO_2 emissions are affected significantly by changes in the US economy.

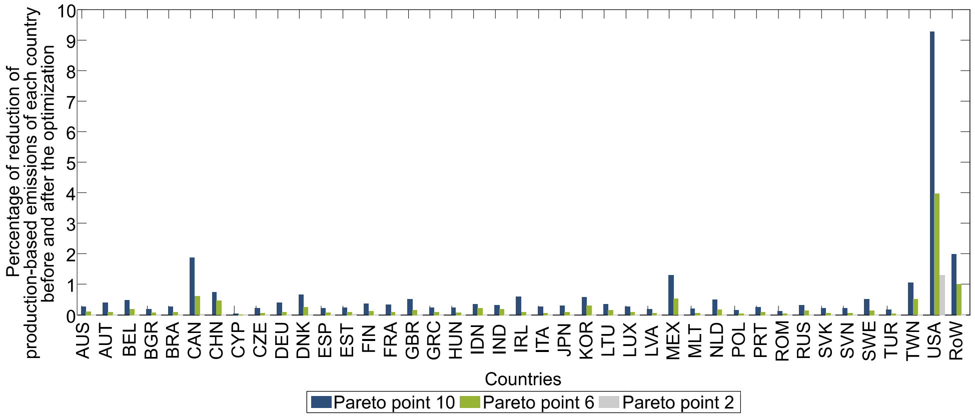


Fig. 9. Total percentage reduction of production-based emissions before and after the optimization. Each bar represents a different Pareto point: the minimum impact solution (blue bar), an intermediate Pareto point (green bar) and the maximum ratio solution (grey bar) (solutions 10, 6 and 2 of Table 4, respectively).

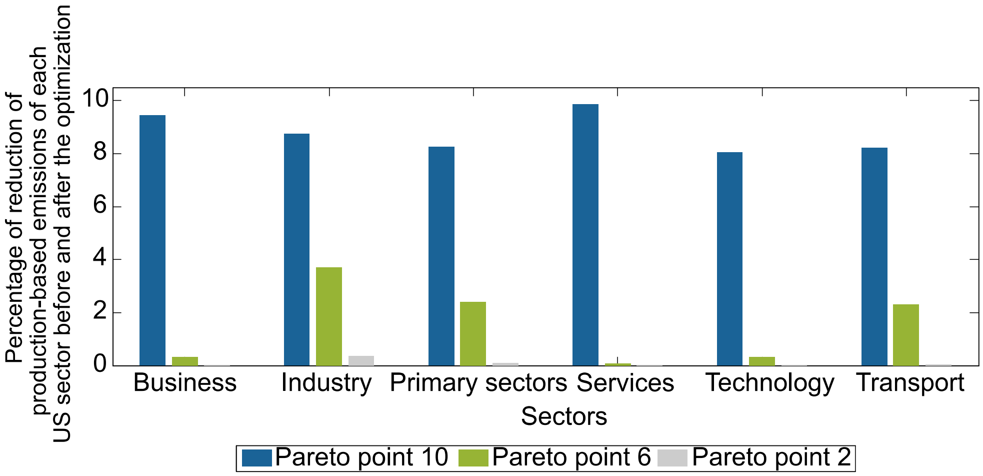


Fig. 10. Total percentage reduction of production-based emissions of US sectors before and after the optimization. Each bar represents one Pareto point: the minimum impact solution (blue bar), an intermediate Pareto point

(green bar) and the maximum ratio solution (grey bar) (solutions 10, 6 and 2 of Table 4, respectively).

Fig. 10 shows how the US sectors reduce their emissions during the optimization (see Fig. A.6. in Supplementary material for the disaggregated results). As observed, as we move from the maximum impact solution to the minimum impact one (Pareto point 2), the first sector that is cut is industry (0.36%), which shows a high ratio CO_2 emissions/demand satisfaction (see Eq.14). An increasing number of sectors are then gradually cut until the minimum impact solution is reached, in which the emissions reductions in all sectors are above 8%. A more disaggregated analysis shows that the first sector affected by the optimization is electricity, gas and water supply (2.6%). In addition, the emissions associated with chemical engineering activities are reduced by 8.2% in the minimum impact solution.

Finally, Fig. 11 is similar to Fig. 10, but shows the changes in emissions of the sectors at a global scale rather than the changes taking place only in US.

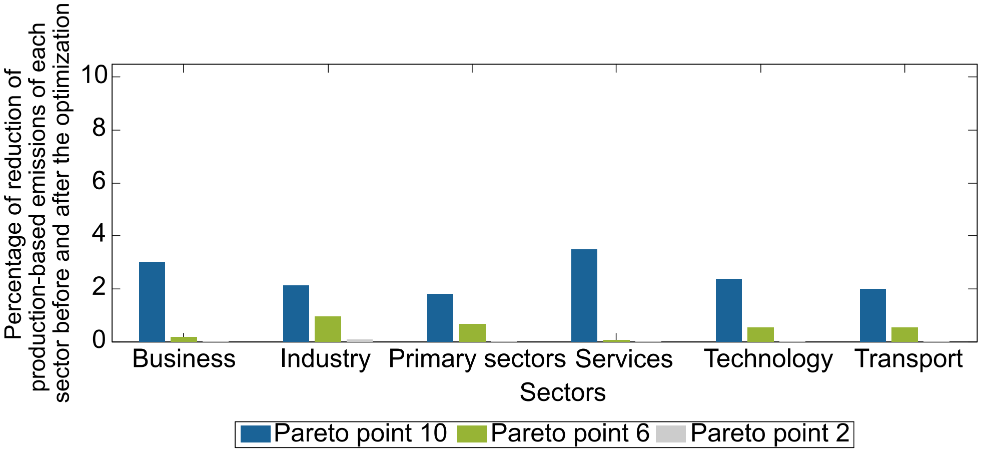


Fig. 11. Total percentage reduction of production-based emissions of global sectors before and after the optimization. Each bar represents one Pareto point: the minimum impact solution (blue bar), an intermediate Pareto point

(green bar) and the maximum ratio solution (grey bar) (solutions 10, 6 and 2 of Table 4, respectively).

As seen in Fig. 11, the model regulates first those sectors with a high ratio impact/total output, with the sector industry being the first to be modified. The analysis of the minimum impact solution shows also that the most affected sector is services (3.5%) followed closely by the business sector (3.0%) (see Fig. A.7. of the Supplementary material for the disaggregated results).

5.4.4 Impact of Shale Gas

The interest in shale gas as an available source of natural gas has grown rapidly in the US, where it has become one of the major sources of energy. This trend in the US is motivated by different factors, including the existence of large reserves and the fact that it is cleaner than standard fossil fuels in terms of contribution to global warming (see Table 5)[23].

Table 5. Pollution intensity of electricity technologies in US[24].

Energy Source	Pollution intensity (kgCO_2/kWh)
Coal	1.001
Petroleum	0.840
Shale Gas	0.479
Natural Gas	0.469
Geothermal	0.045
Solar	0.042
Nuclear	0.016
Wind	0.012
Hydroelectric	0.004

Bearing this in mind, this section aims to analyze the effect that increasing the share of shale gas in the electricity grid of US will have on its overall environmental performance. Specifically, this section analyzes several plausible scenarios, each entailing a different replacement ratio of coal by shale gas (i.e., percentages of replacement of coal by shale gas: 15% scenario Shale +, 25% scenario Shale ++, and 50% scenario Shale +++).

To model these scenarios, we proceeded as follows. The pollution intensity parameter of the US sector *Electricity, gas and water supply* (subsector S17 belonging to the sector industry, as shown in Table A.1. of disaggregated sectors provided as supplementary material) was modified, keeping the remaining parameters constant. The amount of energy required per unit of money traded (denoted by parameter $energy(s17, US)$) was first obtained as follows:

$$energy(s17, US) = \frac{e(s17, US)}{\sum_n PI(n) \cdot w(n)} \tag{15}$$

where *PI(n)* is the pollution intensity of technology *n* (i.e., CO_2 emissions per kWh), *w(n)* is the share of technology *n* in the electricity grid of US (that falls in the interval 0-1) and *e(S17,US)* is the pollution intensity factor of the sector *Electricity, gas and water supply* (S17) of US, expressed in $kgCO_2/\$$.

After determining the amount of energy required per monetary unit traded in sector S17, we next modified the share of coal and shale gas (*w(coal)* and *w(shale gas)*) according to the forecasted scenarios displayed in Table 6. The modified impact per monetary unit traded in sector S17 was then obtained as follows:

$$e'(s17, US) = energy(s17, US) \sum_n PI(n) \cdot w'(n) \tag{16}$$

Table 6. Electricity grid of US for the base case, scenario Shale+, scenario Shale++ and scenario Shale+++. The pollution intensity of sector 17 for every scenario is shown in the last row of the table.

Energy Source	Base case % of use[23]	Shale+ % of use	Shale++ % of use	Shale+++ % of use
Coal	44.5	37.8	33.4	22.3
Geothermal	0.4	0.4	0.4	0.4
Hydroelectric	7.0	7.0	7.0	7.0
Natural Gas	23.6	23.6	23.6	23.6
Nuclear	20.2	20.2	20.2	20.2
Petroleum	1.3	1.3	1.3	1.3
Shale Gas	0.0	6.7	11.1	22.3
Solar	0.5	0.5	0.5	0.5
Wind	2.0	2.0	2.0	2.0
e(s17,US) (kgCO$_2$/\$)	5.25	4.93	4.71	4.18

The LP was then solved again for the new modified environmental coefficients of sector 17 (Eq. 9).

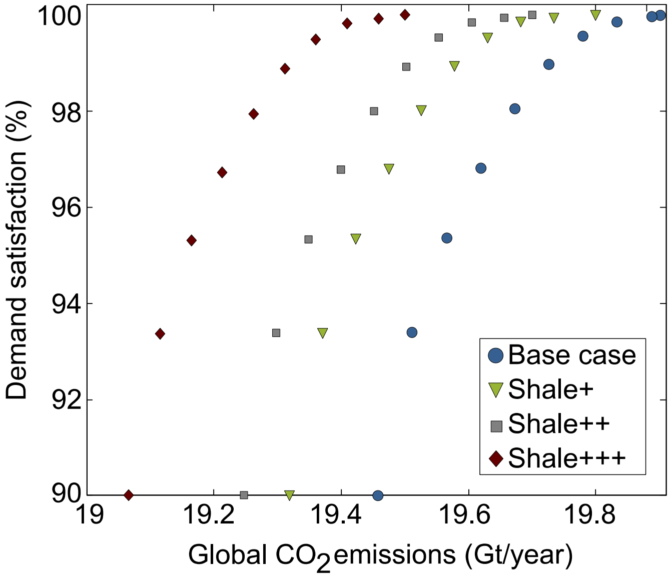


Fig. 12A. Pareto optimal frontier for global CO$_2$ production-based emissions (Gt/year) vs US demand satisfaction (%) in 2009 for the base case, scenario Shale+ (15% of coal replaced by shale gas), scenario Shale++ (25% of coal replaced by shale gas) and scenario Shale+++ (50% of coal replaced by shale gas).

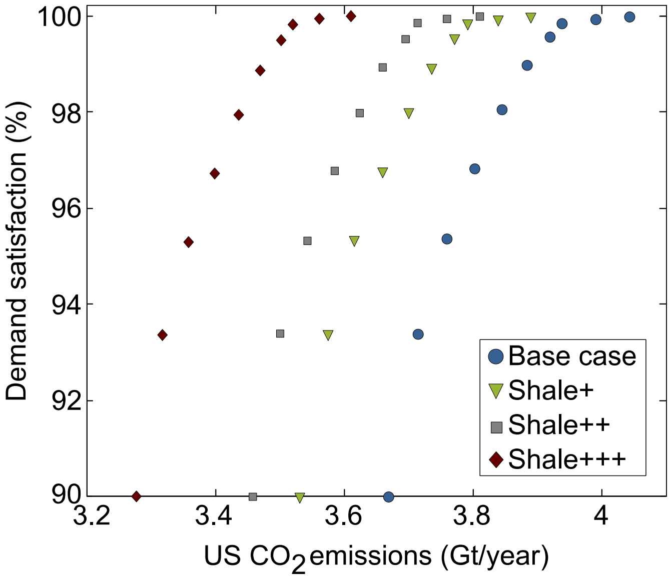


Fig. 12B. Pareto optimal frontier for production-based CO_2 emissions in US (Gt/year) vs US demand satisfaction (%) in 2009 for the base case, scenario Shale+ (15% of coal replaced by shale gas), scenario Shale++ (25% of coal replaced by shale gas) and scenario Shale+++ (50% of coal replaced by shale gas).

Fig.12A shows the 10 Pareto points (CO_2 emissions worldwide vs demand satisfaction) for the base case, scenario Shale+ (15% of coal replaced by shale gas), scenario Shale++ (25% of coal replaced by shale gas) and scenario Shale+++ (50% of coal replaced by shale gas). These points were solved following the same procedure as before, that is, maximizing the demand satisfaction for different targets on the emissions. Fig. 12B is equivalent to Fig.12A, but it shows the US production-based emissions instead of the world production-based emissions. Note that the points have been projected here onto the subspace "US emissions vs demand satisfaction", despite the fact that they were generated in the subspace "Global emissions vs demand satisfaction".

The analysis of the extreme scenario Shale+++ (50% of coal replaced by shale gas) shows that US CO_2 production-based emissions can drop by more than 10% compared to the base case, while the world emissions can drop by up to 2% in all the Pareto points (the Pareto frontier shifts to the left).

An in-depth analysis of the Pareto frontier shows that the most affected countries and sectors are the same that in the base case (Figs. 9-11). However, when the shale gas is included in the electricity grid, the CO_2 emissions reductions are significantly larger.

5.5 Conclusions

This work has presented an approach for minimizing the CO_2 emissions at a macroeconomic level by modifying the sectors of an economy. Our approach combines multi-objective optimization and multi-regional input-output models within a single unified framework that allows identifying key economic sectors whose regulation leads to larger reductions in impact at a minimum change in demand satisfaction. The tool introduced was applied to the US economy in order to identify the best policies to be implemented in practice for mitigating global warming.

A preliminary analysis of the IO data reveals that consumption-based US emissions are higher than production-based, evidencing that part of its impact is currently being masked by displacing the manufacturing tasks to other countries. This happens as well on a sectorial basis, where the life cycle emissions of several sectors exceed their emissions taking place within the limits of US. More than half of the production-based emissions belong to the sector industry, while sectors related to chemical engineering activities represent 9% of the total emissions (i.e., sectors *Coke, Refined Petroleum and Nuclear Fuel, Chemicals and Chemical Products and Rubber and*

Plastics shown in the supplementary material). Most of these emissions, however, are ultimately associated with sectors that differ from the one that releases them (i.e., the emissions are originated in one sector, but are required to cover the demand of a different sector). As for the spatial distribution of emissions, we found that the trade of emissions is larger with China, Canada, Russia, Japan, Mexico and Great Britain.

The optimization algorithm identified the sectors that should be regulated in order to attain a given environmental target while maximizing the demand satisfaction. The global sectors that would be more affected by a potential environmental regulation of the US economy would be services and business, with a reduction of 3.5% and 3.0%, respectively, in the minimum impact solution. These changes in the economy would also have a significant impact on Mexico and Canada, countries with which US maintains a more intense commercial activity.

Finally, replacing fossil fuels by shale gas can lead to reductions of up to 2% in global CO_2 emissions and up to 10% in production-based US CO_2 emissions.

Our analysis provides valuable insight for decision makers during the development of more effective environmental regulations. This approach can be easily extended to deal with other economic regions and environmental impacts, and opens new avenues for the application of process systems engineering tools in macroeconomic problems.

5.6 Acknowledgements

The authors wish to acknowledge the financial support received from the University Rovira i Virgili and from the Spanish Ministry of Education and Science (CTQ2012-37039 and DPI2012-37154).

5.7 Nomenclature

Acronyms

EEIO	Environmentally extended input-output
EEMRIO	Environmentally extended multi-regional input-output
EU	European Union
GHG	Greenhouse Gas Emissions
IO	Input-output
LP	Linear programing
Shale+	Case study 1: 15% of coal replaced by shale gas
Shale++	Case study 2: 25% of coal replaced by shale gas
Shale+++	Case study 3: 50% of coal replaced by shale gas
US	United States
WIOD	World Input-Output Database

Index

i	Economic sector
j	Economic sector
n	Energy technology
r	Region
r'	Region

Parameters

a(i,j)	Amount (in US$) of output of sector *i* necessary to produce one dollar of output of sector *j*
a(i,j,r,r')	Amount (in US$) of output of sector *i* of region *r* necessary to produce one dollar of output of sector *j* of region *r'*

$e(i)$	Environmental pollution intensity for sector *i* (i.e., impact per monetary unit traded)
$e(i,r)$	Environmental pollution intensity for sector *i* of sector *r* (i.e., impact per monetary unit traded)
$energy(s17,US)$	Amount of energy required per unit of money traded
$Imp^C(i,r)$	Consumption-based CO_2 emissions
$Imp^P(i,r)$	Production-based CO_2 emissions
$PI(n)$	Pollution intensity of technology *n*
$w(n)$	Share of energy technology n in the electricity grid of US
$X^C(i,r)$	Economic transactions required to fulfill the demand of sector *i* of region *r*
$X^P(i,r)$	Sales of sector i of region r
$x(i,j)$	Output of sector *i* acting like an input for sector *j*
$\bar{x}(i,j)$	Current output of sector *i* acting like an input for sector *j*
$x(i,j,r,r')$	Output of sector *i* of region *r* acting like an input for sector *j* of region *r'*
$\bar{x}(i,j,r,r')$	Current output of sector *i* of region *r* acting like an input for sector *j* of region *r'*
$\bar{X}(j)$	Current total output in currency units (e.g. US$) of sector *j*
$\bar{X}(j,r)$	Current total output in currency units (e.g. US$) of sector *j* in region *r'*

Variables

$DSat$	Demand satisfaction
$Imp(i)$	Environmental impact (i.e., global warming potential) associated with sector *i*
$Imp(i,r)$	Environmental impact (i.e., global warming potential) produced by sector *i* of region *r*

RATIO	Ratio between the optimal reductions in CO_2 emissions for a given change in demand satisfaction for every point of the Pareto frontier
Timp	Total environmental impact generated by all of the sectors of the economy
X(i)	Total output in currency units (e.g. US$) of sector *i*
X(i,r)	Total output in currency units (e.g. US$) of sector *i* in region *r*
X(j)	Total output in currency units (e.g. US$) of sector *j*
X(j,r')	Total output in currency units (e.g. US$) of sector *j* in region *r'*
y(i)	Final demand (end user) of the sector *i*
y(i,r)	Final demand (end user) of the sector *i* of region *r*
$y_0(i,r)$	Current final demand (end user) of the sector i of region *r*

5.8 References

1. Davis SJ, Caldeira K. Consumption-based accounting of CO_2 emissions. Proc Natl Acad Sci U S A. 2010;107:5687-5692.
2. Davis SJ, Peters GP, Caldeira K. The supply chain of CO_2 emissions. Proc Natl Acad Sci U S A. 2011; 108:18554-9.
3. Hertwich EG, Peters GP. Carbon footprint of nations: A global, trade-linked analysis. Environ Sci Technol. 2009;43:6414-6420.
4. Peters GP. From production-based to consumption-based national emission inventories. Ecol Econ. 2008;65:13-23.
5. Peters GP, Hertwich EG. CO_2 embodied in international trade with implications for global climate policy. Environ Sci Technol. 2008;42:1401-1407.

6. Wiebe KS, Bruckner M, Giljum S, Lutz C, Polzin C. Carbon and materials embodied in the international trade of emerging economies: A multiregional input-output assessment of trends between 1995 and 2005. J Ind Ecol. 2012;16:636-646.
7. Leontief W. Quantitative Input and Output Relations in the Economic Systems of the United States. Rev Econ Stat. 1936;18:105-125.
8. Wiedmann T, Lenzen M, Turner K, Barrett J. Examining the global environmental impact of regional consumption activities - Part 2: Review of input-output models for the assessment of environmental impacts embodied in trade. Ecol Econ. 2007;61:15-26.
9. Leontief W. Environmental Repercussions and the Economic Structure: An Input-Output. Rev Econ Stat. 1970; 52:262-271.
10. Tukker A, Bulavskaya T, Giljum S, et al. The Global Resource Footprint of nations. 2013.
11. Watson D, Moll S. Environmental benefits and disadvantages of economic specialisation within global markets and implications for SCP monitoring. In: Paper for the SCORE! Conference. Brussels, Belgium; 2008.
12. Wiedmann T. A review of recent multi-region input-output models used for consumption-based emission and resource accounting. Ecol Econ. 2009;69:211-222.
13. Cho C-J. The economic-energy-environmental policy problem: An application of the interactive multiobjective decision method for Chungbuk Province. J Environ Manage. 1999;56:119-131.
14. Hsu GJY, Chou FY. Integrated planning for mitigating CO2 emissions in Taiwan: A multi-objective programming approach. Energy Policy. 2000;28:519-523.

15. Oliveira C, Antunes CH. A multiple objective model to deal with economy-energy-environment interactions. In: European Journal of Operational Research.Vol 153.; 2004:370-385.

16. San Cristóbal JR. An environmental/input–output linear programming model to reach the targets for greenhouse gas emissions set by the kyoto protocol. Econ Syst Res. 2010;22(3):223-236.

17. Hristu-Varsakelis D, Karagianni S, Pempetzoglou M, Sfetsos A. Optimizing production with energy and GHG emission constraints in Greece: An input-Output analysis. Energy Policy. 2010;38:1566-1577.

18. Lin C. Identifying lowest-emission choices and environmental Pareto frontiers for wastewater treatment wastewater treatment input-output model based linear programming. J Ind Ecol. 2011;15:367-380.

19. Timmer M, Erumban AA, Gouma R, et al. The World Input-Output Database (WIOD): Contents, Sources and Methods, version 0.9.; 2012.

20. Miller RE, Blair PD. Input-output analysis: foundations and extensions. Prentice-Hall.; 1985.

21. Bérubé J-F, Gendreau M, Potvin J-Y. An exact -constraint method for bi-objective combinatorial optimization problems: Application to the Traveling Salesman Problem with Profits. Eur J Oper Res. 2009;194:39-50.

22. O'Mahony M, Timmer MP. Output, input and productivity measures at the industry level: The EU KLEMS database. Econ J. 2009;119.

23. EIA. Electric Power Annual 2009. Energy. 2011;0348:April. 2010.

24. Edenhofer O, Pichs Madruga R, Sokona Y, Report S, Panel I, Change C. Renewable Energy Sources and Climate Change Mitigation Special Report of the Intergovernmental Panel on Climate Change.; 2012.

5.9 Appendix

Table A.1. List of manufacturing sectors that appear in WIOD-database

S1	Agriculture, Hunting, Forestry and Fishing
S2	Mining and Quarrying
S3	Food, Beverages and Tobacco
S4	Textiles and Textile Products
S5	Leather, Leather and Footwear
S6	Wood and Products of Wood and Cork
S7	Pulp, Paper, Paper , Printing and Publishing
S8	Coke, Refined Petroleum and Nuclear Fuel
S9	Chemicals and Chemical Products
S10	Rubber and Plastics
S11	Other Non-Metallic Mineral
S12	Basic Metals and Fabricated Metal
S13	Machinery, Nec
S14	Electrical and Optical Equipment
S15	Transport Equipment
S16	Manufacturing, Nec; Recycling
S17	Electricity, Gas and Water Supply
S18	Construction
S19	Sale, Maintenance and Repair of Motor Vehicles Retail Sale of Fuel
S20	Wholesale Trade and Commission Trade, Except of Motor Vehicles
S21	Retail Trade, Except of Motor Vehicles ; Repair of Household Goods
S22	Hotels and Restaurants
S23	Inland Transport
S24	Water Transport
S25	Air Transport
S26	Other Supporting and Auxiliary Transport Activities; Activities of Travel Agencies
S27	Post and Telecommunications
S28	Financial Intermediation
S29	Real Estate Activities
S30	Renting of M&Eq and Other Business Activities
S31	Public Admin and Defence; Compulsory Social Security
S32	Education
S33	Health and Social Work
S34	Other Community, Social and Personal Services
S35	Private Households with Employed Persons

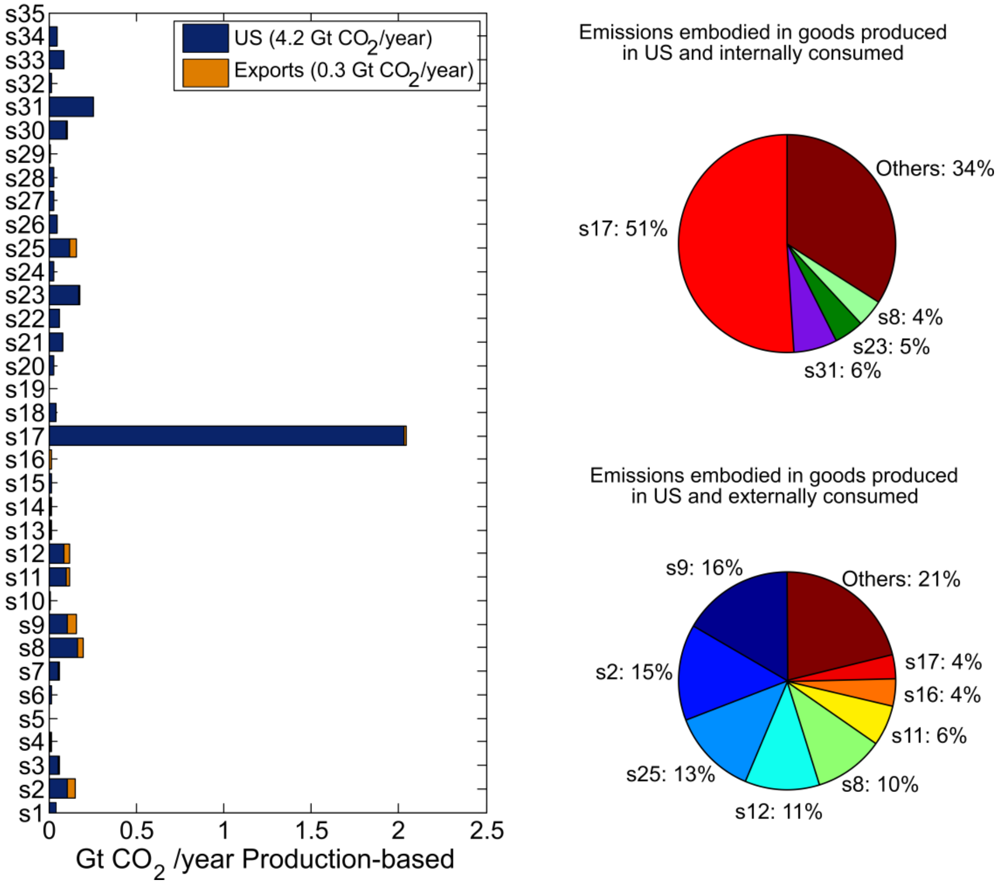


Fig.A.1 Blue bars represent the breakdown of total production-based CO_2 emissions generated within the limits of US (total emissions equal 4.2 Gt CO_2/year). Orange bars are the breakdown of CO_2 emissions exported via trade (total exported emissions equal 0.3 Gt CO_2/year).

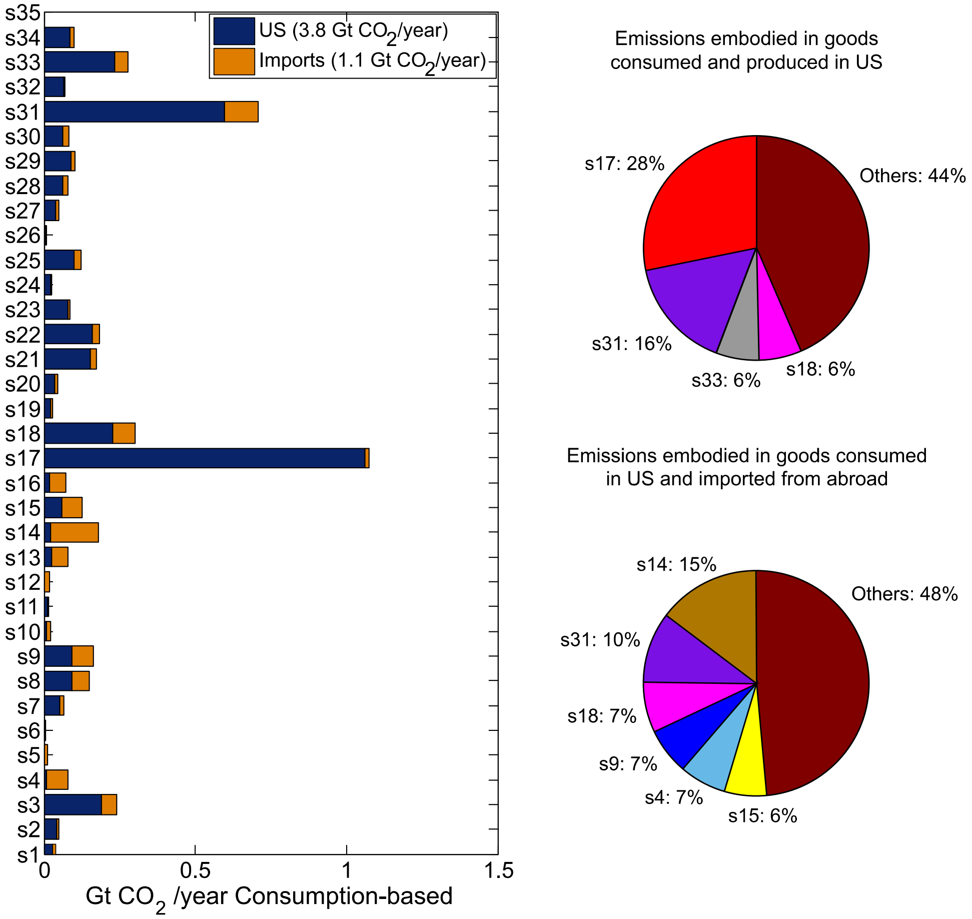


Fig. A.2 Blue bars represent the breakdown of total consumption-based CO_2 emissions generated to satisfy the demand of each US sector (total emissions equal 3.8 Gt CO_2/year). Orange bars are the sectorial breakdown of CO_2 emissions imported via trade (total imported emissions equal 1.1 Gt CO_2/year.

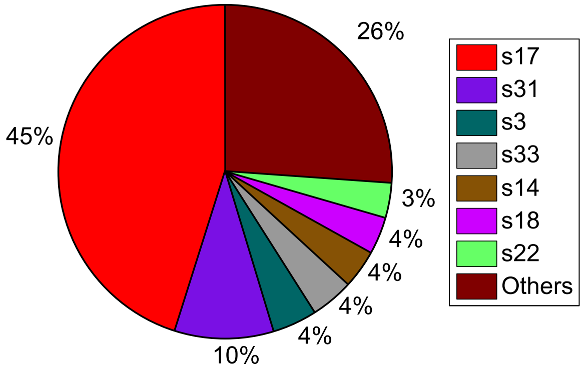


Fig. A.3. Breakdown of the emissions of Electricity, Gas and Water Supply in 2009 according to the final demand of the sectors. Each portion represents the percentage of production-based CO_2 emissions generated by the US sector *Electricity, Gas and Water Supply* (S17) that are attributed to the intermediate demand of each US sector

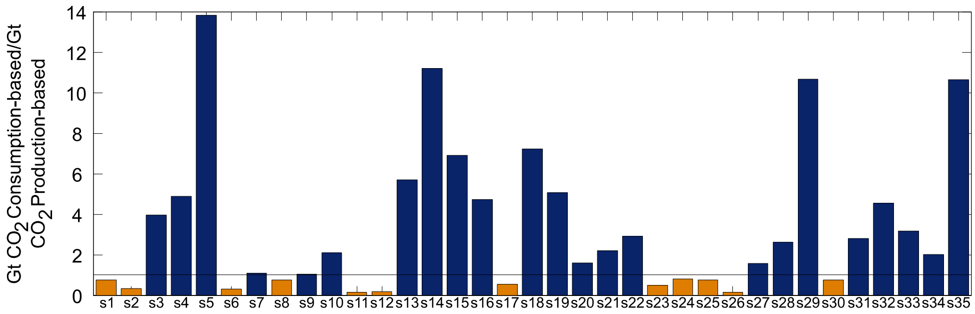


Fig. A.4. Comparison between the consumption (blue bars) and production-based (orange bars) accounting approaches in 2009. Each bar represents one industrial sector.

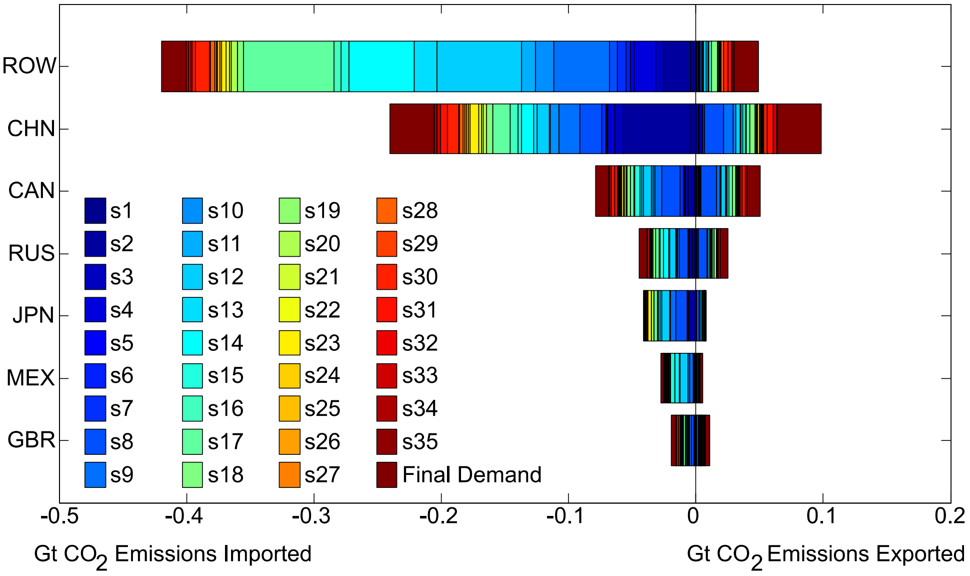


Fig. A.5. Countries with higher trade of CO_2 with US in 2009. ROW = Rest of World; CHN = China; CAN = Canada; RUS = Russia; JPN = Japan; MEX = Mexico; GBR = United Kingdom

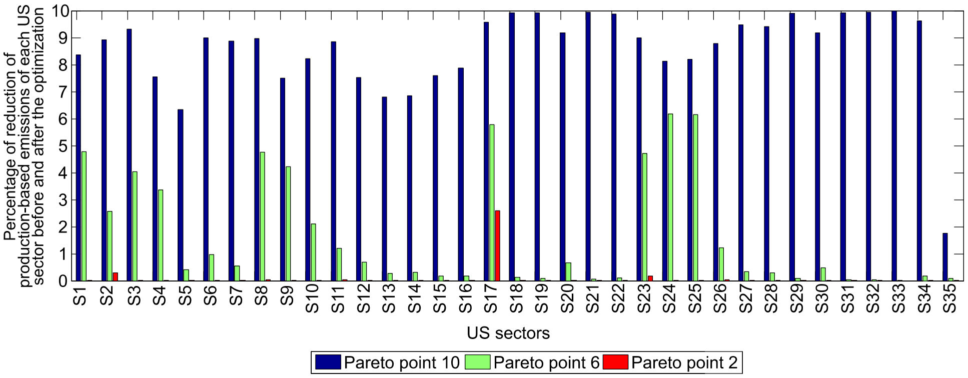


Fig. A.6 Total percentage reduction of production-based emissions of US sectors before and after the optimization. Each bar represents one Pareto point: the minimum impact solution (blue bar), an intermediate Pareto point (green bar) and the maximum ratio solution (red bar) (solutions 10, 6 and 2 of Table 4, respectively).

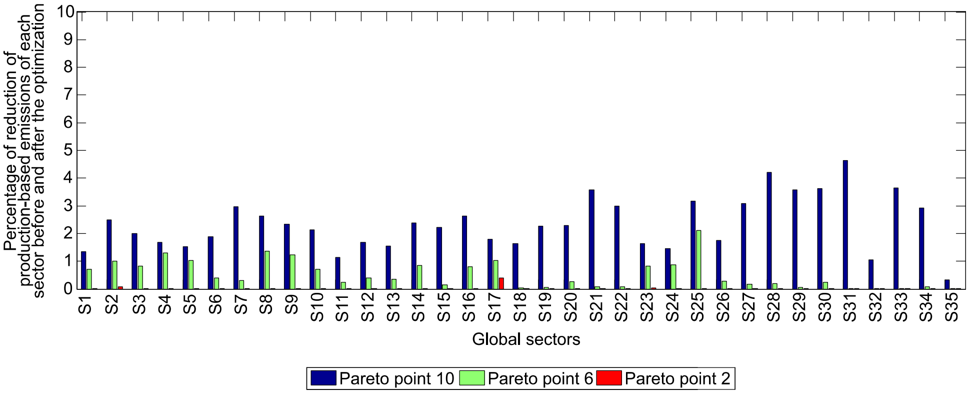


Fig. A.7 Total percentage reduction of production-based emissions of global sectors before and after the optimization. Each bar represents one Pareto point: the minimum impact solution (blue bar), an intermediate Pareto point (green bar) and the maximum ratio solution (red bar) (solutions 10, 6 and 2 of Table 4, respectively).

6 APPENDICES

6.1 Publications

6.1.1 Research articles

Pascual-González, J., Guillén-Gosálbez, G., Mateo-Sanz, J. M., Jiménez-Esteller, L. Statistical analysis of global environmental impact patterns using a world multi-regional input–output database. *Journal of Cleaner Production* 2015, 90, 360-369.

Pascual-González, J., Guillén-Gosálbez, G., Mateo-Sanz, J. M., Jiménez-Esteller, L. Statistical analysis of the EcoInvent database to uncover relationships between life cycle impact assessment metrics. *Journal of Cleaner Production* 2015, doi:10.1016/j.jclepro.2015.05.129.

Pascual-González, J., Pozo, C., Guillén-Gosálbez, G., Jiménez-Esteller, L. Combined use of MILP and multi-linear regression to simplify LCA studies. *Computers and Chemical Engineering* 2015, 82, 34-43.

Pascual-González, J., Guillén-Gosálbez, G., Jiménez-Esteller, L., Grossmann, I., Siirola, J. Macroeconomic minimization of the global warming potential via environmentally extended multi-regional input-output models: Application to the US economy. Ready to be submitted to *AIChE Journal*.

6.1.2 Book chapters

Pascual-González, J., Guillén-Gosálbez, G., Jiménez, L. Multi-objective optimization of US economy via multi-regional input-output analysis. 24th European symposium on C*omputer Aided Process Engineering.* 2014, 33, 1015-1020. Elsevier, B.V. ISBN: 978-0-444-63434-4.

6.2 Scientific conference participations

6.2.1 Oral communications:

Pascual-González, J., Guillén-Gosálbez, G., Jimenez, L., Cortés-Borda, D. Multi-objective optimization of international economies via multi-regional input-output analysis: Application to the US economy. *American Institute of Chemical Engineers (AIChE) Annual Meeting.* November 2013. San Francisco, USA.

Pascual-González, J., Guillén-Gosálbez, G., Jiménez, L. Multi-objective optimization of US economy via multi-regional input-output analysis. *24th European symposium on computer aided process engineering (ESCAPE24).* July 2014. Budapest, Hungary.

Pascual-González, J., Guillén-Gosálbez, G., Mateo-Sanz, J. M., Jiménez-Esteller, L. Statistical analysis of global environmental impact patterns using a world multi-regional input–output database. *American Institute of Chemical Engineers (AIChE) Annual Meeting.* November 2014. Atlanta, USA.

6.2.2 Poster presentations:

Pascual-González, J., Guillén-Gosálbez, G., Jimenez, L. Multi-objective optimization of international economies via multi-regional input-output analysis: Application to the US economy. *13th Mediterranean Congress of Chemical Engineering (13MCCE).* October 2014 Barcelona, Spain.

environmental
economy
procedure
engineering
sustainable
method
objective
impact
policies
methodology
assess
energy
production
multi-regional
optimal
similar
models
impacts
thesis
LCA
emiss
countr
valuable
reduced
problems
extended
changes
cross
developing
optimizat
IO
multi-objective
linear
correlation
single
demand
key
redundant
combined
model
Pareto
performanc