EnviroInfo 2009 (Berlin)
Environmental Informatics and Industrial Environmental Protection: Concepts, Methods and Tools
Copyright © Shaker Verlag 2009. ISBN: 978-3-8322-8397-1

A Data Warehousing and Data Mining Tool for Environmental Accounting

Ting Yu, Manfred Lenzen, Chris Dey and Jeremy Badcock
Centre of Integrated Sustainability Analysis, University of Sydney, NSW 2006, Australia
t.yu@physics.usyd.edu.au

Abstract

This paper presents an integrated intelligent system being capable of automatically estimating and updating large-size input-output tables for the environmental accounting. This system is a comprehensive solution covering from the data collection and integration, querying, analyzing and presentation. The environmental accounting addresses how to correctly measure the greenhouse gas emission of an organization. Among the various environmental accounting methods, the Economic Input-Output Life Cycle Assessment (EIO-LCA) method uses information about industry transactions-purchases of materials by one industry from other industries, and the information about direct environmental emissions of industries, to estimate the total emissions throughout the whole supply chain. The core engine of the EIO-LCA is the input-output table. This intelligent system is able to interpret and follow users' XML-based query scripts, retrieve data from various sources and integrate for the following data mining components. The data mining component bases on a unique mining algorithm which estimates the input-output tables from the historical data and the spatial data simultaneously. This mining algorithm runs over the parallel computer to enable the system to estimate a matrix of the size up to 3000-by-3000.

Keywords: *Input-output Table, Data Mining, Data Warehousing*

1. Introduction

Instead of a particular protection technique such as solar energy, the environmental accounting brings environmental costs to the attention of corporate stakeholders who may be able and motivated to identify ways of reducing or avoiding those costs while at the same time improving environmental quality. In order to report the environmental cost of the activity of an organization, the environmental accounting requires proper methodologies to correctly measure the environmental impact such as the greenhouse gas emission (Peters, Sack et al. 2007). There are several accounting approaches of measuring the emission, such as auditing and triple bottom line methods (TBL) (Foran, Lenzen et al. 2005). The triple bottom line method captures an expanded spectrum of values and criteria for measuring organizational (and societal) success: economic, ecological and social. With the ratification of the United Nations TBL standard for urban and community accounting in early 2007, this became the dominant approach to public sector full cost accounting (Brown, Dillard et al. 2006). The traditional accounting method only measures the success regarding the economic, but ignore other two. A life cycle assessment (LCA) is the investigation and valuation of the environmental impacts of a given product or service caused or necessitated by its existence, and an evaluation of the environmental impacts of a product or process over its entire life cycle. Environmental life cycle assessment is often thought of as "cradle to grave" and therefore as the most complete accounting of the environmental costs and benefits of a product or service (Hendrickson, Lave et al. 2006). Among the various LCA methods, the Economic Input-Output Life Cycle Assessment (EIO-LCA) method uses information about industry transactions - purchases of materials by one industry from other industries, and the information about direct environmental emissions of industries, to estimate the total emissions throughout the supply chain (Hendrickson, Lave et al. 2006). In the EIO-LCA method, the input-output table is the key engine. The input-output table simply uses a matrix representing the intra-industry flows and the flow between industrial sections and consumption or the flow between the value-added section and the industrial section. Because the economic constantly evolves, the input-output table needs to be updated at least annually to reflect the new circumstance. Unfortunately, in most countries such as Australia, the input-output table is released every 3-4 years by governments, because of the large amount of monetary and human cost involved. The Centre of Integrated Sustainability Analysis (ISA) in the University of Sydney has developed an integrated intelligent approach to estimate and update the input-output tables for different level at a regular basis.

While the past decades have seen the booming supply of data from various sources, a large amount of the data regarding the environment and economic can be easily accessed. But the data from various sources has various structures and different ways of represent their underlying meaning. In order to query the available data, it is necessary to integrate the various types of data first. In many cases, this kind of integration and query operation becomes a daily routing task in order to keep the information up to date. Moreover, estimating the input-output table involves both the temporal and spatial data. The data mining algorithm estimating the input-output table must have capacity of reconciliating and utilizing both types of information. The proposal approach automates the whole process and maximally reduces human's involvement.

2. System Architecture

A typical input-output table can be represented as following:

			USA (2)			
			Shoe (1)	Car (2)	Retail (3)	Oil Refinary (4)
Australia (1)	NSW (1)	Sheep (1)	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$
		Oil (2)	$x_{2,1}$	…	…	$x_{2,4}$
	VIC (2)	Sheep (1)	$x_{3,1} = 0.23$	…	…	$x_{3,4}$
		Oil (2)	$x_{4,1}$	…	…	$x_{4,4}$

Table 1. Example of an Input-output Table

The entry represents the commodity flow between the industry sections of different regions. For example, $x_{3,1}$ represents the commodity from the sheep industry at the Victoria (VIC) in Australia to the shoe industry in USA. The purpose of the whole approach is to estimate and update this table.

The whole system consists of a series of functional components: data integration, query, data mining and model presentation (See Figure 1).

As the first step, the data retrieval component acts as interfaces to various types of datasets including the greenhouse gas emission measurement, macro and micro economic data that are stored in various formats such as Excel files, databases etc. The data integration component unifies these heterogeneous datasets to a single format, integrates and restructures the data retrieved by the previous component. At the same time, users' specification of the concept hierarchy is interpreted and translated into hierarchies defining the structure of the input-output table. This concept hierarchy is very similar to the data warehousing (Hobbs, Hillson et al. 2005). This hierarchy allows users to roll up and drill down the data very easily and also introduce the dynamic to the input-output table. The data mining component employs a unique algorithm designed to estimate the input-output table. In order to process extremely large amount of data, this component sits on a parallel optimization algorithm to quickly converge to the optimal estimation of the unknown input-output table. As the final step, the model checking and presentation components check the quality of the estimated model and present it to users in a structured format.

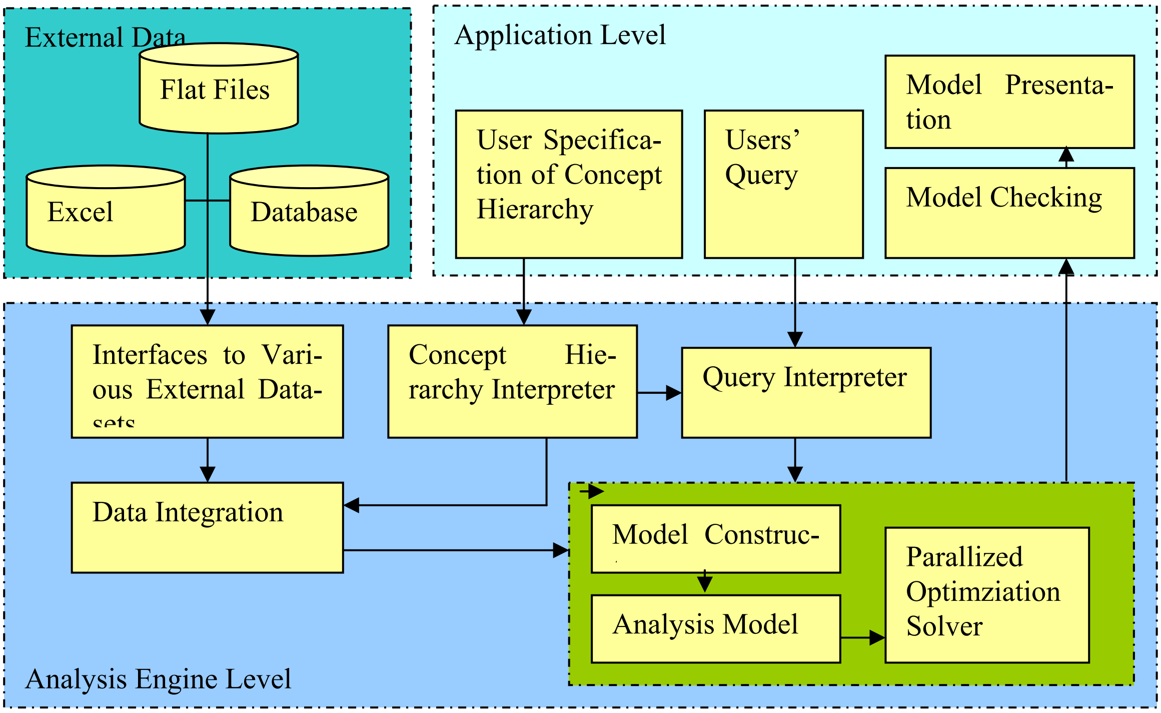

Figure 1. System Architecture

2.1 Model Constructor

The model constructor component communicates with other two components: the concept hierarchy interpreter and query interpreter. The concept hierarchy interpreter constructs the tree-like hierarchy that we will discuss in detail later, and the query interpreter translates the users' query written in a special meta language. The model constructor then 1) require the data integration component to retrieval data from various sources and integrate them, and 2) restructure and assign the meaning to the data according to the previous concept hierarchy and users' query in order to populate the data mining model.

On the process of building a model, the first step is to construct the concept hierarchy. The hierarchy is pre-required for restructuring data from various sources. The hierarchy structure is introduced by a multi-tree structure. For example, the hierarchy representing Australia and USA national economic can be like fig 2.

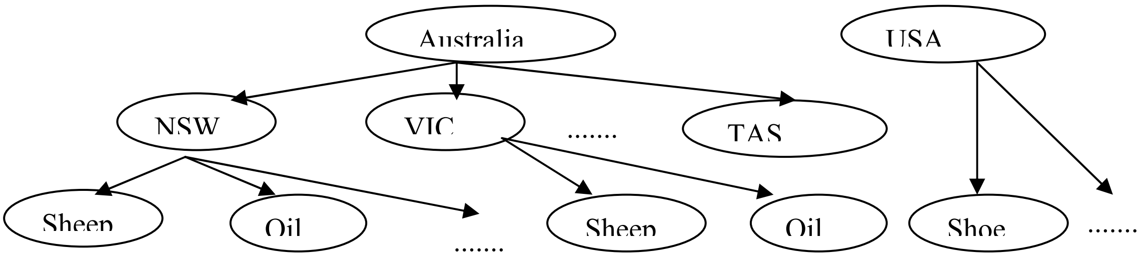

Figure 2: An Example of Concept Hierarchy

The hierarchy brings two major benefits. First, it provides the different levels of abstraction. The flexibility of the concept hierarchy makes users to have snapshots of the table at various levels of abstraction. Users can easily drill down or roll up the table without redefining the table. For example, there is big difference between the structures of input-output tables at the national level and at the corporate level, because the operations within a corporate are much simpler than those of a nation in the most cases.

Secondly, it allows the dynamic structure of the resultant input-output table. Regarding the difference between applications, a dynamical hierarchy provides the flexibility to expand this system to different application with different structures of input-output tables. It is very common to mix two different concept hierarchies for an input-output table. For example, in Table 1, the input-output table is organized by one three-level hierarchy and one two-level hierarchy. The structure of the input-output table in Table 1 is defined by these two concept hierarchies in Figure 2. The row side is defined by the Austrralia concept hierarchy and the column side is defined by the USA concept hierarchy.

Considering the complexity caused by introducing the concept hierarchy, a query language is introduced to provide users' an easy but powerful way to retrieval and organize their data. The query language must be compact and accurate to make the description to be readable and expressive. It is unrealistic to write hundred thousands of codes to describe a single model at a daily base. The query language we create is based on the coordinates of the valuable within the input-output table whose structure is defined by the concept hierarchies. For example, the export from the sheep industry in Victoria, a state of Australia to the shoe industry in USA , $x_{3,1}$, is written as [1, 2, 1 -> 2, 1]. The value of $x_{3,1}$ is indicated as (0.23) [1, 2, 1 -> 2, 1], which means the value of this element is set as 0.23. Some other notations are also included to improve the flexibility and efficiency of the query language. In the system, users' specification is a set of XML-based files including some scripts written in the query language. The concept hierarchy is crucial to assign the meaning to the data retrieved from various sources, since the coordinates of x are completely determined by the hierarchies.

This XML-based file also includes other notation to indicate the data integration component the location of the data source, how to retrieval the desired information from these sources and where to allocate the data into the following data mining model.

2.2 Spatio-Temporal Data Mining with Conflicts

In the data mining component, a unique data mining algorithm is designed to estimate the matrix. This mining algorithm utilizes two types of information: the historical information which contains the temporal patterns between input-output tables of previous years, and the spatial information within the current year. For example, this spatial information can be the total commodity output of the given industry within the current year, or the total greenhouse emission of the given industry. The simplified version of the mining algorithm can be written in the format of an optimization model as below:

$$Min[\frac{dis(X-\overline{X})}{\varepsilon_1}+\sum_{i=1}^{n}\frac{e_i^2}{\varepsilon_{i+1}}] \text{ subject to: } GX+E=C \qquad (1)$$

where: X is the target table to be estimated , $\overline{X}$ is the table of the previous year, E is a vector of the error components $[e_1,...,e_n]^T$, *dis* is a distance metric which quantifies the difference between two matrices, G is the coefficient matrix for the local constraints, C is the right-hand side value for the local constraints. As the *dis* metric has many variety, the one used in this experiment is $\sum(x_i-\overline{x}_i)^2$.

The idea here is to minimize the difference between the target matrix and the matrix of the previous year, while the target matrix satisfies with the local regional information to some degree. For example, if the total export of the sheep industry from Australia to China is known as c_1 , then $GX+E=C$ can be

$$\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}X_1\\X_2\end{bmatrix}+e_1=c_1 \text{.}$$

The element e_i in *E* represents the difference between the real value and estimate value, for example,

$$e_1 = c_1 - \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

The reason why it is introduced is to solve the *conflicting information*. Very often the data collected from different sources is inconsistent between each other, and even conflicting. For example, the number of export of Australian Sheep industry reported by the Australian government may be not consistent with the number of import from Australian Sheep reported by the USA government. Here e_i is introduced to balance the influence between the conflicting information, and reaches a tradeoff between the conflicting information. In some mathematric optimization literatures, this variable is called stochastic variable (Bazaraa, Jarvis et al. 2005).

This data mining algorithm assumes the temporal stability, which assumes the industry structure of a certain region keeps constant or has very few changes within the given time period. This assumption often required to be verified for long time period. Within the short time period, the dramatic change of the industry structure is relatively rare.

Data objects in typical data mining algorithm often can be reduced to points in some multidimensional space without information loss. But the spatio-temporal modelling algorithm can analyze data containing two types of information simultaneously; thereby maximally utilize the available information. In our case, $dis(X - \overline{X})$ models the temporal information of the input-output models between years, and $GX + E = C$ models the spatial or other type of regional or interregional information of the input-output model within the current year.

The reason why the spatio-temporal modelling algorithm is suitable to this system is due to the unique characteristics of the datasets that the system aims to process. The datasets often contain the temporal patterns between years, such as the trend of the carbon emission of certain industry sections, and also much spatial information regarding the total emission within a certain region such as national total emission and state total emission. Even more, the datasets also contains the interrelationship between the industries within a given region or between regions. On the other hand, it is very common that either of datasets is not comprehensive and imperfect and even the conflicts between the datasets exist. Thereby, the modelling algorithm is required to consolidate the conflicted datasets to uncover underlying models, and at the same time, the modelling algorithm is required to incorporate the spatial information and keep the spatial relationship (such as dependency and heterogeneity (Miller 2007)) within datasets.

2.3 Parallel Optimization

In real world practice, the previous modeling algorithm often processes a matrix with dimensions over 1000-by-1000. In the foreseeable future, the size of estimated matrix will increase over 100,000-by-100,000. This requires the algorithm to have extremely outstanding capacity of processing large datasets. In order to address this problem, one parallel optimization algorithm is designed as the solver. The key idea is to divide the constraints into a few subsets of constraints, and then to do optimization against the subset of constraints respectively instead of the whole set of constraints. The simplest case is that the original optimization problem is rewritten as a set of sub-problems

Sub-problem 1 (soft constraints):	Sub-problem 2 (hard constraints):	Sub-problem 3 (nonnegative constraints):
$Min[\frac{(X-\overline{X})^2}{\varepsilon_1}+\sum\frac{e_i^2}{\varepsilon_{i+1}}]$ Subject to: $G_1X+E=C_1$	$Min[\frac{(X-\overline{X})^2}{\varepsilon_1}]$ Subject to: $G_2X=C_2$	$Min[\frac{(X-\overline{X})^2}{\varepsilon_1}]$ Subject to: $X\geq 0$

The results from the sub-problems are combined as a weighted sum which consequently acts as a start point for the next iteration. Suppose the result from the ith sub-problems is $P_i(X_n)$, the weighed sum is written as

$$X_{n+1}=X_n+L[\sum w_iP_i(X_n)-X_n]$$

where L is the relaxation parameter. This method is a special case of the parallel projection method (PPM) (Combettes 2003).

3. Experimental Results

Here we present two methods of checking the quality of the input-output table: direct and indirect checking. The reason why we introduce the indirect method is that the direct evaluation of a large-size matrix is a rather difficult task. A thousand-by-thousand matrix contains up to ten million of numbers. The simple measurements such as the sum do not make too much sense, as the important deviation is submerged by the total deviation which normally is far larger than the individual ones. The key criterion here is the distribution or the interrelationship between the entries of the matrix: whether the matrix reflects the true underlying industry structure, not necessary the exactly right value, at least the right ratios.
First, we create some artificial data, a 12-by-12 input-output table in order to see the performance of this approach. During the experiment, the coefficient $1/\varepsilon_1$ in the equation (1) is tuned to fit the data properly. We change $1/\varepsilon_1$ ranging from 10, 1, 0.1 to 0.01, and the result is below:

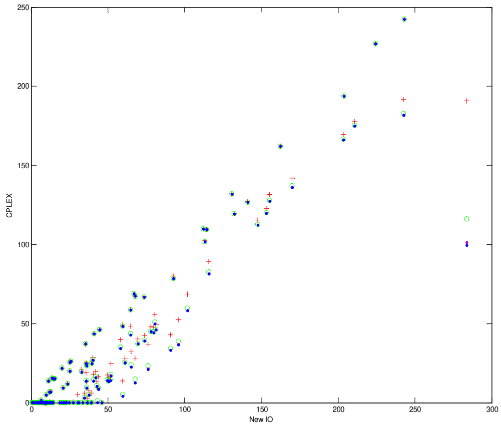

Figure 3. Results with the deviation from 10 to 0.01.

The red cross is the result from the derivation 10, the green circle is the result from the derivation 1, the red dot is the results from the deviation 0.1, and the black dot is the result from the deviation 0.01.

Clearly, with the decrease of value of $1/\varepsilon_1$, the resulted input-output table is moving away from the table of previous year. While $1/\varepsilon_1$ is set smaller, the mining algorithm pushes the model toward the second part of the equation (1).

The multipliers in the input-output framework reflect the impacts of the final demand changes on the upstream industries (Miller and Blair 1985). The information contained by the multipliers is very similar to the sensitivity analysis in the general statistics. The general formula of constructing the multipliers is:

$$M = D(I - A)^{-1}$$

where *M* is the multiplier, *I* is the identity matrix, A is the technique coefficients matrix each element of which $a_{ij} = x_{ij} / \sum x_{ij}$, and *D* is the change of the final demand.

This multiplier indeed counts the impact of the change on the upstream industries, not only the direct supply of the final demand. Any deviation occurring in the upstream industries from the underlying true structure will be amplified and reflected on the multipliers. Thereby, the multipliers send indirect warning signal to imply the deviation occurring on the upstream.

Industry	Direct Intensity	Total Multiplier
Sheep and lambs	0.175306192	0.229353737
Wheat	0.179059807	0.27047284
Barley	0.178962619	0.235765397
Beef cattle	0.175528219	0.265691956
Untreated milk & Dairy cattle	1.233958	1.46699422
Pigs	0.177604301	0.273531332
Poultry & Eggs	0.288054	0.47927187
Sugar cane	3.664307556	3.720540595
Vegetables & Fruit	0.262444	0.3157365
Ginned cotton	0.000836372	0.297221617

Here a part of the multipliers are used to measure the quality of the resulting matrix. This matrix aims to calculate the total water usage of the different industries in Australia. A part of the data is collected from the Water Account reports produced by the Australian Bureau of Statistics (2004-05). At the table above, the direct intensity indicates the direct usage of the water by the industry, and total multipliers indicate the all upstream water usage of the industry. The difference between the direct intensity and total multiplier indicates the upstream consumption. For example, the pig industry has total multiplier 0.273531332. That means each dollar of pork will cost 0.27 litre of water, but the direct usage of water is only 0.177 litre per dollar. The 0.1 litre water is consumed by upstream industries such as some agriculture sections which supply the food for pigs.

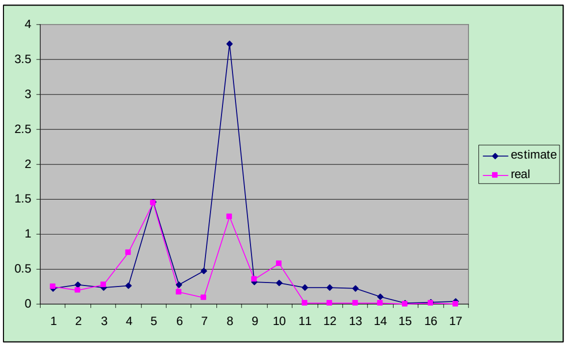

Figure 4: Comparison between two series of multipliers

From the above plot comparing the two series of the multipliers, two series basically follow the same trend, which indicates the industry structure is estimated properly. However the estimated multipliers are more frustrated than the true underlying multipliers, which indicates the estimated multipliers amplifies the impact on the upstream.

Another reason of the difference is the underlying structure change of the given industry. For example, from 1999 to 2004, the Australian rice industry reduces dramatically its rice production due to the continuous draught, but imports more and more rice from other nations such as Thailand and Vietnam. At the same time, the price of rice has been inflated, thereby, the ratio of the water usage by price is dropping, as the price is rising and the water usage is dropping.

4. Conclusion

This system is a data warehousing and data mining tool for estimating and updating input-output tables. The unique characteristics of the data for the input-output tables determine this intelligent system must be capable of dealing the temporal and spatial data simultaneously. At the same time, the large size of the estimated input-output table makes it a difficult task to check the quality of the matrix. This paper presents an integrated system starting from the data collection to the data mining and presentation. According to the result of the experiments, the system successfully produces the input-output tables for the triple bottom line methods (TBL) in the environmental accounting. This system makes estimating input-output tables to become a rather easy task without a huge amount of works for integrating and updating data and model. Before this system, this kind of integration and updating costs months of work, but now it takes only a few days with the consistent quality.

There are still many places to be further investigated. For example, the data mining algorithm can incorporate more historical data by including the data of a few previous years instead of the data of the immediate previous year in the current model. It requires much larger computational ability, and we are investigating more powerful parallal algorithms which will bring enormous extra ability to process extremely large datasets.

Acknowledgement:

Authors would like to give thanks to our colleagues A/Prof Regina Burachik, Dr Yalcin Kaya from the School of Mathematics, University of South Australia, and Jodie Gonzalez Jennings for contribution.

References

(2004-05). 4610.0 - Water Account, Australia. Canberra, The Australian Bureau of Statistics.

Bazaraa, M. S., J. J. Jarvis, et al. (2005). Linear programming and network flows. Hoboken, NJ, Wiley-Interscience.

Brown, D., J. Dillard, et al. (2006). Triple Bottom Line: A Business Metaphor for a Social Construct Understanding the Social Dimension of Sustainability, Taylor & Francis Group.

Combettes, P. L. (2003). "A Block-iterative Surrogate Constraint Splitting Method for Quadratic Signal Recovery." IEEE Transactions on Signal Processing **51**(7): 1771- 1782.

Foran, B., M. Lenzen, et al. (2005). Balancing Act: A Triple Bottom Line Analysis of the Australian Economy, CSIRO and the University of Sydney.

Hendrickson, C. T., L. B. Lave, et al. (2006). Environmental Life Cycle Assessment of Goods and Services: An Input-Output Approach, Resources for the Future.

Hobbs, L., S. Hillson, et al. (2005). Oracle Database 10g Data Warehousing, Elsevier Digital Press.

Miller, H. J. (2007). Geographic Data Mining and Knowledge Discovery. The Handbook of Geographic Information Science. J. Wilson and A. S. Fotheringham, Wiley-Blackwell.

Miller, R. E. and P. D. Blair (1985). Input-output Analysis, Foundations and Extensions. Englewood Cliffs, New Jersey, Prentice-Hall Inc.

Peters, G., F. Sack, et al. (2007). "Towards a deeper and broader ecological footprint." Engineering Sustainability.