A Data Mining System for Estimating a Large-size Matrix in Environmental Accounting

Ting Yu, Manfred Lenzen, Blanca Gallego and John Debenham

Centre for Integrated Sustainability Analysis, University of Sydney, NSW 2006, Australia

Faculty of Information Technology, University of Technology, Sydney, NSW 2007, Australia

Abstract: This paper presents a data mining system being capable of automatically estimating and updating a large-size matrix for environmental accounting. Environmental accounting addresses how to correctly measure 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 model which is in the format of a matrix. This system aims to estimate the large-size input-output model and consists of a series of components with the purposes of data retrieval, data integration, data mining, and model presentation. This unique system is able to interpret and follow users' XML-based scripts, retrieve data from various sources and integrate them for the following data mining components. The data mining component is based on a unique mining algorithm which constructs the matrix from the historical data and the local data simultaneously. This unique data mining algorithm runs over the parallel computer to enable the system to estimate a matrix of the size up to 3700-by-3700. The result demonstrates the acceptable accuracy by comparing a part of the multipliers with the multipliers calculated by the matrix constructed by the surveys. The accuracy of the estimation directly impacts the quality of environmental accounting.

1. Introduction

Environmental protection has caught more and more attention, while the climate is becoming more unpredictable. Instead of a particular protection technique such as water cleaning, 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, environmental accounting requires proper methodologies to correctly measure the environmental impact such as greenhouse gas emission [1]. There are several accounting approaches of measuring emission, such as auditing and *triple bottom line methods (TBL)* [2]. The TBL 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 [3]. The traditional accounting method only measures 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 [4]. 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 [4]. In the EIO-LCA method, the input-output model is the key engine. The input-output model 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 model needs to be updated at least annually to reflect the new circumstance. Unfortunately, in most countries such as Australia, the input-output model is only constructed every 3-4 years, because the large amount of monetary and human cost is involved. The Centre for Integrated Sustainability Analysis (ISA), University of Sydney, is developing a data mining system to estimate and update the input-output model at different level on a regular basis.

The past decades have seen the booming supply of data from various sources, and large amounts of data regarding the environment and economic can be accessed. Unavoidably, data from various sources has various structures and ways of represent their underlying meaning. It is a time-consuming process to restructure the various types of data into a single structure and estimate the matrix. In many cases, this kind of integration and matrix estimation operation becomes a daily routing task in order to keep the information up to date. The proposed system aims to automate the whole process and reduces the manual intervention and much human's involvement.

2. System Design

The whole system consists of functional components: data retrieval, data integration, data mining and model presentation. The row data is retrieved from various data sources, and restructured and integrated into a data mining model. Then the data mining model is fed into the data mining algorithm and consequently solved by the optimization engine. The result from the data mining algorithm is the final result that is an estimated matrix.

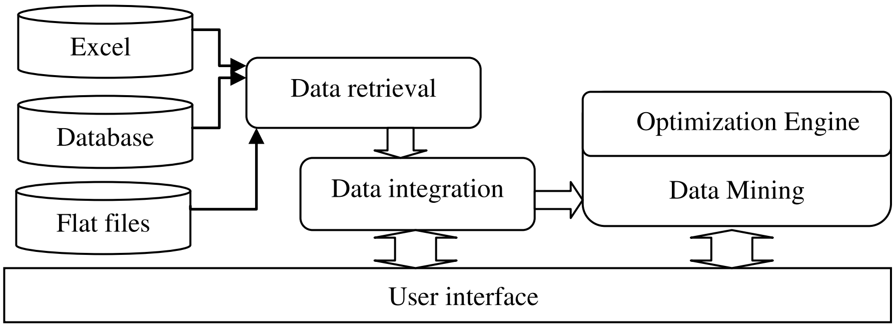


Fig. 1. Structure of the System

The data retrieval component acts as interfaces to all types of datasets including 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 and presents the result for data mining. The data mining component is the core of the whole system. In this component, a unique data mining algorithm is designed to estimate the matrix.

3. Data Integration

The data integration component includes two main sub-modules: the structure builder and the model constructor. The structure builder constructs the tree structure that we will discuss in detail later, and the model constructor constructs the mining model for the following data mining component. Within the model constructor, there are two processes to restructure the data: 1) require the interfaces to retrieval data from various sources and integrate them, and 2) restructure and assign the meaning to the data according to the previous tree structure and users' specification and populate the mining model.

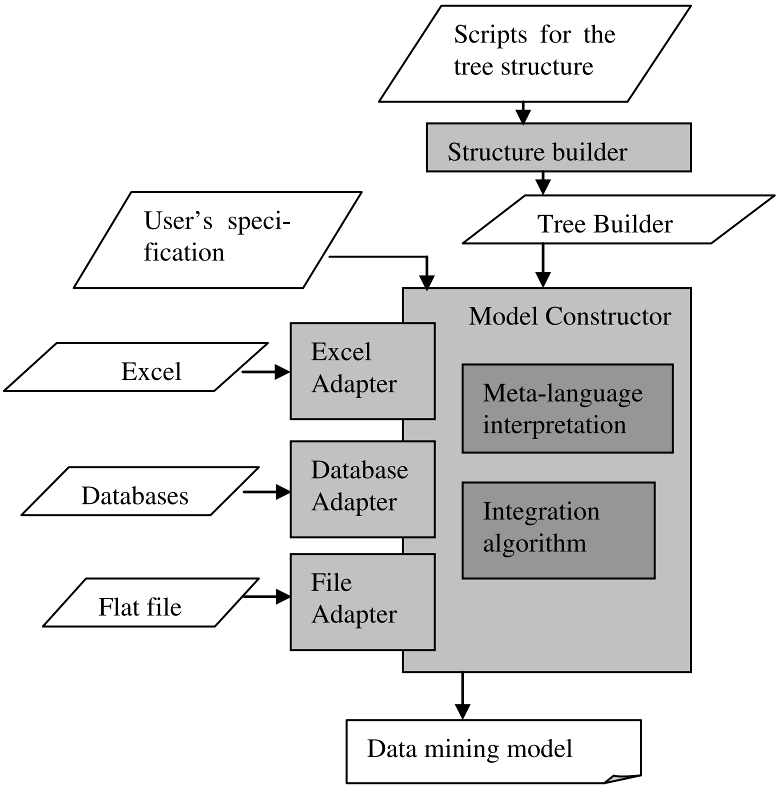


Fig 2. Data Integration Component

The first step is to construct the tree structure. The tree structure is pre-required for restructuring data collected from various sources. An example of the tree structure (See Fig 3) is a three-level tree representing the Australian Economic, one branch of which represents the sheep industry section within the New South Wales, a state of Australia. If the numerical indices are employed instead of their names, the sheep industry section within the New South Wales, a state of Australia can be written in [1,1,1] which means the first leaf in the first branch of the first tree.

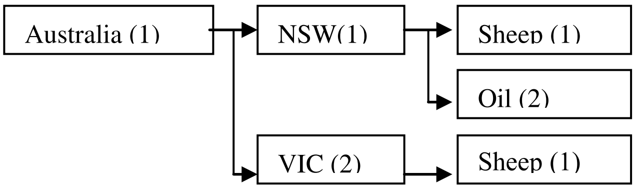


Fig 3, An Example of Tree Structure

The row and column of the matrix is defined by this tree structures, thereby the matrix is defined by the tree structures. The tree structure is unnecessarily with three levels. For example, a matrix (see Figure 4) can be organized by one three-level tree at the row side and one two-level tree at the column side. The coordinate of one entry, say X_1, can be defined as by [1,1,1] at the row side and [1,1] at the column side. That means the entry, X_1, defined by a three-level tree structure and a two-level tree structure at the column side. The tree structure is crucial to assign the meaning to the data retrieved from various sources, since the coordinates of entries are completely determined by it.

			China (1)	
			Shoe (1)	Retail (2)
Australia (1)	NSW (1)	Sheep (1)	$X_1 = 0.23$	X_2
		Oil (2)	X_3	X_4
	VIC (2)	Sheep (1)	X_5	X_6
		Oil (2)	X_7	X_8

Fig 4, An Example of the Matrix Defined by the 3-level Tree and the 2-level Tree

Considering the difference between applications, a dynamical structure of resultant matrix provides the flexibility to expand this software system to different application. On the other hand, the flexibility of the structure makes the system to be available to various level of implementation. For example, there is huge difference between the structures of resultant matrix at the national and at the corporate level, as the operations within a corporate are much simpler than those of a nation in the most cases. The dynamic of the structure is introduced by a multi-tree structure like Figure 3.

Considering the complexity of the model, a Meta language is introduced to provide users' an easy way to organize their data. The Meta language must be compact and accurate to make the description to be readable and useful. It is unrealistic to write hundred thousands of code to describe a single model at a daily base. The meta language we create is based on the coordinate of the valuable in the resultant matrix. For example, the coordinates of one entry is written as [1, 1, 1 -> 1, 1]. The value of this entry X_1 is indicated as the (0.23) [1, 1, 1 -> 1, 1] (See Figure 4). The system will fill the 0.23 in the cell with the coordinate [1, 1, 1] at the row side and [1, 1] at the column side. Consequently, this script indicates that 0.23m dollar worth of sheep products are transferred to the shoe industry in China. Some other notations are also included in order to improve the flexibility and efficiency of the Meta language. Therefore, users' specification is a set of XML-based files including some scripts written in the meta language. This kind of XML-based file indicates the system where to find the data source, how to retrieval the desired information from these sources and where to allocate the data into the optimization model.

4. Temporal-Spatial Mining with Conflict Information

The data mining component is the core engine of the whole system. In this 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 matrices of previous years, and the local information within the current year. For example, this local information can be the total 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:

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

where:

X is the target matrix to be estimated , $\overline{X}$ is the matrix of the previous year, E is a vector of the error components $[e_1,...,e_i]^T$

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 paper is the $\sum(X_i-\overline{X}_i)^2$.

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 $[1,1]\begin{bmatrix}X_1\\X_2\end{bmatrix}+e_1=c_1$. The element e_i in E represents the difference between the real value and estimate value, for example, $e_1=c_1-[1,1]\begin{bmatrix}X_1\\X_2\end{bmatrix}$. 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 by the Chinese government. Here e_i is introduced to balance the influence between the conflicting information, and reaches a tradeoff between the conflicting information.

This 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 is often required to be verified for long time period. Within the short time periods, dramatic change of the industry structure is relatively rare.

Traditional data mining only employs the single data mining techniques, such as the temporal model or spatial model. But two-dimensional data mining algorithm can integrate two types of the data mining models, 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 information of the input-output model within the current year.

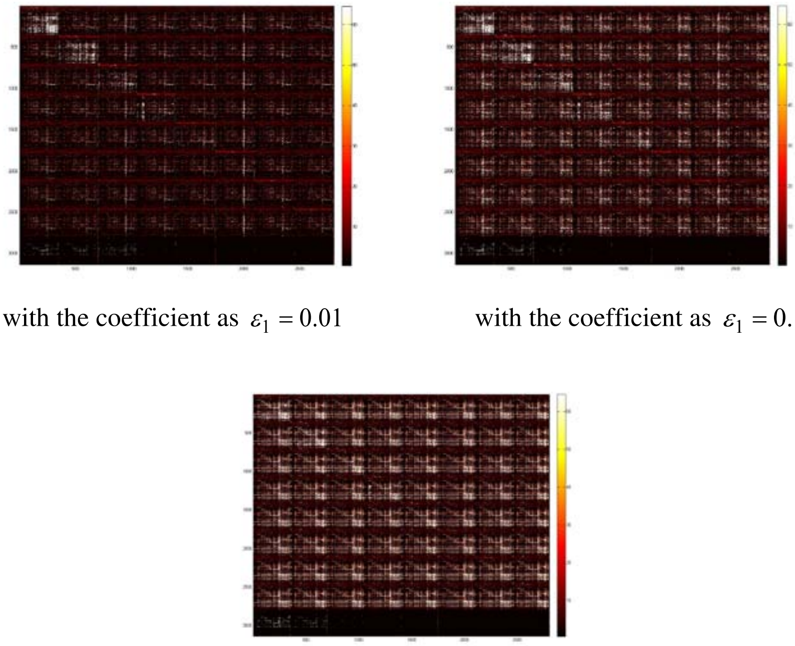
The reason why the temporal-spatial mining algorithm is suitable to this system is due to the unique characteristics of the datasets that the system aims to process. The datasets often contain the temporal patterns between years, such as the trend of the carbon emission of certain industry sections, and also much spatial information regarding the total emission within a certain region such as national total emission and state total emission. 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 mining algorithm is required to consolidate the conflicted datasets to uncover underlying models.

5. Experimental Results

The direct evaluation of a large-size matrix is a rather difficult task. A thousand-by-thousand matrix contains up to ten million numbers. 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 X_i within the matrix: whether the matrix reflects the true underlying industry structure, not necessary the exactly right value, at least the right ratios. During the experiment, the coefficient ε_1 in the equation (1) is tuned to fit the data properly. Here three examples are presented to demonstrate the effects of the tuning.

Darker the color is, the smaller the value of the entry will be. From three pictures (See Figure 5), while ε_1 is set smaller, the mining algorithm pushes the model toward the first part of the equation (1).

Often, we estimate the result of experiments by two methods: direct comparison and indirect comparison between the multipliers of the matrices. The comparison between a part of the resulting input-output table and the available survey data examines the quality of the result of the experiment under the microscope, but it hardly gives the overall quality.



with the coefficient as $\varepsilon_1 = 0.01$

with the coefficient as $\varepsilon_1 = 0.1$

with the coefficient $\varepsilon_1 = 1$

Figure 5: Three pictures by the turning the coefficient ε_1

The multipliers in the input-output framework reflect the aggregated impacts of the final demand changes on the upstream industries [5]. The information contained by the multipliers is very similar to the sensitivity analysis in 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, D is the change in the final demand, and A is the technique coefficients matrix, each entry of which is $X_i / \sum_{i=1}^{n} X_i$. Here, X_i is a value from the matrix estimated by the equation (1). More detailed explanation is available at [5]. This multiplier counts the impact of the change on the whole upstream industries, and 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 an indirect warning signal to imply the deviation occurring on the upstream.

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 [6]. In the following tables, the direct intensity indicates the direct usage of the water by the industry, and total multipliers aggregate the whole upstream water usage of the industry.

Industry	Water Usage (ML)	Final Demand (K$)	Direct Intensity	Total Multiplier
Sheep and lambs	385360	2198275	0.175306192	0.229353737
Wheat	795403	4442246	0.179059807	0.27047284
Barley	193433	1080896	0.178962619	0.235765397
Beef cattle	1557332	8872487	0.175528219	0.265691956
Untreated milk & Dairy cattle	2275602	3687201	1.233958	1.46699422
Pigs	150566	847785	0.177604301	0.273531332
Poultry & Eggs	312984	1811428	0.288054	0.47927187
Sugar cane	1269012	346329	3.664307556	3.720540595
Vegetables & Fruit	862027	3747712	0.262444	0.3157365
Ginned cotton	2120	2534832	0.000836372	0.297221617
Softwoods	141702	809463	0.175060473	0.236982027
Hardwoods	53954	307576	0.175416808	0.239130312
Forestry	150577	860234	0.175046587	0.229861225
Black coal	159409	18603943	0.008572854	0.10262849

Table 1: Estimated Multipliers

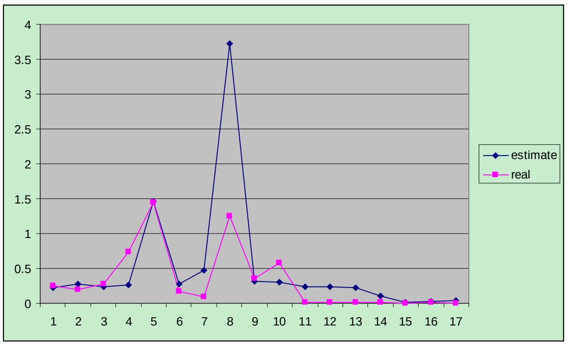


Figure 6: Comparison between two series of multipliers

From the above plot (see Figure 6) comparing the two series of the multipliers, two series basically follow the same pattern, which indicate the industry structure is estimated properly. However the estimated multipliers are more volatile than the true underlying multipliers. This phenomenon indicates the estimated multipliers amplify the errors introduced to the upstream industries.

6. Conclusion

To the best of our knowledge, this system is the first data mining system for estimating the input-output tables for environmental accounting. The unique characteristics of the data for environmental accounting determine the data mining system must be capable of dealing the temporal and spatial data simultaneously. At the same time, the large size of the estimated matrix makes it a difficult task to check the quality of the matrix. This paper presents a completed data mining system starting from data collection to 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 environmental accounting. This system makes environmental accounting a rather easy task without a huge amount of work to collect and update both data and model. Before this system, this kind of collection and updating work 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 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 the algorithms over the parallel computing which will bring enormous power to process extremely large datasets.

Acknowledgement:
Authors would like to thanks our colleagues Dr Chris Dey, Jeremy Badcock and Jodie Gonzalez Jennings for contribution.

Reference:

1. Peters, G., et al., *Towards a deeper and broader ecological footprint.* Engineering Sustainability, 2007.
2. Foran, B., M. Lenzen, and C. Dey, *Balancing Act: A Triple Bottom Line Analysis of the Australian Economy*. 2005: CSIRO and the University of Sydney.
3. Brown, D., J. Dillard, and R.S. Marshall, *Triple Bottom Line: A Business Metaphor for a Social Construct* in *Understanding the Social Dimension of Sustainability*. 2006, Taylor & Francis Group.
4. Hendrickson, C.T., L.B. Lave, and H.S. Matthews, *Environmental Life Cycle Assessment of Goods and Services: An Input-Output Approach.* 2006 Resources for the Future.
5. Miller, R.E. and P.D. Blair, *Input-output Analysis, Foundations and Extensions*. 1985, Englewood Cliffs, New Jersey: Prentice-Hall Inc.
6. *4610.0 - Water Account, Australia*. 2004-05, The Australian Bureau of Statistics: Canberra.