Automatically Estimating and Updating Input-Output Tables

Ting Yu*, Manfred Lenzen, Chris Dey, and Jeremy Badcock

Centre of Integrated Sustainability Analysis, Physics Building A28, 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. The system in this paper consists of a series of components with the purposes of data retrieval, data integration, data analysis, and quality checking. This unique system is able to interpret and follow users' XML-based query scripts, retrieve data from various sources and integrate them for the following data analysis components. The data analysis component is based on a unique modelling algorithm which constructs the matrix from the historical data and the spatial data simultaneously. This unique data analysis algorithm runs over the parallel computer to enable the system to estimate a large-size matrix. The result demonstrates the acceptable accuracy by comparing a part of the multipliers with the corresponding multipliers calculated by the matrix constructed by the surveys.

1 Introduction

In the theoretical economics, the input-output model of economics uses a matrix representation of a nation's (or a region's) economy to predict the effect of changes in one industry on others and by consumers, government, and foreign suppliers on the economy [1]. 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 an integrated intelligent system to estimate and update the input-output model at different level on a regular basis.

The input-output model often consists of a time series of matrices which may have temporal stability or temporal patterns. At the same time, within a given time period, extra information regarding certain parts of the matrix is often available from various government departments or other public or private organizations. However, most of this information is often incomplete and only gives a snapshot of a part of the underlying model. Apart from the massive data, hundreds of years of research has accumulated substantial amount of general knowledge of the national economic. Any

* Corresponding author.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 42–49, 2009.

researcher could utilize this public knowledge to facilitate their discovery. On the contrast, other knowledge discovery activities often do not have such rich resource.

A time series of input-output models represents the evolution of industry structure within and between regions, where the region is defined as a geographic concept. It is a spatio-temporal knowledge discovery process with the help of rich domain knowledge. Including time introduces additional complexity to the geographic knowledge discovery [2]. This paper presents a novel algorithm which estimates and updates the economic matrix for the general equilibrium theory.

2 System Design

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

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

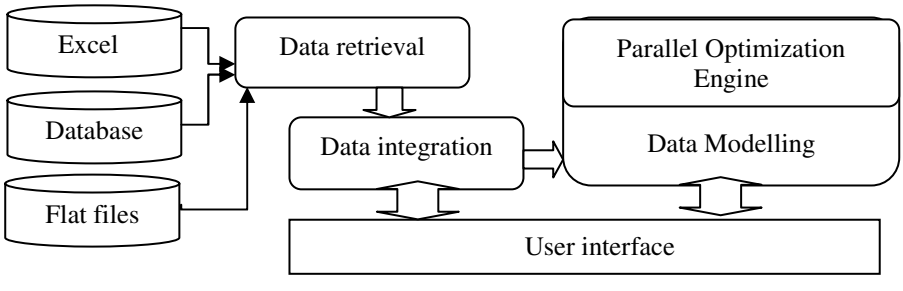

Fig. 1. System Architecture

2.1 Data Integration

The data integration component includes two main sub-modules: the structure builder and the model constructor (See Figure 2). 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.

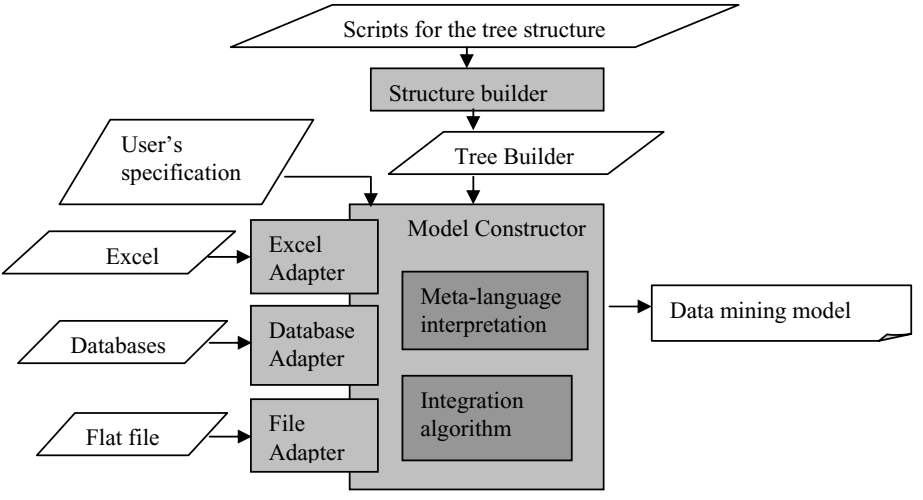

Fig. 2. Data Integration Component

			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. 3. An Example of the Matrix Defined by the 3-level Tree and the 2-level Tree

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.

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 3) 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.

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 in 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 3). 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.

2.2 Spatio-temporal Modelling with Conflict Information

The data analysis component is the core engine of the whole system. In this component, a unique modelling algorithm is designed to estimate the matrix. This modelling algorithm utilizes two types of information: the historical information which contains the temporal patterns between matrices of previous years, and the spatial information within the current year. For example, this spatial information can be the total output of the wool industry in Australia within the current year, or the total greenhouse emission of the car manufacture industry in Australia. The previous tree structure is employed here to represent the geographic concept hierarchies within the spatial information. The modelling algorithm can be written in the format of an optimization model as below:

$$Min[\frac{dis(X-\overline{X})}{\varepsilon_1}+\sum\frac{e_i^2}{\varepsilon_{i+1}}], \text{ Subject to: } G_1X+E=C_1 \quad (1)$$
$$G_2X=C_2$$
$$X\geq 0$$

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, e.g. $\sum(X_i-\overline{X}_i)^2$ in this case. G is the coefficient matrix for the local constraints, and C is the right-hand side value for the local constraints. 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. Here e_i is introduced to balance the influence between the conflicting information, and reaches a tradeoff between the conflicting information. This modelling 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. By this assumption, this algorithm considers the matrix for the previous year as a reasonable proxy to the current year. Within two successive years, dramatic change of the industry structure is relatively rare, and this assumption has a good ground.

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 production of certain industry sections, and also much spatial information regarding the total production 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 [3]) within datasets.

2.3 Parallel Optimization

In real world practice, the previous modeling algorithm often processes 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):
$Min[\frac{(X-\overline{X})^2}{\varepsilon_1} + \sum \frac{e_i^2}{\varepsilon_{i+1}}]$	$Min[\frac{(X-\overline{X})^2}{\varepsilon_1}]$	$Min[\frac{(X-\overline{X})^2}{\varepsilon_1}]$
Subject to: $G_1 X + E = C_1$	Subject to: $G_2 X = C_2$	Subject to: $X \geq 0$

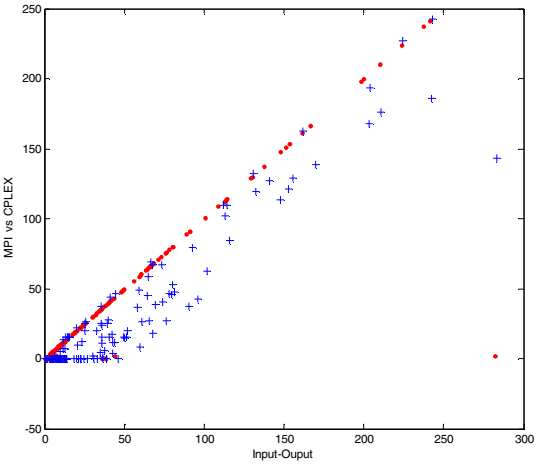

Fig. 4. Results from the CPLEX (Blue Cross Points) vs. results from the Parallel Optimization (Red Dot Points) by comparing with the real input-output data. The x-axis represents the real input-output table.

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 $X_{n+1} = X_n + L*[\sum w_i P_i(X_n) - X_n]$, where L is the relaxation parameter. This method is a special case of the parallel projection method [4]. Because the objective function of this particular problem is quadratic, thereby convex and the constraints are linear thereby convex as well, the optimization process is simpler than general projection methods. This parallel optimization algorithm is implemented over the Message Passing Interface (MPI). For the purpose of demonstration, the performance is compared with a commercial optimization package, the CPLEX by using the same test dataset concluding 12-by-12 entries with 57 constraints.

According to the experiments (see Figure 4), the parallel optimization estimates the underlying the matrix better as the linear relationship between its estimated result and the real matrix are very clear. The drawback is that the parallel optimization does not prevent the estimated value from becoming negative. The first sub-problem requires the estimation to be positive or zero. However in Figure 4, some estimated values are very small negative.

3 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 of the matrix: whether

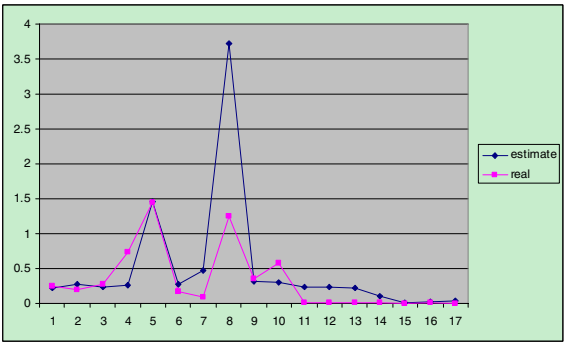

Fig. 5. Comparison between two series of multipliers

the matrix reflects the true underlying structure, not necessary the exactly right value, at least the right ratios.

The multipliers in the input-output framework reflect the aggregated impacts of the final demand changes on the upstream industries [1]. 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 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).

This sensitivity multiplier counts the impact of any change of outputs on the whole upstream inputs, and not only the direct inputs. Any deviation occurring in the upstream inputs from the underlying true structure will be amplified and reflected on the multipliers. Thereby, the multipliers send an indirect warning signal to imply the structural deviation occurring on the upstream inputs.

As a case study, a 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 [5]. The full Australian economy consists of 8 states with 344 industry sections plus7 final demands and 6 value added sections. Totally, the Australian input-output table is a 2808-by-2800 matrix, containing 7,862,400 entries. In order to estimate the matrix, more than 260,000 constraints are included.

From the below plot (see Figure 5) comparing the two series of the multipliers, two series basically follow the same trends, 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. Another reason of the difference is the underlying structure change within a given industry. In Figure 5, the big gap between two series at 8 indeed indicates from 1999 to 2004, the Australian rice industry dramatically reduces its rice production due to the continuous draught, but imports more and more rice from other nations. As the price has been inflated and the water usage is dropping, the ratio of the water usage by price is dropping.

4 Conclusion

This system is an integrated data analysis system for updating a large-scale matrix. The unique characteristics of the data determine the data analysis system must be capable of dealing the temporal and spatial data simultaneously. At the same time, the large size of the estimated matrix requires the system to process a large amount of data efficiently. This paper presents a completed data analysis system starting from data collection to data analysis and quality checking. According to the result of the experiments, the system successfully produces the matrix, and makes it 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.

As the temporal stability is a major assumption, the further developments will emphasis on how to deal with the major structure changes. This research can be applied to broader horizon such as the Markov transient matrix.

References

1. Miller, R.E., Blair, P.D.: Input-output Analysis, Foundations and Extensions. Prentice-Hall Inc., Englewood Cliffs (1985)
2. Miller, H.J., Han, J.: Geographic Data Mining and Knowledge Discovery. CRC, Boca Raton (2001)
3. Miller, H.J.: Geographic Data Mining and Knowledge Discovery. In: Wilson, J., Fotheringham, A.S. (eds.) The Handbook of Geographic Information Science, Wiley-Blackwell (2007)
4. Combettes, P.L.: A Block-iterative Surrogate Constraint Splitting Method for Quadratic Signal Recovery. IEEE Transactions on Signal Processing 51(7), 1771–1782 (2003)
5. 4610.0 - Water Account, Australia. 2004-05, The Australian Bureau of Statistics: Canberra