Methods for estimating input-output models under partial information

Wood, Richard

Industrial Ecology Program, NTNU, 7491, Norway

ph: +4773598938, e-mail: richard.wood@ntnu.no, http://www.ntnu.no/indecol

#### Abstract

Input-output models have been estimated under partial information ever since the postulation of the RAS technique by Stone in 1962. Methods have become much more advanced since this time, such that any type of related data can now be incorporated in the estimation techniques, not just row and column sums. The literature is now at the point where any information in differing pricing types can be fairly easily incorporated into input-output matrix estimation. As a first step, this paper outlines the generic ability of incorporating new data sources.

As a second step, this paper discusses the application of matrix estimation techniques for not only updating input-output models, but also for disaggregating the models. As an example, increased sector detail may be available from physical data such as energy or LCA databases that could be incorporated into an input-output model.

Source data is almost never 100% in agreement, and updating methods generally must be robust enough to handle differences in data sources. Differences can be overcome by a stepped updating method within tables, or by incorporating a specific balancing hierarchy or a hierarchical term within the target function. Again, the choice of form of the term in the target function has different economic meaning. The effect of the different applications of hierarchy can be significant, and the implications are discussed.

Finally, whilst computational ability is far beyond what was available to Stone and colleagues, we are still limited by the ability of solvers to find global minimums, and, less so, in the size of problems. These limitations eventually, flows back to the choice of target function, some being easier to compute than others. Discussion concludes on recent experience in estimating single-region input-output tables within a global multi-regional model (EXIOPOL), with particular reference to the computability of highly conflicting source data.

## Introduction

Since the early days of input-output table creation, there has been a steady advancement of mathematical techniques used to estimate missing information. Beginning with the basic RAS technique (Stone and Brown 1962), extending through modifications to the RAS technique to hold some variables constant(Allen 1974; Lecomber 1975; Cole 1992; Gilchrist and St Louis 1999). Additional reliability information has been further included in the balancing of tables (Lecomber 1975; Lecomber 1975), and negative elements have been included in balancing processes (Junius and Oosterhaven 2003; Lemelin 2009; Lenzen et al. 2009). (Lenzen et al. 2009) summarise a revised method that addresses all the above concerns, including constraint violation, in a single method.

As most are well aware, the derivation of balancing processes as constrained optimisation problems has developed alongside the iterative procedures described above. One of the earlier direct comparisons between the iterative and optimisation approaches was by Bacharach (1970) on the basic RAS technique, although similar work can be found much earlier (see Lemelin 2009). A number of approaches have been put forward such as weighted sum of least squares (Morrison and Thumann 1980), using Lagrangian multipliers (Byron 1978; van der Ploeg 1982).

More recently, we have seen emphasis on using constrained optimisation techniques, dealing with both conflicting information from a variety of data sources (Robinson et al. 2001), flexibility in estimation of inter-regional trade accounts (Canning and Wang 2005),and reconciling large databases (Müller et al. 2009; Wood 2010).

The basis of all these approaches comes down to reconciling different data sources. In the case of the updating an IO table with basic RAS, a prior table is provided as the starting values, and this data is balanced to data on row totals and column totals. All modifications of the RAS process, and through the various constrained optimisation approaches follow this same basic principle: Minimise a target function *t* of a distance between an estimated matrix **Z**, from a prior matrix **Z0:**

|  |  |  |
| --- | --- | --- |
| Eq  | $$Min t\left(Z,Z0\right)$$ |  |

According to a set of additional constraints (such as row/column):

|  |  |  |
| --- | --- | --- |
| Eq  | $$C\_{w}=\sum\_{i^{}\in b\_{1}\left(w\right)}^{}\sum\_{j^{}\in b\_{2}\left(w\right)}^{}\sum\_{k^{}\in b\_{3}\left(w\right)}^{}Z\_{i,j,k}$$ |  |

Row/column totals are usually expressed as:

|  |  |  |
| --- | --- | --- |
| Eq  | $$C\_{i}=u\_{i}=\sum\_{j}^{}Z\_{i,j,1}$$$$C\_{i}=v\_{i}=\sum\_{i^{}}^{}Z\_{i,j,1}$$ |  |

Here, a 3-dimensional **Z** is used, across dimensions *i,j,k* corresponding to rows, columns and tables type, but equally valid under any n-dimensions and classifications. The row/column constraints are sums purely over the basic price table (here *k*=1).The list of constraints (e.g. row/column totals) **C** is of length *w*, and relate to **Z** by concordance matrices **b1,b2,b3**.

The problem can be set up using scaling factors on **Z0**, such that

|  |  |  |
| --- | --- | --- |
| Eq  | $$Z=S∙Z0$$ |  |

Where the “$∙$” signifies element by element multiplication. The problem can then enforce sign preservation easily.

|  |  |  |
| --- | --- | --- |
| Eq  | $$S\geq 0$$ |  |

## Disaggregation

Most literature has historically focussed on using these balancing techniques as a means to update input-output tables from prior to later years. A number of studies have also used the techniques to help disaggregate tables (Robinson et al. 2001; Wiedmann et al. 2008; Müller et al. 2009; Wood 2010). The principle is the same, but rather than using a prior year as the basis for the initial estimate **Z0**, an alternate dataset(s) or assumptions must be used in order to arrive at a full description of all flows within **Z0**. For the agricultural case, as shown in (Müller et al. 2009), a full database of agricultural requirements and sales is linked into a disaggregated subset of the European supply-use tables. Similarly in (McDougall 2009), agricultural and non-agricultural datasets are used to help disaggregate the agricultural and non-agricultural sectors, although the method used in balancing the tables here is not clear.

Ideally, real data is available by both products and sales of each flow in **Z0**. Alternatively, assumptions must be made on the applicability of one subset of data to another subset of data. At a high level, for example, product sales might be known in disaggregate, but purchases only known in aggregate. The technical co-efficients of a disaggregate sector would remain unchanged if the gross sales is then used to split all purchasers. Such assumptions must be treated carefully, and it is not the purpose of this paper to dwell on these issues, suffice to say that some information, real or assumed, must be available as a basis for flows within **Z0**.

The disaggregation problem then becomes

|  |  |  |
| --- | --- | --- |
| Eq  | $$Min t\left(Z,Z0\right)$$ |  |

According to a set of additional constraints **C***w1*, derived from an aggregate matrix $Z\_{i1,j1,k1}^{1}$, where *i1,j1,k1* represent the aggregate classifications used in the aggregated matrix, and the aggregation between each classification is shown through the concordance matrices **b1, b2, b3** :

|  |  |  |
| --- | --- | --- |
| Eq  | $$C\_{w1}=Z\_{i1,j1,k1}^{1}=\sum\_{k}^{}b\_{k1,k}^{3}\*\left(\sum\_{j}^{}b\_{j1,j}^{2}\*\left(\sum\_{i}^{}b\_{i1,i}^{1}\*Z\_{i,j,k}\right)\right)$$ |  |

Additionally, row/column constraints, or any other additional constraint can be included as before:

|  |  |  |
| --- | --- | --- |
| Eq  | $$C\_{w}=\sum\_{i^{}\in b\_{1}\left(w\right)}^{}\sum\_{j^{}\in b\_{2}\left(w\right)}^{}\sum\_{i^{}\in b\_{3}\left(w\right)}^{}Z\_{i,j,k}$$ |  |

Accounting constraints such as balances between supply and use, row/column totals, margin totals, etc must also be specified (if not implicit in the constraint set **C***w*), but for simplicity, will not be defined here.

Examples of such approaches can be found in (Wiedmann et al. 2008; Müller et al. 2009; Wood 2010).

##  Consistency between datasets

Consistency between a dataset used to provide the starting values **Z0** and the control totals $C\_{w1}$and $C\_{w}$ is not required, and is of course expected (otherwise **Z0** =**Z**). Differences between $C\_{w1}$and $C\_{w}$are much more common, however, particularly when $C\_{w1}$and $C\_{w}$are obtained from different databases. This consistency issue has been addressed through entering an error term $δC\_{w}$into Eq 7 and Eq 8.

|  |  |  |
| --- | --- | --- |
| Eq  | $$C\_{w1}=Z\_{i1,j1,k1}^{1}=\sum\_{k}^{}b\_{k1,k}^{3}\*\left(\sum\_{j}^{}b\_{j1,j}^{2}\*\left(\sum\_{i}^{}b\_{i1,i}^{1}\*Z\_{i,j,k}\right)\right)+δC\_{w}$$ |  |

|  |  |  |
| --- | --- | --- |
| Eq  | $$C\_{w}=\sum\_{i^{}\in b\_{1}\left(w\right)}^{}\sum\_{j^{}\in b\_{2}\left(w\right)}^{}\sum\_{i^{}\in b\_{3}\left(w\right)}^{}Z\_{i,j,k}+ δC\_{w}$$ |  |

And adjusting the target function to include minimisation of a function *f* of this error term:

|  |  |  |
| --- | --- | --- |
| Eq  | $$Min (t\left(Z,Z0\right)+ f(δC\_{w}))$$ |  |

##  Overlapping datasets

Even in the case of row and column constraints (Eq 3), there is overlap between constraint datasets. This is accentuated when including constraints from an aggregate table (Eq 7) and/or other control totals (Eq 8). As such, having datasets that overlap is not necessarily problematic. If the datasets used for these constraints come from different sources, however, it is likely that the data is not reconciled and that the subjective reliability of the different data sources compared to each other and the initial estimate becomes important. (Robinson et al. 2001) cite (Golan et al. 1996) in experiments on weighting the functional terms *t* and *f*, finding that an equal weighting gives reasonable results. Such experiments are highly subjective to the quality of both Z0 and Cw however, as well as the functional terms used in *t* and *f*. Equal weighting is a matter of convenience, but generally, most alternatives are subjective, (or mathematically complex and hence difficult to solve).

Alternatively, a weighting structure can be set up such that $ C\_{w}$ is always met in preference to Z0, by choosing exceptionally large weights on *f*. Again, the mathematical form of *f* and *t* influence what is an ‘exceptionally large weight’, where a change from 1-1 to 1-10 is small in absolute terms, but is large in relative terms.

##  Alternative hierarchies

Alternatively, with a number of datasets being implemented, a hierarchy of constraint values can be implanted on the estimation process from **Z0** to **Z**. The basic idea behind the hierarchical balancing is that datasets are implemented so that the starting values used to estimate a improved ***Z01*** are calculated from the scaling of the most disaggregate dataset to the next most aggregate dataset. This can be implemented so that the effect of each dataset is known more concisely. The secondary advantage is that the error term can be defined more concisely.

 The implementation is thus:

1. Establish $Z0\_{i,j,k}$ according to representative matrix.
2. Balance initial estimate to aggregate data.
	* This has the effect of scaling up/down the initial estimate/default values to match the aggregate data. This then provides the structure of the default values, correct to aggregate.
3. Iterate through datasets and aggregate data.
	* This has the effect of scaling up/down the previous estimate to the next dataset, and then scales the adjusted dataset to match the aggregate data.
4. Final balance to aggregate data
	* This ensures aggregate is being met.

##  Solvability

When using an optimization approach, correct scaling of the problem can greatly enhance the possibility of finding global optima. Model scaling refers to a linear scaling of the objective function that has no impact on the shape of the target function. The reason for the model scaling is that solvers are much more efficient at finding solutions to functions closer to unity. An incorrectly scaled problem can result in two things, 1) a non-optimal but almost optimal solution, 2) complete failure of the optimization.

In our problem, we experienced large shifts in values due to the poor quality of a lot of the input data used to disaggregate. Also, related to the problem of handling zeros, near-zero values undergo very large relative shifts when only undergoing fairly small shifts in absolute terms. We also face the problem of large variances between countries. In the initial set up of this problem, a common scaling factor was used to reduce the magnitude of the objective function close to unity. In some cases, it was found that a more precise scaling factor was needed. As such, a two step process was undertaken in order to firstly determine the rough magnitude of the scaling factor required by doing an initial run, before adjusting the scaling factor to enable a more precise answer to be reached.

Finally, two target functions were explored – the cross entropy (CE) target function which RAS uses as a basis, and the weighted least squares (QP) target function. As CE is a non-linear function, the ability to solve the problem under the datasets utilized exceeded the capacity of CPLEX to find a reasonable solution in a reasonable time. The use of a QP target function greatly simplified the computational requirements, as it can be solved much easier by quadratic solvers.

## Automation

The system was automated from input data to final balance. Handling data in a number of different classifications or including external data points can quickly become messy. In order to keep this standard, concordance matrices were utilized between every dataset and the final EXIOPOL classification. These concordance matrices become the backbone of the equations that are generated as constraints. The three components of the automation include the initial estimate of the full system in all its generality; a variable length list of constraints from different datasets and concordance matrices describing relationships between sub-components of the system, and between sub-components of the system and external data sources in any classification; and the overall balancing routine. Such a framework is very flexible and easily expandable to incorporate additional data. Time series applications, adding trading partners, including additional primary inputs, etc, all become straightforward within the general architecture of the framework. Such an approach has become beneficial when more and more data has become available over the course of projects. The main benefit is the ability to work in any classification desired.

## Experiences from EXIOPOL

EXIOPOL (Tukker et al. 2009) is a global multi-regional input-output database. A key feature of the database is consistent and detailed classifications across 44 world regions. A requirement of the database is to maintain the aggregate (source) input-output matrices of each region. The total data on transactions made available in the project is:

1. Supply, use, and input-output tables provided by Eurostat (for EU member states)
2. Supply, use, and input-output tables provided directly by statistical agencies
3. More detailed data contained in datasets made publicly available by statistical agencies
4. Agriculture social accounting matrix (AgroSAM) (also to be included in newest GTAP database)
5. Energy and electricity supply, use, and generation mix provided by the International Energy Agency (IEA)
6. Metals and non-metal minerals extraction and production and monetary values provided by the British Geological Survey
7. Metals and non-metal minerals extraction and production and monetary values provided by the US Geological Survey
8. Metal price data provided by the London Metals Exchange, British Geological Survey, and US Geological Survey in that order
9. Trade data provided by the trade linking section of this project, sourced from CommTRADE
10. Data on royalties and rents compiled from statistical agencies
11. Default values in two parts
	* Expected zero flows (e.g. nuclear industry production of wind power)
	* Near country estimate (see later discussion on initial estimates)

This provides both a huge quantity of data, as well as considerable gaps. The gaps relate particularly to areas where no data at all exists for the unique flow – i.e. where the detail of the EXIOPOL classification exceeds that of any of the other datasets, including the ‘default’ or representative matrix. This occurs most particularly in the detail sought for in value added, the electricity sector and the refinery products sector. In these cases, where auxiliary data was not readily available (e.g. for refinery products, detailed Australian IO data was included), assumptions were necessarily made on the relative uses of products purely from possible exclusion of flows, and by proportioning aggregate data via gross production. Whether these assumptions should be maintained in the final aggregation is contentious.

A balancing process seeking to make use of such volumes of data is at the mercy of the accuracy of the data. A balancing process can resolve differences in data, but when partial datasets are used, the possibility of leakages is large. Further a strict hierarchy was adhered to, such that default values (representing technological matrices) would be superseded by any of the above datasets. The purpose of this process was to keep transparency between each dataset. Difficulties arose, however, due to the extent of inaccuracies in the source data. This arose on two levels. Firstly, highly different classification structures were reconciled in a manner that was overly specific. This mainly occurred with IEA data both products and industries being inappropriately mapped. Secondly inconsistent price data was used, both between datasets, which could be expected, but also within datasets, such that some subsectors were highly under/over estimated relative to other subsectors. Inconsistent price data affects balances in much the same way as any difference between the values of one dataset and another. The difficulty arises when the inconsistency relates to only a partial subset of data.

One of the larger challenges of the balancing process was to maintain zero flows under highly conflicting data sources without over-constraining the balancing process. Generic methods of dealing with probable zeros by using small values (Robinson et al. 2001) were insufficient to maintain under balancing. Trying to maintain zero values of datasets ended up in over-constraining the problem, i.e. constraints could not be met because of too many zeros. This came about again by too many inaccurate values or mappings from the original datasets to the EXIOPOL classification.

## Conclusion

Over the last 50 or so years, considerable advances have been made in the generic applicability of optimisation process to estimating input-output tables. No longer are input-output tables updated solely by row and column constraints, but significant additional information is used in arriving at estimated tables. Demands on balancing processes have also increased, with more and more information being required to be met. One basis still remains, that the quality of the end product will only be as good as the inputs to a balancing process.

Allen, R. I. G. (1974). "Some experiments with the RAS method of updating input-output coefficients." Oxford Bulletin of Economics and Statistics **36**: 217-228.

Bacharach, M. (1970). Biproportional matrices & input-output change. Cambridge, UK, Cambridge University Press.

Byron, R. P. (1978). "The estimation of large Social Account Matrices." Journal of the Royal Statistical Society Series A **141**(3): 359-367.

Canning, P. and Z. Wang (2005). "A flexible mathematical programming model to estimate interregional input-output accounts." Journal of Regional Science **45**(3): 539-563.

Cole, S. (1992). "A note on a Lagrangian derivation of a general multi-proportional scaling algorithm." Regional Science and Urban Economics **22**: 291-297.

Gilchrist, D. A. and L. V. St Louis (1999). "Completing input-output tables using partial information, with an application to Canadian data." Economic Systems Research **11**(2): 185-193.

Golan, A., G. Judge, et al. (1996). Maximum Entropy Econometrics: Robust Estimation with Limited Data, Wiley.

Junius, T. and J. Oosterhaven (2003). "The solution of updating or regionalizing a matrix with both positive and negative entries." Economic Systems Research **15**(1): 87-96.

Lecomber, J. R. C. (1975). A critique of methods of adjusting, updating and projecting matrices. Estimating and Projecting Input-Output Coefficients. R. I. G. Allen and W. F. Gossling. London, UK, Input-Output Publishing Company**:** 1-25.

Lecomber, J. R. C. (1975). A critique of methods of adjusting, updating and projecting matrices, together with some new proposals. Input-Output and Throughput - Proceedings of the 1971 Norwich Conference. W. F. Gossling. London, UK, Input-Output Publishing Company**:** 90-100.

Lemelin, A. (2009). "A GRAS variant solving for minimum information loss." Economic Systems Research **21**(4): 399 - 408.

Lenzen, M., B. Gallego, et al. (2009). "Matrix balancing under conflicting information." Economic Systems Research **21**(1): 23-44.

McDougall, R. A. (2009). GTAP 7 Data Base Documentation - Chapter 8.D: Disaggregation of Input-Output Tables, Center for Global Trade Analysis.

Morrison, W. I. and R. G. Thumann (1980). "A Lagrangian multiplier approach to the solution of a special constrained matrix problem." Journal of Regional Science **20**(3): 279-292.

Müller, M., I. Pérez Domínguez, et al. (2009). Construction of Social Accounting Matrices for EU27 with a Disaggregated Agricultural Sector (AgroSAMs). I. f. P. T. S. JRC Scientific and Technical Reports, EUR 24010 EN.

Robinson, S., A. Catteano, et al. (2001). "Updating and estimating a Social Accounting Matrix using Cross-Entropy methods." Economic Systems Research **13**(1): 47-64.

Stone, R. and A. Brown (1962). A Computable Model of Economic Growth. London, UK, Chapman and Hall.

Tukker, A., E. Poliakov, et al. (2009). "Towards a global multi-regional environmentally extended input-output database." Ecological Economics **68**(7): 1928-1937.

van der Ploeg, F. (1982). "Reliability and the adjustment of sequences of large economic accounting matrices." Journal of the Royal Statistical Society A **145**(2): 169-194.

Wiedmann, T., R. Wood, et al. (2008). Development of an Embedded Carbon Emissions Indicator – Producing a time series of input-output tables for the UK by using a MRIO data optimisation system, Report to the UK Department for Environment, Food and Rural Affairs by Stockholm Environment Institute at the University of York and Centre for Integrated Sustainability Analysis at the University of Sydney. Ref.: EV02033, July 2008, Defra, London, UK, Executive summary: <http://randd.defra.gov.uk/Document.aspx?Document=EV02033_7333_EXE.pdf>. Full report: <http://randd.defra.gov.uk/Document.aspx?Document=EV02033_7331_FRP.pdf>.

Wood, R. (2010). "Construction, stability and predictability of an input-output time series for Australia." Economic Systems Research: In revision.