INCORPORATING DATA QUALITY IMPROVEMENT INTO SUPPLY-USE TABLE BALANCING

MARTIN C. SERPELL

Department of Computer Science and Creative Technologies, University of the West of England, Coldharbour Lane, Bristol, BS16 1QY, UK

+44(0)117 32 83357

Martin2.Serpell@uwe.ac.uk

Abstract

This paper investigates the benefits of using a boundary tightening algorithm to improve the quality of the data used in Supply and Use Table (SUTs) Balancing, building on similarities with certain approaches to Statistical Disclosure Control. Boundary tightening was shown to significantly improve the quality of the finally balanced SUTs well beyond that of existing techniques. Most notably, improvements occurred when boundary tightening was applied prior to the balancing process - showing that it can be used as a valuable preliminary to other approaches. It also multiplied the improvement in SUTs quality when more accurate updated information was added to the SUTs. The findings of this paper strongly suggest that this boundary tightening algorithm will improve the quality of the output of the balancing process and it is equally likely to be useful when applied to other processes that handle uncertain data.

Keywords: Supply and Use tables; National economic and social accounts; Balancing; improving data quality

1. INTRODUCTION

The National Statistical Agencies collect a wide range of data from a variety of sources to populate the SUTs. The data tends to be collected by industry with a product dimension (e.g. output by product) or a functional heading with a product dimension (e.g. household consumption by functional heading and by product). An overview of the framework is shown in Figure 1. Once the SUTs framework is populated, there is a need to balance three key identities with a time dimension:

- Sum of the industry outputs equals sum of the industry inputs.
- Supply of each product equals use of each product.
- Production based estimate of GVA equals the income based estimate of GVA.

This paper focuses on the first two aspects.

Agencies such as the UK's Office for National Statistics (ONS) practice Statistical Disclosure Control to ensuring the confidentiality of individuals' data within published statistical tables. A common approach is to identify, and then not publish, a "suppression pattern" of cells- chosen so that values that could reveal confidential information are not calculable. To ensure that the vulnerable cells are adequately protected, the resulting tables are attacked using tools that try to calculate any suppressed values. It is this final process that corresponds to boundary tightening.

The following example makes this more concrete. The ONS publishes information on the economy in an aggregated form, typically showing various measures broken down by type of industry and geography. Internally, the data may be held as a three or four dimensional dataset. However, it will be published as a series of interrelated two dimensional statistical tables, with marginal (row and column) totals, as well as the grand total.

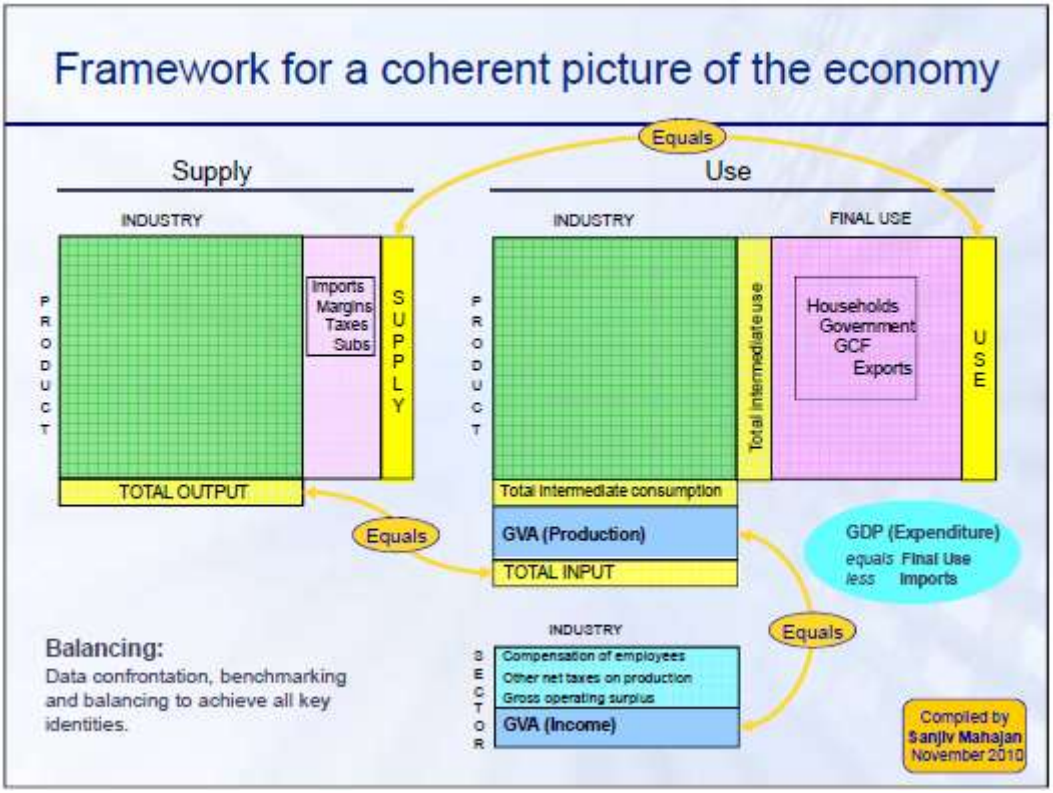

Figure 1: Overview of a Supply and Use Table

A statistical table with marginal totals can be represented as a set of cells, $\mathbf{a}_i$, i = 1...n, satisfying m linear constraint equations such that $\mathbf{Ma} = 0$, where each $\mathbf{M}_{ij}$ has one of the values {0, +1, -1}. These linear constraint equations represent the relationships between the table cells and the marginal totals. If the suppressed cells reside in set **P** then the range of their possible values can be determined by solving the following linear programs.

$$\text{Lower bound } (\mathbf{a}_p) = \min \mathbf{x}_p \text{ such that } \mathbf{Mx} = 0;\ \mathbf{x}_i \geq 0,\ i \text{ in } \mathbf{P};\ \mathbf{x}_i = \mathbf{a}_i,\ i \text{ not in } \mathbf{P}. \quad (1)$$

$$\text{Upper bound } (\mathbf{a}_p) = \max \mathbf{x}_p \text{ such that } \mathbf{Mx} = 0;\ \mathbf{x}_i \geq 0,\ i \text{ in } \mathbf{P};\ \mathbf{x}_i = \mathbf{a}_i,\ i \text{ not in } \mathbf{P}. \quad (2)$$

It was postulated that these linear programs, or something similar, might help with Supply and Use Table balancing (Brodie, 2012). Many of the values within a SUT are not known with certainty but

instead are known to reside between upper and lower bounds with a given level of confidence. Solving the SUTs balancing problem also requires invoking known relationships between cells, which is a clear counterpart to the constraint matrix **M** above. Applying linear programs (1) and (2) to the upper and lower bounds of the SUTs cell values and using the constraint equations that joins them, as with statistical disclosure control, the upper and lower bounds for each cell can be tightened, reducing the size of their confidence interval.

Our hypothesis is that if the SUTs were initialised with the lower and upper bounds within which the "true" cell values were believed to lie, and then the linear programs (1) and (2) were used to tighten those bounds, this would lead to better predictions of the cell values than starting from estimates alone. If this hypothesis is correct, it will both save time (by reducing the amount of balancing that is required) and lead to more accurately balanced tables. To investigate this hypothesis the rest of this paper is set out as follows: Section 2 looks at the background of SUTs balancing, Section 3 the experimental set up used in this investigation, Section 4 the results and Section 5 the conclusions.

2. BACKGROUND

2.1. The Importance of Supply and Use Tables

Supply and Use tables (SUTs), and the Input–Output tables (IOTs) that are derived from them, provide decision-makers with detailed information about the workings of the economy (Temurshoev et al, 2011). This is not only used for economic planning but also to understand pollution (Druckman et al, 2008), sustainability (Dietrich et al, 2012; den Boer et al, 2014) and much more. The more detailed this information is, the better it is for decision making (Wood, 2011). As IOTs are derived from SUTs, the more accurate the SUTs the more accurate the IOT. Collecting the data and constructing the SUTs is both expensive and time consuming. This means that the publication of the current SUTs (and IOTs) cannot take place in the current year but more often in the following year (Lahr and Mesnard, 2004; Temurshoev et al, 2011; Temurshoev and Timmer, 2011).

There has been a move towards providing SUTs (and IOTs) for the separate regions of a nation. Creating regional SUTs (and IOTs) is difficult, and often requires the use of surveys to estimate regional trade flows (Piispala, 1999). The effort is, however, appreciated by many organisations, and is used to improve regional economic development. There has also been a move towards combining national SUTs (and IOTs) to create global and global regional SUTs (and IOTs). This has been driven by globalisation and the need to create even more efficient global value chains (GVCs) (Bo et al, 2013). Such tables typically require the national SUTs (and IOTs) to be harmonised (WIOD, EXIOPOL and IDE-JETRO), but another approach is to balance the combined raw national data (EORA (Lenzen et al, 2013)) in one operation (Tukker and Dietzenbacher, 2013). The demand for both regional, national, global regional and global SUTs (and IOTs) is likely to continue to grow.

Two types of IOTs are produced; "product by product", and "industry by industry". Producing simple versions of either relies on assumptions: the former about technology, and the latter about sales structure. Producing consistent versions of either is more complex, as each table is affected by both technology (and sales structure) assumptions (Smith and McDonald, 2011). A comparison of different techniques used to create IOTs has been carried out by Temurshoev and Timmer (2011). Smith and McDonald (2011) have developed techniques whereby practitioners can create their own technology (and sales structure) assumptions.

2.2. Current Approaches to Supply and Use Table Balancing

The compilation and balancing of SUTs at current prices and in volume terms for a sequence of years will also help to balance the changes in volumes, values and prices in the best possible way and recommended as the best approach for the production of SUTs. Figure 2 shows an overview of the "H-Approach" which is also reflected in the UN Guidelines on Integrated Economic Statistics. The "H-Approach" is the recommended compilation approach which brings these forms together and provides an overview of an integrated SUTs approach as well as linking to IOTs. This is both in current prices and in previous years' prices as well as the links between basic prices and purchasers' prices.

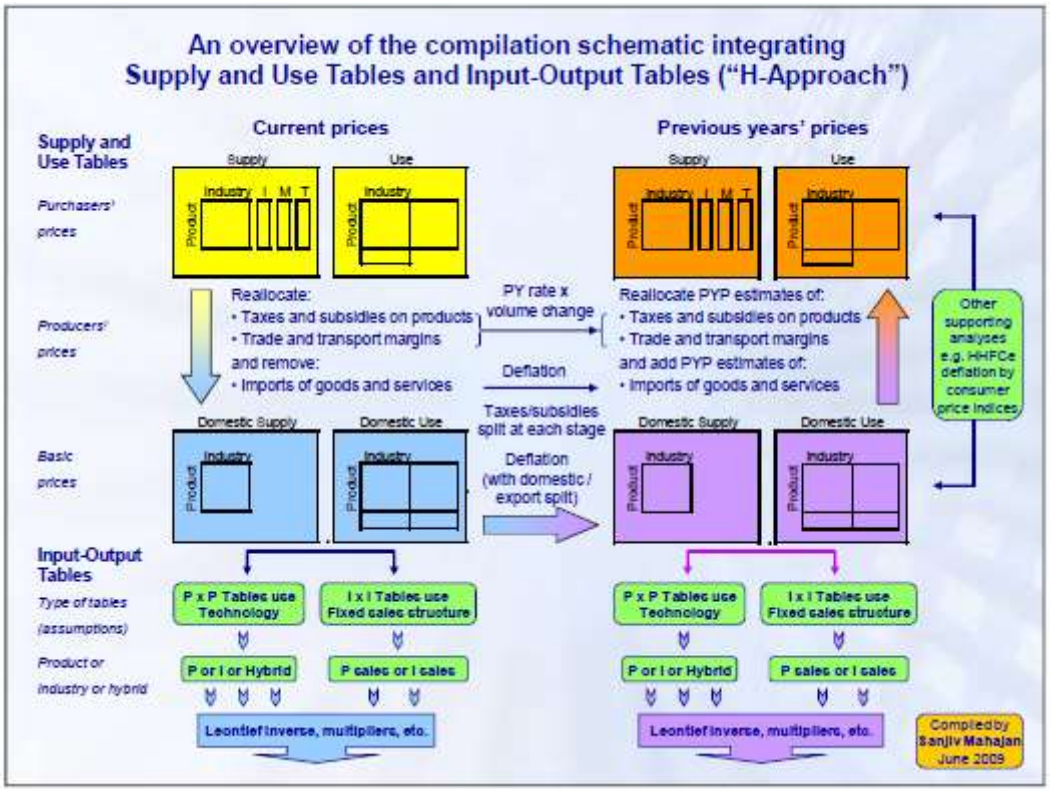

Figure 2: An overview of the compilation schematic integrating SUTs and IOTs

Once initial data has been entered into a SUT from various sources, it must then be balanced, so that all columns and rows sum to their given totals. Many different approaches have been described in the literature. Bi-proportional methods are often used as they are quick and produce reasonable results (Lahr and Mesnard, 2004). The bi-proportional method known as RAS was developed by Stone (1961) and Stone & Brown (1962). In this method the row (column) values are repeatedly adjusted by the ratio of the row (column) total to the sum of the row (column) values. There are many alternatives to RAS, both bi-proportional and optimisation-based methods. All methods tend to minimise the distance between the newly balanced table and the original data prior to balancing. A major failing is that many SUTs balancing methods do not take into account the reliability of the data being balanced. For example, Lahr and Mesnard (2004) stated that a drawback of RAS was that it was not able to take into account the relative reliability of the data within the SUTs being balanced. This problem has since been addressed with the development of the KRAS method by Lenzen et al (2009).

Comparisons of different SUTs balancing methods have been carried out by Lahr and Mesnard (2004), Jackson and Murray (2004), Huang et al (2008) and Temurshoev et al (2011). Lahr and Mesnard (2004) found RAS to perform best, particularly with regard to speed. They also reported a simple way for RAS to deal with fixed cell values within the SUT: by setting their value to zero in the cell and subtracting the true value from the row and column totals. Jackson and Murray (2004) also found RAS to perform best when all the cell values were positive. Junius and Oosterhaven (2003) generalised the RAS procedure (GRAS) so that it could handle negative as well as positive numbers.

Huang et al (2008) reported improvements to GRAS and other balancing techniques. They also reported a simple method for dealing with negative cell values when using a minimising objective function. Temurshoev et al (2011) compared eight different SUTs balancing techniques, and found that GRAS performed equal best with two others. They also reported that, when tested against real SUTs data from Spain, the Euro method used by Eurostat performed badly. Importantly, they noted that comparing different balancing algorithms was problematic as it was not reasonable to use one of the algorithms as a benchmark and no exact (not estimated) data for comparison exists; to carry out their comparisons they used SUTs data from the Netherlands and Spain. This prompted us to develop a SUTs generator with tuneable characteristics, where known "true" cell values permit rigorous algorithm comparisons.

It has been shown that the addition of known information, such as better estimates of cell values due to domain knowledge, or of totals due to aggregation, improves the quality of the SUTs balancing (Lahr and Mesnard, 2004; Temurshoev and Timmer, 2011). Moyer et al (2004) reported that in the United States there is a rich source of data from which to construct SUTs and that this data is ranked by quality so that the highest quality data can be used.

An improvement was introduced by balancing the Supply and Use tables simultaneously rather than separately. This removes the need to adjust them post balancing, in order to make them consistent. This simultaneous balancing producing better results than balancing the tables sequentially. Temurshoev and Timmer (2011) developed such a technique named SUT-RAS which applies a GRAS like algorithm to the combined SUT. In an extension of this idea, National Statistics Agencies in many advanced economies are (or soon will be) expected to publish SUTs consistent both the current and previous years' prices. Nicolardi (2013) has reported that the best way to achieve this is to balance the two simultaneously; however this is complex as the tables are linked by a set of deflators (Ahmed, 1999).

SUTs balancing is done in such a way as to minimise the differences between the same cells of the SUTs from one year to the next (Wood, 2011; Temurshoev and Timmer, 2011) both forward and backwards in time (Lenzen et al, 2012). Published SUTs are regularly revised. For example, the values in a SUT published for the year 2000 would be regularly updated over the following years as more accurate economic information became available. These revisions are caused by factors such as new sources of information, correction of errors, or changes to methodology, industrial/product classifications, international guidelines, etc. This implies that the initially published SUTs represent the least reliable economic information. It also means that any economic forecasting based on the published SUTs will need to be regularly revised. Jacobs and van Norden (2011) developed a state-space model that incorporated measurement errors (news, noise and spill overs) into the revision process that accepted the fact that published estimates may never converge to "true" values.

In reality National Statistics Agencies may not employ fully automated systems when balancing their SUTs. At the UK Office for National Statistics (ONS) balancing is a painstaking process that takes approximately ten staff many weeks. This is because the balancing process typically involves approximately fifty iterations. In each of these, new domain knowledge is added, and the SUTs manually (re)balanced taking into account existing boundary knowledge. When conflicts with input data arise a negotiation is entered into with the data providers. Only when this process has been completed the final phase of balancing is carried out using RAS. In this way ONS ensures that the finally balanced SUTs is of the highest achievable quality.

This paper investigates the impact of adding boundary tightening as a precursor to the SUTs balancing process. In the light of the review above, supply tables and use tables are balanced simultaneously, and the RAS balancing algorithm has been used for its speed and ease of implementation.

3. EXPERIMENTAL SETUP

3.1. Creating Test SUTs

Comparing balancing algorithms is difficult as there are very few SUTs where the "true" cell values are known. For this reason it was decided that synthetic SUTs would be generated with known "true" cell values, known errors and known uncertainties (reliabilities). These SUTs were created in a variety of sizes and with a variety of different cell values. Therefore one SUT may describe an economy of a country that has a large fishing fleet and agricultural business, whereas another might have a large coal and steel industry. In this way the boundary tightening algorithm can be tested on a wide variety of possible SUTs. This ensures that the algorithm does not only work with one given set of data but is generally applicable to any SUT. The structure of the synthetic SUTs is shown in Figure 3.

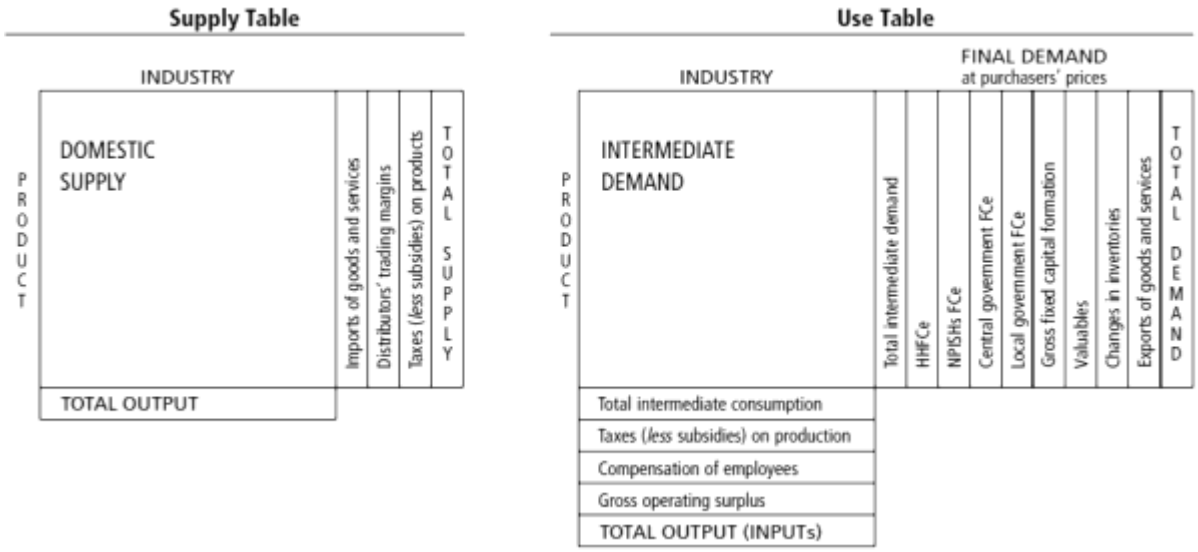

Figure 3: Modelled Supply and Use Table (picture source ONS)

Each cell in a SUT consists of a true value (used to measure estimation error), lower and upper bounds (within which the cell value is believed to reside), an estimated value and an indicator as to

whether the estimated value was allowed to be improved. The tool used to create these SUTs will be made available.

Twenty different SUTs were created with numbers of products and industries drawn randomly between 100 and 150. For each cell in a SUT the following process is used. First, the true value is set to a positive random value. Next the lower (upper) bound is initialised as the true value, minus (plus) some uncertainty, plus a displacement uniformly randomly drawn from the range [-MAXERROR, MAXERROR]. The value of MAXERROR is always less than the uncertainty and therefore the true value always lies between the lower and upper bounds. It should be noted though that this is not a requirement and does not always happen in real data, since the estimates for the bounds may be wrong. Finally, the initial estimated value is set to the half-way point between the lower and upper bounds. The indicator as to whether the estimated value was allowed to be improved is initially set to true. It is used to fix the estimated value so that the different balancing algorithms will not attempt to change it.

3.2. Quality Measures

SUTs error was measured as the sum of the absolute differences between the true and estimated values for all cells in the SUT. SUTs uncertainty was measured as the sum of the upper bounds minus the lower bounds for all the cells in the SUT. Using these SUTs error and uncertainty measurements allowed the performance of different SUTs balancing algorithms to be compared.

3.3. Statistical Tests

Pearson's correlation test was used to test if a relationship exists between two variables; if one does exist then the test will also indicate if it is a positive or negative relationship. The Mann-Whitney U test was used to determine if there is a significant difference in the performance between the different algorithms under test. This test does not rely on the data under test being normally distributed. A good description of these statistical tests can be found in Field (2013).

3.4. The Algorithms to be compared

The six different SUTs balancing approaches consisted of the following components.

1. Estimation, RAS
2. Estimation, RAS with lower and upper bounds (RASwB)
3. Estimation, Add domain knowledge, RAS (DK+RAS)
4. Estimation, Add domain knowledge, RAS with lower and upper bounds (DK+RASwB)
5. Estimation, Boundary tightening, Estimation, RAS with lower and upper bounds (BT+RASwB)
6. Estimation, Add domain knowledge, Boundary tightening, Estimation, RAS with lower and upper bounds (BT+DK+RASwB)

- **Estimation** sets each estimated value to the mean of the cell's lower and upper bounds.
- **RAS** iteratively cycles through rows and columns, updating each estimated value so the rows and columns sum to their totals. RAS was used as it was simple to implement and produces good

results when balancing positive numbers. We implemented a slight modification to allow for fixed cell values in a way similar to Lahr and Mesnard (2004). Cells with fixed values are left out of the adjustment process, their values are not included when the row and column cell values are summed, and are also temporarily removed from the row and column totals.

- RAS with lower and upper bounds (**RASwB**) works similarly to RAS, except that if the new estimated value lies below the lower bound the estimated value is set to the lower bound and likewise for the upper bound. When this happens the SUTs remains unbalanced and so the process was repeated until balance was achieved.
- Adding new domain knowledge (**DK**) to a SUTs cell moves its estimated value, lower bound and upper bound nearer to its true value.
- Boundary tightening (**BT**) only affects the lower and upper bounds; unlike the other balancing components it does not change the estimated value. Boundary tightening works by finding the lower and upper bounds of each SUTs cell by comparing it with the lower and upper bounds of each of the cells in the same row and column. In this investigation boundary tightening was implemented using a shuttle algorithm similar to that described by Dobra and Fienberg (2008). The shuttle algorithm executes considerably faster and can work on considerably larger tables than equivalent mathematical models; however it is not as thorough as the mathematical models when calculating the new lower and upper bounds. When tightening a table cells upper and lower bounds mathematical models act globally, they consider the whole table, whereas the shuttle algorithm acts locally only considering the row and column the cell resides in. This difference has been previously exploited in a pre-processing optimisation applied to the cell suppression problem (Serpell, 2013).

The six different SUTs balancing approaches were chosen to allow any improvement in the balancing process by adding boundary tightening to be evaluated. Comparing RAS with RAS with lower and upper bounds (RASwB) allows any improvement caused by limiting the RAS process such that it does not break the lower and upper limits set on each cell in the SUTs to be evaluated. The improvement in accuracy due to the addition of domain knowledge (DK) can be compared when balancing is carried out with RAS and with RAS with boundary tightening (BT). This allows the improvement caused by boundary tightening (BT) to be evaluated as domain knowledge (DK) is added to the SUT.

3.5. Experimental Procedure

Each of the six algorithms was used to balance the same twenty synthetic SUTs. The initial SUTs error and uncertainty was recorded before balancing and the final SUTs error and uncertainty was recorded after balancing took place. The SUTs error and uncertainty were converted into percentages, by dividing them by the sum of the SUTs cell true values and multiplying by 100, to provide a measurement that was not biased by the size of the values in the SUTs.

Often during the balancing process industry experts will improve on the values that they have estimated for the unbalanced SUT. The impact of this improved domain knowledge can be examined experimentally. The balancing algorithms that included the addition of new domain knowledge were each executed four times with increasing amounts of new domain knowledge added to their domestic supply and intermediate use matrices. The percentages of domain knowledge added were 5%, 10%, 15% and 20%. This allows the effect of adding new domain knowledge whilst balancing to

be measured. This domain knowledge was added by the tightening of the lower and upper bound values associated with given cells in the domestic supply and intermediate use matrices.

3.6. Hardware and Software Used

The experiments were run on a desktop PC which was running the Microsoft Windows 7 operating system. The PC had a four core Intel processor running at 3.30 GHz and 16 GB of RAM installed. The software for the experiment was written in the C programming language and was written to use only one of the processor cores.

4. RESULTS

4.1. Adding Domain Knowledge

During and after the balancing process more accurate information may become available. This information (or domain knowledge) may come from industry or product (domain) experts, from revised statistics produced by other departments or from external bodies. In the case of it being from an industry expert it could be that the expert knows that a particular plant that makes a particular product has recently closed down. The new domain knowledge may apply to a single cell in the SUTs or impact an entire year. The improved domain knowledge will be included in the next balancing round and will improve the accuracy of the balanced SUT.

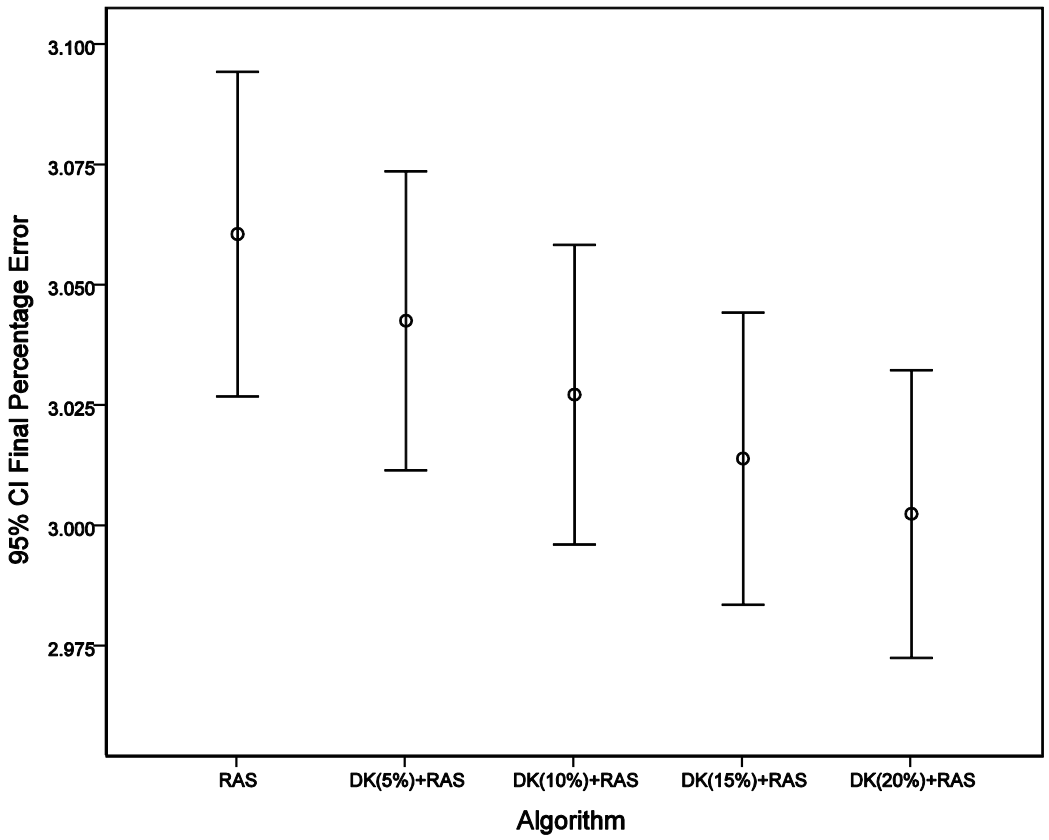

Figure 4: Adding domain knowledge to RAS balancing. Domain knowledge (DK) is shown increasing by 0, 5, 10, 15 and 20%.

Algorithm	Final % Error			Final % Uncertainty		
	Mean	N	Std Dev	Mean	N	Std Dev
RAS	3.061	20	0.072	19.480	20	0.032
DK(5%)+RAS	3.043	20	0.066	19.280	20	0.038
DK(10%)+RAS	3.027	20	0.067	19.091	20	0.044
DK(15%)+RAS	3.014	20	0.065	18.912	20	0.048
DK(20%)+RAS	3.003	20	0.064	18.737	20	0.055

Table 1: Adding domain knowledge (DK) to RAS balancing

Table 1 shows that as domain knowledge was added to the SUTs (5%, 10%, 15% and 20%) the quality of the estimates in the SUTs improved (0.588%, 1.111%, 1.535% and 1.895%). The Pearson's Correlation test showed a weak negative correlation between the final SUTs percentage error and the amount of domain knowledge added to the SUTs (r=-0.300, sig = 0.002). This agrees with the findings of Lahr and Mesnard (2004) and Temurshoev and Timmer (2011). However the improvement in accuracy was not as large as expected and this was probably because the row and column totals were left unimproved which meant that any following balancing simply distributed their error to the other cells in the SUT. Regression analysis for the final SUTs percentage error and the amount of domain knowledge added gave a slope of -0.290, see Figure 4.

4.2. RAS with Lower and Upper Bounds

The Pearson's Correlation test also showed that there was a very strong negative correlation between the final SUTs percentage uncertainty and the amount of new domain knowledge added to the SUTs (r=-0.986, sig < 0.001). This however is to be expected as the addition of new domain knowledge removes uncertainty. Domain knowledge was added to the intermediate use and domestic supply matrices (5%, 10%, 15% and 20%) which lowered the overall SUTs uncertainty (1.027%, 1.997%, 2.916% and 3.814%).

The RAS with lower and upper bounds gave a lower final percentage error in the SUTs than did traditional RAS, see Table 2 and Figure 5. The quality of the estimated values in the SUTs improved by 1.535% when lower and upper bounds were taken into account when doing RAS balancing. This improvement in performance was shown to be significant using the Mann-Whitney U test (Z = -2.137, sig = 0.033).

Algorithm	Final % Error			Final % Uncertainty		
	Mean	N	Std Dev	Mean	N	Std Dev
RAS	3.061	20	0.072	19.480	20	0.032
RASwB	3.014	20	0.054	19.480	20	0.032

Table 2: Comparing traditional RAS and RAS with lower and upper bounds (RASwB)

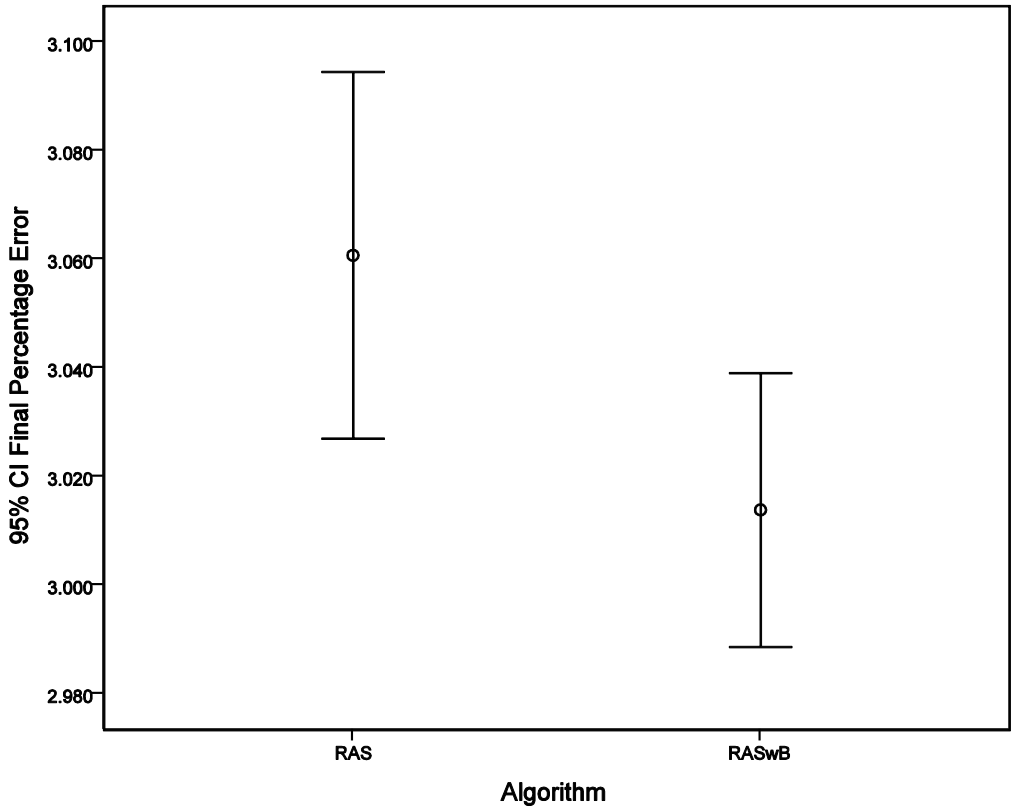

Figure 5: Comparing traditional RAS and RAS with lower and upper bounds (RASwB)

Neither RAS nor 'RAS with lower and upper bounds' changed the uncertainty within the SUTs.

Table 3 shows that, as before, when domain knowledge was added to the SUTs (5%, 10%, 15% and 20%) the quality of the estimates in the SUTs improved (0.597%, 1.161%, 1.692% and 2.157%). This improvement was slightly better than when domain knowledge was added to traditional RAS (0.588%, 1.111%, 1.535% and 1.895%). The Pearson's Correlation test showed a weak negative correlation between the final SUTs percentage error and the amount of domain knowledge added to the SUTs (r=-0.442, sig < 0.001). Again, this agrees with the findings of Lahr and Mesnard (2004) and Temurshoev and Timmer (2011). Linear regression showed that the final SUTs percentage error decreased as domain knowledge was added with a slope of -0.482, this was steeper than when traditional RAS was used (-0.290).

	Final % Error			Final % Uncertainty		
Algorithm	Mean	N	Std Dev	Mean	N	Std Dev
RASwB	3.014	20	0.054	19.480	20	0.032
DK(5%)+RASwB	2.996	20	0.052	19.280	20	0.038
DK(10%)+RASwB	2.979	20	0.054	19.091	20	0.044
DK(15%)+RASwB	2.963	20	0.053	18.912	20	0.048
DK(20%)+RASwB	2.949	20	0.053	18.737	20	0.055

Table 3: Adding domain knowledge (DK) to RAS with lower and upper bounds (RASwB)

The reduction in uncertainty was identical to that of adding domain knowledge to traditional RAS.

4.3. Adding Boundary Tightening

Table 4 shows that the addition of boundary tightening improved the quality of the SUTs by 46.815%. The Mann-Whitney U test has confirmed that the addition of boundary tightening significantly improves the quality of the final balanced SUTs (Z = -5.410, sig < 0.001). The addition of boundary tightening to RASwB greatly reduced the final percentage error in the SUTs, see Figure 6.

	Final % Error			Final % Uncertainty		
Algorithm	Mean	N	Std Dev	Mean	N	Std Dev
RASwB	3.014	20	0.054	19.480	20	0.032
BT+RASwB	1.630	20	0.054	16.148	20	0.256

Table 4: Adding boundary tightening (BT) to RAS with lower and upper bounds (RASwB)

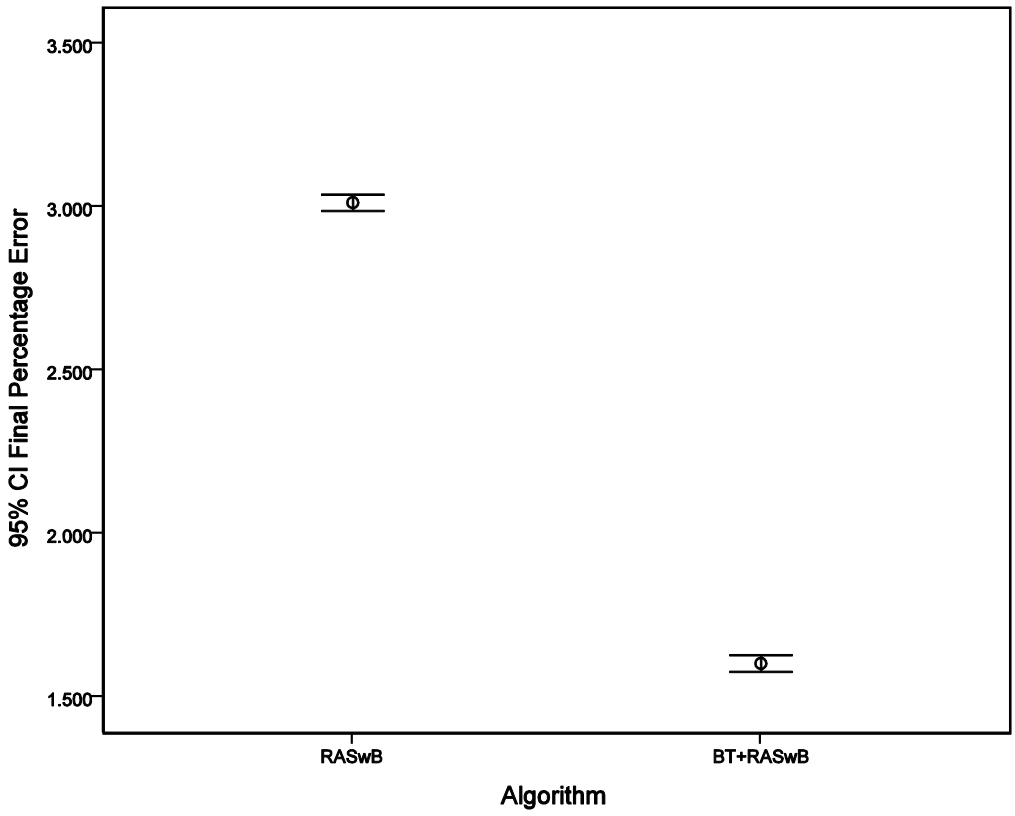

Figure 6: Adding boundary tightening (BT) to RAS with lower and upper bounds (RASwB)

Investigation found that the boundary tightening reduced the uncertainty, almost exclusively, in the column and row totals. This would have led to better estimated values for the row and column totals and this improvement would have been propagated throughout the SUTs by the subsequent balancing process. Boundary tightening has reduced the set of possible solutions to the SUTs balancing process by removing a set of poor quality solutions in which boundary contradictions exist.

The addition of domain knowledge to boundary tightening improves the quality of the balanced SUTs even further, see Table 5. When domain knowledge is added (5%, 10%, 15%, 20%) the quality

of the estimates in the SUTs improves (5.490%, 10.480%, 15.783% and 20.524%). Regression analysis for the final SUTs percentage error and the amount of new domain knowledge added gave a slope of -1.648, see Figure 7. This was much steeper than when new domain knowledge was added without any boundary tightening (-0.290) which implies that the boundary tightening process magnifies the effectiveness of adding domain knowledge. As shown previously, when boundary tightening is used in the absence of the addition of new domain knowledge the lower and upper bounds of the row and column totals are tightened and those of the rest of the SUTs are mostly left unchanged. However, once new domain knowledge was added to the body of the SUTs the boundary tightening had an effect on all of the cells in the SUTs.

	Final % Error			Final % Uncertainty		
Algorithm	Mean	N	Std Dev	Mean	N	Std Dev
BT+RASwB	1.603	20	0.054	16.148	20	0.256
BT+DK(5%)+RASwB	1.515	20	0.057	15.723	20	0.240
BT+DK(10%)+RASwB	1.435	20	0.061	15.292	20	0.245
BT+DK(15%)+RASwB	1.350	20	0.065	14.851	20	0.244
BT+DK(20%)+RASwB	1.274	20	0.069	14.380	20	0.262

Table 5: Adding domain knowledge (DK) to boundary tightening (BT) plus RAS with lower and upper bounds (RASwB)

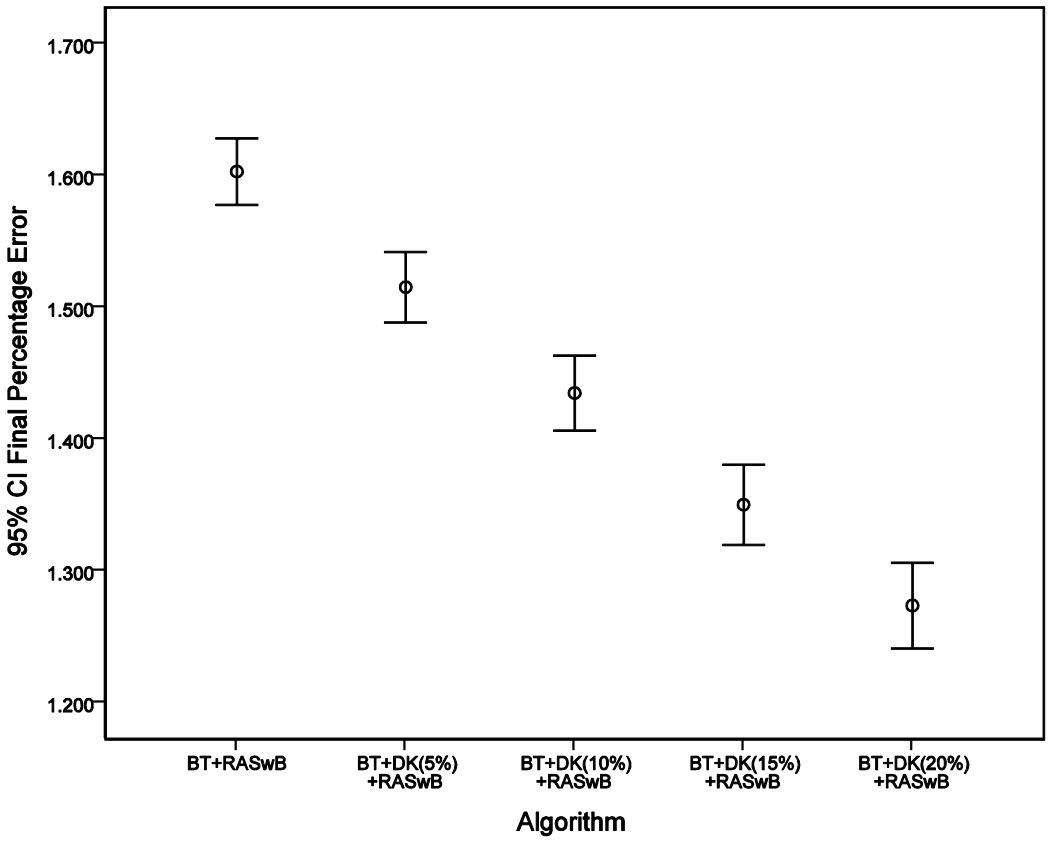

Figure 7: Adding domain knowledge (DK) to boundary tightening (BT) plus RAS with lower and upper bounds (RASwB)

The uncertainty in the SUTs is reduced (2.632%, 5.301%, 8.032% and 10.949%) when domain knowledge was added (5%, 10%, 15% and 20%). The reduction in the amount of uncertainty was much greater than for when balancing was carried out without boundary tightening. Again this implies that boundary tightening magnifies the effect of adding new domain knowledge.

4.4. The Effect of Balancing on Accuracy

By comparing the initial SUTs percentage error prior to balancing with the final SUTs percentage error after balancing the effect of balancing on SUTs accuracy can be measured.

Table 6 shows that with traditional algorithms the error in the SUTs increases when balancing is carried out. This should not be surprising; SUTs are initially filled with the best estimated values that can be provided and are then moved away from those values by the balancing process. In reality the National Statistics Agency will not balance all cells equally; it will preserve those estimated values that it trusts and tend to push more of the balancing onto those cells where the estimated values are less trusted. In this way the Agency will minimise the amount of error that is introduced during the balancing process.

		Initial % Error		Final % Error		% Improvement	
Algorithm	N	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
RAS	20	2.571	0.032	3.061	0.072	-19.047	2.421

Table 6: The effect of balancing on accuracy

4.5. Comparing the Balancing Algorithms

Table 7 shows the effect on SUTs error when different balancing algorithms are applied. Traditional balancing algorithms (RAS/RASwB, DK+RAS,DK+RASwB) increase the SUTs error as described above. In stark contrast to this, when Boundary Tightening is used, balancing the tables reduces their error.

This is because tightening the lower and upper bounds allows more accurate cell estimates. Even when no new domain knowledge was added a large improvement in the quality of the estimated values in the balanced SUTs was achieved using boundary tightening. In-depth analysis revealed that boundary tightening led to improved estimates in the marginal totals, specifically the Supply Table fields 'Total Output' and Total Supply' and the Use Table fields 'Total Intermediate Consumption', 'Total Intermediate Demand', 'Total Outputs' and 'Total Demand'. The balancing process then propagated these improvements throughout all cells in the SUTs. When new domain knowledge was added cell estimates across the whole SUTs improved.

Figure 8 clearly shows that boundary tightening improves the quality of the estimates in the balanced SUTs. It also clearly shows that the addition of new domain knowledge combined with boundary tightening provides the best quality balanced SUTs.

Algorithm	N	Initial % Error		Final % Error		% Improvement	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
RAS	20	2.571	0.032	3.061	0.072	-19.047	2.421
RASwB	20	2.571	0.032	3.014	0.032	-17.224	1.632
DK(5%)+RAS	20	2.571	0.032	3.043	0.066	-18.347	2.191
DK(10%)+RAS	20	2.571	0.032	3.027	0.067	-17.748	2.134
DK(15%)+RAS	20	2.571	0.032	3.014	0.065	-17.232	2.056
DK(20%)+RAS	20	2.571	0.032	3.003	0.064	-16.784	1.994
DK(5%)+RASwB	20	2.571	0.032	2.996	0.052	-16.519	1.597
DK(10%)+RASwB	20	2.571	0.032	2.979	0.054	-15.861	1.661
DK(15%)+RASwB	20	2.571	0.032	2.963	0.053	-15.262	1.668
DK(20%)+RASwB	20	2.571	0.032	2.949	0.053	-14.686	1.676
BT+RASwB	20	2.571	0.032	1.603	0.054	37.646	2.065
BT+DK(5%)+RASwB	20	2.571	0.032	1.515	0.057	41.058	2.160
BT+DK(10%)+RASwB	20	2.571	0.032	1.435	0.060	44.184	2.279
BT+DK(15%)+RASwB	20	2.571	0.032	1.350	0.065	47.481	2.457
BT+DK(20%)+RASwB	20	2.571	0.032	1.274	0.069	50.458	2.643

Table 7: Comparing the balancing algorithms; RAS, RAS with lower and upper bounds (RASwB), with the addition of domain knowledge (DK) and boundary tightening (BT)

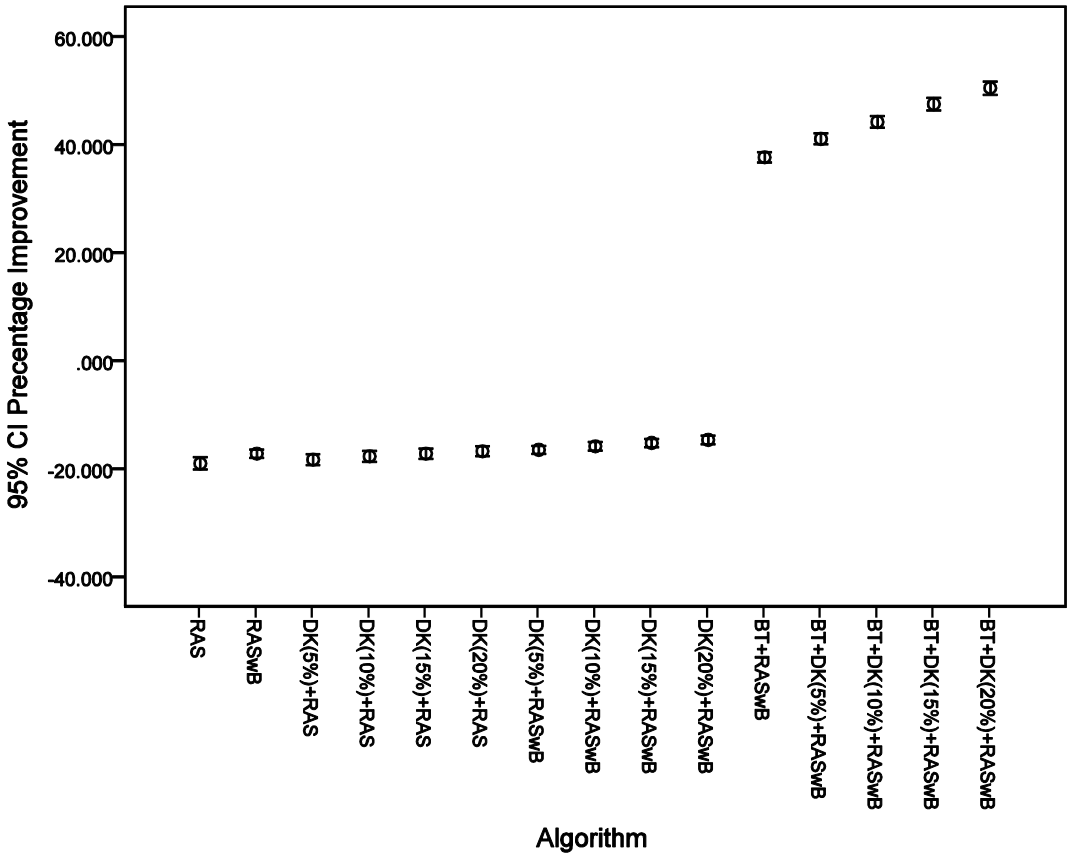

Figure 8: Comparing the balancing algorithms; RAS, RAS with lower and upper bounds (RASwB), with the addition of domain knowledge (DK) and boundary tightening (BT)

5. CONCLUSIONS AND FUTURE WORK

We have demonstrated that the addition of a simple pre-processing step, to improve the quality of the initial data, can lead to massive improvements in the balancing of Supply and Use Tables. We have also described a novel process for creating artificial SUTs that permits us to compare balancing algorithms according to the error they induce.

The addition of boundary tightening to SUTs balancing improved the quality of the final balanced SUTs by between 40 and 50%. This improvement is achieved in less than a second on a standard desktop computer. We also showed that lesser improvements could be achieved via the addition of new domain knowledge and applying lower and upper bound constraints to RAS. Importantly, there was a synergistic effect between boundary tightening and adding domain knowledge.

Experience of using boundary tightening in Statistical Disclosure Control suggests that using the linear programming model would be limited to SUTs no bigger than described in this paper. The variation of the shuttle algorithm used here could easily tighten the boundaries on much larger SUTs, providing the possibility to balance SUTs with much more detailed industry and product information. This algorithm used here is simpler than that described by Dobra and Fienberg (2008) as it only considers the lower and upper bounds on table cells in its calculations and does not go through a process of evaluating candidate table cell values. The algorithm also operates on floating point numbers instead of integers as did the original. There is scope for future improvements to this algorithm, as it acts locally - only considering the row and column each table cell resides in. This leads to some boundary tightening being missed, which a comparable linear programming model would have found.

This boundary tightening procedure will work on SUTs in current prices or on SUTs in previous years prices but will need to be developed further to work on SUTs in current and previous years prices simultaneously. The addition of boundary tightening to the SUTs balancing process was relatively easy as the information that it requires (the lower and upper bounds of the SUTs cell values) was already available. With little effort it may be possible to significantly improve the quality of the SUTs balancing process for real–world SUTs. The future focus of this research will be the application of boundary tightening to real–world SUTs and finding new areas to apply this data quality improvement technique.

6. ACKNOWLEDGEMENTS

The author would like to thank staff at the Office of National Statistics (ONS) in the UK for all the help that they have provided and particularly Pete Brodie for recognising the correspondence between Cell Suppression and Supply and Use Table balancing.

The author would like to thank the anonymous reviewers for their helpful suggestions.

7. FUNDING

This work was supported by the University of the West of England.

References

Ahmad, N. (1999). Experimental Constant Price Input-Output Supply-Use Balances: An approach to improving the quality of the national accounts. *Economic Trends*, (548), 29-36.

Brodie, P. (2012). Personal communication.

den Boer, E., Williams, I. D., Curran, A., & Kopacek, B. (2014). Briefing: Demonstrating the circular resource economy–the Zerowin approach. Proceedings of the ICE-Waste and Resource Management, 167(WR3), 97-100

Dietrich, J., Becker, F., Kast, G., Nittka, T., Kopacek, B., Schadlbauer, S., & Regenfelder, M. (2012, September). Practical demonstrator "ReUse ICT Network". In Electronics Goes Green 2012+(EGG), 2012 (pp. 1-5). IEEE.

Dobra, A., & Fienberg, S. E. (2008). The generalized shuttle algorithm. Algebraic and Geometric Methods in Probability and Statistics, Cambridge University Press.

Druckman, A., Bradley, P., Papathanasopoulou, E., & Jackson, T. (2008). Measuring progress towards carbon reduction in the UK. Ecological Economics, 66(4), 594-604.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage.

Huang, W., Kobayashi, S., & Tanji, H. (2008). Updating an input–output matrix with sign-preservation: some improved objective functions and their solutions. *Economic Systems Research*, *20*(1), 111-123.

Jackson, R., & Murray, A. (2004). Alternative input-output matrix updating formulations. *Economic Systems Research*, *16*(2), 135-148.

Jacobs, J. P., & Van Norden, S. (2011). Modeling data revisions: Measurement error and dynamics of "true" values. *Journal of Econometrics*, *161*(2), 101-109.

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

Piispala, J. (1999). Constructing Regional Supply and Use Tables in Finland. *Statistics*.

Lahr, M., & De Mesnard, L. (2004). Biproportional techniques in input-output analysis: table updating and structural analysis. *Economic Systems Research*, *16*(2), 115-134.

Lenzen, M., Gallego, B., & Wood, R. (2009). Matrix balancing under conflicting information. *Economic Systems Research*, 21(1), 23-44.

Lenzen, M., Pinto de Moura, M. C., Geschke, A., Kanemoto, K., & Moran, D. D. (2012). A Cycling Method for Constructing Input–Output Table Time Series from Incomplete Data. *Economic Systems Research*, *24*(4), 413-432.

Lenzen, M., Moran, D., Kanemoto, K., & Geschke, A. (2013). Building Eora: a global multi-region input–output database at high country and sector resolution. *Economic Systems Research*, *25*(1), 20-49.

Mahajan, S. (2006). Development, compilation and use of input–output supply and use tables. *Economic Trends*, 634, 28–46.

Mahajan, S. (2012). Figures 1 and 2.

Meng, B., Zhang, Y., & Inomata, S. (2013). Compilation and Applications of IDE-JETRO's International Input–Output Tables. *Economic Systems Research*, *25*(1), 122-142.

Moyer, B. C., Planting, M. A., Fahim-Nader, M., & Lum, S. K. (2004). Preview of the comprehensive revision of the annual industry accounts. *Survey of Current Business*, *84*(3), 38-51.

Nicolardi, V. (2011). Supply-Use Tables: Simultaneously Balancing at Current and Constant Prices. A new Procedure. In *19th International Input-Output Conference,* pp. 13-17.

Nicolardi, V. (2013). Simultaneously balancing supply-use tables at current and constant prices: a new procedure. *Economic Systems Research*, *25*(4), 409-434.

Serpell, M., Smith, J., Clark, A. & Staggemeier, A. T. (2013) A pre-processing optimization applied to the cell suppression problem in statistical disclosure control. Information Sciences, 238. pp. 22-32.

Smith, N. J., & McDonald, G. W. (2011). Estimation of Symmetric Input–Output Tables: An Extension to Bohlin and Widell. *Economic Systems Research*, *23*(1), 49-72.

Stone, R. (1961). Input-output and national accounts. Paris, France: *Organisation for European economic co-operation.*

Stone, R., & Brown, A. (1962). A computable model of economic growth (Vol. 1). London: Chapman and Hall.

Temurshoev, U., & Timmer, M. P. (2011). Joint estimation of supply and use tables. *Papers in Regional Science*, *90*(4), 863-882.

Temurshoev, U., Webb, C., & Yamano, N. (2011). Projection of Supply and Use tables: methods and their empirical assessment. *Economic Systems Research*, *23*(1), 91-123.

Tukker, A., & Dietzenbacher, E. (2013). Global multiregional input–output frameworks: an introduction and outlook. *Economic Systems Research*, *25*(1), 1-19.

Wood, R. (2011). Construction, stability and predictability of an input–output time-series for Australia. *Economic Systems Research*, *23*(2), 175-211.