Australian Bureau of Statistics

Making Optimisation Balancing Intuitive

Geoffrey Brent, ABS National Accounts Branch
geoffrey.brent@abs.gov.au

www.abs.gov.au

Background

ABS produces a range of social and economic statistical tables e.g.:

- Economic accounts
- Environmental accounts
- Employment figures
- Population estimates (used to determine electoral representation)
    - etc. etc. etc.

Economic statistics

- Economic game: players produce and use a range of resources.

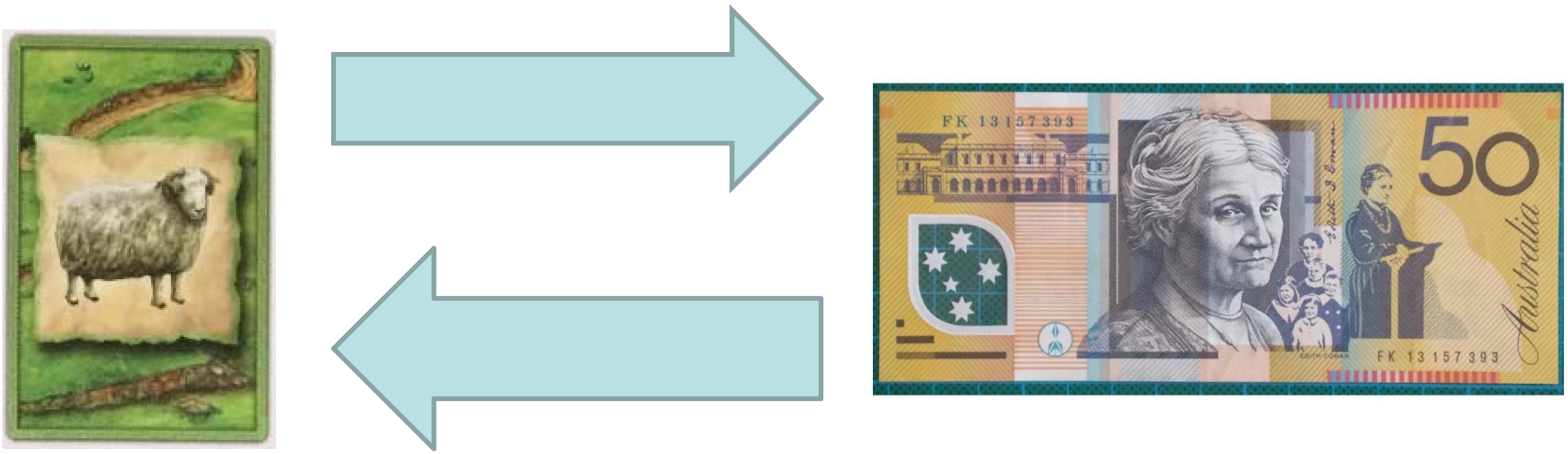


- Players sell the resources they produce to get the resources they need.

Australian Bureau of Statistics

Economic statistics (2)

- Game involves a series of transactions.



Economic statistics (3)

- Economists want to understand:
  - Who's producing what goods & services ("products") over a given period
  - What products they consume to do it.
- Size of economy (GDP) = *net* production.
  - Don't double-count stuff that gets used up making other stuff.
- Divide economy into *sectors* (household, government, industries, …) and *products.*

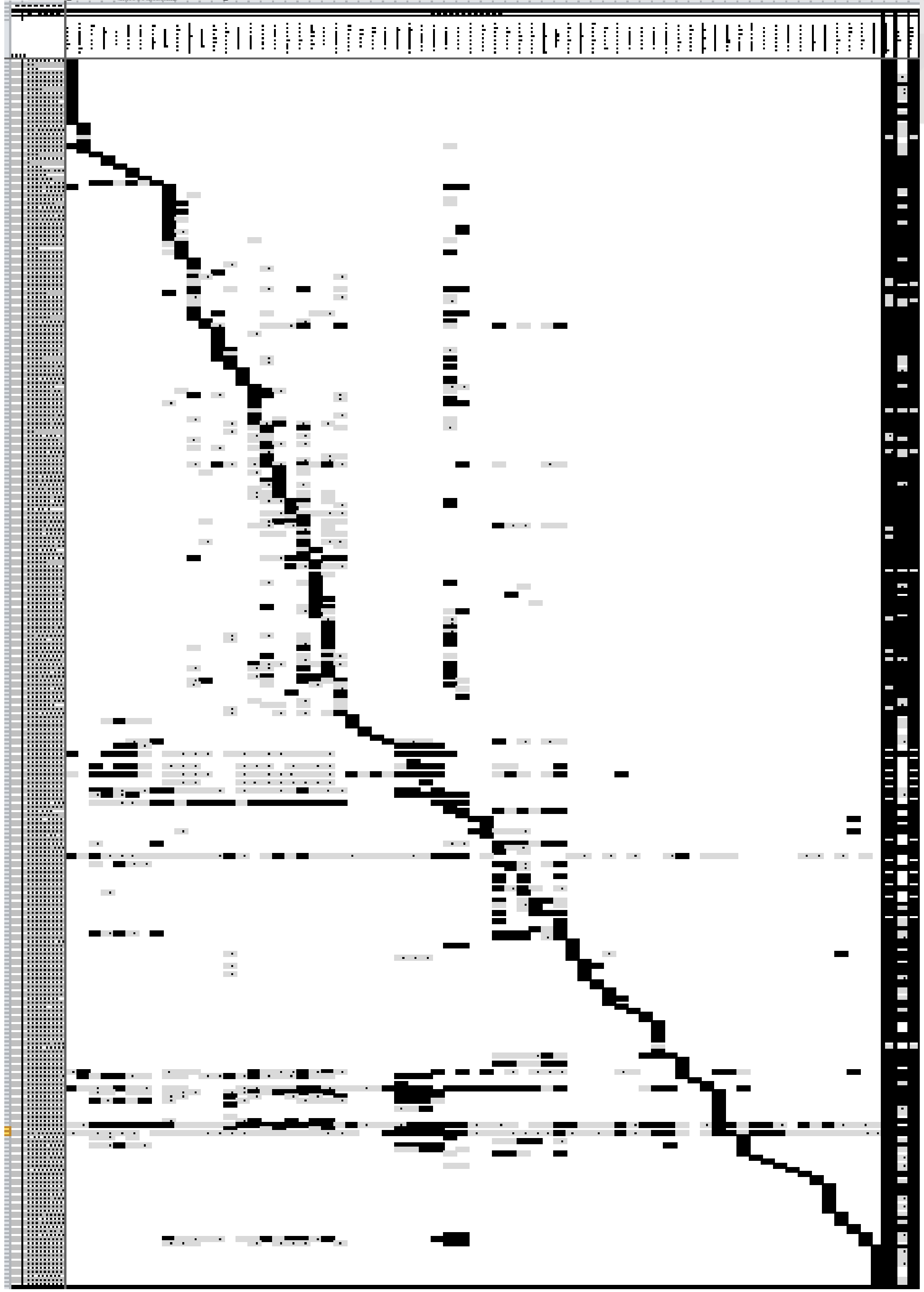
Example: Supply-Use

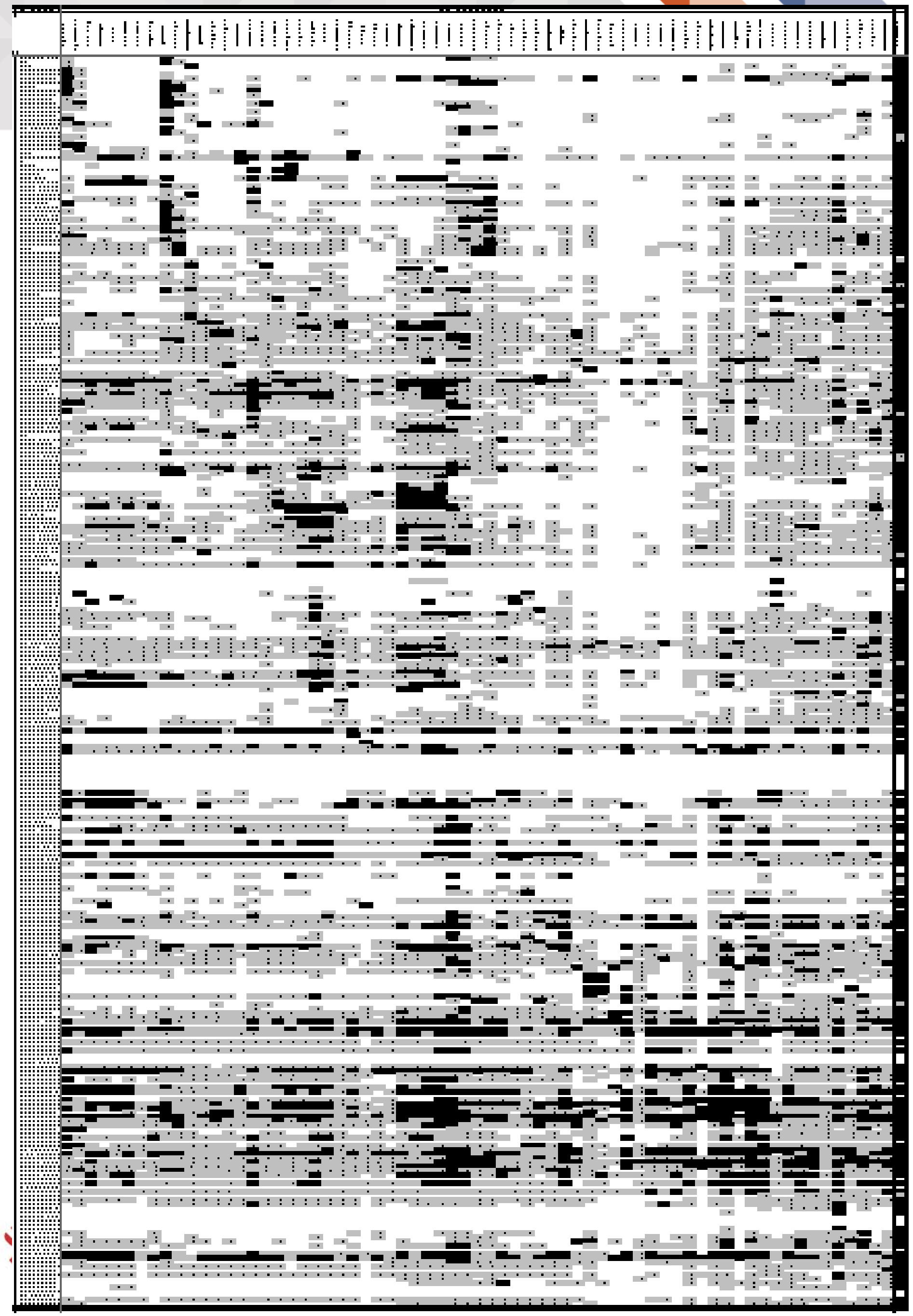
Supply-Use tables provide annual (and quarterly) summary of the Australian economy:

- Measures production and consumption of 301 products by 67 industries, household, and government sectors + exports/imports.
- Used to measure gross domestic product.
- Used as starting point for economic modelling.

Supply-Use (2)

Use ($M)	Agriculture	Food mfg.	Telecommunications	Household sector
Sugar/ confectionery	5	2000	2	5000
Clothing	50	40	40	20000
Petrol	1300	100	900	20000
Financial services	1500	200	500	25000





Multiple sources

The same economic transactions can be measured in several different ways, e.g.:

- Survey households and ask about spending on sugar/sweets.
- Survey retailers and ask about revenue from sugar/sweets.
- Get tax data from ATO and use to estimate sales etc.

Balancing

- Many theoretical identities that *should* hold within these tables.
  - Total value of sugar bought = total value sold.
  - Total sales by retail industry = total costs + profits. etc. etc.
- Sum of Supply rows/columns should match corresponding Use rows/columns.
- Most sources have some degree of error.
- Need to adjust ("balance") for consistency.

Balancing (2)

- Big discrepancies are reviewed and adjusted by experts.
- Infeasible to completely balance via manual processes.
  - Multi-dimensional: balancing a row unbalances columns & v.v.
  - Too big: ~ 100,000 non-zero cells in SU.
- New automated process: quadratic optimisation with AMPL/Gurobi.

Balancing method

$\hat{\boldsymbol{x}}$ = unbalanced (input) data

$\tilde{\boldsymbol{x}}$ = balanced (output) data (DV)

$\Delta\boldsymbol{x}$ = $\tilde{\boldsymbol{x}}$-$\hat{\boldsymbol{x}}$ (balancing adjustment)

S = index set for $\hat{\boldsymbol{x}}$, $\tilde{\boldsymbol{x}}$

Subject to constraints, minimize least-squares objective function:

$$\sum_{i \in S} (\Delta x_i)^2 w_i$$

Weighting

- Most constraints are straightforward additive identities: sum of Row A = sum of Row B etc.
- How do we set the weights?
  - Theory: if we have estimates of variance σ_i^2 for the error on each of our sources, then we should set $w_i = 1/\sigma_i^2$.
  - Usually we don't have these estimates.

Weighting (2)

- Balancing experts know a good outcome when they see it.
- Need to use this experience to design & iteratively improve the OF.
- Want to make weighting & debugging as intuitive as possible.
- Minimise number of design iterations required to get an acceptable OF.
- Want consistency with previous balancing.

Weighting function

General principles:

- Experts rate quality of cells.
- More trustworthy sources should get smaller adjustments.
- Cells with larger values should get larger adjustments.
- Set cell weight as function of magnitude and quality.
- For example…

Weighting function (2)

Other agencies' methods for weighting:

$$1/w_i = |\hat{x}_i|^{\theta} h_i$$

w_i = weight on cell i

$|\hat{x}_i|$ = unbalanced magnitude of cell i

θ = parameter, typically between 0 and 2

h_i = parameter for quality of sources for cell i.

Weighting function (3)

$$1/w_i = |\hat{x}_i|^{\theta} h_i$$

Some questions:

- What should θ be?
- How do we make choice of h_i as meaningful as possible to subject matter experts?

Weighting function (4)

One approach:

- Identify reasonable adjustment magnitude for each cell (SMEs or past data).
- Choose weights that will keep adjustments consistent with these expectations.
  - Easier said than done!
- Supply-Use is large and complex.
  - ~100k cells, each involved in ~ 3 linear constraints and 1 nonlinear.

Weighting function (5)

- System is too complex to quantify exactly how weighting choices will affect adjustments.
  - Depends also on inputs.
- Instead, consider a much simpler system with just one linear constraint…

Simplified problem

Minimise OF:

$$OF(\tilde{\boldsymbol{x}}) = \sum_{i \in S} (\tilde{x}_i - \hat{x}_i)^2 w_i$$

Subject to a single additive constraint:

$$g(\tilde{\boldsymbol{x}}) = \sum_{i \in S} \tilde{x}_i = c$$

Simplified problem (2)

Lagrange multipliers tell us that the solution will satisfy:

$$\left(\frac{\partial OF}{\partial \tilde{x}_1}, \frac{\partial OF}{\partial \tilde{x}_2}, \dots, \frac{\partial OF}{\partial \tilde{x}_N}\right) = \lambda \left(\frac{\partial g}{\partial \tilde{x}_1}, \frac{\partial g}{\partial \tilde{x}_2}, \dots, \frac{\partial g}{\partial \tilde{x}_N}\right)$$

i.e.

$$\Delta x_i = (\tilde{x}_i - \hat{x}_i) = k/w_i$$

for some constant k.

Implications

- Assume this relationship *approximately* holds for the larger problem.
- Implies that $\theta = 2$ will lead to peculiar adjustment behaviour:

$$\Delta x_i \cong k / w_i = k|\hat{x}_i|^2 h_i$$

- Larger values get *quadratically* larger adjustments.
- This has undesirable consequences…

Implications (2)

- If a value of \$10M is adjusted by -\$1M to \$9M:
  - \$90M will be adjusted by -\$81M to \$9M.
  - \$100M will be adjusted by -\$100M to zero.
- This turned out to be a known (but not published) issue for systems using $\theta = 2$.

Implications (3)

- ABS occasionally merges/splits products & industries to reflect changes in structure of economy.
- Suppose we merge “ice cream, vanilla” and “ice cream, other” into single product “ice cream”.
- Using $\theta = 2$, merging these products means higher % adjustment here.
- This is bad – want consistency.

Implications (4)

- Instead, this relationship implies we can use $\theta = 1$ and set h_i to equal expected % adjustment:

$$\Delta x_i \cong k|\hat{x}_i|h_i$$

- Simple to apply and interpret.
- Despite simplifications, this seems to work pretty well in practice.
- Slight modification specific to this problem extends to nonlinear constraints.

Anomaly detection

- Sometimes expectations for data accuracy are unrealistic.
- Want to identify cases where accuracy expectations or input numbers require expert attention.
- Too much data for exhaustive checks – need to filter/prioritise.
- How do we identify “anomalous” adjustments?

Anomaly detection (2)

- Obvious approach, used elsewhere: focus on largest contributors to the objective function:

$$(\tilde{x}_i - \hat{x}_i)^2 w_i$$

- Lagrange-multiplier analysis for simple scenario suggests that this is a bad criterion...

Anomaly detection (3)

- LM approach suggests we should expect adjustments proportional to $1/w_i$.
- Hence expected OF contribution by cell will be proportional to $w_i/w_i^2 = 1/w_i$.
- Hence this approach will emphasise cells with smallest weights and may miss problem cells with larger weights.

Anomaly detection (4)

- LM implies that $(\tilde{x}_i - \hat{x}_i)w_i$ is a better indicator for anomalous adjustments.
- Heat-map plots based on this indicator are very useful in spotting problems.
- Visualising the whole table can help identify patterns of anomalous adjustment…

Anomaly detection (5)

Horizontal/vertical stripes show problems across a product/industry, not just one cell.

Anomaly detection (6)

Pattern of large positive adjustments with one large negative adjustment: probably driven by that exceptional cell.

0	[illegible]	[illegible]	0	0	0	0	0	0	0
0	0	-0.3831	0	0	0	0	0	0	0
0	0	0	0.69869	0	0	0	0	0	0
0	0	0	0.18145	0	0	6.66E-14	0	0	0
0	0	0	0.01664	0	0	0	0	0	1.85E-15
0	0	0	0.55556	0	0.9434	0	0	0	0
0	0	0	0.06452	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0.34483	0	0	0	0	0	0
0	0	0	0.75269	0	0	0.96154	0	0	[illegible]
0	[illegible]	0	0.86587	0	0	0	0	0	0
0	0	0	1.71E-14	0	0	0	0	0	0
0	0	0	0.17241	0	0	0	0	0	0
0	0	0	[illegible]	0	0	0	0	0	0
0	0	0	-1.6087	0	0	0	0	0	0
0	0	0	0.09331	0	0.35714	0	0	0	0.23585
0	0	0	[illegible]	0	0	0	0	0	0
0	0	0	0	0.0281	0	0	0	0	0

Closing notes

- Now evaluating this method to balance data for the 2016-17 financial year.
- Old method requires ~120 staff-weeks of work every year.
- Hoping to cut this by about 75% while improving turn-around time and consistency.

Questions?