Australian Bureau of Statistics

statistics for informed decision making

A macro significance editing framework to detect and prioritise anomalous estimates

Keith Farwell and Paul Schubert

Work Session on Statistical Data Editing,
(Ljubljana, Slovenia, 9-11 May 2011)

The problem

- Macro editing
  - many estimates to quality assure
  - typically a hierarchy of estimates
    - e.g. national, regional, subregional
  - how do we know where to focus our effort?
- Statistical data editing
  - detect
  - resolve
  - treat

Scores used for macro editing

- $Score = \frac{Observed\ estimate - Expected\ estimate}{Scaling\ factor}$
- Comments:

1. two aspects affecting quality of scores
   a. quality of expected estimates
   b. choice of scaling value
2. not a significance score
3. prone to the *size masking effect*
4. not yet taking account of hierarchy of estimates

Significance scores

- $Score = 100 \times \frac{Measure\ of\ predicted\ impact\ of\ editing}{Scaling\ value}$
- *editing impact =*
  Adjusted expected target estimate
  – Expected target estimate

Micro editing example:

$$score = 100 \times \left| w_i \frac{(y_i - y_i^*)}{\hat{Y}} \right|$$

Macro significance scores

- $Score = 100 \times \frac{Measure\ of\ predicted\ impact\ of\ editing}{Scaling\ value\ for\ target\ level}$
- *editing impact =*

  Adjusted expected target estimate

  – Expected target estimate

Macro editing example: Base scores

$$S_{est,base}(Y_i) = 100 \times \frac{\Delta(Y_{i,base})}{Y^*_{i,base}}$$

where

$$\Delta(Y_{i,base}) = Y_{i,base} - Y^*_{i,base}$$

Hierarchical levels and scores

- Say we have three levels of estimates of interest: national, regional, subregional
- We can define
  - base level: subregional
  - target 1 level: regional
  - target 2 level: national
- We can create three scores for each subregional estimate
  - base score: SR score
  - base_target 1 score: SR_Reg score
  - base_target 2 score: SR_Nat score

SR scores

- all scores have the form

$$score = 100 \times \frac{|Current\ estimate - Expected\ estimate|}{Scaling\ value}$$

- for estimates

$$SR\ score = 100 \times \frac{|Current\ SR\ est - Expected\ SR\ est|}{Expected\ SR\ est}$$

- if using the previous estimate as the expected estimate, then

$$SR\ score = 100 \times \frac{|Current\ SR\ est - Previous\ SR\ est|}{Previous\ SR\ est}$$

Hierarchical scores

- Use previous regional and previous national estimates as expected target 1 and expected target 2 estimates:

$$SR\_Reg\ score = 100 \times \frac{|Current\ SR\ est - Previous\ SR\ est|}{Previous\ Reg\ est}$$

$$SR\_Nat\ score = 100 \times \frac{|Current\ SR\ est - Previous\ SR\ est|}{Previous\ Nat\ est}$$

- We have a three-level hierarchy
  - three scores are produced
  - three cutoffs are needed

Hierarchical estimate and ratio scores

- The hierarchical estimate score is:

$$S_{est,base,target}(Y_i) = 100 \times \frac{\Delta Y_{i,base}}{Y^*_{i,target}}$$

- The hierarchical ratio score is:

$$S_{ratio,base,target}(R_{i,j}) = 100 \times \frac{R^*_{i,j,target|base} - R^*_{i,j,target}}{R^*_{i,j,target}}$$

with expected target ratio

$$R^*_{i,j,target} = \frac{Y^*_{i,target}}{Y^*_{j,target}}$$

and adjusted expected target ratio

$$R^*_{i,j,target|base} = \frac{Y^*_{i,target} + \Delta Y_{i,base}}{Y^*_{j,target} + \Delta Y_{j,base}}$$

Example application

- ABS Agricultural collection
- use previous estimates as expected estimates
- Subregion: Statistical Division (SD) – 1646 estimates
- Region: State – 290 estimates
- National: Australia – 49 estimates
- Calculate the three scores with SD as base level, State and Aust as target 1 and target 2 levels

Exclude extreme values

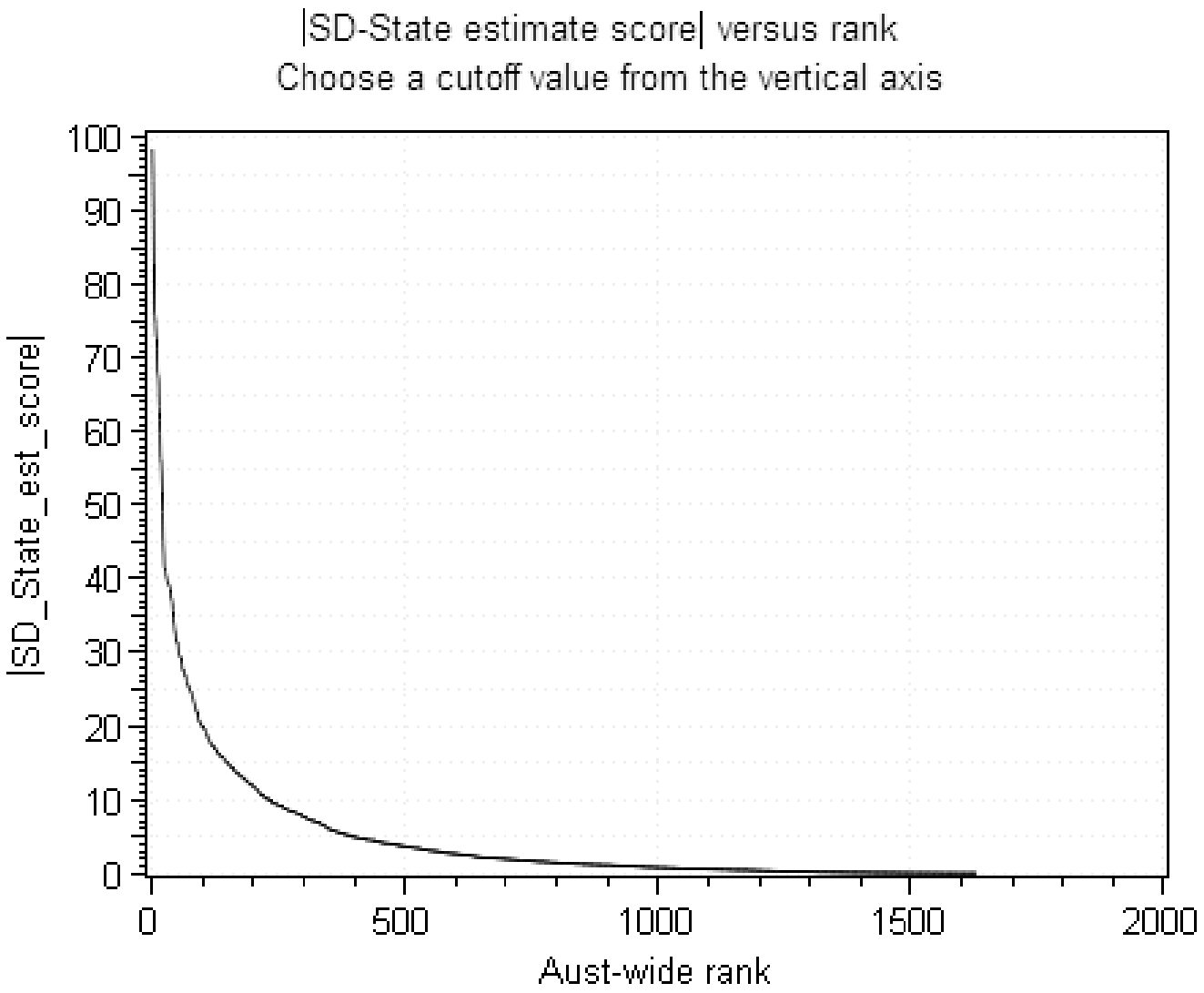
Count of absolute SD-State scores > 100 %

count	Frequency
1	16

absolute SD-State estimate scores > 100 %
These have been excluded from the SD-State estimate score graph in order to make the graph more readable

Obs	item	state	abs_sd_state_est_score1
1	4304603	1	16958.71
2	4304603	5	7614.14
:	:	:	:
15	1510801	8	115.00
16	1500801	3	110.89

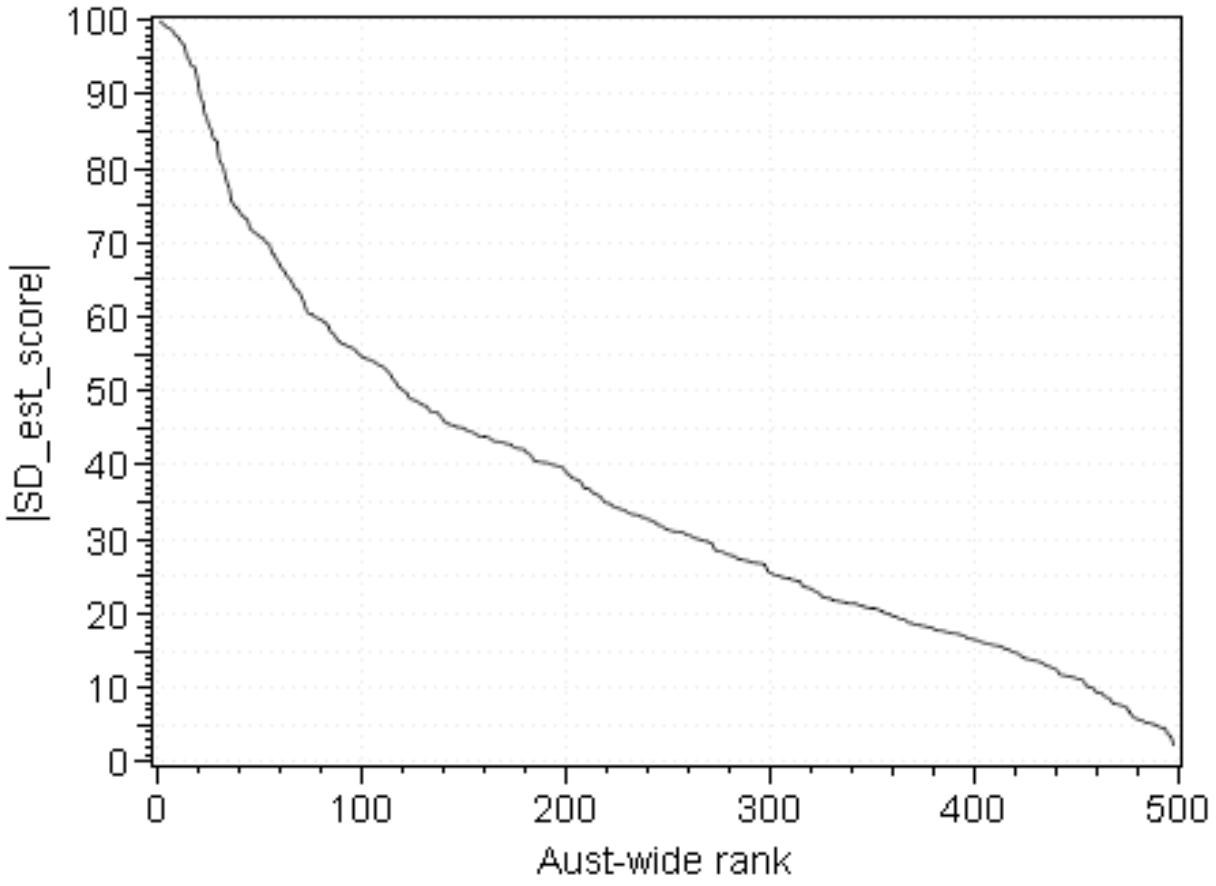
Select SD_State cutoff



Selecting SD cutoff

statistics for informed decision making

|SD estimate score| versus rank
for scores with |SD-State estimate score| > 1.75 % and |SD-Aust estimate score| > 0.25 %
absolute SD estimate scores > 100% have been excluded to enhance readability
Choose an SD estimate score cutoff value from the vertical axis



Results

Hierarchical macro editing categories	Number of SD estimates	Number of anomalous estimates
(0,0,0)	367	.
(0,0,1)	407	.
(0,1,0)	42	.
(0,1,1)	135	.
(1,0,0)	61	.
(1,0,1)	66	.
(1,1,0)	80	.
(1,1,1)	493	493
Total	1,651	493

Cut-offs:
|SD_Aust estimate score| > 0.25
|SD_State estimate score| > 1.75
|SD estimate score| > 15.0

Australian Bureau of Statistics

More information

statistics for informed decision making

- ABS Methodology Advisory Committee paper (longer version of this work session paper)
  - abs.gov.au, select ‘Methods and Standards’ page
- email
  - keith.farwell@abs.gov.au
  - paul.schubert@abs.gov.au