Australian Bureau of Statistics

1352.0.55.104

Research Paper

The Use of Scores to Detect and Prioritise Anomalous Estimates

New Issue

Research Paper

The Use of Scores to Detect and Prioritise Anomalous Estimates

Keith Farwell

Tasmanian Methodology Unit
Statistical Services Branch

Methodology Advisory Committee

6 November 2009, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) TUE 23 NOV 2010

ABS Catalogue no. 1352.0.55.104

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Keith Farwell, Statistical Services Branch on Hobart (03) 6222 5889 or email <statistical.services@abs.gov.au>.

THE USE OF SCORES TO DETECT AND PRIORITISE ANOMALOUS ESTIMATES

Keith Farwell
Tasmanian Methodology Unit

QUESTIONS FOR THE COMMITTEE

1. Does the Committee have any general views on:
   - (i) the proposal to not develop a movement score;
   - (ii) the usefulness of sensitivity measures;
   - (iii) two-sided versus one-sided cut-offs (and is one-sided suitable for significance editing);
   - (iv) the incorporation of Hidiroglou–Berthelot macro-edits into ABS macro-editing tools;
   - (v) appropriate methods to analyse the effectiveness of macro-edits; and
   - (vi) the general elements of the proposed macro significance editing framework?
2. Does the Committee wish to make specific comments on:
   - (i) the general definition of significance and macro-editing impact;
   - (ii) what would be good scaling values (e.g. should we use expected standard errors as default scaling values for estimate scores?);
   - (iii) the hierarchical macro-edit approach;
   - (iv) the applicability of ellipsoidal distance for combining scores;
   - (v) the usefulness of the Hidiroglou–Berthelot edit variants explored in this paper; and
   - (vi) the performance of the hierarchical macro-edits compared with the Hidiroglou–Berthelot macro-edits?
3. Does the Committee have any general observations or advice (such as areas to explore or develop)?

CONTENTS

THE USE OF SCORES TO DETECT AND PRIORITISE ANOMALOUS ESTIMATES

Keith Farwell
Tasmanian Methodology Unit

ABSTRACT

This paper provides an overview of scores used in macro-editing and presents some new scores based on significance criteria. Problems with macro-editing scoring methodologies due to the effect of swamping and masking are discussed. A review of the well-known Hidiroglou–Berthelot edit is provided within a significance editing context. After a brief summary of work done by the U.S. Census Bureau on several score-based methods, the paper introduces the concept of significance for macro-editing and outlines a macro significance editing framework based on an extension of the existing micro significance editing framework used within the Australian Bureau of Statistics (ABS). Some results from empirical comparisons between a proposed macro significance editing application called *hierarchical macro-editing* and several variants of the Hidiroglou–Berthelot macro-edit are discussed. The paper finishes with a summary of findings and recommendations for developing score-based macro-editing for business surveys conducted by the ABS.

1. INTRODUCTION

The ABS Editing Guide (ABS, 2007) defines editing as the activity aimed at detecting, resolving, and treating anomalies in data to help make the data 'fit for purpose'. Micro-editing involves the editing of collection inputs such as unit records (or micro-data). The micro-data are made fit for purpose by reducing errors in the reported data. The first task involves selecting micro-data considered to be anomalous. More specifically, it involves finding unit record values which appear to be erroneous. The next step involves determining if each anomalous value is, in fact, erroneous. If the value is erroneous, the last step involves taking a course of action to correct the error. The typical action is to replace the erroneous value with a more acceptable value using manual or automatic techniques. In any case, the last step includes documenting the decisions and actions taken (including recording that data failed a micro-edit but was found to be correct and left unmodified).

Macro-editing involves the editing of collection outputs such as estimates, ratios of estimates, and standard errors (or macro-data) rather than the editing of unit records. The macro-data are made fit for purpose by either correcting questionable macro-data or validating and explaining it. Accordingly, the first step in macro-editing involves detecting anomalous estimates (which includes estimates of standard error) rather than anomalous unit records.

As with micro-editing, the second step involves determining the nature of the anomaly. The second step is more complex for macro-editing than it is for micro-editing. The kinds of anomalies found with estimates differ greatly from those found with unit records and the causes can be far more varied and complex. A unit record is anomalous if the reported data appears to be incorrect whereas an estimate appears to be anomalous if it does not accord sufficiently to expectations. For the micro-editing case, the incorrect reported data is dealt with by correcting it. For the macro-editing case, a questionable estimate could be affected by processing or estimation errors, important reported data errors, the presence of outliers, etc.; or it could be correct and requires justification. The macro-editor must firstly determine whether the anomaly is the result of processing and estimation errors or reported data errors. Processing and estimation problems can have many causes. Some examples include problems with inappropriate processing flags and codes, weighting errors, the impact of outliers, incorrect input files, missing strata, faulty frames, incorrect benchmarks, poor or incorrect imputation, unacceptable response rates and death rates, and inappropriate macro-data adjustment factors. Macro-editors need to think about what is happening in the data. If no processing or estimation errors are found, macro-editing attention tends to turn towards the micro-data. In this sense, macro-editing involves a component of micro-editing. However, if macro-editors investigate micro-records before checking for the presence of processing and estimation errors, there is the risk that editors will spend too much time checking unit records and lose the macro-editing focus.

The third step involves amendment to data or processes as required and associated documentation. This may include accepting the anomalous estimate as correct and documenting the justification.

The ABS is looking to introduce more objectivity into the macro-editing process. One area of interest is the use of scores for detecting anomalous estimates. A score-based anomalous estimate detection process will add rigour and repeatability to the overall anomalous estimate detection process (which may also contain a subjective detection element).

The ABS has built a tool which uses scores to detect and prioritise anomalous unit record data, called the *Significance Editing Engine* (SEE), which is used for micro-editing business survey data (Farwell, 2004; Farwell, 2005; Australian Bureau of

Statistics, 2011). This paper looks to extend the significance editing concepts currently used in the detection phase of micro-editing to the detection phase of macro-editing.

A measure of significance for macro-editing is used to develop a *macro significance score*. Each estimate within a domain of study can be ordered and ranked by score size. A cut-off method can be applied to the distribution of scores to divide the estimates into those considered acceptable and those considered anomalous. The higher the score, the more likely it is that the estimate may have been affected by important processing or estimation errors, important reported data errors, or the presence of outliers.

A macro significance editing approach has the advantage that it is based on similar concepts to those currently used in micro significance editing. It requires the identification of anomalous data (in this case, macro-data) through the calculation of scores, the ranking of the anomalous data by score size, and the application of editing cut-offs (that is, an editing cost-benefit analysis) based on comparisons of observed data with editor expectations of them.

This paper commences with an overview of basic scores in Section 2 highlighting their relationship to the general form of a significance score. Problems with these scores, when they are used as stand-alone scores for macro-editing, are discussed. An effective macro-editing scoring methodology needs to be able to deal with these problems and Section 3 outlines a score developed by Hidiroglou and Berthelot (1986) which was designed to address them. Section 4 provides a brief summary of a series of investigations by the U.S. Census Bureau on several score-based anomalous estimate detection methods (Sigman, 2005; Thompson, 2007; and Thompson and Ozcoskun, 2007). Section 5 introduces the concept of significance for macro-editing and Section 6 extends the existing micro significance editing framework into a framework which can cover macro-editing. Various new macro significance editing scores and applications are suggested including a new method called *hierarchical macro-editing*. Section 7 presents some empirical comparisons between hierarchical macro-editing and several variants of the macro-edit developed by Hidiroglou and Berthelot. Section 8 concludes the paper with a summary and recommendations.

2. OVERVIEW OF SCORES USED FOR MACRO-EDITING

Many basic scores that have been used in macro-editing have the following general form:

$$\text{Score} = \frac{\text{Observed estimate} - \text{Expected estimate}}{\text{Scaling value}} \tag{1}$$

Various types of scores can be derived from (1) by substituting different choices for the expected estimate and the scaling value. Table 2.1 below displays some examples.

2.1 Examples of basic macro-editing scores

Score	*Expected estimate*	*Scaling value*
Percentage movement	Previous estimate	Previous estimate
Z-score	Mean of the estimates	Standard deviation of the estimates
Non-parametric version of the Z-score	Median of the estimates	Interquartile range of the estimates
'Estimate' score	Mean of the estimates	Mean of the estimates
Non-parametric version of estimate score	Median of the estimates	Median of the estimates

Note that the term 'estimate' in table 2.1 may also include a rate (calculated as a ratio of two estimates of total) or an estimate of standard error (or coefficient of variation for a census). If the estimates are rates in table 2.1, we obtain many typical *ratio scores* (which are a popular choice for macro-editing scores). This paper will refer to the varieties of scores as:

(i) *estimate scores* (for estimates of total);

(ii) *ratio scores* (for ratios of two estimates of total);

(iii) *standard error scores* (for estimates of standard error); or

(iv) *coefficient of variation scores* (for coefficients of variation in censuses).

In fact, if we use a relationship between two estimates of total to form an expected estimate and use the expected estimate as the scaling value in (1), the basic estimate and ratio scores are identical. To demonstrate, let $R = Y/Z$ (where Y and Z are estimates of total) and let R^* be the expected value for R. We calculate the expected value for Y with $Y^* = R^*Z$. The score for an estimate of total is:

$$\text{Estimate score} = \frac{Y - Y^*}{Y^*} = \frac{Y - R^*Z}{R^*Z} = \frac{R - R^*}{R^*} = \text{Ratio score}$$

The ratio scores can be for ratios of current and previous estimates for the same variable (called *historical* ratios) or ratios of two different variables from the same collection (called *current* ratios). Regardless of the type of score, there are two major aspects that fundamentally affect the quality of scores based on (1) which are:

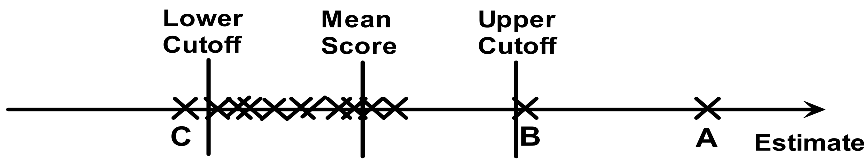
(i) the quality of the expected estimates; and

(ii) the choice of scaling value.

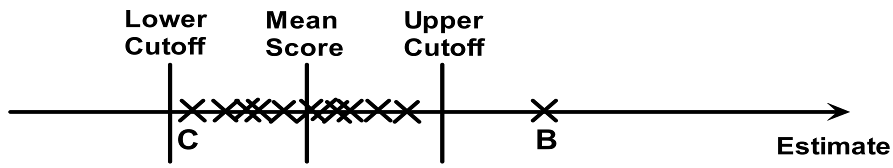
From a technical viewpoint, the anomaly identification process can be subject to two kinds of identification errors sometimes referred to as *swamping* and *masking*. Swamping is said to occur when estimates which are not anomalies are declared as anomalies. Masking is said to occur if actual anomalies are not detected as anomalies. For further details refer to Gather and Becker (1997); Samprit, Hadi, and Price (1999); and Maimon and Rokach (2005).

Consider the following two sets of examples of swamping and masking. In figures 2.2(a) and 2.2(b) below we apply a Z-score approach. Estimates are defined as anomalous if they fall outside the upper or lower cutoffs. In figure 2.2(a), we select A, B and C as anomalous. In figure 2.2(b), we remove A and repeat the process resulting in the non-selection of C. We can say that swamping occurred in figure 2.2(a) because the selection of C was a false negative decision due to the influence of A on the mean and standard deviation.

2.2(a) Initial distribution of estimates

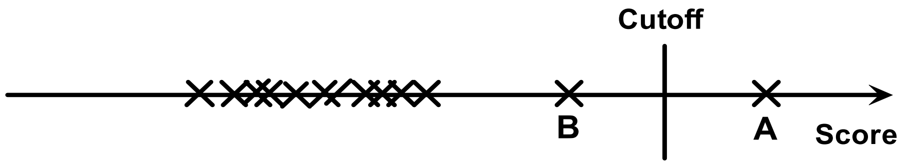


2.2(b) Distribution of estimates following removal of A

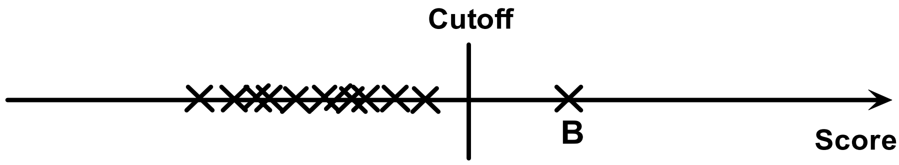


In figures 2.3(a) and 2.3(b) below we choose a one-sided cutoff based on score size and score distribution. We select estimates with large scores which differ markedly from the other scores. In figure 2.3(a), we might select A as anomalous but accept B. In figure 2.3(b), A is removed and we reassign a cut-off resulting is the selection of B. We can say that masking occurred in figure 2.3(a) because B was not selected in the presence of A.

2.3(a) Initial distribution of scores



2.3(b) Distribution of scores following removal of A



Swamping and masking are concepts that have a clear meaning when used in a strict technical context but these concepts can be less clear when used in a macro-editing context. However, this paper will use these terms since they indicate a basic set of problems which would be tedious to describe on a case-by-case basis.

Scores using means and standard errors are prone to swamping caused by the presence of extreme values. Scores based on methods resistant to extreme values such as medians and quartiles can be prone to swamping and masking when very asymmetric distributions of estimates are involved. Sometimes, a transformation of the estimates prior to scoring them may alleviate the problems. However, the transformation needs to be carefully assessed prior to applying it and this makes the use of transformations difficult to manage.

Hidiroglou and Berthelot (1986) point out, within a micro-editing context, that Sugavanam (1983) found that the variability of historical ratios (defined as the ratio of the current and previous reported value for a unit) is greater for small businesses than for larger businesses. This has also been observed in ABS business data and the same phenomenon occurs with estimates. The variability of historical ratios for ‘small’ domains is greater than the variability of those for ‘large’ domains. The distribution of historical or current ratios is often skewed with a long right tail containing many large ratios for estimates from small domains. In fact, the distribution of any basic score derived from (1), when the expected estimate is also the scaling value, will suffer from the same problem since the distribution of such scores will be the same as the distribution of ratios of observed and expected estimates. Such scores result in an anomaly identification procedure that tends to select too many small estimates and not enough large estimates. Hidiroglou and Berthelot (1986) call this the *size masking effect*. Percentage movements of estimates of total and percentage movements of ratios of estimates of total are typical examples of scores affected by size masking.

3. THE HIDIROGLOU–BERTHELOT (H–B) MACRO-EDIT

Hidiroglou and Berthelot (1986) developed a scoring and cut-off approach, called the *H–B macro-edit* in this paper, which was designed to address the problems outlined in Section 2. Although the original H–B edit was designed for micro-editing, the overall approach can be applied to macro-editing by substituting estimate values for unit record values. The macro-editing version of the H–B score can be applied to current or historical ratios but the ratios must be strictly positive or strictly negative. It is useful for the development of this paper to review the development of the H–B score.

Starting with a basic ratio score using the median ratio as both the expected and scaling value in (1), Hidiroglou and Berthelot attend to swamping and masking problems by introducing three key steps. The first two steps involve transformations which Sigman (2005) calls a *centering* transformation and a *magnitude* transformation. The centering transformation is applied to the ratios to even out the differing lengths of the tails of the ratio distribution. The magnitude transformation is then applied to the scores for the 'centered' ratios to control the impact of the size masking effect. The third step involves the use of dynamic two-sided cut-offs based on non-parametric measures.

For simplicity, we outline the H–B macro-edit development using historical ratios. Let $Y_{i,d,t}$ and $Y_{i,d,t-1}$ be estimates of total for variable i within domain d for period t such that $Y_{i,d,t} > 0$ and $Y_{i,d,t-1} > 0$.

$$R_{i,d} = \frac{Y_{i,d,t}}{Y_{i,d,t-1}}$$

is the *historical* ratio for estimate Y_i within domain d. If we use the median of the historical ratios (within domain d) as the expected ratio and as the scaling value in (1), we obtain the following initial ratio score:

$$S^{init}(R_{i,d}) = 100 \times \frac{R_{i,d} - \text{median}(R_{i,d})}{\text{median}(R_{i,d})} \tag{2}$$

which tends to be skewed, so Hidiroglou–Berthelot apply the following centering transformation to the original ratios:

$$S^*_{\text{HB}}(R_{i,d}) = \begin{cases} 100 \times \dfrac{R_{i,d} - \text{median}(R_{i,d})}{R_{id}} & \text{if } 0 < R_{i,d} < \text{median}(R_{i,d}) \\ 100 \times \dfrac{R_{i,d} - \text{median}(R_{i,d})}{\text{median}(R_{i,d})} & \text{if } R_{i,d} \geq \text{median}(R_{i,d}) \end{cases} \tag{3}$$

Score (3) transforms the ratios to ensure that anomalous ratios are selected from both sides of the distribution when using a cutoff methodology based on *asymmetric fences* which is defined in (5) below. Refer to Thompson (1999) for more details concerning asymmetric fences. Hidiroglou and Berthelot found that the centered score $S^*_{\mathrm{HB}}(R_{i,d})$ tends to be prone to swamping due to too many estimates for large domains receiving high scores. In order to control the importance associated with the size of estimates, Hidiroglou and Berthelot apply a multiplicative adjustment (or magnitude transformation) to $S^*_{\mathrm{HB}}(R_{i,d})$ to create the final H–B score:

$$S_{\mathrm{HB}}(R_{i,d}) = S^*_{\mathrm{HB}}(R_{i,d}) \times \max\left(Y_{i,d,t}, Y_{i,d,t-1}\right)^U \tag{4}$$

where U $(0 \le U \le 1)$ is used to 'tune' the score by placing more importance on a small change associated with a large estimate pair compared to a large change associated with a small estimate pair. The magnitude adjustment can range from no effect (when $U = 0$) to maximum effect (when $U = 1$).

Hidiroglou and Berthelot apply the following two-sided dynamic cut-offs (based on asymmetric fences) to the final scores:

$$\begin{aligned} \text{Upper cutoff} &= \mathrm{median}\left(S_{\mathrm{HB}}(R_{i,d})\right) + \alpha \max\left(D_{\mathrm{Q3S}}, \beta\, \mathrm{median}\left(S_{\mathrm{HB}}(R_{i,d})\right)\right) \\ \text{Lower cutoff} &= \mathrm{median}\left(S_{\mathrm{HB}}(R_{i,d})\right) - \alpha \max\left(D_{\mathrm{Q1S}}, \beta\, \mathrm{median}\left(S_{\mathrm{HB}}(R_{i,d})\right)\right) \end{aligned} \tag{5}$$

where $D_{\mathrm{Q1S}} = \mathrm{median}\left(S_{\mathrm{HB}}(R_{i,d})\right) - SQ1$

and $D_{\mathrm{Q3S}} = SQ3 - \mathrm{median}\left(S_{\mathrm{HB}}(R_{i,d})\right)$

$SQ1$ is the 25th percentile of the final scores;

$SQ3$ is the 75th percentile of the final scores;

α controls the fence width (that is, the width of the acceptance region); and

β controls the minimum allowable width of the acceptance region (by disallowing quartiles that are too narrow).

The estimate is considered anomalous if its score is greater than the upper cutoff or less than the lower cutoff.

The H–B macro-edit (which includes the cut-off methodology) can be expressed in the following *quartile-transformed* form which simplifies the cut-off definition:

$$S_{\text{QHB}}(R_{i,d}) = \begin{cases} \dfrac{\text{median}\left(S_{\text{HB}}(R_{i,d})\right) - S_{\text{HB}}(R_{i,d})}{\max\left(D_{\text{Q1S}}, \beta\,\text{median}\left(S_{\text{HB}}(R_{i,d})\right)\right)} & \text{if } S_{\text{HB}}(R_{i,d}) < \text{median}\left(S_{\text{HB}}(R_{i,d})\right) \\ \dfrac{S_{\text{HB}}(R_{i,d}) - \text{median}\left(S_{\text{HB}}(R_{i,d})\right)}{\max\left(D_{\text{Q3S}}, \beta\,\text{median}\left(S_{\text{HB}}(R_{i,d})\right)\right)} & \text{if } S_{\text{HB}}(R_{i,d}) \geq \text{median}\left(S_{\text{HB}}(R_{i,d})\right) \end{cases} \tag{6}$$

and the ratio is declared anomalous if $S_{QHB}(R_{i,d}) > \alpha$.

The user must provide values for the parameters U, α and β. The value chosen for U tends to be very subjective. Several investigations (Hidiroglou–Berthelot, 1986; Banim, 2000; Sigman, 2005; Thompson, 2007; and Thompson and Ozcoskun, 2007) indicate typical 'default' values of $\beta = 0.05$ and $U = 0.3$ while the choice of α is more open-ended depending on each collection.

The H–B macro-edit cannot be used (without modification) for estimates that can be both positive and negative due to constraints for the cutoff methodology. The H–B historical ratio macro-edit cannot be used when previous estimates are zero and the H–B current ratio macro-edit cannot be used when current estimates are zero. There must be a suitable number of estimates within each class in the domain of interest to allow for the calculation of sufficiently robust medians and quartiles. As a minimum requirement there must be at least three estimates available within each domain class and, for robust measures, there should be many more. This places restrictions on which collections are suitable candidates for the H–B macro-edit. For example, there tends to be fewer classes within the domains of interest for a small country such as Australia than for larger countries such as Canada or North America. The H–B macro-edit does not encourage macro-editors to interact with the data and users may find it difficult to explain and understand. On the other hand, the method is very robust and can work with a variety of collections. Given a tolerance width, the method generates dynamic two-sided cut-offs based on the observed data and does not require the input of expected values. However, the method is reliant on there being a relationship between the numerator and denominator variables used in the ratios.

4. A SUMMARY OF THE U.S. CENSUS BUREAU'S ECONOMIC CENSUS FINDINGS ON THE USE OF SCORES FOR MACRO-EDITING

Sigman (2005), Thompson (2007) and Thompson and Ozcoskun (2007) have examined a number of scoring methods based on ratios that attempt to alleviate the problems identified in Section 2. They examined the Z-score, the H–B score, the use of resistant fences (Hoaglin, Mosteller, and Tukey, 1983) and asymmetric fences (and other methods involving trimmed means, Winsorised variances, and robust regression which are resistant to the effect of extreme values). They noted that the Census has a specific macro-editing problem in that many thousands of estimates must be examined and validated within a short period of time. A key component of their macro-editing strategy is the efficient detection of anomalous estimates using scores.

Z-scores were created for historic and current cell ratios (Sigman 2005). A cut-off of 1.78 was used and an *initial* anomaly was declared if its score was greater than the cut-off value. Initial anomalies were declared as *final* anomalies only if various auxiliary conditions were satisfied. For example, very small cells (defined as cells where all cell values were less than specified cut-offs, which varied across sectors) could not be labelled as final anomalies. For example, cut-offs of 10 for the number of reporting units and 20 for the total number of employees were used for the Wholesale sector.

Sigman modified the H–B score defined in (6) to make it applicable to current ratios by replacing the multiplicative factor used for historical ratios with:

$$\max\left(Y_{i,d,t},\ \text{median}(R_{i,j,d})Y_{j,d,t-1}\right)^{U} \qquad (7)$$

where i indicates the numerator variable and j the denominator variable.

Sigman (2005) reported that no one method was proven to outperform the others, though the results were subjective. Although $S_{QHB}(R_{i,d})$ with $U = 0.3$ and the Z-score were most popular (for historical ratios) with the subject matter testers, those who preferred the Z-score had used it previously in the 1992 Census. It was felt that this may have influenced their preferences and a compromise version was settled on where an initial anomaly was declared if:

$$\left(\left|S_{\text{QHB}}(R_{i,d})\right| \text{ with } U = 0.3\right) > 4$$

and

$$\left(\left|S_{\text{QHB}}(R_{i,d})\right| \text{ with } U = 0\right) > 4$$

A final anomaly was declared if the output cell passed very-small-cell conditions. This hybrid edit was used because editors felt that $S_{QHB}(R_{i,d})$ with $U = 0.3$ tended to identify too many anomalous ratios involving large estimates. Note that $S_{QHB}(R_{i,d})$ with $U = 0$ takes no account of the size of the estimates.

Thompson (2007) and Thompson and Ozcoskun (2007) continued and extended the investigation. Thompson (2007) found that approaches based on robust regression were relatively free of masking but more prone to swamping than other methods tested. The resistant fences approach requires that the ratios are reasonably symmetrically distributed. Cut-offs should be dynamic rather than preset and depend on the estimates studied. The H–B macro-edit appeared to be the best of the methods tested (though it did not outperform asymmetric fences for high correlation ratios).

Thompson and Ozcoskun (2007) tested the methods on a greater variety of collections, concentrating on the H–B macro-edit and asymmetric fences approach for historical and current ratios. They found that, when the estimates are strictly positive and there is some statistical association between the estimates involved in the ratios, the H–B macro-edit was generally effective. Asymmetric fences did not perform well with the poorly-correlated estimates from current ratios. H–B with $a = 10$, $\beta = 0.05$ and $U = 0.3$ or 0.5 produced the best outcomes.

The authors highlight that each technique had varying levels of success depending on the characteristics of the data within and between each survey. They warn that extrapolation of their results to other situations is risky and reinforce the earlier conclusion that predetermined cut-offs are not successful and that methods that dynamically identify anomalous estimates are needed.

5. THE CONCEPT AND DEFINITION OF SIGNIFICANCE FOR MACRO-EDITING

This section develops the concept of *significance* for macro-editing. The scores outlined in Section 2, which are prone to size masking, are not set up within a significance context. They take no account of the importance of the estimate being scored with respect to the totality of estimates requiring macro-editing. The H–B score attempts to deal with the problem through the centering and magnitude transformations but it does not make use of macro-editor expectations.

Significance editing is based on predicting the impact of editing actions on the outcomes considered important. Significance scores have the following general form:

$$\text{Score} \;=\; 100\times\frac{\text{Measure of predicted impact of editing}}{\text{Scaling value}} \qquad (8)$$

In micro significance editing, impact is measured as the bias in a set of chosen estimates caused by reported data errors (Farwell, Poole and Carlton, 2002). The basic measure of predicted impact for micro significance editing is:

$$\begin{aligned}\text{Micro-editing impact} \;=\; &\text{Adjusted expected target estimate}\\ &-\text{Expected target estimate}\end{aligned} \qquad (9)$$

where the *adjusted expected target estimate* is calculated in the following way. The expected target estimate is a function of the expected unit values and the estimation methodology. When a reported value is obtained, we remove the contribution of its associated expected value from the calculation of the expected target estimate and replace it with the contribution from the reported value. That is, we replace the expected unit value with the reported unit value and we recalculate a new expected target estimate. This is done on a value by value basis. Accordingly, there is an adjusted expected target estimate value for each reported value requiring a score. For Horvitz–Thompson estimates of total, this is equivalent to multiplying the difference between the reported and expected value by the unit weight. Although the impact on an estimate of total is obvious, (9) is needed to support scores for more complex estimates such as estimates of rates, standard errors, and indexes.

The scaling value in (8) may be an expected estimate or an expected standard error. The score can be considered to ‘target’ a set of estimates. These are referred to as *target estimates* while the domain containing them is referred to as the *level of significance* or *target domain* in this paper.

The definition and measurement of predicted impact for macro-editing is far less straightforward than that for micro-editing. There are often conflicting macro-editing priorities which change as macro-editing progresses. For example, consider a situation where we have previous and current estimates for several variables at State level. Our first priority is to produce high quality Australian level estimates (with high quality movement estimates as a by-product). However, we also want to produce good quality State level estimates. How can we address both sets of priorities?

A percentage movement score (see table 2.1) is useful because the individual movements at State level are important. If we ordered the State estimates by the magnitude of the scores we would tend to rank estimates for small States higher than estimates for large States due to the size masking effect (and the same would happen if we used a historical ratio score).

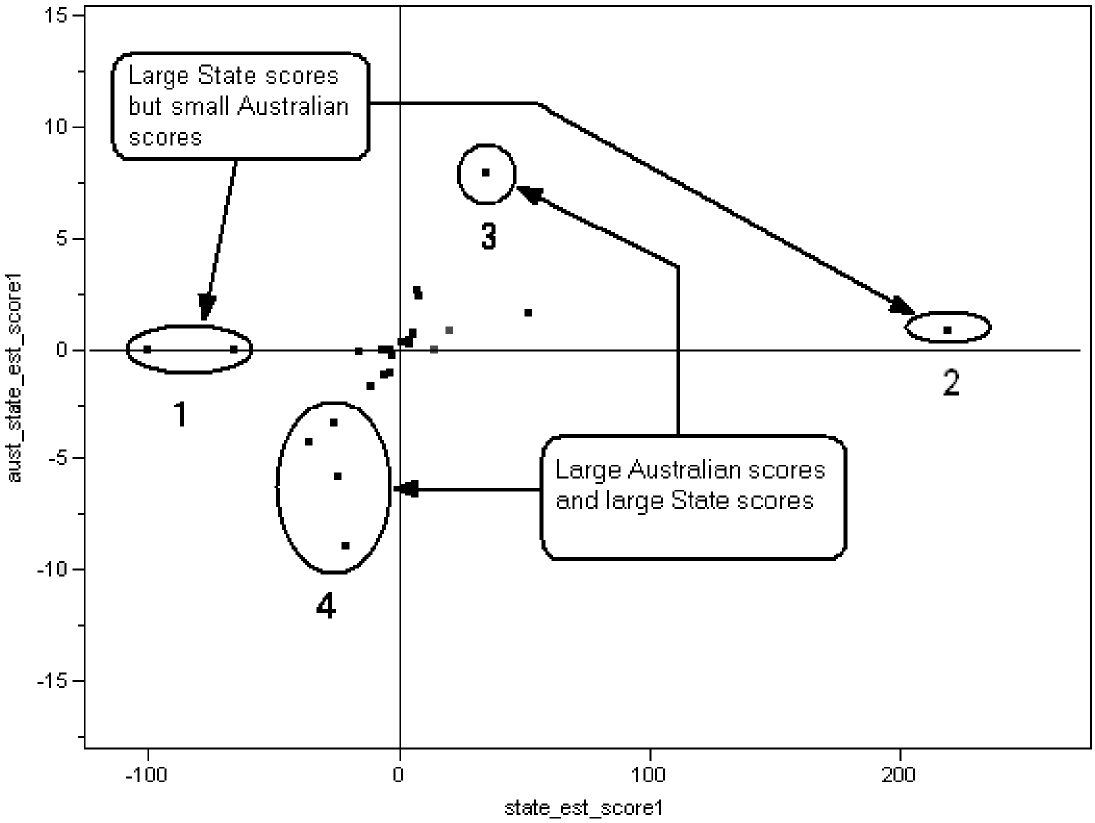
If we were not in a position to edit every State estimate with a large movement, it would be logical to assess the importance of the State movements in terms of their impact on the Australian movements. The following two scores, which measure the State movement for variable *i* as a percentage of the previous State and Australian estimates, can be used to analyse the problem:

$$State\_est\_score1 = 100 \times \frac{Y_{i,t,\text{State}} - Y_{i,t-1,\text{State}}}{Y_{i,t-1,\text{State}}} \qquad (10)$$

$$Aust\_state\_est\_score1 = 100 \times \frac{Y_{i,t,\text{State}} - Y_{i,t-1,\text{State}}}{Y_{i,t-1,\text{Aust}}} \qquad (11)$$

Figure 5.1 below plots *State_est_score1* against *Aust_state_est_score1*. It can be seen that some State estimates have very large scores at State level but very small scores at Australian level which indicates that their movements are relatively unimportant at Australian level. Other State estimates have relatively small State movements which are important from an Australian perspective. The difficulty with using scores in macro-editing is the need to balance conflicting priorities. What is more important from a macro-editing perspective? Is it the estimate points contained within regions 1 or 2 in figure 5.1 or those contained in regions 3 or 4? We can say, at least, that the highlighted sets of estimates appear more important than the remaining estimates in the display.

5.1 Aust_state_est_score1 versus State_est_score1 (for State estimates)



There are often more than two levels of estimates to deal with. For example, in an ABS census of Australian agriculture, there may be up to 30 variables in any one of 65 statistical divisions (SDs) which subdivide States, up to 300 variables in any one of 8 States, and over 900 variables Australia-wide. As the numbers of variables increase, and as the levels become finer, we find ourselves faced with a macro-editing dilemma. It is not uncommon to be faced with thousands of estimates and movements across several levels which need to be assessed and prioritised during macro-editing.

This paper proposes that the predicted measure of impact for macro-editing be the following extension of (9):

$$\text{Macro-editing impact} = \text{Adjusted expected target estimate} - \text{Expected target estimate} \quad (12)$$

where the definition of the adjusted expected target estimate is the natural extension of the definition used for micro-editing. When a base estimate is scored, we remove the contribution of its associated expected estimate from the calculation of the expected target estimate and replace it with the contribution from the actual base estimate. That is, we replace the expected base estimate with the observed base estimate requiring a score and we recalculate a new expected target estimate. This is done on a base estimate by base estimate basis. Accordingly, there is an adjusted expected target estimate value for each base estimate requiring a score. For estimates of total, this is equivalent to using the difference between the observed and expected base estimate as the measure for macro-editing impact. Although the macro-editing impact is obvious for estimates of total, (12) is needed to derive more complex scores involving estimates such as rates, standard errors, and indexes. Refer to the Appendix for an example of this concept applied to a ratio estimate.

A macro-editing score can be developed using (12). However, unlike the micro-editing version, there may be several target levels and several scores associated with a single observed estimate. In fact, the predicted impact will depend on the level of significance and a scaling value for each estimate at that level will be required. This idea is developed in this paper by incorporating the concept of a *hierarchy* of targets and *hierarchical* scaling values.

The general form of a macro significance score is:

$$\text{Score} = 100 \times \frac{\text{Measure of predicted macro-editing impact}}{\text{Scaling value for target level}} \tag{13}$$

and scores (10) and (11) are an example of a two-level hierarchy involving two estimate scores.

6. OUTLINE OF THE MACRO SIGNIFICANCE EDITING FRAMEWORK

In this Section, we outline macro significance concepts such as the domain of study, the level of significance, base and target estimates and scores, expected estimates, sensitivity measures, hierarchical scores, hierarchical macro-edits, combined scores, cost/benefit curves and Gini indexes. Macro significance scores will be defined for estimates of total, ratios of estimates of total, and standard errors of estimates using (12) and (13) and a framework will be proposed which will allow scores to be combined. For example, a current ratio score can be combined with estimate scores for the numerator and denominator estimates. It will be possible to rank estimates using functions of scores, functions of ranks when several ranks are involved, or functions of both scores and ranks. The use of expected estimates in the measure of predicted impact leads to more complex scores such as a current ratio score which uses historical estimates as expected values for the numerator and denominator estimates.

6.1 The study domain and base scores

Base scores are scores where the scaling value in (13) is from the study domain. These scores require observed and expected estimates (or expected standard errors) at the base level. The expected estimates may be based on historical estimates, modelled estimates, current medians or averages, or as a last resort, guesses. *State_est_score1* (10) is an example of a base score.

The base estimate score is:

$$S_{\text{est,base}}(Y_i) \ = \ 100 \times \frac{\Delta Y_{i,\text{base}}}{Y^*_{i,\text{base}}} \tag{14}$$

where $Y_{i,\text{base}}$ and $Y^*_{i,\text{base}}$ are the observed and expected estimates of total for variable i within the base domain, and

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

The base ratio score is:

$$S_{\text{ratio,base}}(R_{i,j}) \ = \ 100 \times \frac{\Delta R_{i,j,\text{base}}}{R^*_{i,j,\text{base}}} \tag{15}$$

where:

$Y_{i,\text{base}}$ is the observed numerator base estimate of total for variable i;

$Y_{j,\text{base}}$ is the observed denominator base estimate of total for variable j;

$Y^*_{i,\text{base}}$ is the expected value for $Y_{i,\text{base}}$;

$Y^*_{j,\text{base}}$ is the expected value for $Y_{j,\text{base}}$; and

$$R_{i,j,\text{base}} = \frac{Y_{i,\text{base}}}{Y_{j,\text{base}}}$$

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

$$\Delta R_{i,j,\text{base}} = R_{i,j,\text{base}} - R^*_{i,j,\text{base}}$$

Base scores using expected estimates as scaling values cannot be defined when expected base estimates are zero and expected standard errors should be used instead. If expected standard errors are used, replace the expected base estimates in the denominators of (14) and (15) with $\alpha_{\text{base}}\text{SE}^*(Y_{i,\text{base}})$ and $\alpha_{\text{base}}\text{SE}^*(R_{i,j,\text{base}})$ respectively, where $\text{SE}^*(Y_{i,\text{base}})$ and $\text{SE}^*(R_{i,j,\text{base}})$ are expected standard errors. The parameter α_{base} has been incorporated to allow expected standard error to be converted to a *bias tolerance* (with $\alpha_{\text{base}} = 1$ suggested as the default value).

The base standard error score for an estimate of total is:

$$S_{\text{se,base}}(Y_i) \;=\; 100 \times \frac{\Delta\text{SE}(Y_{i,\text{base}})}{\alpha_{\text{base}}\text{SE}^*(Y_{i,\text{base}})} \qquad (16)$$

where $$\Delta\text{SE}(Y_{i,\text{base}}) = \text{SE}(Y_{i,\text{base}}) - \text{SE}^*(Y_{i,\text{base}})$$

The base standard error score for an estimate of rate is:

$$S_{\text{se,base}}(R_{i,j}) \;=\; 100 \times \frac{\Delta\text{SE}(R_{i,j,\text{base}})}{\alpha_{\text{base}}\text{SE}^*(R_{i,j,\text{base}})} \qquad (17)$$

where $$\Delta\text{SE}(R_{i,j,\text{base}}) = \text{SE}(R_{i,j,\text{base}}) - \text{SE}^*(R_{i,j,\text{base}})$$

and i and j represent two different variables. (Note that equivalent scores for censuses can be created based on observed and expected coefficients of variation.)

The standard error score is interesting in that it is usually only those observed to be larger than the expected standard errors that are usually considered as anomalous. However, one could argue that a standard error that is much smaller than expected could also indicate a macro-editing problem (such as a systematic processing error).

If we are only interested in standard errors that are too large, we add the following conditions to (16) and (17):

$S_{\text{se,base}}(Y_i) = 0$ when $\text{SE}(Y_{i,\text{base}}) < \text{SE}^*(Y_{i,\text{base}})$ for standard error scores for estimates;

and

$S_{\text{se,base}}(R_{i,j}) = 0$ when $\text{SE}(R_{i,j,\text{base}}) < \text{SE}^*(R_{i,j,\text{base}})$ for standard error scores for rates.

If expected estimates are used as scaling values, replace the expected base standard errors in the denominators of (16) and (17) with $Y^*_{i,\text{base}}$ for estimate scores or $R^*_{i,j,\text{base}}$ for ratio scores. Any standard error base score using expected estimates as scaling values cannot be defined when expected base estimates are zero and expected standard errors should be used.

Movement scores are not developed in this paper (though they could be a consideration for a collection designed specifically to measure accurate movements). They are not needed for most collections because movement scores are very similar to estimate scores which use previous estimates as expected estimates.

6.2 Sensitivity measures

To manage problems affecting the quality of the scores, we introduce the concept of *sensitivity measures* which are an auxiliary layer of conditions imposed on the anomaly detection process. They can be used to exclude specific estimates from the anomaly selection process or to modify the scores themselves. The conditions that Sigman (2005) placed on labelling initial anomalies as final anomalies in Section 4 are an example of *conditional* sensitivity measures. The magnitude transformations used in the historical and current ratio H–B macro-edits, (4) and (7), are examples of *multiplicative* sensitivity measures. Multiplicative sensitivity measures tend to be implicit since they are generally included in the definition of the score. If we were to use base scores only for anomaly detection, some form of sensitivity measure would be needed to control size masking. Some examples of sensitivity measures are:

$$d_{i,\text{base}} = \max\left(\left|Y_{i,\text{base}}\right|, \left|Y^*_{i,\text{base}}\right|\right)^U \quad \text{and } 0 < U \le 1 \tag{18}$$

$$d_{i,\text{base}} = \left(\frac{\max\left(\left|Y_{i,\text{base}}\right|, \left|Y^*_{i,\text{base}}\right|\right)}{\left|Y^*_{i,\text{base}}\right|}\right)^U \quad \text{and } U \ge 1 \tag{19}$$

where each could be combined with a base score (such as a standard error score) and anomalous estimates are those with $s_{i,\text{base}} > c_1$ and $d_{i,\text{base}} > c_2$ (c_1 and c_2 are cut-offs).

Relative sensitivity measures allow a single sensitivity cut-off to be applied to many variables. Other examples of conditional sensitivity measures include restrictions on the minimum number of respondents allowed for base estimates, restrictions on the minimum number of estimates contributing to target estimates, and the exclusion of estimates of zero.

6.3 The level of significance, target estimates and hierarchical scores

Hierarchical scores are a specific multiplicative application of sensitivity measure (19) with $U = 1$ to base scores where the end result is a score which uses a *target* estimate as the scaling value. That is, hierarchical scores are scores for base estimates where the level of significance is a higher level than the base level.

The hierarchical estimate score is:

$$S_{\text{est,base,target}}(Y_i) \;=\; 100\times\frac{\Delta Y_{i,\text{base}}}{Y^*_{i,\text{target}}} \tag{20}$$

The hierarchical ratio score is:

$$S_{\text{ratio,base,target}}(R_{i,j}) \;=\; 100\times\frac{R^*_{i,j,\text{target}|\text{base}} - R^*_{i,j,\text{target}}}{R^*_{i,j,\text{target}}} \tag{21}$$

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

is the expected target ratio;

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

is the adjusted expected target ratio. (Refer to the Appendix for details on how the adjusted expected target ratio is calculated.)

Hierarchical scores using expected estimates as scaling values cannot be defined when expected target estimates are zero. Difficulties can also arise with hierarchical scores when base estimates can be both positive and negative. For example, as the sum of expected base estimates approach zero the hierarchical score becomes increasingly erratic. It is recommended to use the standard error as the scaling value in such cases. If standard errors are used as scaling values, replace the expected target estimates in the denominators of (20) and (21) with $a_{\text{target}}\text{SE}^*(Y_{i,\text{target}})$ and $a_{\text{target}}\text{SE}^*(R_{ij,\text{target}})$ respectively (using $a_{\text{target}} = 1$ as the default).

Hierarchical scores for standard errors are affected by the independence of the base estimates. When they are not independent, target variances are not the sum of base variances leading to complicated scores. The following specifications assume that the base estimates are independent estimates. The hierarchical standard error score for an estimate of total is:

$$S_{\text{se,base,target}}(Y_i) = 100 \times \frac{\text{SE}^*(Y_{i,\text{target|base}}) - \text{SE}^*(Y_{i,\text{target}})}{\alpha_{\text{target}} \text{SE}^*(Y_{i,\text{target}})} \tag{22}$$

where $\Delta\text{Var}(Y_{i,\text{base}}) = \text{SE}(Y_{i,\text{base}})^2 - \text{SE}^*(Y_{i,\text{base}})^2$

and $SE^*(Y_{i,\text{target|base}}) = \sqrt{\text{Var}(Y_{i,\text{target}}) + \Delta\text{Var}(Y_{i,\text{base}})}$

is the adjusted expected (drop–1) target standard error.

If the user chooses to only look for standard errors which are larger than the expected standard error, add the condition that:

$$S_{\text{se,base,target}}(Y_i) = 0 \quad \text{if SE}(Y_{i,\text{base}}) < \text{SE}^*(Y_{i,\text{base}})$$

Further research is needed on standard error scores when dependent base estimates are involved to account for the impact of the covariances that dependent base estimates generate. Also, standard errors for ratios are affected both by estimate dependence and by the non-linearity of variance formula. Using the Jackknife drop–1 method to calculate $\text{SE}^*(R_{i,j,\text{base|target}})$ is complicated and various approximations under certain conditions may be needed. The incorporation of a sensitivity measure such as (18) or (19) may be a possible compromise solution.

Aust_State_est_score1 (11) is an example of a hierarchical estimate score. It is common to have several levels of significance and, therefore, several hierarchical scores. Hierarchical scores are used to develop hierarchical macro-edits which are described in the following section.

6.4 Hierarchical macro-edits

We introduce the concept of ***hierarchical*** macro-editing in this Section. Hierarchical macro-edits can be used to detect anomalous base estimates while taking into account the importance of the base estimate deviations from their expectations in terms of their impacts on the chosen target levels. They involve a combination of base and hierarchical scores and cut-offs where a cut-off is chosen for each of the base and hierarchical scores. Although each cut-off can be chosen independently, the preferred method is to select the hierarchical cut-offs first and apply these to the base estimates. The distribution of those base estimates above the hierarchical cut-offs is then examined and a base cut-off is chosen. The hierarchical and base cut-offs are then applied to the full set of base estimates to select the anomalous estimates.

A value from (0,1) is assigned for each base estimate indicating whether it passed or failed each of the base and hierarchical edits (where '1' indicates the estimate failed the chosen cut-off). Each base estimate is assigned an n-dimensional point where n is the number of hierarchies. For example, a three-level hierarchy results in points (or categories) such as (1,1,1), (1,1,0),......., through to (0,0,1) and (0,0,0). The first coordinate relates to the highest hierarchical level, the second coordinate to the next highest hierarchy, and so on. The last coordinate relates to the base level. For example, (1,0,1) indicates that the base estimate failed the highest level hierarchical edit, passed the second highest level hierarchical edit, and failed the base level edit.

The user can choose the hierarchical category or group of categories they feel is most appropriate to investigate. The anomalous estimates within each group can be ordered by the size of the base score or one of the hierarchical scores. For example, the category (1,1,1) would be top priority and typically ordered by base score size while those in (1,0,0) would tend to be ordered by the top level hierarchical score size. Hierarchical edits have the ability to address conflicting macro-editing priorities while giving some flexibility to the macro-editor. They can be combined with sensitivity measures if necessary.

Four types of prototype hierarchical macro-edits are currently being tested in the ABS. One is for macro-editing estimates of total and one for ratios. Both use an optional conditional sensitivity measure similar to (19) based on benchmarks. The third is also for ratios but combines the ratio and estimate score results using ellipsoidal distance (defined in Section 6.9). These require the user to provide expected base and target estimates while the fourth is designed for use without expected estimates. It generates expected estimates through the use of a median. It uses an implicit multiplicative sensitivity measure similar to (19) based on benchmarks to create the hierarchical scores. Refer to (Farwell, 2009a) and (Farwell, 2009b) for more details.

These hierarchical macro-edit prototypes revolve around six basic steps which are:

(a) create macro-data;

(b) create scores and ranks;

(c) select hierarchical score cut-offs;

(d) apply hierarchical score cut-offs and select the base score cut-off;

(e) select hierarchical outlier categories; and

(f) select anomalous base estimates.

As an interlude to this outline of the significance editing framework, Section 6.5 below presents an example of hierarchical macro-editing before returning to the framework in Section 6.6.

6.5 An example of a three-level hierarchical macro-edit

In this example, we demonstrate hierarchical macro-editing on estimates of total from an ABS Agricultural collection using previous estimates as expected estimates. It involves a three-level hierarchy consisting of statistical division (SD) as the base level, State as the next highest level, and Australia as the highest level. The example dataset consists of 1646 SD estimates which aggregate to 290 State estimates and 49 Australian estimates involving 49 variables. We use (14) and (20) to calculate the following three estimate scores:

$$\begin{aligned} SD\_est\_score &= 100\times\frac{\text{Current SD estimate} - \text{Previous SD estimate}}{\text{Previous SD estimate}} \\ SD\_State\_est\_score &= 100\times\frac{\text{Current SD estimate} - \text{Previous SD estimate}}{\text{Previous State estimate}} \\ SD\_Aust\_est\_score &= 100\times\frac{\text{Current SD estimate} - \text{Previous SD estimate}}{\text{Previous Aust estimate}} \end{aligned} \qquad (23)$$

As outlined in Section 6.4, the SD-State and SD-Aust hierarchical cut-offs are the first to be chosen using graphs displaying score size versus rank (based on descending score size). Extreme scores are excluded from the graphs (to improve readability) by applying user-defined *graph cut-off* values (a default value of 100 is used). The excluded scores are separately listed. For example, figure 6.2 below was used to choose an *SD_State_est_score* cutoff of 1.75 (16 SD estimates were excluded from the graph as shown in figure 6.1 below). Similarly, an *SD_Aust_est_score* cut-off of 0.25 was selected using a graph of *SD_Aust_est_score* size versus rank.

Figure 6.3 below displays the distribution of *SD_est_score* size prior to application of the two hierarchical cut-offs (158 estimates were excluded by the graph cut-off). Figure 6.4 below shows the distribution after the two hierarchical cut-offs have been applied (75 estimates were excluded by the graph cut-off) and was used to select 15.0 as the *SD_est_score* cut-off.

6.1 Application of a graph cut-off

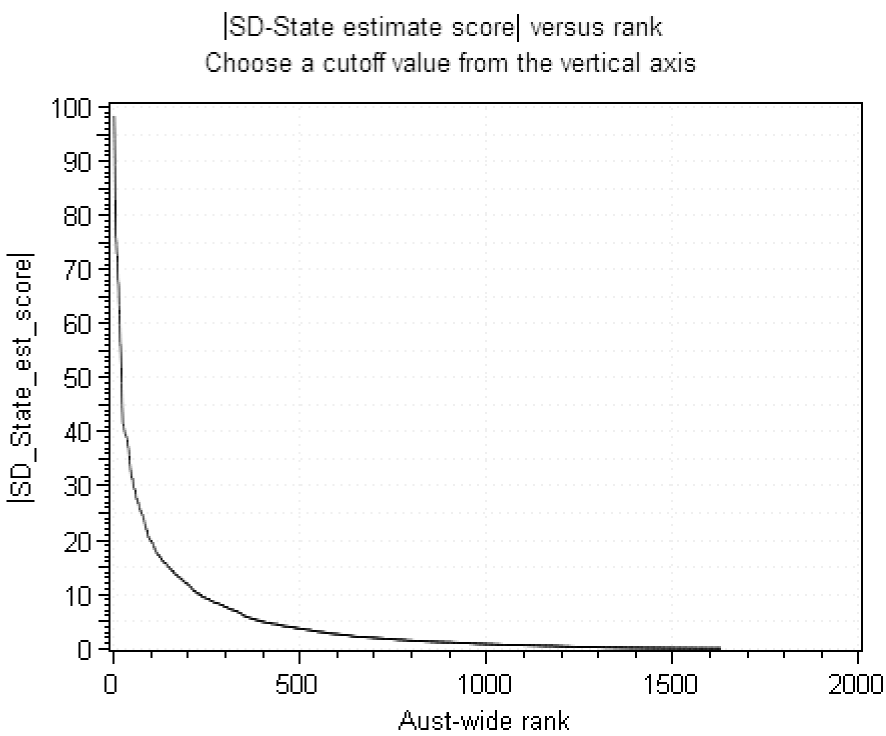
Count of absolute SD-State scores > 100 %

count	Frequency
1	16

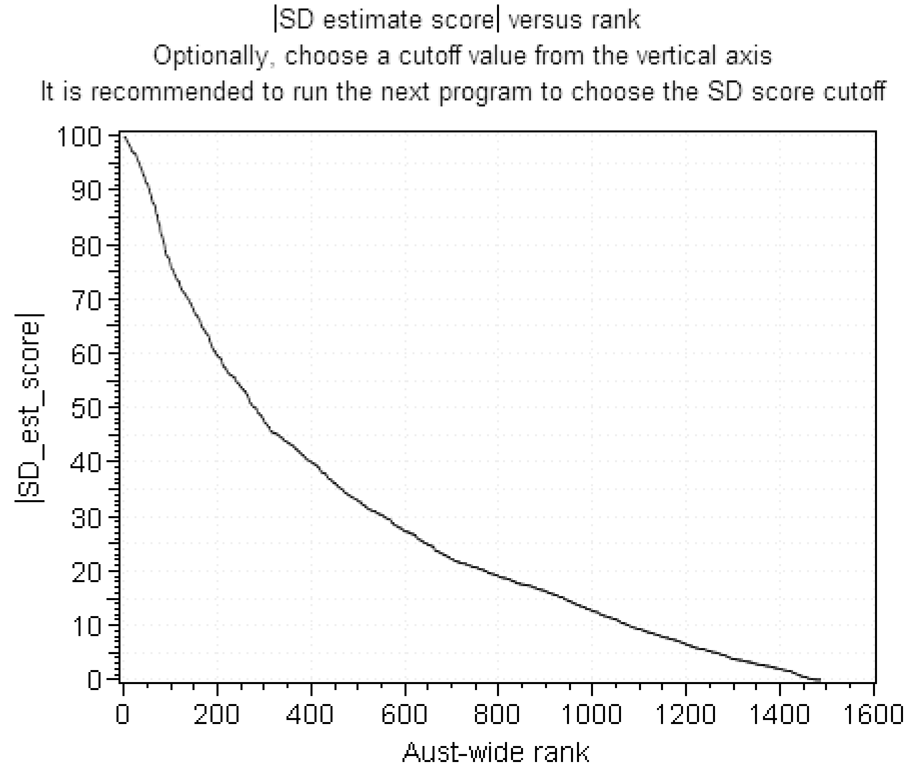
absolute SD-State estimate scores > 100 %
These have been excluded from the SD-State estimate score graph in order to make the graph more readable

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

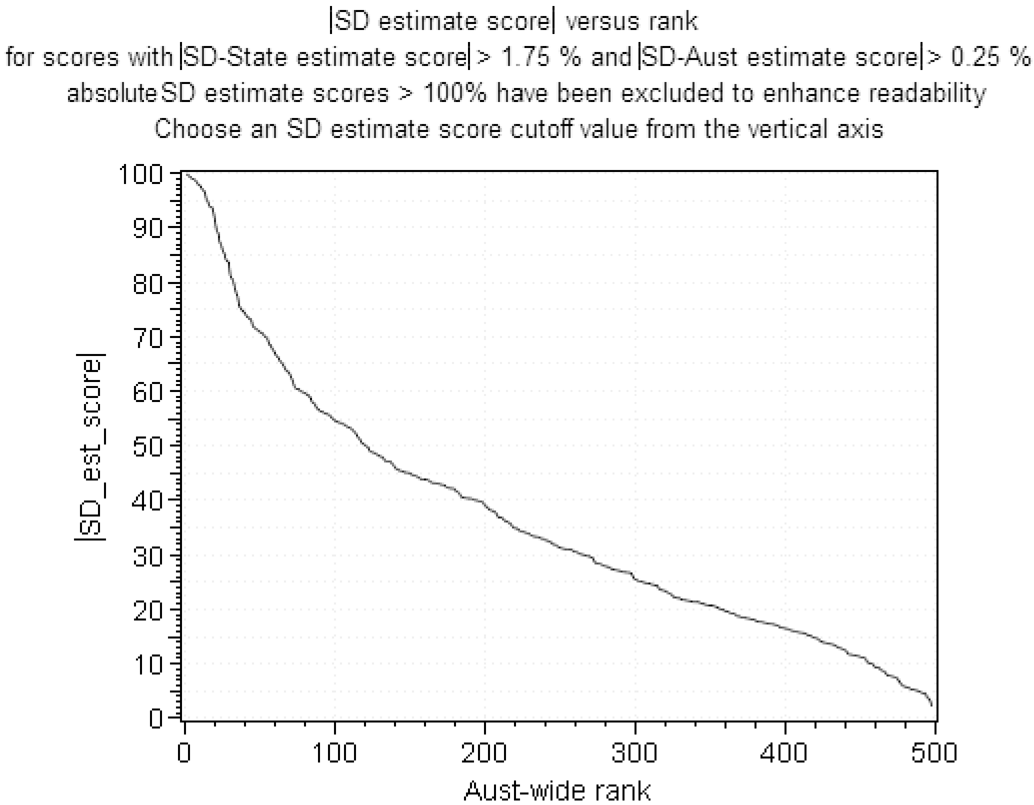
6.2 SD_State_est_score size versus rank



6.3 SD_est_score size versus rank prior to applying hierarchical cut-offs



6.4 SD_est_score size versus rank after applying hierarchical cut-offs



Using 0.25, 1.75, and 15.0 as the three cut-offs, table 6.5 below displays results for the hierarchical macro-edit categories. Macro-editors can choose the hierarchical categories they feel are most appropriate to investigate. The estimates within each group can be ordered by the magnitude of one of the three scores in (23). For example, category (1,1,1) should be top priority and the estimates within it should be ordered by descending *SD_est_score* size. Editors may wish to examine estimates within other categories with a view to augment the selections from (1,1,1). A subset of these estimates can optionally be selected and added to the existing selections. For example, after examining categories (1,1,0), (1,0,1), (1,0,0) and (0,1,1) using various orderings, it is apparent that some extra selections can be made from category (1,1,0).

Tables 6.6 and 6.7 below display the top estimates in this category ordered by descending *SD_Aust_est_score* size and descending *SD_State_est_score* size respectively. The SD estimates ranked 1st, 2nd, 3rd, 7th, 8th, 9th, 11th, 12th and 13th in table 6.6 could be selected due to their impact mainly on State level. Somewhat subjectively, SD estimates ranked 4th, 5th, 6th and 10th could be excluded as they come from smaller States where a higher implicit *SD_State_est_score* cut-off could be applied. The SD estimates ranked 4th, 6th and 7th in table 6.7 could be selected due to their impact on Australian and State levels. This would result in 10 selections being added to the 493 in category (1,1,1). However, we choose to keep this example simple because the results will be used as part of the analysis in Section 7. For this example, we choose only the estimates in (1,1,1) as our set of anomalous estimates resulting in the selection of 493 anomalous SD estimates. Details of the first 15 selections are shown in table 6.8 below.

6.5 Hierarchical macro-edit results for SD estimates

Hierarchical macro-edit categories	*Number of SD estimates*	*Number of anomalous SD estimates*
000	367	.
001	407	.
010	42	.
011	135	.
100	61	.
101	66	.
110	80	.
111	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

6.6 The top 15 SD estimates in the (1,1,0) category (ordered by SD_Aust score size)

				SD Estimates		*Estimate scores*		
Rank	*State*	*SD*	*Item*	*Current unedited*	*Previous*	*SD*	*SD_State*	*SD_Aust*
1	2	240	3608812	178,859	204,445	–12.5	–9.5	–5.4
2	1	150	4200713	156,023,306	138,166,232	12.9	9.8	4.5
3	3	320	1504102	729,971	805,169	–9.3	–5.8	–3.7
4	5	535	1807001	243,035	212,070	14.6	4.6	3.6
5	4	420	3606102	66,640	58,532	13.9	10.0	3.5
6	4	415	1809101	77,780	87,944	–11.6	–6.5	–2.9
7	1	105	3605802	13,127,431	12,109,911	8.4	7.5	2.2
8	1	150	4200712	2,753,087	2,607,380	5.6	4.1	2.1
9	1	135	1900301	30,129	26,479	13.8	3.4	1.8
10	5	525	1807001	272,129	287,492	–5.3	–2.3	–1.8
11	1	140	1500801	149,968	168,308	–10.9	–4.1	–1.7
12	2	205	3605801	54	52	4.6	4.1	1.5
13	2	240	3608811	2,703	2,573	5.1	3.9	1.5
14	3	330	1504101	191,290	180,266	6.1	2.1	1.5
15	2	215	1008102	439,608	488,382	–10.0	–2.6	–1.4

6.7 The top 15 SD estimates in the (1,1,0) category (ordered by SD_State score size)

				SD Estimates		*Estimate scores*		
Rank	*State*	*SD*	*Item*	*Current unedited*	*Previous*	*SD*	*SD_State*	*SD_Aust*
1	5	510	7004801	104,458	120,540	–13.3	–12.1	–0.5
2	4	410	8003911	3,401,215	2,965,009	14.7	10.8	0.6
3	5	510	3606101	179	155	14.8	10.5	0.4
4	4	420	3606102	66,640	58,532	13.9	10.0	3.5
5	4	405	3605802	3,467,689	3,848,559	–9.9	–9.9	–0.8
6	1	150	4200713	156,023,306	138,166,232	12.9	9.8	4.5
7	2	240	3608812	178,859	204,445	–12.5	–9.5	–5.4
8	1	155	3608811	716	831	–13.8	–8.5	–1.4
9	1	155	3608812	38,315	43,971	–12.9	–8.4	–1.2
10	4	425	1005101	15,955	18,528	–13.9	–7.9	–1.1
11	1	105	3605802	13,127,431	12,109,911	8.4	7.5	2.2
12	3	305	3605801	11	12	–10.2	–7.2	–0.8
13	3	320	1500101	421,315	467,952	–11.7	–7.0	–0.4
14	4	415	1809101	77,780	87,944	–11.6	–6.5	–2.9
15	3	320	1504102	729,971	805,169	–9.3	–5.8	–3.7

6.8 The top 15 anomalous SD estimates selected from the (1,1,1) category

				Estimates		*Estimate scores*		
Rank	*State*	*SD*	*Item*	*Current*	*Previous*	*SD*	*SD_State*	*SD_Aust*
1	1	120	0100101	174,541,373	401,460	43,376.7	273.9	39.6
2	2	225	3606101	4,421	15	29,026.7	906.7	79.3
3	1	125	4304603	3,685,906	13,255	27,707.6	16,958.7	1,427.7
4	5	505	0100101	16,253,704	112,051	14,405.6	16.0	3.7
5	5	535	4304603	609,654	7,007	8,600.5	7,614.1	234.3
6	3	325	3606101	100	2	4,900.0	14.0	1.8
7	6	610	1918101	1,762	39	4,373.3	1,025.6	36.6
8	3	330	0100101	486,430,959	13,234,790	3,575.4	327.9	107.5
9	2	240	1005101	467,876	14,516	3,123.1	691.8	197.8
10	3	310	1918301	6,359	199	3,097.1	726.3	111.5
11	3	325	3606102	1,800	68	2,545.1	7.3	0.7
12	3	345	4304603	6,701	261	2,469.3	2.9	2.5
13	6	605	0100101	1,464,775	58,052	2,423.2	80.6	0.3
14	5	525	1005102	20,326	976	1,982.5	39.6	1.6
15	5	505	1005102	8,349	422	1,879.0	16.2	0.7

We now return to the outline of the macro significance editing framework.

6.6 Ranks and cut-offs

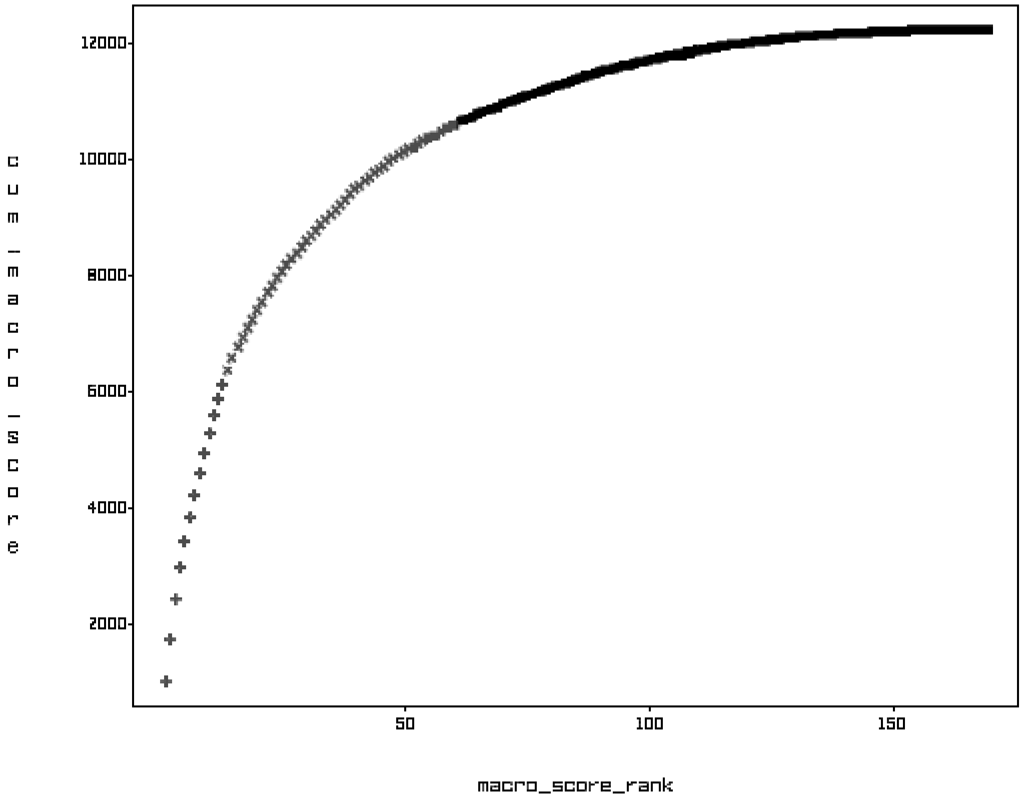
Various ranking methods are available within the *Significance Editing Manual* (ABS, 2011) and this paper will not detail them. Cut-offs may be two-sided or one-sided. Two-sided cut-offs can be used when separate cut-offs are needed for each tail of the score distribution. This paper proposes that one-sided cut-offs be the default with an option for using two-sided cut-offs for combined scores.

6.7 Cost/benefit graphs and the Gini index

Figure 6.9 below shows cumulative score size versus rank which is sometimes referred to as a *cost/benefit* curve in significance editing. The points shown as '+' and '×' in the graph represent the top 60 (possibly) anomalous estimate and standard error combinations. These are divided into primary (indicated with '+') and secondary anomalies (indicated with '×'). A Gini index can be calculated for a cost/benefit curve since it is a form of Lorenz curve and may have some application for macro-editing.

For the example in Section 6.5, a Gini index can be calculated for each State estimate using the cumulative |*SD_State_est_score*| and rank. A very large index value would indicate that a small proportion of SD estimates contribute disproportionately to the State score for that item. State estimates can be ordered by the Gini index.

6.9 Cost/benefit graph and anomaly selections



6.8 Combined scores and macroscores

SEE currently provides the following three choices for combining scores (*ABS Significance Editing Manual,* 2011): the maximum of the scores; the weighted Euclidean distance of the scores; and the weighted root mean square distance of the scores. An estimate or ratio score can be combined with a standard error score to create a *macroscore*. Each estimate and standard error pair can be assigned a single macroscore covering divergence from expectations in both estimate value and standard error. Macroscores for key variables may be a useful way to commence macro-editing, when many key variables are involved, since they allow major errors in processes or data to be quickly found prior to more detailed macro-editing.

The base and hierarchical macroscores for an estimate of total (Y_i), using weighted Euclidean distance, are:

$$S_{\text{macro,base}}(Y_i) = \sqrt{\left(w_{\text{est,base}} S_{\text{est,base}}(Y_i)\right)^2 + \left(w_{\text{se,base}} S_{\text{se,base}}(Y_i)\right)^2} \tag{24}$$

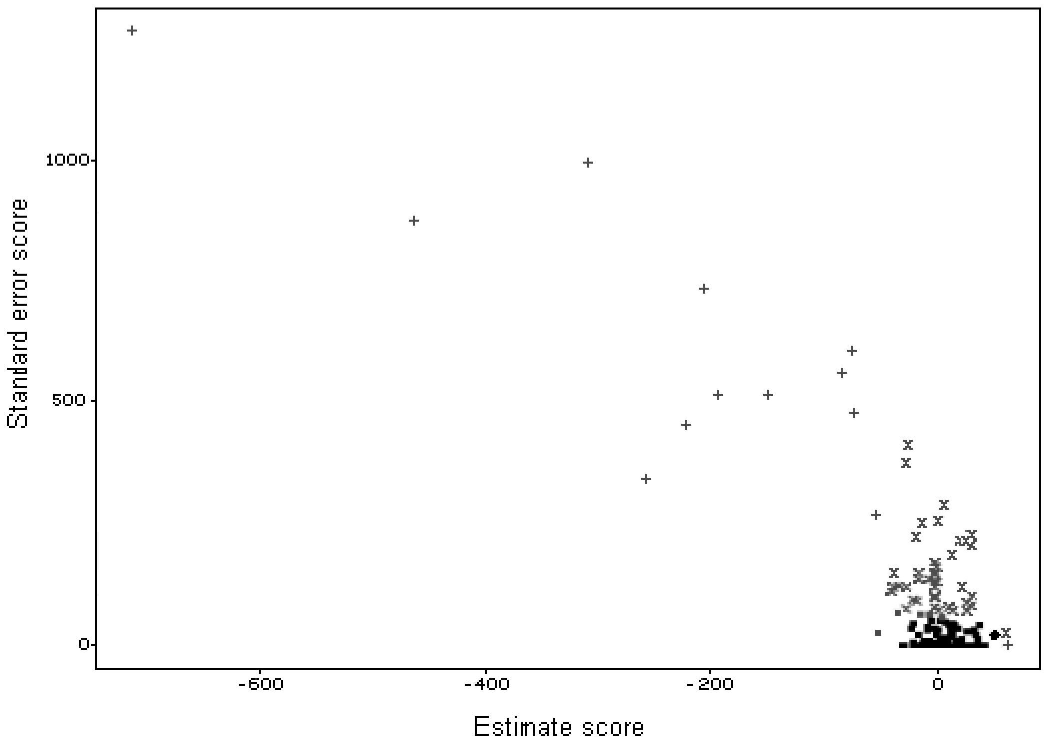
$$S_{\text{macro,base,target}}(Y_i) = \sqrt{\begin{array}{l}\left(w_{\text{est,base,target}} S_{\text{est,base,target}}(Y_i)\right)^2 + \\ \left(w_{\text{se,base,target}} S_{\text{se,base,target}}(Y_i)\right)^2\end{array}} \tag{25}$$

The user-defined score weights $w_{\text{est,base}}$, $w_{\text{se,base}}$, $w_{\text{est,base,target}}$ and $w_{\text{se,base,target}}$ have default values of 1. The equivalent versions for an estimate of rate ($R_{i,j}$) are as above with $R_{i,j}$ replacing Y_i in (24) and (25).

A combined score based on estimate or standard error scores could be useful for detecting problem output cells. For example, we could choose a set of variables for a given level of significance and create a combined score. Output cells could be ordered by the combined score size and size masking could be controlled by incorporating hierarchical scores or sensitivity measures (where estimates with a sensitivity score below the sensitivity cut-off receive a zero final score).

The score used in figure 6.9 above is a macroscore based on (24). The same set of primary and secondary anomaly selections that are displayed in figure 6.9 (the '+' and '×' points) are also displayed in figure 6.10 below which plots estimate score against standard error score. Note that the data displayed in figures 6.9 and 6.10 is for illustration purposes only and is not the same data that was used in Section 6.5.

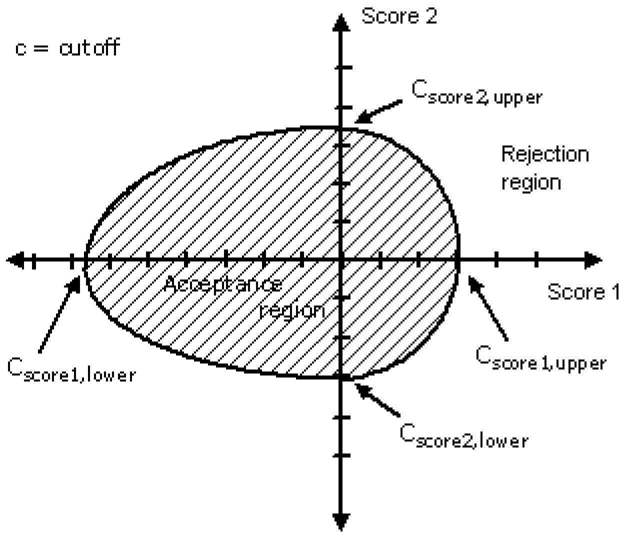
6.10 Estimate score versus standard error score with macro-edit selections



6.9 Combining scores using ellipsoidal distance

Scores can be combined using n-dimensional ellipsoidal distance, a generalisation of the weighted Euclidean distance. There are several ways that the ellipsoidal radii may be created. For example, in figure 6.11 below we have two scores that we wish to combine. We can use the individual score cut-offs as ellipse radii and points outside the ellipse are considered anomalous.

6.11 Example of combined score using ellipsoidal distance



We propose that ellipsoidal distance be the default distance for macroscores since standard error scores may be more variable than estimate scores. For example, the scatter plot in figure 6.10 above displays a cluster of points elongated along standard error score axis (compared to the estimate score axis) and a set of estimate scores with a long left tail. To account for differing spread of the two scores, we could apply a one-sided cut-off to the standard error score and a two-sided cut-off to the estimate score. We can use the three cut-offs to create a macroscore using ellipsoidal distance as follows:

$$Macroscore = \begin{cases} \sqrt{\left(\dfrac{S_{\text{est}}(Y_i)}{c_{\text{est,upper}}}\right)^2 + \left(\dfrac{S_{\text{se}}(Y_i)}{c_{\text{se}}}\right)^2} & \text{if } S_{\text{est}}(Y_i) \geq 0 \text{ and } S_{\text{se}}(Y_i) \geq 0 \\ \sqrt{\left(\dfrac{S_{\text{est}}(Y_i)}{c_{\text{est,lower}}}\right)^2 + \left(\dfrac{S_{\text{se}}(Y_i)}{c_{\text{se}}}\right)^2} & \text{if } S_{\text{est}}(Y_i) < 0 \text{ and } S_{\text{se}}(Y_i) \geq 0 \end{cases} \qquad (26)$$

where $c_{\text{est,upper}}$, $c_{\text{est,lower}}$ and c_{se} are the cut-offs.

For one-sided cut-offs (c_{est} and c_{se}), the ellipsoidal combined score is:

$$Macroscore = \sqrt{\left(\frac{S_{\text{est}}(Y_i)}{c_{\text{est}}}\right)^2 + \left(\frac{S_{\text{se}}(Y_i)}{c_{\text{se}}}\right)^2} \tag{27}$$

and estimate and standard error pairs with $Macroscore > 1$ are selected as anomalous.

There are several ways to create the ellipsoidal distance. For example, we could use a multiple of the median of the absolute value of the scores as a cutoff. That is:

$$Macroscore = \frac{1}{\alpha}\sqrt{\left(\frac{S_{\text{est}}(Y_i)}{\text{median}\left|S_{\text{est}}(Y_i)\right|}\right)^2 + \left(\frac{S_{\text{se}}(Y_i)}{\text{median}\left|S_{\text{se}}(Y_i)\right|}\right)^2} \tag{28}$$

where

$a \geq 1$,

$\alpha(\text{median}\,|S_{\text{est}}(Y_i)|)$ is the cutoff value for the absolute values of the estimate score,

$\alpha(\text{median}\,|S_{\text{se}}(Y_i)|)$ is the cutoff value for the absolute values of the standard error, and

estimate and standard error pairs with $Macroscore > \alpha$ are selected as anomalous.

7. AN EMPIRICAL COMPARISON OF HIERARCHICAL AND HIDIROGLOU–BERTHELOT MACRO-EDITS

This Section presents some comparisons between the 493 anomalous SD estimates selected with the three-level hierarchical macro-edit in Section 6.5 (refer to table 6.5) and a similar number of selections made using several versions of the H–B macro-edit. The study dataset is the same dataset which was used in Section 6.5 except that final estimates have been added. For brevity, we will refer to the hierarchical macro-edit as the *estimate score* edit in these results.

The study dataset consists of *initial* and *final* current estimates and *previous* (historical) estimates. The initial estimates were created using the unit record values that were first recorded in the system. These were very unrefined because they had not been modified by auto-correction, micro-editing, or imputation. Final estimates were created using the file of unit records that generated the final published estimates. Previous estimates were those previously published. The study dataset was constructed by combining the initial, final, and previous SD estimates. Those with missing values or non-matches of current or previous estimates at SD, State, or Australian level were excluded from the dataset. Also, those with previous or current values of zero were excluded. Previous estimates that were zero were removed because the macro-edits under study cannot deal with them. In any event, they are identified in a preliminary macro-editing step within the hierarchical macro-edit suites and are separately listed prior to applying the hierarchical macro-edits. Current estimates that are zero were removed to facilitate the measurement of pseudo-bias (as explained in Section 7.1, relative pseudo-bias is not defined when the final estimate is zero). Their removal did not appear to alter the results in this paper. The preliminary macro-edits also identify target estimates with a single contributing base estimate. The final study dataset consisted of 1646 SD estimates which aggregate to 290 State estimates and 49 Australian estimates involving 49 variables.

The results presented here are indicative only and should be used only to 'get a feel' for the two styles of macro-edits for the following reasons. Firstly, as Thompson and Ozcoskun (2007) advise, one should not extrapolate results based on one collection. The data used in this paper appear to be of particularly poor quality. That is, there are many very large differences between initial and final SD estimate values and only about 25% of them were altered (that is, about 25% of SD estimates in the study dataset had final values which differed from their initial values). Estimate change in the study dataset is reasonably sparse. Secondly, there are many ways both types of macro-edits could be applied and the results presented here, although representing our best attempt at using the macro-edits, are dependent on how they were applied. Thirdly, it is extremely difficult to prove technically that one set of results is better than another due to the complex, and often conflicting, objectives of macro-editing.

Of the 1646 initial SD estimates, 411 final SD estimates were different from their initial equivalents resulting in 156 of the 290 final State estimates and 36 of the 49 Australian final estimates being different. Table 7.1 below indicates where the 411 altered SD estimates fell amongst the hierarchical macro-edit categories from table 6.5. It is interesting to note that the greatest concentrations were in (1,1,0), (1,0,0) and (1,1,1).

7.1 Count of altered SD estimates amongst the hierarchical categories

Hierarchical macro-edit categories	*SD estimate altered by macro-editing*		
	No	*Yes*	*Total*
000	314	53	367
001	357	50	403
010	35	7	42
011	112	23	134
100	36	25	61
101	51	15	66
110	46	34	80
111	289	204	493
Total	1,240	411	1,646

This investigation developed three main variants of H–B edits (for historical ratios) depending on the multiplicative adjustment used. The multiplicative adjustments used were sensitivity measures (18) and (19) with historical estimates used as expected estimates (State level was used for the relative sensitivity measure). The H–B macro-edit variants are denoted as:

(i) *HB1* which uses (18), the normal H–B edit multiplicative adjustment;

(ii) *HB2* which uses (19), a relative multiplicative adjustment; and

(iii) *HB all_items* which uses (19) and does not distinguish between variables.

The *HB1* and *HB2* cut-offs are by calculated on an item-by-item basis while the *HB all_items* cut-offs are calculated ignoring item. *HB1* and *HB2* use medians for $S^{*}_{\mathrm{HB}}(R_{i,d})$ calculated at either State or Australian level. State level would normally be the preferred choice for Agricultural estimates since they are most affected by location and environmental conditions. However, the option to use Australian medians was included because there were too many cases with too few estimates at State level to calculate robust medians and quartiles. Since quartiles are more affected than medians, it is the H–B cut-offs that are most affected by small numbers of contributing estimates. There were 23 State estimates with only one contributing SD estimate and 45 State estimates with only two contributing SD estimates. Only State level was needed for the *HB all_items* median ratios because there are sufficient estimates available once item is ignored. The resulting nine H-B variants are outlined in table 7.2 below. All the H–B macro-edits use $\beta = 0.05$. Various values for U were tried and we settled on $U = 0.3$ for *HB1* edits and $U = 3$ for *HB2* and *HB all_items* edits.

Tables 7.2 and 7.3 provide some comparisons of anomalous estimate selections (where each edit was tuned to select approximately 493 anomalous SD estimates). The results are compared with the 493 anomalous SD estimates previously selected by the *estimate score* edit in Section 6.5.

Table 7.3 indicates that the *HB all_items* edit (b) is closest to the *estimate score* edit (a) in terms of shared selections. The *HB2* edits (which use a relative magnitude adjustment) consistently achieved higher overlap with both the *estimate score* edit (a) and the *HB all_items* edit (b) than with the *HB1* edits (which use an absolute magnitude adjustment). Amongst the *HB2* edits, the choice of State or Australian level for calculation of medians and quartiles for the centering transformation and cut-off levels appears to have a limited effect on the level of overlap with the *estimate score* edit (a) selections. However, the best results in terms of overlap use Australian medians in the centering transformation and Australian medians and quartiles in the cut-offs.

The *estimate score* edit (a) achieved the highest level of overlap with the 411 estimates which were altered by the macro-editors (with 204 selections). The *HB all_items* edit (b) was next best performed, followed by the *HB2* edits (h) and (d).

7.2 Variants of the H–B macro-edit

Edit	*Median used for* $S^*_{\mathrm{HB},R_{id}}$	*Median and quartiles used for cut-offs*	α	*U*	*Number of SD estimate selections*
(b) HB all_items	State by item	State all items	14.50	3.0	495
(c) HB1	State by item	State by item	1.23	0.3	491
(d) HB2			1.80	3.0	491
(e) HB1	Aust by item	State by item	1.24	0.3	490
(f) HB2			1.45	3.0	492
(g) HB1	State by item	Aust by item	2.13	0.3	492
(h) HB2			10.00	3.0	494
(i) HB1	Aust by item	Aust by item	1.91	0.3	494
(j) HB2			10.20	3.0	492

7.3 Comparisons of anomaly SD estimate selections

Edit	*Number of SD estimate selections*	*Number of selections in common with the estimate score edit (a)*	*Number of selections in common with HB all_items edit (b)*	*Number of selections in common between selected HB edit pairs*	*Number of selections in common with each HB edit pair and the estimate score edit (a)*	*Number of selections that were changed by editors*
(a) Estimate score	493	493	306	N/A	N/A	204
(b) HB all_items	495	306	495	N/A	N/A	193
(c) HB1	491	205	207	255	160	131
(d) HB2	491	225	324			185
(e) HB1	490	208	198	244	155	131
(f) HB2	492	260	312			181
(g) HB1	492	227	233	228	149	139
(h) HB2	494	250	362			158
(i) HB1	494	265	211	217	175	146
(j) HB2	492	271	312			181

7.1 Relative pseudo-bias comparisons

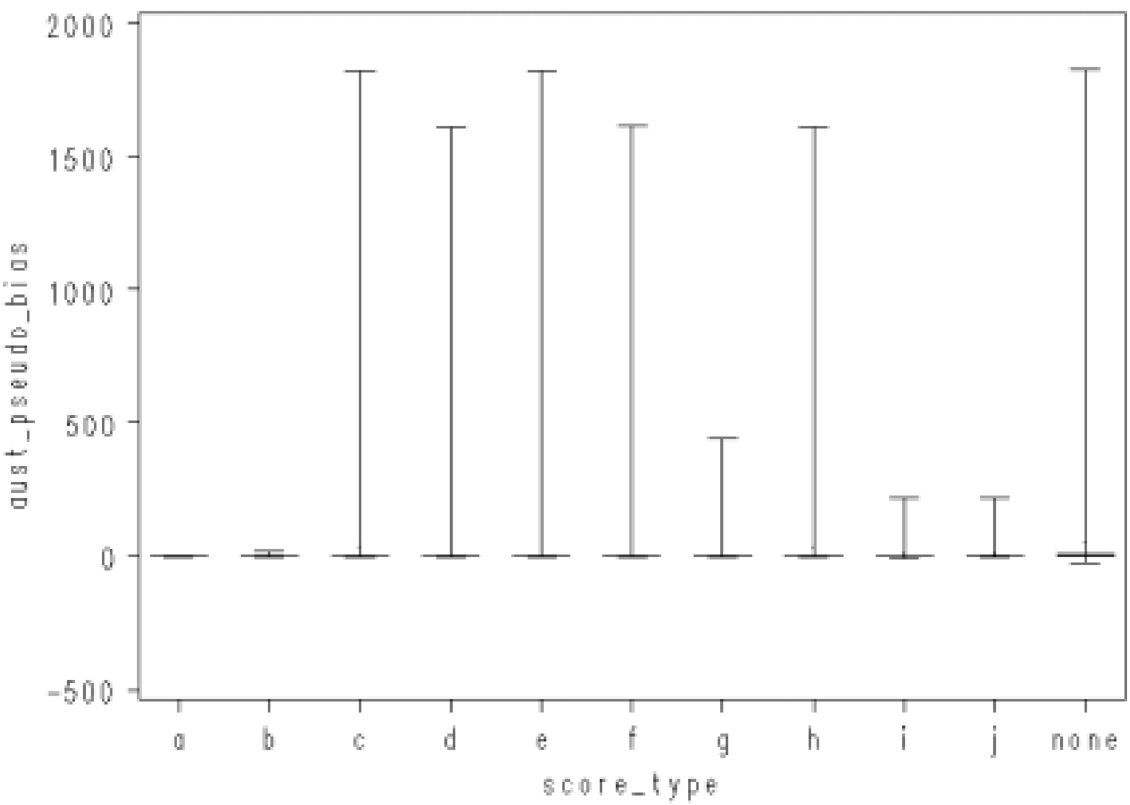
This Section examines the effectiveness of the edits in terms of changes made to the SD estimates by the macro-editors. *Relative pseudo-bias* for an edit is defined as the difference between the edited initial estimate and the final estimate expressed as a percentage of the final estimate (where the initial SD estimate is replaced by the final SD estimate if it is selected by a particular edit). *Unedited* relative pseudo-bias refers to the relative pseudo-bias associated with initial estimates when no editing has been performed. For brevity, unless a distinction is needed, "relative pseudo-bias" will be referred to simply as "pseudo-bias' in this Section.

Before discussing results, it is worth noting that macro-editing involves more than finding and correcting erroneous estimates. It also involves validating and justifying questionable correct estimates. The study dataset does not contain a flag to indicate that an estimate was examined by the macro-editors. We know that estimates with differing initial and final values have been edited but we do not know which of the remaining estimates were selected for macro-editing. Therefore, this pseudo-bias analysis only examines a component of the macro-editing scenario. The second issue requiring careful consideration is the size masking effect. For example, a large reduction in SD pseudo-bias may not be a good outcome. The reduction needs to be assessed in terms of the impact the change in SD estimates due to editing on the State and Australian estimates. More specifically, the main focus should be on the reduction in pseudo-bias at the State and Australian levels while assessing the distribution of SD pseudo-bias. A more comprehensive analysis would have incorporated criteria which takes into account the impact of the expected change to an initial SD estimate on the resulting final State and Australian estimates. However, such criteria are part of the make-up of the edits under assessment and incorporating the criteria into this assessment could distort the results towards a particular edit. For practicality, we decided to keep the SD assessment brief and simple.

7.2 Australian pseudo-bias results

Figure 7.4 below displays the spread of pseudo-bias amongst the 49 Australian estimates. It shows that the *estimate score* edit (a), *HB all_items* edit (b), HB1 edit (g), HB1 edit (i) and HB2 edit (j) gave the best results with the *estimate score* edit (a) and *HB all_items* edit (b) the standouts. Table 7.5 below lists the edit results for the top 10 Australian estimates ordered by descending absolute unedited Australian pseudo-bias values. It can be seen, for the top 10, that the *estimate score* edit (a) performs best, followed by the *HB all_items* edit (b).

7.4 Skeletal box plots of Australian pseudo-bias by edit type

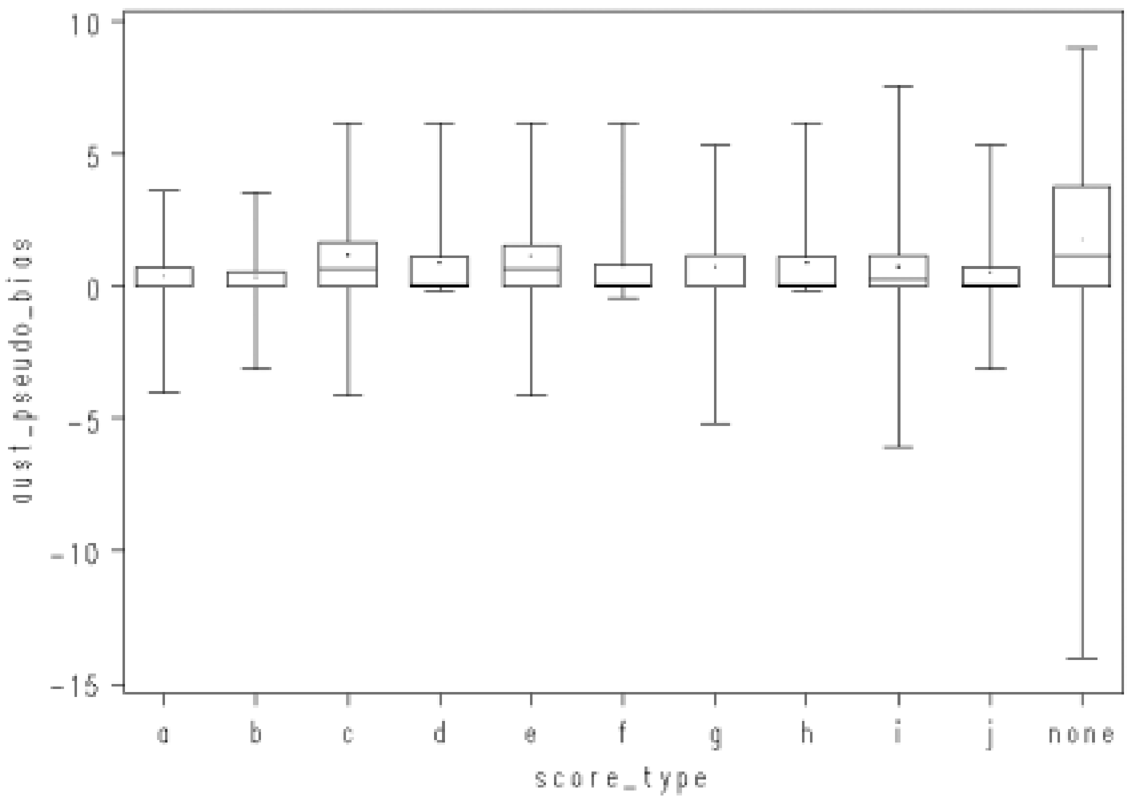


7.5 Australian pseudo-bias for the top 10 Australian estimates ordered by absolute unedited Australian pseudo-bias

		Unedited Aust pseudo-bias	Edit type (as defined in tables 7.2 and 7.3)									
Rank	*Item*		*(a)*	*(b)*	*(c)*	*(d)*	*(e)*	*(f)*	*(g)*	*(h)*	*(i)*	*(j)*
1	4304603	1,826.7	0.0	5.7	1,824.2	1,615.8	1,824.2	1,618.3	442.7	7.4	215.8	214.1
2	0100101	204.5	4.8	3.7	17.4	12.8	19.7	13.1	9.6	17.9	11.5	16.0
3	1005101	191.4	1.2	4.1	2.3	4.5	2.3	4.5	1.7	4.7	1.7	6.5
4	1918301	127.0	0.1	1.4	1.0	1.0	1.0	1.0	1.9	3.2	0.1	3.2
5	3606103	92.2	0.0	15.3	17.9	17.9	17.9	17.9	15.7	15.4	17.9	2.4
6	1918101	81.1	0.1	0.1	0.6	0.1	0.5	0.1	–0.1	0.1	–0.1	0.1
7	1902101	35.9	0.0	0.3	7.0	7.0	7.0	7.0	7.0	35.9	0.3	7.0
8	1900902	28.2	0.8	0.9	1.3	1.1	1.3	1.2	1.3	1.2	1.2	1.1
9	3605801	–24.2	0.0	0.0	–0.5	–0.5	–0.5	–0.5	0.0	0.0	0.0	0.0
10	1501701	15.0	0.6	0.7	0.6	0.4	0.7	0.7	0.7	0.4	0.7	0.7

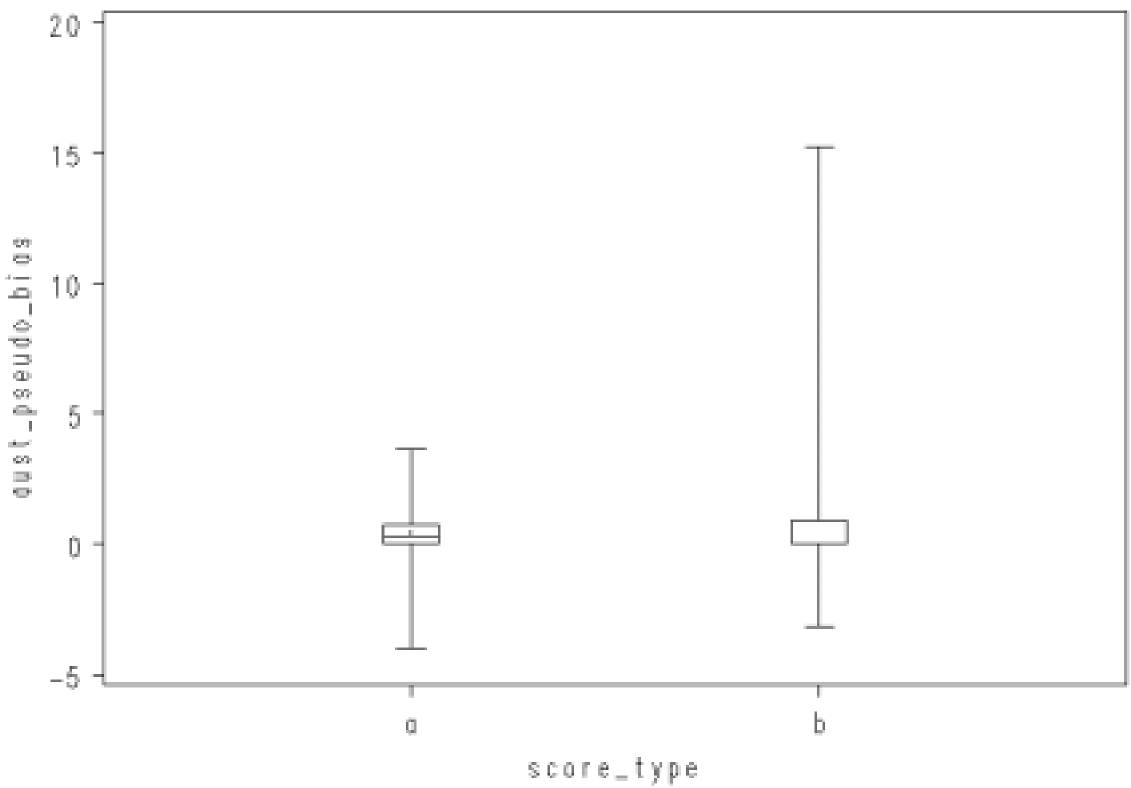
Figure 7.6 below, after removing the top 10 listed in table 7.5, displays the spread of Australian pseudo-bias (including unedited pseudo-bias) amongst the remaining 39 Australian estimates. It gives a feel for the relative sizes of the reductions in unedited pseudo-bias amongst the edits.

7.6 Skeletal box plots of Australian pseudo-bias by edit type (top 10 removed)



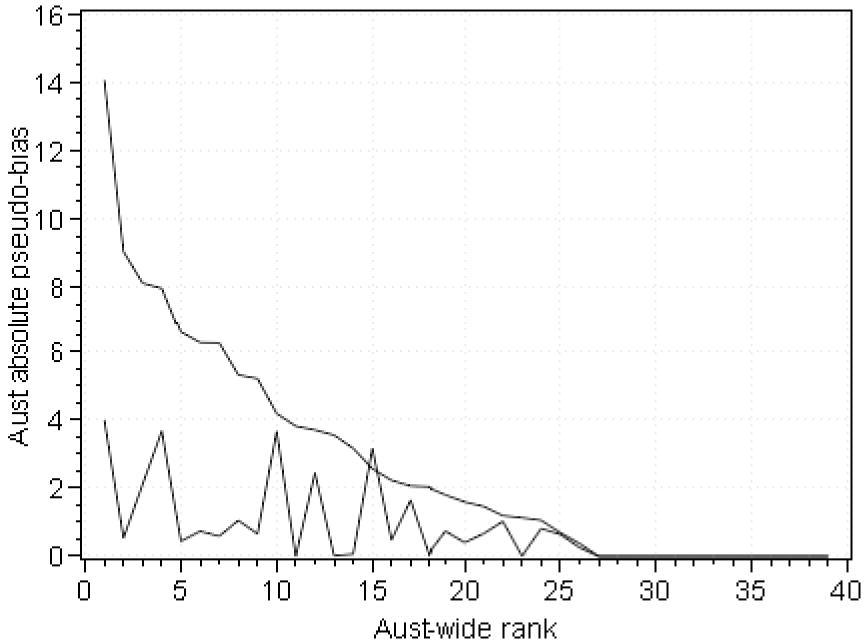
The information in figure 7.6 above needs careful consideration due to the removal of the 10 estimates with largest unedited Australian pseudo-bias. For example, figure 7.7 below gives the same display for the *estimate score* edit (a) and *HB all_items* edit (b) except that all 49 Australian estimates are included. It can be seen that the *estimate score* edit (a) performs slightly better than the *HB all_items* edit (b) at the Australian level.

7.7 Skeletal box plots of Australian pseudo-bias for the estimate score edit (a) and HB all_items edit (b) using all 49 Australian estimates

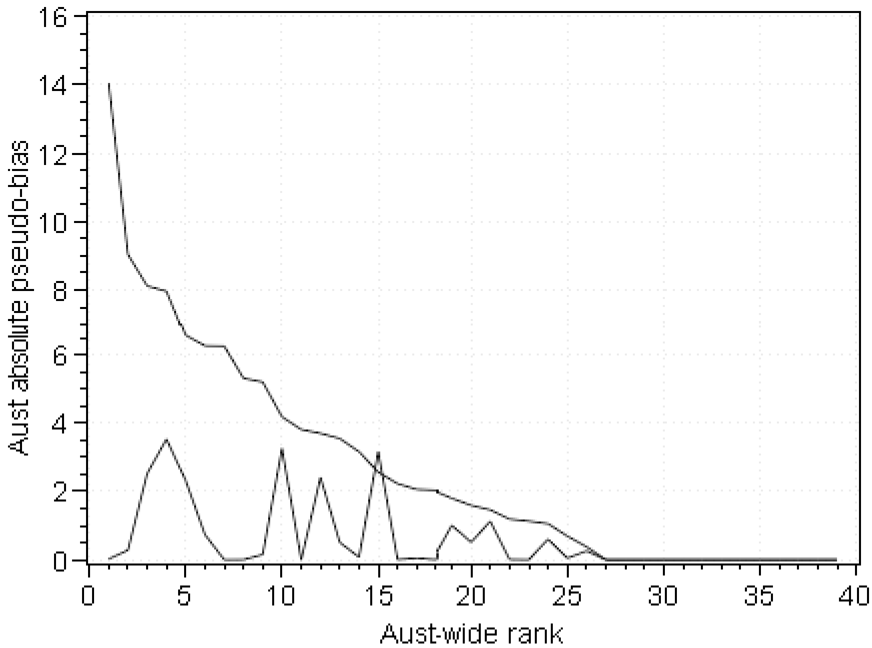


Finally, figures 7.8 and 7.9 below provide a comparison of the performances of the *estimate score* edit (a) and *HB all_items* edit (b) for the 39 Australian estimates after removing the top 10 Australian estimates listed in table 7.5. The sawtooth line represents the pseudo-bias after editing while the curved line represents the unedited pseudo-bias (that is, pseudo-bias before editing). Note that for the 15th ranked estimate, macro-editing increased pseudo-bias. The Australia-wide rank is based on a descending ordering of the absolute unedited Australian pseudo-bias value. It can be seen how the selective nature of the edits (when selecting anomalous SD estimates) influences the changes at the Australian level. The increasingly 'selective' behaviour can be seen at State and SD levels in figures 7.12, 7.13, 7.15 and 7.16 below.

7.8 Truncated absolute Australian pseudo-bias for the estimate score edit (a)



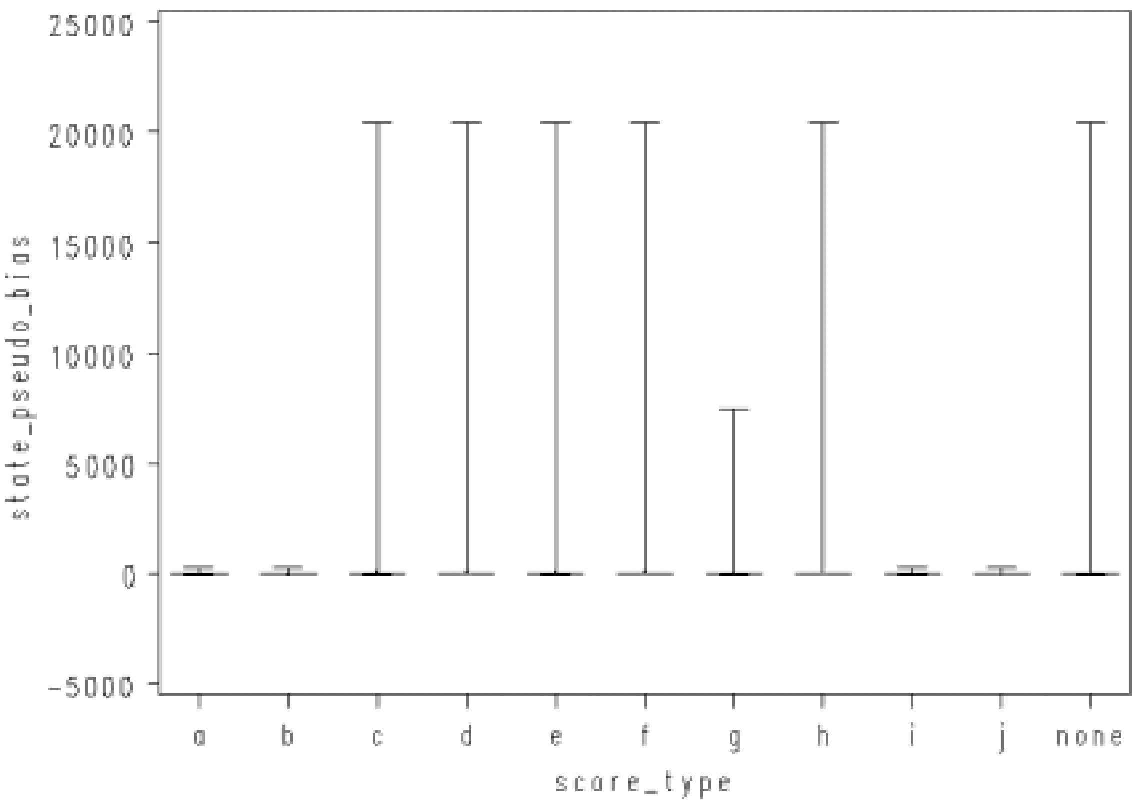
7.9 Truncated absolute Australian pseudo-bias for the HB all items edit (b)



7.3 State pseudo-bias comparisons

As discussed in the introduction to this Section, the State dimension is very difficult to assess and summarise. However, figure 7.10 below indicates that the *estimate score* edit (a), the *HB all_items* edit (b), the HB1 edit (i) and the HB2 edit (j) give the best results in terms of achieving State estimate change.

7.10 Skeletal box plots of State pseudo-bias by edit type



Tables 7.11(a) and 7.11(b) below list the edit results for the top 10 State estimates ordered by descending absolute unedited State pseudo-bias. It can be seen, for the top 10, that the *estimate score* edit (a) performs best, followed by the *HB all_items* edit (b). HB2 edit (j) is the best of the rest.

The performance of the *HB1* and *HB2* edits at State and Australian levels is related to the way they were tuned and some performed well at State level but not well at Australian level and vice versa.

7.11(a) State pseudo-bias for the top 10 State estimates ordered by absolute unedited State pseudo-bias

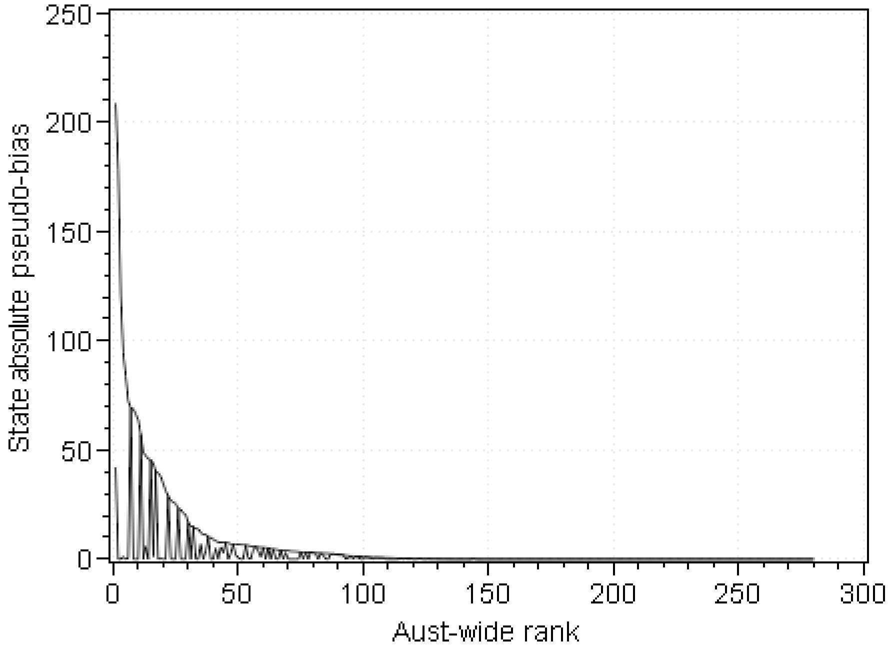
		Unedited State	*Edit type (as defined in tables 7.2 and 7.3)*				
Rank	*Item*	*pseudo-bias*	*(a)*	*(b)*	*(c)*	*(d)*	*(e)*
1	4304603	20,433.1	0.0	0.0	20,433.1	20,433.1	20,433.1
2	4304603	7,427.2	0.0	0.0	7,428.2	7,428.2	7,428.2
3	3606103	1,110.4	0.0	0.5	0.5	0.5	0.5
4	1918301	989.6	0.0	0.0	0.0	0.0	0.0
5	1918101	944.5	0.0	0.0	0.0	0.0	0.0
6	1005101	607.7	0.0	7.6	2.9	7.7	2.9
7	0100101	462.6	4.2	4.6	31.4	27.3	31.4
8	0100101	319.2	1.5	5.8	30	9.3	32.9
9	1510801	241.0	241.0	241.0	241.0	241.0	241.0
10	4304603	293.4	0.0	6.4	236.7	3.7	236.7

7.11(b) State pseudo-bias for the top 10 State estimates ordered by absolute unedited State pseudo-bias

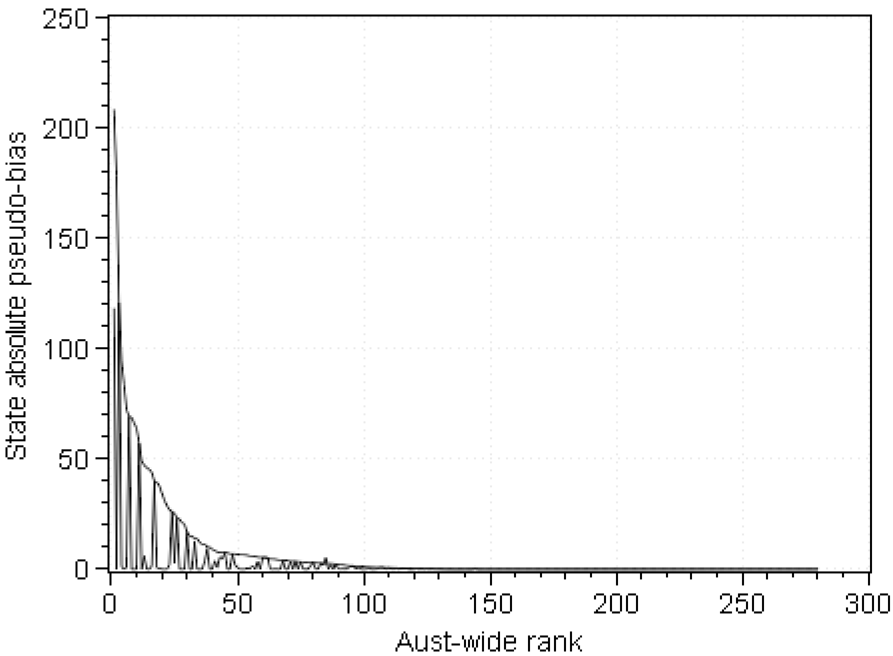
		Unedited State	*Edit type (as defined in tables 7.2 and 7.3)*				
Rank	*Item*	*pseudo-bias*	*(f)*	*(g)*	*(h)*	*(i)*	*(j)*
1	4304603	20,433.1	20,433.1	24.2	20,433.1	24.2	0.0
2	4304603	7,427.2	7,428.2	7,428.2	7,428.2	7,428.2	0.0
3	3606103	1,110.4	0.5	0.5	0.5	0.5	0.5
4	1918301	989.6	0.0	0.0	0.0	0.0	0.0
5	1918101	944.5	0.0	0.0	0.0	0.0	0.0
6	1005101	607.7	7.7	0.8	7.7	0.8	8.6
7	0100101	462.6	27.3	7.4	27.3	8.8	31.4
8	0100101	319.2	12.2	30.9	9.3	33.2	30.5
9	1510801	241.0	241.0	241.0	241.0	241.0	0.0
10	4304603	293.4	6.4	239.4	3.7	239.4	239.4

Figures 7.12 and 7.13 below provide a comparison of the performances of the *estimate score* edit (a) and *HB all_items* edit (b) for the 280 State estimates after removing the top 10 State estimates listed in tables 7.11(a) and 7.11(b). The Australia-wide rank is based on a descending ordering of the absolute unedited State pseudo-bias values ignoring State.

7.12 Truncated absolute State pseudo-bias for the estimate score edit (a)



7.13 Truncated absolute State pseudo-bias for the HB all items edit (b)



7.4 SD pseudo-bias comparisons

As discussed in the introduction to this Section, the SD results presented are very basic and need careful consideration due to the problem of size masking. The following results are designed only to provide some insights from a few different angles.

Table 7.14 below indicates edit performance for SD estimates with an absolute unedited SD pseudo-bias greater than 300%. It can be seen that the *estimate score* edit (a) performed best, followed by the *HB all_items* edit (b) and *HB 2* edit (j) in terms of reducing the largest of the SD pseudo-biases.

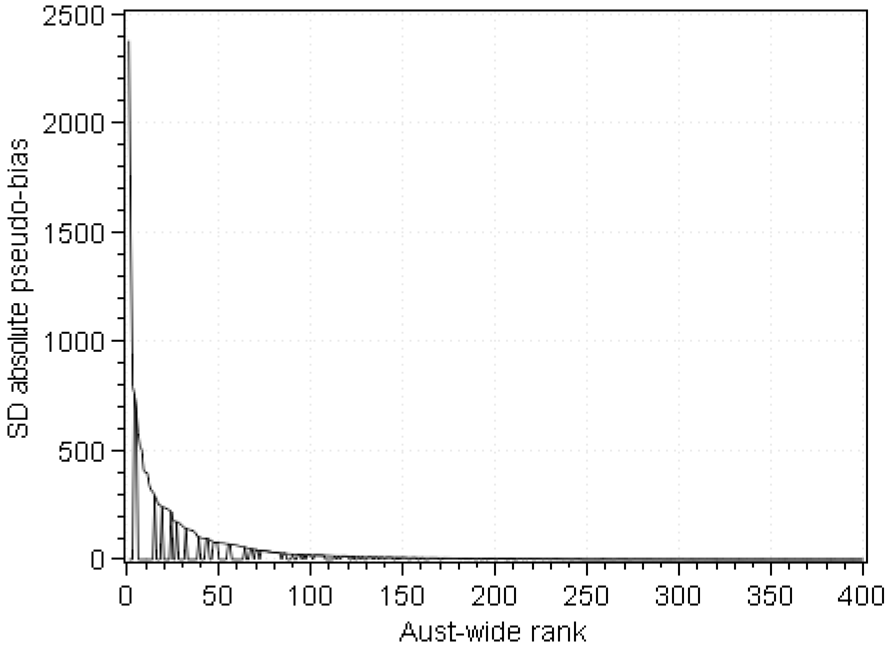
7.14 Edit results for SD estimates with absolute unedited SD pseudo-bias above 300%

					Edit type (as defined in tables 7.2 and 7.3)									
Rank	*Item*	*SD*	*State*	*SD relative pseudo-bias (%)*	*(a)*	*(b)*	*(c)*	*(d)*	*(e)*	*(f)*	*(g)*	*(h)*	*(i)*	*(j)*
1	3606101	225	2	499,903										
2	0100101	120	1	54,369										
3	4304603	125	1	35,090			X	X	X	X				X
4	0100101	505	5	21,597										
5	4304630	535	5	8,520			X	X	X	X	X			
6	1918301	310	3	5,006										
7	1918101	610	6	4,095										
8	0100101	330	3	3,763										
9	4304603	345	3	3,112		X				X	X	X	X	X
10	0100101	605	6	3,016										
11	1005101	240	2	2,376										
12	4304603	310	3	1,502		X	X	X	X	X	X	X	X	X
13	3606101	155	1	794		X	X	X	X	X	X	X	X	
14	1005101	105	1	758	X	X		X		X		X		X
15	1005102	340	3	683	X	X	X	X	X	X	X	X	X	X
16	0100101	355	3	580										
17	1918101	310	3	518										
18	1005102	505	5	504			X	X	X	X		X		
19	3606102	505	5	409			X	X	X	X				
20	0100101	320	3	399			X	X	X	X		X		
21	1900902	410	4	398										
22	1005101	330	3	343										
23	1809101	535	5	318				X		X		X		X
24	1900902	225	2	311										
Number of SD estimates uncorrected					2	5	8	10	8	11	5	8	4	6

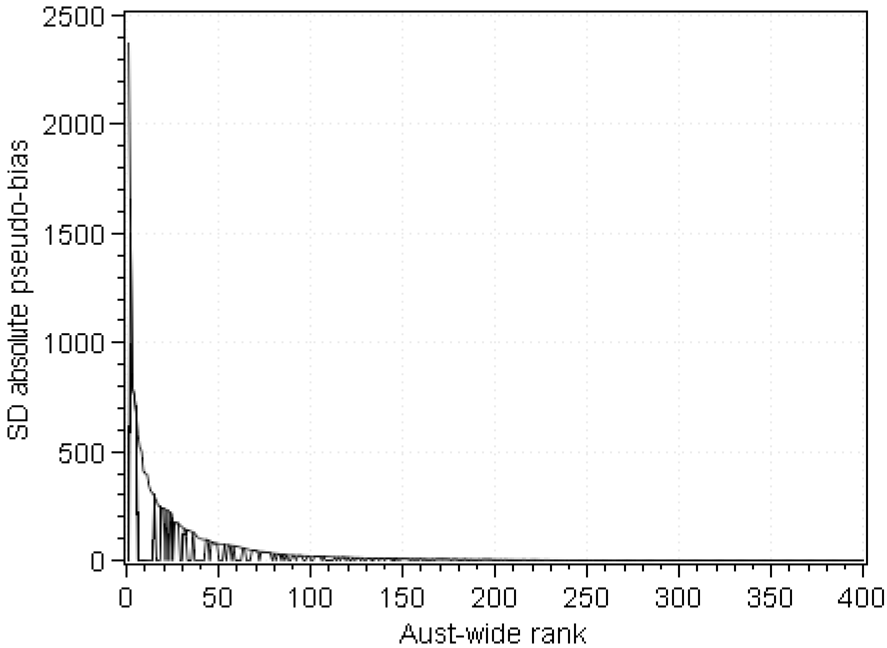
'X' indicates that the edit did not select (and correct) the particular erroneous SD estimate.

Figures 7.15 and 7.16 below provide a comparison of the performances of the *estimate score* edit (a) and *HB all_items* edit (b) for the 401 SD estimates which were altered by the macro-editors (after removing the top 10 SD estimates listed in table 7.14). The Australia-wide rank is based on a descending ordering of the absolute unedited SD pseudo-bias values.

7.15 Truncated absolute SD pseudo-bias for the estimate score edit (a)



7.16 Truncated absolute SD pseudo-bias for the HB all items edit (b)



8. SUMMARY AND CONCLUSIONS

This paper demonstrates that the macro significance editing framework can be used to develop a score-based macro-editing methodology for business surveys. It has the advantage that it dovetails with the existing micro significance editing framework already in operation in the ABS, forming a general significance editing framework. The functionality of SEE can be extended to cover the objective detection component of macro-editing. The general definition of significance and the framework developed in this paper allows for new scores such as ratio scores which incorporate historical estimates for calculating current ratio scores. The framework allows for estimates such as standard errors for sample surveys and coefficients of variation for censuses to be included in scores and for scores to be combined. This allows scores such as macroscores and combined estimate and ratio scores to be developed. The scores make use of macro-editor expectations for the data when available, though scores can also be developed when expectations for the data are not available.

Hierarchical scores and macro-edits provide very useful tools for addressing swamping and masking problems and appear to be viable alternatives to the H–B macro-edit variants for both historical and current ratios. They are easy to understand and encourage editors to interact with the data (particularly for movements in estimates). They use explicit manually-chosen edit boundaries (cut-offs) which allow for flexibility in dealing with conflicting macro-editing priorities.

The significance framework encourages efficient use of editor resources by allowing editing managers to make informed decisions about what to edit and how much to edit. The use of ranks in the framework is an important element in this regard. The framework supports the use of simple manually-chosen interactive cut-offs and graphical displays such as graphs of score versus rank and cost/benefit graphs. These help to visualise the macro-editing workload and have the advantage that they can be characterised and ordered by their GINI index value.

The H–B macro-edits are a viable alternative to some of the macro significance scores and approaches and can be applied in many situations. The H–B macro-edits would be a useful option to include in ABS macro-editing tools. They use dynamic two-sided cut-offs which are automatically generated (once the user defines fence widths) and provide an alternative to the interactive one-sided cut-off approach generally used with macro significance editing. If H–B macro-edits are to be used, it is recommended that the three variants analysed in this paper be explored further (particularly the *HB all_items* version).

However, H–B macro-edits have some limitations. In their current form, they can only be applied to strictly positive or strictly negative estimates. They do not rank anomalous estimates and do not encourage examination of the data. They need to be tuned and it is not clear how to decide that a specific tuning is optimal. They have a black box feel and are difficult to explain. It is possible that macro-editors may be reluctant to accept them. They are less robust as the number of estimates within the domain of study decrease and this could be a problem for some Australian collections. They do not use macro-editor estimate expectations and, when expected estimates are available, significance scores should provide more powerful alternatives.

The results in this paper suggest that the estimate and ratio scores from macro significance editing provide excellent alternatives to the H–B ratio scores when expected estimates are available. In fact, the estimate score covers the H–B historical and current ratio scores while the ratio score extends the scoring of ratios beyond the H–B macro-edits. This paper recommends that macro significance editing be developed in the ABS (particularly the hierarchical macro-edits). This paper also recommends that H–B macro-edits be implemented as an alternative to macro significance editing for ABS macro-editors.

ACKNOWLEDGEMENT

The author wishes to express his gratitude to Peter Rossiter from the Australian Bureau of Statistics for his invaluable assistance and advice in arranging this report into its current form.

REFERENCES

Australian Bureau of Statistics (2007) *The Editing Guide: The Key to Validating Data from Businesses*, Beta version 4, ABS, Canberra.

Australian Bureau of Statistics (2011, to appear) *Significance Editing Manual*, ABS Corporate Manuals, ABS, Canberra.

Banim, J. (2000) "An Assessment of Macro Editing Methods", UN/ECE Work Session on Statistical Data Editing, Cardiff, United Kingdom, 18–20 October, 2000. (last viewed on 24 September 2010) <http://www.unece.org/stats/documents/2000/10/sde/7.e.pdf>

Farwell, K. (2004) "The General Application of Significance Editing to Economic Collections", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.066, Australian Bureau of Statistics, Canberra.

Farwell, K. (2005) "Significance Editing for a Variety of Survey Situations", Paper presented at the 55th Session of the International Statistical Institute, Sydney, 5–12 April.

Farwell, K. (2009a) "Hierarchical Macro-editing", Technical Report, Australian Bureau of Statistics, Canberra.

Farwell, K. (2009b) "Macro Significance Editing Technical Specifications", Technical Report, Australian Bureau of Statistics, Canberra.

Farwell, K.; Poole, R. and Carlton, S. (2002) "A Technical Framework for Input Significance Editing", Conference paper for DataClean2002, Jyvaskyla, Finland.

Gather, U. and Becker, C. (1997) "Outlier Identification and Robust Methods", in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics, Volume 15: Robust Inference*, pp. 123–143, Elsevier, Amsterdam.

Hidiroglou, M.A. and Berthelot, J.M. (1986) "Statistical Editing and Imputation for Periodic Business Surveys", *Survey Methodology*, 12(1), pp. 73–83.

Hoaglin, D.C.; Mosteller, F. and Tukey, J.W. (eds.) (1983) *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

Maimon, O. and Rokach, L. (eds.) (2005) *The Data Mining and Knowledge Discovery Handbook*, Springer.

Samprit, C.; Hadi, A.S. and Price, B. (1999) *Regression Analysis by Example*, Edition 3, John Wiley and Sons Inc..

Sigman, R.S. (2005) "Statistical Methods Used to Detect Cell-Level and Respondent-Level Outliers in the 2002 Economic Census of the Services Sector", *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Sugavanam, R. (1983) "A Statistical Edit for Change", Technical Report, Statistics Canada.

Thompson, K.J. (1999) "Ratio Edit Tolerance Development Using Variations of Exploratory Data Analysis (EDA) Resistant Fences Methods", *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. (last viewed on 24 September 2010) <http://www.fcsm.gov/99papers/thompson.pdf>

Thompson, K.J. (2007, to appear) "Investigation of Macro Editing Techniques for Outlier Detection in Survey Data", *Proceedings of the Third International Conference on Establishment Surveys*, American Statistical Association.

Thompson, K.J. and Ozcoskun, L. (2007) "An Empirical Investigation into Macro Editing", *2007 Federal Committee on Statistical Methodology Research Conference Papers*, (last viewed on 24 September 2010) <http://www.fcsm.gov/07papers/Thompson.III-B.pdf>

APPENDIX

A. THE JACKKNIFE METHOD FOR DERIVING HIERARCHICAL IMPACT

Hierarchical ratio scores are derived using a Jackknife *drop–1* approach. The hierarchical macro significance score attempts to measure the impact of the difference between what was expected and what was observed at the base level with respect to what was expected at target level. For example, say we have two estimates of total for base class B, $Y_{i,B}$ and $Y_{j,B}$, for variables i and j (and that the base estimates are strictly positive). The drop–1 approach involves removing, for a given base estimate, its expected contribution to the expected target estimate and replacing it with the observed contribution to create an adjusted expected target estimate. We then calculate the difference between the original expected target estimate and the adjusted expected target estimate. This is used as the measure of hierarchical impact and it is expressed relative to the expected target estimate or standard error multiple. For example:

The hierarchical estimate score for $Y_{i,B}$ is:

$$\begin{aligned} S_{B,\text{target}}(Y_{i,B}) &= 100 \times \frac{\left(Y^*_{i,\text{target}} - Y^*_{i,B} + Y_{i,B}\right) - Y^*_{i,\text{target}}}{Y^*_{i,\text{target}}} \\ &= 100 \times \frac{\left(Y^*_{i,\text{target}} + \Delta Y_{i,B}\right) - Y^*_{i,\text{target}}}{Y^*_{i,\text{target}}} \\ &= 100 \times \frac{\Delta Y_{i,B}}{Y^*_{i,\text{target}}} \end{aligned}$$

where, for variable k, $$\Delta Y_{k,B} = Y_{k,B} - Y^*_{k,B}$$

and the hierarchical ratio score for $R_{i,j}$ is:

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

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

The drop–1 method provides a more accurate measure of impact than alternatives based on Taylor Series linearising approximations when there are few contributors.

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au