1351.0.55.038

Australian Bureau of Statistics

Research Paper

Socio-Economic Indexes For Areas: Robustness, Diversity Within Larger Areas and the New Geography Standard

www.abs.gov.au

New Issue

Research Paper

Socio-Economic Indexes For Areas: Robustness, Diversity Within Larger Areas and the New Geography Standard

Peter Radisich and Phillip Wise

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 05 MAR 2012

ABS Catalogue no. 1351.0.55.038

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Dr Phillip Gould, Analytical Services Branch, on Canberra (02) 6252 5315 or email <analytical.services@abs.gov.au>.

CONTENTS

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ASGC	Australian Standard Geographical Classification
ASGS	Australian Statistical Geography Standard
CD	Collection District
CED	Commonwealth Electoral Division
Census	Australian Census of Population and Housing
IEO	Index of Education and Occupation
IER	Index of Economic Resources
IRSAD	Index of Relative Socio-economic Advantage and Disadvantage
IRSD	Index of Relative Socio-economic Disadvantage
LGA	Local Government Area
MB	Mesh Block
NHS	National Health Survey
PCA	Principal Component Analysis
POA	Postal Area
SEIFA	Socio-Economic Indexes For Areas
SA1	Statistical Area Level 1
SED	State Electoral Division
SLA	Statistical Local Area
SSC	State Suburb

SOCIO-ECONOMIC INDEXES FOR AREAS: ROBUSTNESS, DIVERSITY WITHIN LARGER AREAS AND THE NEW GEOGRAPHY STANDARD

Peter Radisich and Phillip Wise
Analytical Services Branch

ABSTRACT

Socio-economic indexes for areas (SEIFA) summarise the socio-economic advantage and disadvantage of areas using information from the Census of Population and Housing. A new set of indexes are released after each Census, with the most recent being the SEIFA 2006 indexes released in March 2008. Since the latest release, the ABS has conducted additional research to increase understanding of the indexes and to make improvements for SEIFA 2011.

This paper has three main purposes: to explore the robustness of the SEIFA indexes to the influence of specific variables and areas; to investigate ways of representing the diversity of socio-economic characteristics within larger areas; and to explore the impact of the new geography standard on the indexes. The analysis in this paper is based on 2006 Census data.

The ABS has produced a number of papers which provide more general information about SEIFA. This paper seeks to build on this discussion of SEIFA, and the reader should consider the issues in this paper in conjunction with previous ABS publications on the subject.

1. INTRODUCTION

Socio-economic indexes for areas (SEIFA) are a suite of four indexes designed to measure the relative socio-economic advantage and disadvantage of areas using data from the Census of Population and Housing (which will be referred to as the Census throughout this paper). The ABS has been producing SEIFA in its current form for every Census since the 1986 Census, and the ABS is planning to continue this for the 2011 Census. The notion of relative socio-economic advantage and disadvantage used in SEIFA is broadly defined as:

> People's access to material and social resources and their ability to participate in society; relative to what is commonly experienced or accepted by the wider community.

ABS (2008a, 2008b) and Adhikari (2006) provide a much more extensive analysis and explanation of the notion of disadvantage used in SEIFA. The reader is encouraged to read these papers prior to conducting an analysis which uses SEIFA.

The four indexes which comprise SEIFA are:

- Index of Relative Socio-economic Disadvantage (IRSD),
- Index of Relative Socio-economic Advantage and Disadvantage (IRSAD),
- Index of Economic Resources (IER),
- Index of Education and Occupation (IEO).

Each index captures a different aspect of the notion of advantage and disadvantage, and each index is composed of slightly different variables. Appendix A provides a brief description of the indexes, together with the variables used in their construction.

Historically, the indexes were calculated first at the Census Collection District (CD) level of geography, which is the smallest unit of geography for Census data. Each CD has a population of approximately 500 people, and there are approximately 38,000 CDs for the 2006 Census. The use of CDs is changing for the 2011 Census, where a new geographical standard is being used, the Australian Statistical Geography Standard (ASGS). The impact of this change on SEIFA is discussed in Section 4.

Each index combines the information from its component variables using Principal Components Analysis, described in Section 2. This information is then used to create a score for each CD, which can be used to rank and compare areas in terms of their relative socio-economic advantage and disadvantage. To aid interpretation, the CD level index scores have been standardised to have an average score of 1000 and a standard deviation of 100 across Australia.[1] The ABS calculates the index scores for

[1] Standardisation does not affect the ranking of the CDs.

larger geographic units by taking a population weighted average of the constituent CDs scores.[2]

In this paper, frequent reference will be made to quantities called deciles, such as "the CD had a SEIFA score in decile 2". To explain, each CD is ranked according to their SEIFA score from lowest to highest, and placed into one of 10 ordered groups, called deciles, such that decile 1 contains the lowest 10 per cent of CD scores, decile 2 contains the next lowest 10 per cent of CD scores, and so on up to decile 10 which contains the highest 10 per cent of CD scores.

The remainder of this paper is divided up into three main sections which cover distinct issues, so that the paper does not need to be read sequentially. The analysis in this paper is based on 2006 Census data. Section 2 presents some of the research the ABS has conducted and also examines future directions being considered in creating measures of robustness for SEIFA. Section 3 discusses the diversity of socio-economic characteristics within larger areas (larger than CDs), and presents an alternative way of measuring disadvantage for larger areas which is more suited to capturing disadvantaged subpopulations in diverse areas. This section is intended to build on the examples of how to appropriately use and interpret the SEIFA indexes already available in Adhikari (2006) and ABS (2008a and 2008b). Section 4 discusses the impact of the ASGS on SEIFA 2011 which will be implemented for the 2011 Census – see ABS (2011a; 2011b; 2008c). The section includes an experimental analysis comparing one of the CD level SEIFA indexes to a corresponding index calculated for an approximation of the 2009 Statistical Area Level 1 (SA1) geographic unit.

[2] The index scores for larger areas are not re-standardised, and so will not have a mean of 1000 and standard deviation 100 across Australia.

2. MEASURES OF ROBUSTNESS FOR SEIFA

Since the release of the SEIFA 2006 indexes, the ABS has conducted additional research into the indexes. In this section we summarise this research, and focus on assessing the robustness of SEIFA by investigating the sensitivity of the indexes with respect to the inclusion of particular variables and areas, and also considering the impact of small perturbations to the data.

In order to measure the robustness of SEIFA, this paper presents three approaches. The extent to which removing an area from the index affects the ranking of CDs is explored in Section 2.2, while the impact of removing a variable from the index, and how this affects the ranking of CDs, is explored in Section 2.3. Section 2.4 discusses the impact of confidentialisation of Census data on a user-created index.

Before proceeding further, it is necessary to first understand how the indexes are constructed.

2.1 Constructing SEIFA

This section contains a brief description of how the SEIFA indexes are constructed – refer to ABS (2008a; 2008b) for more information. The SEIFA indexes are calculated at the CD level using a technique called principal components analysis (PCA) – see Jolliffe (1985) for details. All of the variables in each index are proportions. Each variable has a numerator count and a denominator count at the CD level. Before employing the PCA method we standardise the variable proportion by subtracting the average proportion from each CD, and then dividing the result by the standard deviation of the proportion. PCA is used to determine the weights for loading the standardised variables on the index. The index itself is defined as the first principal component[3] of the data – readers should see ABS (2008b) for more details. The PCA derived weights are used to create a raw index score for each CD[4], which can be expressed as follows

$$Y_i = x_{i1}w_1 + x_{i2}w_2 + \ \ldots \ + x_{iP}w_P ,$$

where:

- Y_i is the raw index score for the i-th CD,[5]
- x_{ik} is the standardised variable value of the k-th variable for the i-th CD,

---

3 The first principal component is the weighted linear combination designed to capture the maximum amount of variation present in the standardised variables (ABS 2008b).

4 See Appendix A for the weights for each variable.

5 This raw score is then re-scaled so that it has an average score of 1000 and a standard deviation of 100 across Australia. Note that this raw score is also equal to the first principal component score.

- w_k is the weight for the k-th standardised variable, determined from the first principal component in the principal components analysis, and
- P is the number of variables in the index.

One of the benefits of using PCA is that it allows the complex relationships between the variables in each index to determine the relative importance of each variable. However, the weights derived from PCA can be sensitive to atypical CDs (Croux and Haesbroeck, 2000); the extent of this sensitivity is explored in Section 2.2. Additionally, the raw score and weights determined by PCA depend on the combination of variables used. For instance, if one variable is removed from an index, then the weights for the remaining variables will change.

2.2 Robustness with respect to influential and atypical areas

Census data contains information about every person, family, and household in Australia.[6] Thus it is certain to contain numerous areas which are atypical compared to Australia as a whole. In order to deepen understanding of the SEIFA indexes, and establish confidence in how they can be used in particular applications, the question of whether atypical CDs have a large impact on the SEIFA variable weights needs to be answered.

One way to consider this is to remove a CD, re-calculate the SEIFA variable weights, and then measure the difference between the new variable weights and the weights including that CD. However, we can make very accurate approximations to the changes in variable weights using a linearization technique known as the 'influence function' (Critchley, 1985; Jolliffe, 1986; Brooks, 1994; Shi, 1997; Croux and Haesbroeck, 2000). The influence function is a linear approximation to the actual change in variable weights resulting from removing an area from the index. Appendix B contains further information on the mathematics underpinning the influence function. The approximation given by the influence function is accurate to a constant divided by the number of observations, and for CD level SEIFA (with a large number of observations N = 37, 457) is effectively equal to the exact change. The sum of squared changes to each weight for each CD was calculated, and this was used to rank the CDs in terms of their influence. Jolliffe (1986) gives a geometrical interpretation of the sum of squared changes.[7]

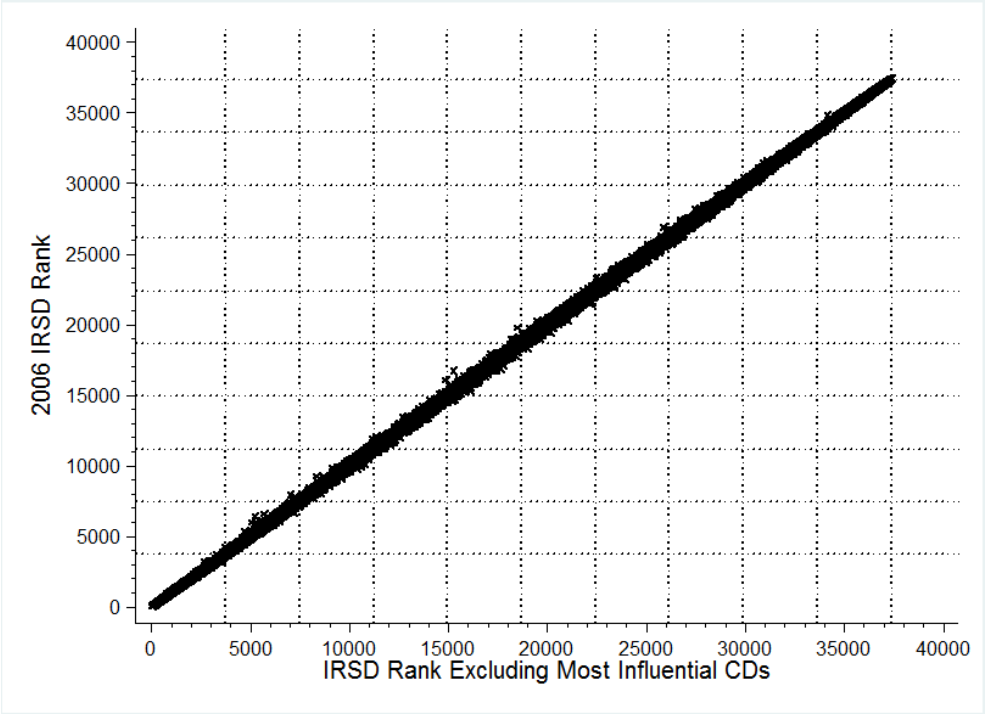
---

6 In theory, every person present in Australia on Census night, excluding foreign diplomats and their families, should have been included on a Census form at the place where they stayed. However, in practice there may be some people who are missed and some who are doubled counted; this is called the Census undercount.

7 If the weights are considered as describing a straight line, then the sum of squared changes in the weights is an increasing function of the angle between the two lines described by each set of weights.

Our analysis of the most influential areas showed that due to the large number of CDs covered by SEIFA, it is difficult for any single CD or small group of CDs to exert a high influence on the weights. To illustrate this point, figure 2.1 shows the 2006 IRSD rank against corresponding IRSD rank after the 100 most influential CDs (as determined using the influence function established previously) were removed from the calculation of the IRSD variable weights. Decile lines have been included in figure 2.1 to aid interpretation. Practically all points lie close to the 45 degree line, indicating that the IRSD is robust with respect to atypical and outlying areas. Similar results to that shown in figure 2.1 were observed for the other three indexes.

2.1 Comparison of IRSD ranks with and without the 100 most influential observations used to calculate variable weights



2.3 Robustness with respect to variable inclusion

SEIFA indexes can be sensitive to the particular combination of variables used to create them. Each variable must have a conceptual relationship to the notion of relative socio-economic advantage and disadvantage to be considered for inclusion. Additionally, ABS (2008a) states "*any measure of disadvantage will only reflect the information which was put into it*" (p. 18). This section is geared towards answering the question: how much do the rankings of CDs depend on particular variables being included in the index?

Our approach to assessing the impact of the variables on each index was to drop one variable, then use the remaining variables to rebuild the index from scratch, conducting PCA to obtain new weights from which the index rankings could be recalculated. The different sets of rankings were then compared.

To illustrate, consider the Index of Education an Occupation (IEO). There are nine variables in the IEO, one of which is *percentage of people aged 15 years and over at university or other tertiary institution* (ATUNI).[8] To assess the impact of this variable on the IEO, the variable ATUNI was removed and then an index ranking was re-calculated based on the remaining eight variables. So each CD now has two rankings: one rank using all nine variables and one rank where ATUNI has been removed from the index. We took the difference in the rankings to be our measure of how much the variable ATUNI impacted the index. So, if a particular CD has an IEO rank of 10,000 when all variables are included; and a rank of 11,425 when the variable ATUNI is removed, then our measure of influence would be 1,425. For easier interpretation, the change in ranks is expressed as a percentage of the total number of ranks (37,457), so this would be a change of 1,425 / 37,457 = 3.8 per cent of ranks.

The process outlined above was repeated for each variable in the index. The above CD will now have a set of changes in addition to 1,425 such as (1,425, 1,048, –20, 56, –369, 35, 80, –534, 2,056). So each CD has a ranking using all variables, and one difference in ranking due to dropping each variable. To measure the effect of this, we focus on the maximum difference in ranks for a CD caused by removing a variable. For the above example, the CD would have a maximum difference of 2,056 or 5.5 per cent of ranks.

Interpreting the results of such an analysis requires a degree of caution. The SEIFA indexes are intended to be general measures of disadvantage, so the ranks should be reflecting only multiple indicators of advantage or disadvantage. Ideally, the index should not be too sensitive to any single indicator, nor should it be insensitive to any single indicator.

Table 2.2 shows the maximum deviation from the published rank for each index. This represents an 'extreme case' scenario of the sensitivity of each index to removing a single variable. It shows the proportion of CDs which have maximum changes from their published rank within certain ranges. For example, under the IRSAD column and the second row is 51.4, indicating that 51.4 per cent of CDs had a maximum change in ranks due to removing a variable of between 2 per cent and 5 per cent of ranks, or between 749 and 1,873 ranks.

---

8 See Appendix A for a list of the variables in each index, as well as their associated mnemonics.

2.2 Maximum change in ranks due to removing a single variable

	SEIFA Index			
Maximum change in ranks due to dropping a variable	*IRSAD (% of CDs)*	*IRSD (% of CDs)*	*IER (% of CDs)*	*IEO (% of CDs)*
≤ 2% of ranks	37.1	20.4	13.3	20.1
> 2% and ≤ 5% of ranks	51.4	48.5	35.2	43.2
> 5% and ≤ 10% of ranks	10.5	26.3	36.1	28.7
> 10% and ≤ 20% of ranks	1.0	4.3	14.2	7.2
> 20% of ranks	0.1	0.4	1.1	0.8

According to table 2.2, each of the indexes is generally robust to removing a variable from the index, with the vast majority of CDs having maximum changes under 10 per cent of ranks. It also shows that there are a small percentage of CDs whose rankings are quite sensitive to the removal of particular variables from the SEIFA indexes. These CDs tended to lie around the middle of the distribution, in deciles 3–8.

There are differences between each of the indexes, with IRSAD and IRSD being the least sensitive. This is to be expected, as the IRSAD and IRSD are more general indexes and have a greater number of variables compared to IER and IEO.

2.4 Effects of confidentiality on user-created indexes

Users can create their own socio-economic indexes, but if they want to use the Census data then they must use confidentialised data. The SEIFA indexes are created within the ABS using unconfidentialised data.

When Census data is confidentialised, small perturbations are made to the Census data items. This analysis examines the sensitivity of a SEIFA index to the small perturbations introduced as part of confidentialising Census data. Specifically, the questions of interest we ask are:

- How closely can a confidentialised user-created index mimic one based on unconfidentialised data?
- How sensitive are the indexes to small perturbations in the underlying variable proportions?

This section presents an assessment of the impact on an index of using confidentialised data.

2.4.1 The construction of a SEIFA index using confidentialised Census data

In order to carry out the analysis in this section a process was developed to replicate the experience users would have in reconstructing a SEIFA index using confidentialised Census data. The index creation process directly followed that of a standard SEIFA index, except that the counts of the numerators and denominators of Census data were confidentialised.[9] See ABS (2008a, Chapter 3) for further information on SEIFA index construction.

The IRSAD index was chosen as the basis for the investigation. The analysis was performed at the Census Collection District (CD) level, as per the SEIFA index construction. So that we could compare the results to the index based on unconfidentialised data, only the 37,457 CDs with a SEIFA 2006 score were included in the analysis. Section 2.4.2 presents some of the results of this comparison.

2.4.2 Index comparison results

After the index was calculated using the confidentialised Census data, the CDs were re-ranked into deciles according to this new index. Each CD now has two decile rankings; one based on unconfidentialised data, and one based on confidentialised data. Table 2.3 provides a comparison of these two rankings. If there are minor changes then most of the frequencies should lie across the diagonal part of the table, from top left to bottom right.

2.3 IRSAD CD decile from unconfidentialised Census Data by IRSAD CD deciles from confidentialised Census data

Original IRSAD CD level decile	*IRSAD CD level decile using confidentialised Census data*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Total*
1	3,549	192	4	0	0	0	0	0	0	0	3,745
2	192	3,274	279	1	0	0	0	0	0	0	3,746
3	1	270	3,137	332	6	0	0	0	0	0	3,746
4	2	9	314	3,036	374	11	0	0	0	0	3,746
5	1	1	10	365	3,014	344	9	0	1	0	3,745
6	0	0	0	12	341	3,076	312	4	1	0	3,746
7	0	0	2	0	10	306	3,175	251	2	0	3,746
8	0	0	0	0	0	8	243	3,288	207	0	3,746
9	0	0	0	0	0	1	7	203	3,380	155	3,746
10	0	0	0	0	0	0	0	0	155	3,590	3,745
Total	3,745	3,746	3,746	3,746	3,745	3,746	3,746	3,746	3,746	3,745	37,457

[9] All variables in SEIFA are *proportions*. Each variable has a numerator count and a denominator count which is derived directly from the Census data. See ABS (2008a, 2008b) for details on how the SEIFA variables were created.

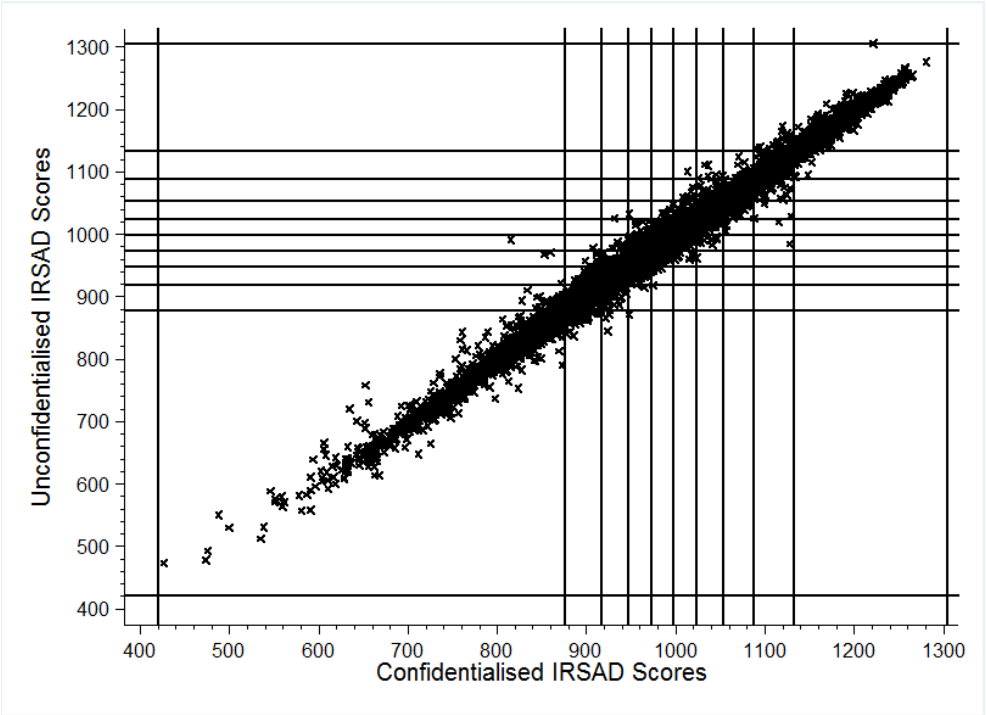
The comparisons of the two indexes revealed that most CDs remained within 1 decile of the original IRSAD decile based on unconfidentialised Census data. However, in a few cases there was a degree of discrepancy between the IRSAD deciles based on confidentialised Census data when compared to the original deciles. For example, table 2.3 shows that a CD with a published IRSAD score in decile 7 could potentially have a score between decile 3 to 9 using the index based on confidentialised Census data.

Figure 2.4(a) and table 2.4(b) together show the impact on the CDs of the confidentialisation process. Decile lines are included on figure 2.4(a) to aid interpretation. Figure 2.4(a) shows that CDs in the most advantaged and most disadvantaged deciles were not affected as much as the CDs across the middle of the CD distribution. This indicates that the tails of the distribution are more robust to the changes brought about by using confidentialised data. The bow shape evident in the figure highlights the effect of the confidentialisation process across the middle deciles of the score distribution. Table 2.4(b) uses percentiles to provide further information about the CD ranking changes. A percentile is one of 100 ordered groups based on the ranks of the CD scores; percentile 1 contains the lowest 1 per cent of CD scores, whilst percentile 100 contains the highest 1 per cent of CD scores. The reason for reporting CD percentile changes is because deciles can be susceptible to a boundary effect, whereby a CD could lie in decile 1 in the unconfidentialised index and decile 2 in the confidentialised index only because it was originally in percentile 19 and moved to percentile 21. The difference of two percentiles more accurately reflects the magnitude of the change in rankings. Table 2.4(b) shows that approximately 97 per cent of CDs moved less than five percentiles, with only 0.6 per cent of CDs moving more than ten percentiles.

The characteristics of the CDs with the largest score differences after confidentialisation were investigated. These CDs tended to be small in population, or have a small number of dwellings. This is intuitive as confidentiality makes similarly 'small' adjustments to counts.

As noted in Section 1, the variables used in SEIFA are proportions, which means that each variable has a count for the denominator, and a corresponding count for the numerator. When calculating proportions, the random error introduced can be ignored except when very small numerators and denominators are involved, in which case the impact on proportions can be significant (ABS, 2006). Aside from these effects of the confidentiality process, possible respondent and processing errors also have a relatively large impact on small numerator and denominator counts (ABS, 2009).

2.4(a) Original IRSAD scores by IRSAD scores based on unconfidentialised Census data



2.4(b) IRSAD CD score percentile changes from using confidentialised Census data to construct an index

IRSAD CD score percentile change due to using confidentialised Census data	*CD Frequency (% of CDs)*
≤ 2 percentiles	85.4
> 2 and ≤ 5 percentiles	11.3
> 5 and ≤ 10 percentiles	2.6
> 10 and ≤ 20 percentiles	0.6
> 20 percentiles	0.1

This indicates that users creating their own index using confidentialised Census data may need to have more stringent requirements for a CD to be included in the analysis. For example, considering the CD exclusion rules for SEIFA 2006, the population requirements may need to be increased to 20 or 30, and minimum dwelling requirements may need to be increased to 10 or 15.[10] An alternative is for users to create their own indexes at a larger level of geography.

[10] For 2006, a CD requires at least 10 usual residents and at least five occupied private dwellings to receive a SEIFA score. See the technical paper (ABS 2008b) for a list of all requirements for a CD to receive a SEIFA score.

The analysis of the impact of using confidentialised Census data to construct the IRSAD indicates that the index is moderately robust around the tails of the distribution of scores. The most disadvantaged and advantaged two deciles are less affected by using confidentialised data. The middle six deciles (3–8) contained a greater degree of error that was introduced from the confidentialisation process. As expected, CDs with smaller populations tended to be affected more than CDs with larger populations. It is reasonable to expect that similar results would be obtained for other user-created indexes based on confidentialised Census data.

3. SOCIO-ECONOMIC DIVERSITY WITHIN LARGER AREAS

This section discusses some of the issues associated with using SEIFA to analyse geographic units larger than CDs (for example, Local Government Areas), and gives some suggestions for the proper use of SEIFA in these circumstances. The reader is encouraged to refer to Wise and Mathews (2011) for further discussion on socio-economic diversity within areas, specifically focusing on the extent to which individual level advantage and disadvantage is heterogeneous within areas.

3.1 Comparing areas in the presence of diversity

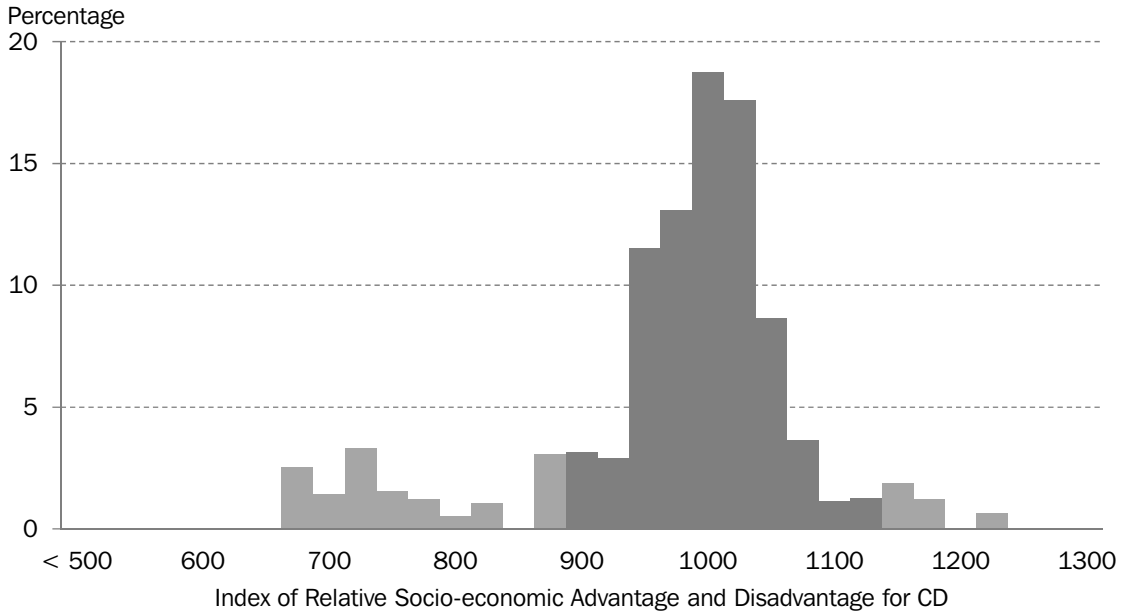
A SEIFA score allows users to compare areas in terms of the average socio-economic advantage and disadvantage of the people and households in each particular area. CDs are the smallest geographic unit available for SEIFA 2006. For larger geographies, such as Local Government Areas (LGAs), a SEIFA score is created from the population weighted average of the CD scores within the larger area. However, a single score for an area does not take into account the socio-economic diversity within that area. The diversity may be important to consider when comparing two larger areas.

This issue is best illustrated with an example. Consider the LGA of Campbelltown, located in on the outskirts of Sydney in NSW. This LGA received a score of 959 (LGA decile 7) on the 2006 IRSAD. Figure 3.1 shows the distribution of CD scores within this LGA.[11] The score in the range 975–1000 indicates 18.7 per cent of *usual residents* in the LGA of Campbelltown live in a CD with an IRSAD score between 975–1000. In this LGA, approximately 14.7 per cent of usual residents live in CDs with a IRSAD score in decile 1 (between 650 and 875), 81.6 per cent live in CDs with a decile 2–9 (between 850 and 1125) and 3.7 per cent live in CDs with a decile 10 (1125 and above).

The chart shows that almost all of the 25 point ranges of CD scores are represented in this LGA between 650 and 1225. The entire range of CD scores for the national IRSAD is approximately 450 to 1325, suggesting that the LGA of Campbelltown is not dissimilar to the rest of the country in terms of the diversity of IRSAD scores. This means that information on distribution provides additional insights on the socio-economic diversity within the LGA.

[11] Obtained from the SEIFA population distributions for Local Government Areas, ABS (2008d).

3.1 Population distribution of CDs within the LGA of Campbelltown



The usefulness of the information on distribution becomes clear when the IRSAD scores at the CD level are compared between the LGA of Campbelltown and that of Kilcoy located to the north west of Brisbane. The LGA of Kilcoy received an IRSAD score of 904, which places it in LGA decile 3, which is four LGA deciles lower than the LGA of Campbelltown. Based on the IRSAD LGA score, one would conclude that Kilcoy is more disadvantaged than Campbelltown.

3.2 Population distribution of CDs within the LGA of Kilcoy

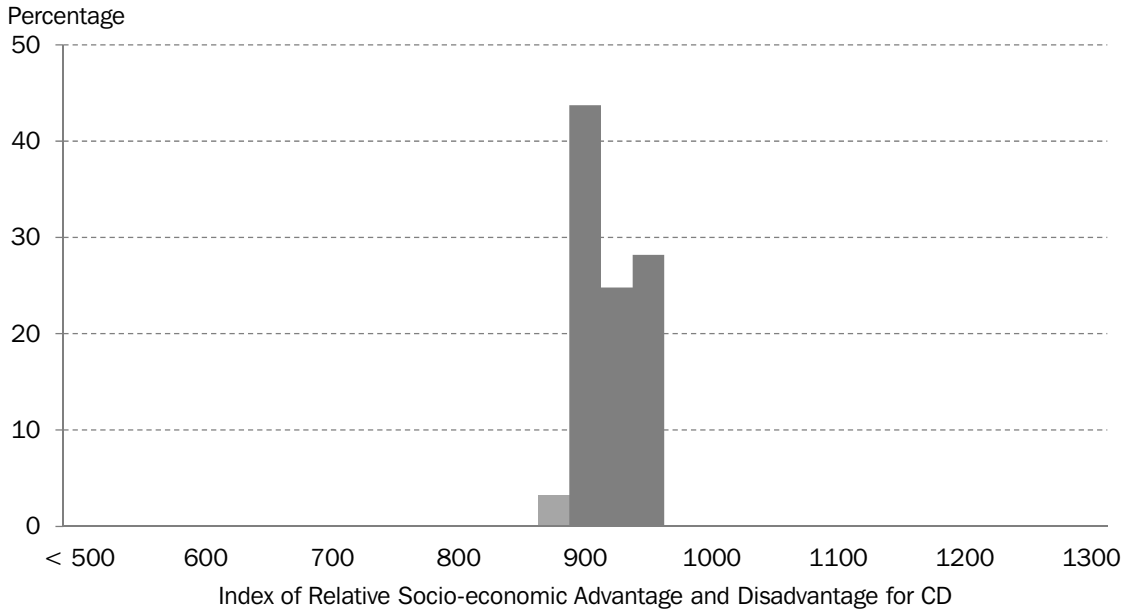


Figure 3.2 shows the distribution of the CD level IRSAD scores for the LGA of Kilcoy. Kilcoy has a lower average score compared to Campbelltown. However, in the LGA of Campbelltown, there is a greater proportion of its population living in more disadvantaged CDs. If the focus of an analysis is in the sub-population of extremely disadvantaged people, Campbelltown is more likely than Kilcoy to be an area of interest. In this circumstance, the single average IRSAD LGA score can be misleading. This raises the risk of the ecological fallacy, which is where errors are made when drawing conclusions about individuals based on the average characteristics of the area in which they live. This issue is addressed more fully in both Wise and Mathews (2011) and Baker and Adhikari (2007).

One alternative measure is the distribution of CDs. For example, the proportion of CDs in each LGA that fall into decile 1 (i.e. the most disadvantaged 10 per cent of CDs) may be a good indicator for the analysis. This measure suggests that Campbelltown has a higher proportion (16.2 per cent) of CDs in the most disadvantaged decile compared with Kilcoy (12.5 per cent). According to the 2006 Census, 15.2 per cent of the resident population in Campbelltown live within these CDs compared with 3.2 per cent in Kilcoy. In other words, Campbelltown has a more disadvantaged subpopulation measured in terms of number of CDs and resident population, although on average it has a higher IRSAD LGA score.

The difference in conclusions occurs because these two measures are summarising different pieces of the CD level information. The LGA level SEIFA score is based on an average so that, within an LGA, a relatively advantaged CD offsets a relatively disadvantaged CD when CD level information is aggregated to form the IRSAD at the LGA level. The proportion of CDs in the most disadvantaged decile provides some of the information that is lost in the aggregation. However, both are valid measures for 'relative disadvantage' and they are useful in answering different questions. Users are encouraged to consider the purpose of their analysis, and to distinguish the statistical properties of different measures when they choose between using the IRSAD at the LGA level and the distribution of CDs within the LGA.

3.2 CD-concentration scores for larger areas

An alternative to the SEIFA score for a larger area, such as LGAs, is to use the distributional information of the CDs within each area. One approach involves choosing a 'cut-off' for the distribution of CD level SEIFA scores, such as the most disadvantaged 10 or 20 per cent of CDs. We then calculate the proportion of CDs within each larger area which falls within this cut-off. We will call this measure the *CD-concentration score* throughout the remainder of this paper. This measure represents the *concentration of CDs within the area that fall within the chosen proportion of the CD level SEIFA distribution*. In the example discussed in Section 3.1, the LGA of Campbelltown has 210 CDs covered by an IRSAD SEIFA score, and 43

of these CDs are in the two most disadvantaged CD deciles. So the IRSAD CD-concentration score for the 20 per cent cut-off is given by 43/210 = 0.21. This proportion can then be compared to other areas.

If we choose the whole of Australia as our area, then the CD-concentration score will be equal to the percentage of the CD distribution that is measured. So if the cut-off is chosen to be the two most disadvantaged CD deciles, then by definition the CD-concentration score for Australia will be 0.2 or 20 per cent. This provides a natural 'benchmark' for this measure in terms of defining what is a 'high' and 'low' concentration of CDs. A CD-concentration score for a particular area that is less than 0.2 indicates an *under-representation* of the CDs in the most disadvantaged 20 per cent of the Australia-wide CD distribution. Similarly, values above 0.2 for specific areas indicate an *over-representation* of the CDs in the most disadvantaged 20 per cent of the Australia-wide CD distribution. For the LGA of Campbelltown, the CD-concentration score of 0.21 is slightly larger than 0.2. This indicates that Campbelltown has a very slight over-representation of CDs in the two most disadvantaged deciles of the Australia-wide CD distribution.

Continuing to use 20 per cent as the cut-off, table 3.3 shows a comparison of the IRSD CD-concentration score to the LGA-level IRSD score. The row indicates the CD-concentration score, and the column indicates the IRSD decile of the LGA. For example, the count of 63 under decile 1 for CD-concentration score of 0.9 to 1.0 indicates there are 63 LGAs which have a LGA decile 1 and have between 90 and 100 per cent of their CDs fall within the most disadvantaged 20 per cent of CD level IRSD scores.

3.3 IRSD LGA decile by CD-concentration score

	CD-concentration score	*IRSD LGA level decile (most disadvantaged)*				→					*(least disadvantaged)*	
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Total*
(most disadv.) → (least disadv.)	0.0 to 0.1	0	0	0	1	2	10	15	28	44	66	166
	0.1 to 0.2	0	0	0	4	14	24	28	30	17	0	117
	0.2 to 0.3	0	2	11	24	28	26	21	9	4	0	125
	0.3 to 0.4	0	8	19	28	17	6	2	0	1	0	81
	0.4 to 0.5	0	11	21	8	4	0	1	0	1	0	46
	0.5 to 0.6	1	18	11	2	1	1	0	0	0	0	34
	0.6 to 0.7	0	14	3	0	0	0	0	0	0	0	17
	0.7 to 0.8	1	4	2	0	0	0	0	0	0	0	7
	0.8 to 0.9	1	5	0	0	0	0	0	0	0	0	6
	0.9 to 1.0	63	5	0	0	0	0	0	0	0	0	68
	Total	66	67	67	67	66	67	67	67	67	66	667

Table 3.3 indicates that the least disadvantaged and most disadvantaged LGA deciles are by far the most homogenous with respect to the CD-concentration score. The other LGA deciles show a less consistent relationship with the CD-concentration score. For example, an LGA with an IRSD score in decile 4 may have a higher or lower CD-concentration score compared to an LGA with an IRSD score in decile 5, 6, or 7. Thus, each measure contains *different information about the socio-economic disadvantage of CDs within the larger area*. Using either measure in isolation will not give a complete picture.

Only using an index score may oversimplify reporting of an area's relative socio-economic disadvantage. The investigations based around the CD-concentration score provide extra information about an LGA *in addition to the index score*. The CD-concentration score is designed to *enhance* the description of socio-economic advantage and disadvantage given by the LGA index score, not to replace the large area score.

In the analysis presented above, the cut-off value chosen for the CD-concentration score was the most disadvantaged 20 per cent of CDs. However, many other choices of cut-off value could be taken depending on the focus of users.

To facilitate the calculation of CD-concentration scores (and other diversity measures), the ABS intends to release decile distribution spreadsheets for SEIFA 2011 for selected larger areas. An example of the type of output, using 2006 Census data, is shown in table 3.4 for the two LGAs discussed in Section 3.1.[12] The row indicates the LGA, and the column indicates the deciles of the CDs which are within that LGA. For example, the count of 26 under decile 3 for Campbelltown indicates that there are 26 CDs within the Campbelltown LGA which sit in decile 3. The 'no score' column shows the number of CDs within the LGA which did not receive a CD level SEIFA score. See ABS (2008a, 2008b) for details on why a CD may not have received a score.

3.4 Distribution of CDs within the LGAs of Campbelltown and Kilcoy – IRSD[13]

	IRSD CD level decile (most disadvantaged)				→					*(least disadvantaged)*	
LGA	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*No score*
Campbelltown	36	19	26	27	37	26	21	7	6	5	2
Kilcoy	0	3	2	2	1	0	0	0	0	0	0

12 Note that although the areas are the same, this section is using the IRSD, and *not* the IRSAD, so the results are slightly different from what was reported in Section 3.1.

13 For SEIFA 2011, SA1 will replace CD as the base level of geography. See Section 4 for further details.

3.3 Using and interpreting the CD concentration score in analysis

In this section, both SEIFA IRSD deciles and the CD-concentration scores will be implemented to illustrate how they can be used to analyse data from the 2004–05 National Health Survey (NHS, cat. no. 4364.0). ABS (2008b) goes through a more general analysis of the relationship between health indicators and SEIFA. The purpose of this section is not to analyse the relationship between SEIFA and health per se, but to show how to interpret the results in similar situations. This section analyses one particular indicator of health – self-reported health status – and its relationship with area level socio-economic disadvantage as measured by the Index of Relative Socio-economic Disadvantage (IRSD) at the Statistical Local Area (SLA) level.

The 2004–05 NHS survey comprised approximately 26,000 individuals across Australia, and survey responses contain a wealth of health information. The data from the survey was matched to the SLA IRSD decile and the SLA IRSD CD-concentration score (using a 20 per cent cut-off) for the area the respondent resided in.

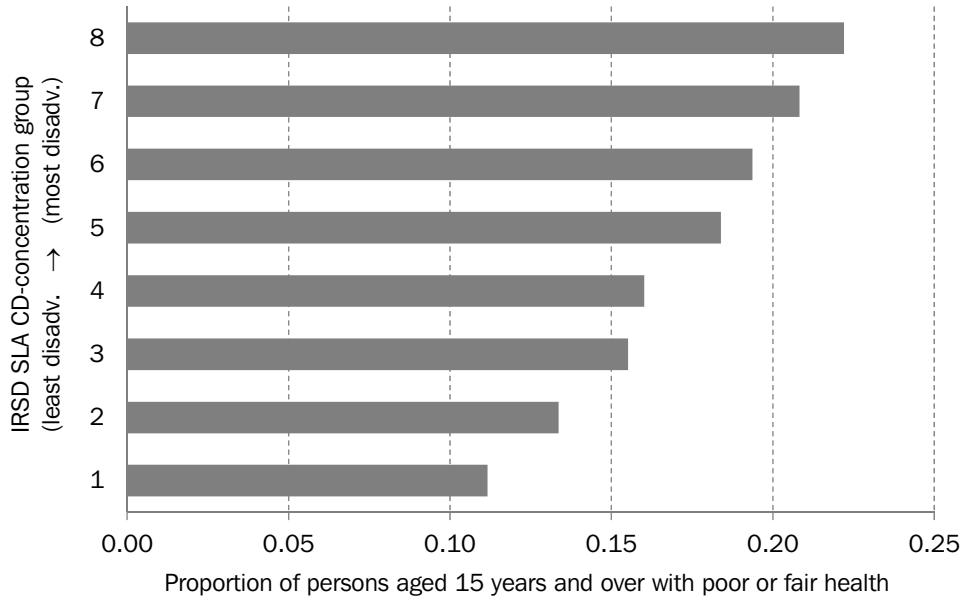
For the purposes of this analysis, the CD concentration scores were further classified into groups. This was done as follows. For the SLA-level there is a large number of SLAs with a CD-concentration score of 0 – approximately 30 per cent of all SLAs. These SLAs were allocated into the CD-concentration group 1. Consequently, the groups are numbered from 1 to 8; group 1 contains approximately 30 per cent of SLAs, and groups 2–8 each contain approximately 10 per cent of SLAs.

Figure 3.5(a) and 3.5(b) present the relationship between SEIFA and the estimated proportion of people who have self-reported health status of 'fair' or 'poor' at the SLA level. Figure 3.5(a) gives the relationship when SLAs are ranked using the CD-concentration score group and figure 3.5(b) shows the relationship when SLAs are ranked using the IRSD SLA index decile.

It is important to note that the interpretation of 'decile 1' and 'group 1' is different for the two analyses. An IRSD index decile equal to 1 indicates that the SLA lying in this decile is relatively disadvantaged as indicated by the IRSD index. Conversely, a CD-concentration group of 1 indicates that the SLA has a relatively low concentration of disadvantaged sub populations; these tend to be the relatively less disadvantaged SLAs. The vertical axes on both of the plots have been adjusted so that it reads from the least disadvantaged SLAs at the bottom of the plot to the most disadvantaged SLAs at the top.

3.5 SLA level: proportions of persons aged 15 years and over who reported 'poor' or 'fair' health by IRSD

(a) SLA level CD-concentration group



(b) SLA level index decile

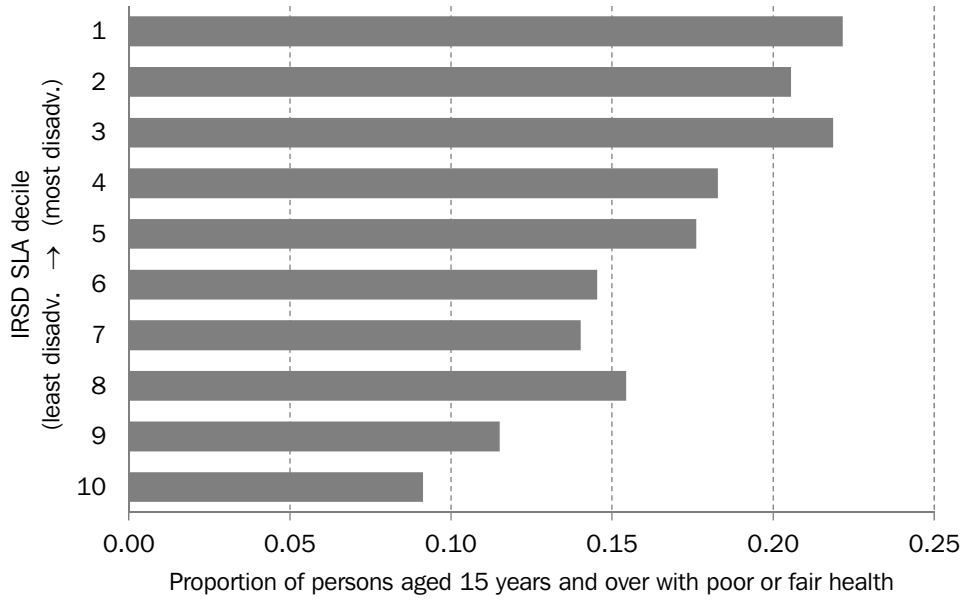


Figure 3.5(a) shows that SLAs with a higher IRSD CD-concentration group tend to have a higher proportion of people who reported their health status as 'fair' or 'poor'. Similarly, figure 3.5(b) shows that SLAs with a lower IRSD index decile tended to have a higher proportion of people who reported their health status as 'fair' or 'poor'. On comparing figure 3.5(a) with figure 3.5(b) the relationship is smoother for the CD-concentration group in comparison to the IRSD SLA index decile. Thus, the relationship between the CD-concentration group and self-reported health status appears more consistent than the corresponding relationship between the SLA index decile and self-reported health status.[14] In this case, the CD concentration score is a more relevant indicator measure to use than the raw SEIFA score. This may be due to the ability of the CD-concentration score to pick up SLAs which have high numbers of both advantaged and disadvantaged CDs, which would tend to offset each other in the index score.

3.4 Diversity within CDs

The above discussion focused on the diversity of the CD level SEIFA scores within larger areas, such as LGAs and SLAs. The main point was to show the loss of important information when one only considers the SEIFA score for the larger area in isolation. However, the CD scores themselves have limitations. For example, they are a summary of individual, family, and household level characteristics. Therefore, there may potentially be socio-economic diversity present within CDs which the CD level SEIFA scores do not take into account. This issue is currently an area of research within the ABS. For a more thorough discussion of this issue, see Wise and Mathews (2011).

[14] For example, a simple linear regression shows that the CD-concentration decile explains 98.7 per cent of the variation in the proportion of people who report 'poor' or 'fair' health, whereas the IRSD index decile explains 92.1 per cent of the variation.

4. PLANS FOR SEIFA 2011 AND THE NEW GEOGRAPHY STANDARD

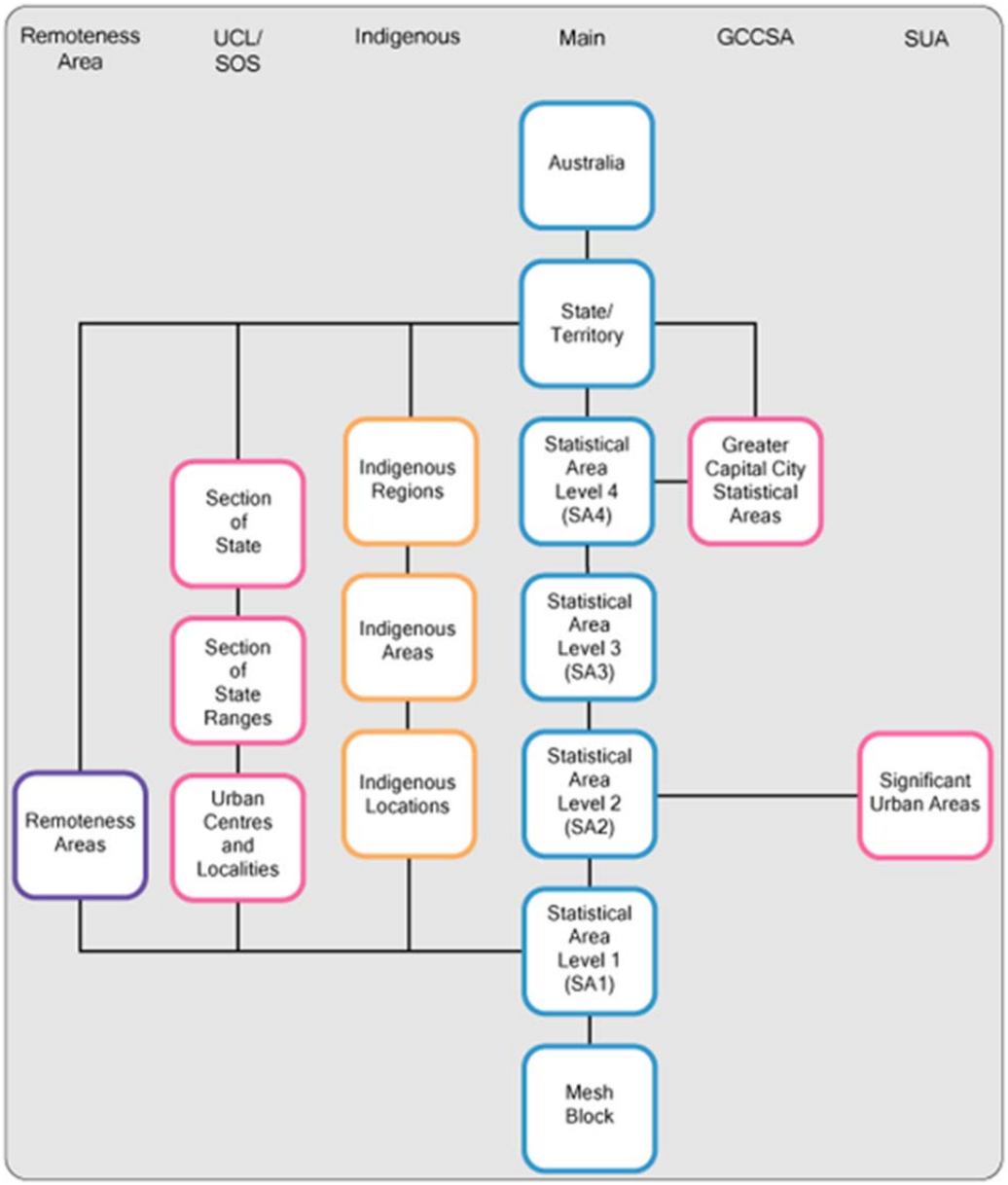
In 2006, SEIFA scores and rankings were created at the Census Collection District (CD) level, and the CD was the primary unit of analysis. Scores and rankings were also created for some larger geographic areas by aggregating the compositional CD scores within these areas. In addition to releasing score and rank information, population distribution spreadsheets were released for larger areas to allow users to look at the distribution of CD level SEIFA scores within these areas.

In 2011, the ABS is planning to release the same set of four indexes that were released in 2006. The same methodology will be used to summarise the Census variables. However, the ABS will replace the existing Australian Standard Geographical Classification (ASGC) with a new geographic system. From July 1st 2011 onwards, the new Australian Statistical Geography Standard (ASGS) will be implemented. Figure 4.1 shows a basic picture of what the structure looks like.

The ABS regions are defined as follows:

- Mesh Blocks (MB) 347,627 across Australia, only limited Census data, total population and dwelling counts will be released at the mesh block level.
- Statistical Areas Level 1 (SA1). Built from whole MBs, 54,805 in total. Will be the smallest region for which most Census data is released. Populations are in the range 200 – 800.
- Statistical Areas Level 1 (SA2). Built from whole SA1s, 2,214 in total. Population in the range 3,000 – 25,000.
- Statistical Areas Level 3 (SA3). Built from whole SA2s, 351 in total. Population in the range 30,000 – 130,000.
- Statistical Areas Level 4 (SA4). Built from whole SA3s, 106 in total. Population in the range 100,000 – 500,000.

4.1 Structure of the new geography



Non-standard ABS geographies will be built using a 'best-fit' allocation of either SA1s or MBs, with the exception of Census Collection Districts (CDs), which will no longer be produced. ABS (2007; 2010a; 2010b) provides general information on the ASGC and the ASGS, and ABS (2011a; 2011b) provides general information regarding the 2011 Census. The remainder of this section will focus on the implications of the change from the ASGC to the ASGS for SEIFA, and is split into the following parts:

- Geographic output of SEIFA information for 2011.
- Issues associated with a Mesh Block level SEIFA.
- Experimental analysis comparing CD level and SA1 level SEIFA.

4.1 Geographic output of SEIFA information for 2011

This section outlines ABS intentions for releasing SEIFA at various geographies. As mentioned earlier, CDs will not be continued under the ASGS, and consequently the primary geographic unit of analysis for SEIFA will need to change. The new primary unit of analysis for SEIFA in 2011 will be the SA1 geography.

In addition to creating the SA1 level scores, the ABS will create SEIFA scores for other larger geographies by aggregating the compositional SA1 scores within these geographies. For larger geographies not in the ASGS, a best fit correspondence of SA1s to the larger geography will be used to aggregate the SA1 scores. A very similar method was used in SEIFA 2006 to allocate CDs to Postal Areas (POAs) and State Suburbs (SSCs).

The ABS will also release a similar product to the distribution spreadsheets published for SEIFA 2006 (cat. no. 2033.0.55.001) that highlight the diversity of socio-economic advantage and disadvantage of the SA1s within a particular area. Section 3.3 provides an example of how these distributions can be used in an analysis.

Furthermore, the ABS is intending to release a new product which gives the decile distribution of SA1s within larger geographies, similar to that shown in table 3.4, and in table 4.9. It presents the information contained in the SA1 level SEIFA indexes so that it is easier for users to assess the distribution of SA1s within larger areas. These spreadsheets will facilitate the easy construction of the SA1-concentration scores described in Section 3 (referred to as CD-concentration scores in Section 3).

Table 4.2 shows the intended output that SEIFA 2011 information will be released at for each geographic unit.

4.2 Geographic output summary for SEIFA 2011

Geographic unit	*Index score*	*SA1[15] distribution spreadsheet*	*SA1[16] decile distributions*
Statistical Area level 1 (SA1)	Yes	N/A	N/A
Statistical Area level 2 (SA2)	Yes	No	Yes
Statistical Area level 3 (SA3)	No	Yes	Yes
Statistical Area level 4 (SA4)	No	Yes	Yes
Statistical Local Area (SLA)	Yes	No	Yes
Local Government Area (LGA)	Yes	Yes	Yes
State Suburb (SSC)	Yes	No	Yes
Postal Area (POA)	Yes	No	Yes
Commonwealth Electoral Division (CED)	No	Yes	Yes
State Electoral Division (SED)	No	Yes	Yes

15 See figures 3.1 and 3.2 for examples.

16 See table 3.4 for example.

4.2 Issues associated with a Mesh Block level SEIFA index

SEIFA indexes will be released at the SA1 level, which conforms to the Census Output dissemination strategy (ABS 2011b). The Census information used to create SEIFA will not be output at the MB level; hence SEIFA will not be output at the MB level for 2011.

A further consideration when creating a Mesh Block level SEIFA is whether the quality of the Census data can support such a small unit of analysis. Some areas are not given a SEIFA score for two broad reasons:

- Low usual resident population or occupied private dwellings; and
- High non-response rate for important Census questions such as: income, occupation, labour force status, and education status.

If the unit of analysis is made too fine, we may need to exclude a large part of the population. For example, consider a small SA1 with a usual resident population of 15, which comprises two Mesh Blocks of usual resident populations 7 and 8. Based on the SEIFA 2006 exclusion rules, an area is excluded if it has a usual resident population of 10 or less. In this example, the SA1 would not get excluded based on this rule, but the two MBs which comprise it would get excluded. Thus, the Mesh Block level index would exclude the two MBs, whereas the SA1 level index would not. This indicates that a price is paid for going to a finer level of geography; a greater proportion of the total usual resident population is excluded from the analysis.

A brief investigation was performed to see how many MBs and SA1s would not have received a SEIFA score in 2006, using the same criterion that was used for 2006 CDs (ABS, 2008b, Section 4.2.3). Table 4.3 shows the 2006 usual resident population living in areas with and without a SEIFA score using the 2006 criteria. It considers three geographic units separately; CDs, SA1s, and experimental 2006 MBs. The second, third, and fourth columns show, in thousands, the number of usual residents residing in MBs, SA1s, and CDs respectively, that would not receive a SEIFA score using the 2006 Census.

This shows that MB level data would exclude areas containing almost four times as much of the population compared to SA1 and CD levels. This table also shows that the proportion of the population within areas without a score is slightly lower for the SA1 level compared to the CD level. This is interesting, since SA1s are typically smaller than CDs, but are designed differently to better capture the population.

Table 4.3 suggests that even if producing a MB level SEIFA index was a part of the Census output strategy in 2011, there will be a trade off in terms of the proportion of the population without a score. Using SA1s would give a SEIFA score to a larger proportion of the usual resident population, compared to both CDs and MBs. The next section contains a more in depth comparison of the SA1 unit with the CD unit, including a comparison of the areas which would not receive a SEIFA score.

4.3 2006 usual resident population without a SEIFA score using 2006 criterion (a) (b)

	Geographic unit		
	Mesh Block (000s)	*Statistical Area 1 (000s)*	*Census Collection District (000s)*
2006 usual resident population in areas without a SEIFA score	401	112	116
2006 usual resident population in areas with a SEIFA score	19,454	19,743	19,739
Total 2006 usual resident population	19,855	19,855	19,855

(a) Figures for SA1s were derived from analysis that is described in Section 4.3.
(b) Figures for MBs were derived using the 2006 experimental MB data.

4.3 Experimental analysis comparing Census Collection District level and Statistical Area level 1 SEIFA

The ABS has performed some initial investigations into the differences between a SEIFA index calculated at the CD level and at the SA1 level. To some extent this can be seen as an analysis of the so-called 'modifiable area unit problem' (Openshaw, 1984). For this analysis, the 2006 data was used to calculate SEIFA under two different geographies, the 2006 CD, and an approximation to the 2009 SA1 geography using 2006 experimental Mesh Blocks. This analysis is of an approximate nature, and is only intended to give users an indication of the possible changes to expect in SEIFA due to the changes in geographic units. There will be additional changes for the SEIFA 2011 indexes which cannot be assessed in this analysis due to the differences in the population, as measured by the Census, between 2006 and 2011.

The approximation to the SA1 used here is based on allocating a 2006 Mesh Block to a 2009 SA1 if the centroid of that Mesh Block lies within the SA1 boundary.[17] This method will give accurate SA1s if the MB boundaries do not change. Most of the MB boundaries did not change between 2006 and 2009, indicating that most of the corresponding SA1 boundaries would also not have changed. We do not have any quality measures on the accuracy of how well *each individual* 2009 SA1 is approximated by the allocation of 2006 MBs. However, this is an experimental analysis intended to give users a general indication for how the new geography will impact the SEIFA indexes. Hence, detailed quality measures, although useful, are not necessary for this analysis. The CD level SEIFA results shown in this section are the same as those in ABS (2008b).

[17] This method is also known as the 'point in polygon' method.

We now investigate the SA1s which would be excluded from SEIFA using the 2006 criteria. Table 4.4 shows how many CDs were excluded from SEIFA 2006 and how many SA1s would have been excluded, due to each exclusion rule. The first two columns show the number of CDs and SA1s falling under each exclusion category respectively. The total number of CDs and SA1s excluded (1,256 and 1,834) is not equal to the sum of the entries in the first two columns, because each area can satisfy multiple criteria. The third and fourth columns show, respectively, the number of CDs and SA1s excluded by each category that have not been excluded by one of the above categories. The sum of the third and fourth column entries equals the total number of CDs and SA1s excluded.

4.4 Number of CDs and SA1s excluded by each exclusion rule, 2006

	Number of units excluded by each rule		*Number of units excluded by each rule (hierarchical)*	
Exclusion rule	*CD*	*SA1*	*CD*	*SA1*
Population = 0	616	952	616	952
Offshore, shipping, no usual address(a)	101	9(a)	47	9(a)
Population > 0 and ≤ 10	188	548	163	548
Employed persons ≤ 5	894	1,611	102	135
Classifiable occupied private dwellings ≤ 5	1,004	1,704	127	134
Household equivalised Income not stated ≥ 70%	22	43	1	4
Occupation not stated ≥ 70%	3	2	1	0
Level of education (non-school qualification) not stated ≥ 70%	284	98	182	28
Labour force status not stated ≥ 70%	23	54	0	2
Type of educational institution attending not stated ≥ 70%	44	98	0	0
People in non-private dwellings ≥ 80%	147	164	17	20
Total number of areas excluded			1,256	1,834

(a) The offshore and shipping SA1s were excluded from the original geographic allocation of mesh blocks, and are thus unavailable for this analysis. The count of 9 indicates the 9 no usual address SA1s.

Table 4.3 shows that there are fewer areas which get excluded due to non-response – 34 SA1s (4+0+28+2+0) compared to 184 CDs (1+1+182+0+0). This is mostly due to the level of education not stated criterion. In both cases, the vast majority of areas are excluded because of a small number of occupied private dwellings.

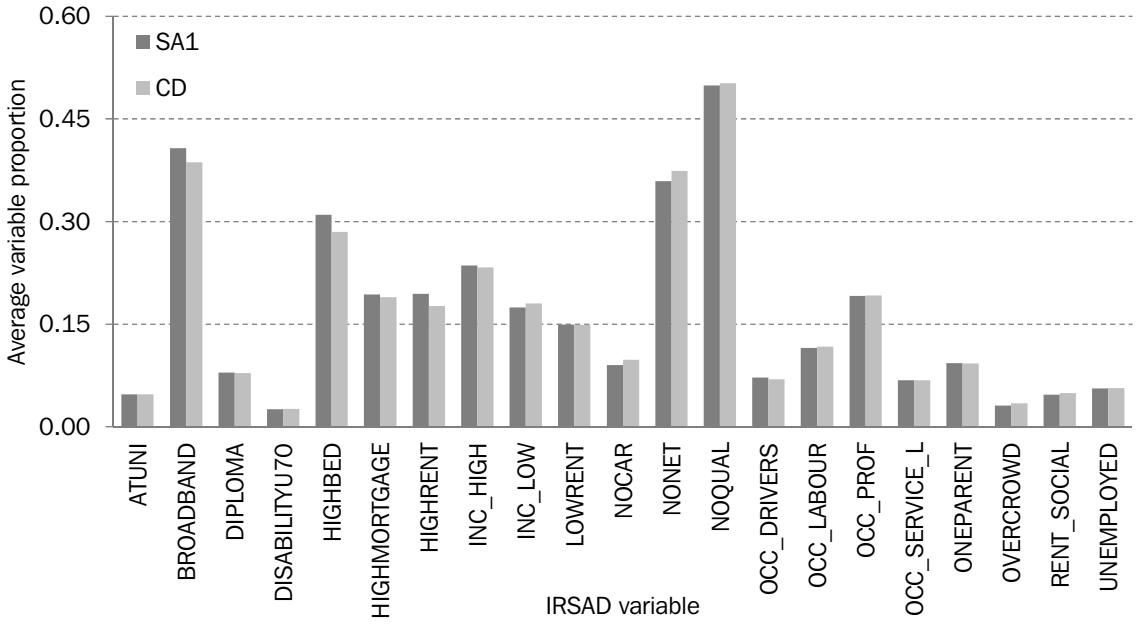
The remaining analysis in this section is based on all CDs and SA1s which were not excluded by the above criteria. Only the IRSAD will be considered in the comparison.

One of the simplest ways to compare the CDs to the SA1s is to compare the mean variable proportions of the IRSAD variables across Australia.[18] Figure 4.5 shows the

[18] Each CD/SA1 contains a different number of people and dwellings. Thus, the average proportions across CDs/SA1s do not necessarily correspond to the proportion across Australia.

average proportion for each variable in the IRSAD across Australia. The variables have been listed according to their mnemonics. They are reasonably self-explanatory, and are listed with the full variable definitions in Appendix A. Figure 4.5 shows that the average variable proportions are quite similar across Australia. The variables *percentage of households with broadband internet connection* (BROADBAND); *percentage of households with 4 or more bedrooms* (HIGHBED); and *percentage of rented households paying rent over $290 rent per week* (HIGHRENT) show a slightly higher average proportion using SA1s compared to using CDs. The variables *percentage of households without a car* (NOCAR) and *percentage of households with no internet connection* (NONET) show a slightly lower average proportion using SA1s compared to using CDs.

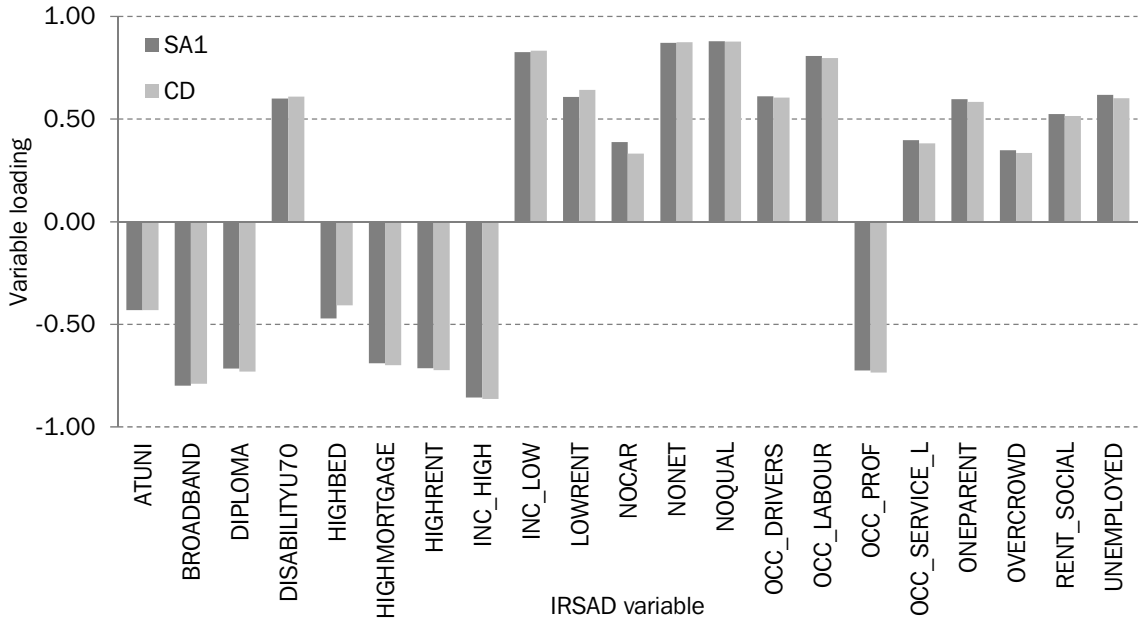
4.5 Comparison of CD level and SA1 level average proportions (a)



(a) The variable mnemonic has been used in this graph.
For a description of what each mnemonic refers, see Appendix A.

Figure 4.6 shows the variable loadings for each variable in IRSAD using SA1s compared to using CDs. The variable loading represents the relative importance of each variable in determining the index score. A variable with a higher loading will have a greater influence on the index score compared to a variable with a lower loading. This plot shows little difference between the CD and SA1 unit of analysis, with the variables HIGHBED and NOCAR having higher loadings, while *percentage of households paying rent who pay less than $120 per week, excluding $0 per week* (LOWRENT) received a smaller loading.

4.6 Comparison of CD level and SA1 level variable loadings (a)



(a) The variable mnemonic has been used in this graph.
For a description of what each mnemonic refers, see Appendix A.

The analysis presented so far has been broad based. Using this analysis alone, one can be reasonably confident that the changes due to the new geographic unit will not induce a major change on the SEIFA indexes in terms of the correlations underpinning the indexes.

In terms of how an SA1 level index will report the distribution of socio-economic advantage and disadvantage across Australia, there may be some changes to *particular areas*. For example, a particular 2006 CD might have quite a diverse range of socio-economic advantage and disadvantage. However, under the new geography, it is possible for the CD to be split into multiple SA1s; which 'separated out' this diversity, creating more homogenous areas. Section 3 discusses in greater detail the issue of diversity of socio-economic characteristics within larger areas.

CDs and SA1s are not directly comparable geographic units. Our approach is to analyse the distribution of SA1s and CDs within Local Government Areas (LGAs). However, to avoid drawing erroneous conclusions due to the error in the approximate SA1s and the LGA boundaries, only LGAs with 20 CDs or more will be analysed. There are 352 out of 667 LGAs which meet this criterion. To assess the distributions, the CD- and SA1-concentration scores for each LGA will be compared using various cut-offs. See Section 3.2 for a discussion of the concentration scores.

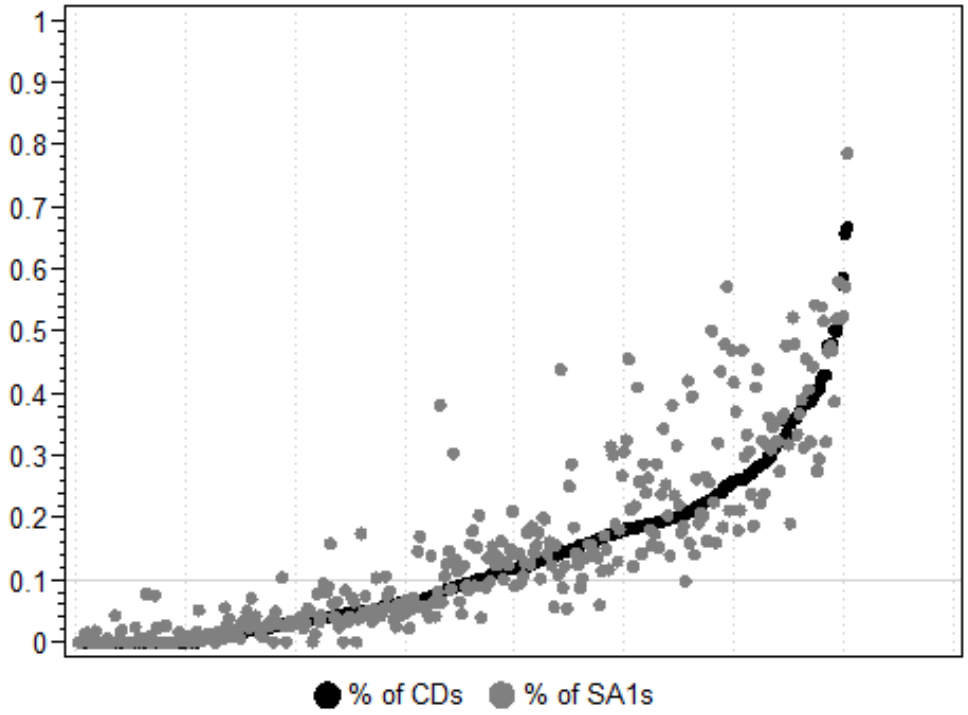
Figure 4.7 shows the CD and SA1 decile 1 concentration scores within each LGA. The CD-concentration scores are represented with a black circle, and the SA1-concentration scores have been represented with a grey circle. For ease of interpretation, the LGAs were sorted by their CD decile 1 concentration score, so the graph reads from lowest concentration of decile 1 CDs on the left to the highest concentration of decile 1 CDs to the right. 10 per cent is the relevant 'benchmark' percentage: points above this percentage have an over-representation of the most disadvantaged CDs (or SA1s), and points below this percentage have an under-representation of the most disadvantaged CDs (or SA1s). Figure 4.8 shows a similar plot for the concentration of CD and SA1s in the most *advantaged* decile of CDs and SA1s. A reference line at 10 per cent has been drawn on both plots to aid interpretation.

Figures 4.7 and 4.8 both show that the distribution of the most advantaged and most disadvantaged CDs stays roughly the same when compared to the corresponding distribution of SA1s. In fact, the distribution of the most advantaged 10 per cent of SA1s is nearly identical to the corresponding distribution of the most advantaged 10 per cent of CDs. The distribution of the most disadvantaged 10 per cent of SA1s is less close to the corresponding distribution of the most disadvantaged 10 per cent of CDs, but there appears to be no systematic affects. Some LGAs have a SA1 concentration score higher than the corresponding CD concentration score, and some have a lower CD concentration score.

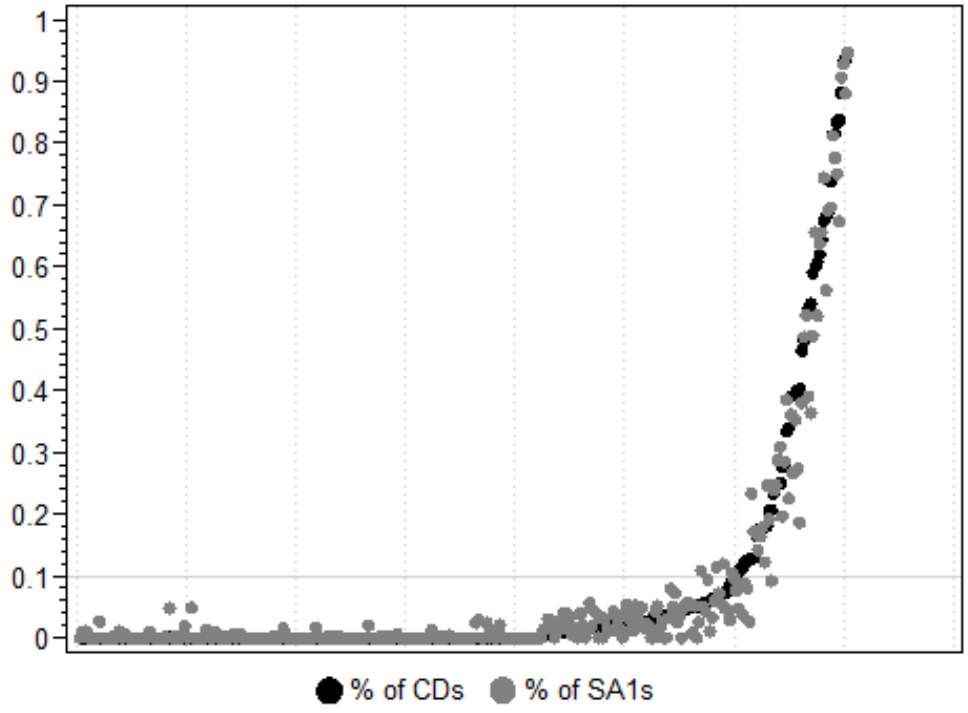
There are a few reasonably large changes shown in figure 4.7 which warranted some additional analysis. These are small in proportion, but have a tendency to distort the plot, making it appear as if there are a lot of LGAs which change by a large amount.

The LGA of Buloke, located in the North West of Victoria, had the largest difference between the distribution of the most disadvantaged CDs and SA1s. The changes to this particular LGA will be investigated in more detail. It is important to note that in the LGA of Buloke there are 26 CDs and 21 SA1s; this helps contextualise the differences presented in this investigation. 7.7 per cent of the CDs within Buloke were located in the most disadvantaged CD decile, whereas 38.1 per cent of the SA1s were located in the most disadvantaged SA1 decile. Table 4.9 presents the distribution of CDs and SA1s in each decile within the LGA of Buloke, showing that there are eight SA1s in the most disadvantaged SA1 decile, compared to two CDs in the most disadvantaged CD decile. Additionally, there are twelve CDs in CD deciles 4–7 whereas there are no SA1s in SA1 deciles 4–7.

4.7 SA1- and CD-concentration scores for LGAs: percentage of CDs (black) and SA1s (grey) within each LGA that are in CD decile 1 (black) and SA1 decile 1 (grey)



4.8 SA1- and CD-concentration scores for LGAs: percentage of CDs (black) and SA1s (grey) within each LGA that are in CD decile 10 (black) and SA1 decile 10 (grey)

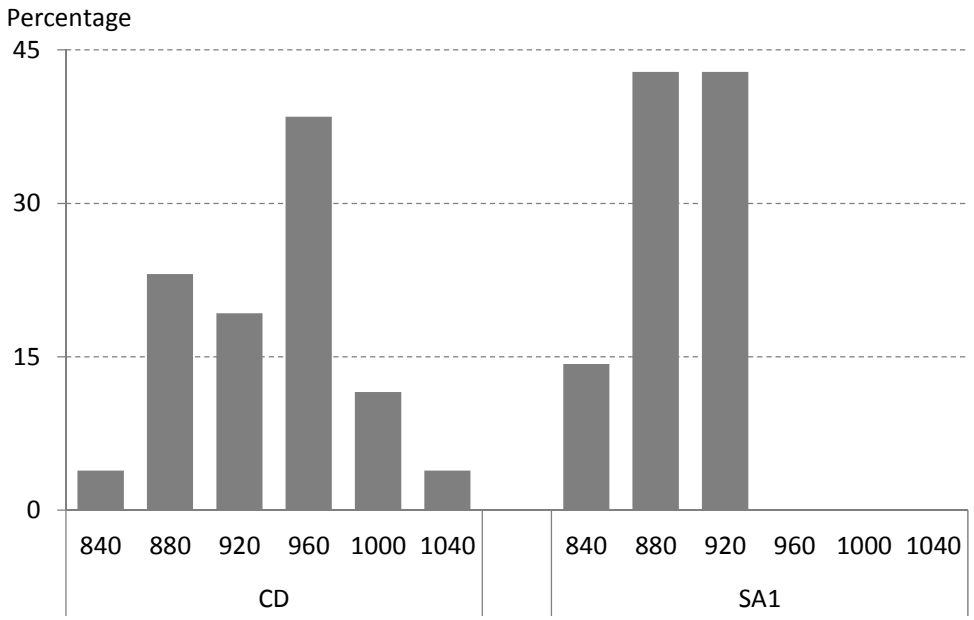


4.9 Distribution of CDs and SA1s within the LGA of Buloke – IRSAD

Geographic unit	IRSAD decile										No score
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	
No. of CDs	2	8	4	7	2	2	1	0	0	0	0
No. of SA1s	8	9	4	0	0	0	0	0	0	0	0

Figure 4.10 shows the distribution of CD and SA1 IRSAD index scores for the LGA of Buloke. This shows that there are no SA1 scores above 940, but there are quite a large number of CD scores above 940. The peak of the distribution has moved back from around 960 to around 900, indicating that the distribution of SA1s has shifted to lower scores. Additionally, the CD distribution has a longish tail and appears more diverse, whereas the SA1 distribution has no tail and appears less diverse.

4.10 Distribution of CD and SA1 IRSAD scores within the LGA of Buloke



Finally, figures 4.11(a) and 4.11(b) show maps of Buloke and the surrounding area based on CDs and SA1s. The same colouring scheme was used in each map: it runs from black to colour decile 1 areas (relatively most disadvantaged areas), and gradually gets lighter up to a very light grey for the areas in decile 10 (relatively most advantaged areas). These maps show an apparently large difference between the two geographic units, at least for the larger, more rural SA1s of this area.

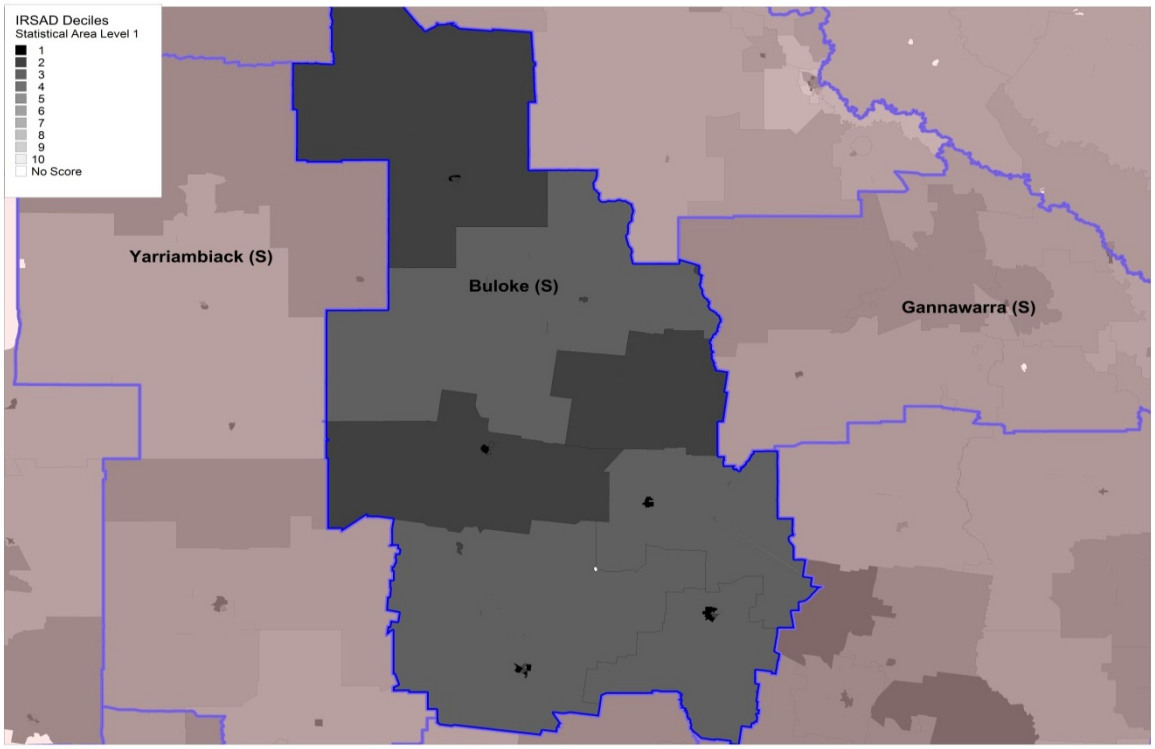
Deeper investigations into this change revealed that CDs in deciles 3–7 were relatively small in population – about 130 – compared to the CDs in deciles 2 – about 540.[19] The SA1 geographic units have more uniform populations of about 340, and no SA1 population is below 200 in Buloke. The relatively less disadvantaged, smaller CDs are being split and combined with relatively more disadvantaged, larger CDs. This can be seen from the significantly different boundaries for the rural areas in figure 4.11(a) and 4.11(b).

The point to note is that in some instances, the change from CDs to SA1s will alter the reporting of the socio-economic characteristics of an LGA. However, it should be noted that Buloke is the most extreme case identified in this analysis. Taking a broader perspective, consider figures 4.12(a) and 4.12(b) that show the city of Melbourne using SA1s and using CDs respectively. The CDs and SA1s are giving very similar pictures of Melbourne. In cases where there are differences (not just for Melbourne, but all Australia), the general expectation is that SA1s will better define areas for statistical reporting purposes, since they more clearly define urban and rural areas, small towns, and discrete Indigenous communities.

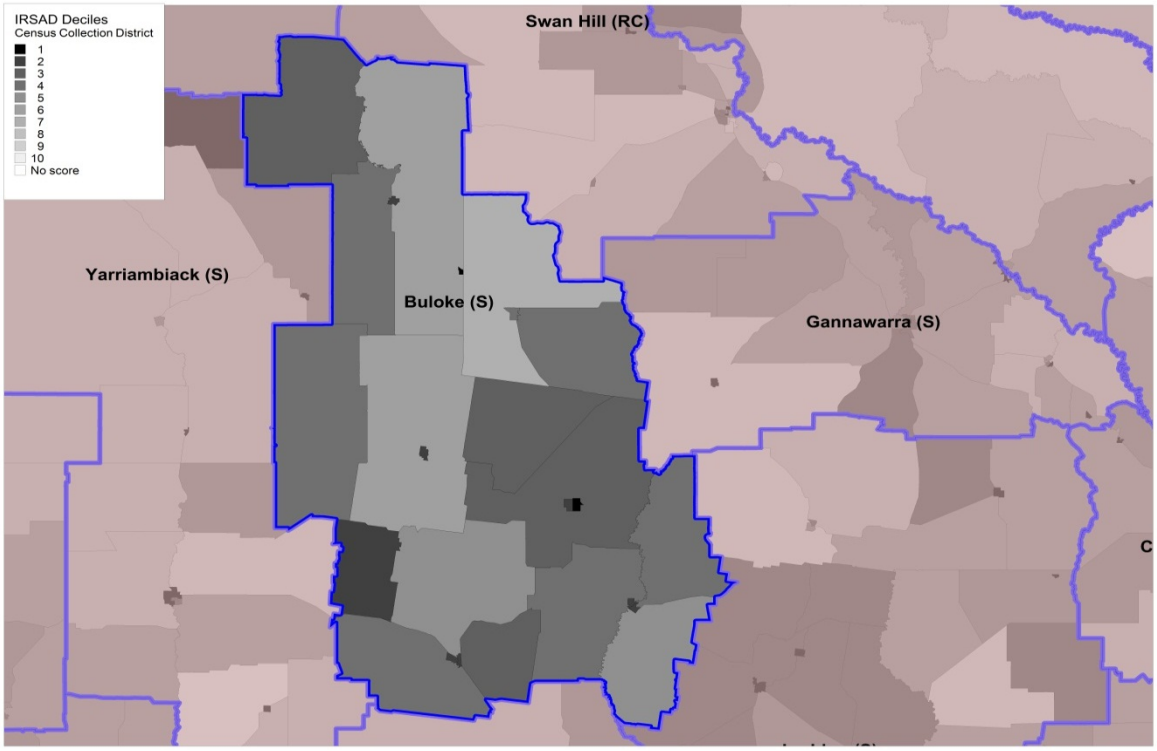
Concluding this section of the paper, the comparative analysis between the CD and approximate-SA1 indexes has yielded three main findings: the change to SA1s will not have a big impact either way on the amount of population not receiving a score due to exclusion rules; the correlations between variables underpinning the index are not too different when using SA1s; and there will be some changes to the way socio-economic characteristics are reported across larger areas, with the expectation that it will provide more useful information.

[19] Note that there is still a large change; if SA1s and CDs are weighted by population we get, instead of 7.7 and 38.1 per cent, a change from 8.0 to 37.7 per cent.

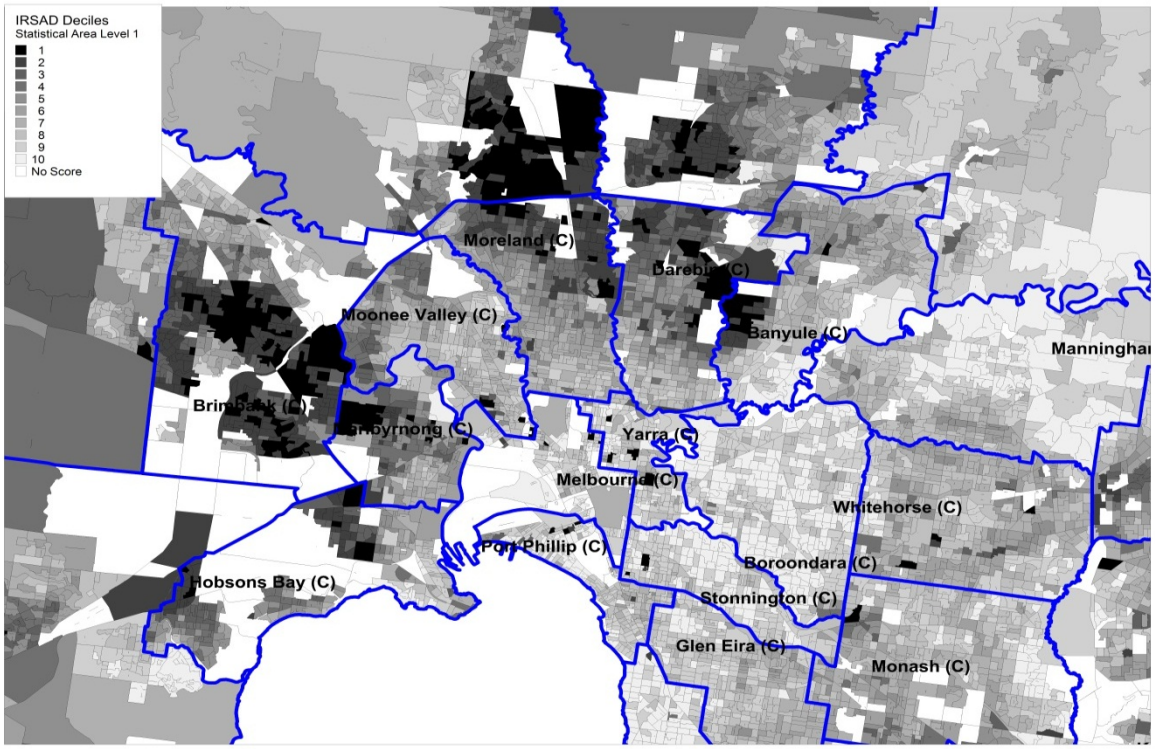
4.11(a) Maps of SA1 IRSAD scores for the LGA of Buloke and surrounding area



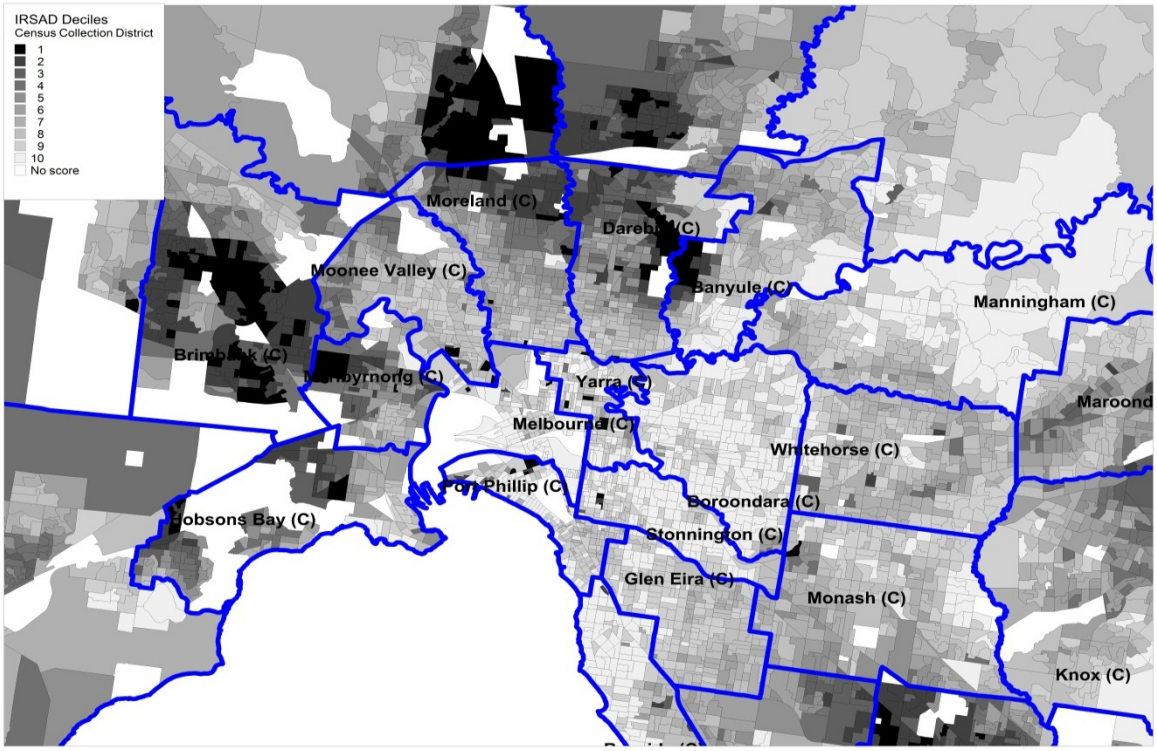
4.11(b) Maps of CD IRSAD scores for the LGA of Buloke and surrounding area



4.12(a) Maps of SA1 IRSAD scores for Melbourne



4.12(b) Maps of CD IRSAD scores for Melbourne



5. CONCLUDING REMARKS

This paper has presented an overview of recent research activities and future directions for SEIFA at the ABS. Three important areas have been covered: exploring the robustness of the SEIFA indexes to the influence of specific variables and areas; investigating ways of representing the diversity of socio-economic characteristics within larger areas; and estimating the impact of the new geography standard on the indexes.

The SEIFA indexes were found to be generally robust against removing an area or a variable from the indexes, when assessing the effects on the ranking of CDs. However, some particular areas were more affected than others, although the most disadvantaged and least disadvantaged areas tended to be more resistant to ranking changes. Similarly, when considering the impact of confidentialisation of Census data on a user-created index, it was found that the tails of the distribution of CD IRSAD scores were moderately robust, whilst the middle six deciles (3 – 8) contained a greater degree of difference. As expected, CDs with smaller populations tended to be affected more than CDs with larger populations.

The diversity of the CD level SEIFA scores within larger areas, such as LGAs and SLAs, was explored using a measure named the CD-concentration score. For our analysis, this measure represented the concentration of CDs within the area that fall within the bottom 20 per cent (our chosen proportion) of the CD level SEIFA distribution. Using this CD-concentration score showed that there can be a loss of important information when one only considers the SEIFA score for the larger area in isolation.

The final section of the paper considered a comparative analysis between the 2006 CD IRSAD and approximate-SA1 IRSAD indexes, which yielded three main findings: the use of SA1s as the base unit of SEIFA will not have a large impact on the excluded population; the correlations between variables underpinning the IRSAD index are not too different when using SA1s; and there will be some changes to the way socio-economic characteristics are reported across larger areas, with the expectation that it will be an improvement on the old standard.

The ABS is planning to release SEIFA 2011 on 28 March 2013. SA1 will be the base unit of geography. SEIFA will not be released at the MB level for SEIFA 2011; however the ABS recognises the demand for finer level information and will continue to assess what is feasible when looking towards the future – see for example Wise and Matthews (2011). The ABS will also continue to release distribution information for larger geographic units to enrich the analysis and understanding of socio-economic advantage and disadvantage within these areas.

REFERENCES

Adhikari, P. (2006) "Socio-Economic Indexes for Areas: Introduction, Use and Future Directions", *Methodology Research Papers*, cat. no. 1351.0.55.015, Australian Bureau of Statistics, Canberra.

Australian Bureau of Statistics (2007) *The Review of the Australian Standard Geographical Classification*, cat. no. 1216.0.55.001, ABS, Canberra.

—— (2006) *Census Dictionary, 2006*, cat. no. 2901.0, ABS, Canberra.

—— (2008a) *Information Paper: An Introduction to Socio-Economic Indexes for Areas (SEIFA), 2006*, cat. no. 2039.0, ABS, Canberra.

—— (2008b) *Socio-Economic Indexes for Areas (SEIFA) – Technical Paper, 2006*, cat. no. 2039.0.55.001, ABS, Canberra.

—— (2008c) *Information Paper: Outcome from the Review of the Australian Standard Geographical Classification*, cat. no. 1216.0.55.002, ABS, Canberra.

—— (2008d) *Census of Population and Housing: Socio-Economic Indexes for Areas, Australia – Data only*, cat. no. 2033.0.55.001, ABS, Canberra.

—— (2009) *TableBuilder 2006 – User Manual*, cat. no. 2065.0, ABS, Canberra.

—— (2010a) *Australian Standard Geographical Classification (ASGC), July 2010*, cat. no. 1216.0, ABS, Canberra.

—— (2010b) *Australian Statistical Geography Standard: Design of the Statistical Areas Level 4, Capital Cities and Statistical Areas Level 3, May 2010*, cat. no. 1216.0.55.003, ABS, Canberra.

—— (2011a) *Census of Population and Housing: Nature and Content, 2011*, cat. no. 2008.0, ABS, Canberra.

—— (2011b) *Discussion Paper: Census of Population and Housing – ABS Views on 2011 Census Output Geography, 2011*, cat. no. 2911.0.55.003, ABS, Canberra.

Baker, J. and Adhikari, P. (2007) "Socio-Economic Indexes for Individuals and Families", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.086, Australian Bureau of Statistics, Canberra.

Brooks, S.P. (1994) "Diagnostic for Principal Components: Influence Functions as Diagnostic Tools", *The Statistician*, 43(4), pp. 483–494.

Critchley, F. (1985) "Influence in Principal Components Analysis", *Biometrika*, 72(3), pp. 627–636.

Croux, C. and Haesbroeck, G. (2000) "Principal Components Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies", *Biometrika*, 87(3), pp. 603–618.

Jolliffe, I.T. (1986) *Principal Components Analysis*, Springer Series in Statistics.

Openshaw, S. (1984) "The Modifiable Areal Unit Problem", *Concepts and Techniques in Modern Geography*, 38, GeoBooks, Norwich, England.

Shi, L. (1997) "Local Influence in Principal Components Analysis", *Biometrika*, 84(1), pp. 175–186.

Wise, P. and Matthews, R. (2011) "Socio-Economic Indexes for Areas: Getting a Handle on Individual Diversity within Areas", *Methodology Research Papers*, cat. no. 1351.0.55.036, Australian Bureau of Statistics, Canberra.

ACKNOWLEDGEMENTS

The authors would like to thank Marcus Blake, Jeffrey Wright, Phillip Gould and Jenny Myers for their helpful input with this paper. We would also like to thank Peter Rossiter for providing comments and feedback. The content and presentation of the paper are much improved as a result of their input. Responsibility for any errors or omissions remains solely with the authors.

APPENDIXES

A. LIST OF SEIFA 2006 VARIABLES AND WEIGHTS

Index of Relative Socio-economic Disadvantage (IRSD)

The IRSD is a general index which measures the relative disadvantage of an area. It is based only on variables which are related to disadvantaging aspects of an area. A relatively low score or rank on this index indicates the area is relatively more disadvantaged, and a relatively high score or rank indicates that the area is relatively less disadvantaged. This is the only index which contains no indicators of relative advantage. Table A.1 shows a list of the variables used in IRSD and their weights.

A.1 List of variables used for the IRSD and their weights

Mnemonic	*Variable*	*Weight*
NONET	% Occupied private dwellings with no Internet connection	–0.33
INC_LOW	% People with stated annual household equivalised income between $13,000 and $20,799 (approx. 2nd and 3rd deciles)	–0.30
NOQUAL	% People aged 15 years and over with no post-school qualifications	–0.30
OCC_LABOUR	% Employed people classified as Labourers	–0.30
UNEMPLOYED	% People (in the labour force) unemployed	–0.27
RENT_SOCIAL	% Households renting from a Government or Community organisation	–0.27
LOWRENT	% Households paying rent who pay less than $120 per week (excluding $0 per week)	–0.26
ONEPARENT	% Families that are one parent families with dependent offspring only	–0.26
DISABILITYU70	% People aged under 70 who need assistance with core activities due to a long-term health condition, disability or old age	–0.24
NOCAR	% Occupied private dwellings with no car	–0.22
OCC_DRIVERS	% Employed people classified as Machinery Operators and Drivers	–0.20
OVERCROWD	% Occupied private dwellings requiring one or more extra bedrooms (based on Canadian National Occupancy Standard)	–0.20
INDIGENOUS	% People who identified themselves as being of Aboriginal and/or Torres Strait Islander origin	–0.20
SEP_DIVORCED	% People aged 15 years and over who are separated or divorced	–0.20
NOSCHOOL	% People aged 15 years and over who did not go to school	–0.17
OCC_SERVICE_L	% Employed people classified as Low Skill Community and Personal Service Workers	–0.17
ENGLISHPOOR	% People who do not speak English well	–0.13

Index of Relative Socio-economic Advantage and Disadvantage (IRSAD)

The IRSAD is a general index which measures the relative advantage and disadvantage of an area. It is based on variables which are related to both advantaging and disadvantaging aspects of an area. It has many disadvantaging variables in common with the IRSD. A relatively low score or rank on this index indicates the area is relatively more disadvantaged and less advantaged, and a relatively high score or rank indicates that the area is relatively less disadvantaged and more advantaged. Additionally, an advantaging characteristic of an area will offset a disadvantaging characteristic. So, if an area has both relatively high disadvantaging and high advantaging characteristics, it will tend to get an IRSAD score around the middle of the distribution. Table A.2 shows a list of the variables used in IRSAD and their weights.

A.2 List of variables used for the IRSAD and their weights

Mnemonic	*Variable*	*Weight*
NONET	% Occupied private dwellings with no Internet connection	–0.29
NOQUAL	% People aged 15 years and over with no post-school qualifications	–0.29
INC_LOW	% People with stated annual household equivalised income between $13,000 and $20,799 (approx. 2nd and 3rd deciles)	–0.28
OCC_LABOUR	% Employed people classified as Labourers	–0.26
LOWRENT	% Households paying rent who pay less than $120 per week (excluding $0 per week)	–0.21
UNEMPLOYED	% People (in the labour force) unemployed	–0.20
OCC_DRIVERS	% Employed people classified as Machinery Operators and Drivers	–0.20
DISABILITYU70	% People aged under 70 who need assistance with core activities due to a long-term health condition, disability or old age	–0.20
ONEPARENT	% Families that are one parent families with dependent offspring only	–0.19
RENT_SOCIAL	% Households renting from a Government or Community organisation	–0.17
OCC_SERVICE_L	% Employed people classified as Low Skill Community and Personal Service Workers	–0.13
OVERCROWD	% Occupied private dwellings requiring one or more extra bedrooms (based on Canadian National Occupancy Standard)	–0.11
NOCAR	% Occupied private dwellings with no car	–0.11
HIGHBED	% Occupied private dwellings with four or more bedrooms	0.13
ATUNI	% People aged 15 years and over at university or other tertiary institution	0.14
HIGHMORTGAGE	% Households paying mortgage who pay more than $2,120 per month	0.23
HIGHRENT	% Households paying rent who pay more than $290 per week	0.24
DIPLOMA	% People aged 15 years and over with an advanced diploma or diploma qualification	0.24
OCC_PROF	% Employed people classified as Professionals	0.24
BROADBAND	% Occupied private dwellings with a broadband Internet connection	0.26
INC_HIGH	% People with stated annual household equivalised income greater than $52,000 (approx. 9th and 10th deciles)	0.29

Index of Economic Resources (IER)

The IER reflects the economic resources of households within an area. It is based on variables which are related to both relatively higher and relatively lower economic resources within an area. Thus, it has a similar interpretation to the IRSAD, and it also shares the "offsetting" property; an indicator of relatively higher economic resources will offset an indicator of relatively lower economic resources within an area. A relatively low score or rank on this index indicates the area has more households with relatively less economic resources and fewer households with relatively high economic resources. A relatively high score or rank indicates that the area has more households with relatively higher economic resources and fewer households with relatively less economic resources. Table A.3 shows a list of the variables used in IER and their weights.

A.3 List of variables used for the IER and their weights

Mnemonic	*Variable*	*Weight*
INC_LOW	% People with stated annual household equivalised income between $13,000 and $20,799 (approx. 2nd and 3rd deciles)	–0.31
ONEPARENT	% Families that are one parent families with dependent offspring only	–0.30
NOCAR	% Occupied private dwellings with no car	–0.30
RENT_SOCIAL	% Households renting from a Government or Community organisation	–0.29
LOWRENT	% Households paying rent who pay less than $120 per week (excluding $0 per week)	–0.28
UNEMP_POP_RATIO	% People aged 15 years and over who are unemployed	–0.27
LONE	% Households that are lone person households	–0.25
OVERCROWD	% Occupied private dwellings requiring one or more extra bedrooms (based on Canadian National Occupancy Standard)	–0.20
OWNING	% Households owning the dwelling they occupy (without a mortgage)	0.14
UNINCORP	% Occupied private dwellings with at least one person who is an owner of an unincorporated enterprise	0.20
HIGHMORTGAGE	% Households paying mortgage who pay more than $2,120 per month	0.23
HIGHRENT	% Households paying rent who pay more than $290 per week	0.24
MORTGAGE	% Households owning the dwelling they occupy (with a mortgage)	0.24
INC_HIGH	% People with stated annual household equivalised income greater than $52,000 (approx. 9th and 10th deciles)	0.27
HIGHBED	% Occupied private dwellings with four or more bedrooms	0.29

Index of Education and Occupation (IEO)

The IEO reflects the general level of education and occupation related skills of people within an area. It is based on variables which are related to both relatively higher and lower levels of skills and education levels within an area. The IEO also exhibits the 'offsetting' property that the IER and IRSAD share. It has a similar interpretation to these indexes, but it is more specific than the IRSAD, and does not include any direct measures of income or housing like the IER. A relatively low score or rank on this index indicates the area has more people with relatively less education and lower skilled occupations and relatively less people with higher education and higher skilled occupations. A relatively high score or rank on this index indicates the area has less people with relatively less education and lower skilled occupations and relatively more people with higher education and higher skilled occupations. Table A.4 shows a list of the variables used in IEO and their weights.

A.4 List of variables used for the IEO and their weights

Mnemonic	*Variable*	*Weight*
NOYEAR12	% People aged 15 years and over whose highest level of schooling completed is Year 11 or lower	–0.41
NOQUAL	% People aged 15 years and over with no post-school qualifications	–0.40
OCC_SKILL5	% Employed people who work in a Skill Level 5 occupation	–0.36
OCC_SKILL4	% Employed people who work in a Skill Level 4 occupation	–0.31
UNEMPLOYED	% People (in the labour force) who are unemployed	–0.23
CERTIFICATE	% People aged 15 years and over with a certificate qualification	–0.23
ATUNI	% People aged 15 years and over at university or other tertiary institution	0.26
DIPLOMA	% People aged 15 years and over with an advanced diploma or diploma qualification	0.35
OCC_SKILL1	% Employed people who work in a Skill Level 1 occupation	0.39

B. THEORY UNDERLYING THE INFLUENCE FUNCTION

This appendix outlines the more technical aspects of the influence function used to identify atypical and influential areas. Our goal is to make an accurate estimate of the difference between the weights using all areas in the calculation (i.e. the published weights) and the weights obtained from removing a single area. However, we want to avoid doing approximately 37,000 re-calculations of the variable weights (this will become approximately 55,000 with the introduction of SA1s).

Denote $S(F)$ as a statistic which depends on the distribution function F.[20] For the purposes of SEIFA, F is the SEIFA variable values for all areas and $S(F)$ is the function which converts the SEIFA variables values into the SEIFA variable weights. Now we can represent removing an area as a perturbation of the distribution F as follows:

$$F_Z = (1-\varepsilon)F + \varepsilon\delta_Z ,$$

where:

F_Z is the distribution of the data after perturbing it by another distribution δ_Z,

ε is the weight given to the perturbation, and

δ_Z is a distribution giving point mass on the value z, where z is a vector.

In the above expression, if we set z equal to the SEIFA variable values for the area being removed and set $\varepsilon = -(N-1)^{-1}$, where N is the number of areas, then F_Z is the distribution of the data after removing that area from the data.

Define $S\left(F_Z, \varepsilon\right)$ as the function used to calculate the SEIFA variable weights applied to the distribution F_Z instead of F. Notice that we have $S(F) = S\left(F_Z, 0\right)$. We would like to find $S\left(F_Z\right)$ with z equal to the SEIFA variable values for the area being removed and ε set equal to the negative inverse of the number of areas less one. Because ε is small and our statistic $S\left(F_Z, \varepsilon\right)$ satisfies certain properties – see Critchley (1985, Section 4) – we can perform a Taylor series expansion of $S\left(F_Z\right)$ about the point $\varepsilon = 0$ as follows:

$$S\left(F_Z, \varepsilon\right) = S\left(F_Z, 0\right) + \varepsilon\, S'\left(F_Z, 0\right) + O(\varepsilon^2) ,$$

---

[20] Note that we mean 'distribution' here in the sense of a cumulative distribution function:
$F = F\left(x_1, \ldots, x_p\right) = \Pr\left(X_1 \le x_1, \ldots, X_p \le x_p\right)$.
For SEIFA we take F as the empirical distribution function. This is calculated as the proportion of times the expression inside the probability is true over the SEIFA data set.
For example, if $x_1 = 0.2$, $x_2 = x_3 = \ldots = x_p = 0.5$, then F would be equal to the proportion of areas with the first variable (X_1) less than or equal to 0.2 and the other SEIFA variables less than or equal to 0.5.

where

- $S(F_Z, \varepsilon)$ is the SEIFA variable weights applied to the perturbed distribution F_Z instead of F;
- $S(F_Z, 0)$ is the SEIFA variable weights applied to the unperturbed F (i.e. the published SEIFA variable weights);
- $S'(F_Z, 0)$ is the influence function for the area. It is equal to the partial derivative of the function $S(F_Z, \varepsilon)$ with respect to ε evaluated at the point $\varepsilon = 0$;
- $O(\varepsilon^2)$ is a constant times ε^2 and is negligibly small as $\varepsilon \to 0$.

The approximation used is to ignore the $O(\varepsilon^2)$ term. We will be evaluating the above expression at $\varepsilon = (37,456)^{-1} \approx -0.00003$, which indicates that the error is equal to a constant times $\varepsilon^2 \approx 10^{-10}$. Thus, the change in variable weights can be approximated very accurately by:

$$S(F_Z, \varepsilon) - S(F_Z, 0) \approx \varepsilon S'(F_Z, 0).$$

Note that the above difference is a vector, and so does not allow a direct comparison between two areas. Hence, the sum of squared changes to each weight for each CD was calculated, and this was used to rank the CDs in terms of their influence. Jolliffe (1986) gives a geometrical interpretation of the sum of squared changes.[21]

[21] If the weights are considered as describing a straight line, then the sum of squared changes in the weights is an increasing function of the angle between the two lines described by each set of weights.

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