1351.0.55.036

Australian Bureau of Statistics

Research Paper

Socio-Economic Indexes For Areas: Getting a Handle on Individual Diversity Within Areas

www.abs.gov.au

New Issue

Research Paper

Socio-Economic Indexes For Areas: Getting a Handle on Individual Diversity Within Areas

Phillip Wise and Rosalynn Mathews

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 12 SEP 2011

ABS Catalogue no. 1351.0.55.036

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 6832 or email <analytical.services@abs.gov.au>.

CONTENTS

ABSTRACT 1

1. INTRODUCTION 2

2. SOCIO-ECONOMIC INDEXES 4

2.1 Notion of relative socio-economic advantage and disadvantage 4

2.2 Area level indexes of relative socio-economic advantage and disadvantage 6

2.3 Individual level indexes of relative socio-economic advantage and disadvantage 7

2.4 Observations excluded from individual level index construction 9

3. IMPORTANT CAVEATS PERTAINING TO THIS PAPER 10

3.1 A new geography standard 10

3.2 Consideration of the 15–64 year old population 10

3.3 Selection of variables in this index 11

4. PRINCIPAL COMPONENTS ANALYSIS AND INDEX CONSTRUCTION 12

4.1 Background to PCA and methodology for constructing an index score 12

4.2 Variable correlations results 15

4.3 Variable loadings results 16

5. ANALYSIS OF INDIVIDUAL LEVEL DIVERSITY 18

5.1 Index score distribution 18

5.2 Decile and grouping analysis 21

5.3 Potential for misclassification when using SEIFA as a proxy for individual-level advantage and disadvantage 27

5.4 Identifying areas with diverse patterns of individual-level advantage and disadvantage 33

5.5 An alternative measure of disadvantage – the number of indicators of disadvantage per person 36

6. CONCLUDING REMARKS 38

ACKNOWLEDGEMENTS 40

REFERENCES 41

APPENDIXES

A. VARIABLES CONSIDERED FOR INCLUSION IN SEIFI IRSAD, WITH PREVALENCES 43

B. CORRELATION ANALYSIS FOR SEIFI IRSD AND SEIFI IRSAD 44

C. VARIABLE LOADINGS FOR SEIFI IRSD AND SEIFI IRSAD 50

D. SEIFI IRSD AND SEIFI IRSAD UNIQUE SCORES 52

E. POTENTIAL FOR MISCLASSIFICATION OF PERSONS IN THE LEAST DISADVANTAGED / MOST ADVANTAGED GROUPS 54

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ARIA	Accessibility / Remoteness Index of Australia
ASGC	Australian Standard Geographical Classification
ASGS	Australian Statistical Geography Standard
CD	Collection District
Census	Australian Census of Population and Housing
IRSAD	Index of Relative Socio-economic Advantage and Disadvantage
IRSD	Index of Relative Socio-economic Disadvantage
LGA	Local Government Area
OT	Other Territories
PCA	Principal Component Analysis
SEIFA	Socio-Economic Indexes For Areas
SEIFI	Socio-Economic Indexes For Individuals
SA1	Statistical Area Level 1
SLA	Statistical Local Area

SOCIO-ECONOMIC INDEXES FOR AREAS: GETTING A HANDLE ON INDIVIDUAL DIVERSITY WITHIN AREAS

Phillip Wise and Rosalynn Mathews
Analytical Services Branch

ABSTRACT

Socio-economic indexes for areas (SEIFA) seek to summarise the socio-economic conditions of an area using relevant information from the Census of Population and Housing. The SEIFA indexes are widely used measures of relative socio-economic advantage and disadvantage at the Census Collector District level.

The indexes provide contextual information about the area in which a person lives, but within any area there are likely to be individuals with different characteristics to the overall population of that area. If inferences are made about these individuals based purely on the characteristics of the area in which they live, they could be misleading and there is potential for error in any conclusions – this is referred to as the ecological fallacy.

Using 2006 Australian Census of Population and Housing data, this paper explores individual level diversity within areas by creating and analysing two person-based socio-economic indexes: one of relative disadvantage and the other of relative advantage and disadvantage. The conceptual and methodological basis for these indexes was established by Baker and Adhikari (2007).

The primary purpose of this paper is to illustrate how individual level index scores can be used to illustrate and measure the diversity of socio-economic advantage and disadvantage within area level SEIFA. Secondary to this analysis of diversity, this paper serves to highlight the advantages of SEIFA when compared with individual level indexes of relative socio-economic advantage and disadvantage, including maximising the proportion of the population receiving an index score.

1. INTRODUCTION

Socio-Economic Indexes for Areas (SEIFA) is a set of indexes produced by the ABS after every Census. SEIFA utilises relevant Census data on education, income, employment and housing to produce index scores that rank areas based on their relative socio-economic advantage and disadvantage. One of the long-term research interests surrounding SEIFA, both within the ABS and amongst external users, has been to examine the diversity of individual level socio-economic advantage and disadvantage within areas. This paper seeks to explore this individual level diversity by calculating individual level indexes using a method introduced by Baker and Adhikari (2007).

SEIFA provides contextual information about the average level of socio-economic advantage and disadvantage across a geographical area, as opposed to the relative advantage and disadvantage experienced by a person within that area. When the area level SEIFA information is applied as a proxy for individuals or subgroups residing within an area, in order to make inferences about the socio-economic characteristics of these individuals or subgroups, then a researcher is open to the risk of the ecological fallacy. This well-established concept is further discussed in Kennedy and Firman (2004) and Baker and Adhikari (2007). The ecological fallacy is most likely to be an issue in areas where the characteristics of particular individuals or other population subgroups are too diverse to be meaningfully represented by the average characteristics of people in the area.

The ABS has previously conducted research to assess how much individual diversity within areas and the ecological fallacy can affect some uses of SEIFA. For example, Baker and Adhikari (2007) constructed individual and family level indexes of socio-economic disadvantage using 2001 Census data for the state of Western Australia. The authors found that both individual and family level relative socio-economic disadvantage was diverse within areas, and that consequently there was a high risk of an ecological fallacy if the SEIFA indexes were used as a measure of individual level disadvantage. The method established in Baker and Adhikari (2007) is extended in this paper to create an individual level index of disadvantage (SEIFI IRSD), and an individual level index of advantage and disadvantage (SEIFI IRSAD), for all Australian states and territories using 2006 Census data, to further assess the extent to which there is individual diversity within areas.

The focus of this paper is to determine how well the SEIFA indexes capture the socio-economic advantage and disadvantage of individuals within areas. Consequently, emphasis is placed on better understanding SEIFA so that it is used well. At the same time, the analysis of individual level diversity within an area is intended to inform users both of the risks associated with only considering the SEIFA score of an area, and to highlight the advantages of area level SEIFA indexes over individual level indexes of relative socio-economic advantage and disadvantage.

Baker and Adhikari (2007) list a number of issues that need addressing in moving towards an individual level index, and it is important to recognise these when considering the benefits of SEIFA over individual level indexes. This paper has addressed the issue of inclusion of advantaging variables in individual level index creation by creating both SEIFI IRSD and SEIFI IRSAD. Other issues not investigated in this paper, however, include: a means for validating the individual level indexes; the need to review the definition of individual level advantage and disadvantage; the minimisation of population exclusions; the creation of indexes for different age groups; and the selection of the best individual level Census variables. This paper therefore falls short of producing a well-established individual level index. For the purposes of this paper, the method used in Baker and Adhikari (2007) has been adopted, although it is important to recognise that the above points require consideration in the broader context of index calculation.

This paper is structured as follows. Section 2 introduces the background, underlying concepts and framework surrounding the notion of socio-economic advantage and disadvantage used for SEIFA, and how they relate to an individual level index. Section 3 discusses some caveats associated with the derivation and application of the individual level indexes in this paper. Section 4 outlines the data and methodology used in the construction of the individual level indexes. In Section 5, we investigate the diversity of individual socio-economic advantage and disadvantage within areas by comparing population groups across the individual and area level indexes. A further investigation using the number of indicators of disadvantage as a basis for analysis is also detailed in Section 5. The conclusion contains a summary of findings and a reflection on the possible future directions for research in this sphere of work.

2. SOCIO-ECONOMIC INDEXES

This section covers the conceptual basis of the research, and provides details on the key methodological issues feeding into the creation of individual level indexes of socio-economic advantage and disadvantage. It covers the notion of advantage and disadvantage used to create the individual level indexes, the contextual differences between area and individual level advantage and disadvantage, the implementation of a procedure for creating variables related to the concept of individual level advantage and disadvantage, and the overall implications for the scope of the indexes.

2.1 Notion of relative socio-economic advantage and disadvantage

The notion of relative socio-economic advantage and disadvantage used for creating the 2006 SEIFA indexes is established in ABS (2008a). The concept of advantage and disadvantage underpinning the SEIFA methodology can broadly be 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.

The concept has been extensively discussed in ABS (2008a, 2008b) and Adhikari (2006), and the reader is encouraged to read these references for a better understanding of area level socio-economic advantage and disadvantage.

There is a long history of measuring the concept of advantage and disadvantage at the individual level. The work of Townsend (1979, 1987) helped develop key concepts which have been employed in major studies such as 'Breadline Britain' and 'Poverty and Social Exclusion 1999' (Mack and Lansley, 1985; Gordon and Pantazis, 1997; Gordon, Adelman *et al.*, 2000).

Conceptualising an individual's socio-economic disadvantage begins with the consideration that they are 'unable to participate fully in society' (Vinson, 2007). Further, this concept requires an appraisal of whether the socio-economic conditions experienced by individuals can be considered disadvantaged relative to the wider community (Townsend, 1987). The conceptual aspect of socio-economic advantage, on the other hand, is sometimes inferred as a lack of disadvantage (ABS, 2008b), but is also captured directly through measures appropriate for the analysis purpose (ABS, 2011c).

It is important to highlight the distinction between individual and area level socio-economic advantage and disadvantage, because the difference in contexts can cause significant errors to be made when making inferences about an area or population (Lim and Gemici, 2011). For example, the ABS (2008a) stresses that CD level SEIFA information should not be used for individual analysis. To emphasise the need for care when constructing measures of socio-economic status, ABS (2011c) discusses at length the important issues that require consideration, such as conceptual relevance, appropriateness of analysis units, what aspects of socio-economic status are relevant for

use and data availability. Other in-depth appraisals of available individual and area level socio-economic measures can be found in Bailey *et al.* (2003) and Morris and Carstairs (1991), whilst a discussion about the issues associated with moving between individual and area level socio-economic measures is contained in Marks *et al.* (2000). For example, Marks *et al.* (2000) highlights research presented in Ainley and Long (1995) reporting correlations of 0.36 to 0.45 between individual and CD level socio-economic status scores for a sample of secondary school students. Similarly, Lim and Gemici (2011) find that SEIFA misclassifies the socio-economic status of 40% of 15–25 year old individuals using the 2003 cohort of the Longitudinal Surveys of Australia's Youth.

For the purpose of this paper, the area level definition aids understanding of the concept of individual level socio-economic advantage and disadvantage. This approach was also taken in the derivation of the individual level New Zealand index of socio-economic deprivation (Salmond *et al.*, 2006), which used the same theoretical basis as the national New Zealand census-based area level indexes of relative socio-economic deprivation. Baker and Adhikari (2007) similarly used the SEIFA concept of advantage and disadvantage as a basis for individual level socio-economic advantage and disadvantage. This means that the individual level concept of advantage and disadvantage developed here is very similar to the area level concept used for 2006 SEIFA (ABS, 2008a).

Since relative socio-economic advantage and disadvantage is a complex and multi-dimensional concept, it is difficult to condense into a single index with a manageable, accessible framework. The limitations of the data collected in the Census also place restrictions on the scope of the notion of advantage and disadvantage available to be used. The most important dimensions covered by the SEIFA indexes, and other socio-economic indexes developed around the world, include occupation, income and education.

An important point to consider at all times is both what is measured by a socio-economic index, and what is *not* measured. This helps clarify the scope for measuring relative socio-economic advantage and disadvantage. To illustrate, SEIFA indexes could indicate a particular area is relatively less disadvantaged than another area; however, expanding the scope of disadvantage to include pollution and crime rates (for example), may change this interpretation. Also, the 2006 SEIFA index of relative socio-economic disadvantage contains only variables capturing aspects of disadvantage, and hence only classifies areas from most disadvantaged to least disadvantaged. Bailey *et al.* (2003) considered the theory of deprivation indexes and found that individual level approaches tended to have a narrow focus on the 'necessities of life', whereas area based measures encompassed a wider range of issues concerned with concentrations of deprivation. The issue here is context, and how the notions captured by the socio-economic indexes reflect the concept of advantage and disadvantage analysed.

Despite the fact that area and individual level advantage and disadvantage are separate concepts, there are clear commonalities between the two. For the purposes of this paper, the working definitions of area level and individual level advantage and disadvantage are the same as those employed in the research paper by Baker and Adhikari (2007). Specifically:

- 'Area level disadvantage is related to the characteristics of the community or neighbourhood as reflected in the attributes of the people living in that area' (page 5).
- 'Individual level socio-economic disadvantage is a more personal concept relating to a person's own ability to access resources and participate in society' (page 5).

Hence, socio-economic advantage and disadvantage at the individual level for this paper is considered to be defined in terms of an individuals' access to resources, and their ability to participate in society. This is to be measured using the same scoping list of variables considered for the 2006 SEIFA index of disadvantage, and the index of advantage and disadvantage, to produce two individual level indexes reflecting the notion of relative advantage and disadvantage.

2.2 Area level indexes of relative socio-economic advantage and disadvantage

SEIFA is a suite of four indexes released at the Census Collector District (CD) level (ABS, 2006), and each index is designed to capture slightly different aspects of the notion of advantage and disadvantage employed by the ABS. To achieve this, each index is composed of different variables derived from Census data.

Among the four SEIFA indexes, the most commonly used is the Index of Relative Socio-economic Disadvantage (IRSD). The IRSD was designed as a general measure of relative socio-economic disadvantage at the area level. A low score on this index reflects an area with relatively high levels of socio-economic disadvantage; however a high score on this index indicates a relative lack of disadvantage. Hence, the IRSD is only appropriate for comparing areas in terms of relative disadvantage.

The Index of Relative Socio-economic Advantage and Disadvantage (IRSAD) was designed as a general measure of both relative socio-economic advantage and disadvantage at the area level, and hence offsetting of advantaged and disadvantaged characteristics is possible. A low score on this index reflects an area with relatively high levels of socio-economic disadvantage, whilst a high score on this index indicates an area with high levels of advantage.

In this paper, these two SEIFA indexes are used as the basis for constructing SEIFI IRSD and SEIFI IRSAD.

It is worth noting the advantages of SEIFA at this point so that the challenges facing the construction of individual level indexes can be better understood. SEIFA is an important, robust product with an established methodology and a long history of use in research. To begin with, 2006 SEIFA covered 99.4% of the Australian population. This is a very high proportion, which is one of the central goals of SEIFA. The area level nature of SEIFA was also theoretically and conceptually tested, and was proven to be sound in practical applications. The aggregate nature of the data, and stringent exclusion rules, work to ensure that missing data and nonresponse are minimised, confidentiality is upheld, and there is sufficient meaningful data in an area to support index construction.

2.3 Individual level indexes of relative socio-economic advantage and disadvantage

The construction process for the two individual level indexes began with the same initial scoping lists of variables related to socio-economic advantage and disadvantage that was used to construct the 2006 SEIFA IRSD and IRSAD (see tables 2.1 and A.1 for details of these variables and their prevalence within the included population).

Two variables from the scoping lists for the 2006 SEIFA IRSD and IRSAD, '% of occupied private dwellings requiring one or more extra bedrooms' (variable mnemonic: OVERCROWD) and '% occupied private dwellings with one or more bedrooms spare' (variable mnemonic: SPAREBED), were not considered. It was determined to be infeasible to calculate these variables at the individual level and so they were omitted from this research. However, these variables both capture important aspects of socio-economic advantage and disadvantage not covered by the other variables, and this issue should be considered in any future work in this sphere.

Each area level SEIFA variable from the scoping lists needed to be transformed to the individual level for the purposes of the individual level index construction. However, the individual level indexes are based on personal records whilst SEIFA was derived from summary statistics at the CD level. This means that the variables cannot be mapped from the area to the individual level without careful transformation and adjustment. The method used for this paper involved transforming all area level variables into binary indicators at the individual level. For example, the continuous area level variable '% People who do not speak English well' became a binary variable with value 1 if the person could not speak English well, and 0 otherwise. This method was also used to derive the individual level New Zealand Index of socio-economic Deprivation (Salmond *et al.*, 2006).

2.1 List of variables considered for the individual level index of disadvantage, with prevalence (%)

Individual level variable	*Code*	*(%)*
Persons aged 15 years and over with no post-school qualifications	noqual	45.97
Persons aged 15 years and over who left school after year 11 or lower	noyear12	49.32
Person has stated annual household equivalised income between $13,000 and $20,799	inc_low	13.01
Person is employed in the sector classified as low skill clerical and administrative workers	occ_admin_l	11.38
Person is separated or divorced	sep_divorced	11.60
Person is employed in the sector classified as labourers	occ_labour	10.62
Person is employed in the sector classified as low skill sales workers	occ_sales_l	7.55
Person is employed in the sector classified as machinery operators and drivers	occ_drivers	6.76
Person is employed in the sector classified as low skill community and personal service workers	occ_service_l	6.76
Person in the labour force is unemployed	unemployed	5.29
Person does not speak English well	englishpoor	2.32
Person aged 15 years and over did not go to school	noschool	0.66
Person identified themselves as being of Aboriginal and/or Torres Strait Islander origin	indigenous	2.02
Person under the age of 70 has a long-term health condition or disability and needs assistance with core activities	disabilityU70	2.47
Person in in a one parent family with dependent offspring only	oneparent	10.30
Person resides in an occupied private dwelling with no internet connection	nonet	29.90
Person resides in an occupied private dwelling with no car	nocar	7.00
Person resides in a household renting from Government or community organisations	rent_social	4.46
Person resides in an occupied private dwelling with one or no bedrooms	fewbed	4.33
Person resides in an occupied private dwelling paying less than $120 rent per week (but not $0)	low_rent	13.25

This raises a practical difficulty in individual level variable creation: applicability. Many of the Census variables relating to advantage and disadvantage address factors such as employment, education and economic resources. These aspects of advantage and disadvantage are not necessarily relevant for all persons in the population. For instance, young people under the age of 15 will most likely not have completed their education or be employed, and so will have socio-economic characteristics largely determined by their parent or guardian, whilst retirees over the age of 64 find their accumulated wealth is a better indicator of their economic standing than their income. Therefore, because of the different stages of the life cycle, it is not practical to have a unique individual level index for all demographics of the population, and only those individuals between the ages of 15 and 64 are included in this research. This is in line with the previous research on individual level indexes performed in Baker and Adhikari (2007), and reflects a recommendation made in Bailey *et al.* (2003) to derive separate individual level indexes of adult and child deprivation.

2.4 Observations excluded from individual level index construction

In order to maintain data quality and clarify the conceptual meaningfulness of the included variables, certain population exclusions were made to the 2006 Census data. The criteria used to determine these exclusions were selected with two goals in mind: one was to stay in line with the research undertaken by Baker and Adhikari (2007), and the other was to align the population inclusions with 2006 SEIFA.

The specific rules for excluding individuals are detailed following:

- *Non-response:* People were excluded if they did not respond to all relevant person, family and dwelling level Census variable questions.

  No persons were excluded from the analysis based on this criterion.

- *Consistency with 2006 SEIFA:* This aids comparisons of results once the index is finalised, and involves including only those people found in CDs that were included in the original 2006 SEIFA analysis.

  There were 157,491 persons excluded from the analysis based on this criterion.

- *Age restrictions:* For applicability (as discussed in Section 2.3), persons below 15 years of age and above 64 years of age were excluded.

  There were 6,541,598 persons excluded from the analysis based on this criterion.

The final dataset for the individual level index analysis that is the subject of this research includes 13,156,201 persons. This dataset has been analysed against the scoping list of variables for the 2006 SEIFA IRSD, and table 2.1 contains results for the prevalence of the different disadvantage variables considered for inclusion in the index. For example, from table 2.1, approximately 10% of the in-scope population is employed in the sector classified as labourers. The corresponding prevalence table, table A.1, for the scoping list of variables considered for inclusion in the individual level index of advantage and disadvantage is contained in Appendix A. In both cases the prevalence is a proportion of the total 13,156,201 persons included in the analysis.

One of the goals of SEIFA is to maximise the proportion of the Australian population to which a SEIFA score is given. However, this research excludes approximately one-third (33.15% or 6,699,089 persons) of the total population. This means that when interpreting the results in this paper, the reader should keep in mind that this only applies to a very specific subset of the population - the 15–64 year old population. Identified areas of further research into this issue include the formulation of age-specific indexes drawing on appropriate information for the different age groups (0–14, 15–64, 65+), and not restricting the analysis to be consistent with 2006 SEIFA. These options have been left for further research.

3. IMPORTANT CAVEATS PERTAINING TO THIS PAPER

This section seeks to put into perspective and highlight some important caveats that require consideration when interpreting the index construction and analysis presented in Sections 4 and 5.

3.1 A new geography standard

The ABS is planning to release SEIFA 2011 in early 2013. The next release of SEIFA will involve the implementation of the new Australian Statistical Geography Standard (ASGS), which replaces the existing Australian Standard Geographic Classification (ASGC) from 1 July 2011. The main impact this will have on SEIFA is that the CD will no longer be the base geographical unit for SEIFA analysis; the new ASGS structure to be used in its place is the Statistical Area Level 1 (SA1).

Given the design criteria for SA1s compared to CDs, there is a general expectation that SA1s will better capture the socio-economic gradient within areas. That is, because SA1s more clearly define urban and rural areas, small rural towns and discrete Indigenous communities, the amount of diversity within SA1 areas may be reduced. It is possible that the shift from CDs to SA1s could alter the findings presented in this paper. For this reason it could be misleading to apply conclusions from this analysis (based on CDs) to SEIFA 2011 (which will be based on SA1s).

For more general information on the ASGC, the ASGS and regarding the 2011 Census, refer to ABS (2007, 2008c, 2010, 2011a and 2011b).

3.2 Consideration of the 15–64 year old population

The individual level indexes in this paper are derived for the 15–64 year old population. As described in Sections 2.3 and 2.4, concerns with conceptual issues surrounding variable construction lead to substantial population exclusions, limiting the scope of the indexes to the 15–64 year old population. This resulted in approximately one-third of the population counted in the 2006 Census being excluded from the analysis.

The population exclusion figure at the individual level is in stark contrast to 2006 SEIFA, where only 0.6% of the population was excluded. This vast difference reflects an advantage of SEIFA for index construction, namely that most of the population receives an index score.

When considering comparisons of the individual level indexes to SEIFA, it is important to remember that SEIFA uses data for all age groups, but the individual level indexes are for the 15-64 year old population only.

3.3 Selection of variables in this index

The issue of relevance of different measures of socio-economic status, including comparisons of area and individual level measures, is discussed in depth in ABS (2011c). The discussion centres around how the suitability of a measure of socio-economic status depends on the aims of an analysis and the data being used, and this model has been used to inform the construction of the individual level indexes in Sections 2 and 4.

For the purposes of this paper, the individual level indexes draw together both area based concepts and individual level considerations to contextualise the available Census information at the individual level. Whilst Section 2.1 explains the concept and basis for the individual level indexes constructed in this paper, it is important to recognise that the individual level indexes addressed in this paper are not based on a commonly agreed framework for individual level socio-economic advantage and disadvantage. The creation of individual level indexes from scratch, rather than based on a method designed for areas, might involve different, more representative variables for index construction.

This paper presents some individual level indexes that have been constructed for the specific aims of this analysis – they are not an attempt to establish a generic set of individual level indexes.

4. PRINCIPAL COMPONENTS ANALYSIS AND INDEX CONSTRUCTION

This section presents a brief introduction to Principal Components Analysis (PCA), the statistical technique used to construct the individual level indexes and SEIFA, with an overview of the associated outputs and their interpretation. The correlation and loading results of the individual level index creation process are then detailed.

4.1 Background to PCA and methodology for constructing an index score

SEIFA indexes are calculated using a statistical analysis technique called principal components analysis (PCA). The reason PCA is used to construct SEIFA indexes is because it is a very effective technique at reducing vast quantities of data (such as the data available through the Census) into manageable segments. The basic means by which PCA achieves this is by summarising a large number of correlated variables into a smaller set of transformed variables, called the principal components. Each principal component is thus a weighted linear combination of the original variables.

It is possible to have as many principal components as there are variables in a dataset. However, the individual level indexes and SEIFA utilise only the first principal component of the standardised variables because this is the one component designed to explain the maximum amount of variation in a dataset. The first principal component from PCA gives a line of best fit[1] for each index (summarising the common trend in the underlying set of variables). This line is a weighted linear combination of the variables comprising each index. The raw index score for each area is created from the weights, which can be expressed as follows:

$$Y_i = x_{i1}w_1 + x_{i2}w_2 + \cdots + x_{ip}w_p$$

where Y_i is the raw index score for the i-th CD; x_{ip} is the standardised variable value of the p-th variable for the i-th CD; and w_p is the weight for the p-th standardised variable, determined by the PCA.

The details for PCA are the same for the individual level index, but instead of area level results for Y_i, these values represent the raw scores for each individual in the included dataset.

The individual level indexes deal with binary variables in the form of indicators, thus limiting the number of unique scores possible in an index of disadvantage to 2^P. If P is 8, then we can expect 256 unique scores from the output distribution. This limited split of scores restricts the differentiation available between individuals, especially considering the included population totals over 13 million, theoretically giving an

1 The line of best fit minimises the sum of squared errors between the standardised variable values and their estimated value from using the first principal component. The estimate for the p-th variable in the i-th area is given by the first principal component as $Y_i w_p$.

average of 51,391 per unique score. The index score will largely be driven by the number of indicators an individual has, with the weights determining the relative importance, or contribution to the index, of particular combinations of indicators. This consideration suggests a high degree of clumping of individuals on scores will be present in the indexes.

Constructing an individual level index score

The basic process for constructing a socio-economic index using PCA (ABS, 2008b) is summarised following:

1. *Create initial variable scoping list.*

   This is a listing of the variables derived from the Census data that are determined to relate to the overall concept of socio-economic advantage and disadvantage captured by the index in question.

2. *Construct the variables.*

   For this research into individual level indexes, the variables are individual level binary indicators.

3. *Remove highly correlated variables.*

   This prevents instability in the variable weights and over-representation of any specific socio-economic characteristic. When two variables have a correlation greater than |0.8|, discretion is employed: one variable is generally removed if the correlated variables capture similar aspects of socio-economic disadvantage (for example, persons who did not go to school and persons who left school before completing year 12); if the correlated pair of variables measures different socio-economic characteristics, then depending on the size of the correlation and the particular variables involved, one variable may be dropped.

4. *Conduct initial PCA to obtain loadings.*

   This step is used to obtain the loading for each variable on the first principal component.

5. *Remove low loading variables.*

   Variables with loadings below |0.3| were excluded since this was taken as an indication that they were not strong indicators of relative advantage or disadvantage. Variables were removed one at a time, starting with the lowest loading variable, and the PCA was re-run. The threshold of |0.3| is an accepted level in the PCA literature (see Joliffe, 1986, pp. 108, 111).

6. *Conduct PCA on the reduced variable list.*

   After removing variables, the process in step 5 is repeated to ensure no remaining variable loadings fall below $|0.3|$, or in other words that each included variable contributes significantly to the final index.

7. *Standardise component/index scores.*

   The index is standardised to have a mean of 1000 and a standard deviation of 100. This step is done for presentation purposes, since no alteration is made to the underlying rankings during the standardisation process.

8. *Reverse signs of loadings and weights.*

   Reversing the signs of the loadings and weights ensures that advantage indicators now have positive weights and loadings, again for ease of interpretation when the results are presented.

Special considerations for the construction of an individual level index

Whilst the basic outline of the index construction process, and the description of the PCA outputs accompanying the process, remains unchanged for the individual level index researched in this paper, there are a few key points to declare before proceeding with the analysis.

PCA is most typically performed on continuous variable data, and when ordinal variables are used they are treated as if they were continuous. As discussed, the calculation of a correlation matrix for these binary variables takes place before running the PCA; the correlations are derived using Pearson's correlation coefficient. Utilising this type of calculation to get the correlation matrix for the binary variable data leads to biased PCA results (Rigdon and Ferguson, 1991).

The method employed to work around this bias is to conduct the PCA on a tetrachoric correlation matrix (also known as polychoric correlation if ordinal variables are used) (Olsson, 1979), as was performed for the previous research paper Baker and Adhikari (2007). As they describe in their paper, tetrachoric correlation calculates the correlation between latent variables (which are assumed to underlie the binary variables). So whilst only binary characteristics are being observed for each individual, it is assumed that there is an underlying continuous variable determining this outcome.

Outputs from PCA

PCA produces the following outputs, associated with the first principal component:

- *Variable loadings* representing the correlation between the latent variable upon which the correlations are based and the principal component;
- An *eigenvalue* indicating the variance of the component. The eigenvalue divided by the number of variables gives the percentage of the variation in the dataset explained by the principal component;
- *Variable weights* are the coefficients used in the linear transformation that produces the principal component.

These outputs are valuable statistics that help determine the component fit and ensure the best variables are selected for inclusion in the final index.

4.2 Variable correlations results

As discussed in the previous section, one of the first analytical stages of the index construction process involves calculating the correlations between each pair of identified variables. Highly correlated variables (that is, variables with correlations greater than $|0.8|$) can introduce instability to the subsequent variable weights (ABS, 2008b) and indicate 'double counting' of specific aspects of socio-economic advantage and disadvantage, so these variables are reviewed for inclusion in the final index.

Appendix B contains the detailed results on the tetrachoric correlations for the SEIFI IRSD and SEIFI IRSAD indexes, including the full correlation tables and specific tables for those variable pairs with correlations greater than $|0.8|$, as well as discussion surrounding the inclusion of highly correlated variable pairs.

The correlation analysis for the creation of the SEIFI IRSD index revealed that the variables *person did not go to school* (NOSCHOOL) and *person left school before year 12* (NOYEAR12) had a high correlation (0.999). This clearly indicates that the number of people who did not go to school is well captured by the number of people who did not complete year 12. Therefore the NOSCHOOL variable was dropped, since the prevalence of this variable in the population (0.66%) was much lower than the prevalence of the NOYEAR12 variable (49.32%) (see table 2.1 for further details).

Whilst the correlation analysis for the creation of the SEIFI IRSAD index revealed several highly correlated variable pairs, each variable involved captured a different aspect of socio-economic advantage or disadvantage. Therefore, no variables were dropped from the correlation analysis of the variables considered for the SEIFI IRSAD index.

4.3 Variable loadings results

The next stage of index construction is to run an initial PCA and examine the variable loadings on the first principal component. If the loading is below the |0.3| threshold prescribed in the literature (for more information, see Joliffe, 1986) then the contribution made by the variable to the first component is minor and does not significantly improve the ability of the overall index to explain the variation in the dataset. Therefore the variables with loadings less than |0.3| are excluded. An iterative process is used whereby the variable with the lowest loading is dropped first, and then the PCA is repeated until all variables have a loading above |0.3|.

Following this procedure, several variables were dropped from both the initial scoping list for SEIFI IRSD and SEIFI IRSAD. For further details on those variables dropped, and the iterative order in which they were dropped, refer to Appendix C.

PCA was run on the reduced variable list for each index to give the ultimate loadings and weights. The results from this analysis for the SEIFI IRSD index are displayed in table 4.1, along with a comparison to the corresponding 2006 SEIFA IRSD loadings and weights for the variables included in the individual level index. The results for the SEIFI IRSAD index are displayed in table 4.2, along with a comparison to the corresponding 2006 SEIFA IRSAD loadings and weights.

4.1 List of included variables for individual level index of disadvantage, and corresponding SEIFA IRSD analytics

	SEIFI IRSD		*SEIFA IRSD*	
Variable	*Loading*	*Weight*	*Loading*	*Weight*
rent_social	–0.75	–0.52	–0.70	–0.27
lowrent	–0.72	–0.50	–0.67	–0.26
nonet	–0.50	–0.34	–0.85	–0.33
nocar	–0.49	–0.34	–0.57	–0.22
indigenous	–0.44	–0.30	–0.52	–0.20
noqual	–0.35	–0.24	–0.76	–0.30
noyear12	–0.33	–0.23	N/A	N/A
inc_low	–0.32	–0.22	–0.76	–0.30

Comparing the loadings and weights in table 4.1, the area based SEIFA IRSD index had higher variable loadings generally compared to the individual level SEIFI IRSD index. The variables with the highest loadings for the SEIFI IRSD were RENT_SOCIAL and LOWRENT; the loadings for these variables were interestingly similar to the 2006 SEIFA IRSD results. The loadings for these two variables are also much higher than the next largest loading (for NONET), indicating that these two variables contribute significantly to the overall notion of disadvantage in the SEIFI IRSD constructed for this research.

4.2 List of included variables for individual level index of advantage and disadvantage, and corresponding IRSAD analytics

	Individual		*IRSAD*	
Variable	*Loading*	*Weight*	*Loading*	*Weight*
noqual	–0.54	–0.34	–0.88	–0.29
noyear12	–0.54	–0.34	N/A	N/A
nonet	–0.41	–0.32	–0.87	–0.29
rent_social	–0.40	–0.25	–0.51	–0.17
lowrent	–0.37	–0.23	–0.64	–0.21
inc_low	–0.34	–0.21	–0.83	–0.28
broadband	0.45	0.28	0.79	0.26
inc_high	0.50	0.31	0.86	0.29
occ_prof	0.60	0.38	0.73	0.24
degree	0.68	0.43	N/A	N/A

The final PCA for SEIFI IRSD that was carried out on the eight variables revealed that the first principal component had an eigenvalue of 2.19. The eigenvalue is used to determine the variation of the dataset that is explained by the first component. Based on this metric, the first component explains 27.38% of the variation in the dataset. Referring to ABS (2008b), this figure is lower than the corresponding result for the area level 2006 SEIFA IRSD (39%).

Table 4.2 shows that the variables with the highest loadings for the SEIFI IRSAD PCA were DEGREE and OCC_PROF. The 2006 SEIFA IRSAD variables had higher loadings then the SEIFI IRSAD variables, for each included variable in SEIFI IRSAD.

The final PCA for SEIFI IRSAD was carried out on the ten variables included in table 4.2, and the first component was found to have an eigenvalue of 2.54. Hence, the first component explains 25.40% of the variation in the dataset. Again, this is a much lower value than was calculated for the area level 2006 SEIFA IRSAD (44%).

Now that both individual level indexes have been constructed, analysis and investigation can proceed into both the properties of the two indexes and how they can be used to illustrate diversity within areas.

5. ANALYSIS OF INDIVIDUAL LEVEL DIVERSITY

The individual level indexes of both disadvantage (SEIFI IRSD), and advantage and disadvantage (SEIFI IRSAD), were constructed following the process described in Section 4.1. Before introducing any analytical output, for clarification in interpreting the results:

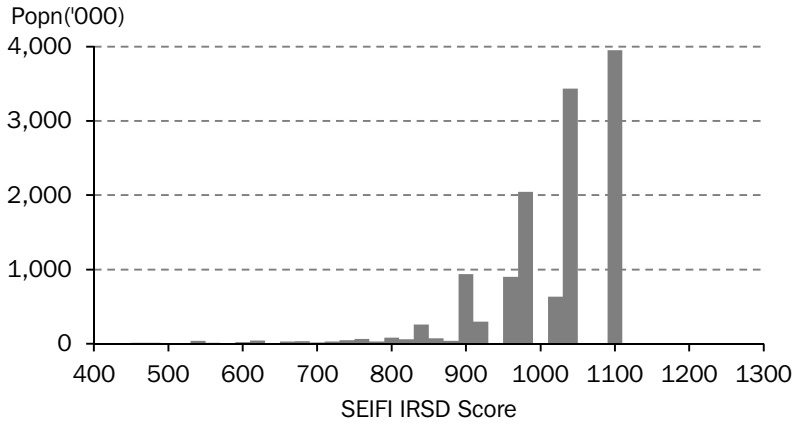
- A low score on the SEIFI IRSD index represents an individual with relatively more socio-economic disadvantage, whilst a high score on this index represents relatively less socio-economic disadvantage.
- A low score on the SEIFI IRSAD index represents an individual with relatively more socio-economic disadvantage, whilst a high score on this index represents relatively more socio-economic advantage.

The scores themselves are not continuous but ordinal, and thus require some care in interpretation. To illustrate, someone with a score of 400 on one index is not twice as disadvantaged as a person with a score of 800. However, we can say they are relatively more disadvantaged.

5.1 Index score distribution

The first thing that we examine is the distribution of the SEIFI IRSD and SEIFI IRSAD scores. Figure 5.1 reveals the distribution of the SEIFI IRSD scores. The scores are presented on the horizontal axis, with the vertical axis containing the corresponding 15–64 year old population counts against the different scores.

5.1 SEIFI IRSD score distribution



What is most striking about the distribution is the high degree of clumping at the least disadvantaged end of the spectrum. This is very similar to what was observed in Baker and Adhikari (2007). There are very few unique scores above the mean of 1000, with a long tail of low SEIFI IRSD scores. The clumping visible on the high index scores indicates that the majority of the 15–64 year old population do not have more than

one indicator of disadvantage from the variables selected for constructing the SEIFI IRSD index.

The distribution clearly indicates that the SEIFI IRSD index is very good at delineating between the relative socio-economic disadvantage of the most disadvantaged persons in the 15–64 year old population. However, as expected, the SEIFI IRSD distribution has virtually no capacity to provide a means to compare persons with few indicators of disadvantage.

The SEIFI IRSD index is limited to 256 unique scores because of the interactions available through the eight indicator variables related to socio-economic disadvantage selected in the index creation process. As figure 5.1 shows, the clumping in the distribution indicates that there are large proportions of the population with certain combinations of these indicators. For instance, the large spike on the right hand side of the distribution represents approximately 30% of the population, all of whom hold none of the indicators of disadvantage feeding into the SEIFI IRSD index. Each SEIFI IRSD score reflects a certain number indicators of disadvantage and their relative loading.

Figure 5.2 reveals the distribution of the SEIFI IRSAD scores. Again, the scores form the horizontal axis, whilst the vertical axis presents the corresponding 15–64 year old population counts against the different scores.

5.2 SEIFI IRSAD score distribution

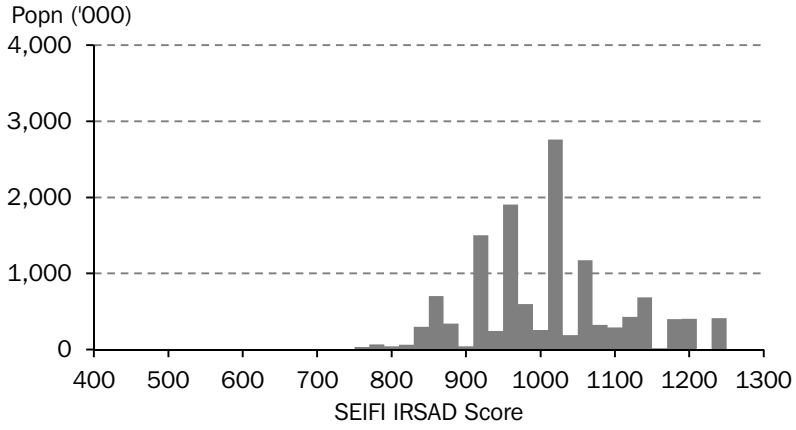


Figure 5.2 shows a range of scores with a lower degree of clumping than was observed for SEIFI IRSD, a result of the inclusion of advantaging variables in the index construction. The most prevalent scores lie across the middle of the distribution, with the frequency of 15–64 year old persons tapering off towards the tails of the distribution. There is noticeably less clumping at the advantaged end of the SEIFI IRSAD distribution compared to the SEIFI IRSD, which exhibited severe negative skewness.

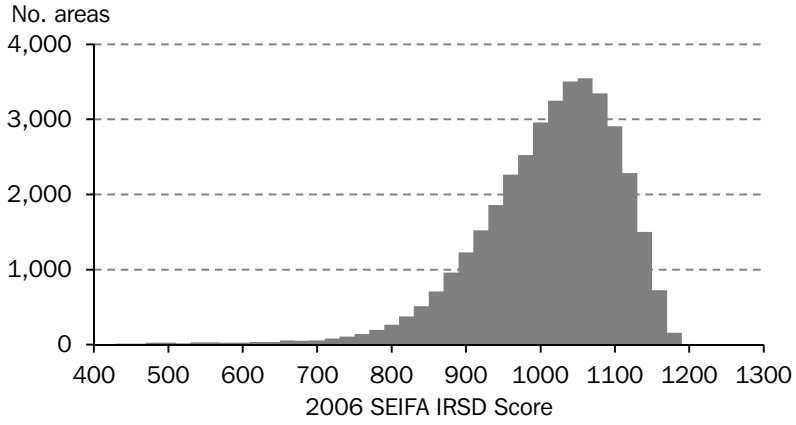
The addition of advantaging variables results in a smoother distribution, as well as more unique scores and thus a greater capacity to compare relative socio-economic advantage and disadvantage. However, it should be noted though that the SEIFI IRSD distribution provides greater means for comparing the most disadvantaged persons in the 15–64 year old population.

Appendix D contains a more in-depth examination of the number of unique scores, and the most prevalent scores, in both the SEIFI IRSD and SEIFI IRSAD distributions.

Area level score distributions

It will be instructive to compare the 2006 SEIFA IRSD distribution to the corresponding individual level SEIFI IRSD score distribution (figure 5.1), to observe any similarities or differences and what these could represent. Figure 5.3 shows the 2006 SEIFA IRSD score distribution.

5.3 2006 SEIFA IRSD score distribution

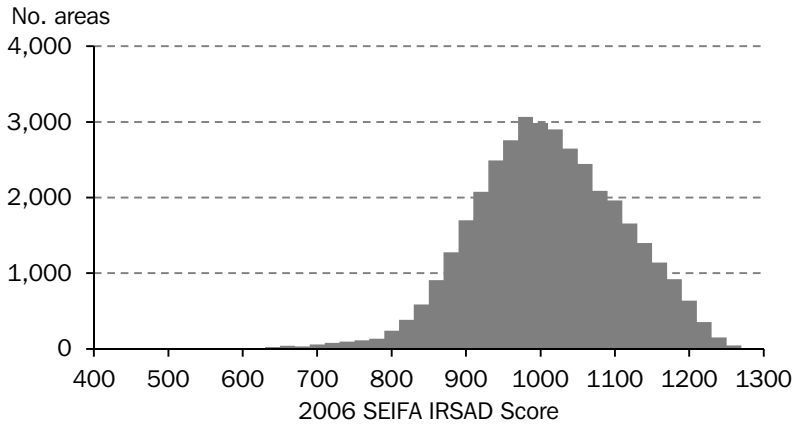


As with the SEIFI IRSD, the 2006 SEIFA IRSD only seeks to capture socio-economic disadvantage. There is little scope for the indexes to distinguish between CD areas or individuals with lower levels of disadvantage. This results in the scores being negatively skewed for both indexes. However, the concentration of scores and the peak in the distribution towards the least disadvantaged end of the spectrum are both smoother and wider for SEIFA IRSD than the obvious clumping present at the corresponding end of the SEIFI IRSD score distribution.

Similarly, the individual level SEIFI IRSAD distribution (figure 5.2) can be compared to the corresponding 2006 SEIFA IRSAD score distribution. Figure 5.4 shows the IRSAD score distribution.

The SEIFA IRSAD score distribution is characteristically more reminiscent of a Normal distribution, with a distinct peak and roughly equivalent tails. There appears to be a slightly longer tail of low SEIFA IRSAD score frequencies, but overall the distribution is clearly less skewed than the corresponding individual level SEIFI IRSAD distribution.

5.4 SEIFA IRSAD score distribution



It is clear from this comparison of the area and individual level score distributions that area level indexes give much clearer distributions without significant clumping and dispersion of scores. Hence, area level indexes provide better quality information for rankings.

5.2 Decile and grouping analysis

A decile is a segment containing 10% of the population formed on an auxiliary variable: for example deciles can be based on the SEIFI IRSD score, with decile 1 representing the most disadvantaged persons in the 15–64 year old population and decile 10 the least disadvantaged. We recommended users of 2006 SEIFA look at the decile of an area to help understand the average relative socio-economic disadvantage of that area, and also to aid in comparative analyses (ABS, 2008a).

SEIFI IRSD groups

Due to the high degree of clumping in the SEIFI IRSD score distribution, formulating deciles is clearly not possible. However, we want a small set of meaningful groups to enable more simple analyses. For reporting purposes, an appraisal of the cumulative 15–64 year old population proportion against the corresponding SEIFI IRSD score showed that the most consistent split would involve using four groups: the first two groups represent approximately the most and second most disadvantaged 20% of the 15–64 year old population, whilst the third and fourth groups represent approximately the least and second least disadvantaged 30% of the 15–64 year old population. The frequency distribution of the group assignation is summarised in table 5.5. Splitting the included population into groups was also performed in Baker and Adhikari (2007) because of the clumping present in the score distribution.

5.5 Frequency distribution of SEIFI IRSD groups

	15–64 year-old population		*SEIFI IRSD score*	
SEIFI IRSD group	*Frequency*	*Percentage*	*Minimum*	*Maximum*
1	2,547,876	19.37	388	942
2	2,587,970	19.67	943	975
3	4,067,623	30.92	1004	1035
4	3,952,732	30.04	1094	1094

SEIFI IRSAD groups

The SEIFI IRSAD distribution is less clustered than the SEIFI IRSD distribution, however with the tendency of individuals to again cluster on specific scores, it is not possible for deciles to be formed. This is why figures in the frequency table of SEIFI IRSAD groups do not round off conveniently to 10%, and is also why the terminology of groups is retained for SEIFI IRSAD analysis. The distribution of the 15–64 year old population into groups is displayed in table 5.6.

5.6 Frequency distribution of SEIFI IRSAD groups

	15–64 year-old population		*SEIFI IRSAD score*	
SEIFI IRSAD group	*Frequency*	*Percentage*	*Minimum*	*Maximum*
1	1,261,475	9.59	744	871
2	1,316,942	10.01	872	908
3	1,514,895	11.51	909	949
4	1,135,596	8.63	950	961
5	1,498,832	11.39	962	1003
6	1,023,098	7.78	1004	1011
7	1,579,330	12.00	1012	1052
8	1,198,331	9.11	1053	1079
9	1,251,001	9.51	1080	1138
10	1,376,701	10.46	1139	1234

Roughly, the most disadvantaged 10% of the 15–64 year old population falls into group 1, whilst group 10 contains the most advantaged 10%. The smallest group in terms of 15–64 year old population proportion is group 6 with 7.78%, compared to group 7 with the largest percentage at 12% due to clustering at this point in the distribution of scores.

Now that groups have been established for the SEIFI IRSD and SEIFI IRSAD scores in the 15–64 year old population, we can use this information to explore different aspects of socio-economic advantage and disadvantage and how they interact within the different states and territories.

SEIFI IRSD state and territory comparisons

Table 5.7 shows the frequency table of SEIFI IRSD group by state/territory 15–64 year old population, with the percentage figures representing the proportion of the state/territory 15–64 year old population in each group. The cells highlighted in light grey are the highest proportions for each group, whilst the cells highlighted in dark grey are the lowest proportions for each group.

5.7 Frequency table of SEIFI IRSD group by state or territory, with state / territory 15–64 year old population percentage distributions

SEIFI IRSD group	*NSW*	*Vic.*	*Qld*	*SA*	*WA*	*Tas.*	*NT*	*ACT*	*OT*	*Aust.*
					Frequency ('000)					
1	826.6	595.8	511.7	229.9	230.6	84.7	39.3	28.6	0.7	2,547.9
2	800.8	640.2	542.8	215.5	266.8	75.2	20.3	26.0	0.2	2,588.0
3	1,321.2	1,037.4	803.7	298.7	413.9	85.7	32.3	74.5	0.3	4,067.6
4	1,362.9	1,016.1	728.1	249.3	399.9	63.7	34.1	98.3	0.4	3,952.7
Total	4,311.5	3,289.5	2,586.3	993.3	1,311.2	309.3	126.1	227.4	1.6	13,156.2
					Percentage					
1	19.2	18.1	19.8	23.1	17.6	27.4	31.2	12.6	42.8	
2	18.6	19.5	21.0	21.7	20.4	24.3	16.1	11.4	15.4	
3	30.6	31.5	31.1	30.1	31.6	27.7	25.6	32.8	20.0	
4	31.6	30.9	28.2	25.1	30.5	20.6	27.1	43.2	21.8	

The Australian Capital Territory has the lowest proportion of its 15–64 year old population in each of the most disadvantaged SEIFI IRSD groups (groups 1 and 2), and subsequently the highest proportion in each of the least disadvantaged SEIFI IRSD groups (groups 3 and 4). The proportion of the ACT 15–64 year old population in group 4 is higher than any other state, indicating the prevalence of least relative disadvantage in this territory. This is consistent with SEIFA results, which point to a high proportion of the least disadvantaged areas being in the ACT.

Tasmania has the highest proportion of its 15–64 year old population in group 2 of any state or territory, and the lowest proportion in group 4, reflecting a general level of more relative disadvantage than the other states and territories. Similarly, the Northern Territory has the highest proportion of its 15–64 year old population in group 1 and the lowest proportion in group 3.

This information however does not shed light on the distribution of socio-economic advantage and disadvantage within the different SEIFA CD areas in each state or territory, instead giving broad stroke results at the state/territory level. Section 5.3 refines the analysis to look at diversity of socio-economic advantage and disadvantage at the CD level.

SEIFI IRSAD state and territory comparisons

Table 5.8 shows the proportion of the state 15–64 year old population that falls into each SEIFI IRSAD group. The percentage figures highlighted in light grey again represent the highest 15–64 year old population percentage for each group, whilst the lowest 15–64 year old population percentage for each group is highlighted in dark grey.

5.8 Frequency table of SEIFI IRSAD group by state or territory, with state / territory 15–64 year old population percentage distributions

SEIFI IRSAD Group	*NSW*	*Vic.*	*Qld*	*SA*	*WA*	*Tas.*	*NT*	*ACT*	*OT*	*Aust.*
					Frequency ('000)					
1	401.0	287.4	252.2	122.8	114.3	48.2	24.3	11.0	0.4	1,261.5
2	396.6	314.3	278.4	129.2	129.6	44.6	12.4	11.7	0.2	1,316.9
3	482.6	386.0	314.7	118.7	145.0	43.8	10.4	13.7	0.1	1,514.9
4	349.5	263.7	234.5	106.0	123.2	31.6	11.9	15.0	0.2	1,135.6
5	472.3	396.5	314.3	103.2	151.2	28.1	9.8	23.4	0.1	1,498.8
6	337.8	257.8	203.1	70.9	105.9	22.7	8.7	16.0	0.1	1,023.1
7	527.0	379.9	316.1	107.1	171.9	29.0	19.7	28.5	0.2	1,579.3
8	409.7	297.4	235.8	75.9	124.7	19.2	9.8	25.8	0.1	1,198.3
9	439.6	337.2	213.3	78.3	117.6	20.4	9.6	35.0	0.1	1,251.0
10	495.5	369.5	224.0	81.3	127.8	21.6	9.5	47.4	0.1	1,376.7
Total	4,311.5	3,289.5	2,586.3	993.3	1,311.2	309.3	126.1	227.4	1.6	13,156.2
					Percentage					
1	9.3	8.7	9.8	12.4	8.7	15.6	19.3	4.8	24.6	
2	9.2	9.6	10.8	13.0	9.9	14.4	9.8	5.1	15.2	
3	11.2	11.7	12.2	12.0	11.1	14.2	8.2	6.0	8.4	
4	8.1	8.0	9.1	10.7	9.4	10.2	9.4	6.6	12.5	
5	11.0	12.1	12.2	10.4	11.5	9.1	7.8	10.3	5.2	
6	7.8	7.8	7.9	7.1	8.1	7.4	6.9	7.1	5.5	
7	12.2	11.6	12.2	10.8	13.1	9.4	15.7	12.5	13.7	
8	9.5	9.0	9.1	7.6	9.5	6.2	7.8	11.3	5.1	
9	10.2	10.3	8.3	7.9	9.0	6.6	7.6	15.4	6.4	
10	11.5	11.2	8.7	8.2	9.8	7.0	7.6	20.8	3.4	

The Australian Capital Territory has the highest proportion of its 15–64 year old population in the most advantaged groups (8, 9 and 10), and the lowest proportion in the most disadvantaged groups (1, 2, 3 and 4). These figures reflect the results observed for 2006 SEIFA IRSAD, where the Australian Capital Territory was observed to have significant proportions of its total territory 15–64 year old population residing in relatively advantaged areas.

The difference in table 5.8 between the Australian Capital Territory and the remaining states and territories is quite noticeable. For example, the ACT has 20.8% of its 15–64 year old population in group 10, whilst the next highest proportion is New South Wales with 11.5%; this is less than half of the 15–64 year old population proportion observed for the ACT. Similarly, for the lower groups, the ACT has 4.8% of its 15–64 year old population in group 1, compared to the next lowest proportion of 8.7% in Western Australia. This indicates that most of the ACT 15-64 year old population are relatively advantaged, especially in comparison to Tasmania which has the lowest proportion of its 15–64 year old population in the highest three groups (19.8%), compared to 47.6% in the ACT.

The Northern Territory has almost 20% of its 15–64 year old population in group 1, the largest proportion of any of the states or territories, reflecting the high level of relative disadvantage observed in this territory. On the other hand, New South Wales has a very even population distribution throughout the groups, with most groups having very close to 10% of the New South Wales 15–64 year old population in them. This is also the case with Western Australia, and to a slightly lesser extent, Queensland and Victoria.

These results are similar to those found in the 15–64 year old population proportion distributions through the SEIFI IRSD groups (table 5.7). The ACT in that table had a high proportion of its 15–64 year old population in the least disadvantaged group, and a low proportion in the most disadvantaged group. Tasmania, on the other hand, had the reverse: a high proportion in the most disadvantaged group and a low proportion in the least disadvantaged group. These similarities between the two indexes indicate a degree of robustness to the findings.

SEIFI IRSD and SEIFI IRSAD comparison

To further investigate the link between the two indexes and the impact of the addition of advantaging variables on the distribution of people within the group classifications, a frequency table of classifications for each person of both their SEIFI IRSD and SEIFI IRSAD groups was created. Broadly speaking, we would expect persons to be classified into similar socio-economic groups based on each index. Table 5.9 contains the frequency table for this comparison.

5.9 Frequency table of SEIFI IRSD group against SEIFI IRSAD group

SEIFI IRSAD group	*SEIFI IRSD Group*				
	1	*2*	*3*	*4*	*Total*
1	1,261,475	0	0	0	1,261,475
2	715,046	601,896	0	0	1,316,942
3	304,175	1,210,720	0	0	1,514,895
4	73,870	189,934	871,792	0	1,135,596
5	133,036	457,397	908,399	0	1,498,832
6	6,656	37,123	979,319	0	1,023,098
7	33,111	45,579	500,237	1,000,403	1,579,330
8	7,440	22,981	455,079	712,831	1,198,331
9	12,268	16,824	294,254	927,655	1,251,001
10	799	5,516	58,543	1,311,843	1,376,701
Total	2,547,876	2,587,970	4,067,623	3,952,732	13,156,201

There are many interesting conclusions that can be drawn from table 5.9. Firstly, considering just those persons who were in SEIFI IRSAD groups 1–3, it can be seen that the most disadvantaged persons based on the SEIFI IRSAD fall into the two most disadvantaged SEIFI IRSD groups. This result seemingly points to an ability in both indexes to identify and classify the most relatively disadvantaged persons similarly when compared to the remaining 15–64 year old population. On the flip side, however, it can be seen that approximately one-tenth of persons falling into SEIFI IRSD group 1 lie in SEIFI IRSAD groups 4 – 10, or approximately 2.4% in SEIFI IRSAD groups 6–10. This may only represent a fraction of the total 15–64 year old population but it can clearly be seen that persons identified as the most relatively disadvantaged by the SEIFI IRSD could potentially be identified as the most relatively advantaged by the SEIFI IRSAD. However, these figures are very small and overall the largest proportions lie across the diagonal of the table.

SEIFI IRSD group 2 and group 3 have progressively lower proportions of their 15–64 year old population in the lower SEIFI IRSAD groups; group 2 has no persons classified in SEIFI IRSAD group 1 and group 3 has no persons classified in SEIFI IRSAD groups 1 to 3. The least disadvantaged 30% of the 15–64 year old population (SEIFI IRSD group 4) has been split into SEIFI IRSAD groups 7 to 10. This demonstrates the benefits of including advantaging variables; it allows for those who were previously not differentiable to be classified according to their relative advantage and disadvantage (recalling from table 5.5 that the least disadvantaged 30% of the 15–64 year old population received the same SEIFI IRSD index score).

5.3 Potential for misclassification when using SEIFA as a proxy for individual-level advantage and disadvantage

There are two interesting investigations that can be performed on the SEIFI IRSD groups and SEIFI IRSAD groups that involve examining the characteristics of individuals compared to the SEIFA deciles of the area they reside in. SEIFA scores are CD level measures summarising average relative socio-economic advantage and disadvantage, and when interpreted correctly provide a great deal of information to the user. If the SEIFA scores however are used as a proxy for individual level socio-economic advantage and disadvantage, then there is a risk of misclassification – the area level score is not reflective of an individual's score. It is important to note that SEIFA can still be used for individual level analysis provided the interpretation is correct; namely, the index score represents the average socio-economic characteristics of the area in which a person lives.

The analysis in this section thus compares the individual level SEIFI IRSD and SEIFI IRSAD groups to the area level SEIFA IRSD and SEIFA IRSAD deciles. For example, consider a person with a SEIFI IRSD score in group 2 who lives in an area with a SEIFA IRSD score in decile 9. The interpretation of this comparison would indicate that the average socio-economic characteristics of the area in which this persons resides indicate a relatively low level of disadvantage (decile 9); however, the socio-economic disadvantage of the individual is relatively high, because their score places them in the second most disadvantaged group, group 2.

The analysis in this section provides insight into the extent to which relatively disadvantaged persons reside in areas with different levels of relative disadvantage. Table 5.10 contains a comparison between the individual and area level measures: the percentages represent the proportion of the 15–64 year old population in each SEIFA IRSD CD decile split by SEIFI IRSD group. Careful attention must be paid to the interpretation of this table because of the clash in conceptual grounding between the area and individual level indexes involved.

In table 5.10 we are most interested in the spread across the rows and down the columns: if the frequency of persons is high across the spectrum then this indicates that many individuals live in areas with SEIFA scores not necessarily representative of their individual socio-economic disadvantage. Whilst there are many interesting pieces of information illustrated in table 5.10, some of the most important conclusions to draw relate to the rate at which persons live in areas that do not necessarily reflect their individual socio-economic disadvantage.

5.10 Frequency table of SEIFI IRSD group against SEIFA IRSD CD decile, with decile-based 15–64 year old population percentages

SEIFI IRSD group	*SEIFA IRSD decile*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Total*
	Frequency ('000)										
1	508.8	378.9	327.0	287.9	255.3	221.1	195.6	163.9	131.1	78.3	2,547.9
2	213.9	270.5	283.9	283.2	286.4	282.8	275.6	256.9	243.4	191.3	2,588.0
3	220.8	311.7	351.4	381.7	405.0	428.1	455.5	474.0	514.9	524.7	4,067.6
4	178.5	240.5	271.7	314.7	344.5	377.8	429.3	487.6	586.4	721.7	3,952.7
Total	1,122.0	1,201.7	1,234.0	1,267.5	1,291.2	1,309.8	1,356.0	1,382.3	1,475.8	1,515.9	13,156.2
	Percentage										
1	45.4	31.5	26.5	22.7	19.8	16.9	14.4	11.9	8.9	5.2	
2	19.1	22.5	23.0	22.3	22.2	21.6	20.3	18.6	16.5	12.6	
3	19.7	25.9	28.5	30.1	31.4	32.7	33.6	34.3	34.9	34.6	
4	15.9	20.0	22.0	24.8	26.7	28.9	31.7	35.3	39.7	47.6	

For instance, approximately 35% of persons who live in an area classified as the most disadvantaged (SEIFA IRSD decile 1) are actually in SEIFI IRSD group 3 or 4. However, recall that 60% of the 15-64 year old population are in SEIFI IRSD groups 3 and 4, and thus represent the least disadvantaged individuals based on SEIFI IRSD. Hence, these persons in SEIFI IRSD group 3 or 4 are relatively less disadvantaged than 40% of the 15–64 year old population for the individual analysis (from table 5.5), and yet they live in areas classified as the most disadvantaged. Similarly, approximately 45% of persons living in areas in SEIFA IRSD decile 2 lie in SEIFI IRSD group 3 or 4. These results show that disadvantaged areas have significant proportions of 15-64 year olds with low levels of relative socio-economic disadvantage. *Vice versa*, approximately 18% of persons living in the least disadvantaged areas (SEIFA IRSD decile 10) are classified as the most relatively disadvantaged according to SEIFI IRSD because they lie in SEIFI IRSD group 1 or 2.

So, whilst the majority of the 15–64 year old population in each SEIFA IRSD decile has a SEIFI IRSD group reflecting the level of socio-economic disadvantage of the area in which they reside, there are large proportions in each decile that have dissimilar individual characteristics.

To investigate this issue further, the area based SEIFA IRSAD decile is compared to the individual level SEIFI IRSAD group in table 5.11. The percentages represent the proportion of the 15–64 year old population in each SEIFA IRSAD decile split by SEIFI IRSAD group.

5.11 Frequency table of SEIFI IRSAD group against SEIFA IRSAD CD decile, with decile-based 15–64 year old population percentages

SEIFI IRSAD group	*SEIFA IRSAD decile*										*Total*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	
	Frequency ('000)										
1	297.3	206.0	169.4	143.0	121.9	104.9	86.2	66.6	45.3	21.0	1,261.5
2	161.9	173.9	168.0	159.2	149.1	140.6	126.3	108.3	82.6	46.8	1,316.9
3	164.7	185.7	180.7	174.3	169.4	165.1	156.7	139.8	110.4	68.0	1,514.9
4	74.0	106.0	117.4	125.6	130.1	135.6	134.6	126.2	108.6	77.5	1,135.6
5	102.4	131.0	135.7	140.8	148.7	159.6	171.8	178.8	178.3	151.7	1,498.8
6	48.5	75.3	86.4	95.2	105.9	118.4	128.1	133.7	126.1	105.4	1,023.1
7	103.1	123.9	131.2	137.2	146.4	159.3	173.5	186.3	202.9	215.4	1,579.3
8	42.2	68.5	81.5	94.3	108.3	127.9	147.9	167.1	179.1	181.7	1,198.3
9	31.7	54.6	67.9	80.9	97.7	119.1	148.5	181.8	217.2	251.7	1,251.0
10	21.3	39.4	52.6	66.3	84.5	108.6	148.5	200.2	271.7	383.7	1,376.7
Total	1,047.1	1,164.3	1,190.8	1,216.7	1,262.0	1,339.2	1,422.0	1,488.7	1,522.3	1,503.1	13,156.2
	Percentage										
1	28.4	17.7	14.2	11.8	9.7	7.8	6.1	4.5	3.0	1.4	
2	15.5	14.9	14.1	13.1	11.8	10.5	8.9	7.3	5.4	3.1	
3	15.7	16.0	15.2	14.3	13.4	12.3	11.0	9.4	7.3	4.5	
4	7.1	9.1	9.9	10.3	10.3	10.1	9.5	8.5	7.1	5.2	
5	9.8	11.3	11.4	11.6	11.8	11.9	12.1	12.0	11.7	10.1	
6	4.6	6.5	7.3	7.8	8.4	8.8	9.0	9.0	8.3	7.0	
7	9.9	10.6	11.0	11.3	11.6	11.9	12.2	12.5	13.3	14.3	
8	4.0	5.9	6.8	7.8	8.6	9.6	10.4	11.2	11.8	12.1	
9	3.0	4.7	5.7	6.7	7.7	8.9	10.4	12.2	14.3	16.8	
10	2.0	3.4	4.4	5.5	6.7	8.1	10.4	13.5	17.9	25.5	

If no differences between the individual and area level measures exist, then the diagonal (from top-left to bottom-right of the table) should have the highest frequency percentages because they represent similar levels of relative socio-economic advantage and disadvantage in the area and individual level index deciles and groups.

This table shows that a large proportion of the 15–64 year old population live in areas that have different average characteristics to those observed at the individual level. The largest percentage of persons whose individual and area level measures align is at the lowest and highest deciles/groups respectively with 28.4% and 25.5%. This is not surprising as generally the extreme deciles and groups capture the least sensitive populations to different measures of socio-economic advantage and disadvantage, as has been observed in previous SEIFA diagnostic studies detailed in Radisich and Wise (2011). These results were also expected because of the high probability of a person in the lowest 10% of the individual level socio-economic spectrum residing in the most disadvantaged areas, and vice versa for the most advantaged 10% of persons living in the most advantaged areas.

Conversely, the table also reveals a significant degree of spread across the groups. For instance, 23.6% of the 15–64 year old population of the most disadvantaged areas lie in the five most advantaged SEIFI IRSAD groups (6–10). Similarly, 24.3% of the 15–64 year old population of the most advantaged areas lie in the five most disadvantaged SEIFI IRSAD groups (1–5).

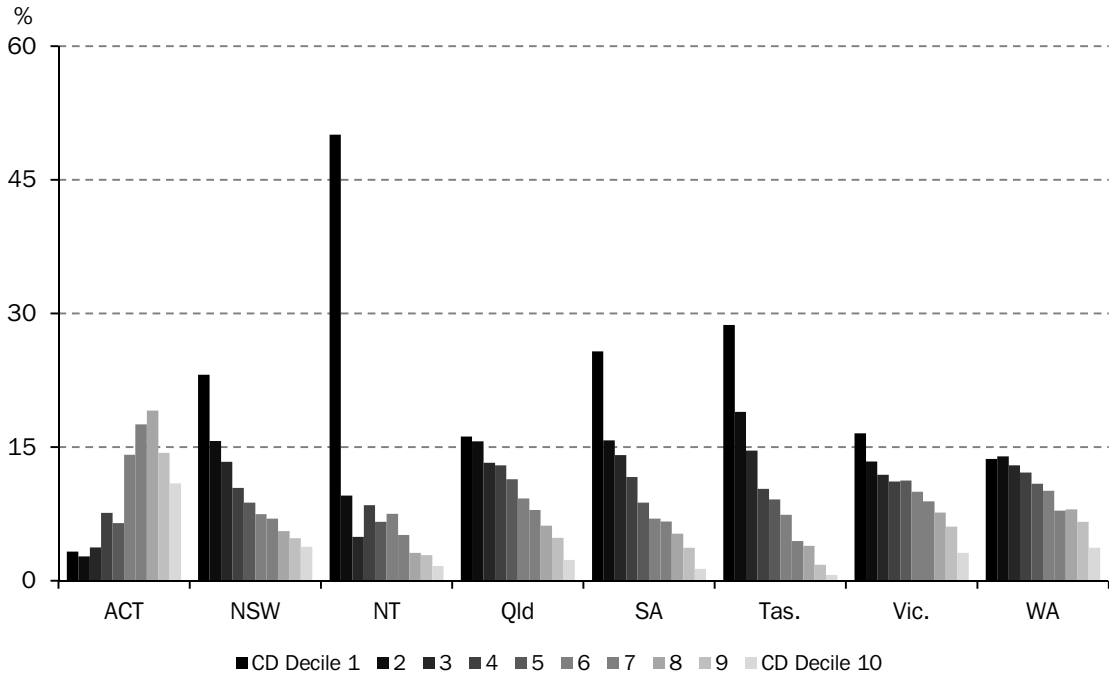
These observations are similar to those seen in the comparison between SEIFI IRSD groups and SEIFA IRSD deciles, especially for what can be seen in the lower deciles. For example, table 5.10 shows that approximately 35% of the 15–64 year old persons who live in SEIFA IRSD decile 1 (considered the most disadvantaged areas) received SEIFI IRSD scores in group 3 or 4, which are the least disadvantaged groups. The corresponding figure is approximately 33% for SEIFI IRSAD groups 5–10, because SEIFI IRSD groups 3 and 4 account for approximately 60% of the 15–64 year old population (which roughly equates to SEIFI IRSAD groups 5–10).

Focussing on the most disadvantaged SEIFI IRSD group and SEIFI IRSAD group

This section focusses on only the most disadvantaged persons in the 15–64 year old population; that is, those persons in SEIFI IRSD group 1 or SEIFI IRSAD group 1. Graph 5.12 plots individuals in SEIFI IRSD group 1 (the most disadvantaged group) by their corresponding SEIFA IRSD CD decile, further split by state or territory. That is, this graph shows the most disadvantaged persons from the individual level index analysis and the type of areas they reside in. The percentage figures that form the y-axis represent the proportion of the most disadvantaged SEIFI IRSD group 1 state or territory 15–64 year old population that resides in an area classified into the different SEIFA IRSD deciles. For example, 50% of the Northern Territory 15–64 year old population in SEIFI IRSD group 1 resides in an area classified by SEIFA IRSD as decile 1. Note that the 15–64 year old population proportions plotted sum to 100%.

There are three notable features in graph 5.12. The first notable feature is the spike for the Northern Territory at decile 1. This suggests that the most disadvantaged persons in the Northern Territory reside mainly in the most disadvantaged areas: 60% of group 1 persons in the Northern Territory reside in a CD area with a score in SEIFA IRSD decile 1 or 2.

5.12 Percentage of individuals from SEIFI IRSD group 1 residing in areas classified by SEIFA IRSD CD decile, by state and territory

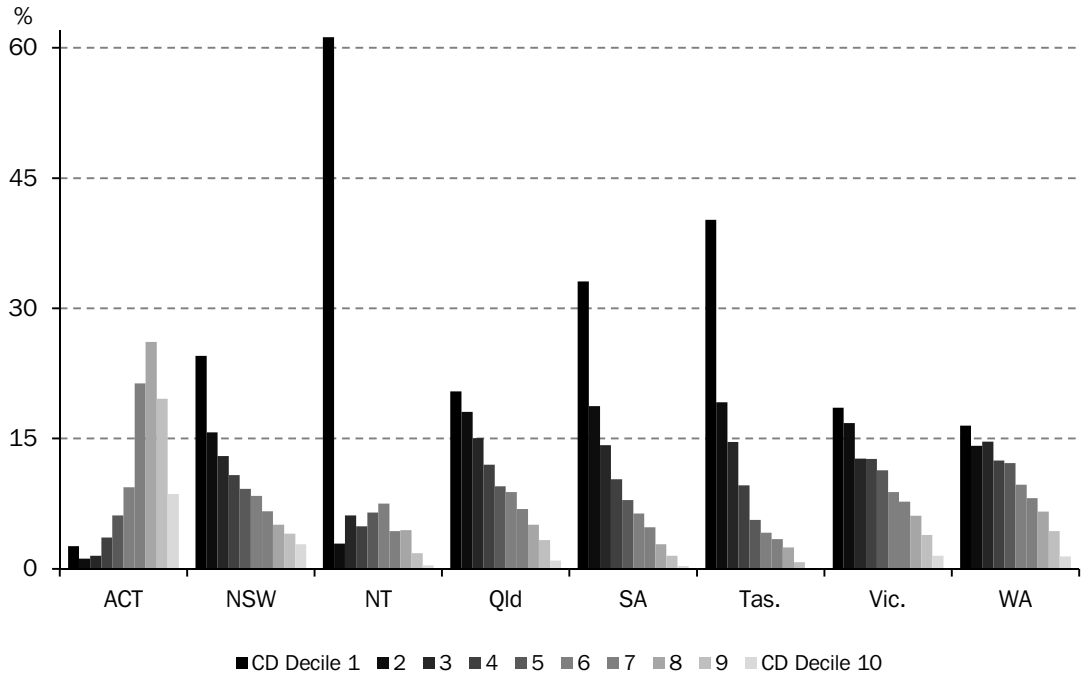


The second feature is that the Australian Capital Territory appears to have a larger proportion of its most disadvantaged persons residing in areas with low levels of relative disadvantage, as evidenced by the hump on the right hand side of the graph. In other words, the most disadvantaged persons in the ACT proportionally reside in the least disadvantaged areas – well over half of the ACT group 1 15–64 year old population resides in an area classified with a SEIFA IRSD decile greater than 6. This result points to a high degree of diversity within the ACT CD areas.

The third feature is the consistent overall trend amongst the remaining states and territories evident in the graph. There is a close alignment as the SEIFA IRSD decile increases, highlighting the uniqueness of the ACT and Northern Territory features. The consistency amongst the results for the remaining states also highlights the extent to which individual level socio-economic advantage and disadvantage is diverse in areas. The downward linear trend in each state and territory except for the ACT reflects decreasing prevalence of the most relatively disadvantaged individuals residing in less disadvantaged areas.

The corresponding graph for the SEIFI IRSAD group 1 15–64 year old population is presented as graph 5.13. This graph shows the distribution of people in group 1 of the SEIFI IRSAD score distribution across the states and territories, split by the different SEIFA IRSAD deciles. In other words, the most disadvantaged 10% of the individual level 15–64 year old population and the type of areas in which they live. Note again that the 15–64 year old population proportions plotted sum to 100%.

5.13 Percentage of individuals from SEIFI IRSAD group 1 residing in areas classified by SEIFA IRSAD CD decile, by state and territory



The overall pattern of trends presented in graph 5.13 (SEIFI IRSAD) is very similar to graph 5.12 (SEIFI IRSD). There is a clear hump seen for the Australian Capital Territory across the highest SEIFA IRSAD deciles; over 54% of the ACT 15–64 year old population in group 1 lives in areas classified in SEIFA IRSAD deciles 8–10. This distinct similarity shows that the pattern of disadvantaged persons in the ACT is the same regardless of the inclusion of advantaging variables. The other states and territories have the same pattern as graph 5.12 as well, including the Northern Territory, which has a large spike in SEIFA IRSAD decile 1, indicating over 60% of its relatively most disadvantaged persons live in the most disadvantaged areas – a similar result was reached using the SEIFI IRSD results. The consistency amongst the patterns for the remaining states again indicates that the issue of individual level diversity is important when interpreting SEIFA index scores.

A further investigation of the misclassification rate, this time however considering the 15–64 year old population of the least disadvantaged group for SEIFI IRSD and most advantaged group for SEIFI IRSAD, is detailed in Appendix E. The results from this analysis highlight that the ACT, New South Wales and Western Australia have a high incidence of the most relatively advantaged persons residing in more advantaged areas, as classified by SEIFA, compared to the other states and territories.

5.4 Identifying areas with diverse patterns of individual-level advantage and disadvantage

A further investigation aims to identify diversity within CD areas using the two individual level indexes' groups to highlight CDs with higher than average proportions of most disadvantaged and most advantaged persons; that is, CDs with a socio-economically diverse 15–64 year old population. Referring back to table 5.5 for the frequencies of persons in each SEIFI IRSD group, a CD has been defined to be 'diverse' if: the proportion of its 15–64 year old population in group 1 was greater than 19.37% *and* the proportion of its 15–64 year old population in group 4 was greater than 30.04%. These frequencies were used as proxies for the average rate of group-based disadvantage in CDs across Australia. This diversity measure thus highlights CDs with a greater than average proportion of both the most disadvantaged and least disadvantaged persons; that is, there are many people living at both ends of the spectrum in the same area. The measure can be aggregated to the state/territory level to determine the proportion of the CD areas in each state or territory that are diverse, according to the definition above.

Table 5.14 contains the summary of the state/territory based results.

5.14 Percentage of 'diverse' CDs categorised by state, based on SEIFI IRSD group

		'Diverse' CDs	
*State**	*Total number of CDs+*	*Number*	*Percentage*
New South Wales	11,811	720	6.1
Victoria	9,095	432	4.8
Queensland	7,458	385	5.2
South Australia	3,178	128	4.0
Western Australia	3,980	234	5.9
Tasmania	1,045	24	2.3
Northern Territory	356	53	14.9
Aust. Capital Territory	522	59	11.3

* Other Territories have been omitted from this analysis due to a small number of CDs.
\+ Total number of CDs that received a SEIFA score in 2006.

Table 5.14 is very effective at highlighting the extent of underlying individual level diversity within areas for the different states and territories. Broadly, table 5.14 highlights that no more than 15% of the areas in any state or territory are diverse, with the average proportion of 6.8% reflecting a general concentration of diversity within a small subset of the total number of areas, as indicated by the measure used in this section.

From table 5.14, a higher proportion of the CDs in the Northern Territory (14.9%) and the ACT (11.3%) are classified as diverse, compared to the remaining states. The rate of diverse areas is nearly twice as much as the next highest state or territory (New South Wales at 6.1%), reinforcing the notion that the ACT and the Northern Territory have a high rate of incidence of CDs with vastly different socio-economic characteristics within their usual resident 15–64 year old populations.

On the flip side, Tasmania emerges as having a noticeably lower percentage of diverse CDs (2.3%), indicating that Tasmanian 15–64 year old populations captured in CDs are more socio-economically homogeneous than what is observed in the other states and territories.

For SEIFI IRSAD, the definition of diversity is slightly different to that employed in the SEIFI IRSD diversity analysis, due to the differences in the index concepts and grouping analysis involved. However, the premise is still the same: in this case, the SEIFI IRSAD groups are used instead of the SEIFI IRSD groups. A CD is thus classified as diverse if it has a greater than average proportion (9.6% from table 5.6) of its included 15–64 year old population in the most disadvantaged group (group 1) *and* a greater than average proportion (10.5%) of its included 15–64 year old population in the most advantaged group (group 10). The results are shown in table 5.15 below.

Table 5.15 highlights lower proportions of diverse areas for the states and territories, with the average proportion of 3.5%, based on SEIFI IRSAD groups, compared to the results in table 5.14 based on SEIFI IRSD groups. From table 5.15, the least diverse state or territory is Queensland, with only 1.7% of its included CDs being classified as diverse. The most diverse states or territories are Northern Territory (4.2%) and ACT (7.9%), and these two territories are clearly more diverse than the remaining states: this supports the results in table 5.14, highlighting an underlying trend of higher incidence of diversity at the area level in these territories. However, in this case, the ACT is noticeably more diverse than the Northern Territory and any other state.

The results presented in table 5.15 show smaller proportions of 'diversity' compared to the SEIFI IRSD based results in table 5.14, however this can easily be accounted for in the differing definitions of diversity employed in each analysis. The definition used to classify a CD as diverse in SEIFI IRSD was based on groups (which are approximately 20–30% of the 15–64 year old population), compared to basing the diversity definition on the SEIFI IRSAD groups (which represent close to 10% of the 15–64 year old population).

5.15 Percentage of 'diverse' CDs categorised by state or territory, based on SEIFI IRSAD group

State*	Total number of CDs+	'Diverse' CDs	
		Number	Percentage
New South Wales	11,811	269	2.3
Victoria	9,095	218	2.4
Queensland	7,458	130	1.7
South Australia	3,178	103	3.2
Western Australia	3,980	115	2.9
Tasmania	1,045	35	3.4
Northern Territory	356	15	4.2
Aust. Capital Territory	522	41	7.9

* Other Territories have been omitted from this analysis due to a small number of CDs.
\+ Total number of CDs that received a SEIFA score in 2006.

As much as the SEIFI IRSD and SEIFI IRSAD results are similar, there are differences. For example, Tasmania is the only state or territory to see an increase in its diversity proportion moving from the SEIFI IRSD to the SEIFI IRSAD results. Additionally, the Australian Capital Territory is seen to have noticeably more diversity than the Northern Territory using SEIFI IRSAD groups. However, using SEIFI IRSD, the Northern Territory is more diverse than the ACT.

It is important to clarify here that the findings presented in tables 5.14 and 5.15 are consistent with what was presented in graphs 5.12 and 5.13. The two sets of figures represent different measures of diversity. Using the Northern Territory as an example, graphs 5.12 and 5.13 indicate this territory has a higher proportion of its most disadvantaged 15–64 year old population residing in the most disadvantaged areas (as identified by the 2006 SEIFA IRSD and IRSAD indexes), compared to the remaining states and territories. Tables 5.14 and 5.15 then go on to reveal that the Northern Territory has a high proportion of CDs with greater than average proportions of their 15–64 year old population in the most disadvantaged *and* least disadvantaged (using SEIFI IRSD) or most advantaged (using SEIFI IRSAD) groups. This can occur when CDs encapsulate both high levels of disadvantage and advantage (or lack of disadvantage).

Sensitivity analysis was performed on the choice of cut-off for the definition of diverse in both individual level indexes. The investigation revealed that the ranking of the states and territories largely remained the same, with the ACT and Northern Territory still clearly the states or territories with the greatest proportion of 'diverse' areas.

5.5 An alternative measure of disadvantage – the number of indicators of disadvantage per person

It is of interest to examine the number of indicators of disadvantage as a proxy for individual level socio-economic indexes. In the context of the research into individual level indexes in this paper, it is important to consider the validity of alternative measures, especially one that is accessible and easily understood. In the index creation process, there were eight indicators of disadvantage used in the SEIFI IRSD PCA. These indicator variables were NOQUAL, INC_LOW, INDIGENOUS, NOYEAR12, NONET, RENT_SOCIAL, LOWRENT and NOCAR (refer to table 2.1 for the definitions of these variables). If the simple method of summing the number of indicators of disadvantage can produce comparable output to individual level indexes, then it is worth considering the potential of this approach for appraising individual level socio-economic advantage and disadvantage.

Table 5.16 contains the frequency information for the number of indicators of disadvantage a person can potentially have.

5.16 Frequency table of indicators of disadvantage

Number of disadvantage indicators	*15–64 year-old population*	
	Frequency	*Percentage*
0	3,952,732	30.04
1	4,101,459	31.18
2	3,050,035	23.18
3	1,343,012	10.21
4	447,531	3.40
5	149,408	1.14
6	75,043	0.57
7	29,536	0.22
8	7,445	0.06

Note: For four or more indicators of disadvantage,
sum = 708,963 and % of 15–64 year old population = 5.39.

It is possible to perform many exploratory analyses using the number of indicators of disadvantage as the primary classification, however for space considerations only one has been included in this report because the others generally provide no further information not already highlighted by the SEIFI IRSD score and group analyses.

5.17 Percentage of 'Diverse' CDs, categorised by state or territory, based on number of disadvantage indicators

State*	Total number of CDs+	'Diverse' CDs	
		Number	Percentage
New South Wales	11,811	664	5.6
Victoria	9,095	384	4.2
Queensland	7,458	364	4.9
South Australia	3,178	128	4.0
Western Australia	3,980	261	6.6
Tasmania	1,045	30	2.9
Northern Territory	356	56	15.7
Aust. Capital Territory	522	98	18.8

* Other Territories have been omitted from this analysis due to a small number of CDs.
\+ Total number of CDs that received a SEIFA score in 2006.

Table 5.17 contains the results from an analysis using the indicators of disadvantage to recreate the diversity measure discussed at the end of Section 5.3. In this case, a CD is diverse if it has more than 30.04% of its included 15–64 year old population with no indicators of disadvantage (representing the least disadvantaged persons) and more than 5.39% of its included 15–64 year old population with four or more indicators of disadvantage (representing the most disadvantaged persons). The definition of a diverse CD now makes reference to the frequencies of number of indicators of disadvantage in table 5.16.

Table 5.17 shows the average proportion of diverse CDs, 8.8%, is similar to the results in table 5.14 based on SEIFI IRSD groups. As was also observed in table 5.14, the results from table 5.17 highlight that the Northern Territory and the ACT have higher proportions of diverse CDs compared to the remaining states. The sharp discrepancy between the proportions for these two territories and the other states highlights the extent to which socio-economic disadvantage can be heterogeneous within an area. Tasmania again has the lowest proportion of diverse CDs.

The percentage of diverse CDs in each state and territory is very similar comparing the results from table 5.17 to 5.14 (based on SEIFI IRSD), with the only large difference in percentage point terms observed for the ACT. This indicates that the number of indicators of disadvantage approach can provide a comparable level of insight into diversity within areas when considered against more complex individual level indexes.

6. CONCLUDING REMARKS

This paper has explored the extent of individual level diversity for the 15–64 year old population within area level SEIFA indexes through the creation of individual level indexes of socio-economic advantage and disadvantage. The methodology used to create the indexes was established in Baker and Adhikari (2007).

The individual level indexes revealed varying degrees of clumping: the index of disadvantage was significantly negatively skewed, with substantial clumping on the highest scores (corresponding to persons with least relative disadvantage), whilst the index of advantage and disadvantage exhibited a more even spread of scores with some clumping around the midpoint of the distribution. The addition of advantaging variables allowed for greater differentiation to be achieved across the whole socio-economic spectrum, when comparing the two index distributions. However, the clumping observed in both individual level index score distributions considerably reduced their capacity to rank individuals. The two SEIFA score distributions were compared to highlight their comparative advantages and desirability.

The individual level indexes were used to explore the diversity of socio-economic advantage and disadvantage within CDs, and how the rates of diversity differed between the states and territories of Australia. An underlying level of individual level diversity was observed within CDs in each state and territory, and was highlighted as similar by different measures for most states and territories. The Australian Capital Territory and Northern Territory however were consistently highlighted as having a greater proportion of areas with a high incidence of diversity. Moreover, the ACT had a high proportion of the most relatively disadvantaged persons residing in areas classified by SEIFA as being less disadvantaged, whilst New South Wales and Western Australia had high proportions of the most relatively advantaged persons residing in more advantaged areas, when compared with the remaining states and territories. These observations highlight the care that needs to be taken when using SEIFA information to draw conclusions about individuals who reside in those areas.

Analysing the two individual level indexes of socio-economic advantage and disadvantage created in this paper has facilitated the appraisal of diversity within areas, something that has until now been not possible except in Baker and Adhikari (2007). The results presented in this paper are important illustrations of how diversity of advantage and disadvantage within an area can exist, and the extent to which individuals with differing levels of socio-economic advantage and disadvantage reside in the same area.

However, this paper has made clear some of the shortcomings of individual level indexes, and why SEIFA remains an important, robust product. Firstly, concerns with substantial population exclusions limit the applicability of the analysis; approximately one-third of the population counted in the 2006 Census were excluded from the individual level index construction process for applicability, compared to 0.6% of the population excluded for 2006 SEIFA. This vast difference reflects the robustness of the SEIFA indexes, namely that it maximises the proportion of the population that receives an index score. SEIFA is also more theoretically and conceptually sound because it is based on variables chosen for applicability in an area-based index, it is externally validated, and the aggregate nature of the data and stringent exclusion rules both work to ensure that there is sufficient meaningful data in an area to support index construction.

Future directions

Given the analysis presented in this paper, users of SEIFA will understandably be wondering if a product can be released that enables them to appropriately tackle the issues involved. However, before attempting this, the following critical issues would have to be resolved:

- consensus on the definition of individual level advantage and disadvantage, the best set of variables to measure it, a means for validating individual level indexes and an appropriate method for setting weights (if at all) – this would need to be considered for different age brackets.
- how such a product could be integrated with the existing SEIFA product to ensure that it is a useful addition, rather than creating confusion amongst users – this is critical to the chance of such a product being released.

These issues were already identified to some extent by Baker and Adhikari (2007).

There is still a need to address the issue of individual level diversity within areas once the new ASGS geography standard is introduced, as discussed in Section 3.1. This is why it is recommended that the analysis presented in this paper be repeated after the release of SEIFA 2011. In the meantime, we recommend using the SEIFA indexes for socio-economic analysis, bearing in mind the caveats relating to these measures not being attributable to individuals, but only to the average relative socio-economic advantage and disadvantage in an area.

ACKNOWLEDGEMENTS

The authors would like to thank Jeffrey Wright and Stephen Collett for their helpful input with this research project, and Stevie Broadfoot for developing the diversity measure. We would also like to thank the following members of ABS staff who provided comments and feedback: Gemma Van Halderen, Phillip Gould, Peter Radisich, Peter Rossiter, Kate Bond, Andrew Webster, Caroline Daley and Shiji Zhao. 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.

This project was completed using funding from the Annual Statistics Consultancy Fund (ASCF). The ASCF is funding from the ABS ACT Regional Office resources that is used for projects initiated by the ACT Government for work that assists them in taking their part in the National Statistical Service (NSS). While the ACT Government proposes projects and the Chief Minister's Department is involved in the decision making process, the final decision on the project rests with the ABS.

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.

Ainley, J. and Long, M. (1995) "Measuring Student Socio-economic Status" in Ainley, J., Graetz, B., Long, M. and Batten, M., *Socioeconomic Status and School Education*, Australian Government Publishing Service, Canberra, pp. 52–76.

Australian Bureau of Statistics (2006) *Statistical Geography: Volume 2 – Census Geographic Areas, Australia*, cat. no. 2905.0, ABS, Canberra.

—— (2007) *The Review of the Australian Standard Geographical Classification*, cat. no. 1216.0.55.001, 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.

—— (2010) *Australian Standard Geographical Classification (ASGC), July 2010*, cat. no. 1216.0, 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. 2911.0.55.002, ABS, Canberra.

—— (2011c) *Information Paper: Measures of Socioeconomic Status*, cat. no. 1244.0.55.001, ABS, Canberra.

Bailey, N.; Flint, J.; Goodlad, R.; Shucksmith, M.; Fitzpatrick, S. and Pryce, G. (2003) *Measuring Deprivation in Scotland: Developing a Long-Term Strategy,* Scottish Executive Central Statistics Unit, (last viewed on 11 August 2011) <http://www.scotland.gov.uk/publications/2003/09/18197/26538>

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.

Gordon, D.; Adelman, L.; Ashworth, K.; Bradshaw, J.; Levitas, R.; Middleton, S.; Pantazis, C.; Patsios, D.; Payne, S.; Townsend, P. and Williams, J. (2000) *Poverty and Social Exclusion in Britain: Report of the Poverty and Social Exclusion Survey of Britain*, Joseph Rowntree Foundation, York.

Gordon, D. and Pantazis, C. (eds.) (1997) *Breadline Britain in the 1990s*, Aldershot: Ashgate.

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

Kennedy, B. and Firman, D. (2004) *Indigenous SEIFA – Revealing the Ecological Fallacy*, Paper presented to the 12th Biennial Conference of the Australian Population Association, 15–17 September 2004, Canberra.

Lim, P. and Gemici, S. (2011) *Measuring the Socioeconomic Status of Australian Youth*, National Centre for Vocational Education Research, Adelaide.

Mack, J. and Lansley, S. (1985) *Poor Britain*, Allen and Unwin, London.

Marks, G.N.; McMillan, J.; Jones, F.L. and Ainley, J. (2000) "The Measurement of Socioeconomic Status for Reporting of Nationally Comparable Outcomes of Schooling", Draft Report for the National Education Performance Monitoring Taskforce, ACER, Melbourne and Research School of the Social Sciences, Sociology Program, ANU, Canberra.

Morris, R. and Carstairs, V. (1991) "Which Deprivation? A Comparison of Selected Deprivation Indexes", *Journal of Public Health Medicine*, 13, pp. 318–326.

Olsson, U. (1979) "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient", *Psychometrika*, 44 (4), pp. 443–460.

Radisich, P. and Wise, P. (2011 – to appear) "Socio-Economic Indexes for Areas (SEIFA): Recent Developments and Plans for 2011", *Methodology Research Papers*, Australian Bureau of Statistics, Canberra.

Rigdon, E.E. and Ferguson, C.E., Jr. (1991) "The Performance of the Polychoric Correlation Coefficient and Selected Fitting Functions in Confirmatory Factor Analysis with Ordinal Data", *Journal of Marketing Research*, 28(4), pp. 491–497.

Salmond, C.; King, P.; Crampton, P. and Waldegrave, C. (2006) "NZiDep: A New Zealand Index of Socioeconomic Deprivation for Individuals", *Journal of Social Science and Medicine*, 62, pp. 1474–1485.

Townsend, P. (1979) *Poverty in the United Kingdom, A Survey of Household Resources and Standards of Living*, London.

Townsend, P. (1987) "Deprivation", *Journal of Social Policy*, 16, pp. 125–146.

Vinson, T. (2007) *Dropping Off the Edge: the Distribution of Disadvantage in Australia*, Jesuit Social Services, Melbourne.

APPENDIXES

A. VARIABLES CONSIDERED FOR INCLUSION IN SEIFI IRSAD, WITH PREVALENCES

A.1 List of variables considered for the individual level index of advantage and disadvantage, with prevalence (%)

Individual-level variable	*Code*	*(%)*
Persons aged 15 years and over with no post-school qualifications	noqual	45.97
Persons aged 15 years and over who left school after year 10 or lower	noyear12	49.32
Person has stated annual household equivalised income between $13,000 and $20,799	inc_low	13.01
Person is employed in the sector classified as low skill clerical and administrative workers	occ_admin_l	11.38
Person is employed in the sector classified as labourers	occ_labour	10.62
Person is employed in the sector classified as low skill sales workers	occ_sales_l	7.55
Person is employed in the sector classified as machinery operators and drivers	occ_drivers	6.76
Person is employed in the sector classified as low skill community and personal service workers	occ_service_l	6.76
Person in the labour force is unemployed	unemployed	5.29
Person does not speak English well	englishpoor	2.32
Person under the age of 70 has a long-term health condition or disability and needs assistance with core activities	disabilityU70	2.47
Person in in a one parent family with dependent offspring only	oneparent	10.30
Person resides in an occupied private dwelling with no internet connection	nonet	29.90
Person resides in an occupied private dwelling with no car	nocar	7.00
Person resides in a household renting from Government or community organisations	rent_social	4.46
Person resides in an occupied private dwelling paying less than $120 rent per week (but not $0)	low_rent	13.25
Person resides in an occupied private dwelling with a broadband connection	broadband	40.55
Person resides in a household owning the dwelling they occupy (without a mortgage)	owning	34.02
Person resides in an occupied private dwelling paying greater than $290 per week	highrent	17.94
Person resides in an occupied private dwelling with three or more cars	highcar	15.53
Person resides in an occupied private dwelling with four or more bedrooms	highbed	28.69
Person resides in a household paying mortgage greater than $2 120 per month	highmortgage	6.14
Person is employed in the sector classified as professionals	occ_prof	20.20
Person is employed in the sector classified as managers	occ_manager	13.45
Person aged 15 years and over is at university or other tertiary institution	atuni	5.10
Person aged 15 years and over has an advanced diploma or diploma qualification	diploma	8.04
Person aged 15 years and over has a degree or higher qualification	degree	17.65
Person has stated annual household equivalised income greater than $52 000 (approx. 9th and 10th deciles)	inc_high	23.77

B. CORRELATION ANALYSIS FOR SEIFI IRSD AND SEIFI IRSAD

Table B.1 presents the results from the tetrachoric correlation analysis of the individual level variables identified in the 2006 SEIFA IRSD scoping list as contributing to the defined notion of socio-economic disadvantage. From the calculations of the correlations, it can be seen that there are numerous pairs with correlations greater than $|0.8|$. These are listed in table B.2 following table B.1.

B.1 Tetrachoric correlation matrix for individual level SEIFI IRSD index

	oneparent	*nonet*	*rent_social*	*lowrent*	*nocar*	*fewbed*	*occ_drivers*
oneparent	1	.	.	.	.	.	.
nonet	0.14	1	.	.	.	.	.
rent_social	0.35	0.44	1	.	.	.	.
lowrent	0.31	0.47	0.90	1	.	.	.
nocar	0.24	0.44	0.56	0.57	1	.	.
fewbed	–0.18	0.32	0.28	0.46	0.53	1	.
occ_drivers	–0.13	0.13	–0.07	–0.08	–0.12	–0.02	1
occ_labour	–0.01	0.18	0.07	0.06	0.03	0.02	–0.92
occ_sales_l	0.07	–0.07	–0.10	–0.11	–0.07	–0.07	–0.95
occ_admin_l	–0.02	–0.07	–0.20	–0.20	–0.11	–0.03	–0.92
occ_service_l	0.08	0.01	–0.02	–0.04	–0.01	0.01	–0.97
indigenous	0.25	0.36	0.61	0.54	0.41	0.05	–0.03
unemployed	0.16	0.15	0.23	0.25	0.26	0.12	–0.99
sep_divorced	0.46	0.20	0.18	0.22	0.15	0.16	0.05
disabilityU70	–0.02	0.21	0.35	0.29	0.24	0.13	–0.24
noschool	0.04	0.23	0.27	0.18	0.22	0.06	0.01
noyear12	0.15	0.28	0.26	0.23	0.06	–0.05	0.25
noqual	0.06	0.29	0.26	0.23	0.16	0.01	0.28
englishpoor	0.06	0.18	0.16	0.12	0.23	0.04	0.01
inc_low	0.32	0.29	0.31	0.29	0.25	0.10	–0.05

	occ_labour	*occ_sales_l*	*occ_admin_l*	*occ_service_l*	*indigenous*	*unemployed*	*sep_divorced*
occ_labour	1	.	.	.	.	.	.
occ_sales_l	–0.99	1	.	.	.	.	.
occ_admin_l	–0.99	–0.99	1	.	.	.	.
occ_service_l	–0.92	–0.95	–0.92	1	.	.	.
indigenous	0.10	–0.11	–0.10	0.03	1	.	.
unemployed	–0.93	–0.98	–0.93	–0.99	0.18	1	.
sep_divorced	0.01	–0.10	0.02	0.04	–0.01	0.05	1
disabilityU70	–0.07	–0.26	–0.25	–0.25	0.13	–0.08	0.11
noschool	0.08	–0.19	–0.30	–0.17	0.14	0.04	0.04
noyear12	0.24	0.04	–0.03	–0.01	0.25	0.10	0.14
noqual	0.29	0.18	0.16	0.04	0.18	0.15	0.05
englishpoor	0.12	–0.19	–0.32	–0.16	–0.08	0.11	0.01
inc_low	0.07	0.01	–0.16	0.01	0.19	0.23	0.11

	disabilityU70	*noschool*	*noyear12*	*noqual*	*englishpoor*	*inc_low*
disabilityU70	1	.	.	.	.	.
noschool	0.45	1	.	.	.	.
noyear12	0.25	0.99	1	.	.	.
noqual	0.23	0.46	0.40	1	.	.
englishpoor	0.24	0.66	0.10	0.31	1	.
inc_low	0.29	0.15	0.23	0.23	0.20	1

As table B.2 clearly demonstrates, all of the occupation variables (OCC_ADMIN_L, OCC_SERVICE_L, OCC_SALES_L, OCC_LABOUR, OCC_DRIVERS) and the UNEMPLOYED variable were correlated above the prescribed cut off of |0.8|. These variables are highlighted by a grey background. These strong negative correlations are unsurprising because, for example, a person who is employed in one of the industries captured by the variables included for this index cannot by definition be unemployed. Discretion was used to henceforth decide that each of these variables measured different aspects of disadvantage (that is, they represent different types of employment and unemployment), and so all of these variables were retained at this stage in the index construction process.

B.2 List of highly correlated (greater than |0.8|) variables

Variable 1 code	*Variable 2 code*	*Correlation*
rent_social	lowrent	0.8962
noschool	noyear12	0.9990
occ_drivers	occ_labour	–0.9231
occ_drivers	occ_sales_l	–0.9493
occ_drivers	occ_admin_l	–0.9210
occ_drivers	occ_service_l	–0.9676
occ_drivers	unemployed	–0.9990
occ_labour	occ_sales_l	–0.9990
occ_labour	occ_admin_l	–0.9990
occ_labour	occ_service_l	–0.9231
occ_labour	unemployed	–0.9334
occ_sales_l	occ_admin_l	–0.9990
occ_sales_l	occ_service_l	–0.9493
occ_sales_l	unemployed	–0.9823
occ_admin_l	occ_service_l	–0.9210
occ_admin_l	unemployed	–0.9290
occ_service_l	unemployed	–0.9990

The variables NOSCHOOL (% of people who did not go to school) and NOYEAR12 (% of people who left school before year 12) were also found to have a high correlation (0.999): this clearly indicates that the number of people who did not go to school is

well captured by the number of people who did not complete year 12. Therefore the NOSCHOOL variable was dropped, since the prevalence of this variable in the 15–64 year old population (0.66%) was much lower than the prevalence of the NOYEAR12 variable (49.32%) (see table 2.1 for further details).

The variables LOWRENT (% Households paying rent less than $120 per week) and RENT_SOCIAL (% Households renting from Government or Community organisations) were also highly correlated (0.8962). Upon consideration, it was decided that these variables measure two different aspects of socio-economic disadvantage, and therefore both variables were kept for further analysis.

Table B.3 following now presents the results from the tetrachoric correlation analysis of the individual level variables identified in the 2006 SEIFA IRSAD scoping list as contributing to the notion of socio-economic advantage and disadvantage. The pairs with correlations greater than |0.8| are listed in table B.4 following table B.3.

B.3 Tetrachoric correlation matrix for individual level SEIFI IRSAD index

	oneparent	*nonet*	*broadband*	*rent_social*	*owning*	*lowrent*	*highrent*
oneparent	1	.	.	.	.	.	.
nonet	0.14	1	.	.	.	.	.
broadband	–0.05	–0.99	1	.	.	.	.
rent_social	0.35	0.44	–0.27	1	.	.	.
owning	–0.18	0.08	0.03	–0.99	1	.	.
lowrent	0.31	0.47	–0.32	0.90	–0.99	1	.
highrent	0.01	–0.09	0.22	–0.13	–0.98	–0.99	1
nocar	0.24	0.44	–0.24	0.56	–0.24	0.57	0.22
highcar	–0.28	–0.21	0.24	–0.32	0.25	–0.35	–0.11
highbed	–0.08	–0.26	0.34	–0.25	0.18	–0.33	–0.06
highmortgage	–0.18	–0.25	0.31	–0.93	–0.97	–0.97	–0.99
occ_drivers	–0.13	0.13	–0.07	–0.07	–0.03	–0.08	–0.11
occ_labour	–0.01	0.18	–0.09	0.07	–0.02	0.06	–0.09
occ_sales_l	0.07	–0.07	0.10	–0.10	–0.01	–0.11	0.01
occ_prof	–0.12	–0.26	0.23	–0.32	–0.02	–0.27	0.15
occ_manager	–0.18	–0.16	0.13	–0.34	0.03	–0.27	0.08
occ_service_l	0.08	0.01	0.03	–0.02	–0.03	–0.04	0.05
unemployed	0.16	0.15	–0.05	0.23	–0.06	0.25	0.04
disabilityu70	–0.02	0.21	–0.17	0.35	0.09	0.29	–0.13
noyear12	0.15	0.28	–0.12	0.26	0.13	0.23	–0.22
atuni	0.04	–0.29	0.24	–0.16	–0.07	–0.11	0.24
noqual	0.06	0.29	–0.10	0.26	0.09	0.23	–0.07
diploma	–0.04	–0.15	0.13	–0.17	0.02	–0.14	0.06
degree	–0.16	–0.29	0.25	–0.33	–0.01	–0.25	0.22
inc_low	0.32	0.29	–0.16	0.31	0.10	0.29	–0.10
inc_high	–0.40	–0.26	0.33	–0.47	0.02	–0.40	0.21

	nocar	*highcar*	*highbed*	*highmortgage*	*occ_drivers*	*occ_labour*
nocar	1	.	.	.	.	.
highcar	–0.99	1	.	.	.	.
highbed	–0.38	0.45	1	.	.	.
highmortgage	–0.33	0.07	0.29	1	.	.
occ_drivers	–0.16	0.07	–0.02	–0.06	1	.
occ_labour	0.03	0.07	–0.02	–0.11	–0.92	1
occ_sales_l	–0.07	0.16	0.10	–0.02	–0.95	–0.99
occ_prof	–0.15	–0.05	0.05	0.21	–0.99	–0.99
occ_manager	–0.24	0.09	0.11	0.19	–0.99	–0.99
occ_service_l	–0.01	0.07	0.02	–0.05	–0.97	–0.92
unemployed	0.26	–0.08	–0.07	–0.13	–0.99	–0.93
disabilityu70	0.24	–0.15	–0.10	–0.22	–0.24	–0.07
noyear12	0.06	0.07	0.04	–0.16	0.25	0.24
atuni	0.15	0.09	0.04	0.02	–0.25	–0.09
noqual	0.16	0.08	0.01	–0.14	0.28	0.29
diploma	–0.08	–0.01	0.05	0.09	–0.19	–0.19
degree	–0.04	–0.11	0.03	0.23	–0.38	–0.38
inc_low	0.25	–0.21	–0.11	–0.29	–0.05	0.07
inc_high	–0.27	0.15	0.14	0.40	–0.07	–0.20

	occ_sales_l	*occ_prof*	*occ_manager*	*occ_service_l*	*unemployed*	*disabilityu70*	*noyear12*
occ_sales_l	1	.	.	.	.	.	.
occ_prof	–0.99	1	.	.	.	.	.
occ_manager	–0.99	–0.97	1	.	.	.	.
occ_service_l	–0.95	–0.99	–0.99	1	.	.	.
unemployed	–0.98	–0.99	–0.92	–0.99	1	.	.
disabilityu70	–0.26	–0.36	–0.33	–0.25	–0.08	1	.
noyear12	0.04	–0.52	–0.09	–0.01	0.10	0.25	1
atuni	0.26	0.10	–0.14	0.19	0.08	–0.26	–0.57
noqual	0.18	–0.54	–0.06	0.04	0.15	0.23	0.40
diploma	–0.10	0.19	0.12	0.11	–0.06	–0.11	–0.27
degree	–0.26	0.77	0.18	–0.19	–0.13	–0.26	–0.72
inc_low	0.01	–0.35	–0.21	0.01	0.23	0.29	0.23
inc_high	–0.10	0.46	0.30	–0.09	–0.29	–0.34	–0.26

	atuni	*noqual*	*diploma*	*degree*	*inc_low*	*inc_high*
atuni	1	.	.	.	.	.
noqual	0.13	1	.	.	.	.
diploma	0.04	–0.98	1	.	.	.
degree	0.20	–0.99	–0.97	1	.	.
inc_low	–0.02	0.23	–0.09	–0.27	1	.
inc_high	0.03	–0.24	0.15	0.46	–0.99	1

B.4 List of highly correlated (greater than |0.8|) variables

Variable 1	*Variable 2*	*Correlation*
nonet	broadband	–0.9964
rent_social	owning	–0.9990
rent_social	lowrent	0.8962
rent_social	highmortgage	–0.9314
owning	lowrent	–0.9990
owning	highrent	–0.9770
owning	highmortgage	–0.9692
lowrent	highrent	–0.9990
lowrent	highmortgage	–0.9681
nocar	highcar	–0.9913
diploma	degree	–0.9708
noqual	diploma	–0.9803
noqual	degree	–0.9990
inc_low	inc_high	–0.9990
occ_drivers	occ_labour	–0.9231
occ_drivers	occ_sales_l	–0.9493
occ_drivers	occ_prof	–0.9990
occ_drivers	occ_manager	–0.9990
occ_drivers	occ_service_l	–0.9676
occ_drivers	unemployed	–0.9990
occ_labour	occ_sales_l	–0.9990
occ_labour	occ_prof	–0.9920
occ_labour	occ_manager	–0.9990
occ_labour	occ_service_l	–0.9231
occ_labour	unemployed	–0.9334
occ_sales_l	occ_prof	–0.9990
occ_sales_l	occ_manager	–0.9990
occ_sales_l	occ_service_l	–0.9493
occ_sales_l	unemployed	–0.9823
occ_prof	occ_manager	–0.9739
occ_prof	occ_service_l	–0.9990
occ_prof	unemployed	–0.9990
occ_manager	occ_service_l	–0.9990
occ_manager	unemployed	–0.9232
occ_service_l	unemployed	–0.9990

Similarly to the SEIFI IRSD variable correlations, all of the occupation variables (OCC_DRIVERS, OCC_LABOUR, OCC_SALES_L, OCC_PROF, OCC_MANAGER, OCC_SERVICE_L) and the UNEMPLOYED variable were correlated above the prescribed cut off of |0.8|. Because all of these variables were considered to measure different aspects of disadvantage (different types of employment, unemployment), all were retained at this stage of the index construction process.

Of the remaining variables, each pair was deemed to measure a different aspect of socio-economic disadvantage, so no variables were dropped for the subsequent PCA analysis. On a case-by-case basis:

- INC_LOW and INC_HIGH are highly negatively correlated because a person cannot have equivalised household income that is simultaneously in the bottom 2 deciles (low) and the top two deciles (high). These variables measure different aspects of advantage and disadvantage and so are retained. Other variable pairs that are similarly diametrically opposed include NOCAR and HIGHCAR, LOWRENT and HIGHRENT and NONET and BROADBAND; all these variables were retained.
- The education variables DIPLOMA, DEGREE and NOQUAL are all highly negatively correlated because persons with such indicators typically hold different levels of education. Each variable is capturing a different aspect of socio-economic advantage or disadvantage (a person with no qualifications, a university degree or a diploma), and so they are all retained.
- The household rental/ownership variables OWNING, LOWRENT, HIGHRENT, HIGHMORTGAGE and RENT_SOCIAL are all highly negatively correlated (except RENT_SOCIAL and LOWRENT) because they all measure inter-related, but different, aspects of a person's living arrangements. Each variable is capturing a different aspect of socio-economic advantage or disadvantage, and so they are all retained.

This results in no variable exclusions from the scoping list for the creation of a SEIFI IRSAD index based on correlation analysis.

C. VARIABLE LOADINGS FOR SEIFI IRSD AND SEIFI IRSAD

Table C.1 contains the summary of the variables that were dropped from the SEIFI IRSD index construction due to low loadings (less than |0.3|), and the order in which they were dropped (from first at the top of the table to last at the bottom).

C.1 List of variables dropped from SEIFI IRSD index construction due to low loadings

Individual-level variable	*Loading*	*Description*
occ_service_l	0.007	% Employed people classified as Low Skill Community and Personal Service Workers
occ_sales_l	0.032	% Employed people classified as Low Skill Sales Workers
occ_admin_l	0.085	% Employed people classified as Low Skill Clerical and Administrative Workers
occ_drivers	–0.026	% Employed people classified as Machinery Operators and Drivers
occ_labour	–0.118	% Employed people classified as labourers
englishpoor	–0.132	% People who do not speak English well
unemployed	–0.188	% People who are (in the labour force) unemployed
disabilityu70	–0.223	% People aged under 70 who have a long-term health condition or disability and need assistance with core activities
sep_divorced	–0.225	% People aged 15 years and over who are separated or divorced
oneparent	–0.251	% One parent families with dependent offspring only
fewbed	–0.271	% Occupied private dwellings with one or no bedrooms

In line with the procedure adopted for the construction of the 2006 SEIFA IRSD (see the *2006 Technical Paper* for more details), the variables OCC_SERVICE_LN, OCC_SALES_LN and OCC_ADMIN_LN were dropped from the index first because they had positive loadings, and thus represent variables of advantage rather than disadvantage. This works to ensure the index of disadvantage captures just those variables that indicate socio-economic disadvantage. It is interesting that the occupation variables in general had clearly the lowest loadings overall.

Table C.2 contains the summary of the variables, in order, that were dropped from the SEIFI IRSAD index construction due to low loadings.

The occupation variables again are generally the lowest loading on the first principal component. This suggests across the two individual level indexes that occupation did not describe a significant amount of the variation in the dataset. The NOCAR variable had a loading very close to |0.3|, however strictly adhering to the prescribed cut-off was deemed the appropriate course of action to take and so this variable was dropped for the SEIFI IRSAD index construction.

C.2 List of variables dropped from SEIFI IRSAD index construction due to low loadings

Individual-level variable	*Loading*	*Description*
owning	0.008	% Households owning dwelling they occupy (without a mortgage)
occ_sales_l	–0.029	% Employed people classified as Low Skill Sales Workers
occ_service_l	–0.031	% Employed people classified as Low Skill Community and Personal Service Workers
occ_drivers	–0.115	% Employed people classified as Machinery operators and Drivers
highrent	0.129	% Households paying rent greater than $290 per week
atuni	0.134	% people aged 15 years and over at university or other tertiary institution
diploma	0.163	% People aged 15 years and over with an advanced diploma or diploma qualification
occ_manager	0.167	% Employed people classified as Managers
unemployed	–0.164	% People (in the labour force) unemployed
highcar	0.175	% Occupied dwellings with three or more cars
oneparent	–0.194	% One parent families with dependent offspring only
disabilityu70	–0.202	% People aged under 70 who have a long-term health condition or disability and need assistance with core activities
occ_labour	–0.212	% Employed people classified as Labourers
highbed	0.231	% Occupied private dwellings with four or more bedrooms
highmortgage	0.276	% Households paying a mortgage greater than $2,120 per month
nocar	–0.297	% occupied private dwellings with no car

D. SEIFI IRSD AND SEIFI IRSAD UNIQUE SCORES

An examination of the number of unique SEIFI IRSD scores revealed a total of 255 scores (or 2^8–1 from eight variable indicators), ranging from 388 to 1094. To begin with, consider the top five most prevalent scores in the SEIFI IRSD score distribution, as contained in table D.1.

D.1 List of highly prevalent SEIFI IRSD scores

	15–64 year old population		
SEIFI IRSD score	*Frequency*	*Percentage*	*Indicator/s of Disadvantage*
881	702,541	5.34%	nonet, noqual, noyear12
971	1,714,253	13.03%	noqual, noyear12
1032	1,437,973	10.93%	noqual
1034	1,803,715	13.71%	noyear12
1094	3,952,123	30.04%	–

The five scores presented in table D.1 represent over 70% of the total included 15–64 year old population. Further exploration of the data revealed that approximately 85% of the 15–64 year old population has a SEIFI IRSD score over 900 (in other words between 900 and 1094), meaning that the remaining 15% has a range of scores between 388 and 900. This vast discrepancy highlights the highly skewed score distribution and confirms that the individual level index does not discriminate particularly well across the least disadvantaged persons in the 15–64 year old population, making it difficult to differentiate between different levels of disadvantage.

Because some of the distribution clumps at the least disadvantaged end of the spectrum are larger than 10% in terms of the total 15–64 year old population proportion, forming deciles is not possible. The highest score in the distribution is shared by 30.04% of the 15–64 year old population, so an alternative is required to form the population into segments to aid further analyses and comparability – this led to the groupings established in Section 5.2.

The SEIFI IRSAD distribution, on the other hand, has 432 unique scores, ranging from 744 to 1234. Theoretically, a greater number of unique scores work to ensure that there is a greater scope for differentiation between each person based on their relative advantage or disadvantage. Of these unique scores, there are smaller percentages of people with the same score compared to SEIFI IRSD. To reinforce this, the five most prevalent scores are shown in table D.2.

D.2 List of highly prevalent SEIFI IRSAD scores

	15–64 year old population	
SEIFI IRSAD score	*Frequency*	*Percentage*
854	691,394	5.26%
949	763,970	5.81%
1002	644,490	4.90%
1003	739,666	5.62%
1012	1,022,209	7.77%

It can be seen that fewer than 30% of the total included 15–64 year old population have one of the above five scores; the most prevalent score alone in the SEIFI IRSD distribution was held by 30.04% of the 15–64 year old population. These five scores are located in the mid-range of the distribution. This compares favourably with the corresponding figure for the SEIFI IRSD score distribution, which resulted in over 73% of the included 15–64 year old population having one of the top five most prevalent scores.

A comparison between the score distributions for SEIFI IRSD and SEIFI IRSAD reveals the two distributions to be markedly different when comparing the number of unique scores and the range of scores. SEIFI IRSD was observed to have a range of 707, which is larger than the SEIFI IRSAD range of 490. However, the larger range does not necessarily mean that there is greater comparability between the different included persons based on their relative disadvantage. The clumping at the least disadvantaged end of the SEIFI IRSD distribution, which indicates an inability to differentiate between the least disadvantaged persons in the included 15–64 year old population, illustrates this point. The larger range, rather, indicates a greater dispersion of socio-economic disadvantage in the SEIFI IRSD, since both distributions are standardised for presentation purposes to a mean of 1000 and a standard deviation of 100.

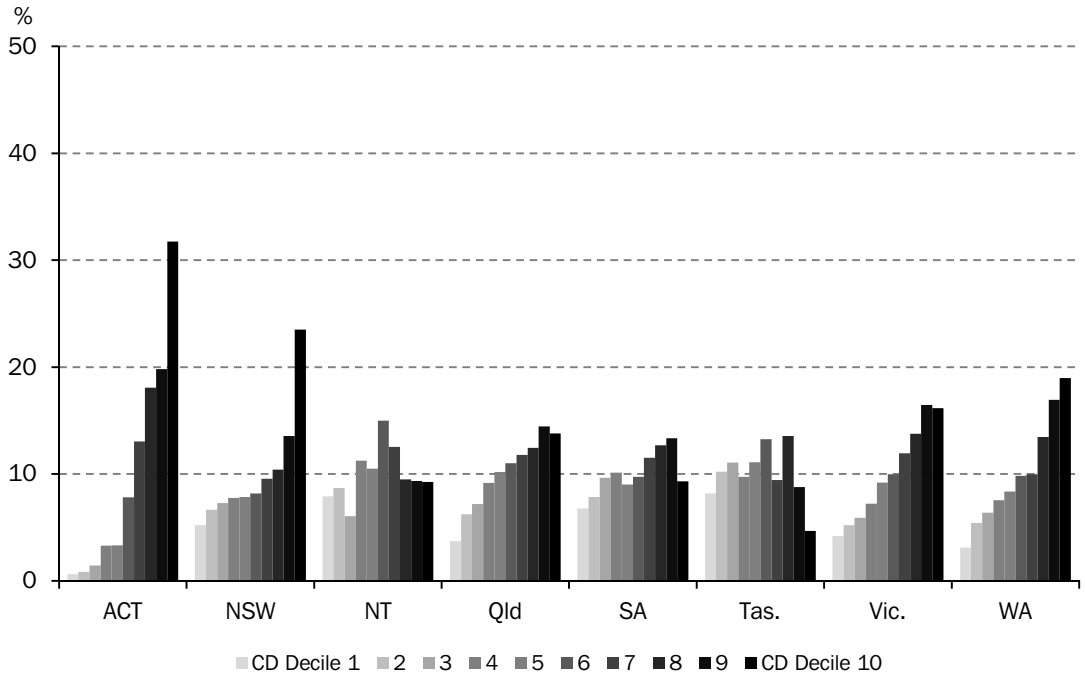
These differences show how the addition of advantaging variables noticeably increases the differentiation available on the socio-economic spectrum between included persons, especially in the most advantaged/least disadvantaged proportion of the 15–64 year old population.

E. POTENTIAL FOR MISCLASSIFICATION OF PERSONS IN THE LEAST DISADVANTAGED / MOST ADVANTAGED GROUPS

This section focusses on the least disadvantaged/most advantaged persons in the 15-64 year old population; that is, those persons in SEIFI IRSD group 4 or SEIFI IRSAD group 10. This information is provided to complement the analysis performed on the most disadvantaged persons identified in both indexes, as discussed in Section 5.3.

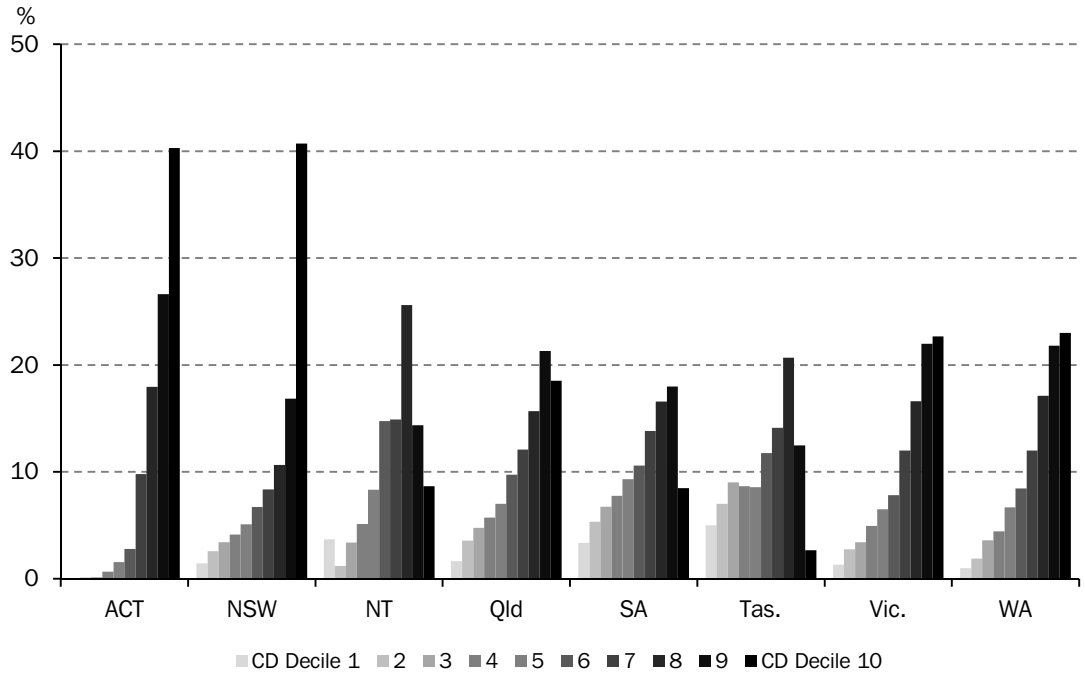
The first graph, graph E.1, plots individuals in SEIFI IRSD group 4 (least disadvantaged) by their corresponding SEIFA IRSD CD decile, further split by state or territory.

E.1 Percentage of individuals from SEIFI IRSD group 4 residing in areas classified by SEIFA IRSD CD decile, by state and territory



There are a few points of interest in the plot. There is an overall positive linear trend amongst the states and territories in graph E.1, reflecting an increasing prevalence of the least relatively disadvantaged individuals residing in less disadvantaged areas. We can see especially that the Australian Capital Territory has a different trend to the rest of the states, with a sharp increase where the other states and territories have a more gradual upward trend. New South Wales and Western Australia also have upward spikes at the least disadvantaged end of the SEIFA IRSD decile spectrum, but follow the same pattern as the other states at the remaining points. This seems to suggest that the ACT, New South Wales and Western Australia have the highest proportion of their least disadvantaged persons residing in the least disadvantaged areas. The ACT has the lowest percentage of its least disadvantaged 15–64 year old population (as classified by SEIFI IRSD group) living in the most disadvantaged SEIFA IRSD deciles.

E.2 Percentage of individuals from SEIFI IRSAD group 10 residing in areas classified by SEIFA IRSAD CD decile, by state and territory



The second graph, graph E.2, plots individuals in SEIFI IRSAD group 10 (most advantaged) by their corresponding SEIFA IRSAD CD decile, split by state or territory.

There is an overall positive linear trend amongst the states and territories in graph E.2, however Tasmania, the Northern Territory and South Australia dip sharply for deciles above SEIFA IRSAD decile 8. Similar to graph E.1, we can see that the ACT and New South Wales have sharp upward spikes at the most advantaged end of the spectrum, reflecting higher rates of the most advantaged persons residing in the most advantaged areas in this state and territory.

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