2033.0.55.001

Australian Bureau of Statistics

Technical Paper

Socio-Economic Indexes for Areas (SEIFA)

2011

New Issue

Technical Paper

Socio-Economic Indexes for Areas (SEIFA)

2011

Brian Pink

Australian Statistician

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 28 MAR 2013

ABS Catalogue no. 2033.0.55.001

Produced by the Australian Bureau of Statistics

INQUIRIES

For further information, please contact Dr Phillip Gould, Analytical Services Branch, on Canberra (02) 6252 5315 or email <analytical.services@abs.gov.au>.

CONTENTS

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ASGC	Australian Standard Geographical Classification
ASGS	Australian Statistical Geography Standard
CD	Collection District
CED	Commonwealth Electoral Division
Census	Australian Census of Population and Housing
IEO	Index of Education and Occupation
IER	Index of Economic Resources
IRSAD	Index of Relative Socio-economic Advantage and Disadvantage
IRSD	Index of Relative Socio-economic Disadvantage
LGA	Local Government Area
MB	Mesh Block
PCA	Principal Component Analysis
POA	Postal Area
SEIFA	Socio-Economic Indexes For Areas
SA1	Statistical Area Level 1
SA2	Statistical Area Level 2
SED	State Electoral Division
SLA	Statistical Local Area
SSC	State Suburb

SOCIO-ECONOMIC INDEXES FOR AREAS (SEIFA) – TECHNICAL PAPER

1. INTRODUCTION

1.1 What is SEIFA?

Socio-Economic Indexes for Areas (SEIFA) is a product developed by the ABS that ranks areas in Australia according to relative socio-economic advantage and disadvantage. The indexes are based on information from the five-yearly Census.

SEIFA 2011 is based on Census 2011 data, and consists of four indexes, each focussing on a different aspect of socio-economic advantage and disadvantage and being a summary of a different subset of Census variables.

Some common uses of SEIFA include:

- determining areas that require funding and services,
- identifying new business opportunities, and
- assisting research into the relationship between socio-economic disadvantage and various social outcomes.

The indexes and associated documentation are free of charge on the ABS website.

1.2 Purpose and outline of technical paper

This paper provides information on the concepts, data, and method used to create SEIFA 2011. A large part of this paper is also devoted to providing information on the correct interpretation and appropriate use of the indexes.

This paper can be viewed as a comprehensive reference for SEIFA 2011. Note that a basic user guide – *SEIFA Basics* – has also been prepared as part of this product release (ABS cat. no. 2033.0.55.001) and can be viewed in html format on the product web pages.

This technical paper can be read from start to finish, although a reader may wish to skip to sections of interest.

Section 2 discusses the notion of relative socio-economic advantage and disadvantage and outlines a measurement framework for SEIFA. With this framework in mind, Section 3 describes in detail the available Census variables and how they fit into the framework. Section 3 concludes by providing a final candidate variable list. Section 4 describes the application of the data analysis technique Principal Component Analysis (PCA) to the candidate variable list in order to construct indexes. This section contains much analytical output. Section 5 details the steps taken to validate the index scores. Section 6 provides guidance and advice on the use of SEIFA. Section 7

presents analysis of the relationship between SEIFA and three important classifying variables: age, states/territories, and remoteness.

For interested readers, a step-by-step description of the index construction process can be found in Section 4.3.

1.3 Some historical context

A relative measure of socio-economic disadvantage was first produced by the ABS following the 1971 Census. Socio-Economic Indexes for Areas (SEIFA), in its present form, was first produced from the 1986 Census and consisted of five indexes:

- Urban Index of Relative Socio-Economic Advantage,
- Rural Index of Relative Socio-Economic Advantage,
- Index of Relative Socio-Economic Disadvantage,
- Index of Economic Resources, and
- Index of Education and Occupation.

The same set of indexes was also created from the 1991 and 1996 Censuses.

In developing SEIFA 2001, the ABS undertook a review. The review examined:

- the variables used in SEIFA,
- the method used to calculate the indexes,
- the number and type of indexes released, and
- the validation process.

The review process included a literature search, looking at overseas and Australian indexes of disadvantage, and also involved extensive user input on a number of issues.

Following the review for SEIFA 2001, two of the indexes—Urban and Rural Indexes of Advantage—were replaced by a single Index of Relative Socio-Economic Advantage and Disadvantage, reducing the number of indexes to four. SEIFA 2006 consisted of the same four indexes.

The following section discusses features of SEIFA 2011.

1.4 Features of SEIFA 2011

This section highlights some important features of SEIFA 2011, and how they differ from SEIFA 2006.

SEIFA 2011 consists of the same four indexes as produced for SEIFA 2006 and 2001, each referring to the general population:

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

Since SEIFA is an established product, we have generally attempted to maintain consistency between SEIFA 2011 and the previous release. However, some changes have been made and are listed below.

New geography standard

- SEIFA 2011 is released according to the Australian Statistical Geography Standard (ASGS). This is a change from past versions of SEIFA, which used the Australian Standard Geographical Classification (ASGC). The main implication for SEIFA from this change is that the new base unit of analysis is the Statistical Area Level 1 (SA1), rather than the Census Collection District (CD) used in the past.
- Index scores for larger geographic areas have also been produced by taking population-weighted averages of constituent SA1 scores. For a list of geographic output levels, see Section 4.7.

Methodological

- The methods used are generally the same, however the exclusion rules have been updated to ensure a reliable index score is obtained for as many areas as possible. Exclusion rules determine which areas do not receive an index score because of low populations or poor quality data. Further details are in Section 4.2.

Conceptual framework

- For the purposes of SEIFA, the ABS continues to broadly define relative socio-economic advantage and disadvantage in terms of *people's access to material and social resources, and their ability to participate in society.*

- A review was conducted by the ABS to enhance its understanding of the many related concepts under the general umbrella terms of advantage and disadvantage, so that SEIFA can be presented in the appropriate context and proper advice can be given to users about what it is measuring. A full discussion on this topic is found in Section 2.1.

Variables underpinning the indexes

- Of particular note to users of past versions of SEIFA, the IRSD no longer contains the variable relating to the proportion of people identifying as Indigenous in an area.
- Although Census 2011 collected the same variables as Census 2006, some newly derived SEIFA variables have been considered (children in jobless families, unengaged youth), and a number of variables (related to household tenure, education and internet access) have had some definitional changes. Some variables were also updated in line with updated classification standards. Variables using cut-off values in their definitions, such as high and low income, were updated appropriately. Section 3 contains more information on these variable issues.

Output

- More information on the distribution of SA1 scores within larger areas has been included in the output spreadsheets to enable more informative and detailed analyses.
- Provision has been made for users with limited technical knowledge to generate thematic maps, by releasing KMZ files that can be opened in Google Earth®. Section 6.4 contains more details.
- A short introductory video presentation has also been released as part of the suite of outputs. It provides a basic overview of SEIFA.

1.5 The nature of the indexes

To set some context for the rest of this paper, it is worth briefly touching on some important characteristics of the indexes:

- The indexes are assigned to areas, not to individuals. They indicate the collective socio-economic characteristics of the people living in an area.
- As measures of socio-economic conditions, the indexes are best interpreted as ordinal measures that rank (order) areas. The index scores are based on an arbitrary numerical scale and do not represent a quantity of advantage or disadvantage. For ease of interpretation, we generally recommend using the index rankings and quantiles (e.g. deciles) for analysis, rather than using the index scores. Index scores are still provided in the output, and can be used by more technically adept users.
- Each index is constructed based on a weighted combination of selected variables. The indexes are dependent on the set of variables chosen for the analysis. A different set of underlying variables would result in a different index.
- The indexes are primarily designed to compare the relative socio-economic characteristics of areas at a given point in time. It can be very difficult to perform useful longitudinal or time series analysis, and it should not be attempted flippantly.

Elaboration on each of the above points can be found in Section 6.1.

2. CONCEPTUAL FRAMEWORK

2.1 The notion of relative socio-economic advantage and disadvantage

The IRSD ranks areas in terms of relative socio-economic disadvantage. The IRSAD ranks areas in terms of relative socio-economic advantage and disadvantage. The Index of Economic Resources (IER) and the Index of Education and Occupation (IEO) measure particular aspects of socio-economic advantage and disadvantage. It is therefore important to clarify what we mean by relative socio-economic advantage and disadvantage. It informs both the candidate list of variables to consider for inclusion in the indexes, and also the appropriate use of the indexes once they have been produced.

For SEIFA 2011, the notion of relative socio-economic advantage and disadvantage is the same as that used for SEIFA 2006. That is, the ABS broadly defines relative socio-economic advantage and disadvantage in terms of *people's access to material and social resources, and their ability to participate in society.*

The fact this is described as a 'notion' and is 'broadly defined' is recognition of the many concepts that are emerging in the literature to describe advantage and disadvantage. Popular conceptualisations of disadvantage include poverty, deprivation, and social exclusion. Concepts that also capture indicators of advantage include human capital, social capital, and socioeconomic position. A key thread through all the literature is the move towards multi-dimensional frameworks to capture a person's ability to participate in society in many aspects of life; e.g. economic, social, and political. In this respect, when interpreted broadly, the ABS definition in the paragraph above captures these aspects.

Regarding a multi-dimensional framework, the dimensions that are included in SEIFA are guided by international research, given the constraints of Census data. The Census does collect information on the key dimensions of income, education, employment, occupation, housing, and also some other miscellaneous indicators of advantage and disadvantage. These are the dimensions used for SEIFA to inform variable selection and are discussed further in Section 3.

Another point to note is that SEIFA measures relative advantage and disadvantage at an area level, not at an individual level. Area level and individual level disadvantage are separate though related concepts. Area level disadvantage depends on the socio-economic conditions of a community or neighbourhood as a whole. These are primarily the collective characteristics of the area's residents, but may also be characteristics of the area itself, such as a lack of public resources, transport infrastructure or high levels of pollution. However, it is important to remember that SEIFA is restricted to the information that is included in the Census.

The ABS definition of relative socio-economic advantage and disadvantage is defined for the purposes of SEIFA, and sits amongst many other conceptualisations of advantage and disadvantage, some of which have been listed above. The numerous conceptualisations and their relationships to each other can be quite confusing to the lay person. To successfully navigate this issue, the user is recommended to consider their research interest and what they require, and then in this light, consider the definition of each SEIFA index and the variables included in each index to determine the appropriate index to use. The fact that the ABS produces four indexes, each summarising a different subset of Census variables, is recognition that users may be interested in different aspects of socio-economic advantage and disadvantage. The next section provides more information on each of the four indexes included in SEIFA.

2.2 Defining the concept behind each of the four indexes

The previous section discussed the notion of advantage and disadvantage that underpins all four indexes. This section focusses the discussion and gives a description of the concept behind each of the four indexes. For a list of the variables included in each index, see Section 4.4.5.

2.2.1 The Index of Relative Socio-Economic Disadvantage

The IRSD summarises variables that indicate relative disadvantage. This index ranks areas on a continuum from most disadvantaged to least disadvantaged.

A low score on this index indicates a high proportion of relatively disadvantaged people in an area. We cannot conclude that an area with a very high score has a large proportion of relatively advantaged ('well off') people, as there are no variables in the index to indicate this. We can only conclude that such an area has a relatively low incidence of disadvantage.

2.2.2 The Index of Relative Socio-Economic Advantage and Disadvantage

The IRSAD summarises variables that indicate either relative advantage or disadvantage. This index ranks areas on a continuum from most disadvantaged to most advantaged.

An area with a high score on this index has a relatively high incidence of advantage and a relatively low incidence of disadvantage. Due to the differences in scope between this index and the IRSD, the scores of some areas can vary substantially between the two indexes. For example, consider a large area that has parts containing relatively disadvantaged people, and other parts containing relatively advantaged people. This area may have a low IRSD ranking, due to its pockets of disadvantage. However, its IRSAD ranking may be moderate, or even above average, because the pockets of advantage may offset the pockets of disadvantage.

2.2.3 The Index of Economic Resources

The IER summarises variables relating to the financial aspects of relative socio-economic advantage and disadvantage. These include indicators of high and low income, as well as variables that correlate with high or low wealth.

Areas with higher scores have relatively greater access to economic resources than areas with lower scores.

2.2.4 The Index of Education and Occupation

The IEO summarises variables relating to the educational and occupational aspects of relative socio-economic advantage and disadvantage. This index focuses on the skills of the people in an area, both formal qualifications and the skills required to perform different occupations.

A low score indicates that an area has a high proportion of people without qualifications, without jobs, and/or with low skilled jobs. A high score indicates many people with high qualifications and/or highly skilled jobs.

3. THE DATA UNDERPINNING THE INDEXES

This section looks at the data used to construct the four indexes in SEIFA 2011. All data is from the 2011 Census of Population and Housing.[1]

3.1 Developing a candidate list of variables

Before constructing the indexes, we reviewed the list of Census variables and identified those associated with our definition of socio-economic advantage and disadvantage, as discussed in Section 2.

When developing the candidate list of variables, we considered variables that are either (i) a cause, (ii) a consequence, or (iii) have an association with advantage or disadvantage. We adopted this approach because it was deemed it to provide the best measure to reflect the relative advantage and disadvantage of an area. Variables that are a cause or an association act as proxy measures for consequence variables that are not observed on the Census, but are still important in measuring advantage or disadvantage.

The variables used in SEIFA 2006 provided a starting point for developing a candidate list of variables, particularly considering that the Census questions had not changed from 2006 to 2011. New variables were considered for inclusion by reassessing the list of Census variables in the context of the year 2011, and the notion of advantage and disadvantage we used. The literature on indicators of advantage and disadvantage was also considered to help in this assessment.

As mentioned briefly in Section 2.1, we used a multi-dimensional framework to guide the variable selection process. The dimensions used were:

- income variables,
- education variables,
- employment variables,
- occupation variables,
- housing variables, and
- other miscellaneous indicators of relative advantage or disadvantage.

Variables can relate to persons, families, or dwellings. This reflects the fact that some of the Census variables apply to persons, some to families, and some to dwellings.

1 Quality Statements are available for each Census data item on the ABS website through the Census web portal. See also *Census Dictionary, 2011* (ABS, 2011a).

3.2 Constructing the variables

Before moving onto a discussion about which variables were included in the candidate list, it is useful to consider some general points on how the variables were defined for use in the indexes.

Specifications

To facilitate the construction of the area-based indexes, the variables were expressed as proportion of units in an area with a specific characteristic. Depending on the variable, the unit may be a person, family, or dwelling.

As each variable was expressed as a proportion, a numerator and denominator were required. The numerator for each variable was a subset of the denominator. In most cases, the numerator and denominator specifications were based on SEIFA 2006 specifications. Where variables were new or modified for 2011, we specified numerators and denominators based on our own analysis and research into the relevant literature, as well as consultation with ABS subject matter experts. Appendix A contains detailed descriptions of the numerators and denominators used for all the SEIFA variables.

Note that for convenience of presentation in the following sections, the variable proportions are expressed as percentages.

Place of Usual Residence

A person may or may not be enumerated at their place of usual residence on Census Night. For all variables used in SEIFA 2011, persons were returned back to their usual residence to create SA1 level numerator and denominator counts. SEIFA 2006 was the first release of the indexes to use place of usual residence as the basis for area level counts, with previous editions of SEIFA using place of enumeration counts to create the variables. Counts compiled on a 'place of usual residence' basis are more appropriate for SEIFA, because they are less likely to be influenced by seasonal factors such as school holidays and snow seasons. However, it is important to understand that certain areas, for example SA1s in popular tourist destinations, may receive scores influenced by the specific time at which the Census is conducted. For instance, the 2011 Census was conducted in August 2011, corresponding to the high season for ski resorts and the townships in those areas. This means that these areas may witness higher property rental prices, higher employment figures and greater income levels than if the Census were conducted in the low season.

Not stated and not applicable

We excluded records with 'Not stated' and 'Not applicable' values (for the particular variable) from both the numerator and denominator counts. For details, see Appendix A.

Transformation of skewed variables

We considered transforming some variables that had highly skewed distributions, in order to make the variables behave more realistically in terms of their contribution to an area's index score. We investigated this issue for several variables, and concluded that transforming variables (including truncation) had little effect on the final indexes, yet added an additional layer of complexity (and many decisions) to their calculation. Therefore, for SEIFA 2011 we decided to maintain the practice from SEIFA 2006 and not perform any transformation of variables.

3.3 Description of candidate SEIFA variables

This section contains a description of each variable on the candidate variable list. There is a brief discussion of how each variable relates to our definition of relative socio-economic advantage or disadvantage. We also highlight the variables that have been modified since SEIFA 2006, and those that are new in 2011. The tables containing the variable descriptions also state whether the variable is an indicator of relative advantage (adv) or relative disadvantage (dis).

Each subsection corresponded to one of the socio-economic dimensions listed in Section 3.1.

3.3.1 Income variables

3.1 List of income variables

Variable mnemonic	*Variable description*
INC_LOW	% People with stated annual household equivalised income between $1 and $20,799 (approx. 1st and 2nd deciles) (dis)
INC_HIGH	% People with stated annual household equivalised income greater than $52,000 (approx. 9th and 10th deciles) (adv)

Note – In this table, and subsequent tables, the variable descriptions state whether the variable is an indicator of relative advantage (adv) or relative disadvantage (dis).

Income is an important economic resource, and is a core component of our notion of relative socio-economic advantage and disadvantage (outlined in Section 2.1). Income variables are used in all the SEIFA indexes except the Index of Education and Occupation.

The SEIFA 2006 income variables used the widely accepted practice of equivalising household income. Equivalisation is a process in which household income is adjusted by an 'equivalence scale',[2] based on the number of adults and children in the household. This practice has been retained for income variables in SEIFA 2011.

2 The scale adopted by ABS is the modified OECD equivalence scale. For details, see Appendix 3 in *Household Income and Income Distribution, Australia, 2009-10* (ABS, 2011b).

The low income variable has been defined for SEIFA 2011 to capture approximately the first and second deciles of the equivalised household income distribution, excluding negative and nil income. That is, those people living in dwellings with equivalised household income between $1 and $399 per week ($1 to $20,799 per year). Much of the low income decile was a strong indicator of disadvantage, but people reporting negative and nil incomes tended to have profiles with less association with disadvantage. Further discussion on the definition of the low income variable is provided in Section 3.5.1.

The cut-off of $52,000 for the high income variable was chosen to approximately capture the highest income quintile (top 20%).

One limitation of the SEIFA income variables is that personal income is collected in ranges in the Census. In order to calculate equivalised household income, a dollar value had to be imputed for personal income, based on the range reported. The imputed figure was an estimation of the median income for each income range, based on income data from the *ABS Survey of Income and Housing, 2009–10*.

3.3.2 Education variables

3.2 List of education variables

Variable mnemonic	*Variable description*
ATUNI	% People aged 15 years and over attending university or other tertiary institution (adv)
ATSCHOOL	% People aged 15 years and over attending secondary school (adv)
CERTIFICATE	% People aged 15 years and over whose highest level of educational attainment is a Certificate Level III or IV qualification (dis)
DEGREE	% People aged 15 years and over whose highest level of educational attainment is a bachelor degree or higher qualification (adv)
DIPLOMA	% People aged 15 years and over whose highest level of educational attainment is an advanced diploma or diploma qualification (adv)
NOEDU	% People aged 15 years and over who have no educational attainment (dis)
NOYEAR12ORHIGHER	% People aged 15 years and over whose highest level of educational attainment is Year 11 or lower (includes Certificate Levels I and II; excludes those still at secondary school) (dis)

Education is an important domain when considering socio-economic advantage and disadvantage because the skills people obtain through school and post-school education can increase their own standard of living, as well as that of their community.

The SEIFA 2006 education variables were derived from two Census variables, QALLP (an individual's highest level of non-school qualification) and HSCP (an individual's highest year of school completed). The issue with this approach is that someone can have a high university qualification such as a masters degree while never having completed year 12. The 2006 variable, NOYEAR 12 (% people aged 15 years and over who left school at year 11 or lower), does not capture or account for this possibility.

This is not desirable because the variable is aiming to capture people whose highest level of educational attainment is relatively low.

To remedy the overlap between education categories in SEIFA 2006, the 2011 education variables are based on the Census variable HEAP (an individual's highest level of educational attainment), which is itself derived from the QALLP and HSCP variables. The decision to use the HEAP Census variable was based on a recommendation following the production of SEIFA 2006.

Certificate Levels I and II are regarded as a lower educational attainment than year 12 schooling, and as SEIFA 2011 education variables aim to express highest level of educational attainment, are grouped in the NOYR12ORHIGHER variable, as opposed to the CERTIFICATE variable. This specific educational hierarchy is based on the ABS publication *Education and Work Australia, May 2011* (ABS, 2011c). Note also that the CERTIFICATE variable is an indicator of relative disadvantage in SEIFA. It is true that having a certificate qualification gives a person an advantage over someone with no qualifications. However, at an area level, a high proportion of people with certificate qualifications correlates with other disadvantaging characteristics (e.g. lower skilled occupations).

3.3.3 Employment variables

3.3 List of employment variables

Variable mnemonic	*Variable description*
UNEMPLOYED	% People (in the labour force) who are unemployed (dis)
UNEMP_RATIO	% People aged 15 and over who are unemployed (dis)

For most people, employment is the main source of their income. Employment can also contribute to social participation and self-esteem. An unemployment variable is included in all of the SEIFA indexes.

The standard unemployment variable (UNEMPLOYED) is calculated as the number of unemployed people divided by the number of people in the labour force (the unemployment rate). The variable used in the Index of Economic Resources (UNEMP_RATIO) is the number of unemployed people divided by the entire adult population of the area. This was retained from SEIFA 2006 to distinguish the unemployed from those employed and those not in the labour force, as the latter two groups were found to have significantly higher average wealth.

3.3.4 Occupation variables

3.4 List of occupation variables

Variable mnemonic	*Variable description*
OCC_DRIVERS	% Employed people classified as Machinery Operators and Drivers (dis)
OCC_LABOUR	% Employed people classified as Labourers (dis)
OCC_MANAGER	% Employed people classified as Managers (adv)
OCC_PROF	% Employed people classified as Professionals (adv)
OCC_SALES_L	% Employed people classified as Low-Skill Sales Workers (dis)
OCC_SERVICE_L	% Employed people classified as Low-Skill Community and Personal Service Workers (dis)
OCC_SKILL1	% Employed people who work in a Skill Level 1 occupation (adv)
OCC_SKILL2	% Employed people who work in a Skill Level 2 occupation (adv)
OCC_SKILL4	% Employed people who work in a Skill Level 4 occupation (dis)
OCC_SKILL5	% Employed people who work in a Skill Level 5 occupation (dis)

Occupation plays a significant part in determining socio-economic advantage and disadvantage. The ability to accumulate economic resources varies greatly with occupation type.

The SEIFA 2011 occupation variables have been classified using ANZSCO – *Australian and New Zealand Standard Classification of Occupations, First Edition, Revision 1* (ABS, 2009). Released in 2009, this revision included the addition of 24 new occupations (categories at the 6-digit level) and the deletion/merging of eight occupations. It also included updates to the definitions and titles of some existing occupations and higher categories (that is, the 2-digit, 3-digit and 4-digit levels).

Each occupation in ANZSCO 2006 is assigned a skill level ranging from 1 (highest) to 5 (lowest), which is "a function of the range and complexity of the set of tasks performed in a particular occupation" (ABS, 2006, p. 6). These skill levels were used as the basis of the occupation variables in the Index of Education and Occupation. The aim was to include broad categories of both advantaging and disadvantaging occupations, which complement the education variables by introducing the aspect of vocational skills.

For the IRSD and the IRSAD, we used the ANZSCO major groups in conjunction with the skill levels to construct the occupation variables. This was done to identify occupations, or groups of occupations, which contribute to relative advantage or disadvantage at an area level. Using the major groups as well as the skill levels also helped to maintain consistency with SEIFA 2006.

3.3.5 Housing variables

3.5 List of housing variables (a)

Variable mnemonic	*Variable description*
FEWBED	% Occupied private dwellings with one or no bedrooms (dis)
HIGHBED	% Occupied private dwellings with four or more bedrooms (adv)
HIGHMORTGAGE	% Occupied private dwellings paying more than $2,800 per month in mortgage repayments (adv)
HIGHRENT	% Occupied private dwellings paying more than $370 per week in rent (adv)
LOWRENT	% Occupied private dwellings paying less than $166 per week in rent (excluding $0 per week) (dis)
MORTGAGE	% Occupied private dwellings owning the dwelling they occupy (with a mortgage) (adv)
OVERCROWD	% Occupied private dwellings requiring one or more extra bedrooms (based on Canadian National Occupancy Standard) (dis)
OWNING	% Occupied private dwellings owning the dwelling they occupy (without a mortgage) (adv)
SPAREBED	% Occupied private dwellings with one or more bedrooms spare (based on Canadian National Occupancy Standard) (adv)

(a) All dwelling variables excluded dwellings whose inhabitants all usually resided elsewhere, whose inhabitants were all under 15, or which could not be classified due to insufficient information. For numerator and denominator specifications, see Appendix A.

Having an adequate and appropriate place to live is fundamental to socio-economic wellbeing. There are many aspects to housing that affect the quality of people's lives. Dwelling size, cost and security of tenure are all important in this regard, and are therefore considered in SEIFA.

Housing size is measured by the variables FEWBED, HIGHBED, OVERCROWD and SPAREBED. The variable FEWBED measures dwellings with one or no bedrooms, whilst the variable HIGHBED measures dwellings with four or more bedrooms. The variable OVERCROWD measures dwellings that do not have enough bedrooms for their occupants. The variable SPAREBED measures dwellings conversely that have one or more bedrooms spare for their occupants. These last two variables are calculated using the Canadian National Occupancy Standard.[3]

Housing cost is measured in SEIFA using reported mortgage or rent payments. The cut-offs for the high and low groups were based on the ranges corresponding to the top and bottom quintiles. The high housing cost variables (HIGHMORTGAGE, HIGHRENT) are indicators of relative advantage, because they indicate greater financial capacity, as well as higher quality housing or locational advantage. The low housing cost variable (LOWRENT) is an indicator of relative disadvantage, for similar reasons.

3 The Canadian National Occupancy Standard determines housing appropriateness, using the number of bedrooms and the number, age, sex and relationships of household members. For more information, refer to *Housing Occupancy and Costs, 2009-10* (ABS, 2011d).

Owning a house, with or without a mortgage, is an indicator of advantage. First, owning a house implies security of tenure. For many Australian households, the family home is their most valuable asset. Owning with a mortgage indicates the financial capacity to make repayments, as well as the possession of a future asset.

The way we construct the household tenure variables has changed for SEIFA 2011. The denominator of the mortgage and rent variable proportions has been redefined to be based on all households in an area, instead of just those households with a mortgage or renting. This reduces the volatility of these variables in areas where there are low proportions of rented and mortgaged dwellings.

In SEIFA 2006, people renting from a government or community authority were captured in a variable named RENT_SOCIAL. Provision of public housing is typically means tested, and therefore highly associated with low financial wellbeing, however differing public housing policies across Australian jurisdictions make RENT_SOCIAL complex and difficult to interpret. Additionally, analysis of 2011 Census data revealed a large proportion of households in public housing also appear in the low rent category, and the LOWRENT and RENT_SOCIAL variables are highly correlated. For these reasons, the RENT_SOCIAL variable was not considered for SEIFA 2011.

The Census captures limited household information, and does not for instance capture housing affordability, housing stress, dwelling value and dwelling quality. Although some variables, such as number of bedrooms and amount of rent or mortgage payments, may provide a proxy in some instances, their relationship to dwelling quality and dwelling value is not uniform across all areas. Due to this lack of comparability we have not attempted to construct these variables.

3.3.6 Other indicators of relative advantage or disadvantage

With the information available to us from the Census there are additional variables we can construct related to socio-economic advantage and disadvantage that do not fall into the main domains of education, occupation, housing or employment. These variables are discussed below.

A new variable CHILDJOBLESS has been included for the first time in SEIFA 2011, defined as the proportion of families with children under 15 years old and jobless parents. The variable could be an indicator for entrenched disadvantage since children who grow up in jobless families may be more likely to experience inter-generational unemployment and diminished opportunities to participate in society. This variable is based on one of the Australian government's social inclusion priorities through the Australian Social Inclusion Board.[4]

4 For more information, see the Australian Social Inclusion Board papers *How Australia is Faring* (p. 32) and A *Compendium of Social Inclusion Indicators* (p. 53).

3.6 List of other indicators of relative advantage or disadvantage (a)

Variable mnemonic	*Variable description*
CHILDJOBLESS	% Families with children under 15 years of age and jobless parents (dis)
DIALUP	% Occupied private dwellings with a dialup internet connection (dis)
DISABILITYU70	% People aged under 70 who need assistance with core activities due to a long-term health condition, disability or old age (dis)
ENGLISHPOOR	% People who do not speak English well (dis)
GROUP	% Occupied private dwellings that are group occupied private dwellings (dis)
HIGHCAR	% Occupied private dwellings with three or more cars (adv)
LONE	% Occupied private dwellings that are lone person occupied private dwellings (dis)
NOCAR	% Occupied private dwellings with no cars (dis)
NONET	% Occupied private dwellings with no Internet connection (dis)
ONEPARENT	% Families that are one parent families with dependent offspring only (dis)
SEP_DIVORCED	% People aged 15 and over who are separated or divorced (dis)
UNINCORP	% Occupied private dwellings with at least one person who is an owner of an unincorporated enterprise (adv)

(a) All dwelling variables excluded dwellings whose inhabitants all usually resided elsewhere, whose inhabitants were all under 15, or which could not be classified due to insufficient information. For numerator and denominator specifications see Appendix A.

Having an internet connection allows access to information and services and may demonstrate a certain level of financial capability. In SEIFA 2006, the proportion of people with a broadband internet connection (BROADBAND) was used as an indicator of relative advantage. However, since the 2006 Census there was been a marked uptake in broadband internet and a corresponding decline in dial-up internet. As a result of the changes in the characteristics of internet access, it is no longer sensible to consider broadband internet connections to be an indicator of relative advantage – see *Internet Activity, Australia, June 2012* (ABS, 2012a). The BROADBAND variable has been dropped for SEIFA 2011. The DIALUP variable has been retained as an indicator of disadvantage. Section 3.5.2 contains more details on the internet variables.

The disability variable (DISABILITYU70) provides an indication of the physical or health aspects of socio-economic disadvantage. It is based on the Census question on need for assistance, which was developed to provide an indication of whether people have a profound or severe disability. People with a profound or severe disability are defined as those people needing help or assistance in one or more of the three core activity areas of self-care, mobility and communication, because of a disability, long term health condition (lasting six months or more) or old age.[5] Disability limits employment opportunities, and possibly access to community resources. For the purpose of indicating relative socio-economic disadvantage, we have limited the scope of the SEIFA disability variable to people aged under 70, as was done for SEIFA 2006.

5 Note that the Census measure was designed to indicate the disability status of people in Australia according to geographic area, or for small groups within the broader population. It is not a comprehensive measure of disability. For more information see *Census Dictionary, 2011* (ABS, 2011a).

Lacking fluency in English may limit employment opportunities, and ability to participate in society.

A car is both a material resource and a means of transport that enables greater freedom. A limitation of the NOCAR variable is that the need for a car varies depending on the remoteness of the area and access to public transport.

An analysis of wealth data from the ABS *Survey of Income and Housing, 2007–08*, showed that lone person households have lower average wealth (per person) than other household types. A higher proportion of lone-person households in an area is correlated with lower ability to access economic resources beyond what is measured by the equivalised household income variables. An analysis on group households yielded a similar conclusion – an association with low wealth. A high proportion of unincorporated enterprise owners was found to correlate with high wealth and access to economic resources. These three variables were used only in the Index of Economic Resources.

One parent households are disadvantaged as compared to other household types, because of the need to simultaneously provide and care for dependents. Apart from having lower equivalised household incomes, one parent families also have lower rates of employment and labour force participation, lower rates of home ownership and higher incidence of financial stress, as compared to couple family households – see, for example, *Australian Social Trends, 2007* (ABS, 2007). There are significant correlations at the area level between the number of one parent families and many indicators of relative socio-economic disadvantage. The same patterns are evident for areas with high proportions of people who are separated or divorced.

We considered including new Census data items relating to supported accommodation, improvised dwellings and youth engagement in both education and employment. However, these data items had very skewed distributions and had relatively high levels of non-response. When considered with the exclusion rules framework (see Section 4.2) concerning low denominator counts, these variables excluded significant numbers of additional areas. The types of areas excluded were biased towards areas with high proportions of aged residents. For these reasons none of these variables were included in SEIFA.

One variable included in the IRSD in past releases of SEIFA has been the proportion of people in an area who identified as being of Aboriginal and/or Torres Strait Islander origin. This variable was not included on the final candidate variable list for SEIFA 2011. For more details on this issue see Section 3.5.3.

3.4 Basic exploratory analysis of variables

The Census data was converted into the SEIFA variable proportions, as defined in section 3.3. Summary statistics, distributions, and comparisons with the SEIFA 2006 proportions were analysed in order to better understand the data and identify any changes since 2006.

Overall, there were no unexpected changes to the SEIFA variable proportions. The shape and spread of the distributions changed between the 2006 and 2011 Census for the following variables:

- dwellings with no internet connection,
- dwellings paying low rental payments,
- dwellings paying high rental payments,
- people whose highest level of educational attainment is a bachelor degree or higher, and
- people whose highest level of educational attainment is a certificate I or II qualification.

These findings were unsurprising given the changes to the household rental market, internet affordability, technology improvements and increases in the education of the Australian population that have occurred over the past five years.

To further validate the SEIFA 2011 variable proportions, the areas with the lowest ten and highest ten proportion values were inspected for plausibility. There were no unusual or unexplainable results.

3.5 Exploration of some selected variables

As mentioned previously, many of the potential variables for SEIFA 2011 are based on SEIFA 2006. However, there were some variables that required substantial analysis and thought before deciding on whether to include them or how to define them. This section presents analysis and discussion of three categories of variables that required extra consideration for SEIFA 2011: income variables, internet variables, and an Indigenous variable.

3.5.1 Income variables

The low and high equivalised household income variables used in the SEIFA indexes attempt to capture the lowest and highest quintiles of the stated income distribution from the Census. However, because Census income data is reported in ranges, the population distribution across the income range categories does not always facilitate accurate calculation of quintiles. The 2011 income distribution segmented clearly into a top income quintile for equivalised income greater than $52,000 per year, the same definition as was used for SEIFA 2006, however this was not the case for the bottom income quintile. Further complicating the choice of low income definition is the issue of negative and nil equivalised income.

A broad conclusion is difficult to draw about low equivalised income because of the diverse nature of households with low, negative and nil income – see *Household Wealth and Wealth Distribution* (ABS, 2011e). For instance, a retiree who does not get the age pension may be drawing down on a lump sum superannuation, which does not count as income. Negative income can arise from owning an unincorporated business or from losses on financial investments. However, people with negative incomes generally do not share similar socio-economic characteristics to people in the lowest positive income category; they tend to have enough wealth to cover negative incomes, at least temporarily.

The SEIFA 2006 low income variable captured people with equivalised household incomes between $13,000 and $20,799, corresponding to the second and third deciles of the income distribution. The choice to use the second and third deciles and to exclude the first decile was based on the notion that people in the lowest income decile have varying financial circumstances. However, for SEIFA 2011 we thought this could be refined further, and hence conducted some analysis of alternatives.

The analysis compared some alternative low income definitions with the 2006 low income definition '% people with weekly household equivalised income between $300 and $399' (INC_LOW_OLD). The first alternative definition removed negative and nil income and defined low income as '% people with weekly equivalised household income between $1 and $399' (INC_LOW). The second definition included negative and nil income, framing low income as '% people with weekly equivalised household income between $<0 and $399' (INC_LOW_NILNEG).

In assessing income definitions, we first examined the effect the choice of income variable has on the variable selection process for the 2011 IRSD – see Section 4 for details. The final loading (correlation with the index) for INC_LOW_OLD was found to be 0.79, INC_LOW was 0.90 and INC_LOW_NILNEG was 0.89. Including the lowest income decile in the definition of low income clearly increases the strength of the relationship between low income and the IRSD. Additionally, the choice of low income variable does not alter the order in which the remaining variables are excluded from the index.

Our second line of enquiry was to examine whether people living in households with nil or negative income can be classified as relatively disadvantaged. For these households, we looked at a number of Census variables including the number of vehicles, household mortgage repayments, and highest level of educational attainment. In all the analyses, it was observed that people living in households with nil or negative income tended to have more similar characteristics to those living in higher income households.

Further analysis was conducted to determine whether nil or negative income is a good indicator of disadvantage at an area level. Using the alternative indexes created by the above process, we created plots of the proportion of people living in nil or negative income households within each SA1 against both the IRSD score and the IRSD percentile. The plots indicated that the proportions of nil or negative income households are not a good indicator of disadvantage at an area level.

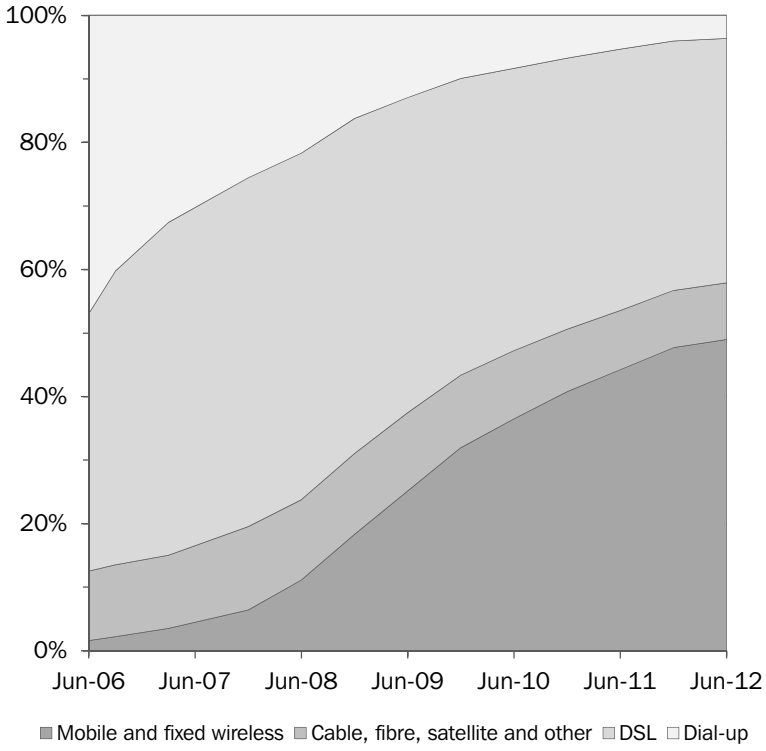
Based on these findings, the decision was made to use the INC_LOW definition ('% people with weekly equivalised household income between $1 and $399) as our low income variable.

3.5.2 Internet variables

Internet access in a household allows access to information and services and can be used to demonstrate certain levels of financial capability. In SEIFA 2001, the proportion of people with any type of Internet connection was used as an indicator of relative advantage. In 2006, the proportion of occupied private dwellings with a broadband internet connection was used to indicate relative advantage, and the lack of any internet connection was used to indicate relative disadvantage. This section discusses analysis to establish the relative merits of considering broadband, dialup (DIALUP), and no internet connections (NONET) to measure relative advantage and disadvantage in SEIFA 2011.

Figure 3.7, derived from the ABS release *Internet Activity, Australia, Jun 2012* (ABS, 2012a), presents the changes in internet connections over time, as a proportion of subscribers by connection type.

3.7 Change in internet access characteristics (June 2006 to June 2011), Proportion of subscribers by connection type

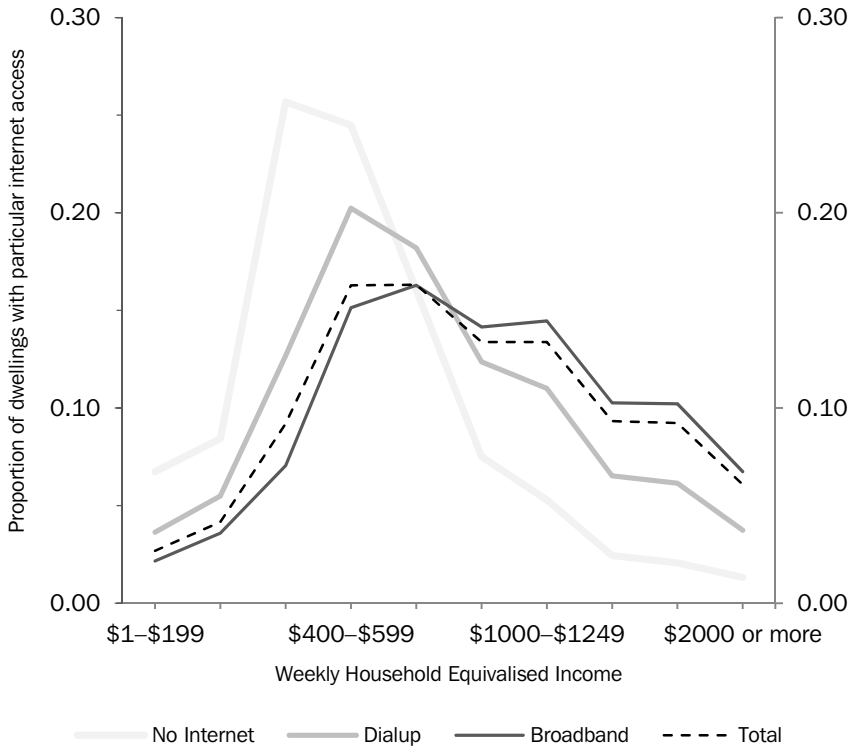


With rapid increases in the accessibility and capability of technology, and given the changes to the distribution of internet connection types presented in figure 3.7, it is no longer appropriate to consider broadband internet connections to reflect relative advantage. Instead, this segment of the population will act as a reference group.

Given the increasing prevalence of internet access, being disconnected or having a very slow internet connection is increasingly limiting access to resources for the most disadvantaged people in the population. However, it was unclear how subscribers to dial-up internet differed from those without internet, and whether it would thus be appropriate to consider both variables for the SEIFA indexes, only one of the variables, or a combination of both. To assess this, we analysed the relationship between internet connection type and household income, as presented in figure 3.8 below.

Figure 3.8 highlights the different distribution of equivalised household income for the DIALUP and NONET populations. Dwellings with no internet connection are more likely to have lower equivalised income than dwellings with internet connections, and likewise for dwellings with dialup internet connections when compared with the general population. Given these differences it was decided to consider both DIALUP and NONET as separate variables for inclusion in SEIFA 2011.

3.8 Household equivalised income by internet access characteristics



3.5.3 Indigenous variable

For past versions of SEIFA, an Indigenous variable (% of people in an area who identified themselves as being of Aboriginal and/or Torres Strait Islander origin) had been included in the IRSD on the basis it reflected measures of disadvantage, although was not a direct measure itself. For example, it may be associated with poor health or living conditions.

For SEIFA 2011, the Indigenous variable has not been included in the IRSD. In short, the reason for this decision is to provide a conceptually clearer index (IRSD) while giving a very similar set of rankings that are arguably no worse in terms of achieving the 'best' ranking of areas.

Catalysts for this change in position include:

- A marked change in the reported Indigenous population in the 2011 Census. There was an approximately 21% increase in the reported Indigenous population (ABS 2012b), mainly in urban areas. Further investigations indicated that the change varied between age groups and education level. For SEIFA purposes, it means we cannot easily assume that the SEIFA Indigenous variable is a consistent indicator (or proxy) of disadvantage across all of Australia (at least not as much as we have in the past). This emerging data issue fed into an existing uneasiness about the lack of a consistent framework for selecting proxy variables in SEIFA, particularly when it is not immediately clear what a variable is acting as a proxy for. There are already many explicit and recognised indicators of disadvantage included in SEIFA (e.g income, education, occupation, and housing).
- Feedback from stakeholders indicated that they want clarity on what the indexes are measuring. This is important so that they can be used properly. Some potential users have opted not to use SEIFA in the past because the inclusion of the Indigenous variable in the IRSD caused some confusion about how the IRSD should be used, particularly when analysing the Indigenous population.
- Some stakeholders have acknowledged that they see the logic of why the ABS has included the Indigenous variable in the past, since it was deemed that a better ranking of areas was achieved. However, analysis for SEIFA 2011 has indicated that the inclusion of the Indigenous variable does not have a substantial impact on the rankings. For further details of a comparison between the IRSD with and without the Indigenous variable, see Appendix B.

As with past versions of SEIFA, the goal for SEIFA 2011 has been to achieve the best ranking of areas in terms of relative disadvantage and advantage. It is with this goal in mind that the decision was made to not consider the Indigenous variable for inclusion in the 2011 IRSD. It was deemed that the conceptual confusion brought into the index by including the Indigenous variable is no longer offset by any potential positive impacts of it acting as a proxy – the rankings do not change meaningfully. This assessment is based on the context of SEIFA 2011, particularly Census 2011 data and feedback from the growing SEIFA user base.

3.6 Candidate variable list for each index

Table 3.9 shows the candidate variable list for each index. The candidate list includes all variables considered for inclusion in an index before conducting principal component analysis (discussed in section 4). The final list of variables included in each index can be found in table 4.10 in section 4.4.5.

3.9 Candidate variable list for each index, by socio-economic dimension

Dimension	*Index of Relative Disadvantage*	*Index of Relative Advantage and Disadvantage*	*Index of Economic Resources*	*Index of Education and Occupation*
Income	INC_LOW	INC_HIGH INC_LOW	INC_HIGH INC_LOW	
Education	NOYR12ORHIGHER NOEDU CERTIFICATE	NOYR12ORHIGHER NOEDU CERTIFICATE ATUNI DIPLOMA DEGREE		NOYR12ORHIGHER NOEDU CERTIFICATE ATUNI DIPLOMA DEGREE ATSCHOOL
Employment	UNEMPLOYED	UNEMPLOYED	UNEMP_RATIO	UNEMPLOYED
Occupation	OCC_LABOUR OCC_DRIVERS OCC_SERVICE_L OCC_SALES_L	OCC_LABOUR OCC_DRIVERS OCC_SERVICE_L OCC_SALES_L OCC_PROF OCC_MANAGER		OCC_SKILL1 OCC_SKILL2 OCC_SKILL4 OCC_SKILL5
Housing	LOWRENT OVERCROWD FEWBED	LOWRENT OVERCROWD FEWBED HIGHBED HIGHRENT HIGHMORTGAGE OWNING SPAREBED	LOWRENT OVERCROWD MORTGAGE HIGHBED HIGHRENT HIGHMORTGAGE OWNING	
Other	CHILDJOBLESS ONEPARENT NOCAR DISABILITYU70 ENGLISHPOOR SEP_DIVORCED NONET DIALUP	CHILDJOBLESS ONEPARENT NOCAR DISABILITYU70 ENGLISHPOOR SEP_DIVORCED NONET DIALUP HIGHCAR	UNINCORP ONEPARENT NOCAR GROUP LONE	

Note – Appendix A contains the definitions of each variable listed in this table.
Note – The variables listed in this table are not the final list of variables included in the indexes. For the final list, see table 4.10 in Section 4.4.5.

4. CONSTRUCTION OF THE INDEXES

This section describes the methods used to construct the indexes, some important technical specifications of each index, and some basic output.

Note that Sections 4.1 and 4.2 provide important contextual information for fully understanding the step-by-step index construction process presented in Section 4.3.

4.1 Principal Component Analysis

Each index is a weighted sum of SEIFA variables. As with past versions of SEIFA, principal component analysis (PCA) is used to determine the weights. This section introduces some technical concepts related to PCA to assist the reader understand the SEIFA index construction process. Some references are given at the end of this section for readers interested in a comprehensive discussion of PCA.

PCA is a technique that involves summarising a large number of correlated variables into a set of new uncorrelated components, each of which is a linear combination of the original variables. There are as many principal components as there are variables. If the original variables are highly correlated, much of the variation can be summarised by a reduced set of components, hence summarising the information and enabling some easier analysis.

The first principal component accounts for the largest proportion of variance in the original dataset, with each following component explaining less of the variance. The principle component used for each SEIFA index is the one that can be interpreted as best explaining the variation in the concept of advantage and disadvantage for that index. For all four of the indexes in SEIFA 2011, the first principal component was used[6].

The PCA procedure gives an eigenvalue for each component, which indicates the amount of variance in the original data explained by the component. The proportion of variance explained by a principal component is its eigenvalue divided by the sum of all the eigenvalues.

Each variable in the analysis will be correlated with each component. This correlation is called the loading. Loadings help to interpret which aspects of advantage and disadvantage a component may represent. The loadings are also useful in comparing results obtained from different sets of original variables (such as for the four indexes in SEIFA). Loadings for each index are presented in the following sections.

In order to generate the component scores (otherwise known as raw scores) the loading is converted to a weight by dividing it by the square root of the eigenvalue. The product of the weight and standardised variable values are summed to produce

6 Component rotation is an optional variant of PCA and has been considered in past versions of SEIFA. After some investigations for SEIFA 2011, the same conclusion was drawn – the first unrotated component was most suitable for forming each of the indexes.

the raw scores. The raw scores for each component will then have variance equal to the eigenvalue for that component. We then rescale (standardise) the raw scores to a mean of 1,000 and standard deviation of 100 to create a new set of scores that are the index scores in SEIFA.

More detailed explanations of PCA can be found in Joliffe (1986) and O'Rourke (2005), among others.

Before moving onto a step-by-step description of the index construction process, it is necessary to describe how we finalise the dataset for analysis and output. This is covered in the next section.

4.2 Areas with no index scores

Some areas (SA1s) do not receive an index score, either due to low populations or poor quality data. The criteria to identify these areas are termed 'exclusion rules'.

For SEIFA 2011, the exclusion rule framework has been updated (from 2006) in order to obtain a reliable index score for as many areas as possible. The changes to the exclusion rule framework provide a small but positive refinement to the final list of areas receiving a score.

The 2011 exclusion rules work under a two-phase system:

- The first phase excludes areas (SA1s) that should not receive a SEIFA score because of the type of area, confidentiality, and reliability concerns (e.g. no address SA1, low population).
- The second phase excludes areas (SA1s) by looking specifically at the variables included in each index. For each SA1, if any of the variables have a low denominator count (< 6), it is deemed that there is not enough data to support a reliable calculation of an index score for that area.

Some additional comments on the exclusion rule framework:

- The first phase rules are applied before PCA, and the second phase rules are applied after PCA and the list of variables is finalised. Section 4.3 provides details on how this is implemented.
- SA1s excluded in the first phase will be excluded for all four indexes. The number of SA1s excluded in the second phase will be different for each index – they have different sets of variables.
- Following on from the point above, an area can receive an index score for one index and not another depending on the make-up of its variables.
- The low denominator cut-off of 6 is chosen based on past practice and a judgement on how many responses are required to calculate a reliable value for an area.

- The exclusion of areas is based on both the confidentialised and unconfidentialised counts for each SEIFA variable to ensure the confidentiality of respondents is upheld and the reliability of the indexes is maintained.

The specific exclusion rules and the number of areas meeting each rule are summarised in table 4.1. Note that areas might fall into multiple categories, and this is why the column sum does not equal the final total number of excluded areas.

4.1 Summary of excluded areas

	Total SA1s excluded – First phase
Exclusion rule	
Population = 0	1,166
0 < Population ≤ 10	673
Employed persons ≤ 5	1,870
Number of classifiable occupied private dwellings ≤ 5	1,986
Proportion of people in private dwellings ≤ 20%	1,540
No address SA1	9
Offshore SA1	24
Total excluded	2,126

Index	*Total SA1s excluded – Second phase*
IRSD	102
IRSAD	103
IER	74
IEO	11

In 2011, the percentage of areas excluded for the four indexes ranges from 3.90–4.07% (2,137 to 2,231 areas out of 54,805) and the percentage of population excluded ranges from 0.71–0.72% (151,700 to 155,109 people out of 21,507,715). These figures compare well to the excluded areas for SEIFA 2006, where 1,256 or 3.2% of all CDs were excluded and 157,491 people or 0.79% of the population were excluded.

The increase in proportion of areas excluded can be attributed to the improved design criteria of SA1s, whereby there are more zero and low population SA1s. The consequence of both the improved design in SA1s and also the new exclusion rules framework is a lower proportion of the population excluded for SEIFA 2011.

If readers are interested, further details on the SEIFA 2006 exclusion rules are available in ABS (2008).

4.3 Step-by-step process

With the preceding two sections providing context, a step-by-step process for constructing the indexes is presented below.

Step 1. Creating the initial variable list

Given the available data, we created a list of variables related to our definition of relative socio-economic advantage and disadvantage.

Step 2. Constructing the variables

We created all variables as proportions at the SA1 level (e.g. '% people aged 15 and over with no post-school qualifications'). We then standardised these proportions to a mean of 0 and a standard deviation of 1. The standardisation was used to prevent variables with larger prevalence, or larger ranges, from exerting excessive influence on the index.

Step 3: Applying first phase exclusion rules

We excluded areas (SA1s) that should not receive an index score because of the type of area, confidentiality, and reliability concerns. See table 4.1 for specific rules.

Step 4: Calculating the correlation matrix

We set to missing any variables that have denominators less than our prescribed cut-off of 6. Note that we did not exclude areas based on this cut-off at this stage in the process – this occurred at step 9.

We calculated the correlation matrix and used pairwise deletion[7] when areas (observations) contain missing values. Given the number of observations in our dataset (approximately 55,000 SA1s) and the low prevalence of missing values, the use of pairwise deletion had very little impact on the correlation matrix, however it did enable a convenient way of implementing our second phase exclusion rules (step 9).

Step 5. Removing very highly correlated variables

We removed highly correlated variables to avoid over-representing any specific socio-economic characteristic. When two variables had a correlation coefficient greater than 0.8 in absolute value, and were measuring conceptually similar aspects of advantage or disadvantage, we generally removed one of them. However, we applied some discretion, depending on the particular variables and the size of the correlation.

7 Pairwise deletion is a method for dealing with missing data. The maximum number of non-missing values for each pair of variables is used in the calculation of the correlation matrix. This is in contrast to listwise deletion in which entire records (areas in our case) are removed from the analysis if any of their variables have missing values.

Step 6. Conducting the initial PCA

Using the correlation matrix, we conducted principal component analysis (PCA) to obtain the loading for each variable on the first principal component.

Step 7. Removing low loading variables

We excluded variables with loadings less than 0.3 in absolute value, on the grounds that they were not strong indicators of relative advantage or disadvantage. This limit is an accepted level in the PCA literature (see Joliffe, 1986) and has been used in past releases of SEIFA. We removed variables one at a time, starting with the lowest loading variable.

Step 8. Conducting PCA on the reduced list of variables

We conducted a PCA on the reduced variable list, and if any other variables loaded below 0.3, we repeated steps 7 and 8.

Step 9. Finalise list of variables in index and apply second phase exclusion rules

Once we knew the final list of variables in the index, we could exclude any areas (SA1s) that had any of their variable denominators less than our prescribed cut-off of 6.

Step 10. Calculating and standardising component/index scores

We derived the first principal component scores for each SA1 by taking the product of each standardised variable with its respective weight, then taking the sum across all variables. Note that the weight for each variable was calculated by dividing the loading by the square root of the eigenvalue.

$$Z_{SA1} = \sum_{j=1}^{p} \frac{L_j}{\sqrt{\lambda}} \times X_{j,SA1}$$

where

Z_{SA1} = raw score for the SA1;

$X_{j,SA1}$ = standardised variable value of the j-th variable for the SA1;

L_j = loading for the j-th variable;

λ = the eigenvalue of the principal component; and

p = total number of variables in the index.

For convenience of presentation, we then rescaled (standardised) the raw scores to a mean of 1,000 and standard deviation of 100 to create a new set of scores that are the SA1 index scores in SEIFA.

Note that the principal components are arbitrary with respect to their sign (positive or negative), so we set the sign of the weights and loadings so that they make intuitive sense. That is, we gave advantage indicators positive weights and loadings, and disadvantage indicators negative weights and loadings. Accordingly, high scores indicate relative advantage, and low scores indicate relative disadvantage. This is consistent with previous editions of SEIFA.

Step 11. Creating higher geographic level indexes

We constructed indexes for geographies higher than the SA1 level using population weighted averages of the constituent SA1s. We used the following formula:

$$INDEX_{AREA} = \frac{\sum_{i=1}^{n}\left(INDEX_{SA1_i} \times POP_{SA1_i}\right)}{POP_{AREA}}$$

where

INDEX = index score for each SA1 or higher level area;

POP = population for each SA1 or higher level area[8] ; and

n = total number of SA1s (with index scores) in the higher level area.

Although the higher level indexes were constructed from standardised SA1 level indexes, they were not standardised themselves. Therefore the higher level area indexes do not necessarily have a mean of 1,000 or standard deviation of 100.

Only SA1s with index scores were used to create the higher level indexes. In a small number of cases, where a higher level area contains a number of SA1s that were excluded, its index score may not be a good representation of its entire population. For this reason, the output spreadsheets provide the proportion of each higher area level population that was in excluded SA1s.

In general, we encourage users conducting analysis at higher level areas to keep in mind that the indexes were constructed at the SA1 level, and to consider using the distribution of SA1s within the higher level areas, rather than just the one index score for each higher level area. This is further discussed in Section 6.3.

8 The higher level area population is the sum of the populations from the constituent SA1s that received an index score. Populations in excluded SA1s are not included in this calculation.

4.4 Technical details of each index: variables and loadings

This section gives the results of the principal component analysis carried out for each index, including variable loadings and percentage of variance explained. We also outline which variables were initially considered for inclusion but removed due to high correlations with other variables or weak loadings.

4.4.1 Index of Relative Socio-Economic Disadvantage

The IRSD summarises variables that indicate relative disadvantage at the SA1 level, according to the concept described in Section 2.2.1. The final variable list and corresponding loadings are shown below in table 4.2.

4.2 Final IRSD variables and loadings

Variable mnemonic	*Variable loading*	*Variable description*
INC_LOW	–0.90	% People with stated annual household equivalised income between $1 and $20,799 (approx. 1st and 2nd deciles)
CHILDJOBLESS	–0.85	% Families with children under 15 years of age who live with jobless parents
NONET	–0.81	% Occupied private dwellings with no internet connection
OCC_LABOUR	–0.75	% Employed people classified as 'labourers'
NOYR12ORHIGHER	–0.75	% People aged 15 years and over whose highest level of education is Year 11 or lower. Includes Certificate I and II
UNEMPLOYED	–0.74	% People (in the labour force) unemployed
LOWRENT	–0.73	% Occupied private dwellings paying rent less than $166 per week (excluding $0 per week)
ONEPARENT	–0.71	% One parent families with dependent offspring only
DISABILITYU70	–0.66	% People aged under 70 who have a long–term health condition or disability and need assistance with core activities
NOCAR	–0.56	% Occupied private dwellings with no cars
SEP_DIVORCED	–0.54	% People aged 15 and over who are separated or divorced
OVERCROWD	–0.52	% Occupied private dwellings requiring one or more extra bedrooms (based on Canadian National Occupancy Standard)
OCC_DRIVERS	–0.52	% Employed people classified as Machinery Operators and Drivers
OCC_SERVICE_L	–0.50	% Employed people classified as Low Skill Community and Personal Service Workers
NOEDU	–0.44	% People aged 15 years and over who have no educational attainment
ENGLISHPOOR	–0.34	% People who do not speak English well

Removal of highly correlated variables

Of the variables considered for the IRSD, there were no two variables that had a correlation coefficient greater than 0.8 in absolute value.

Removal of low loading variables

Table 4.3 shows the variables that were dropped from the IRSD because their loading was below our prescribed cut-off of 0.3 in absolute value. The variables are shown in the order they were removed, with the loadings from the iteration when they were removed.

4.3 IRSD variables removed due to low loadings

Variable mnemonic	*Variable loading*	*Variable description*
DIALUP	–0.04	% Occupied private dwellings with a dialup internet connection
CERTIFICATE	–0.07	% People aged 15 years and over whose highest level of educational attainment is a Certificate III or IV qualification
OCC_SALES_L	–0.19	% Employed people classified as Low Skill Sales
FEWBED	–0.20	% Occupied private dwellings with one or no bedrooms

Variance explained

The eigenvalue for the IRSD was 7.06. The index explained 44% of the total variance of its 16 input variables. This is higher than both the 2006 IRSD (39%) and the 2001 IRSD (32.5%).

4.4.2 Index of Relative Socio-Economic Advantage and Disadvantage

The IRSAD summarises variables that indicate either relative socio-economic advantage or disadvantage, according to the concept described in Section 2.2.2. The final variable list and corresponding loadings are shown below in table 4.4.

4.4 Final IRSAD variables and loadings

Variable mnemonic	*Variable loading*	*Variable description*
INC_LOW	–0.89	% People with stated annual household equivalised income between $1 and $20,799 (approx. 1st and 2nd deciles)
NONET	–0.82	% Occupied private dwellings with no internet connection
NOYR12ORHIGHER	–0.82	% People aged 15 years and over whose highest level of education is Year 11 or lower. Includes Certificate I and II
CHILDJOBLESS	–0.80	% Families with children under 15 years of age who live with jobless parents
OCC_LABOUR	–0.78	% Employed people classified as 'labourers'
ONEPARENT	–0.69	% One parent families with dependent offspring only
UNEMPLOYED	–0.69	% People (in the labour force) unemployed
DISABILITYU70	–0.67	% People aged under 70 who have a long-term health condition or disability and need assistance with core activities
LOWRENT	–0.67	% Occupied private dwellings paying rent less than $166 per week (excluding $0 per week)
SEP_DIVORCED	–0.57	% People aged 15 and over who are separated or divorced
OCC_DRIVERS	–0.57	% Employed people classified as Machinery Operators and Drivers
OCC_SERVICE_L	–0.51	% Employed people classified as Low Skill Community and Personal Service Workers
NOCAR	–0.49	% Occupied private dwellings with no cars
OVERCROWD	–0.45	% Occupied private dwellings requiring one or more extra bedrooms (based on Canadian National Occupancy Standard)
NOEDU	–0.37	% People aged 15 years and over who have no educational attainment
HIGHCAR	0.35	% Occupied private dwellings with three or more cars
ATUNI	0.36	% People aged 15 years and over at university or other tertiary institution
SPAREBED	0.37	% Occupied private dwellings with one or more bedrooms spare
HIGHRENT	0.40	% Occupied private dwellings paying rent greater than $370 per week
OCC_MANAGER	0.42	% employed people classified as Managers
HIGHBED	0.52	% Occupied private dwellings with four or more bedrooms
OCC_PROF	0.62	% Employed people classified as Professionals
DIPLOMA	0.63	% People aged 15 years and over whose highest level of education attainment is a diploma qualification
HIGHMORTGAGE	0.70	% Occupied private dwellings paying mortgage greater than $2,800 per month
INC_HIGH	0.84	% People with stated annual household equivalised income greater than $52,000 (approx 9th and 10th deciles)

Removal of highly correlated variables

The variable DEGREE had high correlations with NOYR12ORHIGHER (–0.85) and OCC_PROF (0.92). This suggested that the proportion of people in an area with a degree was explained by other variables in the index. Therefore DEGREE was dropped.

Removal of low loading variables

Table 4.5 shows the variables dropped from the IRSAD because of weak loadings. The variables are shown in the order they were removed, with the loadings from the iteration when they were removed.

4.5 IRSAD variables removed due to low loadings

Variable mnemonic	*Variable loading*	*Variable description*
DIALUP	–0.08	% Occupied private dwellings with a dialup internet connection
FEWBED	–0.16	% Occupied private dwellings with one or no bedrooms
CERTIFICATE	–0.19	% People aged 15 years and over whose highest level of educational attainment is a certificate III or IV qualification
OWNING	0.22	% Occupied private dwellings owning dwelling without a mortgage
OCC_SALES_L	–0.23	% Employed people classified as Low Skill Sales
ENGLISHPOOR	–0.29	% People who do not speak English well

Variance explained

The eigenvalue for the IRSAD was 9.70. The index explained 39% of the total variance of its 25 input variables. This is slightly lower than both the 2006 IRSAD (44%) and the 2001 IRSAD (41%).

4.4.3 Index of Economic Resources

The IER focuses on the financial aspects of relative socio-economic advantage and disadvantage, according to the concept described in Section 2.2.3. The final variable list and corresponding loadings are shown below in table 4.6.

4.6 Final IER variables and loadings

Variable mnemonic	*Variable loading*	*Variable description*
INC_LOW	–0.79	% People with stated annual household equivalised income between $1 and $20,799 (approx. 1st and 2nd deciles)
NOCAR	–0.77	% Occupied private dwellings with no cars
LOWRENT	–0.72	% Occupied private dwellings paying rent less than $166 per week (excluding $0 per week)
ONEPARENT	–0.66	% One parent families with dependent offspring only
LONE	–0.66	% Occupied private dwellings who are lone person occupied private dwellings
UNEMP_RATIO	–0.57	% People aged 15 years and over who are unemployed
OVERCROWD	–0.54	% Occupied private dwellings requiring one or more extra bedrooms (based on Canadian National Occupancy Standard)
GROUP	–0.31	% Occupied private dwellings who are group occupied private dwellings
OWNING	0.33	% Occupied private dwellings owning dwelling without a mortgage
UNINCORP	0.49	% Dwellings with at least one person who is an owner of an unincorporated enterprise
INC_HIGH	0.63	% People with stated annual household equivalised income greater than $52,000 (approx 9th and 10th deciles)
MORTGAGE	0.66	% Occupied private dwellings owning dwelling (with a mortgage)
HIGHMORTGAGE	0.67	% Occupied private dwellings paying mortgage greater than $2,800 per month
HIGHBED	0.74	% Occupied private dwellings with four or more bedrooms

Removal of highly correlated variables

No variables were dropped based on high correlations.

Removal of low loading variables

Table 4.7 shows the variable dropped from the IER because of a weak loading.

4.7 IER variables removed due to low loadings

Variable mnemonic	*Variable loading*	*Variable description*
HIGHRENT	0.07	% occupied private dwellings paying rent greater than $370 per week

Variance explained

The eigenvalue for the IER was 5.50. The index explained 39% of the total variance of its 14 input variables. This is slightly higher than the 2006 IER (35%).

4.4.4 Index of Education and Occupation

The IEO summarises variables related to educational qualifications and vocational skills, according to the concept described in Section 2.2.4. The final variable list and corresponding loadings are shown below in table 4.8.

4.8 Final IEO variables and loadings

Variable mnemonic	*Variable loading*	*Variable description*
NOYR12ORHIGHER	–0.88	% People aged 15 years and over whose highest level of education is Year 11 or lower. Includes Certificate I and II
OCC_SKILL5	–0.80	% Employed people who work in a Skill Level 5 occupation
OCC_SKILL4	–0.74	% Employed people who work in a Skill Level 4 occupation
CERTIFICATE	–0.54	% People aged 15 years and over whose highest level of educational attainment is a certificate III or IV qualification
UNEMPLOYED	–0.49	% People (in the labour force) unemployed
OCC_SKILL2	0.34	% Employed people who work in a Skill Level 2 occupation
ATUNI	0.57	% People aged 15 years and over at university or other tertiary institution
DIPLOMA	0.68	% People aged 15 years and over whose highest level of education attainment is a diploma qualification
OCC_SKILL1	0.89	% Employed people who work in a Skill Level 1 occupation

Removal of highly correlated variables

DEGREE (% People aged 15 years and over with a degree or higher qualification) was initially considered for inclusion in the IEO. However, it shared strong correlations with NOYR12ORHIGHER (–0.85) and OCC_SKILL1 (0.82). It was decided that the proportion of people with a degree was already well explained by the index, and DEGREE was removed.

Removal of low loading variables

Table 4.9 shows the variables dropped from the IEO because of weak loadings. The variables are shown in the order they were removed, with the loadings from the iteration when they were removed.

4.9 IEO variables removed due to low loadings

Variable mnemonic	*Variable loading*	*Variable description*
ATSCHOOL	–0.02	% People aged 15 years and over who are still attending secondary school
NOEDU	–0.28	% People aged 15 years and over who have no educational attainment

Variance explained

The eigenvalue for the IEO was 4.21. The index explained 47% of the total variance of its nine input variables. This is lower than the 2006 IEO (52%) but slightly higher than the 2001 IEO (46%).

4.4.5 Summary of variables included in indexes

Table 4.10 below shows the final set of variables included in each index. It enables comparison between the four indexes.

4.10 List of variables in each index, by socio-economic dimension

Dimension	*Index of Relative Disadvantage*	*Index of Relative Advantage and Disadvantage*	*Index of Economic Resources*	*Index of Education and Occupation*
Income	INC_LOW	INC_HIGH INC_LOW	INC_HIGH INC_LOW	
Education	NOYR12ORHIGHER NOEDU	NOYR12ORHIGHER NOEDU ATUNI DIPLOMA		NOYR12ORHIGHER CERTIFICATE ATUNI DIPLOMA
Employment	UNEMPLOYED	UNEMPLOYED	UNEMP_RATIO	UNEMPLOYED
Occupation	OCC_LABOUR OCC_DRIVERS OCC_SERVICE_L	OCC_LABOUR OCC_DRIVERS OCC_SERVICE_L OCC_MANAGER OCC_PROF		OCC_SKILL1 OCC_SKILL2 OCC_SKILL4 OCC_SKILL5
Housing	LOWRENT OVERCROWD	LOWRENT OVERCROWD SPAREBED HIGHRENT HIGHBED HIGHMORTGAGE	LOWRENT OVERCROWD OWNING MORTGAGE HIGHBED HIGHMORTGAGE	
Other	CHILDJOBLESS ONEPARENT DISABILITYU70 ENGLISHPOOR NOCAR SEP_DIVORCED NONET	CHILDJOBLESS ONEPARENT DISABILITYU70 HIGHCAR NOCAR SEP_DIVORCED NONET	UNINCORP ONEPARENT LONE GROUP NOCAR	

4.5 Distributions of the indexes

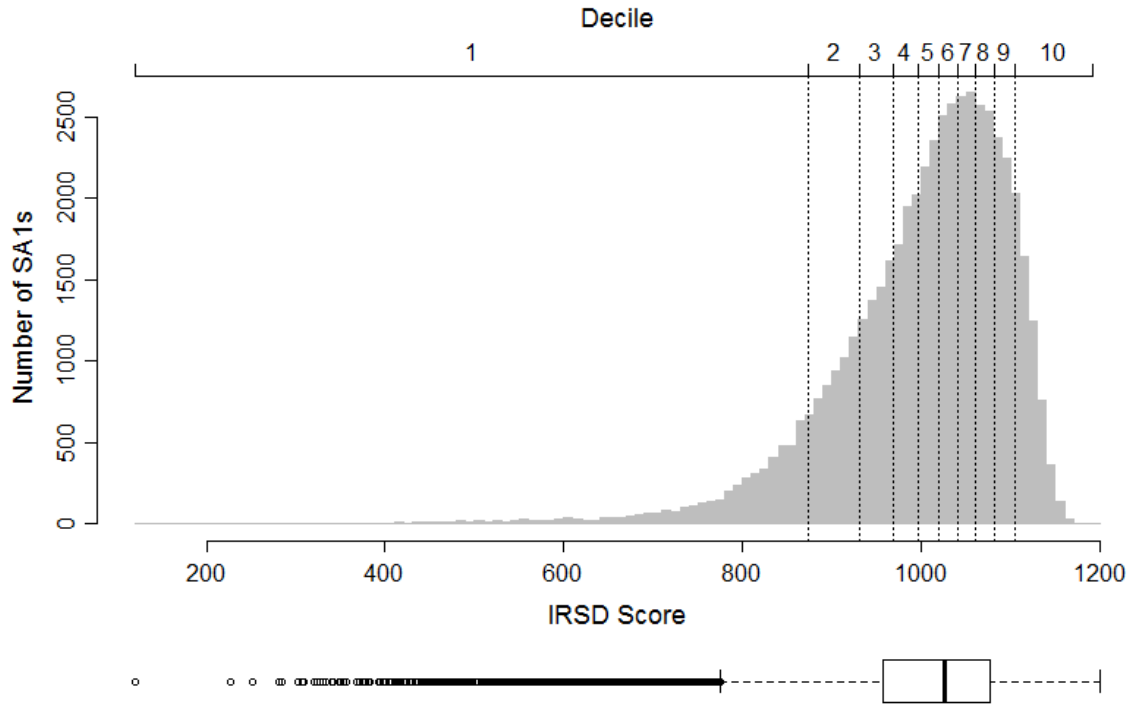
This section presents distribution plots for each index at the SA1 level. Box plots are provided beneath each frequency histogram to add more insight into the distributional features of the index scores. A general observation across all four distribution plots is that they each have longer left tails than right tails. This means that the spread amongst scores is greater for disadvantaged areas than for advantaged areas. All index distributions have a similar shape to the indexes in SEIFA 2006.

Note that a description of how to interpret box plots is provided in Appendix C.

4.5.1 *Index of Relative Socio-Economic Disadvantage*

The IRSD distribution displayed in figure 4.11 has a very long left tail, and is left-skewed. The values range from 120 to around 1200. The left slope is less steep than the right slope, meaning the scores of disadvantaged areas are more spread out than the scores of advantaged areas. This is because the index contains only disadvantage indicators, so there is more scope to distinguish between disadvantaged areas than advantaged areas.

4.11 IRSD score distribution

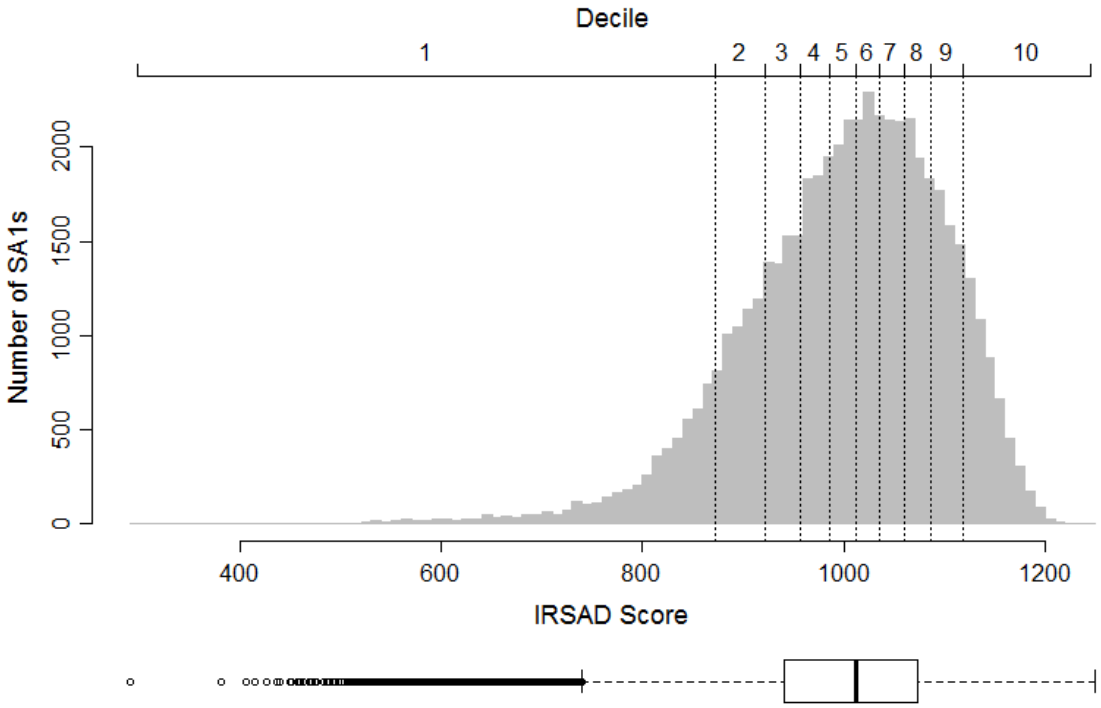


The decile cut-offs (marked along the top axis) show that there is little difference in the scores of SA1s in the middle deciles. This means that the characteristics of SA1s in the middle deciles are not likely to vary much. The discriminating power of this index lies particularly in the lower end, i.e. for identifying relatively disadvantaged SA1s.

4.5.2 *Index of Relative Socio-Economic Advantage and Disadvantage*

Looking at figure 4.12, we can see that the IRSAD has a long left tail, though shorter than the IRSD. The scores range from 300 to around 1250 (a lower range than the IRSD). The right slope is not as steep as the IRSD, meaning the scores of SA1s in the upper deciles are more spread out. This index is more appropriate than the IRSD for users who want to compare the entire range of areas, rather than focussing on disadvantaged areas.

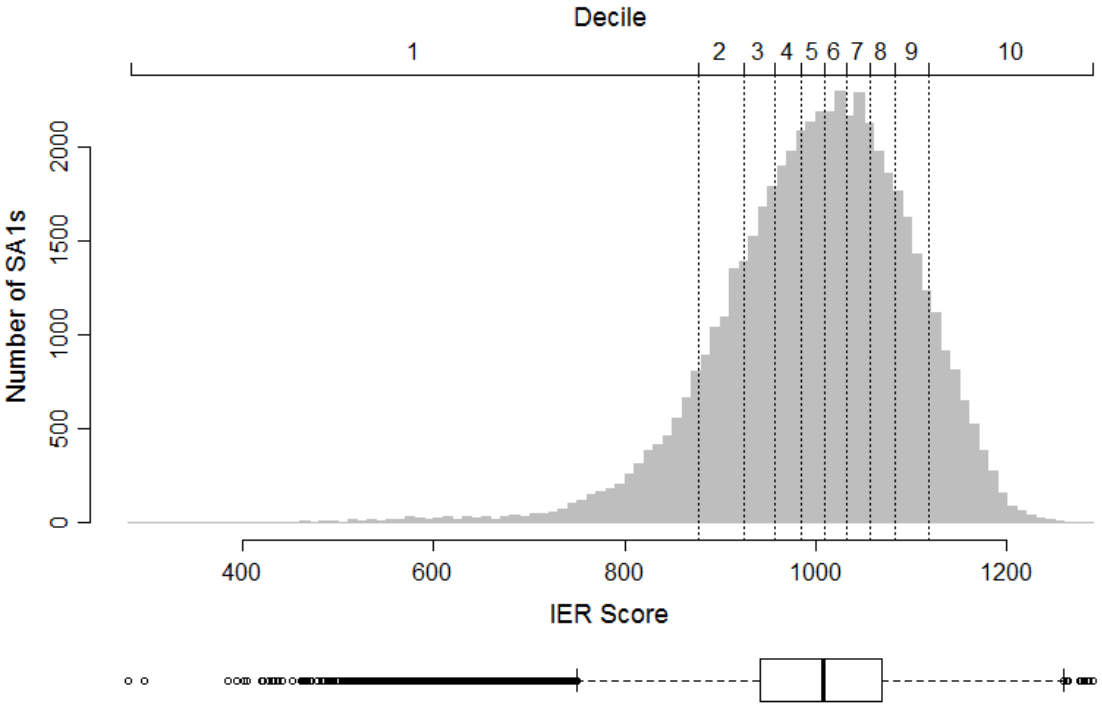
4.12 IRSAD score distribution



4.5.3 *Index of Economic Resources*

Figure 4.13 shows that the IER is the most normally distributed of the four indexes, as was observed for SEIFA 2006. The scores range from around 280 to around 1290. The left tail is very long, similar to the IRSD, and there is also a reasonable spread amongst SA1s in the upper deciles, as evidenced by the gentle right slope. This index can be used to compare all areas in terms of their access to economic resources.

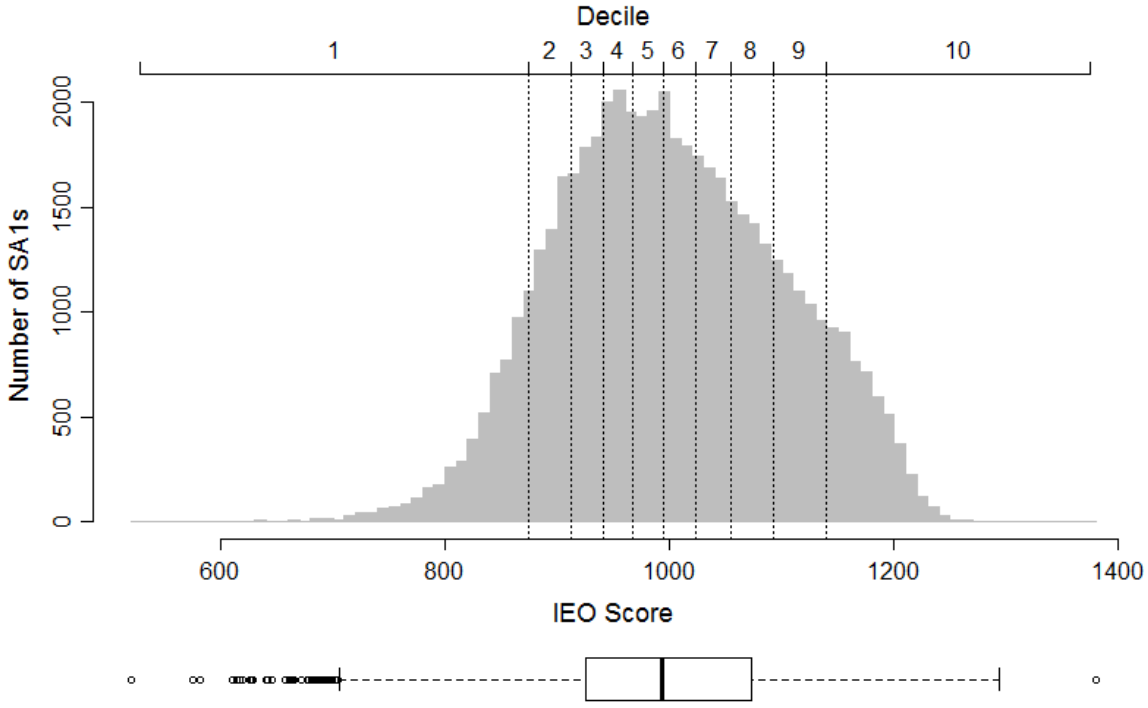
4.13 IER score distribution



4.5.4 *Index of Education and Occupation*

The IEO values range from around 530 to around 1380, as shown in figure 4.14 below. The distribution is slightly right-skewed, and the scores of areas in the upper deciles are more spread out than the scores of areas in the lower deciles. The right slope is the widest of the four indexes. This index can be used to compare the entire range of areas in terms of people's educational qualifications and vocational skills.

4.14 IEO score distribution



4.6 Basic output: scores, ranks, deciles, and percentiles

The output is presented in spreadsheets and is available for free at www.abs.gov.au under the catalogue number 2033.0.55.001.

The basic concepts to grasp before using the spreadsheets are described below. While reading this section, it is helpful to view the spreadsheets simultaneously.

4.6.1 Scores

The scores are a weighted combination of the selected indicators of advantage and disadvantage which have been standardised to a distribution with a mean of 1000 and standard deviation of 100. An area with all of its indicators equal to the national average will receive a score of 1000. The score for an area will increase if an area has: an indicator of advantage that is greater than the national average; or an indicator of disadvantage that is less than the national average. Conversely, the score for an area will decrease if an area has: an indicator of disadvantage that is greater than the national average; or an indicator of advantage that is less than the national average. Indicators which are further away from the national average have a larger impact on the score. As an example, we would expect that an area with an index score of 980 would have most of its indicators closer to the national average compared to an area with an index score of 900.

For areas larger than SA1, the scores are a population weighted average of constituent SA1 scores, as described in Step 11 of Section 4.3.

It is important to remember that the scores are an ordinal measure (discussed in more detail in Section 6.1.2), so care should be taken when comparing scores. For example, an area with a score of 500 is not twice as disadvantaged as an area with a score of 1000, it just had more markers of relative disadvantage.

4.6.2 Ranks, Deciles, and Percentiles

Using the scores, other measures are derived that are easier to interpret and more appropriate to use in many situations. The ABS derives ranks, deciles, and percentiles and includes these in the output spreadsheets. These measures are defined below:

Rank – The areas are ranked in order of their score, from lowest to highest, with rank 1 representing the most disadvantaged area. Note that in the spreadsheets, rankings are provided on a national basis and also a state/territory basis. Note that the same set of scores is used for each ranking – the scores are not recalculated for each state/territory.

Deciles – All areas are ordered from lowest to highest score, the lowest 10% of areas are given a decile number of 1, the next lowest 10% of areas are given a decile number of 2 and so on, up to the highest 10% of areas which are given a decile number of 10. This means that areas are divided up into ten equal sized groups, depending on their score.

Percentiles – All areas are ordered from lowest to highest score, the lowest 1% of areas are given a percentile number of 1, the next lowest 1% of areas are given a percentile number of 2 and so on, up to the highest 1% of areas which are given a percentile number of 100. This means that areas are divided up into one hundred equal sized groups, depending on their score.

Sometimes deciles and percentiles are referred to generally as quantiles. Other commonly used quantiles include quintiles and quartiles, although we have not included these in the output spreadsheets. They can be easily derived using the percentiles.

When deciding which quantile to use in an analysis, it is worth considering the distribution of scores within each quantile. For example, observing figures 4.11 to 4.14, it is clear that decile 1 has a large spread of scores compared to the other deciles. This is worth noting when using deciles, particularly if there is specific interest in the characteristics of areas in decile 1.

4.7 Geographic output levels for SEIFA 2011

The primary unit of analysis and the smallest area for which the indexes are available is the Statistical Area Level 1 (SA1). This is the recommended unit of analysis for SEIFA 2011. We recognise that there are instances where users want index scores for some larger geographic areas, and hence we have produced these larger area scores by taking population-weighted averages of constituent SA1 scores.

For areas larger than SA1, we have also provided information in the output spreadsheets that show the distribution of SA1 index scores within larger areas. This enables users to consider the socio-economic diversity that can exist within a larger area.

Table 4.15 below summarises the output available at the different geographic levels.

4.15 Geographic output summary for SEIFA 2011

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

Note – for the geographies larger than SA1, and not in the ASGS (e.g. SLA), a best fit correspondence of SA1s to the larger geographies was used.

The output spreadsheets contain specific references to the ABS publications from which the geography classifications and correspondences have been sourced.

5. VALIDATION OF THE INDEXES

Once the indexes are calculated, they are checked to ensure that they are measuring the desired concept and that the results generally make sense. This validation is important to establish the credibility of the indexes and identify any issues that may have been missed in the construction of the indexes. This section of the paper describes the methods used to validate the indexes. These methods are:

- visual analysis of thematic mapping tools,
- consultation with ABS Regional Offices to validate the indexes against local knowledge,
- investigation of the correlations between the four indexes,
- identification of the most influential areas and variables in the index creation process, as identified by sensitivity analysis,
- comparison of SEIFA 2011 rankings with 2006 rankings, and
- identification of the drivers of change from SEIFA 2006 to 2011.

The following subsections provide details on each of the points listed above. Note the analysis refers to the SA1 level indexes.

Some validation of index scores for areas larger than SA1 was also conducted. This is described in Section 5.7.

For past releases of SEIFA, the ABS has convened a group of external experts to validate the methodology and variable selection immediately prior to release. For SEIFA 2011, this validation step was omitted because the methodology of SEIFA has been well established over a number of Censuses, and the type of information collected on the Census did not change between the 2006 and 2011 Censuses – effectively the same questions were asked. It should be noted that some informal user consultation was conducted prior to the production phase of SEIFA 2011, and as mentioned previously, there has been much external input into SEIFA in the past.

5.1 Thematic mapping tool

A mapping tool was used that enabled the indexes to be presented as thematic maps that overlay interchangeable backgrounds. The backgrounds could be different types of street maps and also satellite images. This enabled consideration of contextual information when assessing whether index scores were realistic.

The mapping tool was used to determine the extent to which the spatial distribution of relative advantage and disadvantage made sense. It was also used to investigate specific SA1s that required individual attention.

Often, the mapping tool shed light on characteristics of an area that could not be determined from looking at the Census data alone. For instance, satellite imagery revealed that many of the most advantaged SA1s tend to contain new housing developments in expensive areas of major urban centres. This is because such developments are associated with people who have high incomes and skilled jobs.

5.2 ABS Regional Office validation

Lists of the top and bottom ten SA1s for each index in each state/territory were sent to every ABS regional office for a plausibility check, based on local knowledge.

A selection of regions was also inspected to see if the spatial distribution of index scores made sense.

Most of the manually inspected SA1s were assessed as expected or at least understandable. A small number of areas were found to be surprisingly low or high by the regional offices, however these areas were scrutinised with respect to the Census variables and were found to have advantaging or disadvantaging characteristics that justified their calculated index scores.

The ABS Regional Offices made use of the thematic mapping tool discussed in Section 5.1 for their validation tasks.

5.3 Relationships between the indexes

We examined SEIFA for internal consistency by looking at the correlations between the indexes. Table 5.1 shows the rank correlation matrix. All correlations are in the expected directions and show significant relationships. The IRSD is very highly correlated with the IRSAD (0.98). This correlation is higher than was observed for SEIFA 2006 (0.94).

5.1 Spearman's rank correlation matrix

Index	*IRSD*	*IRSAD*	*IER*	*IEO*
IRSD	1.00			
IRSAD	0.98	1.00		
IER	0.85	0.83	1.00	
IEO	0.79	0.85	0.49	1.00

The indexes which measure specific dimensions of advantage and disadvantage (IER and the IEO) have a lower correlation with the other indexes. The IER includes variables chosen to capture high and low wealth, which are not included in the other indexes. The IEO focuses solely on educational qualifications, employment and vocational skills.

The IER and the IEO are positively correlated, but the correlation is much weaker than between the other indexes (0.49). There is a significant difference between the concepts measured by these two indexes, and they do not share any common variables. This correlation has also dropped from the equivalent value in SEIFA 2006.

5.4 Influential areas and variables

Based on recent research conducted by Radisich and Wise (2012), we adopted an additional validation method that assesses the sensitivity (robustness) of the indexes to the exclusion of particular variables and areas. This type of analysis is helpful on a number of fronts:

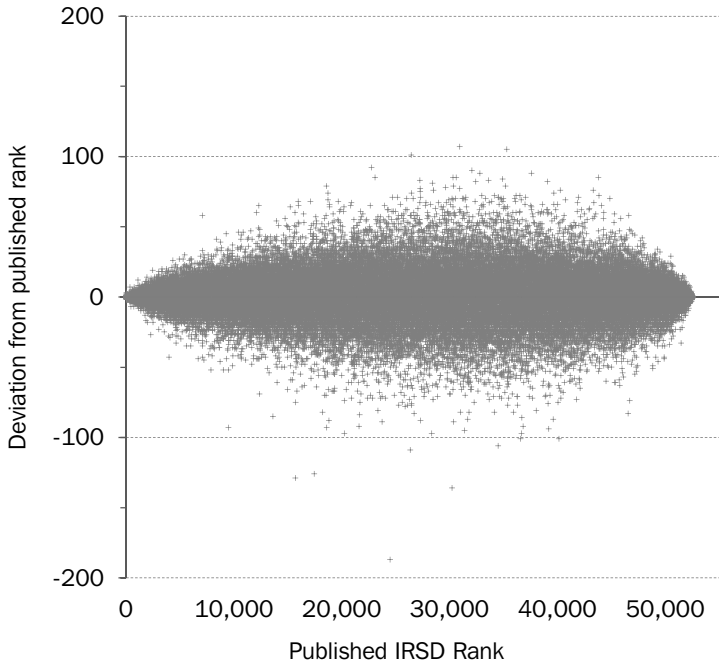
- It can identify issues with the appropriateness of exclusion rules. Influential areas may be those with low populations or high non-response, and sensitivity analysis can help detect whether such areas are being excluded.
- It can identify issues with the way in which variables are defined, checking if they are specified in a way that means their contribution to the index is realistic and not distorted.
- Users are provided with a general idea of the robustness of the indexes, and thus can use the indexes with more confidence.

5.4.1 Robustness with respect to influential and atypical areas

The analysis of the most influential areas showed that, due to the large number of SA1s, it is rare that any single SA1 or group of SA1s exerts a high influence on the variable weights and ultimately the index rankings. To illustrate this point, figure 5.2 presents the deviations from the published 2011 IRSD ranks after the 100 most influential SA1s[9] have been removed from the calculation of the index variable weights. All deviations in rank are within 200 of the published index rank, with most being less than 100. This indicates that the index is robust with respect to atypical and outlying areas. Analysis on the other three indexes yielded the same conclusions.

9 Influential areas are identified using the influence function described in Radisich and Wise (2012). The influence function is a linear approximation to the actual change in variable weights resulting from removing an area from the index.

5.2 Impact on published IRSD ranks of removing the 100 most influential SA1s from the calculation of the variable weights

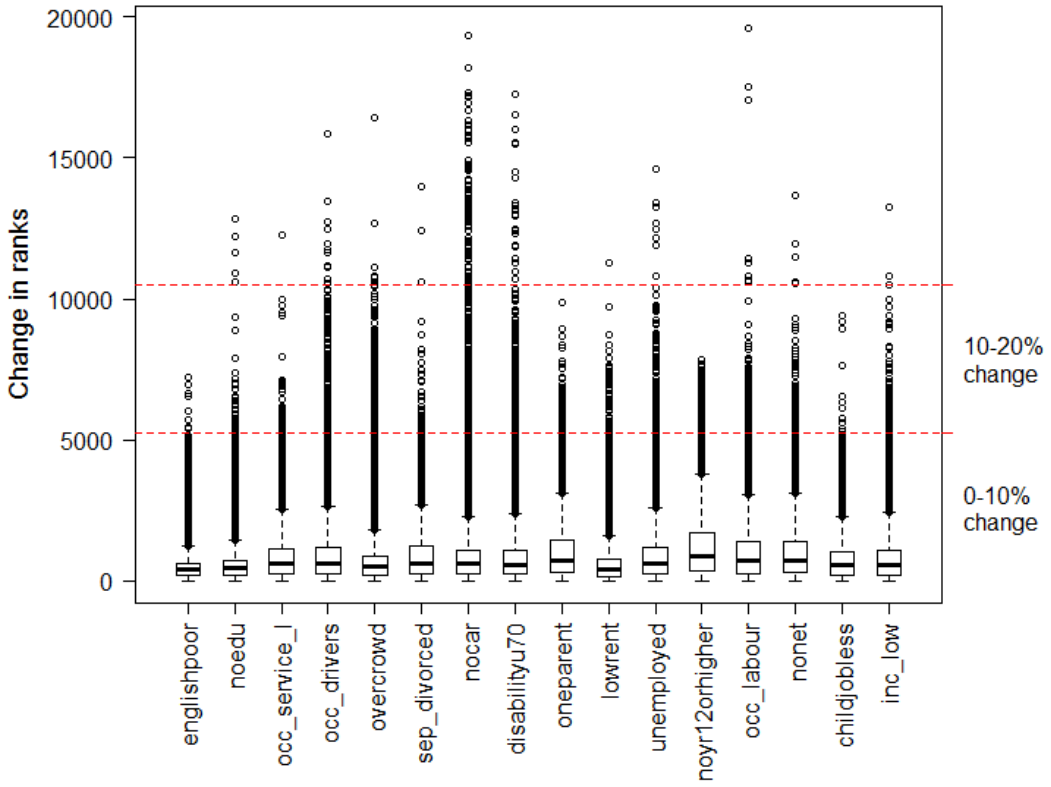


5.4.2 Robustness with respect to variable inclusion

In order to assess the extent to which an index is influenced by a given variable, we developed the following method. For each index, we dropped one variable and created an alternative index (running PCA again) using the remaining variables. We compared the two sets of rankings produced by the alternative indexes. The change in ranks for all SA1s when a particular variable is removed is deemed as a way of gauging how influential a particular variable is on the index. We applied this process for all variables in an index, and for all indexes. Figure 5.3 shows the distribution of change in ranks for each variable for the IRSD.

Figure 5.3 gives the user an indication of how much influence each variable has on the IRSD. When interpreting the box plots, it is important to remember that the black circles represent outliers in the distribution and only make up a small percentage of the change in ranks.

5.3 Distribution of absolute change in ranks by variable for the IRSD



Looking at figure 5.3, the index is generally robust to removing a variable, with the vast majority of SA1s having changes under 10 per cent of ranks. It is apparent that there are a small percentage of SA1s whose rankings are quite sensitive to the removal of particular variables from the index. These SA1s tended to lie around the middle of the index score distribution (in deciles 3 to 8), for each index.

Appendix D shows graphs equivalent to figure 5.3 for the other three indexes.

5.5 Comparing 2006 and 2011 rankings

Direct comparisons between 2006 and 2011 SEIFA rankings are made difficult by the substantial changes to ABS geography coding between the 2006 and 2011 Censuses. In order to compare 2006 CDs and 2011 SA1s, we ascertained which 2006 and 2011 areas are legitimately comparable. SA1s and CDs are independent small area classifications: CDs were based on collector workloads while SA1s are designed for optimising statistical output.

Since CDs and SA1s are so different, it is extremely difficult to concord them to compare areas. There are no finer Census output areas to aggregate for approximate comparisons, but there is a 'poor' quality concordance of 2006 CDs to 2011 SA1s (the poor classification was designated by the ABS Geography section). This concordance was used to select only those areas where the 2006 CD was entirely within a 2011 SA1, and that the SA1 uniquely comprised one entire CD. This resulted in 3,413 comparable areas, or 6.49% of all SA1s and 9.11% of all included 2006 CDs. There was a reasonably even spread across 2006 and 2011 SEIFA index deciles of those comparable CDs and SA1s, with no systematic bias evident in the areas we analysed.

Table 5.4 shows the movements in percentiles of the comparable CDs and SA1s.

5.4 Percentile changes for comparable CD–SA1s from SEIFA 2006 to SEIFA 2011 (a)

Index	*0 to 10 percentiles (% of areas)*	*11 to 20 percentiles (% of areas)*	*21 to 50 percentiles (% of areas)*	*More than 50 percentiles (% of areas)*
IRSD	73.2%	19.3%	7.2%	0.3%
IRSAD	73.9%	18.7%	7.1%	0.2%
IER	66.3%	20.5%	12.7%	0.6%
IEO	77.6%	16.3%	5.7%	0.4%

(a) Analysis limited to those CDs and SA1s identified as comparable (3,413 areas).
Note – Rows may not sum to 100% due to rounding error.

The results in table 5.4 are broadly similar to past Censuses in terms of changes in comparable areas. However, this time we had far fewer comparable areas due to the change from CD to SA1.

Across all indexes, table 5.4 shows that between 87% (IER) and 94% (IEO) of comparable areas changed by less than 20 percentiles. This suggests that the vast majority of similar geographic areas only changed a small amount relative to their 2006 ranking. Some of the outliers in the comparable area percentile movement analysis were inspected for validation purposes. General observations on the top ten percentile differences for each index are listed following:

- IRSD – large decreases in several variables often led to the differences observed in the IRSD. Other big differences were observed for areas with new housing developments on previously unoccupied land.
- IRSAD – increases in the proportion of INC_LOW and drops in the proportions for HIGHMORTGAGE and HIGHRENT (either because of the new denominator and/or intercensal changes) contributed to the large changes observed.

- IER – redefining the household tenure variables to reduce volatility has resulted in the large changes observed for the IER. Dropping the HIGHRENT variable due to a low loading on the index in SEIFA 2011 also affected the areas with the largest percentile differences. HIGHRENT had a loading of 0.55 in the 2006 IER.
- IEO – shifts in occupation from low to high skill appears to be the major factor in most IEO changes. The combination of NOYEAR12 and NOQUAL into NOYEAR12ORHIGHER would also have an effect.

5.6 Drivers of change from SEIFA 2006 to 2011

Apart from the direct comparison of areas described in the previous section, it is worth highlighting some general factors that contributed to areas possibly moving rank between 2006 and 2011. This is briefly discussed below for each index.

IRSD

In terms of variable composition, there have been some changes from 2006. A new variable relating to children living in jobless families was included in the 2011 IRSD for the first time. The low equivalised income variable is the highest loading variable in the 2011 index, and is also the variable whose loading has changed the most since 2006. Note that we made an alteration to the definition of this variable.

Proportions of households paying low rent has increased markedly and is also more important in 2011 for measuring our concept of disadvantage. A variable measuring the proportion of dwellings renting from a government organisation (RENT_SOCIAL) was not considered for SEIFA 2011 because of a high correlation with households paying a low amount of rent after redefining the household tenure variables. It is noted that the variable was beholden to differences in state and territory government policy on social housing. Further analysis of 2011 Census data revealed a large proportion of households that rent from a government or community organisation are in the low rent category anyway.

In general, the volatility in values for the household tenure variables has been reduced by re-considering them in light of a common denominator based on all applicable households in an area rather than only those households renting or with a mortgage (see Section 3.3.5).

The use of the HEAP highest level of educational attainment Census variable when deriving the education variables is new for SEIFA 2011. This has eliminated overlap between people who may have, for example, left school before year 12 but obtained a degree later in life.

IRSAD

In terms of variable composition, the IRSAD has undergone the largest change since 2006. The variable changes mentioned above in the IRSD also all apply to the IRSAD, namely the inclusion of the variable concerning children in jobless families and the re-definition of the education and household tenure variables. Additionally, the broadband internet access variable has been dropped due to changing prevalence of internet access. Analysis indicated it was no longer a suitable indicator of advantage.

IER

The IER has not been changed from a candidate variable list standpoint. The only difference is the re-consideration of the household tenure variables, as mentioned above.

IEO

The IEO has undergone minimal changes since 2006. The education and occupation variables are similar, with the only difference being the use of the HEAP highest level of educational attainment Census variable when deriving the education variables, as discussed above.

5.7 Validation of higher level area indexes

Most of the effort on validation was focussed on the SA1 level indexes because SA1s are the primary unit of analysis and indexes for higher level areas (e.g. SA2) are population weighted averages of the SA1 scores. However, we conducted basic validation checks on any higher level area indexes that we produced. This mainly comprised of viewing thematic maps of the indexes using the mapping tool described in Section 5.1, and also performing some basic comparisons with SEIFA 2006.

6. USING AND INTERPRETING SEIFA

This section provides some advice and information to assist in the appropriate use of SEIFA, and to help users gain the most value from the product.

6.1 Broad guidelines on appropriate use

Before using SEIFA, it is important to be aware of some issues relating to the interpretation of the indexes – these issues were briefly mentioned in Section 1.5. With this in mind, this section presents some broad guidelines for using SEIFA.

6.1.1 Area level indexes

The indexes are assigned to areas, not to individuals. They indicate the collective socio-economic characteristics of the people living in an area. A relatively disadvantaged area is likely to have a high proportion of relatively disadvantaged people. However, such an area is also likely to contain some people who are relatively advantaged. When area level indexes are used as proxy measures of individual level socio-economic advantage and disadvantage, many people are likely to be misclassified. This is known as the ecological fallacy. Wise and Mathews (2011) conducted an investigation into the extent of this issue as it relates to SEIFA.

6.1.2 Ordinal indexes

As measures of socio-economic level, the indexes are best interpreted as ordinal measures. They can be used to rank (order) areas, and are also useful to understand the distribution of socio-economic conditions across different areas. For example, the distribution of scores shown in section 4.5 shows many areas in the middle of the distribution tend to have similar socio-economic conditions compared to areas in the extremes of the score distribution.

Also, the index scores are on an arbitrary numerical scale. The scores do not represent some quantity of advantage or disadvantage. For example, we cannot infer that an area with an index value of 1000 is twice as advantaged as an area with an index value of 500.

For ease of interpretation, we generally recommend using the index rankings and quantiles (e.g. deciles) for analysis, rather than using the index scores. Index scores are still provided in the output, and can still be used by more technically adept users.

For more information on index scores, rankings, and quantiles, see Section 4.6.

6.1.3 Importance of the underlying variables

Each index is constructed based on a weighted combination of selected variables. The indexes are dependent on the set of variables chosen for the analysis. A different set of underlying variables would result in a different index. At the same time, because of the large number of variables in each index, removing or altering one variable will not usually have a large effect – see Section 5.4.2. Each variable set was selected based on the particular aspect of socio-economic advantage and disadvantage being captured for that index (e.g. economic resources). The list of potential variables was constrained by what was available from the Census data.

Users should consider the aspect of socio-economic advantage and disadvantage in which they are interested, and examine the underlying set of variables in each index (see Sections 3 and 4). This will allow them to make an informed decision on whether an index is appropriate for their particular purpose. Section 6.2 provides some tips on choosing which of the four indexes to use.

6.1.4 Issues with longitudinal or time series analysis

The indexes are designed to compare the relative socio-economic characteristics of areas at a given point in time, not to compare individual areas across time (longitudinal analysis). When considering longitudinal or time series analysis using indexes from different Census years, there are a number of issues that need to be considered and that make the analysis very difficult to interpret:

- The constituent variables and variable weights for the index are likely to have changed.
- The boundaries and numbers of relevant small area(s) may have changed.
- The distribution of the standardised index values will have changed (e.g. a score of 800 does not represent the same level of disadvantage in different years).
- There are changes in the way the variables are defined. For example, from SEIFA 2006 onwards, the indexes have been calculated using the characteristics of an area's usual residents, rather than those of the people in the area on Census Night (as was done in earlier editions of SEIFA).

For these reasons, it can be very difficult to perform useful longitudinal or times series analysis, and it should not be attempted without due consideration of the issues.

If comparisons over time are being made, we recommend the use of quantiles (e.g. deciles) rather than ranks or scores.

6.2 Choice of index

Depending on the aim or context of the analysis, one of the SEIFA indexes may be more appropriate than the others. Below are some considerations to make:

- The concept and variables underlying each index before deciding whether SEIFA is suitable for a particular research question. The concepts behind each index are described in Section 2.2. The final variable lists for each index are in Section 4.4.
- The degree to which the four indexes are correlated with each other – this is discussed in Section 5.2.
- The IRSD ranks areas on a continuum from most disadvantage to least disadvantage, whilst the other three indexes (IRSAD, IER, IEO) rank areas on a continuum from most disadvantaged to most advantaged.
- The IRSD and IRSAD are more general measures in the sense that they are comprised of variables from a wider range of socio-economic dimensions. The IER and IEO are more targeted measures (narrower concepts).
- Simpler measures, such as income or employment status, may be more appropriate than SEIFA for some analysis. For an in-depth discussion on choosing a socio-economic measure, see ABS (2011f).

6.3 Using index scores for areas larger than SA1

As discussed in Section 6.1.1, the fact the indexes are area level measures means that they will mask some diversity at finer levels of disaggregation. In some applications of the indexes, it may be important to identify diversity of socio-economic characteristics within areas.

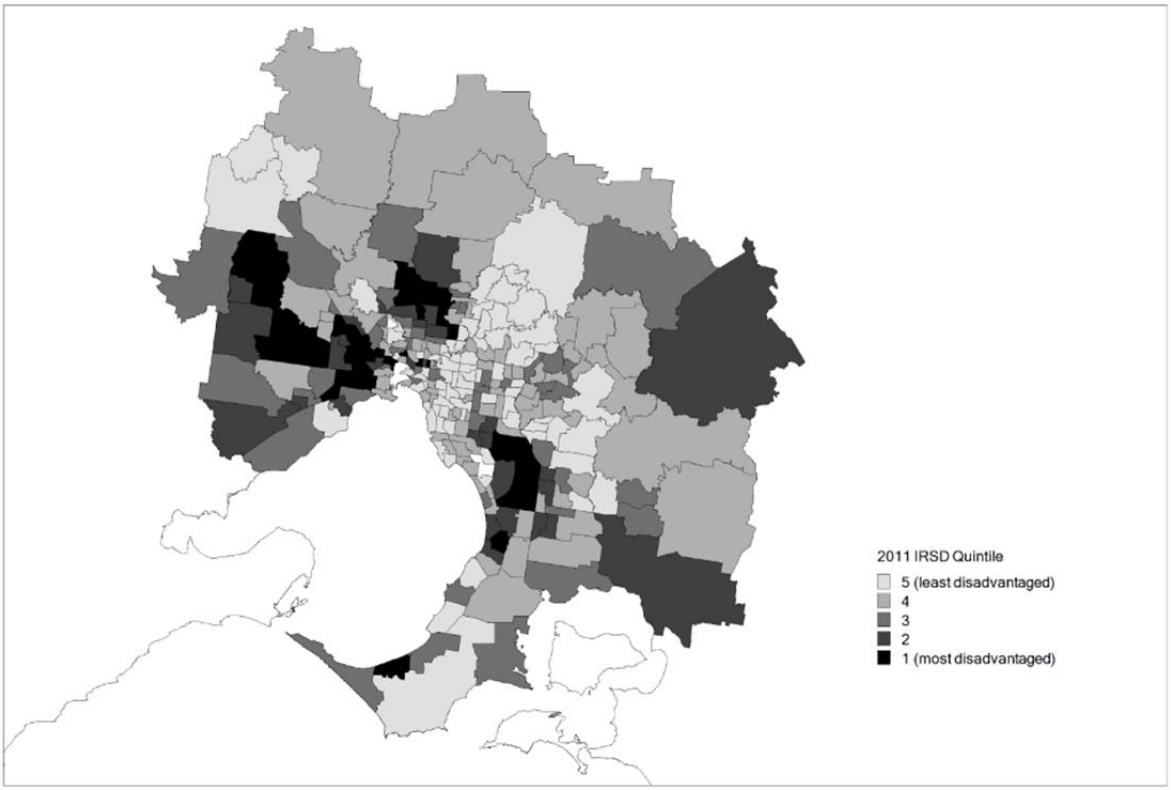
When using an index at a geographic level higher than SA1 (e.g. SA2, LGA), we do have some scope to assess the diversity within that area, by looking at its constituent SA1s. Radisich and Wise (2012) explored these possibilities, and the reader is strongly recommended to consider this reference if using SEIFA at geographic levels higher than SA1. Their paper also proposes an additional measure that can be used to identify diverse larger areas. The measure is called the SA1-concentration score and can identify the presence of disadvantaged SA1s within an overall advantaged larger area.

To enable the analyses described above, an additional type of output has been released for SEIFA 2011. For all geographic levels higher than SA1 for which index scores are released, the corresponding SA1 distributions within those areas have been presented in spreadsheets.

6.4 Mapping the indexes

As previously discussed in Section 5.1, thematic maps of the indexes are an excellent way of observing the spatial distribution of relative socio-economic advantage and disadvantage, and can also add contextual information (street maps, satellite images) to each index score. An example of a thematic map is shown below in figure 6.1.

6.1 Index of Relative Socio-Economic Disadvantage (2011) – SA2s in Greater Melbourne



For users with the appropriate technical skills and software, maps of SEIFA can be generated using geographic information systems. The indexes and appropriate boundary data can be downloaded from the ABS website.

For SEIFA 2011, we have made provision for users with limited technical knowledge to generate thematic maps, by releasing KMZ files that can be opened in Google Earth®, allowing the indexes to be viewed. The SEIFA web pages (on the ABS website) contain instructions on how to view these maps.

6.5 Using the indexes as contextual variables in social analysis

SEIFA index scores are commonly merged onto a person level dataset based on the area in which that person resides. The indexes can then be used to help investigate the relationship between disadvantage (or advantage) and other variables of interest e.g. health status (see ABS (2008)). This type of analysis can yield some very interesting findings, however it is important to interpret the findings correctly. Some interpretive issues are discussed below.

A SEIFA index refers to the area in which a person lives. It is a contextual variable. It is incorrect to say that a person is a very disadvantaged person if they live in a very disadvantaged area. It is true that living in a very disadvantaged area may disadvantage them to a certain extent, but it is possible that they are advantaged in many other respects such as having a good education and earning a high income, and are thus not typical of other residents in that area. The issue of diversity of individuals within areas is further investigated and discussed in Wise and Mathews (2011).

Related to the issue above, it is usually desirable to use the smallest geographic unit possible when merging an index to another dataset. In the case of SEIFA 2011, the SA1 is the smallest unit available, and thus if possible, SA1s should be derived on the dataset to which SEIFA scores are being appended.

6.6 Area-based quantiles versus population-based quantiles

In this paper the word 'quantiles' is used to collectively describe measures such as percentiles and deciles.

In the spreadsheets in which the indexes are presented, quantiles (percentiles and deciles) are presented in addition to the index scores and rankings, as described in Section 4.6. These quantiles are calculated based on dividing the number of areas into equal groups. These are called area-based quantiles.

An alternative way of defining the quantiles is to divide them into equal groups based on the number of people living in those areas. The quantiles would then contain an equal number of people (or at least as can be best achieved) in each group, rather than an equal number of areas. These are called population-based quantiles.

The ABS publishes area-based quantiles because they are easier to interpret, since SEIFA is an area-based measure. They also serve most analytical purposes.

There are some instances in which the use of population-based quantiles is appropriate. Users can create their own population-based quantiles using information already available in the output spreadsheets. As mentioned above, population-based quantiles can be difficult to interpret, so users should take care in how they are applied. The population-based quantiles represent groups of individuals who live in similarly ranked areas, as opposed to groups of similarly ranked individuals.

7. BACKGROUND INFORMATION TO INFORM ANALYSES

This section presents brief analyses on the relationships between SEIFA and some important classifying variables: age, states/territories, and remoteness.

These analyses are included in the technical paper because SEIFA does not directly include any of these variables in its composition, and a broad understanding of how SEIFA relates to these variables is beneficial to analyses using SEIFA.

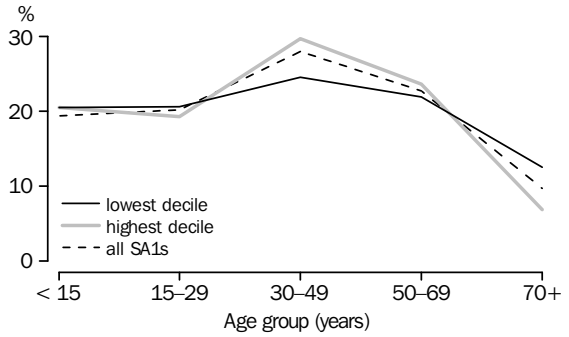
7.1 SEIFA and age

This section discusses the relationship between the indexes for an SA1 and the age of its residents. It also presents some analysis and discussion of some SEIFA variables that are influenced by age.

7.1.1 Comparing SEIFA with Age

Figures 7.1 to 7.4 below show the percentage of residents in five broad age groups, for areas in various SEIFA deciles. Figure 7.1 compares the highest and lowest deciles of the IRSD with all SA1s included in SEIFA. Figures 7.2 to 7.4 are the corresponding comparisons for the other three SEIFA indexes.

7.1 Index of Relative Socio-economic Disadvantage, % people by age group



Similar patterns are evident for the IRSD and the IRSAD. The 30–49 year age group is overrepresented in the highest decile, and underrepresented in the lowest decile, for both indexes. These observations are consistent with findings from SEIFA 2006, and make sense logically when we consider that people in the 30–49 year old age group are more likely to be in the workforce, earning relatively high incomes and with higher levels of education than at other times in the life course. Conversely, people aged 70 years and over are underrepresented in the highest deciles and overrepresented in the lowest deciles, in both indexes. This reflects the fact that this demographic is more likely to have lower incomes, lower mortgage or rental payments and fewer economic resources such as cars and large houses than younger age groups.

7.2 Index of Relative Socio-economic Advantage and Disadvantage, % people by age group

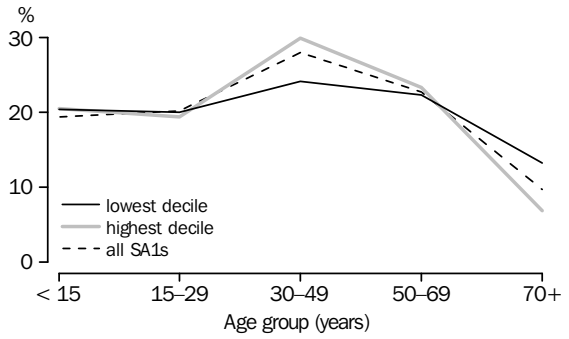
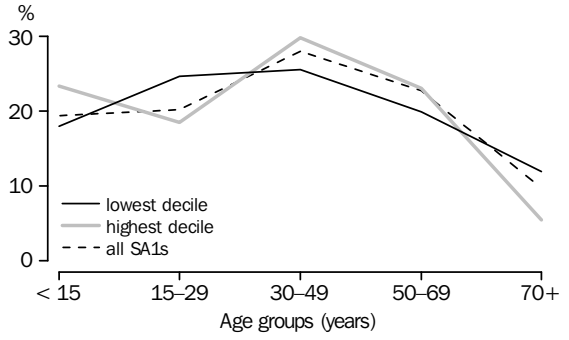


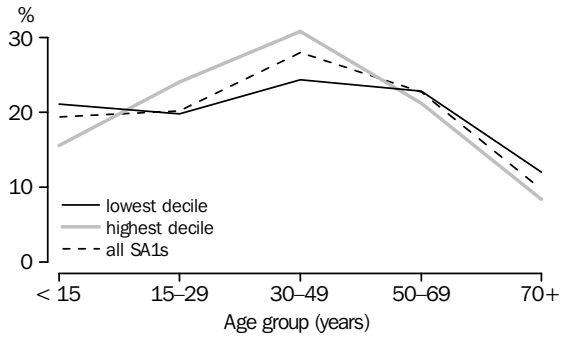
Figure 7.3 provides a comparison of the age structures of residents in the highest and lowest deciles of the IER. Similar to the other indexes, the 30–49 year age group is overrepresented in the most advantaged decile and underrepresented in the most disadvantaged decile. However, the age distribution for this index contains some differences to those of the other indexes. For example, the highest decile has a below average proportion of 15–29 year olds, and an above average proportion of 50–69 year olds. The converse is true for the lowest decile. This can be attributed to the fact that this index has a greater focus on wealth than the other indexes, and since wealth generally accumulates over the working life people around retirement age will generally have greater wealth than people over the age of 70 and younger people.

7.3 Index of Economic Resources, % people by age group



The relationship between age and the IEO is shown in figure 7.4. One notable feature of the age distribution for this index is the high proportion of 15–29 year olds in the highest decile. SA1s with many people in this age group are likely to have a high proportion of people studying at university. People under 15 are overrepresented in the lowest decile of this index, and underrepresented in the highest decile. Previous analysis from SEIFA 2006 has shown that areas with high proportions of families with dependent offspring tend to have more people without school or post-school qualifications, or working in lower skilled occupations.

7.4 Index of Education and Occupation, % people by age group

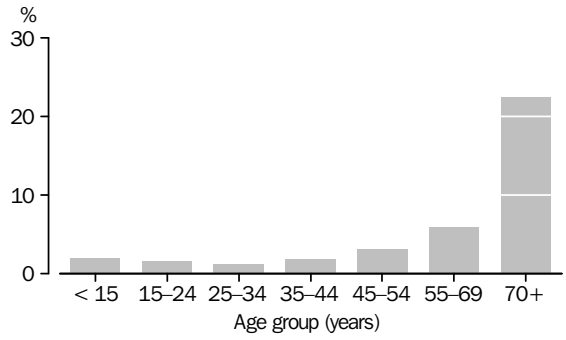


7.1.2 The effect of age on selected SEIFA variables

Some of the socio-economic indicators used in SEIFA are influenced by age or life cycle effects. For instance, if we consider the life course we generally expect to see education levels increase between the ages of 15 to 24 and subsequently remain fairly steady. Incomes usually increase with age up to retirement, where other income streams replace full-time work. Material resources like the number of vehicles we own and the number of bedrooms in our houses will be at their highest for families in the 35–44 age group, when they need larger dwellings and more cars to run a household. This section will examine the effect of age on some selected SEIFA variables.

As a first illustrative example, the proportion of people in various age groups needing assistance with core activities is shown below in figure 7.5. It is evident that the prevalence of disability for people aged 70 years and over is extremely high, relative to the other age groups. Similar results were observed from the 2006 Census. In practice, a variable measuring the proportion of all people in the SA1 with a disability would primarily indicate the proportion of elderly people. In order to refine our disability measure to capture socio-economic factors beyond age, we limited the SEIFA variable to the population aged under 70, as was done for SEIFA 2006. The choice of cut-off of 70 years of age was re-analysed to confirm that it was still an appropriate break in the age distribution to use.

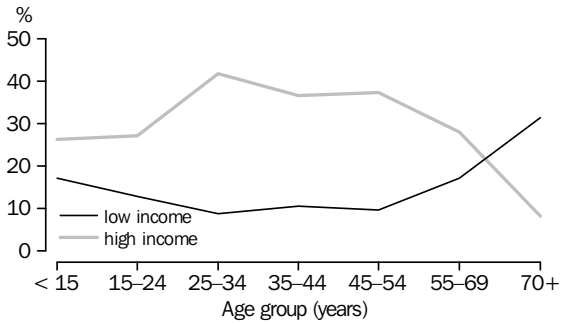
7.5 % People needing assistance with core activities, by age group



The equivalised income variables are also subject to life cycle effects, as shown in figure 7.6. It is evident that people of working age are likely to have higher equivalised incomes than older or younger people. People aged 70 years and over have the lowest equivalised incomes, and this is why we see the low and high income decile lines cross between the 55-69 and 70 years and over age group.

We did not adjust the income variables for age, which is in line with the approach taken for previous SEIFA releases. This is because income is a core aspect of socio-economic advantage and disadvantage for all age groups, and so we did not want to lose the life cycle effects. For example, if a SA1 had a large proportion of older people on low incomes, we wanted this aspect of economic disadvantage to be reflected in SEIFA.

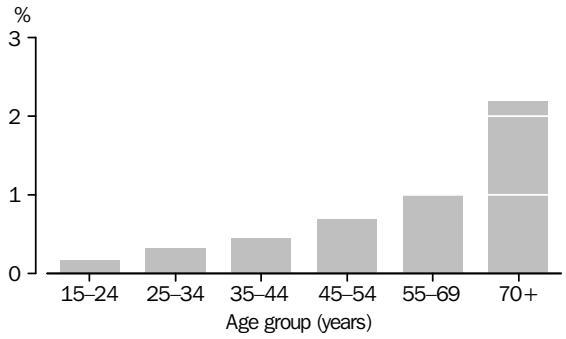
7.6 % People with high and low income, by age group (a)



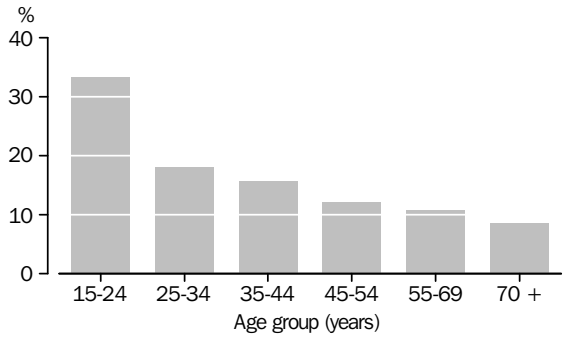
(a) The high income group is approximately the highest equivalised income quintile. The low income group is approximately the lowest equivalised income quintile, excluding negative and nil income.

Level of education is another age-related socio-economic characteristic measured in SEIFA. Figures 7.7 and 7.8 show proportions of people by age group with no educational attainment and whose highest level of educational attainment is year 12 schooling, respectively. The proportion of people with no educational attainment increases with age across all age groups. This is due to changes in social norms regarding school attendance over time. The proportion of people whose highest educational attainment is year 12 schooling is highest amongst those aged 15–24, however some of these people will likely still be studying for their first post-school qualification.

7.7 % People with no educational attainment, by age group



7.8 % People whose highest educational attainment is year 12, by age group



We did not adjust the education variables to account for age. The case for age-adjusting the education variables was not as straightforward as for the disability variable. There is no age after which lack of education drastically increases, as is the case for disability.

Investigations into the effect of age-standardising the education variables were performed before the release of SEIFA 2006, using a number of age ranges. The technique was considered inappropriate for SEIFA because the small population of some CDs meant only broad age ranges could be used, thus limiting the effectiveness of the standardisation. Additionally, CDs with very few people in any particular age range would have to have been excluded from SEIFA for consistency and to ensure sufficient data quality for index construction. Other practical arguments against implementing age standardisation for SEIFA include wanting to keep the SEIFA variables simple where possible, so the indexes are easier to interpret, and none of the previous editions of SEIFA have used standardised variables.

SEIFA is a general measure of relative socio-economic advantage and disadvantage that can be applied in many types of analysis. For some types of analysis, it may be useful to look at the age structure of an area in combination with SEIFA.

7.2 SEIFA and states/territories

This section discusses the relationships between SEIFA and the states/territories.

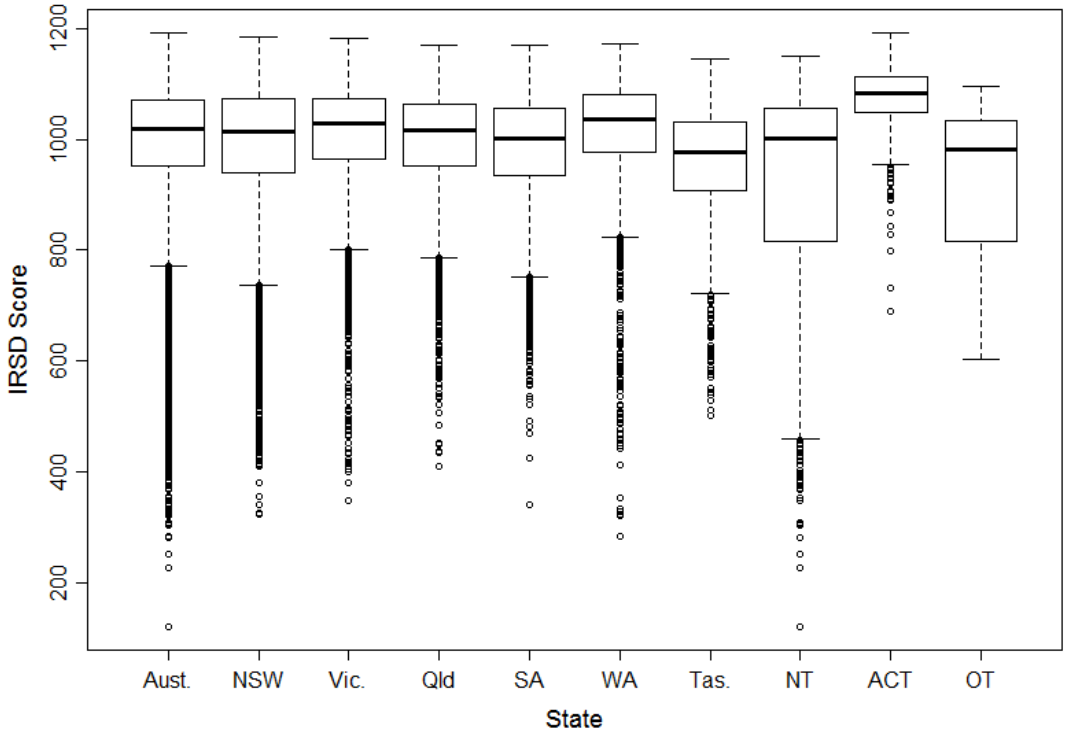
The base level of geography at which SEIFA was constructed was the SA1. Scores for larger geographies have been constructed by taking weighted averages of SA1 scores. However, as areas become larger, aggregated scores become less meaningful. For very large areas, it is more useful to look at the distribution of SA1 scores within each area. This section looks at the distribution of SA1 scores within each state and territory. The distributions of SA1 scores are presented in boxplots. Appendix C contains a description of how to interpret box plots.

Figures 7.9 to 7.12 below compare the distributions of the SA1 scores of the four indexes across the states and territories. It is evident that SA1s in Australian Capital Territory have a much higher median score for all four indexes than any other state or territory. Additionally, the ACT has an extremely high P25 for the IRSD, indicating that based on indicators of disadvantage alone, areas in the ACT rank much higher than areas outside of the ACT.

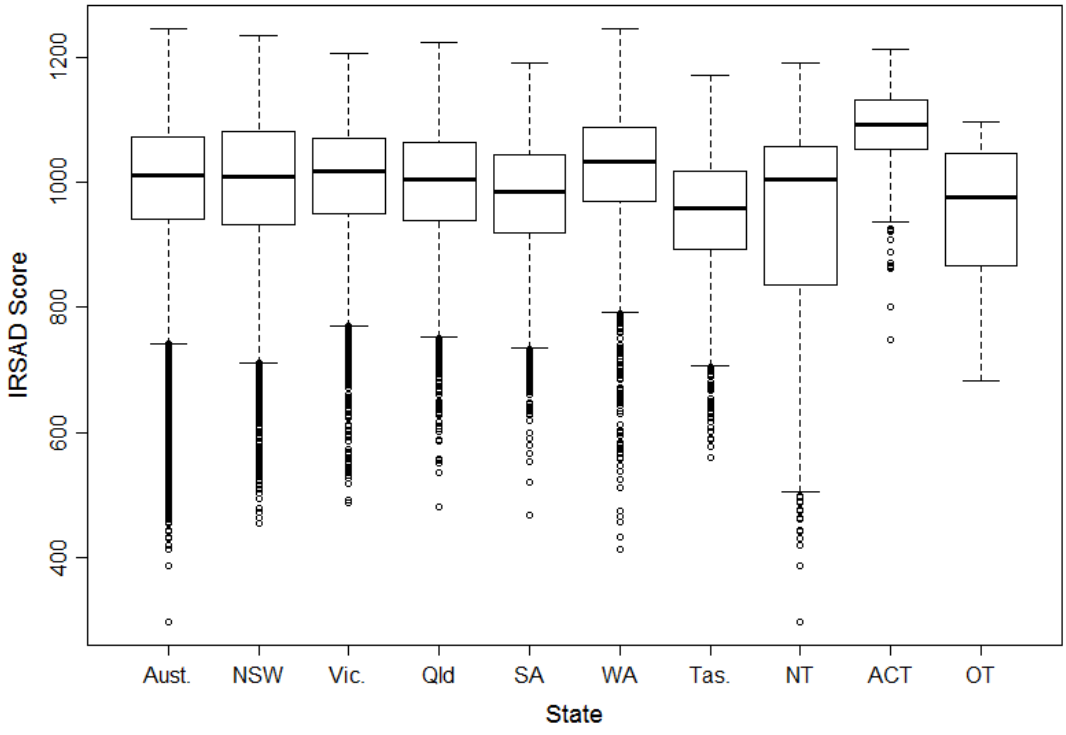
Also noteworthy are the distributions of scores for the Northern Territory for IRSD, IRSAD, and IER. While the medians for these indexes are not noticeably low for the Northern Territory, the P25 value and lower adjacent values are much lower than any other state or territory. This indicates a large skew in the Northern Territory SA1s towards disadvantaged scores.

Apart from the ACT and the Northern Territory, we can see that the distributions of scores for all four indexes are slightly higher in Western Australia and lower in Tasmania, when compared to the other states.

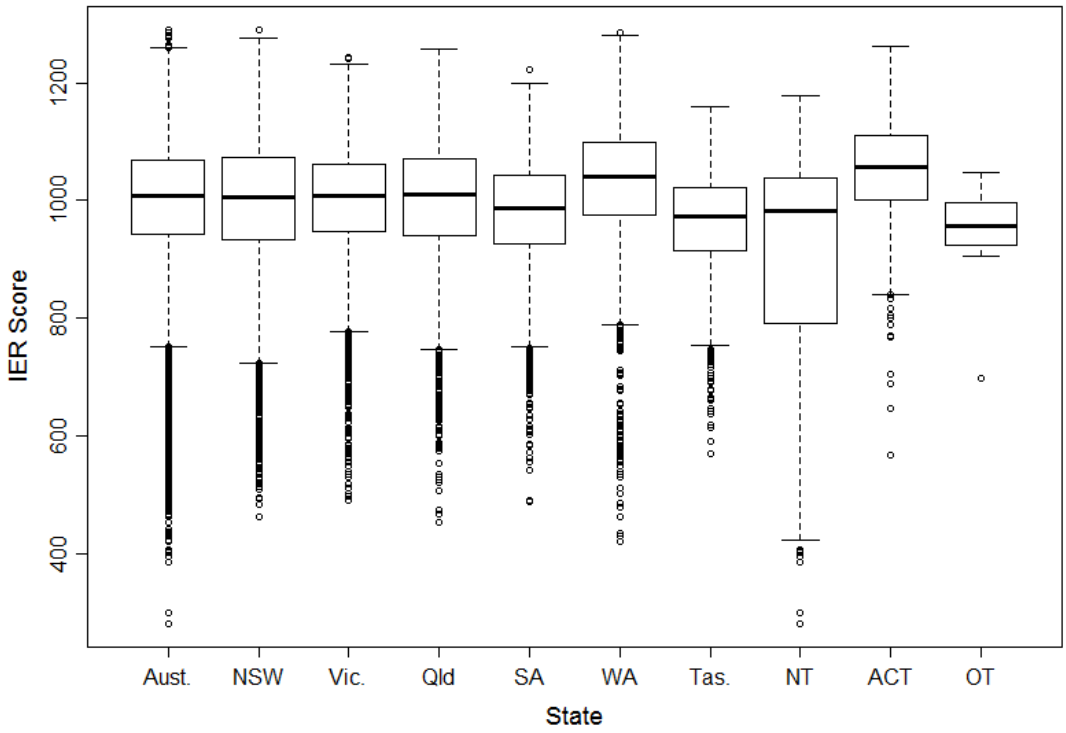
7.9 Distribution of IRSD SA1 scores by state/territory



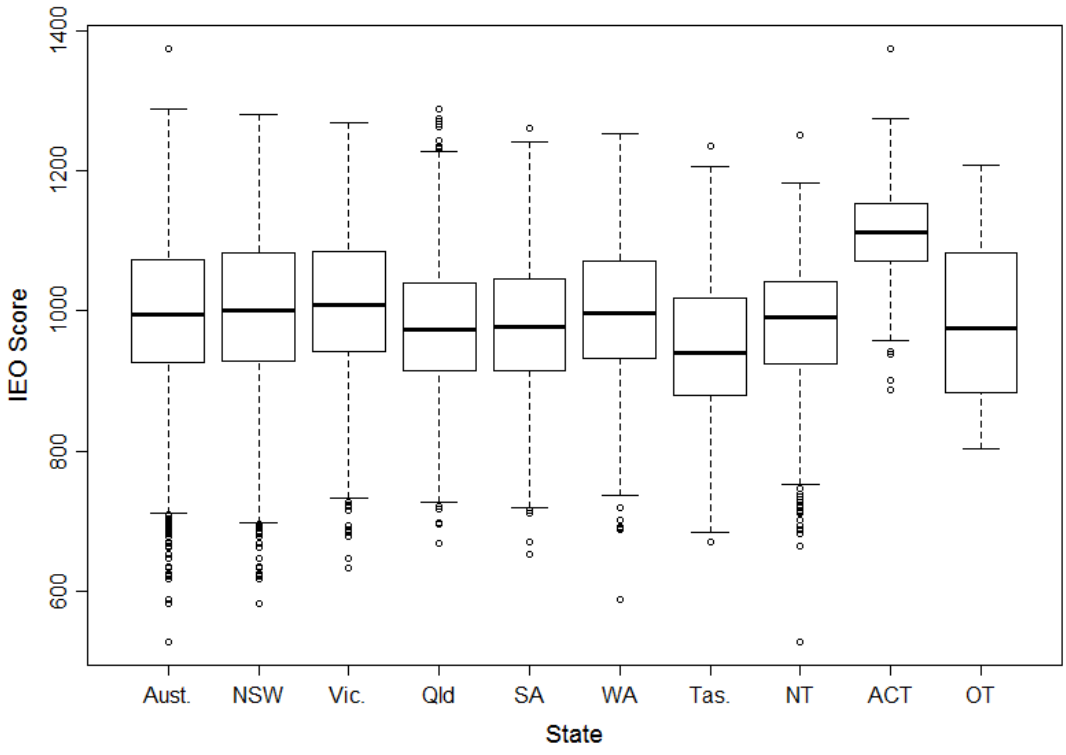
7.10 Distribution of IRSAD SA1 scores by state/territory



7.11 Distribution of IER SA1 scores by state/territory



7.12 Distribution of IEO SA1 scores by state/territory



7.3 SEIFA and remoteness

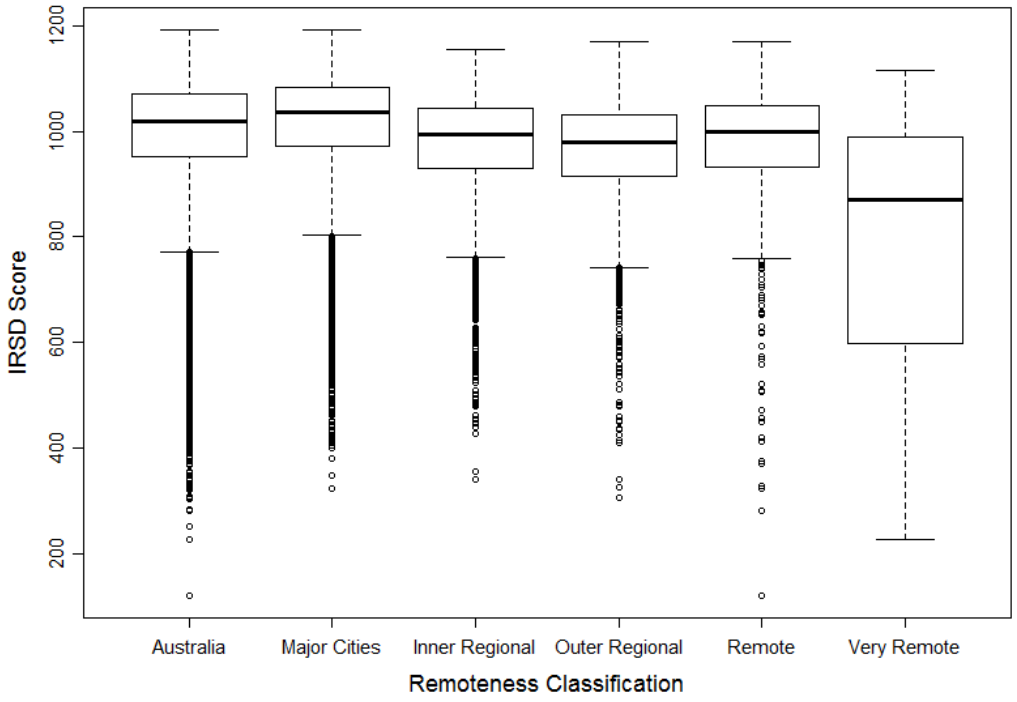
This section discusses the relationship between the SEIFA indexes and the remoteness of an area by looking at the distribution of SA1 scores within each ASGS remoteness category. This analysis gives some idea of whether certain categories experience more variability in the SA1 index scores. For more information on the remoteness classification, see ABS (2013).

Figures 7.13 to 7.16 are boxplots that have been used to show comparisons between the distributions of SA1 index scores for each of the five ABS remoteness categories: Major Cities of Australia, Inner Regional Australia, Outer Regional Australia, Remote Australia, and Very Remote Australia. The overall score distribution for Australia is provided as a reference. Appendix C contains a description of how to interpret box plots.

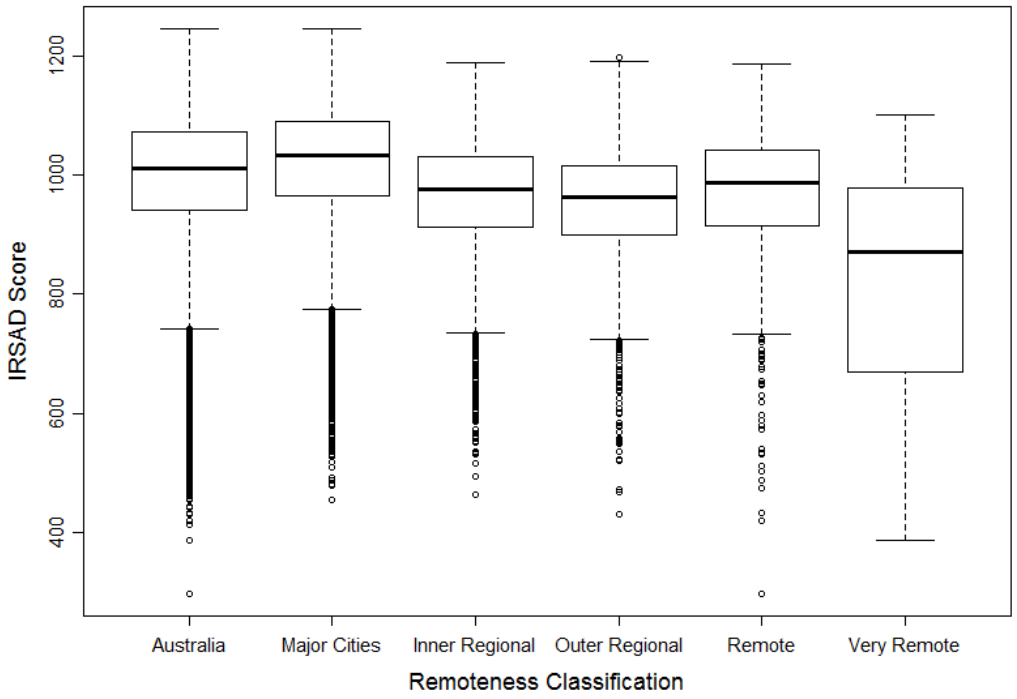
From figures 7.13 to 7.16, it is evident that 'very remote' SA1s have a lower median score for all four indexes, and a wider distribution of scores for the IRSD, IRSAD and the IEO. The long tail of low scores compared to the other remoteness categories is also clearly distinguished for 'very remote' SA1s. There is little distinction between the range and features of the score distribution for inner and outer regional areas, however major cities and remote areas tend to have slightly higher median scores. This is interesting considering the low median values across the four SEIFA indexes for the very remote SA1s. It should be noted that the broad conclusions discussed above relate to the distributions of SA1 scores. Each remote classification exhibits variability in the SA1 scores. Each remote classification contains SA1s that are relatively advantaged and SA1s that are relatively disadvantaged.

SEIFA is a general measure of relative socio-economic advantage and disadvantage that can be applied in many types of analysis. For some types of analysis, it may be useful to look at the remoteness classification of an area in combination with SEIFA.

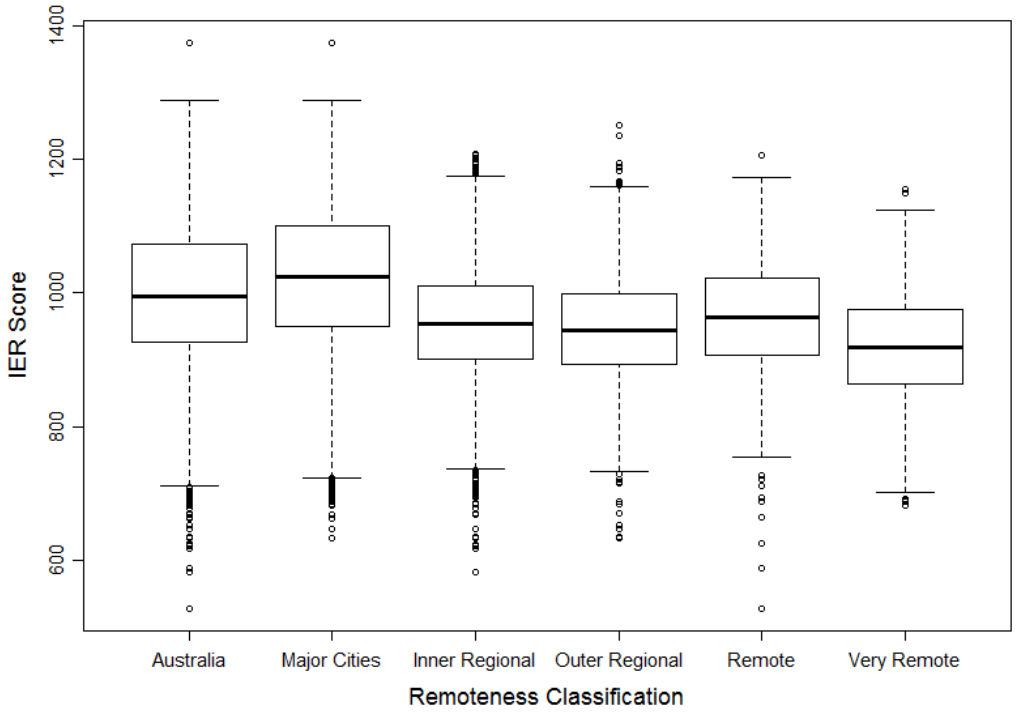
7.13 Distribution of IRSD SA1 scores by remoteness classification



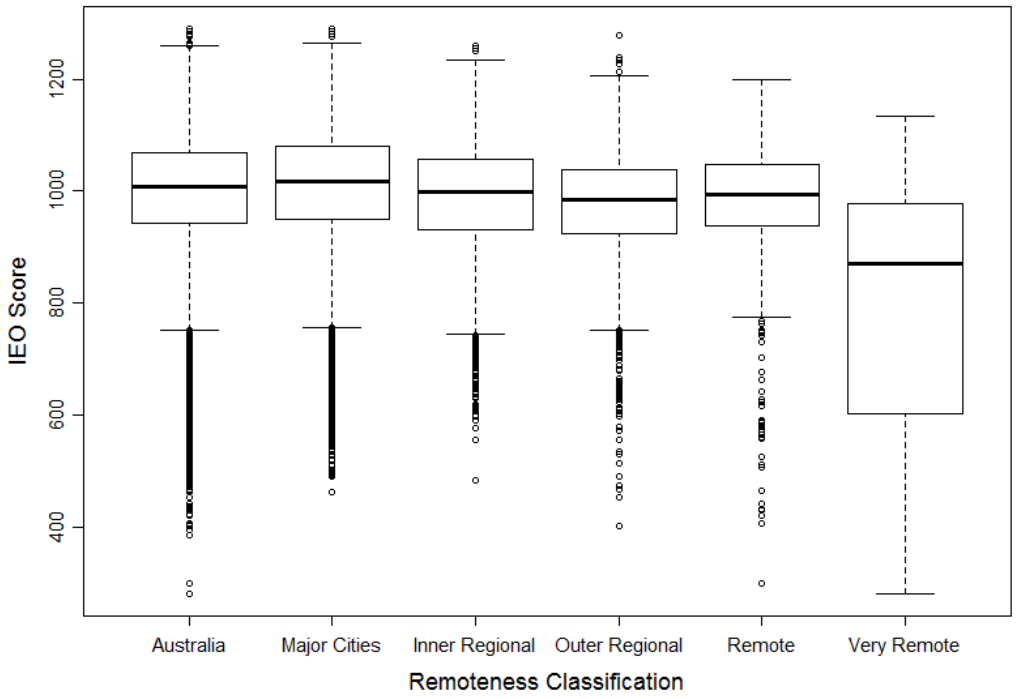
7.14 Distribution of IRSAD SA1 scores by remoteness classification



7.15 Distribution of IER SA1 scores by remoteness classification



7.16 Distribution of IEO SA1 scores by remoteness classification



8. CONCLUDING REMARKS

SEIFA 2011 is the latest version of SEIFA, a product that is released every five years, after the Census. This paper has covered much of the detail associated with the production of SEIFA 2011 and has provided advice on how to use it appropriately. For further information and assistance, please consult www.abs.gov.au.

Information on future releases of SEIFA and associated publications will be added to the website as it arises.

REFERENCES

Australian Bureau of Statistics (2006) *ANZSCO – Australian and New Zealand Standard Classification of Occupations, First Edition*, cat. no. 1220.0, ABS, Canberra.

—— (2007) *Australian Social Trends, 2007,* cat. no. 4102.0 ABS, Canberra.

—— (2008) *Socio-Economic Indexes For Area – Technical Paper, 2006*, cat. no. 2039.0.55.001, ABS, Canberra.

—— (2009) *ANZSCO – Australian and New Zealand Standard Classification of Occupations, First Edition, Revision 1*, cat. no. 1220.0, ABS, Canberra.

—— (2011a) *Census Dictionary, 2011,* cat. no. 2901.0, ABS, Canberra.

—— (2011b) *Household Income and Income Distribution, Australia, 2009–10,* cat. no. 6523.0, ABS, Canberra.

—— (2011c) *Education and Work Australia, May 2011*, cat. no. 6227.0, ABS, Canberra.

—— (2011d) *Housing Occupancy and Costs, Australia, 2009–10,* cat. no. 4130.0, ABS, Canberra.

—— (2011e) *Household Wealth and Wealth Distribution, 2009–10*, cat. no. 6554.0, ABS, Canberra.

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

—— (2012a) *Internet Activity, Australia, Jun 2012*, cat. no. 8153.0, ABS, Canberra.

—— (2012b) *Census of Population and Housing – Counts of Aboriginal and Torres Strait Islander Australians, 2011*, cat. no. 2075.0, ABS, Canberra

—— (2013) *Australian Statistical Geography Standard (ASGS): Volume 5 – Remoteness Structure, July 2011,* cat. no. 1270.0.55.005, ABS, Canberra.

Australian Social Inclusion Board (2009) *A Compendium of Social Inclusion Indicators: How's Australia Faring?*, ASIB, Canberra.

—— (2010) *Social Inclusion in Australia: How Australia is Faring*, ASIB, Canberra.

Joliffe, I.T. (1986) *Principal Component Analysis*, Springer–Verlag, New York.

O'Rourke, N.; Hatcher, L. and Stepanski, E.J. (2005) *A Step-by-Step Approach to Using SAS for Univariate and Multivariate Statistics, Second Edition*, SAS Institute Inc., Cary, NC.

Radisich, P. and Wise, P. (2012) "Socio-Economic Indexes For Areas: Robustness, Diversity Within Larger Areas and the New Geography Standard", *Methodology Research Papers*, cat. no. 1351.0.55.038, Australian Bureau of Statistics, Canberra.

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

APPENDIXES

A. VARIABLE SPECIFICATIONS

This appendix gives descriptions of each variable considered for inclusion in one of the 2011 indexes. The description of the variable proportion is followed by two bullet points; the first is a description of the numerator, the second is a description of the denominator. The square brackets contain specifications for creating the numerator/denominator from Census data items, according to the mnemonics used in the *Census Dictionary, 2011* (ABS, 2011a). The variables are arranged by socio-economic dimension.

Note that for convenience of presentation, the variable proportions are expressed as percentages.

Income variables

INC_LOW	% PEOPLE WITH STATED ANNUAL HOUSEHOLD EQUIVALISED INCOME BETWEEN $1 AND $20,799 (approx. 1st and 2nd deciles) • number of people living in classifiable occupied private dwellings with stated annual household equivalised income between $1 and $20,799 [HIED = 03–05] • number of people living in classifiable occupied private dwellings with stated household equivalised income [HIED = 01–12]
INC_HIGH	% PEOPLE WITH STATED ANNUAL HOUSEHOLD EQUIVALISED INCOME GREATER THAN $52,000 (approx. 9th and 10th deciles) • number of people living in classifiable occupied private dwellings with stated annual household equivalised income greater than $52,000 [HIED = 09–12] • number of people living in classifiable occupied private dwellings with stated household equivalised income [HIED = 01–12]

Education variables

ATSCHOOL	% PEOPLE AGED 15 YEARS AND OVER WHO ARE STILL ATTENDING SECONDARY SCHOOL • number of people aged 15 years and over who are still attending secondary school [AGEP > 14 and TYPP = 31, 32 33] • number of people aged 15 years and over (excluding educational institution attendance not stated) [AGEP > 14 and TYPP ne &&, VV]
ATUNI	% PEOPLE AGED 15 YEARS AND OVER AT A UNIVERSITY OR OTHER TERTIARY INSTITUTION • number of people aged 15 years and over at university or other tertiary institution [AGEP > 14 and TYPP = 50] • number of people aged 15 years and over (excluding educational institution attendance not stated) [AGEP > 14 and TYPP ne &&, VV]
CERTIFICATE	% PEOPLE AGED 15 YEARS AND OVER WHOSE HIGHEST LEVEL OF EDUCATION IS A CERTIFICATE III or IV QUALIFICATION • number of people aged 15 years and over with a certificate III or IV qualification [AGEP > 14 and HEAP = 51] • number of people aged 15 years and over (excluding highest level of education not stated) [AGEP > 14 and HEAP ne 001, @@@, VVV, &&&]

DEGREE — % PEOPLE AGED 15 YEARS AND OVER WHOSE HIGHEST LEVEL OF EDUCATION IS A BACHELOR DEGREE OR HIGHER

- number of people aged 15 years and over whose highest level of education is a bachelor degree or higher [AGEP > 14 and HEAP = 1–3]
- number of people aged 15 years and over (excluding highest level of education not stated) [AGEP > 14 and HEAP ne 001, @@@, VVV, &&&]

DIPLOMA — % PEOPLE AGED 15 YEARS AND OVER WHOSE HIGHEST LEVEL OF EDUCATION IS AN ADVANCED DIPLOMA OR DIPLOMA

- number of people aged 15 years and over whose highest level of education is an advanced diploma or diploma [AGEP > 14 and HEAP = 4]
- number of people aged 15 years and over (excluding highest level of education not stated) [AGEP > 14 and HEAP ne 001, @@@, VVV, &&&]

NOEDU — % PEOPLE AGED 15 YEARS AND OVER WHO HAVE NO EDUCATIONAL ATTAINMENT

- number of people aged 15 years and over whose highest level of education is no educational attainment [AGEP > 14 and HEAP = 998]
- number of people aged 15 years and over (excluding highest level of education not stated) [AGEP > 14 and HEAP ne 001, @@@, VVV, &&&]

NOYEAR12ORHIGHER — % PEOPLE AGED 15 YEARS AND OVER WHOSE HIGHEST LEVEL OF EDUCATION IS YEAR 11 OR LOWER

- number of people aged 15 years and over whose highest level of education is year 11 or lower (includes certificate I and II qualifications; excludes those still at secondary school) [AGEP > 14 and HEAP = 50, 52, 613, 621, 622, 067, 998 and TYPP ne 31, 32, 33]
- number of people aged 15 years and over (excluding highest level of education not stated) [AGEP > 14 and HEAP ne 001, @@@, VVV, &&&]

Employment variables

UNEMPLOYED — % PEOPLE (IN THE LABOUR FORCE) WHO ARE UNEMPLOYED

- number of people aged 15 years and over who are unemployed and looking for work [LFSP = 4–5]
- number of people aged 15 years and over in the labour force [LFSP = 1–5]

UNEMP_RATIO — % PEOPLE AGED 15 YEARS AND OVER WHO ARE UNEMPLOYED

- number of people aged 15 years and over who are unemployed and looking for work [LFSP = 4–5]
- number of people aged 15 years and over (excluding labour force status not stated) [LFSP = 1–6]

Occupation variables

OCC_DRIVERS — % EMPLOYED PEOPLE CLASSIFIED AS MACHINERY OPERATORS AND DRIVERS

- number of employed people classified as Machinery Operators and Drivers [OCCP = 7]
- number of employed people with a stated occupation [OCCP = 1–8]

OCC_LABOUR — % EMPLOYED PEOPLE CLASSIFIED AS LABOURERS

- number of employed people classified as Labourers [OCCP = 8]
- number of employed people with a stated occupation [OCCP = 1–8]

OCC_MANAGER	% EMPLOYED PEOPLE CLASSIFIED AS MANAGERS
	• number of employed people classified as Managers [OCCP = 1]
	• number of employed people with a stated occupation [OCCP = 1–8]
OCC_PROF	% EMPLOYED PEOPLE CLASSIFIED AS PROFESSIONALS
	• number of employed people classified as Professionals [OCCP = 2]
	• number of employed people with a stated occupation [OCCP = 1–8]
OCC_SALES_L	% EMPLOYED PEOPLE CLASSIFIED AS LOW-SKILL SALES WORKERS
	• number of employed people classified as Low-Skill Sales Workers [OCCP = 6 and Skill Level = 5][10]
	• number of employed people with a stated occupation [OCCP = 1–8]
OCC_SERVICE_L	% EMPLOYED PEOPLE CLASSIFIED AS LOW-SKILL COMMUNITY AND PERSONAL SERVICE WORKERS
	• number of employed people classified as Low-Skill Community and Personal Service Workers [OCCP = 4 and Skill Level = 4–5]
	• number of employed people with a stated occupation [OCCP = 1–8]
OCC_SKILL1	% EMPLOYED PEOPLE WHO WORK IN A SKILL LEVEL 1 OCCUPATION
	• number of employed people who work in a Skill Level 1 occupation [Skill Level = 1]
	• number of employed people with a stated occupation [OCCP = 1–8]
OCC_SKILL2	% EMPLOYED PEOPLE WHO WORK IN A SKILL LEVEL 2 OCCUPATION
	• number of employed people who work in a Skill Level 2 occupation [Skill Level = 2]
	• number of employed people with a stated occupation [OCCP = 1–8]
OCC_SKILL4	% EMPLOYED PEOPLE WHO WORK IN A SKILL LEVEL 4 OCCUPATION
	• number of employed people who work in a Skill Level 4 occupation [Skill Level = 4]
	• number of employed people with a stated occupation [OCCP = 1–8]
OCC_SKILL5	% EMPLOYED PEOPLE WHO WORK IN A SKILL LEVEL 5 OCCUPATION
	• number of employed people who work in a Skill Level 5 occupation [Skill Level = 5]
	• number of employed people with a stated occupation [OCCP = 1–8]

Housing variables

FEWBED	% CLASSIFIABLE OCCUPIED PRIVATE DWELLINGS WITH ONE OR NO BEDROOMS
	• number of classifiable occupied private dwellings with one or no bedrooms [BEDD = 0,1 and HHCD = 11–32][11]
	• number of classifiable occupied private dwellings with a stated number of bedrooms [BEDD ne &&, @@ and HHCD = 11–32]
GROUP	% OCCUPIED PRIVATE DWELLINGS THAT ARE GROUP OCCUPIED PRIVATE DWELLINGS
	• number of classifiable occupied private dwellings that are occupied by group households [HHCD = 32 and HHCD = 11–32]
	• number of classifiable occupied private dwellings [HHCD = 11–32]

10 The Skill Level for each occupation can be found in table 5 of the ABS data cube: ANZSCO *First Edition Revision 1 - Structure* (ABS, 2009).

11 Household composition was 'not classifiable' if the household: contained only visitors or persons aged under 15 years on Census night; or was determined to be occupied on Census Night but the collector could not make contact; or could not be classified because there was insufficient information on the Census form.

HIGHBED — % OCCUPIED PRIVATE DWELLINGS WITH FOUR OR MORE BEDROOMS

- number of classifiable occupied private dwellings with four or more bedrooms [BEDD = 4–30 and HHCD = 11–32]
- number of classifiable occupied private dwellings with a stated number of bedrooms [BEDD ne &&, @@ and HHCD = 11–32]

HIGHMORTGAGE — % OCCUPIED PRIVATE DWELLINGS PAYING MORE THAN $2,800 PER MONTH IN MORTGAGE REPAYMENTS

- number of mortgaged classifiable occupied private dwellings with monthly mortgage repayments greater than $2,800 [MRED = 2801–9999 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding those with tenure not stated, mortgage not stated and rent not stated) [TEND ne &, @, MRED ne &&&&, RNTD ne &&&& and HHCD = 11–32]

HIGHRENT — % OCCUPIED PRIVATE DWELLINGS PAYING MORE THAN $370 PER WEEK IN RENT

- number of rented classifiable occupied private dwellings with rent payments greater than $370 per week [RNTD = 371–9999 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding those with tenure not stated, mortgage not stated and rent not stated) [TEND ne &, @, MRED ne &&&&, RNTD ne &&&& and HHCD = 11–32]

LOWRENT — % OCCUPIED PRIVATE DWELLINGS PAYING LESS THAN $166 PER WEEK IN RENT (EXCLUDING $0 PER WEEK)

- number of rented classifiable occupied private dwellings with rent payments less than $166 per week (excluding rent-free and renting from employer) [RNTD = 1–165 and HHCD = 11–32 and LLDD ne 51, 52]
- number of classifiable occupied private dwellings (excluding those with tenure not stated, mortgage not stated and rent not stated) [TEND ne &, @, MRED ne &&&&, RNTD ne &&&& and HHCD = 11–32]

OVERCROWD — % OCCUPIED PRIVATE DWELLINGS REQUIRING ONE OR MORE EXTRA BEDROOMS (BASED ON CANADIAN NATIONAL OCCUPANCY STANDARD)

- number of classifiable occupied private dwellings needing one or more extra bedrooms (based on Canadian National Occupancy Standard[12]) [Housing utilisation[13] = 'One or more extra bedrooms needed' and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding dwellings where housing utilisation cannot be determined or is not stated) [Housing utilisation ne 'Not applicable', 'Unable to be determined', 'Not stated' and HHCD = 11–32]

OWNING — % OCCUPIED PRIVATE DWELLINGS OWNING THE DWELLING THEY OCCUPY (WITHOUT A MORTGAGE)

- number of households owning the dwelling they occupy without a mortgage [TEND = 1 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding tenure not stated) [TEND ne &, @ and HHCD = 11–32]

12 The Canadian National Occupancy Standard determines housing appropriateness, using the number of bedrooms and the number, age, sex and relationships of household members. For more information refer to *Housing Occupancy and Costs, 2009–10* (ABS, 2011d).

13 The 'Housing utilisation' variable was derived from Census data items, according to the Canadian National Occupancy Standard.

MORTGAGE — % OCCUPIED PRIVATE DWELLINGS OWNING THE DWELLING THEY OCCUPY (WITH A MORTGAGE)

- number of mortgaged classifiable occupied private dwellings [TEND = 2, 3, 6 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding tenure not stated) [TEND ne &, @ and HHCD = 11–32]

SPAREBED — % OCCUPIED PRIVATE DWELLINGS WITH ONE OR MORE SPARE BEDROOMS (BASED ON CANADIAN NATIONAL OCCUPANCY STANDARD)

- number of classifiable occupied private dwellings with one or more spare bedrooms (based on Canadian National Occupancy Standard) [Housing utilisation = 'One bedroom spare', 'Two or more bedrooms spare' and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding dwellings where housing utilisation cannot be determined or is not stated) [Housing utilisation ne 'Not applicable', 'Unable to be determined', 'Not stated' and HHCD = 11–32]

LONE — % OCCUPIED PRIVATE DWELLINGS THAT ARE LONE PERSON OCCUPIED PRIVATE DWELLINGS

- number of classifiable occupied private dwellings that are occupied by lone person households [HHCD = 31]
- number of classifiable occupied private dwellings [HHCD = 11–32]

Other indicators of advantage or disadvantage

Cars

HIGHCAR — % OCCUPIED PRIVATE DWELLINGS WITH THREE OR MORE CARS

- number of classifiable occupied private dwellings which had 3 or more registered motor vehicles at or near the dwelling [VEHD = 3–30 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding number of vehicles not stated) [VEHD ne &&, @@ and HHCD = 11–32]

NOCAR — % OCCUPIED PRIVATE DWELLINGS WITH NO CARS

- number of classifiable occupied private dwellings which did not have a registered motor vehicle at or near the dwelling [VEHD = 0 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding number of vehicles not stated) [VEHD ne &&, @@ and HHCD = 11–32]

Internet

DIALUP — % OCCUPIED PRIVATE DWELLINGS WITH A DIALUP INTERNET CONNECTION

- number of classifiable occupied private dwellings with a dialup internet connection [NEDD = 3 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding internet connection not stated) [NEDD ne &, @ and HHCD = 11–32]

NONET — % OCCUPIED PRIVATE DWELLINGS WITH NO INTERNET CONNECTION

- number of classifiable occupied private dwellings with no internet connection [NEDD = 1 and HHCD = 11–32]
- number of classifiable occupied private dwellings (excluding internet connection not stated) [NEDD ne &, @ and HHCD = 11–32]

Other

CHILDJOBLESS — % FAMILIES WITH CHILDREN UNDER 15 YEARS OF AGE AND JOBLESS PARENTS

- number of families with children aged under 15 and jobless parents [FMCF = 21, 31 and LFSF = 16, 17, 19, 25, 26]
- number of families (excluding not applicable and not stated) [FMCF ne @@@@ and LFSF ne 06, 11, 15, 18, 20, 21, 27, @@]

DISABILITYU70 — % PEOPLE AGED UNDER 70 WHO NEED ASSISTANCE WITH CORE ACTIVITIES

- number of people aged under 70 years needing assistance in one or more of the three core activity areas of self-care, mobility and communication, because of a disability, long term health condition (lasting six months or more) or old age [AGEP < 70 and ASSNP = 1]
- number of people aged under 70 years (excluding need for assistance not stated) [AGEP < 70 and ASSNP = 1–2]

ENGLISHPOOR — % PEOPLE WHO DO NOT SPEAK ENGLISH WELL

- number of people aged 5 years and over who speak English either not well or not at all [AGEP > 4 and ENGLP = 4, 5]
- number of people aged 5 years and over (excluding those who did not state their English proficiency or main language) [AGEP > 4 and ENGLP = 1–5]

ONEPARENT — % FAMILIES THAT ARE ONE PARENT FAMILIES WITH DEPENDENT OFFSPRING ONLY

- number of families that are one parent families with dependent offspring only [FMCF = 3112, 3122, 3212]
- number of families [FMCF ne @@@@]

SEP_DIVORCED — % PEOPLE AGED 15 AND OVER WHO ARE SEPARATED OR DIVORCED

- number of people aged 15 years and over who are separated or divorced [MSTP = 3, 4]
- number of people aged 15 years and over [MSTP = 1–5]

UNINCORP — % OCCUPIED PRIVATE DWELLINGS WITH AT LEAST ONE PERSON WHO IS THE OWNER OF AN UNINCORPORATED ENTERPRISE

- number of classifiable occupied private dwellings where at least one usual resident is the owner of an unincorporated enterprise [EMTP = 3, UAICP = 1 and HHCD = 11–32]
- number of classifiable occupied private dwellings [HHCD = 11–32]

B. IMPACT OF REMOVING INDIGENOUS VARIABLE ON IRSD

Table B.1 shows the change in percentile for SA1s when removing the Indigenous variable from the IRSD.

B.1 Percentile differences when including the Indigenous variable in the 2011 IRSD

Percentile Difference	*0–1*	*2–5*	*6–10*	*>10*
Percentage of SA1s	94.98%	4.59%	0.38%	0.05%

Table B.1 shows that 99.57% of SA1s changed by five or fewer percentiles in the 2011 IRSD ranking distribution, and 0.05% (27 SA1s) of areas changed by more than 10 percentiles.

The characteristics of the biggest percentile difference SA1s, that is those areas with a percentile difference greater than 10, were inspected to understand the changes these areas were undergoing when the Indigenous variable was added to the IRSD. In general, it was found that the biggest differences occurred in areas that were otherwise of average ranking, or were less disadvantaged. This indicates that the most disadvantaged areas were already being identified and ranked appropriately by the IRSD without the Indigenous variable included.

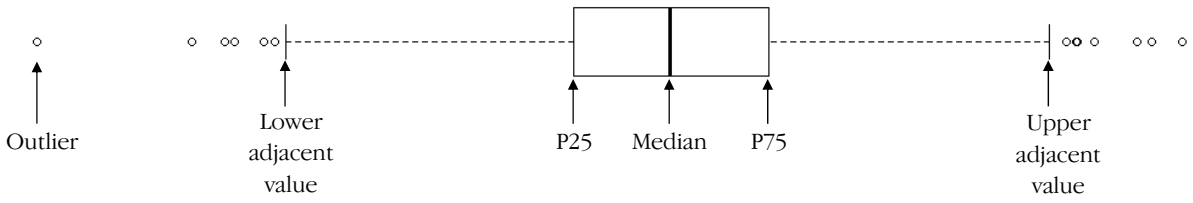
Further comparative analyses of the Indigenous-in and Indigenous-out IRSD highlighted the small effect the variable has on the index:

- There was little impact on the loading of the other variables in the index. The variable loading differences in the IRSD were within ± 0.02.
- We compared the eigenvalues and percentage of variance explained in the underlying data for both the Indigenous-in and Indigenous-out indexes: for the Indigenous-out IRSD, we observed an eigenvalue of 7.06 with a corresponding percentage of variance explained equal to 44.10%; for the Indigenous-in IRSD, we observed an eigenvalue of 7.33 and a corresponding percentage of variance explained equal to 43.10%. The IRSD explains more of the underlying variance in the data without the Indigenous variable, although this is only a marginal difference.
- There was a correlation of 0.999 between the Indigenous-in and Indigenous-out indexes, further highlighting the similarities between the two indexes.
- The influence of the Indigenous variable was also assessed, and was found to be on average the least influential variable of all 2011 IRSD variables (in terms of effect on rankings). More information on the influence function and this type of analysis can be found in Section 5.4.

C. INTERPRETING BOX PLOTS

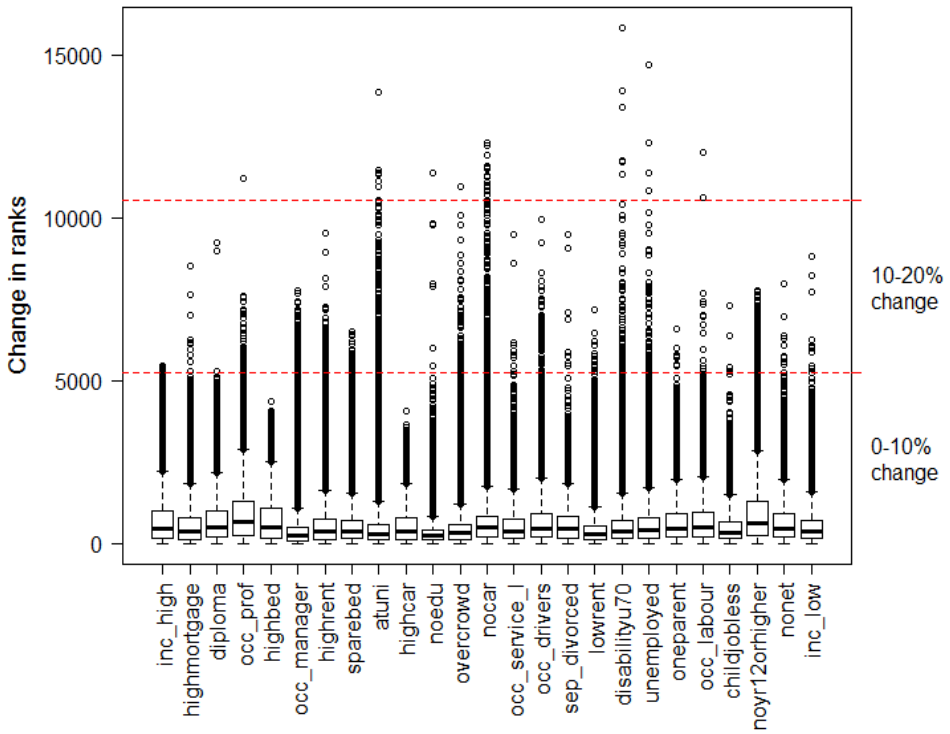
Distributions can be represented by box plots (an example is shown in figure C.1). They are a simple method of presenting distributions. They present the median, upper (P75) and lower quartiles (P25), and range of the distribution. The upper and lower adjacent values are calculated using the interquartile range (IQR), which is the difference between the upper and lower quartile values (P75–P25). For example, the upper adjacent value is the largest value within 1.5 IQRs of the upper quartile. Values outside the upper and lower adjacent values are considered outliers and are represented by small circles.

C.1 Labelled diagram of box plot

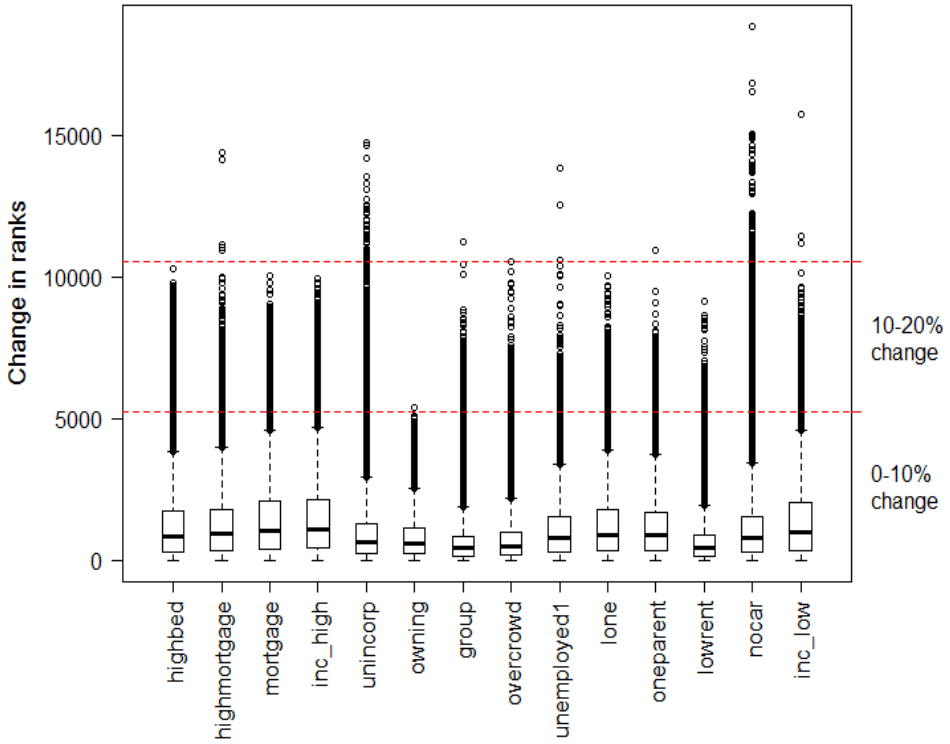


D. GRAPHS OF VARIABLE SENSITIVITY ANALYSIS

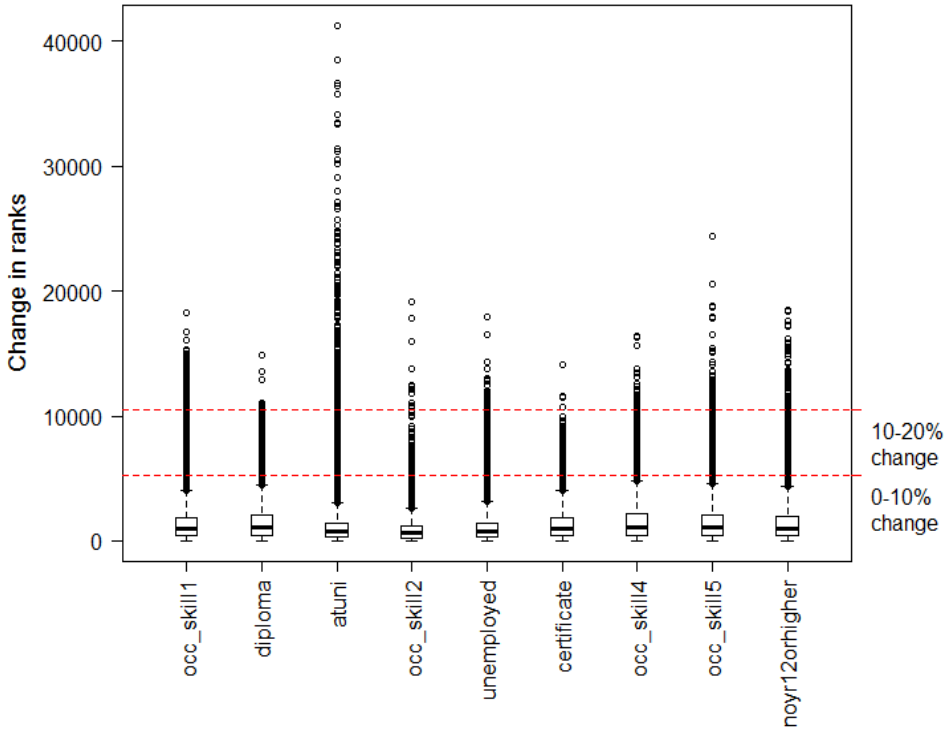
D.1 Distribution of absolute change in ranks by variable for the IRSAD



D.2 Distribution of absolute change in ranks by variable for the IER



D.3 Distribution of absolute change in ranks by variable for the IEO



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.
LIBRARY	A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

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, or purchase a hard copy publication. 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