1352.0.55.135

Australian Bureau of Statistics

Research Paper

Building on SEIFA: Finer Levels of Socio-Economic Summary Measures

www.abs.gov.au

New Issue

Research Paper

Building on SEIFA: Finer Levels of Socio-Economic Summary Measures

Phillip Wise and Courtney Williamson

Analytical Services Branch

Methodology Advisory Committee

7 June 2013, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 25 NOV 2013

ABS Catalogue no. 1352.0.55.135

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Ruel Abello, Analytical Services Branch on Canberra (02) 6252 6307 or email <analytical.services@abs.gov.au>.

BUILDING ON SEIFA: FINER LEVELS OF SOCIO-ECONOMIC SUMMARY MEASURES

Phillip Wise and Courtney Williamson
Analytical Services Branch

QUESTIONS FOR THE COMMITTEE

1. Should ordinal variables be used in the PCA, or are binary variables preferable? Is it preferable to use binary variables, as this avoids the subjective judgement regarding how much more advantaging or disadvantaging categories are compared to others?
2. How should missing items be dealt with – delete entire records, use imputation? What impact will this have on our methodology and variable list used, and does the choice of imputation methodology affect these considerations?
3. If imputation is performed, when should it be carried out? Should imputation be performed on the original Census data or on the variables constructed for household index? Should imputation occur after the PCA?
4. How important is it for users to understand the method? Likewise, how important is it for the variable weights to make intuitive sense to the users?
5. Is it worthwhile using a cut-off of 0.3 to determine whether a variable makes it into the final index? Some variables may not load highly on the summary index but have strong conceptual links with advantage or disadvantage.
6. Can the committee think of any other potential uses of finer-level indexes? Are there any major hindrances to the use of the indexes as presented?
7. Should the development of an index for use on the Basic Address Register be considered separately to an index for public release?

CONTENTS

ABBREVIATIONS

ABS	Australian Bureau of Statistics
BAR	Basic Address Register
CD	Collection District
Census	Australian Census of Population and Housing
MAC	Methodology Advisory Committee
MB	Mesh Block
PCA	Principal Component Analysis
SA1	Statistical Area Level 1
SEIFA	Socio-Economic Indexes For Areas
SES	Socio-economic Status

BUILDING ON SEIFA: FINER LEVELS OF SOCIO-ECONOMIC SUMMARY MEASURES

Phillip Wise and Courtney Williamson
Analytical Services Branch

ABSTRACT

Socio-Economic Indexes for Areas (SEIFA) seek to summarise the socio-economic conditions of an area using relevant information from the Census of Population and Housing. The SEIFA indexes are widely used measures of relative socio-economic advantage and disadvantage at the Statistical Area Level 1 level.

The indexes provide information about the area in which a person lives, but within any area there are likely to be households, families and individuals with different characteristics to the overall population of that area. Constructing socio-economic summary measures for finer units such as households would enable researchers and policy makers in Australia to better differentiate between areas with concentrations of advantage and disadvantage. A household socio-economic index of disadvantage would also enhance analyses by enabling cross-classifications with Census data.

This paper proposes an experimental household level index as an addition to the current suite of SEIFA products. It would complement the area level rankings by adding more depth to the information given by SEIFA, as well as providing its own valuable insights. Producing a household index would also allow users to make more accurate inferences about smaller units, rather than confounding the characteristics of areas with the people living within them.

This paper builds on previous research at the Australian Bureau of Statistics into socio-economic indexes for individuals and families started in Baker and Adhikari (2007) and the individual diversity within areas of socio-economic status in Wise and Mathews (2011). Using 2011 Australian Census of Population and Housing data, this paper focuses on an exploration into the development and dissemination of a socio-economic index for households. It seeks to address issues raised in these two previous research papers.

1. INTRODUCTION

Socio-Economic Indexes for Areas (SEIFA) is an analytical product developed by the Australian Bureau of Statistics (ABS) that ranks areas in Australia according to relative socio-economic advantage and disadvantage. The indexes are based on relevant information from the five-yearly Census and summarise the income, education, occupation, employment and housing characteristics of areas. The SEIFA indexes are assigned to areas, not to individuals, and indicate the collective socio-economic characteristics of the people living in an area. Some common uses of SEIFA include determining areas that require government funding and services, identifying new business opportunities, and assisting research into the relationship between socio-economic disadvantage and various health and social outcomes (ABS, 2013).

A long-term research interest for the SEIFA team in Analytical Services has been the construction and dissemination of a finer level summary measure of socio-economic advantage and disadvantage. This paper catalogues the derivation of an experimental socio-economic index for households, using 2011 Australian Census of Population and Housing data and an appropriate conceptual and methodological basis for this undertaking (Baker and Adhikari, 2007; Wise and Mathews, 2011). More specifically, this paper builds on this previous ABS research into the diversity of socio-economic advantage and disadvantage within areas by discussing practical considerations for developing a household socio-economic summary measure: the choice of households as our finer level output unit; the selection and specification of appropriate Census variables; and a means for disseminating the summary measures.

The motivations behind the push to produce household level measures of socio-economic advantage and disadvantage can be summarised into two key points: unlocking new insights for research and analysis into socio-economic advantage and disadvantage, and providing important contextual information about the diversity within areas of socio-economic advantage and disadvantage. In these ways, the experimental household level index presented in this paper complements SEIFA by adding more depth and context to the area level information, as well as providing its own unique insights.

To elaborate, it is often the case in research and policy contexts that finer level socio-economic measures based at the individual, household or family levels are desired - Scutella and Wilkins (2010) and Lim and Gemici (2011) being two examples. Constructing additional socio-economic summary measures for a finer unit, such as households, would enable researchers and policy makers in Australia to better identify concentrations of advantage and disadvantage within areas. It would also enhance analyses by enabling cross-classifications with Census data. Further, basing our finer level summary measure specifically at the household level opens up possibilities for including such a measure on the ABS Basic Address Register (BAR) as contextual

socio-economic information about a household. This would greatly improve survey sampling to target different socio-economic populations.

Producing a household level index would also allow users to make more accurate inferences about smaller units, rather than confounding the characteristics of areas with the people living within them. Confounding these characteristics can lead to the misclassification of people living within an area, and issues with interpretability of results (Lim *et al* (2011) and McCracken (2001)). This misclassification of an individual's characteristics due to using an area measure as a proxy is called ecological fallacy. The extent to which the SEIFA indexes, as an area level product released for Statistical Areas Level 1 (SA1), can mask finer level diversity of socio-economic disadvantage has been investigated extensively through ABS research catalogued in Baker and Adhikari (2007) and Wise and Mathews (2011).

This paper is structured as follows. Section 2 discusses the conceptual issues related to finer level measures of socio-economic advantage and disadvantage by considering refinements from the main ABS SEIFA product through the Mesh Block, household, family and individual levels. Section 3 covers the concepts and construction of a household level index, and presents reasoning behind the choices made to derive this index relating to variable selection, weighting schemes, a means for validating the index and dealing with missing data. This is followed by a discussion in Section 4 of the issues facing the release of an experimental household level index product, and methods to disseminate aggregate index information to the public. In Section 5 we summarise our findings and outline possible directions for future research into experimental products using finer level indexes of socio-economic advantage and disadvantage.

2. FROM AREAS TO INDIVIDUALS

This section discusses the conceptual issues related to finer level measures of socio-economic advantage and disadvantage, including our motivations for creating finer level measures of socio-economic advantage and disadvantage at the household level.

2.1 The spectrum of output for socio-economic indexes

Whilst previous ABS research into finer level socio-economic indexes has focused on individual and family level data, there has not been a full consideration of the merits of the different approaches. Within a Census Collector District (CD for Census releases including and prior to 2006) or Statistical Area Level 1 (SA1 for Census releases from 2011 onwards), there are four separate structures that the ABS could produce index scores and rankings for. They are:

Statistical Area Level 1 → Mesh Block → Household → Family → Individual

This section discusses the merits of each approach and sets forth arguments for why the household level is preferred as a finer level measure of socio-economic disadvantage. Table 2.1 contains a summary of the advantages and disadvantages associated with producing socio-economic summary measures for different levels of Census data aggregation. Discussion of the points included in the table is structured into sub-headings following.

Population undercoverage

Previous research has highlighted that individual level indexes have issues of applicability across the age spectrum. Baker and Adhikari (2007) and Wise and Mathews (2011) could not feasibly calculate an individual level index for people under the age of 15 or over the age of 64 from Census data due to conceptual issues with occupation and education characteristics. This amounted to approximately one-third of the usual resident population (Wise and Mathews, 2011). Similarly, Bailey *et al.* (2003) recommended that separate individual level indexes for adult and child deprivation be created because of such conceptual issues.

Furthermore, Baker and Adhikari (2007) treated families as only being an identifiable unit if they contained more than one person. This excludes important subsets of the population who experience disadvantage, such as lone person households. The issue also raises the question of applicability to group households, where unrelated single adults live together but would receive unique family identifiers in the Census data.

A household measure would allow for households to be identified as advantaged or disadvantaged based on the characteristics of the house and the people living within it, thus negating the need to only consider a proportion of the total population.

2.1 Advantages and disadvantages of summary measures produced for different data aggregations

Output level	*Advantages*	*Disadvantages*
SA1	User familiarity High quality data Strong confidentiality of data	Tied to geographical output Ecological fallacy (can mask diversity of SES) Not useful for some research and policy applications
Mesh Block	Finer level of output = less diversity within areas Familiarity for users of area level indexes	Tied to geographical output Ecological fallacy still an issue Greater population exclusions than SA1s Low population counts = weaker confidentiality and lower data quality to support index construction
Household	Finer level of output = targeted advantage and disadvantage Strong conceptual basis Wide scope of measurement High population inclusion	More difficult to disseminate to the public while maintaining confidentiality Treatment of missing data items? How to validate the indexes?
Family	Finer level of output = targeted advantage and disadvantage	Ambiguous conceptually (what is family disadvantage?) Difficult to measure with limited family-based Census data More difficult to disseminate to the public while maintaining confidentiality Population undercoverage if excluding single person 'families' Treatment of missing data items? How to validate the indexes?
Individual	Finer level of output = targeted advantage and disadvantage Strong conceptual basis Wide scope of measurement Desired in the research and policy communities	Substantial population exclusions due to applicability of Census data across the age spectrum More difficult to disseminate to the public while maintaining confidentiality Treatment of missing data items? How to validate the indexes?

Data quality

The use of an exclusion rules framework to ensure minimum data requirements for an area to receive a score has been a feature of SEIFA since its inception following the 1986 Census. For example, areas are excluded if they have populations less than 10 or if they have less than 6 relevant respondents for the variables comprising the indexes (ABS, 2013). Mesh Blocks, as a similarly constructed area level summary measure based on proportions of advantaging and disadvantaging characteristics, do not have the same strength of data quality because they are typically much smaller in size than SA1s. Radisich and Wise (2012) contains a theoretical investigation into the effect on Mesh Block output of using similar exclusion rules to the CD level with 2006 data. The results showed almost four times as many population exclusions for Mesh Block output.

Constructing summary measures at the individual, household or family levels presents questions of data quality based on the incidence of 'not stated' responses in the Census. Whilst very low, the level of 'not stated' responses can impact on the rankings. For instance, if 'not stated' is grouped with records that do not fulfil a disadvantaging characteristic, then there is an implicit lack of disadvantage assigned to individuals in these circumstances. The use of imputation to mitigate this issue has been investigated in Section 3 of this paper, however the question of whether imputing is more beneficial than not treating the data in this context remains open.

Conceptual considerations

Household advantage and disadvantage has a stronger conceptual basis than individual or family level measures. The household as a functional unit is the central aspect of modern life. Households are appropriate to consider as the basic unit of analysis for finer level advantage and disadvantage because their members typically pool their income and resources and share similar living characteristics (Zipp and Plutzer, 1996). A household can still be advantaged as a unit if it can support less advantaged members. However it is not straightforward to ascertain whether an individual is advantaged or disadvantaged based on the characteristics of the household in which they live and how this interacts with their personal education and employment situation.

It can also be appropriate to consider the highest level of education or occupation of a person within a household as an indicator of the capacity of that household to support its fellow residents, as we have done in constructing the household index presented in this paper. This approach is common in literature concerning the socio-economic status of students (Lim *et al.*, 2011).

Utility of output

Releasing our household socio-economic index through Census TableBuilder would provide users greater flexibility and allow more detailed analysis than current SEIFA outputs. Households could be categorised into ranked groups based on the index, for example from the most disadvantaged households (group 1) to the most advantaged households (group 4). These groups could then be cross-classified by other Census variables, such as religious affiliation, number of children or age. This would allow users greater freedom to manipulate the data and to produce output that is of most relevance to their analysis. Figure 2.2 shows proposed Census TableBuilder output, which cross-classifies household level socio-economic index groups by language spoken at home. Note that this data is synthetic and is included for illustrative purposes.

2.2 Example of Census TableBuilder output produced for a Household Level Index

Language Spoken at Home by Household Level Socio-Economic Index

Household Level Socio-Economic Index	Group 1	Group 2	Group 3	Group 4
Language Spoken at Home				
Northern European Languages	2,223,946	982,820	555,824	1,737,779
Southern European Languages	50,265	108,790	52,327	112,269
Eastern European Languages	27,716	36,710	20,755	51,381
Southwest and Central Asian Languages	72,040	13,532	5,202	29,104
Southern Asian Languages	64,522	6,948	2,448	15,882
Southeast Asian Languages	61,173	10,879	5,207	29,963
Eastern Asian Languages	69,212	25,252	10,011	47,314
Australian Indigenous Languages	12,758	1,107	463	2,534
Total	2,581,632	1,186,038	652,237	2,026,226

Data Source: 2011 Census of Population and Housing

The data could also be disseminated to provide more detail for SEIFA output by aggregating our household socio-economic index to the SA1 level. This would provide SEIFA users additional information to minimise the extent of the ecological fallacy by identifying household level advantage and disadvantage within each SA1. Figure 2.3 shows a proposed output method that facilitates the release of household level summary measures at the SA1 level. As above, the households are categorised into groups from the most disadvantaged households (group 1) to the most advantaged households (group 4). Similarly to figure 2.2, this data is synthetic and is included for illustrative purposes.

2.3 Example of output produced for a Household Level Index

Australian Bureau of Statistics

1000.0.001 - Household Level Socio-economic Index of Advantage and Disadvantage, Data Cube only, 2011

Released at 11.30am (Canberra time) 01 December 2013

Table 1. Distribution of Households within SA1s, 2011

2011 Statistical Area Level 1 Code (SA1)	2011 SEIFA Decile Ranking	Usual Resident Population of SA1	Household Group 1	Household Group 2	Household Group 3	Household Group 4
1010101	2	52	4	6	20	2
1010102	6	31	15	3	0	0
1010103	8	145	2	3	78	10
1010104	9	208	0	2	5	60
1010105	7	67	5	5	6	3
1010106	1	89	3	3	23	2
1010107	10	115	15	18	9	20
1010108	3	334	3	63	64	5
1010109	3	451	38	22	12	69
1010110	7	90	24	22	2	3

After careful consideration of the pros and cons discussed in this section, we chose to proceed with creating a household level index. Household level output has strong conceptual benefits and also minimises population exclusions for finer level summary measures.

2.2 Conceptual considerations between area and finer level socio-economic indexes

For the purposes of SEIFA, the ABS 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*. Further information on the background of this definition and how it relates to conceptualisations of disadvantage, social exclusion, poverty and deprivation can be found in ABS (2013); Wise and Mathews (2011) contains a discussion of how this definition relates to individual level disadvantage.

This conceptual basis is important because it informs both the candidate list of variables to consider for inclusion in any socio-economic index we wish to construct, and also clarifies the appropriate use of the index once it has been produced (Michalos *et al.*, 2011 and ABS, 2011). For this paper, specifically:

- Area level disadvantage relates to the shared characteristics of a community or neighbourhood, as reflected in the attributes of the people living in that area and the types of households they live in.
- Household level socio-economic disadvantage relates to the individual access to resources of people living within households and their ability to collectively share these resources in order to participate in society.

This is to be measured using a scoping list of variables derived to best represent household socio-economic advantage and disadvantage given the constraints of information available through the 2011 Census.

One way to illustrate the difference in definitions is to consider the case of a high number of motor vehicles at a household. At the area level, a high proportion of households with three or more vehicles reflects relative socio-economic advantage, but at the individual household level having three or more vehicles is a reflection of personal preferences.[1] Common shared characteristics across an area such as number of motor vehicles can reflect aspects of socio-economic advantage or disadvantage, but their meaning at finer levels can be more attuned to personal choice than whether a household is relatively advantaged or disadvantaged.

This approach to use a unique definition to describe socio-economic disadvantage at the household level means that we have not transposed the SEIFA variables and their weights to the household level data. Rather, we are building a separate index that best summarises household level socio-economic advantage and disadvantage. It is important to highlight that users cannot aggregate up from the household measures to areas; the mean and distribution of SEIFA is independent to the mean and distribution of our household index. However, the household level index can provide important contextual information to areas, as is described in further detail in Section 2.1.

1 The relationship between car ownership and socio-economic advantage as changing between different unit levels was identified during the Methodology Advisory Committee discussion and in the review of this paper. Consequently variables relating to ownership of cars still remain in the index.

3. CONSTRUCTION AND CALCULATION

This section discusses practical issues associated with constructing the household level index, including how variables were specified appropriately, how missing data were dealt with, how the weighting scheme was developed and how the index could be validated. The technical details of the calculation of our experimental household level index are discussed as each issue is addressed.

3.1 Variable specification

Before constructing the indexes, we reviewed the list of Census variables and identified those associated with our definition of socio-economic advantage and disadvantage. When developing the candidate list of variables, we considered variables that are (i) a cause, (ii) a consequence, or (iii) have an association with advantage or disadvantage. Variables that are a cause or an association can act as proxy measures for consequence variables, so can be important in measuring advantage or disadvantage when consequence variables are not observed on the Census. We adopted this approach because it was deemed to provide the best measure to reflect relative advantage and disadvantage. This is consistent with the approach adopted for SEIFA (ABS, 2013).

The types of variables considered for use in the household level index can be separated into the following categories:

- SEIFA variables that relate directly to household measurement,
- person and family based SEIFA variables adapted to the household level, and
- new and representative household variables developed from Census data.

Each of the above categories has advantages and disadvantages, and it was decided that the best variables would be selected by using a combination of these three types. The reasons for this were to maintain familiarity for SEIFA users; give sound conceptual grounding of household level advantage and disadvantage; and derive the most relevant indicators of household level advantage and disadvantage from Census data. Although an index created using only household level variables would be conceptually simple to explain, this would severely limit the variables available for selection. The index should utilise individual and family level data, as both can influence the socio-economic characteristics of a household. This infers that a household's level of advantage and disadvantage is not only derived from the dwelling, but also the individuals and families residing within it.

The list of candidate variables is presented in tables 3.1–3.5. All of the variables are binary indicators as this approach was deemed appropriate from previous ABS research into finer level socio-economic indexes. Additionally, references from the health- and asset-based Principal Component Analysis (PCA) literature support the use of binary indicators in instances where we do not have access to ordinal data based on

a scale with roughly equal distances between categories (Kolenikov and Angeles, 2009), such as Likert scales on health outcomes and asset counts. For example, Vyas and Kumaranayake (2006) discuss that it is appropriate when categorical variables have no hierarchical relationship to convert them into binary indicators, since this does not change the relationship between the variables, nor does it add any additional variation or correlation to the dataset. The issue of using binary indicators or a mixture incorporating ordinal scales is explored further in Section 3.3.

3.1 List of household variables*

Variable mnemonic	*Variable description*
NOCAR	Households with no car (dis)
HIGHCAR	Households with three or more cars (adv)
FEWBED	Households with one or no bedrooms (dis)
HIGHBED	Households with four or more bedrooms (adv)
OTHER_HHLD	Households with a structure classified as "other" (e.g. caravan, tent) (dis)
RETIRED_NOT_OWNED	Households with a person aged over 65 years who does not own the home, or occupy it under a like tenure scheme (dis)
NOBROADBAND	Households without broadband internet connection (dis)
MULTI_FAMILY	Households with more than one family living in it (dis)
OVERCROWD	Households requiring one or more extra bedrooms (based on Canadian National Occupancy Standard) (dis)
LOWRENT	Households paying less than $166 a week in rent (excluding $0) (dis)
HIGHRENT	Households paying more than $370 a week in rent (adv)
OWNED	Households where dwelling is owned outright (adv)
MORTGAGE	Households where dwelling is being bought (adv)
LONE	Households that are lone person households (dis)

* Variables are followed by either "adv" or "dis" to indicate whether the variable is advantaging or disadvantaging.

3.2 List of family variables*

Variable mnemonic	*Variable description*
ONEPARENT	Households with a one-parent family (dis)
CHILDJOBLESS	Households with children aged under 15 years and both parents unemployed (dis)

* Variables are followed by either "adv" or "dis" to indicate whether the variable is advantaging or disadvantaging.

3.3 List of education variables*

Variable mnemonic	*Variable description*
NOSCHOOL	Households in which the most educated person has not been to school (dis)
NOYEAR12	Households in which the most educated person left school at year 11 or below (dis)
CERTIFICATE	Households in which the most educated person has a certificate (adv)
DIPLOMA	Households in which the most educated person has a diploma (adv)
DEGREE	Households in which the most educated person has a degree (adv)
ATUNI	Households with a person who is attending university (adv)

* Variables are followed by either "adv" or "dis" to indicate whether the variable is advantaging or disadvantaging.

3.4 List of occupation variables*

Variable mnemonic	*Variable description*
INC_LOW	Households with low annual equivalised income (between $1 and $20,799) (dis)
INC_HIGH	Households with high equivalised income (greater than $52,000) (adv)
UNEMPLOYED	Households where one person aged over 15 years is unemployed (dis)
ALL_UNEMPLOYED	Households where all people aged over 15 years are unemployed (dis)
LOW_SKILL	Households where the most skilled adult is employed in skill level 5 occupation (dis)
SKILL_4	Households where the most skilled adult is employed in skill level 4 occupation (dis)
SKILL_3	Households where the most skilled adult is employed in skill level 3 occupation (–)
SKILL_2	Households where the most skilled adult is employed in skill level 2 occupation (adv)
HIGH_SKILL	Households where the most skilled adult is employed in skill level 1 occupation (adv)

* Variables are followed by either "adv" or "dis" to indicate whether the variable is advantaging or disadvantaging.

3.5 List of miscellaneous variables*

Variable mnemonic	*Variable description*
SEP_DIVORCED	Households with one or more people aged over 15 years separated or divorced (dis)
ENGPOOR	Households with one or more people aged over 15 years who do not speak English well (dis)
UNENGAGED_YOUTH	Households with one or more people aged between 15 and 24 years who are not working or studying (dis)
DISABILITY_UNDER70	Households with one or more people aged under 70 years who require assistance with core activities (dis)
DISABILITY_OVER70	Households with one or more people aged over 70 years who require assistance with core activities (dis)

* Variables are followed by either "adv" or "dis" to indicate whether the variable is advantaging or disadvantaging.

All variables used in this index are based on occupied private dwellings. Dwellings which are classified as unoccupied private dwellings or non-private dwellings are out of scope, which accounts for approximately 960,000 dwellings or 10.5% of all enumerated Census dwellings being excluded. Population classified as migratory, off-shore or shipping is also excluded. Altogether from dwelling and population exclusions, approximately 745,000 out of 21.5 million people (or 3.5% of the population) was excluded from the calculation of the index. This is a significantly smaller proportion of the population when compared with the Socio-Economic Indexes for Individuals, which excluded 33.15% of the population (Wise and Mathews, 2011).

3.2 Missing data

Due to partial non-response from some Census respondents, missing data and how it is treated exists as an issue for creating finer level socio-economic indexes. To illustrate, for the indicator variable for ATUNI (households with a person attending university), households with this specific advantaging characteristic were coded to 1; households without this characteristic were coded to 0. For some households no

response was given for the Census question on type of educational institution attending (TYPP), so it is unclear whether anyone in this household has this characteristic or not. In some studies, such non-response is grouped with the 0 category (Salmond *et al.*, 2006 and Wise and Mathews, 2011). However, this might inappropriately assign such households an implicit lack of advantage.

In Vyas and Kumaranayake (2006), the authors chose to impute the mean response for a variable because there was a low incidence of missing data in their analysis (less than 1%). Hence they expected their choice of action to have little effect on the distribution of socio-economic status (SES). Their paper also provided a comparison of two further alternative approaches. In Cortinovis *et al.* (1993), the authors excluded households with at least one missing value. Such an approach would significantly lower our in-scope population, since approximately 36% of dwellings had at least one missing value. Additionally, Cortinovis *et al.* (1993) suggests that such exclusions could lead to bias towards higher SES households as missing data may occur more frequently in lower SES households. In the other comparison study, Gwatkin *et al.* (2000) used mean imputation to treat missing data. For instances where there is a significant amount of missing data, attributing mean scores will reduce variation among households.

Based on these observations from other practical studies, two actions to deal with missing data were proposed for this paper:

1. remove households with high numbers of non-response, and
2. impute the missing value.

For the candidate variables considered for this index, 64.2% of respondents had no missing responses, and 92.8% of respondents had three or fewer questions that they did not respond to. Furthermore, most of the candidate variables selected for this research had frequencies of non-stated responses less than 5.0%, with the highest at 8.3%. The small proportion of non-response means a reasonable attempt to deal with missing data should have a minimal impact on the index.

We decided to delete records that had ten or more missing responses for the candidate variables. Ten or more missing responses tended to correspond to dwellings where most person based variables were coded as "Not stated", such as for the education and occupation variables. This accounted for 2.0% of the population, or 423,234 people. Due to low levels of missing data and the high computational costs of imputing Census data, we opted to construct our household index without imputation. However we are considering imputation as a useful method to deal with this missing data in future.

3.3 Weight determination

Principal component analysis (PCA) was used to determine the weights for the variables in the household level index. This was done for several reasons. Firstly, PCA was chosen to avoid subjective judgements regarding the variable weights, which helps in developing an objective summary measure for socio-economic characteristics of households. Secondly, this data driven method was chosen as it captures the most variation in the carefully selected candidate variables. Thirdly, this is the same method that was used to determine the weights for SEIFA and for the indexes in Baker and Adhikari (2007) and Wise and Mathews (2011), thus providing familiarity to SEIFA users.

We used the first principal component to determine the household level scores, as this captures the largest proportion of variance in the original dataset. The use of additional components in conjunction with the principal component increases the proportion of variance explained but makes interpretation and dissemination of results more difficult. The correlation between each variable and the component is called the loading, which helps to interpret a component's relationship with the concept of advantage and disadvantage. More information on PCA can be found in the *SEIFA 2011 Technical Paper* (ABS, 2013).

The candidate variables listed in tables 3.1–3.5 were used in the PCA, and removed if their loading was less than 0.3. This process was performed iteratively, until all of the variables had a loading above 0.3. This is the same procedure used to create SEIFA indexes (ABS, 2013). The final variables following this process are shown in table 3.6.

3.6 Comparison list of variable loadings when including ordinal variables

All binary variables		*Mix of binary and ordinal variables*	
Variable	*Loading*	*Variable*	*Loading*
OVERCROWD	–0.50	OVERCROWD	–0.51
CHILD_JOBLESS	–0.44	CHILD_JOBLESS	–0.46
ONEPARENT	–0.41	ONEPARENT	–0.44
SKILL_5	–0.40	MULTI_FAMILY	–0.41
MULTI_FAMILY	–0.40	OCC_SKILL*	–0.40
INC_LOW	–0.36	UNENGAGED_YOUTH	–0.36
UNENGAGED_YOUTH	–0.35	INC_LOW	–0.35
UNEMPLOYED	–0.32	UNEMPLOYED	–0.33
ENGPOOR	–0.30	ENGPOOR	–0.30
LONE	0.37	LONE	0.41
DEGREE	0.42	INC_HIGH	0.48
SKILL_1	0.42	SPAREBED	0.51
SPAREBED	0.49	EDU_ATTAINMENT*	0.53
INC_HIGH	0.50	–	–

* Ordinal variables were used to replace the five 'highest level of occupation skill in the household' indicators and the five 'highest level of education in the household' indicators with one variable each, taking values 1–5.

Table 3.6 shows the index variables ordered from strongest disadvantaging characteristic at the top to the strongest advantaging characteristic at the bottom. The table compares the effect on the loadings of using binary indicators or ordinal scales to represent the education and occupation variables. For both variable specifications, we can see that the OVERCROWD, CHILD_JOBLESS and ONEPARENT variables are strong disadvantaging characteristics at the household level, whilst the SPAREBED variable is a common advantaging characteristic.

Collapsing the education and occupation binary indicators into two respective ordinal scales does not affect the ordering or selection of variables through PCA greatly, as table 3.6 demonstrates. However, it may create confusion for users trying to interpret the loadings associated with these two variables. This is because we are assigning one positive weight to the highest level of educational attainment in the household and one negative weight to the highest occupation skill in the household. In relative terms, low educational attainment will have a lower weight than high educational attainment, since the ordinal scale runs from 1 (low attainment) to 5 (high attainment), however it will still overall have a positive weight. The implication then is that the positive weight represents an advantaging characteristic, as it does for the remaining variables. The use of binary indicators avoids this issue by directly allowing for variables from the same categorical family to have different weights according to their association with advantage or disadvantage. Binary indicators also make no assumptions about equal interpretive distances between the categorical points in an ordinal scale, which we believe is one disadvantage to using ordinal variables to represent the Census skill and education hierarchies. This is why we elected to proceed to construct an index based on binary indicator variables alone.

Figure 3.7 presents the distribution of scores for this household index based on using binary indicators. There is a high degree of clumping in the middle of the distribution on certain unique scores, and a long tail of low index scores.

To assist in comparative analyses, areas were grouped into deciles and percentiles to users understand average relative socio-economic disadvantage of an area and compare between different areas (ABS, 2013). With previous ABS research into finer level measures, a high degree of clumping in the score distribution has been observed, making it difficult to formulate these typical groupings (Wise and Mathews, 2011). Figure 3.7 shows that clumping is present in our household index.

3.7 Distribution of household index scores

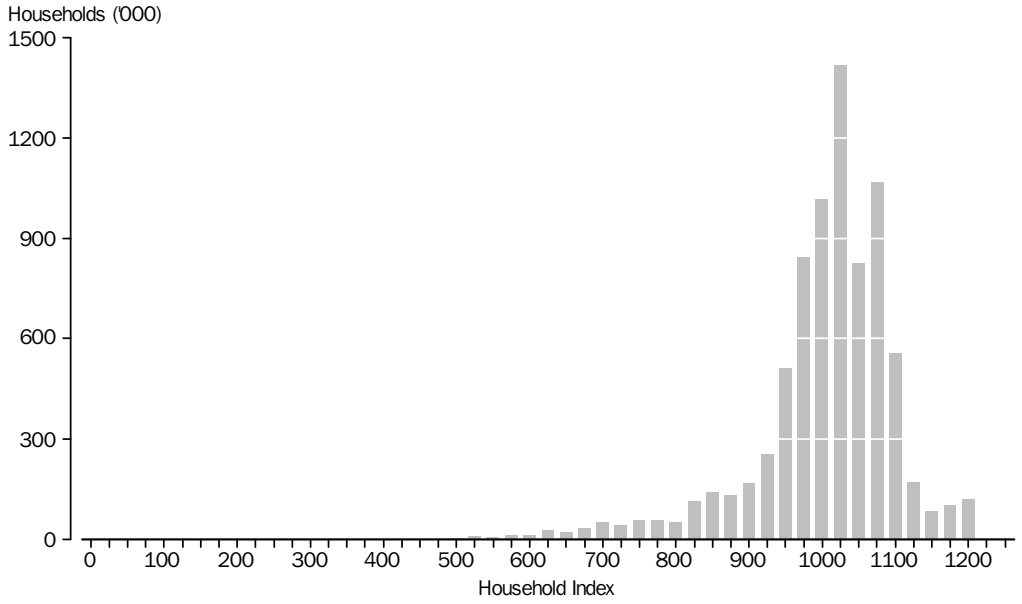

Table 3.8 shows the frequency distribution of households for our experimental index into a grouping structure that creates five groups with approximately 20% of households in each group. The clumping results in a smaller proportion of households forming the most advantaged group (group 5). The long tail of low index scores means caution should be exercised when interpreting the relative socio-economic advantage and disadvantage of households in group 1, because the scores for households range from 218 to 958.

3.8 Frequency distribution of ranked household index groups

Household index group	*Number of households**		*Household index score*	
	Frequency	*Percentage*	*Minimum*	*Maximum*
1	1,581,004	19.89	218	958
2	1,598,315	20.11	959	1004
3	1,724,187	21.69	1005	1019
4	1,733,262	21.80	1021	1063
5	1,312,385	16.51	1064	1190

* Total number of in-scope households for our analysis is 7,949,153.

The presentation of groupings in table 3.8 informs our discussion of dissemination in Section 4.

3.4 Validation

Validation is an important aspect of ABS output products. The SEIFA indexes undergo several stages of validation, including but not limited to: inspection of demographic changes in the intercensal period, analysis of the relationships between the indexes, confirming face validity of the rankings through mapping applications and the determination of influential areas and variables. These tasks are supplemented by the capacity for the SEIFA team to draw on other sources of expertise to confirm our findings. This includes consulting the ABS State and Territory Statistical Service to confirm localised rankings and liaising with Census staff to confirm our input data is derived appropriately (ABS, 2013).

Many of these tasks are not possible for finer level index measures. The primary consideration here is that SEIFA is calculated at the area level on approximately 50,000 areas, so we can check our input data items through other published sources such as Census TableBuilder. Our experimental household index involves processing approximately 8,000,000 records, and is a finer unit than Census publication output, so findings cannot be validated by published sources.

Validation of the index is proposed to be performed by testing the relationships between the indexes and variables known to be correlated with socio-economic advantage and disadvantage which are not captured on the Census, such as health outcomes. Furthermore, we propose to cross-check the programming tasks and processes used to create the index and inspect for face validity the rankings from our index using mapping tools. We also propose to test the sensitivity of our PCA-based weighting scheme by taking multiple random samples of households and re-deriving the weights.

4. DISSEMINATION

This section discusses the issues facing the release of an experimental household level index product, including the institutional responsibilities of the ABS, and methods to disseminate finer level index information to the public.

Confidentiality and the ABS

Releasing a household level index of socio-economic advantage and disadvantage is in line with the ABS goals and strategies to deliver high quality, objective and flexible official statistical solutions (ABS, 2012). Researchers and policy makers in the statistical and public knowledge domains are increasingly calling for access to microdata to support their analyses. Access to confidentialised unit record files for survey data is one way the ABS has been responding to such calls, however it is important to understand that the ABS operates within a clearly defined institutional environment, comprising a legislative framework and quality management practices. The Census and Statistics Act (1905) ensures the statistics the ABS disseminates maintain the confidentiality of information we collect. A key aspect of the practices the ABS employs to uphold respondent confidentiality is avoiding inadvertent disclosure in published statistics.

A wide range of information about employment, occupation, education and housing is collected by the Census. A household level index could be released as part of Census TableBuilder datasets, which would allow users to cross-classify the index by Census variables of interest, such as hours worked, or country of birth. This would provide users more flexibility and finer level outputs than available previously with SEIFA. More detail regarding this proposed output is provided in Section 2, and an example of this type of output is shown in figure 2.1.

A household level index could also be released to provide additional context and detail to SEIFA outputs. This could be achieved by aggregating the household index to the SA1 level. For reasons of privacy, Census data is released at the SA1 level as the finest output geography. More detail regarding this proposed output is provided in Section 2 and an example of this type of output is shown in figure 2.2.

Basic Address Register

Another dissemination possibility is for a household level index to be provided as valuable auxiliary information for adding to the Basic Address Register (BAR) following the 2016 Census. It would be very useful for design, estimation and imputation in survey contexts. In order to implement this, however, we would need to satisfy the following points:

- to know that it worked practically and gave tangible benefits for surveys,
- to change the relevant policies regarding the collection and storage of address identified Census data,
- to give adequate lead-in time before the next Census for this to be implemented.

These are significant obstacles and would require serious consideration before the ABS was to proceed with appending such classification information to the BAR. Issues such as whether storing and then linking addresses to summary Census information fits within the legislative requirements of the ABS to maintain the privacy and confidentiality of respondent information would need to be established. Utility of the information over time is also a concern for the quality of the register, although this could be mitigated somewhat by the assumption that even though a portion of households move between Censuses, they would tend to be replaced with people reasonably similar – socio-economically – to themselves.

5. CONCLUDING REMARKS

This paper has used 2011 Census data and an appropriate conceptual and methodological basis to construct a socio-economic index of advantage and disadvantage for households. Focus was placed on discussions of practical considerations for developing a finer level summary measure: the choice of the household as the most suitable analysis unit, and issues associated with variable specification, weight determination, and dissemination.

We have proposed to disseminate the household level index as part of Census TableBuilder, and as a count of households within Statistical Area Level 1s categorised into appropriate groupings for analysis. The summary measure presented in this paper was derived from binary indicator variables constructed from Census data based on measuring household advantage and disadvantage. We used Principal Component Analysis to specify weights for these variables. We excluded households with more than 10 non-responses to our relevant Census input data items, and only analysed occupied private dwellings. These decisions were deemed appropriate based on conceptual validity, a literature review and to build on user familiarity with previous ABS research into individual level socio-economic indexes.

Comments from the ABS Methodology Advisory Committee

A version of this paper was presented to the ABS Methodology Advisory Committee (MAC) in June 2013. The MAC members were interested to see the ABS continue to pursue the release of an experimental household level index product, since organisations seeking to target services at the moment use SEIFA even when this is not the most appropriate measure for their needs. There was acknowledgment that a finer level summary measure released at the household level would be beneficial to researchers in providing new insights for analysis into socio-economic advantage and disadvantage, and would also shed light on the diversity within areas of socio-economic advantage and disadvantage. MAC members suggested that the paper clarify the definition of advantage and disadvantage at the household level, and include a discussion of the relationship between the household and area based indexes with the view to highlighting that users cannot aggregate up from the household measures to the area measures. There was also some discussion about the extent of the variance that was left unexplained in finer level indexes, and how this compared to SEIFA.

Future directions

Given the discussions presented in this paper, users of SEIFA will understandably be wondering when they can expect a product to be released that enables them to appropriately analyse household level advantage and disadvantage. Before attempting this, the SEIFA team needs to perform critical validation work on a household level index and seek appropriate clearances for release from the ABS confidentiality unit and key internal stakeholders. The validation work we propose to perform includes inspection for face validity of the index using mapping tools, comparisons of results when derived by different people and sensitivity testing of the weights.

REFERENCES

Australian Bureau of Statistics (2011) *Information Paper: Measures of Socio-economic Status*, cat. no. 1244.0.55.001, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1244.0.55.001>

—— (2012) *ABS Corporate Plan, Australia, Jul 2012*, cat. no. 1005.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1005.0>

—— (2013) *Socio-economic Indexes for Areas: Technical Paper, 2011,* cat. no. 2033.0.55.001, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001>

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

Baker, J. and Adhikari, P. (2007) "Socio-Economic Indexes for Individuals and Families", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.086, Australian Bureau of Statistics, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.086>

Cortinovis, I.; Vela, V. and Ndiku, J. (1993) "Construction of a Socio-economic Index to Facilitate Analysis of Health in Data in Developing Countries", *Social Science and Medicine*, 36(8), pp. 1087–1097.

Gwatkin, D.R.; Rustein, S.; Johnson, K. *et al.* (2000) *Socio-economic Differences in Brazil*, Washington, DC: HNP/Poverty Thematic Group of the World Bank.

Kolenikov, S. and Angeles, G. (2009) "Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer?", *Review of Income and Wealth*, 55(1), pp. 128–165.

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

McCracken, K. (2001) "Into a SEIFA SES cul-de-sac?", *Australian and New Zealand Journal of Public Health*, 25(4), pp. 305–306.

Michalos, A.C.; Smale, B.; Labonté, R.; Muharjarine, N.; Scott, K.; Moore, K.; Swystun, L.; Holden, B.; Bernardin, H.; Dunning, B.; Graham, P.; Guhn, M.; Gadermann, A.M.; Zumbo, B.D.; Morgan, A.; Brooker, A.-S. and Hyman, I. (2011) *The Canadian Index of Wellbeing*, Technical Report 1.0. Waterloo, ON: Canadian Index of Wellbeing and University of Waterloo.

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.
<http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.038>

Salmond, C.; King, P.; Crampton, P. and Waldegrave, C. (2006) "NZiDep: A New Zealand Index of Socioeconomic Deprivation for Individuals", *Journal of Social Science and Medicine*, 62(6), pp. 1474–1485.

Vyas, S. and Kumaranayake, L. (2006) "Constructing Socio-economic Status Indices: How to use Principal Components Analysis", *Health Policy and Planning*, 21(6), pp. 459–468.

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.
<http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.036>

All URLs viewed 11 November 2013

ACKNOWLEDGEMENTS

The authors would like to thank Jeffrey Wright and Phillip Gould for their helpful input with this research project, and Peter Radisich, Rosalynn Mathews, Stephen Broadfoot and Peter Rossiter for their suggestions and feedback. The content and presentation of the paper are much improved as a result of their input. Responsibility for any errors or omissions remains solely with the authors.

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