

SOME COMMENTS ON THE SEIFA96 INDEXES

Kevin McCracken

*Department of Human Geography,
Macquarie University
New South Wales 2109
Email: Kevin.Mccracken@mq.edu.au*

Paper Presented to the 10th Biennial Conference
of the Australian Population Association

Melbourne, Australia
29th November to 1st December 2000

SOME COMMENTS ON THE SEIFA96 INDEXES

Kevin McCracken, *Macquarie University*

[Editor's note: see also Excel file, "K McCracken Tables"]

Introduction

Since the 1986 Census the Australian Bureau of Statistics (ABS) has produced a set of multivariate Socio-Economic Indexes for Areas (SEIFA) which have become widely used by government, academic and commercial researchers. In brief, the SEIFA indexes are area scores produced from principal components analyses of selected census-based socio-economic variables. The most recent set of indexes, derived from the 1996 Census of Population and Housing, were released by the ABS late in 1998. There are five indexes in the SEIFA package:

- (1) Index of Relative Socio-Economic Disadvantage (20 constituent variables)
- (2) Urban Index of Socio-Economic Relative Advantage (15 constituent variables)
- (3) Rural Index of Socio-Economic Relative Advantage (13 constituent variables)
- (4) Index of Economic Resources (22 constituent variables)
- (5) Index of Education and Occupation (18 constituent variables).

The index values are the scores on the first unrotated component extracted in each principal components analysis, the first component in all five analyses explaining about 30% of the variability in the respective input variables. An information paper on the indexes (ABS 1998) comes with the package.

This paper addresses several issues arising out of the author's examination and use of the 1996 indexes. The first issue is the broad one of the utility of complex composite measures such as the SEIFA indexes. Second, a number of technical/methodological points about the indexes are briefly discussed: (a) potential confusion over the so-called 'weights' of the constituent variables (b) published advice on interpreting results and (c) the selection of variables. Finally, the paper queries the ABS' policy of information release and non-release regarding the SEIFA measures.

The utility of SEIFA style composite indexes

The ready off-the-shelf availability of the SEIFA indexes is the proverbial case of mixed blessings. For users with neither the knowledge or resources to create their own indexes the SEIFA package is undoubtedly a very welcome product. Widespread use of standard indexes by researchers also carries advantages of research comparability. The downside though, is that the easy availability of the indexes may lead to them becoming uncritically employed, regardless of whether or not they are the most appropriate measures. A recent Sydney public health report in fact states that SEIFA 'has become the de facto standard for SES indexes in recent health publications' (Brnabic *et al.* 1999:82). To be fair to the ABS, users of the indexes are informed in the information paper that the indexes do not provide good measures for all social conditions and are urged to examine them to ascertain whether they are appropriate to their particular problem or analysis (ABS 1998:13). But how widely that advice will be taken on board is uncertain.

For the purposes of the discussion here public health research is used as the focus as particularly frequent use of the indexes has been made in that field. It should be stated to start

with that, as broad summary dimensions of socio-economic condition, the SEIFA indexes (and other composite indicators) can be useful constructs in describing relationships with various dimensions of health status. The Index of Relative Socio-Economic Disadvantage (IRSD) has been especially widely used in this manner. Recent major publications from the New South Wales Health Department (Public Health Division 1997, 2000), the NSW Cancer Council (Lewis *et al.* 1999), the Australian Institute of Health and Welfare (Mathers 1994; Mathers *et al.* 1999) and the National Centre for Epidemiology and Population Health and the Australian Bureau of Statistics (Jain 1994), for instance, all use the IRSD to show associations between area-defined socio-economic disadvantage and health status. Widespread use of the SEIFA indexes is also found in the academic public health research literature.¹ In some research and reporting broad descriptive association may be all that is required, but where the search is for *causes* and suggesting health promotion interventions broad SEIFA-style composite indexes are limited in what they have to offer.

For these latter uses such indexes have two principal shortcomings. First, they are blunt instruments. The central idea of principal components analysis (and other composite index constructing techniques) is to collapse down a set of variables into a summary indicator. This is the attraction and value of the principal components approach. However, this strength of the technique is also its weakness, the resulting summary indicator lacking the specificity to guide and inform etiological investigations and health promotion activities. The use of SEIFA scores has repeatedly confirmed that health status is related to socio-economic status, but offers little direction beyond that.

The second shortcoming is that the scores produced by principal components analysis are not totally unambiguous indicators. Due to the mathematics of the technique, areas sharing a particular score do not necessarily have a virtually identical mix of constituent variable values. In most instances they will share very similar constituent profiles, but this cannot be universally assumed to be the case. In other words, a similar SEIFA score may mask quite significant socio-economic differences between the areas concerned. This is akin to the situation with standardisation in demography where similar standardised rates may mask opposing specific rate patterns².

To illustrate these points the Index of Relative Socio-economic Disadvantage is used here, as this has been the most commonly employed SEIFA index, but the argument could equally be developed with reference to any of the other indexes. As noted in the introduction, the IRSD is based on 20 socio-economic variables. These variables cover family and household social and economic characteristics, plus personal education qualifications, occupation, ethnicity, English language fluency and marital status. Table 1 details the 20 variables, grouped by their 'weights' in the IRSD. The SEIFA information paper informs readers that 'the higher an area's index value for the Index of Relative Disadvantage, the less disadvantaged that area is compared with other areas' (ABS 1998:3).

For this discussion New South Wales local government areas (LGAs) have been chosen as the geographic study units and for each LGA the following variables assembled:

- (a) the IRSD score
- (b) the eleven leading variables comprising the IRSD
- (c) the age-sex standardised premature all causes mortality ratio (deaths under age 75, for 1995-97).

The idea of the data set is to examine the relationship between socio-economic disadvantage and premature mortality.

Correlating the IRSD and premature mortality values for all LGAs in the State (N = 178) reveals a moderately strong negative correlation ($r = -0.61$) between the two variables. For LGAs in the Sydney Statistical Division (N = 45) the association is slightly stronger ($r = -0.67$). Both analyses thus clearly show there is a statistically significant relationship ($p < 0.001$) between the variables, but beyond that tell very little. They do not give any real guidance for health policies and programmes. How, for example, do health promotion agencies run with these findings?

The variables included in the index are listed in the ABS information paper (and shown here in Table 1), with an indication of the contribution of each variable to the index. This however, tells nothing about how closely each of them is related to premature mortality. In short, one really needs to go and calculate the associations of premature mortality with the individual socio-economic variables. Table 2 summarises such an analysis, listing the correlations of the eleven strongest weighting variables in the IRSD with premature mortality. The correlation with the Aboriginal or Torres Strait Islander population, another (lower weighting) variable in the IRSD is also indicated.

In both the statewide and Sydney analyses considerable variation in strength of association with premature mortality is shown by the eleven leading individual variables. In this sense the IRSD, even though having fairly high correlations with most of the variables, is clearly not an across the board explanatory surrogate or guide for health promotion. For the State, the coefficients of the eleven variables range from 0.22 to 0.58. If the Aboriginal/Torres Strait Islander variable is considered the spread is even greater. In the case of Sydney the range is from 0.27 to 0.82, four of the leading eleven variables, plus the Aboriginal/Torres Strait Islander indicator, having closer associations with premature mortality than the composite index. Although in the State analysis the IRSD is the most successful predictor in terms of simple correlation size, it is only marginally so over some of the more focused variables. The low parental income variable's coefficient, for instance, is of almost similar strength ($r = 0.58$) and gives more direction to health planning.

The New South Wales LGA data also demonstrate the second shortcoming of the SEIFA scores claimed above, that is, they are not totally unambiguous indicators. To recap, areas sharing a particular score do not necessarily have virtually identical profiles on the constituent variables. The fact that the principal components used for each of the indexes only account for 'about 30%' of the variation in the respective data matrices by definition means that a good deal of variation is left unexplained. Grouping the LGAs into quintiles based on the IRSD scores shows low intra-quintile variability for the IRSD, but quite substantial variability for some of the constituent variables (Table 3).

Table 4 presents a number of examples of how this can translate to on the ground. Each of the shown pairs of LGAs have nearly identical IRSD values, but differ quite significantly on several of the constituent variables. The message here is the need to combine broad State and regional analyses and initiatives with detailed local area knowledge and planning.

To conclude this section, it should be stated that the above discussion is not designed to dismiss the use of the SEIFA (or similar composite) indexes, but rather to draw attention to their limitations. As stated, in some research contexts they can be useful constructs. However, it is important to recognise their shortcomings and blend their use with more focused indicators of the socio-economic environment.

Technical/methodological issues

Potential confusion over the 'weights' of the variables

As has been explained, the five SEIFA indexes are scores produced from principal components analyses of selected census-based socio-economic variables, the PRINCOMP algorithm of the SAS computing package being used to derive the scores.

Appendix A of the information paper (ABS 1998:18-24) presents a listing of the variables included in each of the indexes, along with the ranges within which the variables' 'weights' fall. It is stated that the weights indicate the contribution of each variable to the index. In the formal language of factor analysis (of which principal components is one method) these weights are in fact eigenvector elements. The term eigenvector however, is never used in the information paper. While not using the technical statistical term probably avoids scaring off some readers, its omission is also likely to cause some confusion.

The specific confusion being envisaged is that between 'weights' and component 'loadings'. Loadings are the correlations of the variables with the components and are central to component interpretation and evaluation. Given this centrality, loadings (unrotated and/or rotated) are virtually always presented in reporting principal components results, certainly in the demographic and other social science research with which the writer is familiar. Eigenvector matrices on the other hand are rarely printed out in published work.

For this reason, it is likely that many SEIFA users with a smattering of principal components knowledge will take the so-called weights to be component loadings, even though on reflection the values are clearly not sufficient to produce the stated 'about 30%' level of explained variance. In fact, many users will probably never have seen an eigenvector matrix as some of the standard computing programs do not give the option of printing out the matrix. Likewise, several of the factor analysis textbook *bibles* (e.g. Harman 1976; Rummel 1970) do not give the eigenvector matrices in any of their worked examples. The potential for this confusion was confirmed in email correspondence with the ABS, weights and loadings in one communication being defined as one and the same. Also, at one point in the information paper (p.26) there is reference to variable loadings.

These comments are not intended to criticise inclusion of the 'weights' (eigenvectors) in the information paper. As the paper states, they are useful in indicating the relative contribution of each variable to an index. The point is simply to argue that they should be clearly defined so readers can be in no doubt about what they are. As well, it would be worthwhile canvassing academic and other researchers around Australia as to which type of information, eigenvectors or loadings (or both), it would be most useful to publish in the next version of SEIFA.³

Advice on interpreting results

Recognising that many of the users of SEIFA will not be experts on principal components analysis the information paper provides a discussion of the derivation and interpretation of the indexes. This discussion is clearly written and should prove useful to those not familiar with the components algorithm. One point in the discussion however, is misleading, namely a confusion between low and negative weights (eigenvectors). Readers are thus told:

'The weights of the variables in each index also displayed face validity, i.e. they made intuitive sense (high income has a high weight, while low income or unemployment have low or negative weights; purchasing a dwelling has a higher weight than renting a dwelling; high rent has a higher weight than low rents; tertiary education has a higher weight than leaving school at 15 and so forth)' (ABS 1998:29).

The equating of low with negative is confusing the strengths of the weights with their directions. The tertiary education versus leaving school at 15 example crops up in the Index of Education and Occupation (p.23). While the signs of the variables' respective weights are different, their indicated strengths are the same (i.e. in the 0.2 to 0.4 and -0.2 to -0.4 bands). The same situation can be seen in the Index of Economic Resources where in the 0 to 0.2 and 0 to -0.2 weights bands there are purchasing and renting dwellings variables. The signs of the two variables' weights are different, but they are of the same strength. Thus readers need to be clear that a negative sign on a weight has nothing to do with magnitude.

Selection of variables

As the information paper notes, the derivation of socio-economic indexes is subjective in nature and thus no two researchers are likely to perfectly agree on which variables to include and exclude. Reading the information paper's discussion and scrutinising the variables finally included clearly shows considerable thought has gone into this process. One index though - the Index of Economic Resources - perhaps could have done with more culling. For example, the listing on p.22 of the information paper of the variables included in the index shows two of the variables to be households purchasing dwelling (%) and households owning dwelling (%). Then, further up the list is another variable measuring households owning or purchasing dwelling (%). The inclusion of three variables about bedrooms in the index could also perhaps be queried.

Information release and non-release

To fully evaluate principal components results it is necessary to have the complete output from all key stages of the computing procedure. In the case of the SEIFA96 indexes that would mean the full correlation and unrotated loadings (and/or eigenvectors) matrices in addition to the component scores. The loadings (and/or eigenvectors) are vital. Whichever is favoured, the loadings or the eigenvectors matrix, the full calculated matrix should be published to allow the delineated components to be closely examined.

As has been noted, the information paper lists the variables included in each index and indicates ranges within which their weights (eigenvectors) fall. While that is of some use, it also leaves a lot unknown. As stated earlier, readers are told the weights indicate the contribution of each variable to the index. However, by only having grouped values it is impossible to know whether a variable just crept across the line into a particular weight group, or whether it was right at the top of the group's interval.

A request was made to the ABS for the exact weights for the five indexes. However, the request was rejected, the ABS replying that there had been several prior requests for the exact values of the weights and that considerable thinking had been done on the arguments for and against releasing them, but that on balance a policy of non release had been adopted. It was explained that there were a number of reasons for this decision, the main one being

commercial. The weights, it was argued, are not essential for manipulation and application of the indexes and thus they were being excluded from the product to protect revenue which may flow from the sale of the indexes. Two points are worth making about this.

First, while the concern to protect revenue is perhaps understandable, the concern in this case would seem to have little foundation. Armed with the full eigenvector matrix any person trying to save him/herself the cost of purchasing the scores would have to go back to square one and create all the variables, standardise them and then multiply them up against the eigenvectors. That would be no small task and anyone up to that would probably do their whole own components analysis in the first place.

The second, and more important point, is the question of external scrutiny and validation of results. Whenever full scrutiny cannot be applied by outside researchers the research involved has to be categorised in a different light to that open to such examination and evaluation. Hopefully the SEIFA packages from future censuses will provide the complete principal components output to permit this scrutiny to occur.

Acknowledgement

The author would like to acknowledge the generous donation of time by the ABS to discussing some of the issues addressed in this paper. The discussions (by email) raised some interesting points and stimulated extra thinking on the subject by the author.

Footnotes

1. For example, see Cantor *et al.* (1995), Hyndman *et al.* (2000), Milligan *et al.* (1998), Smith *et al.* (1996) and Taylor *et al.* (1999).
2. See McCracken (1981).
3. The two sets of values are of course statistically related. If one knows the loadings the eigenvectors can be easily calculated, viz; eigenvectors = loadings divided by the square root of the corresponding eigenvalue. (The eigenvalue is the sum of the squared loadings on a component). In turn, loadings = eigenvectors multiplied by the square root of the corresponding eigenvalue.

References

- Australian Bureau of Statistics (ABS). 1998. *Information Paper, 1996 Census of Population and Housing, Socio-Economic Indexes for Areas*. Canberra: Australian Bureau of Statistics. Cat. No. 2039.0.
- Brnabic, A.J.M., J. Skinner, M. Staff, D. Small, L. March and D. Holt. 1999. *Health From the Harbour to the Hawkesbury: Demographic Update*. Hornsby: Public Health Unit - Northern Sydney Health.
- Cantor, C.H., P.J. Slater and J.M. Najman. 1995. Socioeconomic indices and suicide rate in Queensland. *Australian and New Zealand Journal of Public Health* 19(3): 417-420.
- Harman, H.H. 1976. *Modern Factor Analysis*. Chicago and London: University of Chicago Press.

- Hyndman, J.C.G., C. D'Arcy and J. Holman. 2000. Differential effects on socioeconomic groups of modelling the location of mammography screening clinics using Geographic Information Systems. *Australian and New Zealand Journal of Public Health* 24(3):281-286.
- Jain, S.K. 1994. *Trends in Mortality*. Canberra: National Centre for Epidemiology and Population Health and Australian Bureau of Statistics, Australian Bureau of Statistics. Cat. No. 3313.0.
- Lewis, N., H. Nguyen, D. Smith, M. Coates and B. Armstrong. 1999. *Cancer Maps for New South Wales: Variation by Local Government Area 1991 to 1995*. Sydney: NSW Cancer Council.
- Mathers, C., T. Vos and C. Stevenson. 1999. *The Burden of Disease and Injury in Australia*. Canberra: Australian Institute of Health and Welfare. Cat. No. PHE 17.
- Mathers, C. 1994. *Health Differentials Among Adult Australians Aged 25-64 Years*. Canberra: Australian Institute of Health and Welfare Health Monitoring Series No. 1, Australian Government Publishing Service.
- McCracken, K.W.J. 1981. Analysing geographical variations in mortality: age-specific and summary measures. *Area* 13:203-210.
- Milligan, R.A.K., V. Burke, L.J. Beilin, D.L. Dunbar, M.J. Spencer, E. Balde and M.P. Gracey. 1998. Influence of gender and socioeconomic status on dietary patterns and nutrient intakes in 18-year old Australians. *Australian and New Zealand Journal of Public Health* 22(4):485-493.
- Public Health Division. 2000. *The Health of the People of New South Wales - Report of the Chief Health Officer, 2000*. Sydney: NSW Health Department.
- Public Health Division. 1997. *The Health of the People of New South Wales - Report of the Chief Health Officer*. Sydney: NSW Health Department.
- Rummel, R.J. 1970. *Applied Factor Analysis*. Evanston: Northwestern University Press.
- Smith, D., R. Taylor and M. Coates. 1996. Socioeconomic differentials in cancer incidence and mortality in urban New South Wales, 1987-1991. *Australian and New Zealand Journal of Public Health* 20(2):129-137.
- Taylor, R., T. Chey, A. Bauman and I. Webster. 1999. Socioeconomic, migrant and geographic differentials in coronary heart disease occurrence in New South Wales. *Australian and New Zealand Journal of Public Health* 23(1):20-26.

Table 4 Selected pairs of local government areas with nearly identical SEIFA Index of Relative Socio-Economic Disadvantage scores

<i>LGA Description</i>	Sydney	Gosford	South		Hawkesbury	Windouran	Cowra
			Sydney	Coolamon			
	<i>Central Location</i>	<i>Outer Metropolitan</i>	<i>Central Metropolitan</i>	<i>Central Murrumbidgee</i>	<i>Outer Metropolitan</i>	<i>Central Murray</i>	<i>Central West</i>
<i>Urban/Rural</i>	<i>100% Urban</i>	<i>92% Urban</i>	<i>100% Urban</i>	<i>67% Rural</i>	<i>61% Urban</i>	<i>100% Rural</i>	<i>70% Urban</i>
SEIFA Index of Relative Socio-Economic Disadvantage (IRSD)	1010.4	1010.1	993.3	993.8	1036.0	1035.4	964.9
<i>Variables with heaviest weights on IRSD</i>							
Persons aged 15 and over with no qualifications (%)	57.9	56.0	47.6	67.9	55.6	61.8	64.1
Families with income less than \$15,600 (%)	18.5	13.6	15.7	17.9	9.8	20.7	18.5
Families with offspring having parental income less than \$15,600 (%)	24.8	8.7	18.9	8.4	7.5	16.7	13.2
Females (in labour force) unemployed (%)	7.7	7.7	8.3	7.4	5.3	10.6	8.7
Males (in labour force) unemployed (%)	7.6	9.6	10.1	10.5	5.6	6.3	9.6
Employed females classified as 'Labourer & Related Workers' (%)	3.1	7.1	3.9	9.8	8.3	3.7	12.2
Employed Males classified as 'Labourer & Related Workers' (%)	4.4	9.0	6.3	10.8	8.6	25.4	17.2
Employed Males classified as 'Intermediate Production and Transport Workers' (%)	4.0	10.8	6.9	10.8	14.1	6.9	12.5
Persons aged 15 and over who left school at or under 15 years of age (%)	21.4	43.5	22.9	49.9	37.3	28.9	46.0
One parent families with dependent offspring only (%)	10.7	9.3	11.5	4.1	9.0	2.8	7.9
Households renting (government authority) (%)	13.8	3.8	17.3	0.5	4.2	0.0	6.1

Note: Shaded cells indicate areas whose values differ by $Z \geq 2$ (Z-scores based on N=178 LGAs).