doi:10.1093/heapol/czl029 Advance Access publication 9 October 2006

How to do (or not to do) . . .

Constructing socio-economic status indices: how to use principal components analysis

SEEMA VYAS AND LILANI KUMARANAYAKE
HIVTools Research Group, Health Policy Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK

Theoretically, measures of household wealth can be reflected by income, consumption or expenditure information. However, the collection of accurate income and consumption data requires extensive resources for household surveys. Given the increasingly routine application of principal components analysis (PCA) using asset data in creating socio-economic status (SES) indices, we review how PCA-based indices are constructed, how they can be used, and their validity and limitations. Specifically, issues related to choice of variables, data preparation and problems such as data clustering are addressed. Interpretation of results and methods of classifying households into SES groups are also discussed. PCA has been validated as a method to describe SES differentiation within a population. Issues related to the underlying data will affect PCA and this should be considered when generating and interpreting results.

Key words: socio-economic status, principal components analysis, cluster analysis, methodology

1. Introduction

Common to health research and policy interventions is the concern that there is a differential impact with respect to health outcomes or health service utilization based on socio-economic status (SES) (Deaton 2003; Schellenberg et al. 2003). Thus, information about how households vary by SES, and the extent to which this relates to variables of interest, is central to questions such as how to target the poorest. Standard economic measures of SES use monetary information, such as income or consumption expenditure. However, the collection of accurate income data is a demanding task (Montgomery et al. 2000), requiring extensive resources for household surveys; for example, allowances need to be made for households and individuals drawing income from multiple sources. Also, in some instances, an indicator of income is quite difficult to use (Cortinovis et al. 1993). For example, income information does not capture the fact that people (and especially the poor) may have income in kind, such as crops which are traded, and measuring income can be difficult for the self or transitory employed (e.g. agricultural work), due to accounting issues and seasonality (McKenzie 2003).

By comparison, consumption or expenditure measures are much more reliable and are easier to collect than income, especially in most rural settings (Filmer and Pritchett 2001). However, again a limitation is the extensive data collection required, which is time-consuming and therefore costly. Given the resource constraints to measuring household income or expenditure in low- and middle-income country settings, other methods of developing SES indices are being used which streamline the variables required, enabling data to be collected more rapidly. Rather than income or expenditure, data are collected for variables that capture living standards, such as household ownership of durable assets (e.g. TV, car) and infrastructure and housing characteristics (e.g. source of water, sanitation facility).

While asset-based measures are increasingly being used, there continues to be some debate about their use. Importantly, a key argument revolves around their interpretation. These measures are more reflective of longer-run household wealth or living standards, failing to take account of short-run or temporary interruptions, or shocks to the household (Filmer and Pritchett 2001). Therefore, if the outcome of interest is associated with current resources available to the household, then an index based on assets may not be the appropriate measure.

Falkingham and Namazie (2002) highlight a second issue which is that ownership does not always capture the quality of assets. For example, collecting information on TV ownership does not distinguish between better-off households that are more likely to own a newer or colour TV, and less well-off households that may own an older or black and white one. However, they also point out that in many countries, this would not alter the overall picture of wealth.

A third issue is that some variables may have a different relationship with SES across sub-groups; for example, ownership of farmland may be more reflective of wealth in rural areas.

A final issue is how to aggregate over the range of different variables to derive a uni-dimensional measure of SES, and produce a range of critical points differentiating socio-economic levels. This is because each variable, used individually, may not be sufficient to differentiate household SES. One approach has been to sum the number of assets in households, for example Montgomery et al. (2000), but this assumes that all assets should be weighted equally. More recently, studies have applied principal components analysis (PCA) to such data to derive a SES index (Gwatkin et al. 2000; Filmer and Pritchett 2001; McKenzie 2003), and then grouped households into pre-determined categories, such as quintiles, reflecting different SES levels.

Given the increasingly routine application of PCA using asset data in creating SES indices, we review how PCA-based indices are constructed and how they can be used, and assess their advantages and limitations by presenting a worked example. PCA is explained in section 2, and construction and how to use a SES index is demonstrated in section 3, with data from both urban and rural settings. An evaluation of PCA-based indices is undertaken in section 4.

2. What is PCA?

PCA is a multivariate statistical technique used to reduce the number of variables in a data set into a smaller number of 'dimensions'. In mathematical terms, from an initial set of n correlated variables, PCA creates uncorrelated indices or components, where each component is a linear weighted combination of the initial variables. For example, from a set of variables X_1 through to X_n,

$$\begin{aligned} \mathrm{PC}_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \\ &\vdots \\ \mathrm{PC}_m &= a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mn}X_n \end{aligned}$$

where a_{mn} represents the weight for the mth principal component and the nth variable.

Diagrammatically, the concept of PCA can be shown as in Figure 1. The uncorrelated property of the components is highlighted by the fact they are perpendicular, i.e. at right angles to each other, which mean the indices are measuring different dimensions in the data (Manly 1994).

The weights for each principal component are given by the eigenvectors of the correlation matrix, or if the original data were standardized, the co-variance matrix.

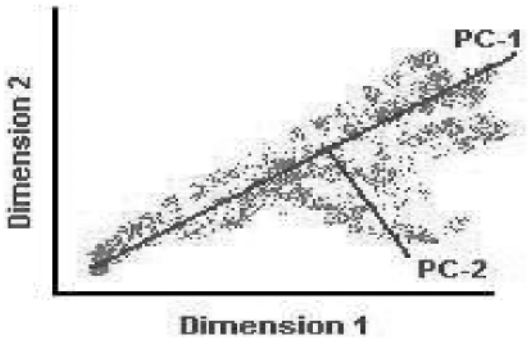

Figure 1. Representation of two sequential components in PCA. *Source*: [http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/PCA_1.htm], accessed 7 September 2005

The variance (λ) for each principal component is given by the eigenvalue of the corresponding eigenvector.[1] The components are ordered so that the first component (PC_1) explains the largest possible amount of variation in the original data, subject to the constraint that the sum of the squared weights ($a_{11}^2 + a_{12}^2 + \cdots + a_{1n}^2$) is equal to one. As the sum of the eigenvalues equals the number of variables in the initial data set, the proportion of the total variation in the original data set accounted by each principal component is given by λ_i/n. The second component (PC_2) is completely uncorrelated with the first component, and explains additional but less variation than the first component, subject to the same constraint. Subsequent components are uncorrelated with previous components; therefore, each component captures an additional dimension in the data, while explaining smaller and smaller proportions of the variation of the original variables. The higher the degree of correlation among the original variables in the data, the fewer components required to capture common information.

3. Constructing a SES index with PCA

Using data from the Demographic Health Survey (DHS) (from [http://www.measuredhs.com]), PCA-based SES measures are derived in this section for two contrasting countries, Brazil and Ethiopia.[2] DHS household surveys have been undertaken in more than 60 countries, focusing on health outcomes and nutrition, and contain data on household characteristics rather than income or expenditure. The World Bank, in its series of 'Socio-economic differences in health, nutrition, and population', has also constructed PCA-based asset indices using DHS data (e.g. Gwatkin et al. 2000), constructing an index for each country as a whole. In our example, we construct a socio-economic index for each site, that is, households in urban and rural locations in both countries, to illustrate some of the issues that arise when using and interpreting PCA-based SES. Standard statistical software can be used and in this instance, STATA (Version 8.1) was used.

We divide this section into four parts to reflect the main steps in constructing a SES index: selection of asset

variables; application of PCA; interpretation of results; and classification of households into socio-economic groups. The first part examines the issues relating to the choice of assets and variables that have been commonly used, in particular, clumping and truncation, stability of household classification and reliability. The second highlights methodological issues such as preparation of data, and identifying the number of principal components to extract that would measure SES. Results of a PCA analysis on asset data are interpreted in the third subsection, and the methods used to classify households into socio-economic groups are presented in the fourth.

3.1 Selection of asset variables

To measure SES, studies have used variables such as ownership of land (Filmer and Pritchett 2001), farm animals and whether living in rented or owner-occupied housing (Schellenberg et al. 2003), literacy or education level of head of household, demographic conditions (e.g. the ratio of number of people to the number of rooms in the household to proxy crowding), and other economic proxies such as occupation of head of household (Cortinovis et al. 1993). Montgomery et al. (2000) identified the absence of a 'best practice' approach of selecting variables to proxy living standards, as, in many studies, variables were chosen on an 'ad-hoc' basis.

In the DHS, information is collected on durable asset ownership, access to utilities and infrastructure (e.g. sanitation facility and source of water), and housing characteristics (e.g. number of rooms for sleeping and building material), which we include in our analysis.

PCA works best when asset variables are correlated, but also when the distribution of variables varies across cases, or in this instance, households. It is the assets that are more unequally distributed between households that are given more weight in PCA (McKenzie 2003). Variables with low standard deviations would carry a low weight from the PCA; for example, an asset which all households own or which no households own (i.e. zero standard deviation) would exhibit no variation between households and would be zero weighted, and so of little use in differentiating SES.

Therefore, as a first step, we carried out descriptive analyses for all the variables, looking at means, frequencies and standard deviations (see Table 1). Descriptive analysis can inform decisions on which variables to include in the analysis, and highlight data management issues, such as coding of variables and missing values. In rural Brazil and urban Ethiopia, indicators of durable asset ownership range from the majority of households owning a radio to a few owning a car. Also, the source of water supply in rural Brazil, and type of floor material in urban Ethiopia, vary across households. In urban Brazil, the vast majority of households owned all or most of the assets listed, and had a tap in residence, though there is variation in type of sanitation facility. However, in rural Ethiopia, few households have assets or any formal sanitation facility, and most have rudimentary types of flooring material (e.g. earth or sand, dung).

McKenzie (2003) highlights that a major challenge for PCA-based asset indices is to ensure the range of asset variables included is broad enough to avoid problems of 'clumping' and 'truncation'. Clumping or clustering is described as households being grouped together in a small number of distinct clusters. Truncation implies a more even distribution of SES, but spread over a narrow range, making differentiating between socio-economic groups difficult (e.g. not being able to distinguish between the poor and the very poor). From the distribution of asset ownership, access to utilities and infrastructure, and housing characteristics in our analysis, clumping and truncation are likely to be issues for the data from rural Ethiopia. This is because many households do not own the durable items, have similar access to utilities and infrastructure, and similar housing characteristics, and so will be grouped together. Also, of the households that do own assets, they have the same ones, which will make differentiating among them difficult. Clumping and truncation may be an issue for urban Brazil due to high levels of ownership of most of the included durable assets, but we do not expect it to be an issue for rural Brazil or urban Ethiopia.

If clumping and truncation are identified as potential problems from the descriptive analysis, as is the case for rural Ethiopia, then one method that could solve this issue is to add more variables to the analysis. The number of variables used in studies has ranged from 10 (Schellenberg et al. 2003) to 30 (McKenzie 2003). Other methods could be to use continuous variables (e.g. the number of acres of land) and using a combination of asset durable ownership, access to utilities and infrastructure, housing characteristics and other variables that appear relevant in assessing household wealth. A preliminary analysis correlating assets and monthly household expenditure was used to inform the choice of indicators to be collected in a study by Hanson et al. (2005). The analysis used the Living Standards Measurement Survey which collected information on both expenditure and asset data. Only asset variables that were significantly correlated with expenditure were included in their subsequent survey.

However, the key is to include additional variables that capture inequality between households. McKenzie (2003) compared SES distributions using housing characteristics only, access to utilities and infrastructure only, durable asset ownership only and all three categories of variables. For both the housing characteristics only and utilities only distributions, there was evidence of clumping and truncation, while durable asset ownership showed some evidence of truncation. The index based on combined assets showed no evidence of clumping or truncation and yielded the smoothest distribution of SES.

Another issue related to selection of asset variables is the stability of household classification into SES groups.

Table 1. Results from principal components analysis

Variable description Brazil/Ethiopia	Brazil urban			Brazil rural			Ethiopia urban			Ethiopia rural		
	Mean	Std. dev.	Factor score	Mean	Std. dev.	Factor score	Mean	Std. dev.	Factor score	Mean	Std. dev.	Factor score
Electricity	0.987	0.114	0.158	0.694	0.461	0.347	0.829	0.376	0.297	0.012	0.107	0.171
Radio	0.881	0.323	0.216	0.765	0.423	0.171	0.689	0.463	0.294	0.139	0.345	0.210
Television	0.716	0.450	0.372	0.314	0.463	0.345	0.215	0.411	0.327	0.000	0.010	0.024
Refrigerator	0.821	0.383	0.363	0.425	0.493	0.397						
Car	0.296	0.456	0.295	0.135	0.341	0.256	0.035	0.184	0.176			
Bicycle										0.003	0.058	0.106
Telephone							0.136	0.343	0.291			
No. of rooms for sleeping	2.150	0.899	0.143	2.175	0.916	0.105						
Source of water supply												
Piped into residence/dwelling	0.760	0.427	0.243	0.200	0.400	0.179	0.007	0.086	0.033			
Piped into yard, plot/compound	0.044	0.204	−0.182	0.051	0.219	−0.033	0.414	0.493	0.367	0.001	0.024	0.105
/Piped outside compound							0.441	0.497	−0.221	0.061	0.239	0.092
Well, spring inside/covered well	0.075	0.264	−0.126	0.381	0.485	0.096	0.012	0.108	−0.077	0.065	0.247	0.103
Well or spring outside/open well	0.054	0.227	0.122	0.276	0.447	−0.154	0.049	0.217	−0.060	0.072	0.259	−0.106
Bottled water/	0.047	0.212	0.062									
/Covered, open spring							0.023	0.151	−0.103	0.427	0.495	0.071
/River							0.048	0.214	−0.152	0.333	0.471	−0.108
Other/and pond, lake, dam, rain	0.019	0.138	−0.135	0.092	0.289	−0.143	0.005	0.070	−0.011	0.041	0.199	−0.033
Sanitation facility												
Toilet to sewer/flush toilet	0.410	0.491	0.277	0.059	0.235	0.109	0.035	0.184	0.147			
Toilet to open space or river/	0.054	0.225	−0.089	0.093	0.290	0.057						
Latrine to sewer/	0.128	0.334	0.062	0.035	0.182	0.098						
Latrine no connection/	0.217	0.411	−0.049	0.176	0.380	0.210						
Traditional latrine/pit	0.138	0.345	−0.184	0.218	0.412	0.133	0.714	0.452	0.218	0.096	0.294	0.580
/Ventilated improved pit latrine							0.033	0.177	0.056	0.001	0.026	0.077
No facility/and bush or field	0.053	0.224	−0.238	0.420	0.493	−0.395	0.218	0.413	−0.328	0.904	0.295	−0.586
Type of floor material												
Earth or sand	0.032	0.175	−0.175	0.191	0.393	−0.294	0.345	0.475	−0.312	0.696	0.460	−0.263
/Dung							0.108	0.310	−0.121	0.283	0.451	0.228
Wood planks/and reed or bamboo	0.070	0.255	0.004	0.059	0.236	0.096	0.028	0.166	0.037	0.002	0.048	0.065
Polished wood/and parquet	0.097	0.296	0.116	0.071	0.256	0.161	0.237	0.425	0.131			
Vinyl/and sheet tiles	0.007	0.082	0.043	0.004	0.063	0.049	0.030	0.171	0.166			
Ceramic tiles/and brick	0.317	0.465	0.275	0.083	0.277	0.192	0.040	0.195	0.095			
Cement	0.436	0.496	−0.309	0.568	0.495	−0.009	0.170	0.375	0.150	0.005	0.074	0.205
Carpet	0.036	0.186	0.090	0.008	0.091	0.062	0.026	0.160	0.071	0.013	0.112	−0.007
Other	0.006	0.076	0.000	0.016	0.124	−0.048	0.016	0.125	0.007	0.000	0.022	0.049

In some studies, this has been found to be closely associated with the choice of variables included in the index. For example, Houweling et al. (2003) compared the relative economic position of households using either durable assets, infrastructure, housing characteristics or a combination of all variables to derive four different PCA-based measures. However, Filmer and Pritchett (2001), in their analysis, concluded the categorization of households was robust to the measure used.

In addition, Houweling et al. (2003) found that variables included in the index that were directly associated with child health outcomes (e.g. sanitation facility) increased inequality among households. Similarly, Lindelow (2002) found including infrastructure variables such as source of water increased socio-economic inequality in health facility utilization. Higher quality infrastructure variables were geographically biased to urban locations where access to health facilities is assumed to be greater. Including infrastructure variables in the index increased the representation of households from urban areas into the richer groups, and subsequently increased inequality. An explanatory analysis should consider an index without direct determinants of the outcome of interest. However, exclusion of variables may make it more difficult to divide households, particularly when considering similar groups, for example in a rural community.

An advantage of collecting asset data, highlighted by McKenzie (2003), is that measurement error is minimized. Onwujekwe et al. (2006) report on the reliability of collecting some asset data commonly used in generating SES indices (e.g. radio, bicycle). Two methods of assessing reliability were used. The first employed two different interviewers to measure observations separated by up to 5 days (inter-rater), and the second employed the same interviewer to measure observations within 1 month of the original survey being administered (test-retest). In both cases, reliability was found not to be high, and resulted in differences in classification of households

into SES groups. Therefore, the user should be aware of issues relating to the accuracy of data collection. A possible way to improve reliability is to include assets that are observable by interviewers, but this may not always be feasible.

3.2 Application of PCA

Data in categorical form (such as religion) are not suitable for PCA, as the categories are converted into a quantitative scale which does not have any meaning.[3] To avoid this, qualitative categorical variables should be re-coded into binary variables. In our example, similar variables with low frequencies were combined together: for example, 'covered spring' and 'spring' were combined for Ethiopia data; 'toilet to open space' and 'toilet to river' for Brazil data. Similar variables with relatively high frequencies were kept as separate variables (spring and river for Ethiopia data). We included all binary variables created from a categorical variable, including those that had low frequencies but were not similar enough to another variable to combine, in order to ensure all the data for each household were measured. We excluded durable assets that were initially binary with very low counts, for instance motorcycle in urban Ethiopia, which was owned by 0.1% of households.

Another data issue is that of missing values. Cortinovis et al. (1993) excluded households with at least one missing value from their analysis to develop socio-economic groups. Gwatkin et al. (2000) replaced missing values with the mean value for that variable. Exclusion of households based on missing socio-economic data could significantly lower sample sizes and the statistical power of study results, and may lead to bias towards higher SES households, as missing data may occur more frequently in lower social classes (Cortinovis et al. 1993). However, attributing mean scores for missing values reduces variation among households, and increases the potential for clumping and truncation. This is more pronounced with high numbers of missing values, though software packages such as STATA offer a range of methods for estimating missing values. In our example, the percentage of households with missing data was small (less than 1% in each site). We expect inclusion or exclusion of these households would have little impact on the distribution of SES, but for variables with missing values, we chose to impute the mean value of that variable.

The analysis of data on household characteristics and asset ownership is complicated by the fact that there are potentially a large number of variables which could be collected, some of which may yield similar information. Thus a natural approach is to use methods such as PCA to try and organize the data to reduce its dimensionality with as little loss of information as possible in the total variation these variables explain (Giri 2004).

In STATA, when specifying PCA, the user is given the choice of deriving eigenvectors (weights) from either the correlation matrix or the co-variance matrix of the data. If the raw data has been standardized, then PCA should use the co-variance matrix.[4] As we did not standardize our data, and they are therefore not expressed in the same units, we ran the analysis using the correlation matrix to ensure that all data have equal weight. For example, the number of rooms for sleeping is a quantitative variable and has greater variance than the other binary variables, and would therefore dominate the first principal component if the co-variance matrix was used.

The number of principal components extracted can also be defined by the user, and a common method used is to select components where the associated eigenvalue is greater than one. However, it is assumed that the first principal component is a measure of economic status (Houweling et al. 2003). McKenzie (2003) considered the use of additional principal components in characterizing household SES, investigating whether they related to non-durable consumption, and concluded that only the first principal component was necessary for measuring wealth. Filmer and Pritchett (2001) also considered the use of additional components in their analysis, and though they found the factor scores for each variable difficult to interpret, they included 'higher order' components in a multivariate regression analysis, and concluded their results were robust to including additional components.

The eigenvalue (variance) for each principal component indicates the percentage of variation in the total data explained. In the studies included in this review, the first principal component accounted for a range from 12% (Houweling et al. 2003) to 27% (McKenzie 2003) of total variation. These percentages are not high, and this could reflect the number of variables included in the analysis or the complexity of correlations between variables, as each included variable may have its own determinant other than SES.

Results from the first principal component for each site are shown in Table 1, and their associated eigenvalues are 4 (rural Brazil and urban Ethiopia), 3.5 (urban Brazil) and 2.2 (rural Ethiopia), accounting for 16.0%, 14.9%, 13.4% and 11.1%, respectively, of the variation in the original data.

3.3 Interpretation of results

The output from a PCA is a table of factor scores or weights for each variable (see Table 1). Generally, a variable with a positive factor score is associated with higher SES, and conversely a variable with a negative factor score is associated with lower SES. It is useful to note that in some studies, ownership of durable assets such as a bicycle have been attributed a negative weight from PCA (Gwatkin et al. 2000; Houweling et al. 2003; McKenzie 2003). This implies, all things being equal, that a household with a bicycle will be ranked lower in terms of SES than a household that does not own a bicycle. The reason for such a result may be due to ownership of a bicycle being more strongly correlated with variables that are expected to be associated with lower SES, for instance

lower quality housing and sanitation conditions. Findings like these can occur when indices have been constructed for combined urban and rural locations, or regions, where the asset represents wealth in some parts of the country but not others. However, in Gwatkin et al. (2000) and McKenzie (2003), the weights for ownership of a bicycle were among the smallest in absolute terms compared with other durable assets, and Houweling et al. (2003) argued their finding was not likely to have influenced their overall conclusions.[5]

As we constructed a separate index for urban and rural locations in both countries, we find for each site the factor scores are positive for all durable assets, as is usage of higher quality source of water and sanitation facility (relative to the alternative available). Low quality type of flooring (e.g. earth or sand) has a negative factor score in all sites.

As a further analysis, we considered an additional principal component. The second principal component showed that for urban Brazil the weights were concentrated on source of water, and on floor type for rural Brazil and rural Ethiopia. For urban Ethiopia, the weights were concentrated on sanitation facility. In all cases, the second principal component explained a sub-group of variables. Therefore, we conclude that the first principal component provided a measure of wealth.

Using the factor scores from the first principal component as weights, a dependent variable can then be constructed for each household (Y_1) which has a mean equal to zero, and a standard deviation equal to one. This dependent variable can be regarded as the households 'socio-economic' score, and the higher the household socio-economic score, the higher the implied SES of that household. The issue of adjusting for household size was raised by McKenzie (2003). As in the study by Filmer and Pritchett (2001), McKenzie (2003) does not adjust for household size, arguing the benefits of indicators used are available at household level.

Interpreting the weights from our example, an urban Brazil household with more assets, piped drinking water to residence, sanitation facility that leads to a sewer, finished floor coverings and higher number of rooms for sleeping would attain a higher SES score. The finding is similar for rural Brazil, except it includes any sanitation facility and a well in residence. In urban Ethiopia, a household with more assets and drinking water piped to compound would attain a higher SES score. In rural Ethiopia, ownership of any asset, or access to infrastructure facilities such as water or sanitation, would lead to a higher SES score.

3.4 Classification of households into socio-economic groups

The constructed household socio-economic score (Y_1) could be included as a continuous independent variable in a regression model, though the estimated coefficient may not be easy to interpret. Other studies have used cut-off points to differentiate households into broad socio-economic categories, and the approaches used were either arbitrarily defined (based on the assumption SES is uniformly distributed), or data driven. Commonly used arbitrary cut-off points are classification of the lowest 40% of households into 'poor', the highest 20% as 'rich' and the rest as the 'middle' group (Filmer and Pritchett 2001), or the division of households into quintiles (Gwatkin et al. 2000). We classified households into quintiles and calculated the mean socio-economic score for each group (Table 2), because if SES is uniformly distributed, the difference in mean socio-economic score between adjoining quintiles should be even. The differences in the average scores were even for rural Brazil and urban Ethiopia. The mean difference is higher between the poorest and second poorest group for urban Brazil than any other adjoining quintile. For rural Ethiopia, the difference is small among the poorest three quintiles, as each group has a similar mean score.

Table 2. Mean socio-economic score by quintile

Site	N	Poorest	Second	Middle	Fourth	Richest
Urban Brazil	10 527	−2.96	−0.82	0.35	1.33	2.14
Rural Brazil	2756	−2.68	−1.44	−0.01	1.40	2.80
Urban Ethiopia	3629	−2.82	−1.17	0.02	1.22	2.83
Rural Ethiopia	10 443	−1.08	−0.72	−0.43	0.20	2.85

Internal coherence compares the mean value for each asset variable by socio-economic group, in our example, quintiles. Filmer and Pritchett (2001) and McKenzie (2003) examined internal coherence of the asset-based index in their studies, and both found mean asset ownership differed by socio-economic group. In our example, ownership of all asset variables, piped water in residence and toilet to sewer increased by socio-economic group in urban and rural Brazil. For example, 31% and 0.2% of households owned a refrigerator in the poorest quintile (urban and rural Brazil, respectively), compared with over 99% in the richest quintiles in both sites (data not shown). In urban Ethiopia, ownership of all assets (except telephone), piped water in residence, tap in compound and use of a flush toilet increased by socio-economic group. In rural Ethiopia, access to a pit latrine increased by socio-economic group, and the proportion of households reporting no sanitation facility decreased by socio-economic group. However, there was no clear trend by socio-economic group of sources of water or most types of floor material (Table 3).

While we find there is evidence for internal coherence for urban and rural Brazil and urban Ethiopia, we cannot conclude the index to be internally coherent for rural Ethiopia.

The assumption that the distribution of SES is quite uniform may not be appropriate in all settings, for example in rural Ethiopia. Histograms of the household socio-economic scores for each site are shown in Figure 2.

Table 3. Ownership of durable assets and housing characteristics by SES quintile

Variable description	Urban Ethiopia					Rural Ethiopia				
	Poorest	Second	Middle	Fourth	Richest	Poorest	Second	Middle	Fourth	Richest
Electricity	0.350	0.824	0.980	0.997	1.000	0.000	0.000	0.000	0.000	0.074
Radio	0.237	0.621	0.741	0.864	0.990	0.000	0.000	0.284	0.178	0.407
Television	0.000	0.007	0.039	0.170	0.869	0.000	0.000	0.000	0.000	0.001
Car	0.000	0.000	0.003	0.010	0.165					
Bicycle						0.000	0.000	0.000	0.000	0.022
Telephone	0.000	0.001	0.110	0.059	0.614					
Source of water supply										
In-residence tap	0.001	0.003	0.001	0.014	0.018					
In-compound tap	0.008	0.055	0.300	0.782	0.949	0.000	0.000	0.000	0.000	0.004
Out-of-compound tap	0.590	0.732	0.643	0.187	0.033	0.000	0.000	0.295	0.027	0.135
Covered well	0.062	0.039	0.014	0.001	0.000	0.000	0.000	0.295	0.020	0.175
Open well	0.034	0.019	0.005	0.000	0.000	0.220	0.003	0.118	0.002	0.027
Covered spring	0.118	0.099	0.022	0.006	0.000	0.000	0.893	0.022	0.599	0.384
River	0.182	0.047	0.007	0.003	0.000	0.780	0.001	0.238	0.320	0.244
Other water	0.004	0.005	0.008	0.007	0.000	0.000	0.104	0.031	0.032	0.030
Sanitation facility										
Flush toilet	0.000	0.000	0.005	0.028	0.144					
Traditional pit latrine	0.190	0.750	0.924	0.917	0.792	0.000	0.000	0.000	0.000	0.615
Ventilated improved pit latrine	0.000	0.028	0.023	0.051	0.061	0.000	0.000	0.000	0.000	0.004
No sanitation facility	0.810	0.222	0.047	0.004	0.003	1.000	1.000	1.000	1.000	0.381
Type of floor material										
Earth floor	0.795	0.614	0.270	0.027	0.003	1.000	0.998	0.877	0.169	0.439
Dung floor	0.194	0.191	0.099	0.050	0.001	0.000	0.000	0.053	0.821	0.498
Wood floor	0.001	0.016	0.043	0.047	0.035	0.000	0.000	0.000	0.000	0.014
Polished wood/parquet floor	0.000	0.005	0.003	0.058	0.135					
Vinyl floor	0.001	0.038	0.218	0.269	0.328					
Ceramic/tiles/brick floor	0.001	0.001	0.007	0.038	0.085					
Cement floor	0.004	0.108	0.305	0.431	0.344	0.000	0.000	0.000	0.000	0.035
Carpet floor	0.000	0.011	0.030	0.045	0.065	0.000	0.002	0.070	0.011	0.010
Other floor	0.001	0.015	0.024	0.035	0.004	0.000	0.000	0.000	0.000	0.003

The distribution of scores tends to follow a normal curve for rural Brazil and urban Ethiopia. For urban Brazil, it is skewed to the left. For rural Ethiopia, it is heavily skewed to the right, highlighting the extent of clumping and truncation which have made it difficult to differentiate between socio-economic groups.

A data driven approach to classifying households is cluster analysis, as used in a study by Cortinovis et al. (1993). Cluster analysis is a statistical procedure that allows for assignment of cases to a fixed number of groups or clusters according to a set of variables. The procedure attempts to group and derive cluster centres. The difference between cluster means is made as large as possible. We used cluster analysis on the household socio-economic score derived for each site to investigate the distribution of 'low', 'medium' and 'high' socio-economic groups (Table 4). Cluster analysis generally fitted the patterns found from the distribution of the household socio-economic scores shown in the histograms. So in our case, applying arbitrary cut-off points, such as the 40–40–20 split as in Filmer and Pritchett (2001), would disaggregate the distribution for, for example, urban Ethiopia, but it would not reflect the clustered nature of the underlying data for rural Ethiopia.

To summarize, for rural Brazil and urban Ethiopia, the distribution of socio-economic scores show little evidence of clumping and truncation, suggesting appropriate and sufficient choice of variables, and the results were found to be internally coherent across quintiles. While the results for urban Brazil were internally coherent, there is some evidence of truncation at the top, suggesting the variables included in the analysis were not sufficient to distinguish households among the rich.

For rural Ethiopia, the distribution of SES was heavily skewed, reflected by almost 60% of households being classified into the low socio-economic group using cluster analysis. The example for rural Ethiopia has highlighted the difficulties of using asset-based indices in some settings. Clumping or truncation can result from using variables which are unable to distinguish households, or it could reflect that households are in fact homogenous in terms of SES.

The decision on whether to construct a socio-economic index at country level (e.g. Gwatkin et al. 2000) or at community level (e.g. Schellenberg et al. 2003) depends on the objectives of the study and the comparisons to be made. Constructing an index at country level risks failing to capture wealth differences in, for example,

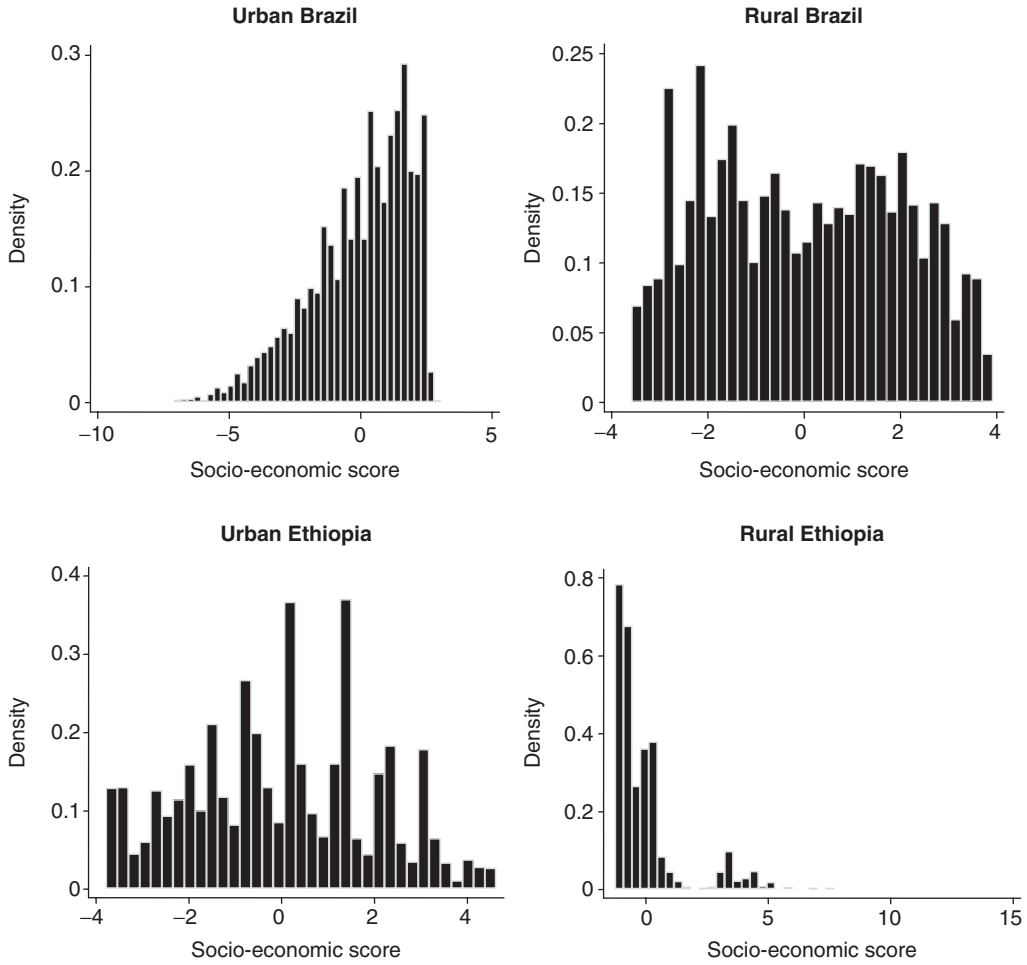

Figure 2. Distribution of socio-economic scores

Table 4. Proportion of households in low, medium and high socio-economic group for entire sample

Site	N	Low (%)	Medium (%)	High (%)
Urban Brazil	10 527	17.77	36.28	45.95
Rural Brazil	2 756	35.92	29.75	34.33
Urban Ethiopia	3629	38.58	40.20	21.22
Rural Ethiopia	10 443	59.26	30.73	10.01

rural or regional communities, and constructing an index at community level increases the risk of clumping and truncation. If the analysis is to be undertaken for a rural community, Houweling et al. (2003) advise including items associated with SES for that location. Planning surveys before hand, and using local knowledge to pick out variables that could discriminate households into groups, could help to determine such a list of indicators. However, there will continue to be a trade-off in terms of the additional expense of obtaining more specialized data for a particular setting, and the simplicity of using asset-based measures.

4. Discussion

This paper describes the process to derive a SES index in the absence of income or consumption data by performing PCA on durable asset ownership, access to utilities and infrastructure, and housing characteristic variables. The main advantage of this method over the more traditional methods based on income and consumption expenditure is that it avoids many of the measurement problems associated with income- and consumption-based methods, such as recall bias, seasonality and data collection time. Compared with other statistical alternatives, PCA is computationally easier, can use the type of data that can be more easily collected in household surveys, and uses all of the variables in reducing the dimensionality of the data (Jobson 1992). Socio-economic categorization is obtained by ranking then classifying households within the distribution into various groupings. The indices derived are relative measures of SES, so while this type of measure is useful for considering inequality between households, it cannot provide information on absolute levels of poverty within a community (McKenzie 2003). It can be used for comparison across countries or settings (such as urban/rural), or over time, provided the separate indices are calculated with the same variables.

Debate about the use of PCA reflects the fact that principal components are artificially constructed indices. Critics of PCA argue that the technique is arbitrary, that the method of choosing the number of components and the variables to include is not well defined. The empirical basis for the technique rests on whether the first principal component can predict SES status. This is entirely dependent on the nature of data and the relationships between variables that are being considered, the validity of the variables included and also their reliability.

The choice of variables included can have an impact on the observed poor-rich difference in health outcomes. For example, Houweling et al. (2003) found variables which were a direct determinant of child health outcomes influenced the classification of socio-economic groups, and Lindelow (2002) found geographic bias with the inclusion of infrastructure variables. For comparative purposes, consideration needs to be given to the variables used.

Many studies using asset-based indices appear to have relied on the 'face validity' of the variables included, i.e. they appear to capture household wealth. Validation of PCA-based SES indices has been undertaken by Filmer and Pritchett (2001) for data from India, Indonesia, Pakistan and Nepal. Their study contained both asset and expenditure information and found coherence between the results of the PCA- and expenditure-based classifications, and also concluded that the index was robust to the variables included. In addition, Lindelow (2002) concluded that consumption expenditure and the PCA-based index are different proxies for the same underlying construct of interest.

Few studies have considered the reliability of collecting asset data. While a study by Onwujekewe et al. (2006) found reliability of collecting some asset variables not to be high, Montgomery et al. (2000) suggest that household income data are also unreliable.

There are alternatives to PCA that can reduce the dimensionality of the data using methods such as correspondence analysis, multivariate regression or factor analysis. Cortinovis et al. (1993) used correspondence analysis to derive a SES measure. However, the analysis can only be used for categorical data (nominal and ordinal); continuous data would need to be reorganized into ranges. With multivariate regression, dimensionality reduction is accomplished by simply choosing which variables to leave out, at the expense of ignoring some dimensions of the data. Factor analysis was used by Sahn and Stifel (2003) and has a similar aim to PCA, in terms of expressing a set of variables into a smaller number of indices or factors. The difference between the two is that while there are no assumptions associated with PCA, the factors derived from factor analysis are assumed to represent the underlying processes that result in the correlations between the variables.

Issues related to the underlying data will affect PCA and this should be considered when creating and interpreting results. Clearly, there are methodological issues that need to be considered when developing PCA-based indices. The recent work on PCA-based SES indices suggests that these can be validated and are robust. McKenzie (2003) states that there are a number of theoretical questions of interest in which wealth inequality is more important than consumption or income inequality, so an asset-based inequality measure may be preferred in empirical tests. However, it is up to the user to bear in mind that PCA is best considered as a summary empirical method.

Endnotes

[1] A vector that results in a scalar multiple of itself when multiplied by a matrix is known as an eigenvector, and the scalar is its associated eigenvalue. Eigenvectors can only be found for square matrices (though not all), and for an $n \times n$ matrix, there are n eigenvectors. For a more detailed description of matrix algebra, and in particular eigenvectors and eigenvalues, see Manly (1994).

[2] Brazil is a lower-middle-income country with a GNI per capita of US$3090. With a GNI per capita of US$110 Ethiopia is one of the world's poorest countries ([http://www.worldbank.org]). The urban population was 83% in Brazil and 16% in Ethiopia in 2003 (UNDP 2005). We used the 1996 Brazil DHS and 2000 Ethiopia DHS.

[3] The construction of a number of binary variables from categorical variables is another way to organize the data, although nominally new variables are created. For example, the categorical variable RELIGION, with the values Christian, Muslim, Jewish, Buddhist, converted to binary form would mean the creation of four new variables CHRISTIAN, MUSLIM, JEWISH, BUDDHIST, all of which took on the value of 0 or 1. As the nature of categorical variables is that there is no hierarchical relationship between the variables (which is why they cannot be converted into a meaningful quantitative scale), their conversion into binary variables and inclusion as additional variables does not change the relationship between the variables nor add any additional variation or correlation in the dataset. Rather, having individual variables, PCA can determine which of the particular religion variables can differentiate between households.

[4] PCA is not invariant to differences in the units of measurement among variables, therefore it is usual to standardize the variables in this instance (Bolch and Huang 1974). Standardization is the process of transforming variables so that the new set of scores has a mean equal to zero and standard deviation equal to one. The correlation matrix is a standardized version of the co-variance matrix.

[5] Factor score for ownership of a bicycle not stated in Houweling et al. (2003).

References

Bolch BW, Huang CJ. 1974. *Multivariate statistical methods for business and economics*. Englewood Cliffs, NJ: Prentice Hall.

Cortinovis I, Vela V, Ndiku J. 1993. Construction of a socio-economic index to facilitate analysis of health in data in developing countries. *Social Science and Medicine* **36**: 1087–97.

Deaton A. 2003. Health, inequality and economic development. *Journal of Economic Literature* **41**: 113–58.

Falkingham J, Namazie C. 2002. *Measuring health and poverty: a review of approaches to identifying the poor*. London: Department for International Development Health Systems Resource Centre (DFID HSRC). Accessed 5 April 2006 at: [http://www.eldis.org/static/DOC11501.htm].

Filmer D, Pritchett LH. 2001. Estimating wealth effect without expenditure data – or tears: an application to educational enrollments in states of India. *Demography* **38**: 115–32.

Giri NC. 2004. *Multivariate statistical analysis*. New York: Marcel Dekker Inc.

Gwatkin DR, Rustein S, Johnson K et al. 2000a. *Socio-economic differences in Brazil*. Washington, DC: HNP/Poverty Thematic Group of the World Bank. Accessed 5 January 2004 online at: [http://www.worldbank.org/poverty/health/data/index.htm#lcr].

Gwatkin DR, Rustein S, Johnson K et al. 2000b. *Socio-economic differences in Ethiopia. Health, Nutrition, and Population in Ethiopia*. Washington, DC: HNP/Poverty Thematic Group of the World Bank. Accessed 5 January 2004 online at [http://www.worldbank.org/poverty/health/data/index.htm#lcr].

Gwatkin DR, Rustein S, Johnson K et al. 2000c. *Socio-economic differences in Nigeria. Health, Nutrition, and Population in Nigeria*. Washington, DC: HNP/Poverty Thematic Group of the World Bank. Accessed 19 March 2002 online at: [http://www.worldbank.org/poverty/health/data/index.htm#lcr].

Hanson K, McPake B, Nakamba P, Archard L. 2005. Preferences for hospital quality in Zambia: results from a discrete choice experiment. *Health Economics* **14**: 687–701.

Houweling TAJ, Kunst AE, Mackenbach JP. 2003. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? *International Journal for Equity in Health* **2**: 8.

Jobson JD. 1992. *Applied multivariate data analysis*. New York: Springer-Verlag.

Lindelow M. 2002. *Sometimes more equal than others: How the choice of welfare indicator can affect the measurement of health inequalities and the incidence of public spending*. CSAE Working Paper Series 2002–15. Oxford: Centre for Study of African Economies, University of Oxford.

Manly BFJ. 1994. *Multivariate statistical methods. A primer*. 2nd Edition. London: Chapman and Hall.

McKenzie DJ. 2003. Measure inequality with asset indicators. *BREAD Working Paper* No. 042. Cambridge, MA: Bureau for Research and Economic Analysis of Development, Center for International Development, Harvard University.

Montgomery MR, Gragnolati K, Burke A, Paredes E. 2000. Measuring living standards with proxy variables. *Demography* **37**: 155–74.

Onwujekwe O, Hanson K, Fox-Rushby J. 2006. Some indicators of socio-economic status may not be reliable and use of indices with these data could worsen equity. *Health Economics* **15**: 639–44.

Sahn D, Stifel D. 2003. Exploring alternative measures of welfare in the absence of expenditure data. *Review of Income and Wealth* **49**: 463–89.

Schellenberg JA, Victora CG, Mushi A et al. 2003. Inequities among the very poor: health care for children in southern Tanzania. *The Lancet* **361**: 561–6.

UNDP. 2005. *Human development report 2005. International cooperation at a crossroads: aid, trade and security in an unequal world*. New York: Oxford University Press for the United Nations Development Programme (UNDP).

Acknowledgements

We would like to thank Kara Hanson, Peter Vickerman and the two anonymous reviewers for their technical input and comments on a draft of the manuscript. We would also like to thank Jo Borghi, Jolene Skordis, Damian Walker and Natasha Palmer for comments on an earlier version of the manuscript.

Biographies

Seema Vyas is a Research Fellow with the Health Policy Unit, Department of Public and Policy, LSHTM. She specializes in quantitative and private health sector analysis.

Lilani Kumaranayake is a Lecturer in Health Economics and Policy, Department of Public Health and Policy, LSHTM. She specializes in the economics of HIV/AIDS, private health sector and quantitative analysis.

Correspondence: Seema Vyas, HIVTools Research Group, Health Policy Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK. Tel: +44 (0) 20 7612 7828; Fax: +44 (0) 20 7637 5391; E-mail: seema.vyas@lshtm.ac.uk