How to measure socioeconomic status using census data in developing countries?

1. Motivation

A measure of socioeconomic status of a household is an important element in most economic and demographic analyses. This measure is useful, not only in terms of estimating poverty and inequality within a society, but it can also be used as a control variable in finding the effects of other variables associated with wealth (Filmer and Pritchett, 2001). Based on theoretical grounds, household income and expenditure levels are often used as measures of household wealth or household socioeconomic status. However, collecting data on income and expenditures can be costly. Moreover, most demographic and household surveys that contain income and expenditures data tend to have small sample sizes.

In contrast, large-scale surveys on population and housing such as census surveys can overcome problems of small sample sizes and underrepresentation of certain population groups in smaller geographical units. Although the main feature of the census data is the enumeration of individuals and households in the country at a particular point in time, it has advantages over other household surveys for at least three reasons. First, census data are more commonly available than nationally representative household surveys. Second, due to the larger scale, census data are more comprehensive when compared to other household surveys, which may not represent all population groups accurately. Third, the larger number of observations in census data can provide more precise estimates for statistical purposes. Given all of these reasons, census data is a promising data source for conducting social and economic research.

To date, the Integrated Public Use Microdata Series (IPUMS) - International, at Minnesota Population Center, University of Minnesota, has collected the world's largest archive of free and publicly available census samples. Currently, the database includes 158 census samples taken during 1960 to present from 55 countries around the world. Furthermore, IPUMS-International data is composed of microdata at individual and household levels. The data includes information on household characteristics as well as a wide range of population characteristics, such as basic demographic, fertility, education, occupation, migration, and others, which are coded and documented in systematically across countries and years.

Nevertheless, despite the availability of census data and its comprehensiveness, most of these census samples, particularly from developing countries, do not collect information on income or expenditures, which are used widely as a measure of socioeconomic status. The lack of this measure limits the ability of researchers to perform analyses using census data. Thus, it is essential to develop a measure of household socioeconomic status based on the other information usually available in censuses. This proposed measure will not only improve the use of census data in social and economic research, but will also give some insights about the relative socioeconomic status of households in a particular country during a specific year.

The asset-based approach to determine socioeconomic status has been widely used in previous studies as an appropriate measure of household wealth (Montgomery et al, 2000; Filmer and Pritchett, 2001; Sahn and Stiefel, 2000 and 2003; McKenzie, 2005). Even though census data is widely available and usually includes information on assets, there are no large scale efforts to date to develop measures of relative household wealth. In this paper, we validate the effectiveness of the asset index using IPUMS-International census samples. Furthermore, we develop separate measures of socioeconomic status in rural and urban areas, to acknowledge for possible differences in wealth accumulation by place of residence. Finally, we explore conditions to assure the robustness of the asset index constructed from census data.

2. Objective

Given the advantages of census data and the lack of socioeconomic measures in most censuses, the goal of this paper is to develop a reliable and robust measure for socioeconomic status at the household level using census data available from IPUMS-International. More specifically, we attempt to use non-monetary indicators including asset ownership, utilities, dwelling characteristics, appliances, and other amenities that are generally available in censuses to compute an asset index. To validate the asset index calculation from census data, we attempt to evaluate the reliability of the proposed index and also to suggest some conditions (or

criteria) for robustness of this index when using different census samples. These justifications will be illustrated by using selected samples from IPUMS-International.

3. Methodology

In this paper, we apply the asset index approach and focus on two separate but interrelated questions. First, we examine how appropriate the asset index measures household socioeconomic status for census data. Within this question, we develop separate measures for urban and rural areas. Second, we identify some conditions to produce a robust indicator, taking into account that the number and type of data available vary widely across censuses.

Calculation of the asset index is performed through Principal Component Analysis (PCA), a data reduction technique, which creates orthogonal linear combinations from a set of variables, and orders them according to their contribution to the overall variability of the variables analyzed. In order to apply PCA to census data, all variables are transformed into a dichotomous version, including the categorical variables housing characteristics (e.g. material of walls or floor) or access to utilities (e.g. type of water or sewage service). In the process of producing the asset index, other methodological alternatives for the weighting procedure will be explored. Based on the asset index, we produce wealth quintiles which reflect the resulting rankings of population by socioeconomic status. In addition, separate rural and urban asset indexes are produced and their agreement with the overall asset index is verified.

The initial research question refers to the reliability of an asset based asset index to measure household wealth applied to census data, which is verified through different strategies. First, we verified the agreement level of results using census data with comparable DHS datasets. For this, we selected censuses coinciding in time with DHS data collection for specific countries (Brazil, Colombia, Egypt, and Senegal). Since both sources of data are nationally representative, we would expect similar distributions of the asset index. In order to verify the agreement between the two, we calculated statistics representing the distribution of each un-standardized index (mean, variance, skewness, and kurtosis) and also compared the standardized indexes graphically using kernel density estimation methods. Two sources of discrepancy can arise between the wealth indices produced using each dataset: the sample design used by DHS (intended to sample households with women during reproductive age) and the specific set of questions available in each case. Therefore, we also examined the consistency of results when using comparable samples of population and restricting calculations to the common subset of variables. Second, following the methodological approach in previous studies (Montgomery et al, 2000; Filmer and Pritchett, 2001; Ferguson et al, 2003; Sanh and Stiefel, 2003), the asset index is compared with predicted household expenditures. Samples were selected for this exercise based on the availability of national household surveys that would allow us to estimate a model for household expenditures. Third, we verified the predictive power of our asset index against selected outcomes, including school attendance and educational attainment.

The second research question is focused on general conditions necessary to produce a robust asset index based on census data. The underlying issue is the variable availability across censuses, which could have any number of assets listed or discrepancies on how data was collected. Even though the general recommendation has been to use the most variables available, as long as those are related to unobserved wealth (Rutstein and Johnson, 2004; McKenzie, 2005), it remains unclear which types of assets have larger contributions to the constructed measure and what the minimum number of necessary variables is. Further, two data problems could arise and restrict the information the asset index provides in term: (i) clumping, if a limited number of values are produced; and (ii) truncation, if there are no indicators available to explain differences at the tails of the wealth distribution (McKenzie, 2005; Minujin & Bang, 2002). First, to check the type of assets that contribute more to the measurement of wealth, we calculate separate indices for asset durables, housing characteristics, and utilities. Second, in order to set a standard for input requirements for the index, we perform similar calculations by restricting the number of variables included in the analysis. For this exercise, we exclude some variables based on contributions to the original asset index and then recalculate the index. We verify the consistency of rankings produced by each subset of variables through Spearman rank correlations and correspondence indices. We then verify how much of the original index is explained by the recalculated indices using subsets of variables through regression analysis. Furthermore, we analyze possible

clumping and truncation problems graphically using kernel density estimation methods and comparing the rank agreement after eliminating tails of the index distribution.

4. Preliminary Results

The asset index was calculated using PCA for four of the IPUMS-I samples: Senegal 2002, Colombia 2005, Brazil 1991 and Egypt 1996. We calculated wealth quintiles based on the asset index and compared asset ownership for each sub-group, to verify internal validity of our measure. For example, for Egypt 1996, we observe significant differences in access to electricity between the richest quintile (99%) and the poorest quintile (78.6%). In addition, we compared the distribution of the asset index based on census and DHS data, based on all variables available in each database. In Figure 1, we show the kernel density distributions for the asset indexes produced for Senegal 2002 (census) and 2005 (DHS); as we observe, the figure indicates a smooth distribution based on census data, closely resembling that of the DHS data. As a final measure of comparison between our asset index and the DHS index, we compared their relative performance in measuring inequalities in school attendance in each of the five quintiles. Table 1 shows gaps in school attendance for Senegal 2002/5 and for Colombia 2005 across datasets. In most cases, the percentages of students attending school are quite similar across DHS and IPUMS-International data, with the largest discrepancy being for the middle population in Senegal with an 8.3 percentage point difference.

In general, our preliminary results indicate that the household asset index calculated on census samples resembles DHS results based on comparable datasets (Egypt, Brazil, Senegal, and Colombia). Ownership of assets across wealth quintiles confirms the internal validity of the measure. Also, the calculated household asset index has a similar performance to the DHS measure when analyzing educational outcomes. As suggested, additional validation of the results will be done using predicted consumption expenditures and through regression analyses for selected educational outcomes (school attendance and attainment). Furthermore, we will examine the conditions for robustness, considering the varying availability of asset variables across census samples.

Appendix

Figure 1: Senegal 2002/5, Kernel density distributions for DHS and Census asset index

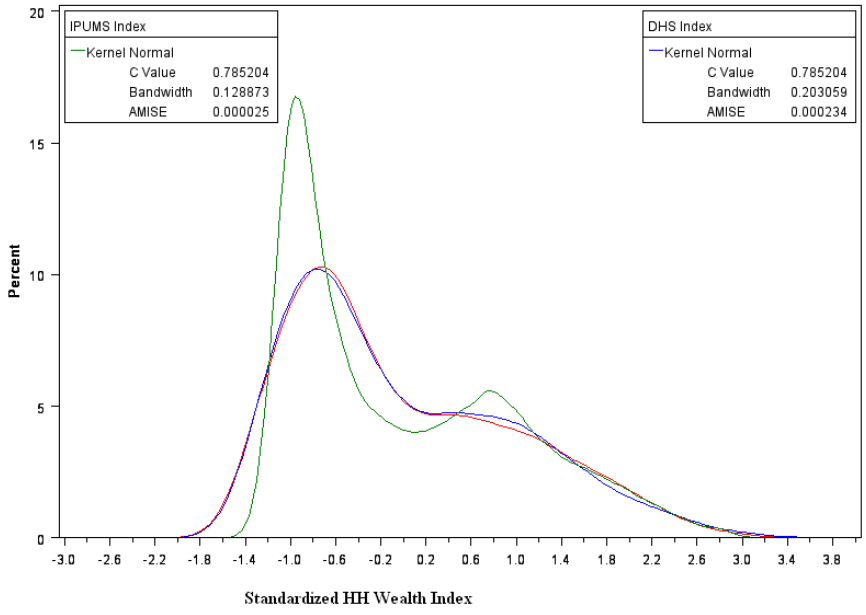

Table 1: School attendance by wealth quintile comparing DHS to IPUMS-I

School attendance	% Population				
	Poorest	Poorer	Middle	Richer	Richest
Senegal (DHS, 2005)	33.3	42.7	55.2	64.3	72.7
Senegal (IPUMS, 2002)	32.2	38.3	46.9	57.2	70.5
Colombia (DHS, 2005)	77.8	84.3	88.2	91.2	94.0
Colombia (IPUMS, 2005)	70.9	82.4	88.0	91.5	94.8

References

Filmer, D., Pritchett, L.H. Estimating wealth effects without expenditure data or tears: an application to educational enrollments in states of India. (2001). *Demography, 38*(1), 115.

McKenzie, D.J. Measuring inequality with asset indicators. (2005). *Journal of Population Economics, 18*(2), 229.

Minujin, Alberto and Joon Hee Bang. Indicadores de inequidad social. Acerca del uso del "índice de bienes" para la distribución de los hogares. *Desarrollo Económico*, 42 (165), 129.

Montgomery M.R., Gragnolati, M., Burke, K.A., Paredes, E. Measuring living standards with proxy variables. *Demography* 2000 , ***37***(2)**,**155.

Rutstein, Shea O. and Kiersten Johnson. 2004. "The DHS Wealth Index*." DHS Comparative Reports No. 6.* Calverton, Maryland: ORC Macro.

Sahn, D.E., Stifel, D.C. Poverty comparisons over time and across countries in Africa. (2000). *World Development, 28*(12), 2123.

Sahn, David E. and David Stiefel. Exploring alternative measures of welfare in the absence of expenditure data (2003). *Review of Income and Wealth*, 49 (4), 463.