# 8

# Other Methodological Considerations

This chapter discusses several necessary methodological issues that have to be dealt with before a proper analysis of global inequality and poverty can be undertaken. All that is needed is that, in addition to basic data on population, per capita income, and consumption, there are data on the distribution of expenditures (or income). These distributions need to be available for the different years in which the surveys were undertaken, and they need to be at a sufficiently low level of aggregation—quintile shares would not do, nor would decile shares. What might just seem right are percentile shares (i.e., mean expenditures for each 1/100th of the population in each economy for which there are survey data). This chapter is about the method developed to achieve this goal, and the accuracy of the method.

## The Individual versus Countries

Concern about inequality, or convergence, or almost any economic phenomenon, is really concern about individuals, not countries. A country is an artificial concept, as has been found out by many societies that have been partitioned, with the USSR being just the latest example. Aggregation has its advantages, both conceptually and individually. But the reason economists have used the country as a unit of analysis to study essentially individual issues (e.g., inequality) is largely because of the *convenience* of data availability. Data are generally not available at an individual level.

For studies of globalization, of inequality, of poverty, and of convergence, *individual*-level data are the most meaningful. For example, when one looks at poverty, a poor person in Bangladesh has more in common with a poor person in Nigeria than with a poor person in the United States. The discussion is generally about the poverty of individuals, not the poverty of nations. The same is true for concerns about globalization; the protests against globalization most likely have more to do with how particular sets of individuals (in different countries) have lost out, rather than particular sets of countries.[1] Analogously, with inequality, surely it matters more how the bottom 20 percent of the *world* has fared with respect to income growth, rather than the bottom 20 percent in China or the United States.

In the analysis of world income inequality, two methods have been prominent: the computationally easy one of inequality, in which a country is the unit of analysis; and the cumbersome yet correct construction of inequality, in which the individual is the unit of analysis. In the former case, each individual in an economy is attributed the mean income for the *entire* country; in the latter case, each individual in the world has her own unique income.

However, the ''easy'' assumption of every person having the same income as the per capita income can be shown to be particularly wrong in the case of global individual inequality. The deduction, from trends in country inequality, that global inequality has worsened is not logical.[2] *Most* patterns of intercountry inequality change are consistent both with improving inequality and with worsening inequality. It all depends where in the overall distribution the change is taking place—if poor economies are getting richer faster than rich economies (and there is no significant overlap in distributions), then world individual income inequality estimates (W3i) will have a tendency toward improvement.

Economists have always recognized these simple facts, but a lack of data has prevented analysis from being done in terms of ''groups of individuals.'' A country has been considered a reasonable shorthand approximation to ''similar groups of individuals,'' and indeed has provided important insights. So the error has not been large; indeed, most of the time, there may not have been any empirical error at all (i.e., the results obtained from an analysis of nations may have yielded exactly the result that an analysis of individual-level data would have provided, if such data had been available).

Most of the existing statements on worsening income inequality are based on an analysis of income distributions at a *country* level. Relatively fewer studies exist on world individual inequality—and the ones that do

1. Obviously, issues such as defense and imperialism are national, not individual, matters; but they are not of concern here.

2. That journalists would make this connection is understandable; that economists would do so is somewhat surprising.

present a mixed picture of trends in the past 20 years. Both the construction of W3i, and the differences of simple accounting procedure (SAP) W3i from other existing estimates, are the subject of this chapter.

## Simple Accounting Procedure for Generating W3i

Data are obviously necessary for an analysis of what has happened to poverty and inequality. Hence, the exhaustive exercise of assembling the data available on populations, income, and consumption for all countries and all the years after 1950 was undertaken. The next step was to assemble the distributions of income and consumption that were available. These distributions were available in quintile form; that is, shares of income or consumption for each 20 percent of the population. Normally, this compilation would be both necessary and sufficient for analysis. However, as has been mentioned several times, a new method of estimating both poverty and inequality has appeared on the scene.

This method uses estimates of survey means, not national account means. So data on survey means were needed. Conveniently, for a large number of household surveys in developing countries, the World Bank posts the means in purchasing power parity (PPP) 1993 prices. But it was observed that these means could not be reproduced using published PPP exchange rates. This meant that one had to go to sources alternative to the World Bank for these data.

The assembly of data on survey means was an involved, but doable, exercise. Analysts for Latin America have long been publishing the ratio of survey to national accounts for the income surveys in those countries. My own analysis of household surveys for India, Malaysia, and Sri Lanka yielded the survey means for several surveys in these countries. Data for the United States were the easiest to obtain—several publications, Web sites, and so on, are available. Milanovic has posted on the Web the survey means, in local currency, for the countries that he used in his analysis of world inequality for 1988 and 1993, including an update for Eastern European economies in the mid-1990s. And a large Asian Development Bank project on poverty in Asia, titled ''RETA 5917—Building a Poverty Database,'' yielded estimates of survey means of more than 50 surveys in Asian countries for the period 1980-2000.[3]

This, then, is a brief description of which data are necessary, and which have been collected. Though several questions can be answered, the most important pertaining to levels and trends in individual inequality, and levels and trends in absolute poverty, cannot be answered with these

3. See Asian Development Bank (2002).

data. Why? Because the distributions are not available at a more disaggregated level than quintiles; the best that is possible is decile data, and even that for only a few countries and also too aggregated for such countries as China, where 125 million people (10 percent) are being attributed the *same* level of average decile income.

What is required is a *percentile* distribution for each country-year distribution. This method would attribute the same income to only 12.5 million people in China, to each 2.8 million people in the United States, and to each 600,000 people in the United Kingdom. Though it may be ''required,'' the larger concern is one of accuracy. Because any such percentile distribution will have to be derived from data on quintile distributions, the interpolation is from knowing the mean income of 5 20-percentile sets to 100 1-percent sets. And if such a derivation is done, how accurate can it be? The accuracy of such a constructed Lorenz curve is an important issue—if the underlying data of global inequality distributions are not accurate, the compilation might be misleading, or even wrong; hence, appendix B is devoted to it.

Toward this end, the SAP method of estimating Lorenz curve distributions was developed, a procedure that yields 100 percentiles,[4] and therefore means, for 100 different and equal (in size) sets of individuals for each distribution year. The difference between the global inequality distributions developed by Berry, Bourguignon, and Morrisson in 1983, Bourguignon and Morrisson in 2001, Milanovic in October 1999, and the SAP global inequality distributions developed by me (the first estimate was in June 2000; Bhalla 2000d) is that the other three do not construct such detailed distributions of income.[5] Bourguignon and Morrisson's method has an average of 11 means per distribution, and Milanovic's combination of rural and urban data for some countries yields an average of about 12 different ''means'' per country-year.

The Lorenz curve method provides an estimate of mean income for each and every percentile *within* a country. Aggregating to obtain the global income distribution is now straightforward. The individual country and world populations for each year are known. Thus, the population of each percentile in each country is also known (e.g., India, with 1 billion people, has 10 million in each percentile, whereas Uganda, with 20 million,

---

4. The developed method allows even greater fine-tuning; e.g., a 200- or 400-point distribution. However, the gains from further disaggregation are probably not that great. Such increased precision is unlikely to change the mean income at each percentile level.

5. In a recent paper, Sala-i-Martin (2002b) also constructs a world income distribution based on quintile data but exaggerates when he contends that ''to our knowledge, this is the first attempt to construct a world income distribution by aggregating individual country distribution'' (p. 2). That credit should rightfully go to the pioneering work of Berry, Bourguignon, and Morrisson (1983) and Bourguignon and Morrisson (2001), both of which, remarkably, are cited and discussed by Sala-i-Martin.

has 200,000 in each percentile), as well as the population in each world percentile (e.g., if the world has 6 billion people, then each percentile has 60 million people).

If the data are ranked according to average percentile income, regardless from which country the data are gathered, then the global distribution is easily formed; that is, each world percentile has an estimate of the number of people and an estimate of the mean income (a weighted average of all the country observations within each percentile). This *is* the world Lorenz curve as estimated by SAP.

## The Kakwani Method of Estimating a Lorenz Curve

If the Lorenz curve can be parameterized (i.e., given a mathematical formulation), then the study of the determinants of inequality can ''begin.'' Unfortunately, there are an infinity of ''patterns'' of distribution, so where can one start? The Lorenz curves can cross, in which case it is a priori ambiguous which distribution is more unequal. In a pioneering study, Kakwani (1980) discusses methods of approximating the Lorenz curve and methods of estimating the same from the limited data (share in income of different income groups and/or shares of individual quintiles).

His preferred formulation, and one used by SAP, is the following:

$$L(p) = p - a^*[p^{\alpha}]^*[(1 - p)^{\beta}] \tag{8.1}$$

where $p$ represents the bottom $p$ percent of the population, and $L(p)$ is the corresponding share in income. Taking logs and rearranging terms, one obtains a form fit for regression:

$$\log\,[p - L(p)] = \log a + \alpha^*\log\,(p) + \beta^*\log(1 - p) \tag{8.2}$$

The parameters obtained from the equation 8.2 can be used to generate the estimated incomes of each percentile of the population. Often, only four independent observations are available for each distribution (the quintile shares—the fifth quintile is derived from the other four and equal to 1 minus the sum of the other four); thus three parameters ($a$, $\alpha$, and $\beta$) are estimated from four observations, leaving only 1 degree of freedom.

The basic equation results are then *filtered* by SAP to satisfy the theoretical boundary constraints (i.e., the sum of the estimated shares of each quintile is actually equal to the *observed* shares, and the share of each percentile is equal to or larger than the share of the previous percentile). The filtering is done through an iterative procedure, whereby at the end of the first round, the shares of each individual percentile in the first quintile get estimated and fixed, then the next quintile, then the next, and so on. (The only somewhat ''arbitrary'' and somewhat ''flexible'' percentiles are the first and the last, and this flexibility shows up in the errors; see below.)

The filtering and consistency tests are the *additions* to the basic structure of the Kakwani method. To reiterate, the ''base'' calculation is made using Kakwani's method; this base prediction is then filtered again and again to obtain the best estimates. The only issue now left is one of accuracy; that is, are the SAP percentile distributions accurate? Can a method that generates percentile distributions from knowledge of five quintile shares be reasonably precise within any reasonable degree of confidence?

## Is SAP Accurate?

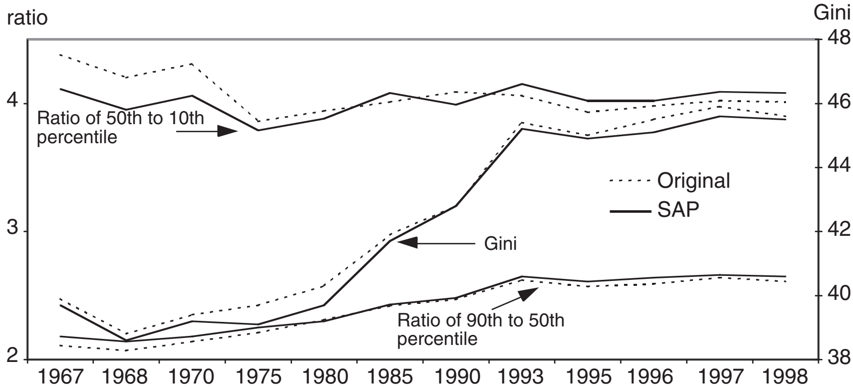
Is SAP accurate? Yes. The SAP method was able to correctly estimate incomes at the percentile level of aggregation for several household distributions for which unit-level data were available (several distributions for India, 1983-99; for states of India, 1983-99; and for Malaysia, 1973-89). All in all, more than 250 ''country''-level distributions were tested. Using Indian household data (see table B.1 in appendix B), it is shown that SAP-estimated percentile levels are within 1 percent of the *actual* (derived from unit-level data) levels for most of the 100 percentiles, except the bottom 1 percent and the top 1 percent. The mean error for the entire distribution as proxied by the Gini coefficient was only 0.4 percent, with the largest such error being 3.7 percent.

Figure 8.1 reports on an SAP reconstruction of household income data for the United States for survey years 1967 to 1998. These unit-level data were not available, but published data on quintiles, selected percentiles, and the Gini were available. The constructed and original Ginis are within a whisker of each other for all the years; further, the SAP method is correctly able to identify the levels of different ratios. For example, the ratio of the mean income of the 90th-percentile household to the 20th-percentile household is reported as 9.22 for 1967 and 10.44 for 1998; the SAP reconstruction suggests that the two ratios are 8.96 and 10.80, respectively.

## Using SAP to Identify Errors in Published Ginis

The tests above suggest that the SAP method is accurate both at the aggregate Gini level (very, very accurate) and at the individual percentile level (very accurate). And if the distributions for an individual country-year are accurate, then it is simple accounting that gets one to (accurate) global estimates of individual inequality and poverty. Some additional,

**Figure 8.1 How good is the SAP method? Evidence from the United States, 1967-98**




Note: The solid line reflects published data; the dashed line represents the approximation obtained for each percentile by the simple accounting procedure (SAP) method; see the text and appendices A and B for details. The SAP method converts quintile-level data into percentile-level data.

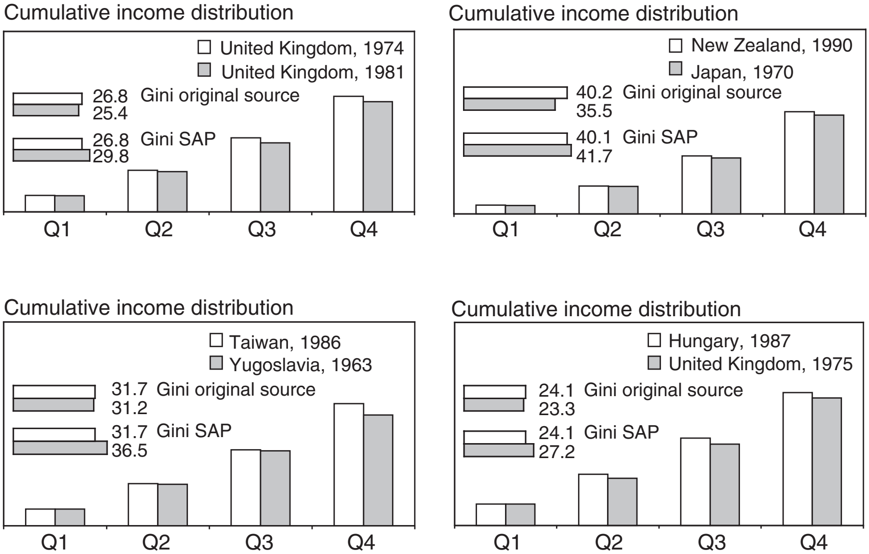
*Sources:* US Census Bureau, *Current Population Survey*, selected March supplements; http://www.census.gov/hhes/www/incineq.html.

albeit indirect, tests also suggest that the SAP method is reliable—so much so that it can help identify errors in published Ginis!

The Gini estimated by SAP is on the basis of the constructed percentile distribution, a distribution that is consistent with the published quintile shares. But some of the published Ginis do not "match" the accompanying published quintile data. What apparently is occurring is that the published quintile shares are for a different distribution than published Ginis; for example, the quintile shares might refer to a household distribution, whereas the Gini refers to a transformed per capita distribution. Atkinson and Brandolini (1999) correctly observe that even the "accepted" quality Ginis contained in the popular Deininger and Squire (1996) dataset are subject to large errors; they document a large number of cases for which the reported Ginis do not accurately reflect the underlying data. Thus, if the foundation is suspect, how credible is the edifice? The SAP cleaning operation was able to identify several such examples of mismatch, and figures 8.2 and 8.3 document some of the most egregious ones.

The figures contain eight pairs of country distributions; the pairs were chosen according to the original published Ginis, and chosen for suggesting that the published distributions are the same, at least according to the Gini. Also plotted in the figures is the cumulative distribution, so that

**Figure 8.2 Ginis compared: Original source and simple accounting procedure (SAP), selected countries**




Q = quintile; Q1 = first quintile, etc.

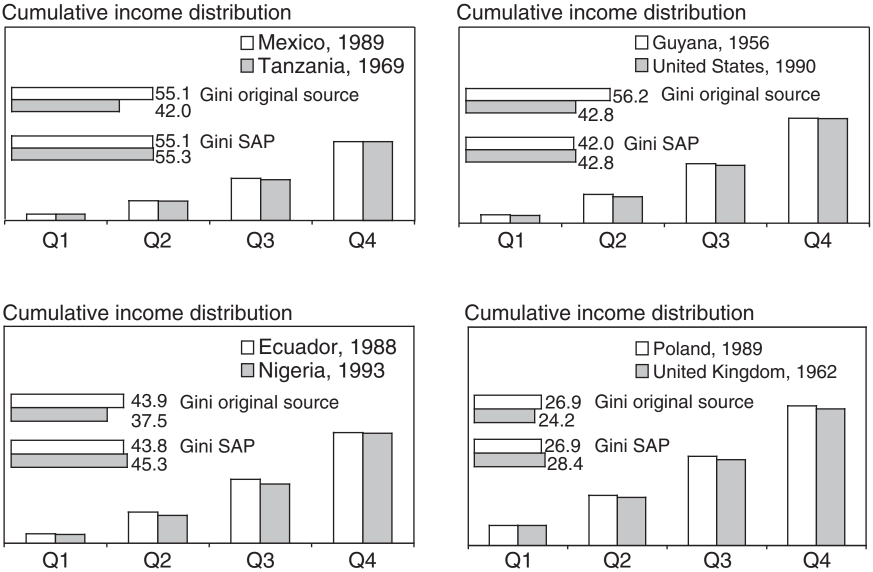
Note: The SAP method converts quintile-level data into percentile-level data and computes the reported Ginis. Q1 (bottom 20 percent) through Q4 (bottom 80 percent) represent the cumulative quintile shares as published; the two Ginis reported are those published and those computed by SAP, along with the quintile shares.

A relatively high Gini value should approximately correspond to relatively lower cumulative income shares at each quintile level, comparing any two countries, or comparing across time for the same country. However, note that, in most cases, relatively higher (original) Ginis correspond to relatively higher income shares for the poorer quintiles. The SAP method yields results that conform far better to the former, expected pattern. For instance, comparing the United Kingdom for 1974 with the United Kingdom for 1981, we find both the (original) Gini and cumulative income shares to be higher in 1974 than in 1981. SAP Gini estimates, though, indicate the reverse, with a relatively higher Gini value in 1981, which corresponds directly to relatively lower cumulative income shares in that year. This, and other comparisons, verifies the accuracy of the SAP method.

There should, in theory, be no difference between the two Ginis; the difference exists because the published Ginis do not accurately reflect the underlying data, possibly because the quintile shares refer to one distribution (e.g., data ordered by households), whereas the published Gini reflects another distribution (e.g., households ordered by per capita expenditure). See also Atkinson and Brandolini (1999).

*Sources:* Deininger and Squire (1996); World Income Inequality Database, available at http://www.wider.unu.edu/wiid; Asian Development Bank (2002); World Bank, *World Development Indicators,* CD-ROMs.

**Figure 8.3 Ginis compared, original source and simple accounting procedure** (SAP)**, other selected countries**




Q = quintile; Q1 = first quintile, etc.

Note: The SAP method converts quintile-level data into percentile-level data and computes the reported Ginis. Q1 (bottom 20 percent) through Q4 (bottom 80 percent) represent the cumulative quintile shares as published; the two Ginis reported are those published and those computed by SAP, along with the quintile shares.

A relatively high Gini value should approximately correspond to relatively lower cumulative income shares at each quintile level, comparing any two countries, or comparing across time for the same country. However, note that, in most cases, relatively higher (original) Ginis correspond to relatively higher income shares for the poorer quintiles. The SAP method yields results that conform far better to the former, expected pattern. For instance, comparing Mexico for 1989 with Tanzania (1969), we find both the (original) Gini and cumulative income shares to be higher for Mexico. SAP Gini estimates, though, indicate the reverse, with a relatively higher Gini value for Tanzania, which corresponds directly to relatively lower cumulative income shares in that year. This, and other comparisons, verifies the accuracy of the SAP method.

There should, in theory, be no difference between the two Ginis; the difference exists because the published Ginis do not accurately reflect the underlying data, possibly because the quintile shares refer to one distribution (e.g., data ordered by households), whereas the published Gini reflects another distribution (e.g., households ordered by per capita expenditure). See also Atkinson and Brandolini (1999).

*Sources:* Deininger and Squire (1996); World Income Inequality Database, available at http://www.wider.unu.edu/wiid; Asian Development Bank (2002); World Bank, *World Development Indicators,* CD-ROMs.

an "eye-inspection" can easily identify whether, and which, distribution is more unequal. The unshaded distribution is always the more equal distribution according to the Lorenz curve (quintile shares), the shaded distribution always the more unequal distribution. Because cumulative distributions are plotted, the unshaded bars should cumulatively be above the shaded bars. Because they are (and were chosen on that basis), the Gini based on the unshaded distribution should be *lower* than that based on the shaded distribution. If it is not, then the reported published Gini is inconsistent with respect to the reported quintile distribution.

Figure 8.3 documents the reality that the Mexican 1989 and the Tanzanian 1969 distributions are both highly unequal (Gini approximately equal to 55). It is not the case that socialist Tanzania had a relatively equal distribution of 42, as was suggested by the original data.

A surprising entrant among the error distributions is that of Nigeria for 1993 (figure 8.3). This distribution is reported in *all* the sources of data (Deininger and Squire 1996; World Institute for Development Economics Research, World Income Inequality Database; *World Development Indicators*; Dollar and Kraay 2000; etc.). This might be more an indication of how errors get transmitted than of anything being wrong with the original data! In any case, the reported Gini is 37.5, with the share of the bottom 40 percent reported as 12.9 percent, and the share of the bottom 80 percent as 50.7 percent. The SAP Gini for this distribution is 45.3, almost 20 percent higher than the official Gini.

Can the "official" estimate be that much in error? The "twin" distribution for Ecuador for 1988 shows that the answer is yes. The *Lorenz curve for Ecuador dominates that for Nigeria for every point on the quintile distribution* (i.e., it is more equal); and the SAP Gini of 43.8 is lower than the SAP Gini for Nigeria (45.3). However, the official estimate for Ecuador is higher by 17 percent! In this instance, it is unambiguously the case that the SAP Gini is correct and that the official reported Gini, and one contained in the various datasets, is in error. The same conclusion holds for all the top 10 errors reported in figures 8.2 and 8.3, and others not reported here.

This reliability of the SAP method is the final step before presentation of the results. If the method can pinpoint errors in published data, it cannot be too far off from being accurate. The next step is a matter of mere accounting, or counting; that is, given distributions for different regions, at a percentile level, it is a simple matter to compute the distributions and the inequality measures.