Global and regional estimation of informal sector employment: A methodology

Farhad Mehran
ILO Statistical Analysis and Development

8 June 2002

The purpose of this note is to describe a methodology for estimating the global number of workers in the informal sector and its breakdown by major regions of the world, based on a limited sample of national data. The description is organized around the four basic steps of the proposed methodology: (1) choice of sample countries; (2) harmonization of data; (3) global estimation; and (4) regional breakdown. In the final section (5) the advantages and drawbacks of the method are evaluated against another methodology based on imputation for missing countries.

1. Choice of sample countries. The idea is to construct a representative sample of countries on the basis of which the world aggregate can be estimated with a given degree of accuracy. This formulation of the problem borrows from sampling theory, and guides the method for selecting the sample countries: they should form, or, at least mimic, a random sample of countries with appropriate design.

It is proposed to select the countries so as to be as close as possible to a stratified random sample with probabilities of selection proportional to the size of the country, measured in terms of the nation's labour force.

The ILO database LABPROJ on *estimates and projections of the economically active population, 1950-2010*, could serve as a sampling frame for this purpose. For example, the 2000 data of LABPROJ could be used to stratify the 178 countries and territories of the database according to six regional groupings and three size-categories - *large* countries, *middle-size* countries and *small* countries. Large countries are included in the sample with probability one, middle-size countries with probability one-third, and small countries with probability one-tenth.

A realization of this process conducted for the purpose estimating world unemployment is shown in Table 1 in the annex. It led to the selection of 33 countries, representing the various regions and different size-categories of countries. In the context of informal sector, the choice of the middle-size and small countries may be modified in light of data availability.

2. Harmonization of data. When the countries are selected and the data obtained, the national data sets should be harmonized with respect to the key differences, for example, with respect to age groups, reference year, and the scope and definition of the informal sector.

Standardization of age groups may be carried by appropriate interpolation and extrapolation techniques. A simple procedure has been developed in the context of world estimation of child labour and reported in appendix 1 of *Every Child Counts*.

The national data may be centred to a common reference year by applying the informal sector employment-population ratio in the year of the national data to the corresponding population in the reference year.

Adjustments for variations in measurement of the informal sector require a systematic examination of the national definitions and identification and quantification of the key differences. Alternatively, a model may be constructed in which the differences are estimated by categorizing the national datasets into a few groups each representing one type of measurement of informal sector employment. The estimated differential "measurement effect" would then be used to adjust the data into a standard measurement method. An example of this harmonization procedure is given in appendix 2 of *Every Child Counts*.

3. Global estimation. Based on the harmonized data, global estimates are obtained by weighting each national data by the inverse of the probability of selection of the corresponding country. The probabilities of selection are computed by assimilating the sample countries as a random sample of countries stratified by regions and selected with probability proportional to size of its labour force.

The global estimates obtained by the procedure described above can be improved by using auxiliary information on the labour force available for every country, not just the sample countries, from LABPROJ. The various calibration methods can be used including, ratio estimates or more elaborate procedures general regression estimators.[1]

The margin of errors of the estimators may be computed by calculating the variance of the estimators under the presumed sampling design. The calculation may be simplified by using generalized regression estimators.[2]

4. Regional breakdown. The national datasets used for producing global estimates contain too few countries in a given region to lead to sufficiently accurate regional estimates. Thus, the regional breakdown of the global estimates should be obtained by an alternative method. A preferred approach consists of disaggregating the world estimate into its regional components in much the same way as national statistics are disaggregated into small area statistics by combining survey estimates with correlated auxiliary data.[3]

A simple procedure uses the so-called synthetic estimate. The percentage of informal sector employment in total employment in a given region is estimated by combining two estimators. One is the percentage of informal sector employment in the total employment in the world irrespective of the region. The other is the percentage obtained for the region on the basis of the limited number of observations in the sample. The two estimates are weighted in an appropriate manner to obtain the final regional percentage. The weights are such that the mean squared error of the final estimate is minimized.

The rationale of the above combined method is that the first estimate (i.e., the world percentage as an estimate of the region) is biased for the region but has relatively low variance as it is based on the full sample, while the second estimate (i.e., regional percentage based on the sample observations) has high variance but low bias. Thus combining the two should lead to a more accurate final estimate.

---

[1] Deville, Jean-Claude, and Carl-Erik Särndal, "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, Volume 87, Number 418, June 1992, pp. 376-382.
[2] Särndal, Carl-Erik, "Efficient estimators with simple variance in unequal probability sampling," *Journal of the American Statistical Association*, Volume 91, Number 435, September 1996, 1289-1299.
[3] Gosh, M., and J.N.K. Rao, "Small area estimation: an appraisal," *Statistical Science*, Volume 9, Number 1, February 1994, pp. 55-76.

5. Evaluation. The methodology proposed here has been applied in two different contexts. One was to estimate global and regional unemployment,[4] the other to estimate the global and regional number of children at work.[5] Alternative methods based on missing value procedures have also been tested for a number of labour market indicators.[6]

The following table compares the results obtained on world and regional unemployment in 1995 from the two approaches.

Table 2. Comparing world and regional estimates of unemployment in 1995 based on two approaches

	Sampling approach	Missing value approach
World unemployed	136 millions	141 millions
World unemployment rate	5.2 %	5.6 %
Developed economies	6.5 %	7.4 %
Transition economies	8.2 %	8.5 %
Latin America and Caribbean Countries	6.2 %	7.6 %
Other countries	3.8 %	4.1 %

The missing value approach has the advantage of making use of available data from all countries, thus maximizing the use of information. It has also the advantage of having a straightforward explanation, easily understood by laymen: a) The world estimate is the sum of country estimates; b) estimates for countries with missing values are obtained through imputation methods, either by substituting averages of the regions they belong or by using data on strongly correlated variables.

One drawback of the missing value approach is the instability of measures of change, particularly, when data from new countries become available or when data from included countries are no longer available. Another drawback of the missing value approach is the lack of appropriate variance estimation.

The drawbacks of the missing value approach are the strengths of the sampling approach and vice versa the advantages of the missing value approach are the drawbacks of the sampling approach.

[4] Mehran, Farhad, "Estimation of world and regional unemployment", Farhad Mehran, *ILO Bulletin of Labour Statistics*, ILO, Geneva, 1999-4.
[5] International Labour Office, *Every Child Counts. New Global Estimates on Child Labour*, ILO, Geneva, 2002.
[6] Schaible, Wes, *Methods for producing world and regional estimates for selected key indicators of the labour market*, ILO, Employment Sector, Geneva, Employment Paper 2000/6. Vijaya, Ramya, and Wesley Schaible, "World and Regional Estimates for Selected Key Indicators of the Labour Market," forthcoming.