**Committee for the Coordination of Statistical Activities**

**Conference on Data Quality for International Organizations**
**Rome, Italy, 7 and 8 July 2008**

**Session 4: Best practices in imputation and estimation**

# Some Quality Issues in Imputation: ILO Experiences

Sylvester Young, ILO

The quality of imputed values is assessed in terms of whether they are 'Fit for purpose'. This paper discusses some issues that should be considered in making such an assessment, based on ILO's own experiences in this regard. The issues relate to:

(a) the purpose for which the imputed values are required;
(b) the approach taken in making the imputations; and
(c) the dissemination policies used.

## 1. Purpose

Imputation at the national level is usually done to permit analysis of the evolution over time of a characteristic of interest, i.e. trend analysis. At the international level, the objective is to permit analysis of the evolution over time of the **distribution** of a characteristic of interest over a group of countries.

In producing estimates and projections of the economically active population, the ILO makes the following statement[1]:

> ***"These estimates and projections are for international comparisons and are neither superior nor necessarily inferior to national estimates and projections, which are produced using country-specific additional information. This additional information is (a) not necessarily available to us for all countries and/or (b) different between countries and so would make international comparability difficult.***"

As a result, the ILO works with countries in the development of models and methods appropriate to their specific needs rather than promote the direct use of its own models.

**Issue 1**: Given these different purposes, should imputed values derived for a country at a point in time through the two processes be expected to be the same?

[1] Economically Active Population Estimates and Projections, http://laborsta.ilo.org/

## 2. Explicit or implicit imputation

Implicit imputation occurs when no specific value is attributed to missing cases. So, only available data is used in the analysis, e.g. find group mean using only those countries with values. In explicit imputation, a specific value is attributed to missing cases, possibly based on a model. Then, both available data and imputed values are used in the analysis.

In the report prepared for a CCSA session on imputations in 2003, Tim Holt[2] stated that explicit imputation is preferable on grounds of 'transparency and data quality'. He concluded as follows:

> **"Suggestion 6: Agencies should seek to establish explicit imputation methods where thorough empirical analyses can demonstrate that these are robust and methodologically sound."**

It must be noted however that Holt included amongst implicit imputations, cases in which a specific value is assigned but not separately identified in the dissemination of the statistics.

In the ILO exercises, explicit imputations are preferred since they are based on assumptions that can be examined to determine whether the values are acceptable (incorrect assumptions lead to biased imputations).

**Issue 2**: Should we accept Holt's suggestion above?

## 3. Analysis of 'missingness'

It is necessary to examine if there is a certain pattern in the way values are missing when making decisions on how to explicitly impute a missing value using some multiple imputation methodology. Since such a pattern, if it exists, is valuable information that should be used to improve the imputations.

Values may be missing completely at random (MCAR). The missing value is independent of all factors associated with the characteristic of interest and the statistical unit. So, the probability of being missing is the same for all units. Then, there is no additional information that can be used to improve the imputation process. This is usually not the case.

---

[2] Holt, T. (2003) *Aggregation of National Data to Regional and Global Estimates*, Report to the Committee for the Coordination of Statistical Activities, 2nd Session, Geneva, 8-10 September 2003

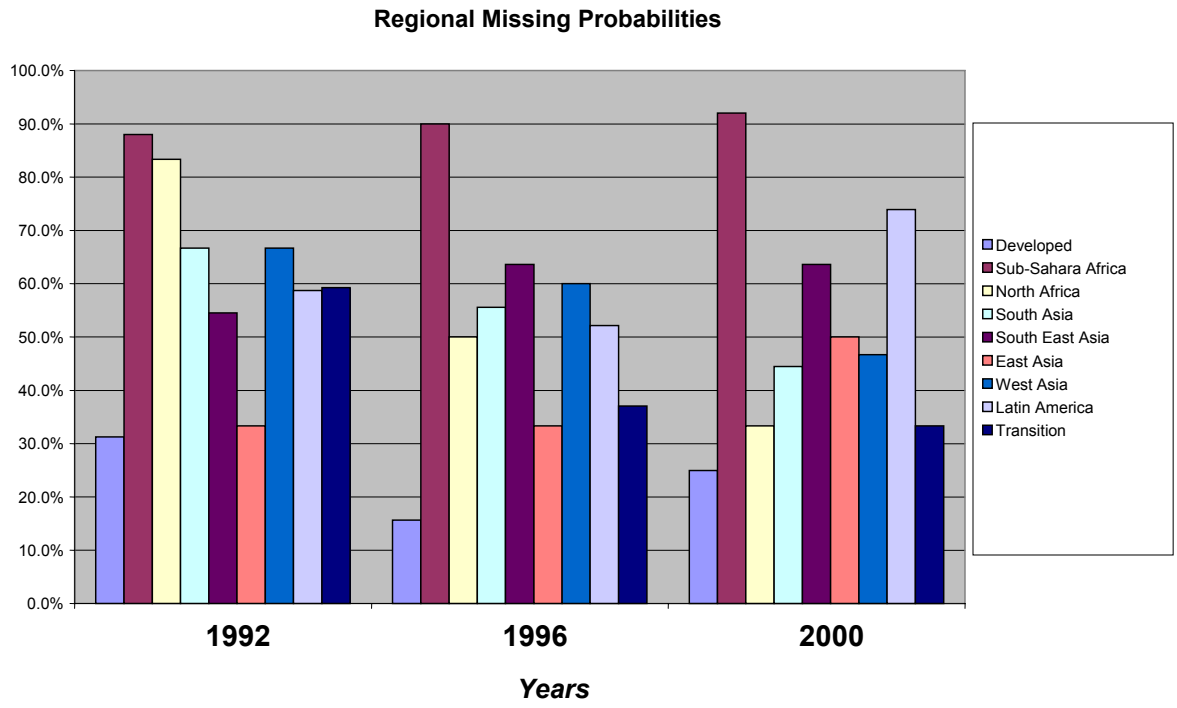
For informative missing values, the probability of being missing depends on available information related to the characteristic of interest or the unit. This is referred to as 'Missing at Random' (MAR). For example, a high probability of being missing is associated with low values of the characteristic or with some factors related to the statistical unit, such as membership of a country in a region. In this case, the information should be taken into account in the imputation. As most imputation methods use regression type models, this could be done through some form of weighting in the estimation of the model or by including the related information directly as variable(s) in the model.

**Example 1**:

In the ILO exercise for estimating labour force participation rates (LFPR) for countries and regions, it was observed that countries with reported LFPR values tended to have much larger per-capita GDP and populations than countries without values. Moreover amongst the countries with reported values those with lower par-capita GDP or smaller population also tended to have larger than average LFPRs. Thus a logistic model for the probability of being missing was fitted using per-capita GDP, population size, membership of the Highly Indebted Poor Country (HIPC) initiative and year dummy variables as independent variables. The inverse probabilities were then used as weights in fitting the imputation model. Thus, reporting countries that are similar to non-reporting countries, with respect to these characteristics, have greater effect on the imputed values for these non-reporting countries.

**Example 2**:

A different approach was taken when imputed values were being determined for the MDG indicator 'Share of women in paid non-agricultural employment'. It was observed that the number of observations available for the indicator from years 1990 to 2002 varied between regions:




Thus countries were stratified into regions and separate imputation models fitted for each region. Alternatively, region could be introduced as dummy variables in a single imputation model for all countries.

**Issue 3:** Is the analysis of 'missingness' important?

## 4. Model Selection and Diagnostics

In choosing the model to be used for imputation, it is important to examine various possibilities. The models tried out could be related on a hierarchical basis (multi-level models) or completely different from each other. They could be based only on time (time series) or include auxiliary variables with or without the time dimension. They may be constructed so that more recent data have greater weight than values that are further away from the missing value, or so that the imputed values respect some second-order consideration of smoothness.

Selection should be based on both goodness-of-fit (e.g. using AIC) and predictive power (e.g. using some jack-knife procedure) of the model. However, as noted by Tabassum and Holt (2004)[3],

> "The model fitting process must not be regarded as automatic and considerable checking and validation of results is required …"

[3] Tabassum, F. and Holt, T. (2004) *Regional and Global Estimates for Indicator 11: Share of Women in Wage Employment in the non-Agricultural Sector, 1990-2001*, Technical Report, ILO

These checks should include testing the reasonableness of the imputations. Many diagnostic techniques are available to do this, such as sensitivity analysis, examination of outliers, comparison of the distributions of the observed and imputed values, graphical displays of the completed data (both observed and imputed) to search for unusual patterns in the values, plausibility checks (using external information or speculation). See Abayomi et al (2008)[4] for some examples.

**Example 3**

Tabassum and Holt (2004)[5] used many of these techniques in checking the optimal model that had been selected in work done for the ILO. Some results from this analysis are as follows:

- A strong negative correlation over time for the imputed indicator values in one region was traced to the many countries with only 2 data points, the first of which (1990) was artificially high. The trend for countries in the region with many more data points was in fact upwards.
- Imputed values were sensitive to the number of countries with missing values as well as the population size of countries with missing values.
- The imputation errors were increased by 2% to derive what they referred to as plausibility ranges (i.e. acceptable ranges) for the imputed values of different countries. Combined with the weights used for regional aggregation, the sensitivity of the regional average to each country's imputation was then computed. For one country, the sensitivity was assessed as a difference of 4.7 to the regional mean.

**Example 4**

In producing estimates of labour force participation rates for a country, using the procedure described in Example 1,

(a) The graph of observed and imputed values for the age group 60 – 64, over the years 1980 to 2006, showed a very steep decline for men from about 70% to about 25% and a rise for women from about 25% to about 64% (See Annex 1). Whilst the trend and values for the women are reasonable according to expectations that persons increasingly stay later in the labour force, the values for the men are inconsistent with such expectations. The anomaly is due to the absence of actual data for all the years except 1980!

(b) The graph of the complete data set over the age ranges 5-9 to 65+ for another country shows a dip in the rates for the age group 55-59 (Annex 2). Normally, one would expect a smooth change in LFPR between ages, although a dip is usual for females in the child-

[4] Abayomi, K., Gelman, A. and Levy, M. (2008) Diagnostics for multivariate imputations, *JR Statist. Soc*., C, 57, 273-291
[5] Ibid

bearing age groups. The observed dip however is after the child-bearing ages, and so unusual.

**Example 6** (Non-ILO)

A hedonic regression was used to estimate imputed rents for owner-occupied dwellings in each of 7 regions in a country. The mean values for both actual and imputed rents for the 7 regions were computed. The mean imputed income was consistently higher than the mean of the actual rents for all the regions. This would suggest that the regression model used for the imputation may be suffering from bias.

Whatever model is selected, there should be periodic re-examination of its performance as new data become available.

**Issue 4**: Should the imputation model be the same for all regions when separate regional imputations are done?

**Issue 5**: What should be done when the imputation model throws up some implausible or unacceptable values? Abayomi et al[6] suggest that "When problems are found, the imputer should refine the imputation model to create improved imputations that are consistent."

## 5. Use of Imputed Values

When should an imputed value of low reliability be used in analysis? An imputed value may have low reliability due to paucity of observed values used to generate it or long intervals in time between the imputed value and the most recent observed value. This can be seen through large imputation errors associated with the value. However when used in aggregations, its influence may be low due its small weight in the aggregation process (low sensitivity of the aggregate to the value) or through compensating errors from other imputed values. It is argued that in these instances, the value could be used in the analysis, but that it should not be identified separately.

**Issue 6**: How many of such values can be tolerated before the aggregate itself reflects the low reliability?

**Issue 7**: How can we test the assumption of canceling errors?

**Issue 8**: Would this view be the same if we are using the entire distribution e.g. looking at the changes over time of the distribution across countries instead of changes over time of the mean of the distribution across countries?

## 6. Dissemination of results

Imputed country values should not be disseminated without proper consultation with the countries involved. This is to ensure acceptance by countries. Holt (2003)[7] proposes that

[6] Ibid

the consultation should be only with respect to the methodology used and not the values themselves.

> "It is important that the methods that underpin the imputation are clearly established on the basis of their statistical properties (rather than the outcome of the particular value) but within this framework it allows the countries concerned to quality assure the process and outcome"

**Issue 9:** Should we accept the above recommendation of Holt?

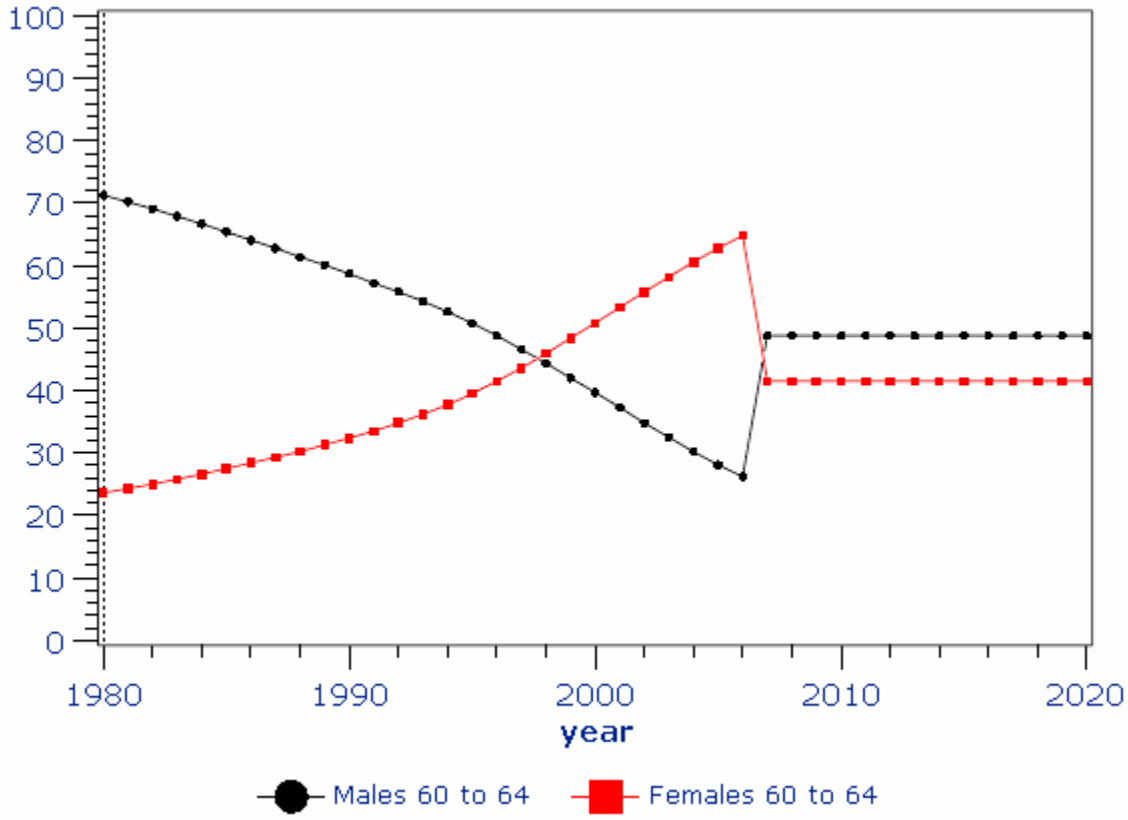
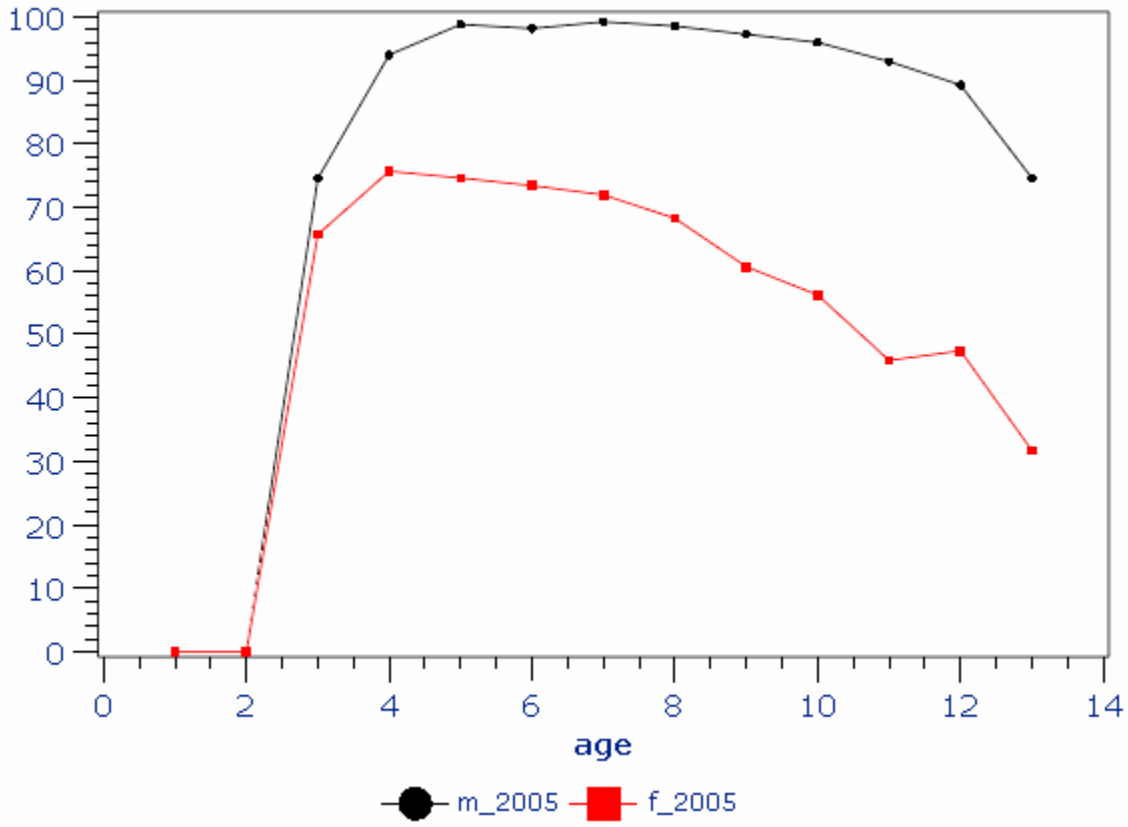
To go further and get acceptance by the community at large, a complete report on the methodology should be publicly available. The report should have details not just of the selected model and its properties, but of the entire process including all models considered, the basis for selecting the one chosen and rejecting the others, any post-fitting adjustments to the values, etc.

Another important issue is that of revision of disseminated values. As new data become available, the imputed values will inevitably change and the analysis may also change.

**Issue 10**: How should the revised values be handled?

[7] Ibid

ANNEX 1: Estimates of labour force participation rates (up to 2006) and projections



ANNEX 2: LFPR Estimates