Review of best practice methodologies for imputing and harmonising data in cross-country datasets

ILO Internal report

Jean-Michel Pasteels

SECOND DRAFT[1] - 28 November 2013

[1] This version has benefited from the comments and suggestions from Patrick Belser, Monica Castillo and Jorge Davalos. The author is very thankful to them.

Contents

I. Review of best practice methodologies for carrying out imputations for missing data in cross-country datasets

I.1 Introduction

Imputation is a statistical technique to estimate missing or aberrant values in a dataset based on collected values from the dataset or comparable data sources.

Initially, imputation techniques have been developed primarily for micro data sets, eg. in the framework of medical studies, non-response in surveys or censuses, dropouts in clinical trials and censored data. Imputation is also frequently used for macro data (eg. cross-country datasets maintained by international organisations). It is also worth mentioning that with cross-country datasets, the number of available records is usually much smaller than with micro-data and the exercise might be delicate when there are many missing data.

Imputation is a general and flexible method for handling problems of missing data, but as highlighted by Dempster and Rubin (1983), it has some pitfalls: *"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to real and imputed data can have substantial biases."*

This review presents a review of the best practice methodologies for carrying out imputations for missing data that could potentially be applied in the case of preparing global and regional estimates on employment, unemployment and labour force by rural and urban breakdowns, if possible, by sex and age.

A similar review has been undertaken recently by Denk and Weber (2011), in the context of a World Bank's project of imputation of labour market indicators in cross-country time series. The aim of that project was to enable the assessment of the labour market situation during the recent financial crisis. The present report is largely based on the complete and excellent work of Denk and Weber (2011). It is also based on textbooks that provide a more detailed introduction to imputation, such as Little and Rubin (2002) and de Waal, Pannekoek, and Scholtus (2011).

The present document also includes an overview of the present methodologies used by the ILO in order to estimate global and regional estimates of various labour market indicators. This includes notably global and regional estimates published in the Global Employment Trends report (ILO 2013b), the Global Wage Report (ILO 2013c), the estimates and projections of the economically active population (ILO 2013a) and the Global estimates of child labour (ILO 2013f).

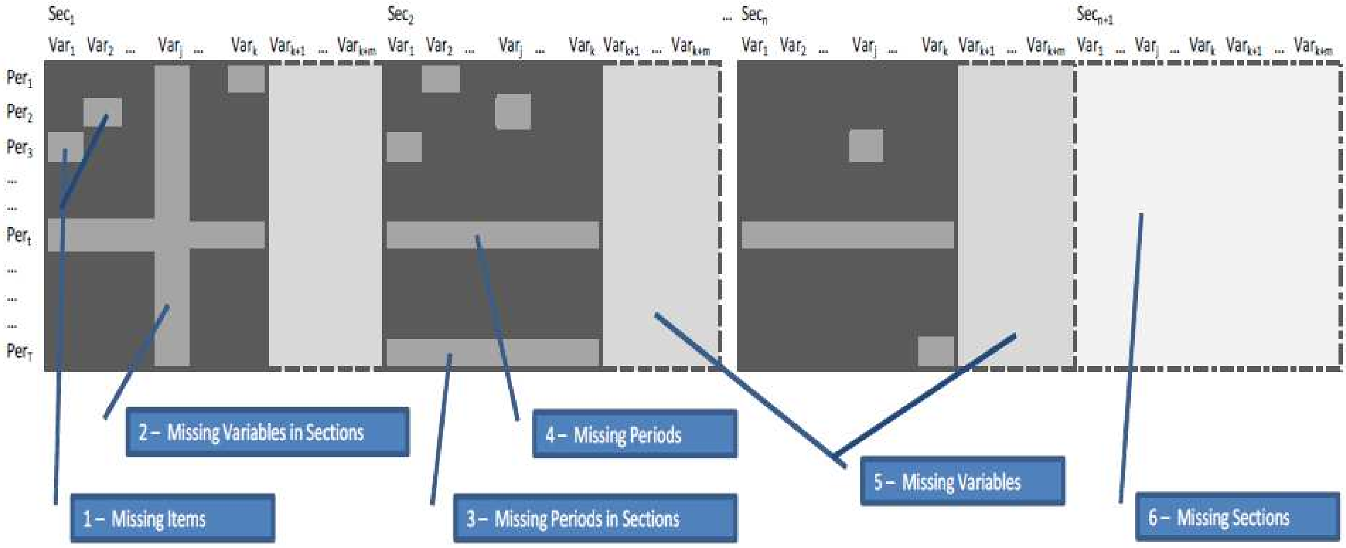
Finally, it is important to highlight that other types of missing data analyses, such as complete record methods or sampling techniques are briefly mentioned in this document but are not examined in detail here.

I.2 TYPES OF MISSING DATA IN CROSS-COUNTRY DATASETS

Different types of missing data may be found. When considering panel data or multivariate cross-sectional time series, the data structure is complex and many missing data patterns may be present as depicted in Figure 1. Sec_k stands for section k, Var_i for variable i and Per_t for period t. The observation (or data point or item) related to section k, variable i and period t is referred as $Var_{i,k,t}$

In the cross-country dataset of our interest, sections represent countries, variables represent labour market indicators (eg. rural unemployment rate, urban unemployment rate, youth rural employment, female labour force participation rate for the total country, …) and periods stand for years.

Figure 1: Missing data types for panel data



Source: Denk and Weber (2011)

As highlighted in Figure 1, different types of missing data can occur:

(1) Missing items or missing data point (eg. rural unemployment rate is not available for 2009 but is available for the other years for country k)

(2) Missing variables in section (eg. rural unemployment rate is not available for any year for country k)

(3) Missing periods in sections (eg. in 2009, labour force survey was not conducted in country k)

(4) Missing periods across all sections (eg. 1985 data for all variables are not available for all countries)

(5) Missing variables across all sections (eg. a given variable is not available for any year for all countries)

(6) Missing sections (there are no data at all for country k, situation that may be referred to as "Statistical No man's land").

One can see that the problem is mild in case (1) and very severe in case (6). For the labour market indicators of our interest, all types of missing data may actually occur. There is still a non-negligible number of countries with no reliable data at all, simply because no properly designed labour force survey were ever conducted or because data for some countries cannot easily be obtained. For EAP broken down by sex and age band, data points are fully imputed for all variables and all years for 17 countries worldwide (ILO 2013a). For employment, unemployment and economically active population, the most frequent cases in the developing world are missing variables and missing periods in sections (see ILO 2013a and ILO 2013b).

The range of tools and techniques used to fill data gaps depend greatly on the missing data pattern. In general, complex methods are used to fill large gaps (eg. missing country) in the dataset while simpler techniques can be used to fill very small gaps (eg. missing item).

I.3 MISSING DATA MECHANISM

As highlighted in all textbooks (eg. Little and Rubin 2002), a critical aspect to be considered when choosing a missing data technique is the underlying missing data mechanism.

Random (ignorable) and non-random (systematic, informative) missing values can be distinguished.

In the case of data missing completely at random (MCAR), the probability that *x* is missing doesn't depend on its value (x) or on the value of other variables (*y*, *z*,…). Therefore, the missing data process is ignorable in imputation. For example, some survey questions may be asked of a simple random sample of original sample.

A weaker assumption that many imputation techniques rely on is data being missing at random (MAR). In that case the missing value for a variable x is independent of the true (but missing) value of that variable after controlling for other variables (say *y* and *z*). In other words, the missingness only depends on other variables which can be taken into account in the imputation procedure. For example, respondents in service occupations (class y) are less likely to report income (x).

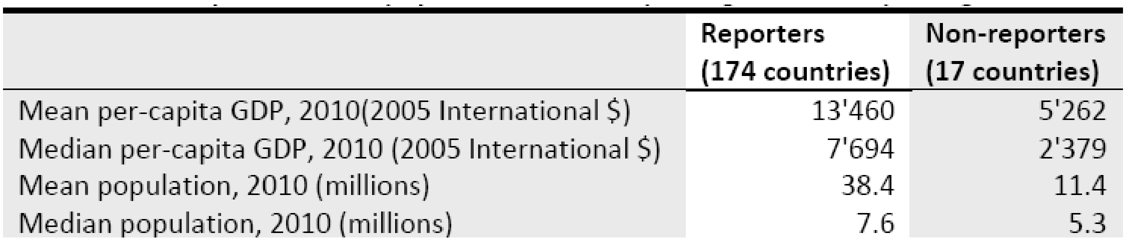
For cross-country data, exogenous shocks such as natural disasters or conflicts are examples of factors that affect the presence of missing data in time series. An imputation method not taking into account this additional information will typically be biased and over- or under-estimate the variable for the missing time period.

If values are missing in a non-random, systematic way (Missing Not at Random, MNAR), the distributions of the variable among complete and missing observations cannot be expected to be the same. This effect is also known as **selection bias**. For MNAR datasets the missing data mechanism may not be ignored.

Missingness at random is relatively easy to handle. Unfortunately, one generally cannot be sure whether data really are missing at random, or whether the missingness depends on unobserved variables or the missing data themselves. The fundamental difficulty is that these potential variables are unobserved, by definition, and so one can never rule them out. Statisticians generally must make assumptions, or check with reference to other studies (for example, similar surveys that include more variables).

In the case of cross-country datasets, data are most often MNAR. There is then a selection problem related to unobservable differences in characteristics among reporting and non-reporting countries. For example, in the EAPEP database (ILO 2013a), there are significant differences between countries that report data on labour force (broken down by age and sex) and non-reporting countries in the sample (see Table 1).

Table 1: Per-capita GDP and population size of reporting and non-reporting countries in the EAPEP database



	Reporters (174 countries)	Non-reporters (17 countries)
Mean per-capita GDP, 2010(2005 International $)	13'460	5'262
Median per-capita GDP, 2010 (2005 International $)	7'694	2'379
Mean population, 2010 (millions)	38.4	11.4
Median population, 2010 (millions)	7.6	5.3

Source: (ILO 2013a)

Table 1 shows that reporting countries have considerably higher per capita GDP and larger populations than non-reporting countries and countries with low per capita GDP also tend to exhibit higher than average labour force participation rates (the variable with missing value), particularly among women, youth and older individuals, suggesting that data are MNAR.

I.4 TYPES OF IMPUTATION METHODS

This section presents in a summarized manner the various types of imputation methods, with a particular focus on the methods applicable to the cross-country dataset of rural and urban labour market indicators. Imputation methods are numerous and there are whole textbooks on the subject (see Little and Rubin 2002 and de Waal *et al*. 2011).

a. Judgmental or deterministic imputation

Judgmental imputation consists in replacing missing values by values that are specified ad-hoc by subject-matter experts, using a manual procedure, or a few rules of thumb.

The problem with this approach is that it does not follow any principle of statistical methodology. It usually distorts the (marginal as well as joint) distributions of the imputed variables. It is however applicable for any type of variable.

Another negative aspect is that any scientific evaluation of its accuracy is hardly possible due to its ad-hoc character. Also, the lack of systematic approach implies the imputation procedures may not replicable in the future (especially as experts come and go).

Nevertheless, if only a few values are missing, the distortion may be negligible. This method is therefore recommended only in the case of very few missing values (missing items).

Deterministic imputation is applicable to all patterns of missing values. For time series data, the "Carry Last Value Forward" is a frequent practice that is simple to apply. It consists in replacing missing values by the most recent available value. This strategy is also used as a simple forecasting method (it is called the naïve method). Actually, this common practice is not so hazardous. It corresponds to the underlying assumption that the time series follows a random walk process or in other words that there is no residual autocorrelation after applying a first difference to the series (see Makridakis *et al.* 1998).

b. Mean and location-based imputation

Mean substitution or more generally “location-based imputation” replaces missing values with a location parameter of the distribution, typically the mean (for metric variables), median (metric or ordinal variables) mode (categorical variables) or any other statistics measuring the center of a distribution.

The mean (or median, mode) is either based on all observed values for a variable (for example using all reported rural unemployment rates) or on all values within a subgroup (stratum) defined by other variables (eg. using all reported rural unemployment rates of low income countries, or of a given region).

This classical method has many drawbacks. According to Graham (2012, p.51), “This is the worst of all possible strategies.” Using overall mean imputation on large parts of a dataset causes serious distortions in the distributions with high peak imputed values and considerably reduced variability of the imputed variables.

Imputing different central values for different subgroups of the data can reduce the distortion, if variables defining the groups are correlated with the variables with missing values.

Further drawbacks of this method are that central values (mean or median), and thus the imputed data point, might take unobservable values and that, with respect to time series, exogenous shocks such as natural disasters or conflicts are not handled satisfactorily, as the missing values for a time period in which a shock occurred will be replaced by the average of time periods without a shock.

This kind of simple approach is not recommended at all in the case of MNAR missing data mechanism as it presumes that data are missing at random.

Location-based imputation should be treated with caution for missing periods in time series. Replacing a missing value for a variable at time *t* (eg. $Rural\ UR_t$) with the mean of the same variable for the same country over all available periods will yield biased and unsatisfactory results in most cases. This problem also applies to substituting a missing value with the mean of the same variable over all countries with available data.

Location-based methods that make more sense and that are frequently used for time series data are interpolation methods and moving averages of the same variable (eg. $Rural\ UR_t$) for the same country over time. There are different methods of interpolation (linear or non-linear such as Splines techniques) and moving averages (symmetric, asymmetric, of different orders).

The danger with interpolation methods and moving averages is that they tend to produce rather smooth curves and are not able to predict exceptional peaks and troughs that may correspond to unusual years (eg. economic crisis, conflict, natural disaster). This is especially critical in case of gaps larger than one period and also for variables that are very sensitive to cyclical or accidental changes. Such techniques are therefore recommended for variables of structural nature with low volatility or that follow long-term trends and for situations when only a few data points are missing.

c. Distribution-based or probabilistic imputation

Instead of using one parameter of the distribution (eg. the mean or the mode), the principle is to use the entire empirical distribution of a variable. The probabilities for the occurrence of observed values of a variable are estimated on the basis of the empirical distribution function (non-parametric) or a parametric distribution based on a distribution assumption (eg. Gaussian or Poisson distribution).

A commonly used probabilistic method is the expectation–maximization (EM) algorithm. The EM algorithm is an iterative method used to find the maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either there are **missing values** among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

d. Model-based imputation

Econometric models are used by economists for a wide range of applications (see Harvey 1989), for example to validate some economic theory, to generate economic forecasts or to undertake business cycle analysis (time series decomposition). Econometric models can also be used to generate imputed values.

There are often used for cross-country data as sample sizes on macroeconomic data (mainly break down statistics, by region, gender, etc.) are often small, which motivates the implementation of model-based approaches.

When working with ordinal variables, the principle is to use correlations between available explanatory variables or predictors (eg. per capita GDP, gross value added in agriculture) and variables with missing values (eg. rural unemployment rate) to estimate a model in order to predict the missing values. The model might be linear or non-linear and it can be estimated by various econometric techniques (Ordinary Least Squares, Generalized Least Squares, see Harvey 1989).

It is worth mentioning that relations between labour related variables are often non linear thus linear approximations might have poor predictive performance. Econometric packages have standard routines (iterative routines) for the estimation of non-linear models.

Also, when working with non-ordinal variables (such as categorical or count variables), specific econometric techniques have been developed. For example, logit or probit models may be used for categorical variables (see Cameron and Trivedi 2010 for a description of the different econometric model corresponding to different types of variables).

Conceptually, model-based imputation is a good way to impute values. It is good in the sense that a great deal of information from the individual (sourced from micro-data) or country (sourced from macro-data) is used to predict the missing values.

For panel data, an imputation model commonly used for labour market variables (see ILO 2013a and 2013b) is a linear model with fixed effects of the following form:

$$Y_{it} = \alpha_i + x_{it}^{'}\beta + e_{it}$$

where *t* represent the period (eg. year) and *i* the country, Y_{it} the dependent variable (ordinal variable), α_i is country-specific fixed effect, χ_{it} is a set of explanatory covariates of Y_{it} and e_{it} is the error term. The missing value is filled by the predicted values Y_{it}, or in other terms by an estimate of the conditional mean E(Y_{it} | χ_{it}).

The fixed-effects are country specific constants (intercepts) that control for all the country-specific factors that influence the dependent variable Y_{it} and that are not captured by the other covariates.

Different variants to this fixed-effect model are often used. For example, allowing parameters β to differ by group of countries (that share the same characteristics) or even to introduce country-specific parameters (β_i). When there are structural breaks in time series, it is also frequent to introduce time dummy variables.

In practice, the higher correlation between the predictors and the missing variable(s), the better the imputation will be. Importantly, as the model is used for predictive purposes and not analytical purposes, all available variables that may improve the prediction should be used. In other words, the overall measure of quality of fit should be maximized (R2 in the case of OLS). Analytical aspects such as causality or colinearity are of secondary importance (see Makridakis *et al*. 1998). The danger is more in leaving out useful predictors than in including too many unimportant variables. Traditional automatic model selection procedures such as stepwise variable selection are therefore not recommended and it is advised to specify the model in a manual way in order to identify the "best" model. More recent model selection techniques, such as the least-angle regression (LARS) algorithm may also be recommended in the context of regression-based imputations. The LARS algorithm performs an automatic model selection based on a cross validation statistic (out-of-sample criteria). Some of the advantages of the LARS method (see Efron et al. 2004 for more detail) are that if two variables are almost equally correlated with the response, then their coefficients should increase at approximately the same rate and also it is effective in contexts where when the number of dimensions is significantly greater than the number of points.

For micro-data, explanatory variables involved in the survey design are of relevance (eg. variables that define the different strata).

For macro-data, the exercise is more difficult as it is often the same countries that lack data for all types of macro-economic indicators (see for example the different variables included in the World Development Indicators database maintained by the World Bank[2]). Therefore, it is often the same variables available worldwide (GDP growth and per capita GDP adjusted for purchasing power parity) that are used to compute missing values for various labour market indicators in cross-country datasets.

Another key element in the choice of explanatory variables is related to the reliability of the explanatory variables in cross-country regressions. Often, explanatory variables that include both real data and imputed values for a non-negligible number of countries are used to impute missing values of other variables. This aspect is often neglected as databases including many imputed values published by international organizations are often perceived as original data for all countries. One can easily see the danger or temptation of "filling holes in Swiss cheese using whipped cream", as stated in the title of the paper of Denk and Weber (2011).

Imputation model building is a time-consuming task, as the quest for the best model implies not only selecting useful explanatory variables but also checking the data sources and reliability.

Model-based imputation better preserves the individual and joint distributions of imputed variables than the previously described techniques. It also usually reduces bias in the estimation of aggregates such as means and totals based on the completed dataset. When well specified, model-based imputation also allows taking into account external shocks.

Model-based methods are most frequently adopted for imputing missing values in incomplete time series. Models may be built across different dimensions (cross-sectional, time series models or panel models). Time series techniques applicable to imputation include a variety of models.

[2] http://data.worldbank.org/data-catalog/world-development-indicators

The most commonly models used in time series analysis are ARIMA models (see Makridakis et al. 1998 and Melard 2008). These models are complex to identify and specific but in recent years most statistical packages include possibilities of automatic ARIMA modelling with a great deal of accuracy. This includes software available to the public domain such as JDemetra+, TRAMO/SEATS and X-13 ARIMA/SEATS[3]. One commonly used technique is to include dummy variables (also called intervention variables) in the model for each missing data point. This implies that the number of missing observations should not be too large in order to obtain enough degrees of freedom and identify adequately the ARIMA model.

More complex methods used for modelling univariate times series exist. They can be used for forecasting purposes but also for imputations. These models are not described here, due to their complexity and as they are used mostly for academic purposes and rarely used in the context of imputations for cross-country data. They include state space models (e.g. Durbin, Koopman, 2004), curve fitting or smoothing algorithms (e.g. He, Yucel, Raghunathan, 2011), Kalman filters (Harvey 1989) or other types of dynamic Bayesian networks (e.g. Pearl 2009).

Another crucial aspect to take into account in model-based imputation is that many labour market variables are proportions or ratios (eg. labour market participation rates or unemployment rate), that vary between 0 and 1. When using ordinary least squares, the risk is that predicted values fall outside this range (eg. resulting in a negative unemployment rate). One approach adopted in several ILO models (notably ILO 2013a and ILO 2013b) is to apply a logistic transformation to the ratio variable *y*: $Y = \ln\left(\frac{y}{1-y}\right)$ and then to run the regression on the transformed variable *Y.* In the next step, the missing values are computed on the basis of the inverse transformation applied to the fitted values $\hat{Y}$: $\hat{y} = e^{\hat{Y}} / (1 + e^{\hat{Y}})$ or $\hat{y} = 1/(1 + e^{-\hat{Y}})$

This process guarantees imputed values *y'* within the 0%-100% range. There is however a problem with this approach as it results in biased estimates of the conditional 'mean'. The reason of this bias is that the estimate of the expected value used for imputation of the variable of interest **is not equal** to the inverse of the logit function[4].

This is a known issue and can be addressed by implementing a non-linear estimate of the expectation. One way to accomplish this is to use a generalized linear model. The theory behind can be found on the generalized linear models literature (see Dobson and Barnett 2008). This approach is already implemented in Stata[5] (using the glm command with a logit link and the binomial family).

It is also important to mention that model-based imputation can be applied to MAR as well as MNAR missing data. In cross-section analysis, one traditional way to handle MNAR data is to use the Heckman selection model. This technique is available in most econometric packages

[3] JDemetra+ is developed by EUROSTAT and the National Bank of Belgium (http://www.cros-portal.eu/content/jdemetra-user-and-developers-documentation) TRAMO/SEATS is developed by the Bank of Spain (http://www.bde.es/bde/es/secciones/servicios/Profesionales/Programas_estadi/Programas_estad_d9fa7f3710fd821.html), Win X-13 is developed by the US Census Bureau (http://www.census.gov/srd/www/winx13/).

[4] The most common example of the same issue is the exponential transformation. Consider $y' = e^{(a+\mu)}$ where μ follows a Normal(0,σ^2) distribution, then the expectation of y is $e^{(a+\sigma^2/2)}$ and not e^a.

[5] See the following example: http://www.ats.ucla.edu/stat/stata/faq/proportion.htm

such as Stata[6]. The Heckman selection model assumes that there exists an underlying regression relationship: $y_i = x_i \beta + u_{1i}$

The variable y_i is observed if $z_i \gamma + u_{2i} > 0$

where

$$\begin{aligned} u_1 &\sim N(0, \sigma) \\ u_2 &\sim N(0, 1) \\ \text{corr}(u_1, u_2) &= \rho \end{aligned}$$

χ_i is a set of explanatory variables of y_i

Z_i is a set of variables thought to determine whether y_i is observed or unobserved (selected or not selected). Z_i includes that all variables in χ_i

When ρ is different from 0, standard regression techniques applied to the first equation yield biased results. The Heckman procedure provides consistent, asymptotically efficient estimates for all the parameters in such models. The procedure also allows to test if the data is MNAR or not (if ρ is statistically different from 0). Note that in Stata, the Heckman model can be estimated using Maximum likelihood (ML) estimation or using two-steps estimation procedure (binomial probit and then a regression), which is recommended for large datasets as ML can be time consuming for such datasets.

In panel data, the extension of the Heckman selection model is however not straightforward (see Wooldrige 1995 for more detail), as the required sample selection correction terms for two dimensional data (panel data) differ from the single dimension ones (cross-section).

e. Donor-based imputation

The principle is to take imputation values from a so-called donor (e.g. a household in a survey or a country for cross-country data) that has complete observations for all variables and also similar characteristics as the incomplete observation (the recipient). The similarity between donor and recipient is determined via matching variables (ordinal or categorical) to be selected based on their correlation with the variable to be imputed. For micro-data (e.g. a household survey), the matching variables may be the level of education, the number of children, the income, etc. For cross-country macroeconomic data, the matching variables may be the GDP per capita, the population density, but also variables of qualitative nature (cultural factors, structure of the economy, labour market legislation, etc.).

Donor-based imputation is used frequently for imputing values in surveys and it is also often used for categorical variables.

Two classical methods are **hot-deck**[7] and **cold-deck** imputations. Hot-deck and cold-deck imputations group all the observations of a dataset into subsets that share the same values in the matching variables. For hot-deck imputation, the donors (households) are selected from the same dataset (the one with missing values) while for cold-deck imputation, the donors are extracted from other comparable data sets (e.g. a similar household survey).

The selection of the donor can be carried out in different manners (sequentially, randomly, based on distances with respect to matching variables, or based on ranks with respect to a common ordinal matching variable.

[6] For its implementation in Stata see, http://www.stata.com/manuals13/rheckman.pdf

[7] The term "hot deck" dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot" because it was currently being processed.

A more recent donor-based imputation method is the **nearest neighbour method**. It consists in measuring the distance between the subset with complete data and the subset of records with missing values. This is done usually on the basis on ordinal matching variables. It is also possible to compute distances in the context of matching variables of mixed types, ordinal and categorical (see Tarsitano and Falcone 2010).

Different procedures exist. The donor record can either be the nearest neighbour (in terms of computed distance) or one of the nearest neighbours selected randomly.

Also, it is possible to adopt a **multi-donor** approach instead of choosing one particular donor from the set of potential donors. In this case, the set of donors (eg. households or countries) can be combined by calculating the imputed value as a (weighted) average or median of the donors' values. The principle of this approach is similar to the model-based approach described previously.

Different weighting schemes are used in practice. For example, the weights can be set on the basis of similarities between donor and recipient or/and to the frequency of a donor already being used for other recipients.

Statistical matching can be regarded as a particular type of donor-based imputation. It is a statistical technique that combines two datasets. It enriches a recipient dataset with variables only available in a donor dataset by combining observations from the two datasets based on the similarity of matching variables that are available in both datasets.

In medicine, it is notably used to evaluate the effect of a treatment by comparing the treated and the non-treated units in an observational study or quasi-experiment (i.e. when the treatment is not randomly assigned). The goal of matching is, for every treated unit, to find one (or more) non-treated unit(s) with similar observable characteristics against whom the effect of the treatment can be assessed. The matching process gives rise to completed observations with variables that were completely missing in the recipient set imputed from the donor set.

There are different variants of statistical matching, such as constrained matching, equivalence class matching or regression-based matching (see De Waal 2011 for more details).

In **constrained matching**, a constraint is put on the weights. Every recipient as well as every donor observation is included in the final dataset with a sample weight identical to its sample weight before matching. In economics this matching technique is employed to estimate counterfactuals i.e. missing variables by definition.

With **equivalence class matching**, the donor and recipient datasets are subdivided into comparable subsets (= equivalence classes) of observations in different possible ways, such as similarity of matching variables or cluster analysis. To each recipient in a subset one or more donors from the same subset can be assigned. Donors may be selected randomly or on the basis of distance measures (when using cluster analysis). Multiple donors can be combined by some aggregation function, e.g. mean, median, or mode, depending on the type of variable.

In **regression-based matching**, the matching between the recipients and donors is based on similarity of additional variables estimated in both datasets (Raessler, 2002). These additional variables are estimated by means of regression models, in which the common matching variables are used as explanatory variables (or regressors). Regression-based matching is comparable to model-based imputation based on two datasets, but instead of using the estimated value as imputation, the value of the nearest potential donor with respect to the estimated variables is used.

Donor-based approaches can be used for all types of missing data (missing variable, missing period, missing items). For the imputation of a missing variable, the donor section can be selected based on the similarity of the trajectories over time of donor and recipient. This donor is also called “time series donor”. For cross-country analyses, this may concern many variables of structural nature (eg. fertility and mortality rates in demographic, female labour force participation rate, etc.).

Like in model-based imputation, the selection of the matching variables is very important. It is of a subjective nature and it affects greatly the quality of the match. Also, if the number of matching variables is large, the number of potential donors may be very small. On the other hand, the selection of very few matching variables may result in a poor match.

Unlike the model-based imputation approaches, donor-based methods can deal with recipient variables of categorical nature. Often, model-based approaches are preferred in case of metric recipient variables, whereas donor-based methods are preferred for variables of categorical nature.

f. Combining inferences from multiple imputations

Imputation, the practice of 'filling in' missing data with plausible values, is an attractive approach to analyzing incomplete data. It apparently solves the missing-data problem at the beginning of the analysis. However, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin (1987) and others.

The question of how to obtain valid inferences from imputed data was addressed by Rubin's (1987) book on multiple imputation (MI). MI is a Monte Carlo technique in which the missing values are replaced by $m>1$ simulated versions, where m is typically small (between 3 and 10).

Multiple imputation allows to reflect the uncertainty about the imputation model. For example, if one impute using a regression model one may want the imputations to reflect not only sampling variability (as random imputation should) but also the uncertainty about the regression coefficients in the model. If these coefficients themselves are modelled, one can draw a new set of missing value imputations for each draw from the distribution of the coefficients.

Multiple imputation does this by creating m (usually about five[8]) imputed values for each missing value, each of which is predicted from a slightly different model and each of which also reflects sampling variability. The simple idea is to use each set of imputed values to form (along with the observed data) a completed dataset. Within each completed dataset a standard analysis can be run. Then inferences can be combined across datasets. For parameters (e.g., means or regression coefficients), the MI estimate is simply the mean of parameter estimates across the imputations. The calculation of the standard errors is a little more complex (see De Waal 2011).

It has been demonstrated on the basis of simulations (that is using full datasets and generating “false” missing values) that multiple imputation provides more accurate statistical inference (Little and Rubin 2002), reducing the biases in regression coefficients and standard errors. It is worth mentioning that the imputed values themselves are similar to the ones obtained from simple imputation.

Performing multiple imputations requires however considerably more steps than single imputations. Recent versions of statistical software such as Stata 11 provide a suite of built-in commands for performing multiple imputations.

[8] Multiple imputation is available as a built in command in version 11 (and onwards) of Stata (See Stata 2013). The default number of multiple values is five.

I.5 ALTERNATIVE MISSING-DATA METHODS

While the present document focuses on imputation techniques, there are other approaches to missing data methods (see Little and Rubin 2002 for a complete review). They can be classified in different families, which are not mutually exclusive.

a. Deletion or Complete Record Methods

An alternative to imputation is the deletion of records with missing data. List-wise deletion refers to the deletion of all records for which there is at least one missing variable (eg. by country). Pair-wise deletion on the other hand is less strict as it seeks to undertake analysis using all cases for which the variables of interest are present.

With these procedures, a large volume of information gets lost and biased estimates are a frequent consequence. In the case of data MNAR, the bias may be serious, in particular with cross-country datasets of labour market indicators as highlighted above.

b. Weighting or Sampling Procedures

These procedures are typically used for surveys (see Little and Rubin 2002). The principle is to assign a weight to an observation (eg. a household) when computing the population mean for the variable of interest. For example, a weight of 2 means that the record counts in the dataset as two identical records.

The two most common types of weights are design weights and post-stratification or non-response weights.

Design weights are normally used to compensate for over- or under-sampling of specific cases or for disproportionate stratification. For example, it is a common practice to over-sample minority group members or persons living in areas with higher percentage minorities. The design weights are used when we want the statistics to be representative of the population.

Post-stratification or non-response weights are used to compensate for that fact that persons with certain characteristics are not as likely to respond to the survey as others. For example, most general population surveys have substantially more female than male respondents (often 60/40) although there are often more males in the population. Because the survey over-represents females and under-represents males in the population, a weight is used to compensate for this bias.

Calculating post-stratification weights or non-response weights is more difficult then design weights. It requires the use of auxiliary information about the population and may take a number of different variables into account. Information usually needed in a survey are population estimates of the distribution of a set of demographic characteristics that have also been measured in the sample. For example, information found in the Census such as gender, age, educational attainment, household size, residence (e.g., rural, urban, metropolitan) or region. Post-Stratification Weights are notably used in the US, for the US Census tabulations, the Current Population Survey and The American Community Survey (ACS).

There are different techniques to estimate the weights (e.g. use separate frequency tables, logistic regression). The problem becomes complex when the weights need to be adjusted for multiple population characteristics (eg. under-representation of young males in rural areas). Therefore, it is crucial to find good estimates for the population characteristics.

The main drawback of weights is that they primarily adjust means and proportions but this may adversely affect inferential data and standard errors. Weights almost always increase the standard errors of the estimates. Finally, very large weights (or very small ones) can also introduce biases and high standard errors.

This type of method may therefore not always be well suited for cross-country macro data, as the population characteristics (whole set of countries) may not well identified or more specifically it is often the same countries that lack data for all types of macro-economic indicators. Using only a few population determinants such as population size or GDP can also lead to poor, extreme and unstable weights.

c. Model-based statistical procedures

Like weighting procedures, model-based statistical procedures do not imply to compute imputed values for missing data.

There is a broad class of complex model-based statistical procedures that are described in detail in Little and Rubin (2002). The principle is to define a model for the observed data and basic inferences on the likelihood or posterior distribution under that model. The parameters are estimated by procedures such as maximum likelihood.

The maximum likelihood method identifies the set of parameter values that produces the highest log-likelihood (most likely to have resulted in the observed data). Conceptually, the process is the same with or without missing data.

According to Little and Rubin (2002, p. 20), "*the advantages of this approach are flexibility, the avoidance of ad hoc methods, in that model assumptions underlying the resulting methods can be displayed and evaluated; and the availability of estimates of variance that take into account incompleteness of the data*."

These procedures are primarily used for micro data. In addition, large samples are needed in order to apply most of these methods. Probably, due to their complexity and the need for large samples, it is difficult to find in the literature applications of these procedures to cross-country datasets.

I.6 Evaluation of the imputation methods

In addition to using an appropriate imputation method and document the underlying assumptions, it is also crucial to include in the documentation indicators assessing the quality of the imputation method(s).

These aspects are discussed in detail in Denk and Weber (2011) and are summarised here.

The quality of the imputation methods can be judged using different criteria:

(i) Indicators of the performance of the applied method

(ii) Indicators of accuracy of the imputed values

(iii) The variability of statistics based on the imputed dataset

(iv) The plausibility of imputed values

(i) Indicators of the performance of the applied method

Performance criteria and indicators are typically method-specific. For some elementary imputation approaches, such as judgmental or mean imputation, performance criteria are not available.

For model-based methods, there are several quality criteria available for each method. For multiple regression models, the usual overall measures used are the coefficient of determination (R^2), the adjusted R^2 or other statistics (Root Mean Squared Error, Fisher Statistics, etc.). In order to analyse the explanatory power of each explanatory variable, the significance probabilities ("p-values") of each regression coefficient are often displayed. For non linear models, alternative measures exist such as the Pseudo R^2.

What is also important is to provide in the documentation the standard error and confidence interval for the predicted (imputed) values, when available (depending on the method used).

For donor-based methods, two types of performance measures are usually used; the distribution parameters of the usage frequency of individual donors and the values of the distance function between donor and recipient. For example the closer the variable of recipients and their donors are (the smallest distance), the higher is the quality of the imputed values.

(ii) Indicators of accuracy of the imputed values

Indicators of accuracy are based on out-of-sample simulations. Since the true values of the imputed data are unknown, the imputed values cannot be compared to their true counterparts.

Hence, accuracy indicators are estimated by treating available values as missing, imputing these artificially missing values, and comparing the imputed values to the ignored true values.

This simulation technique of leaving out observations[9] in an estimation procedure to validate estimation results is known as **cross-validation[10]**. There are different types of cross-validation techniques. K-fold cross-validation, and leaving-one-out cross-validation are the most common types of cross-validation.

According to Chambers (2000), for the validation of imputation results the leaving-one-out approach is typically used. The principle is to drop one value from the sample at a time and repeat the imputation procedure for each artificial missing value. It is a time consuming process, so, in practice, the repetition is only carried out over a selected sub-sample of the available values.

Cross-validation is usually separately conducted for each variable with missing values (eg. unemployment rates). Overall accuracy measures can be derived from these variable-specific accuracy measures by aggregation (e.g. mean error, median error, root mean squared error). Actually, these are the same measures used for assessing the accuracy of forecasting errors (see Makridakis *et al*. 1998), such as the Mean Absolute Error (MAE) which is expressed in the unit of the variable (eg. number of employed) and the Mean Absolute Percentage Error (MAE) which is expressed in relative terms and comparable across variables (eg. percentage error in the number of unemployed).

According to Chambers (2000), four types of imputation accuracy can be discerned:

a) Predictive accuracy or effectiveness: maximum preservation of true values (of each imputed value);
b) Ranking accuracy: maximum preservation of true ordering (ranks) relationship in imputed values;
c) Distributional accuracy: maximum preservation of the distributions of true values;
d) Global estimation accuracy: maximum preservation of analytic results and conclusions.

[9] That is why the "out-of-sample" term is also used.

[10] Cross-validation belongs to the family of resampling techniques, that include Bootstrap, Jackniffe and Permutations tests (of which Monte Carlo) simulations.

This typology constitutes a hierarchy. Fulfilment of predictive accuracy (a), which is the strongest type of accuracy, implies de facto the other three types of accuracy.

The relevance of predictive (a) and ranking accuracy (b) depends on the intended usage of the completed dataset. If the dataset is to be publicly released or used for the development of prediction models, imputation accuracy types (a) and (b) are crucial (this concerns for example ILO estimates of economically active population, see ILO, 2013a).

If the objective is to produce and publish aggregated estimates (like in ILO 2013b or ILO 2013c), aspects (a) and (b) are less important. What matters is more the plausibility of the aggregate (eg. world or regional estimates). There could potentially be a situation of poor accuracy of imputed values at the country level (poor accuracy of types (a), (b) and (c)) but if there is no global bias (systematic overestimation or underestimation), the overall aggregate and its analysis would remain relevant.

Predictive accuracy (a) can be assessed using overall measures of aggregation such as the Mean Absolute Error or the Mean Absolute Percentage Error, which treat each error in percentage terms. For the global estimation accuracy (d), the mean error is recommended (errors of different signs cancel each other out) as it will provide a measure of overall bias.

Correlation coefficients between imputed and true values can be used for assessing accuracies of types (b) and (c). It is preferable to use the rank correlation coefficients, in order to assess the preservation of the ordering relationship of imputed values (accuracy of type (c)). The principle of rank correlation coefficients, such as Spearman's rank correlation coefficient and Kendall's rank correlation coefficient is to replace the observations by their order (ranks) and calculate the correlation coefficient between the ranked values.

(iii) The variability of statistics based on the imputed dataset

It is very important to assess any potential bias in the variability of the estimates based on the dataset completed with imputed values. As highlighted in Denk and Weber (2011), "*the complexity of deriving closed-form solutions for variance and bias increases rapidly with the complexity of the missing data patterns and the imputation method. In general, the scope of theoretical work on the direct calculation of variance and bias is limited to rather simple constellations of missing data*. *This may be one reason for the neglect of the effect of imputation on the variance and bias by many analysts. Thereby, variances are underestimated and the validity of confidence statements is jeopardized*".

Like indicators of imputation accuracy (ii), simulations based on re-sampling techniques, , such as multiple imputations (see section I.4.f) can be used in order to evaluate bias and variance of imputation results of more complex imputation procedures.

(iv) The plausibility of imputed values.

Imputed values must be plausible. This can be checked through statistical data editing, i.e. the process of improving the data quality by detecting and correcting errors. It includes various procedures, either manual (eg. analysis of charts and tables) and automatic (eg. consistency checks and automatic outlier detection).

Outlier detection can be applied to the original dataset (with missing values) and completed dataset (i.e. after imputation) and then compared. Outlier imputed values should be inspected carefully, as they may identify a potential problem in the methodology.

The comparison between the original and completed dataset can be done also in terms of analysis of the dispersion indicators or the variables (eg. range or percentiles) or in terms of aggregated measures (eg. world estimates).

The pattern of imputed values should also be plausible over time (time series plausibility). Outliers (implausible peaks or troughs) may appear in a time series for an imputed variable for a given country, especially if imputation was carried out across countries or across variables instead of over time.

I.7 Missing data and imputations methodologies used by the ILO

For each report or dataset, the method for handling missing data is briefly described. Also, the overview describes the evaluation techniques of the imputation methods used in each report.

1. ILO Estimates and Projections of The Economically Active Population (EAPEP): 1990-2030 (ILO 2013a).

The main objective of this ILO programme is to provide member states, international agencies and the public at large with the most comprehensive, detailed and comparable estimates and projections of the economically active population in the world and its main geographical regions. The 2013 edition covers a sample of 191 countries. The reference period for the estimates is 1990-2012 and for the projections is 2013-2030.

For the 191 countries, the output dataset includes estimates on the economically active population and labour force participation rates by sex and age groups for each year. The estimates are broken down by 11 age groups (15-19, 20-24, ..., 60-64 and 65+).

The input data set consists of labour force participation rates (LFPR) by sex and age group reported by countries. In the 2013 edition, there were no data at all for 17 countries (missing section). Prior to the imputation process, data are harmonised for many countries, in terms of geographical coverage (some countries only report urban data) and age band (many countries report unusual age bands such as 16-19, 20-54,…)[11]. After harmonisation, the proportion of reported (including adjusted) values in the dataset is around 30%.

Imputations

The current methodology contains several steps. First, in order to ensure a realistic imputed LFPR (within the 0%-100%) range, a logistic transformation is applied to the input data file. The formula is the following:

$$Y_{it} = \ln\left(\frac{y_{it}}{1-y_{it}}\right)$$

where y_{it} is the observed labour force participation rate by sex and age in country *i* and year *t*.

Second, a simple linear interpolation technique (location-based) is applied at the country level in order to estimate missing years between two reported years. For instance, a country will report labour force participation rates in 1990 and 1995, but not for the years in between. The imputed values for 1991, 1992, 1993 and 1994 will simply fit on the straight line joining the two observed values (1990 and 1995). This procedure relies on the simple assumption that structural factors are predominant (and follow a linear pattern) as compared to the

[11] Note that the data are not fully harmonised, as they are not adjusted for the exclusion of military forces and prisoners from some surveys, the difference in the treatment of subsistence workers across countries and the use of different reference periods (whole year, month).

cyclical and accidental ones. The proportion of imputed values generated by this procedure is around 19%.

The final phase consists in using panel data in order to impute the remaining proportion of missing values (estimated to 51% in the last exercise). For each of the 8 sub-regions a panel regression is run in order to estimate the missing data along both the country and time dimension (missing items, sections and variables) for all sub-components. There are 22 sub-components for the EAP model (11 age bands by sex).

Importantly, the set of observations includes both real data (more exactly survey-based estimates) as well as imputations obtained at the country level.

Unweighted panel regressions are used for Europe and non-Europe developed countries, for which there are no or very few missing observations.

For the other regions, the problem of the MNAR missing data mechanism is addressed and weighted panel regressions are used to correct for non response bias. Weights are used in panel regressions to diminish the influence of countries that are less similar to non-reporting countries (based on a set of covariates), and to increase the influence of countries that are more similar.

The weights are estimated on the basis of logit regressions. The explanatory variables used in the EAP model include the following country-specific variables: economic growth, population size, per capita GDP and membership in the Heavily Indebted Poor Countries Initiative (HIPC). On average, reporting countries tend to have higher per capita GDP and larger populations than non-reporting countries.

For all regions, a panel regression (weighted or unweighted) is estimated. In order to preserve the unobserved heterogeneity of the various countries, fixed-effects by country are introduced.

The following linear model is constructed (and run on the logistically transformed dependent variable):

$$Y_{it} = \alpha_i + x_{it}^{'}\beta + e_{it}$$

where α_i is country-specific fixed effect, Y_{it} the LFPR (logistically transformed), χ_{it} is a set of explanatory covariates of the LFPR or and e_{it} is the error term.

The fixed-effects model controls for all the country-specific factors that influence the dependent variable (LFPR). The covariates of the EAP model are:

- Per-capita GDP, Per-capita GDP squared

- Real GDP growth rate, Lagged real GDP growth rate

- Share of population aged 0-14,Share of population aged 15-24,Share of population aged 25-64

In the final step, the missing values are computed on the basis of the inverse transformation applied to the fitted values $y' = e^{Y'} / (1 + e^{Y'})$.

Evaluation of imputations

The main evaluation information provided for the EAPEP imputations are indicators of the performance of the applied method. For all regressions and logit regression, overall measures of quality of fit (R^2 and Pseudo R^2) are provided as well as regression coefficients and their statistical significance.

Recently, indicators of accuracy of the imputed values (based on simulated missing values) are also provided for the country level interpolation technique. The mean (non-weighted) absolute errors expressed in percentage points are provided for different age bands, age groups and distance between available data points. For example, the error in time *t* for a distance of 1 can be calculated for the countries which report at least three consecutive years and is obtained as follows:

$$\varepsilon_{t(d=1)} = \left|Y_t - \hat{Y}_{t(d=1)}\right| = \left|Y_t - (Y_{t-1} + Y_{t+1})/2\right|$$

Where Y_t represents here the LFPR (**not** logistically transformed) for year *t*. Similarly, the error in time t for a distance of 10 years can be calculated for the countries which report data at t-10, t and t+10 and is calculated as follows:

$$\varepsilon_{t(d=10)} = \left|Y_t - \hat{Y}_{t(d=10)}\right| = \left|Y_t - (Y_{t-10} + Y_{t+10})/2\right|$$

For example results indicate that the errors at distance 1 year are significantly lower than for subsequent years and that for the prime age (25-54), errors for male LFPRs are the lowest across all regions in the world and do not increase much with distance. The interpolation method is therefore justified for that subgroup of the population, whatever the distance between two available observations. For all the other subgroups of the population, the story is however different and imputation errors increase significantly with the distance.

Such information is very useful. It is however not yet provided for the imputations derived from the other imputation technique, the panel data, which actually generated 51% of the observations in the last exercise. So the accuracy of the imputed values by the panel data remains to be explored, especially for regions with little data.

The plausibility of imputed values is also analysed on the basis of outlier detection and chart analysis. For the 17 countries with no data at all, the imputed values are analysed with care and notably compared to reported values of countries with similar cultural and structural characteristics.

Comments

The methodology paper on the EAPEP (ILO 2013a) includes a very useful section on strengths and weaknesses of the methodology. Regarding the imputation method, the two main limitations that are mentioned are:

- The linear interpolation and the (weighted or not) panel regressions are not based on the same assumptions. The linear interpolation assumes that the changes in LFPR due to cyclical and accidental factors are negligible compared to structural ones. The inclusion of linearly interpolated values in the sample used to estimate the panel regressions introduces linearity artificially in the model.

- For some regions, there are not enough data (e.g. Sub-Saharan Africa), making the estimates far less robust. The parameters of the panel models are run on the sample of historical data plus the interpolated data which might not be the ultimate solution to the problem of scarce reported series. In addition, the GDP per capita is too volatile for a few economies that rely strongly on oil and/or other mineral commodities. In this context, it becomes a poor proxy for what it is meant to capture (wealth per inhabitant, existence of social security schemes, access to education, etc.).

An additional problem not mentioned in the methodology paper on the EAPEP lies in the use of the logistic transformation and its inverse. As mentioned in section I.4.d, this approach results in biased estimates of the conditional 'mean' of the dependent variable.

2. ILO Global Employment Trends 2013 (ILO 2013b).

The objective of the Global Employment Trends (GET) series is to provide each year the latest global and regional information and projections on several indicators of the labour market, such as employment, unemployment, employment-to-population ratio, and unemployment rate, broken down by sex and two age groups (youth 15-24 and adults 25+) when available. It also presents policy considerations in light of the new challenges facing policy makers in the coming year.

Each issue of Global Employment Trends also contains a short term labour market outlook based on projections or scenarios, focusing on unemployment, vulnerable employment and working poverty.

The reports have been published on a yearly basis since 2003, with special editions to analyze labour market trends for segments of the population such as youth and women, or for certain regions.

Imputations at the country level are not published in the report. The GET model (actually a methodology) is used to produce the regional estimates. The methodology is presented in detail in ILO (2010b). The report is released in the beginning of each year (eg. in January 2012 for the 2012 edition) and the full set of data for the previous year (eg. 2011) are not available at that time of the year. However, for 60 countries that publish consistent and timely infra-annual data (monthly or quarterly), the imputation method for the current year is different and uses infra-annual data for the previous year (eg. 2011) and imputes the missing quarters or months using an extrapolative model.

The estimates of economically active population described before are used as an input to produce estimates of employment and unemployment. Also, the methodology used in the GET is very similar to the one used for the EAPEP. Therefore in what follows, only the main differences are highlighted.

Prior to the imputation process, data are harmonised for several countries in terms of geographical coverage and age limits. For example, unemployment rates are adjusted in terms of geographical coverage (some countries only report urban data) and age limits (eg. some countries report 16+ instead of 15+).

There are 178 countries/territories included in the Global Employment Trends (GET) model. In the 2012 GET report, out the 178 countries, there was at least one (reported and selected) data point for 150 countries for the total unemployment rate. This means that time series were fully imputed for 28 countries. Out the 150 countries with statistics, data posterior to 2004 were available (or selected) for 133 countries data posterior to 1999 were available (or selected) for 140 countries.

Imputations methods

There are two phases: **country-level imputations** and **imputations derived from a panel data model**. The later allow notably to generate imputed values for countries with no data at all.

In the **GET model**, there two different types of country-level adjustments:

(i) Imputing missing unemployment sub-components (male, female, youth, adult), when total unemployment is reported (missing variables in sections)

(ii) Imputing missing national and sub-component unemployment rates when these are available for some years only (missing periods).

These two types of adjustments are undertaken twice (on the basis of survey-based data and then on the basis on survey-based plus imputed data). In order to ensure consistency between total unemployment and its different sub-components, the age or sex specific rates are adjusted to match the totals, as the most available information is the total unemployment rate (both sexes). In other words, the methodology goes from general to specific (total to sex-age).

The first adjustment "Imputing missing unemployment sub-components, when total unemployment is reported" concerns many countries, which report total unemployment rates, but do not provide the data disaggregated by sex and by age group. As highlighted in ILO (2010b p. 18), for the 1991-2008 period "*reporting rates for the total unemployment rates are almost 5 times as high as subcomponent reporting rates in the Middle East, almost 3 times as high in North Africa, and approximately twice as high in South Asia and Sub-Saharan Africa*".

Simple imputation methods are used at this stage, which essentially consist of using the relationship between the total unemployment rate and the unemployment rates for the sub-components from years with complete observations **for the same country** (observations on both unemployment rates and sub-components) to fill the gaps for years when sub-components are not reported. It is worth mentioning that the female unemployment rate is calculated as a residual, after analyzing the relationship between the total unemployment rate and the male UR. This ensures consistency between total unemployment and its sub-components.

The second type of adjustment (ii) consists in filling the gaps in a time series, like in the EAPEP model. The previous procedure used to be a simple linear interpolation. It has been revised recently in order to take into account to some extent the cyclical factors.

Therefore, unemployment rate over time is decomposed into a structural component (e.g. sectoral composition of employment) and a cyclical component (related to the economic growth). The cyclical component is based on a simple regression of the unemployment rate and changes in GDP. Note that at least six years of reported unemployment rates are needed in order to run the regression at the country level. For countries with insufficient data points, a regional elasticity (unemployment rate to GDP) is used in the calculation. It is defined as the median of the country-elasticities within each region.

Two imputed values are obtained, each accounting for one of the two components. The final imputed value is then produced as a weighted average of these two values. The value of the weight depends on the quality of the regression of unemployment rate and changes in GDP. For example, for countries where the growth-employment linkages are weaker, such as oil or mineral export-dependent countries, a large discrepancy between the structural and cyclical component is often observed and the relation between GDP growth and the unemployment rate is weak. In such a case, the cyclical component is not accounted for in the final imputed value.

In the 2012 GET report, when considering only the last available year, the corresponding missing unemployment sub-components need to be imputed at the country level as follows:

Country-level imputations of TOTAL UR, by SEX: 24 countries

Country-level imputations of TOTAL YOUTH UR: 47 countries

Country-level imputations of UR by SEX YOUTH: 54 countries

Country-level imputations of UR by TOTAL ADULT: 52 countries

Country-level imputations of UR by SEX ADULT: 57 countries

For the imputations based on **econometric modelling**, the EAP and GET models are based on the same principles. For each of the 8 sub-regions a panel regression is run in order to estimate the missing data along both the country and time dimension for all sub-components. There are 4 sub-components for the unemployment rate (adult and youth, by sex). Importantly, the set of observations includes both survey-based data as well as imputations obtained at the country level.

However, the set of covariates (χ_{it}) used is different in the GET model. In the GET model, the main covariate is the real GDP growth rate, combined with time dummies. Also, the specification differs across sub- regions as some explanatory variables are not significant.

Also note that the GET report includes regional estimates of employment shares broken down by three broad sectors (agriculture, industry and services) and sex. The methodology paper ILO (2010b) does not however describe the underlying model and especially the covariates that are used in the imputation model. Apparently, the model is based on world urbanisation prospects data (UN estimates) in addition to GDP per capita and changes in real GDP.

Evaluation of imputation methods

The main evaluation information provided for the GET methodology are indicators of the performance of the applied method. For all regressions and logit regression, overall measures of quality of fit (R^2 and Pseudo R^2) are provided as well as regression coefficients and their statistical significance.

The plausibility of imputed values is also analysed on the basis of outlier detection and sensitivity analysis.

The ILO (2010b) paper does not present any indicator of accuracy of the imputed values (based on simulated missing values) for either of the two methods (country level and panel data). So the accuracy of the imputed values remains to be explored, especially for regions with little data.

However, a recent document (ILO 2013e) has assessed the post mortem accuracy of forecasts (one year to three years ahead) of unemployment rates generated by the GET model for four reports (2001 to 2013). Even if this analysis does not assess imputation errors (that concern past data), the results are very useful as they shed light on the relevance of the GET model. Actually the GET methodology used for the projection is the same as the panel data, but the GDP growth is based on IMF projections. So the forecasting errors are subject to errors due the model but also to forecasting errors on the projected explanatory variable (GDP).

The results in this post-mortem analysis suggest that, on average across all countries for which real data are available, the GET unemployment rate forecasts are slightly biased; that is, the GET model over-predicts one to two years ahead and under-predicts three to four years ahead. However, this bias is not statistically significant for one to three years ahead.

In general, the tests for accuracy show that the shorter the prediction period, the more accurate the GET forecasts, as indicated by smaller forecast errors for shorter prediction periods and larger forecast errors for longer periods. Table 2 reports some results for the global[12] unemployment rate forecasts. For example, the one year ahead forecast is subjected to a mean absolute error of less than 0.5 percentage point, which is a very good result.

Table 2. Selected accuracy statistics for the global unemployment rate forecasts

Year(s) ahead	**MAE** (Mean Absolute Error)	**RMSE** (Root Mean Squared Error)

[12] More exactly the sample of countries worlwide for which real data are available.

World	1	0.45	0.72
	2	0.94	1.27
	3	1.36	1.98
	4	1.91	3.02

Source: ILO (2013e)

This analysis represents a step forward in the evaluation of the GET methodology. It is however limited to the total unemployment rate. It would be interesting to generalise this type of analysis for the other indicators, notably the subcomponents of UR, for which there is a much higher proportion of imputed values.

Finally, as mentioned in ILO (2010b), "*The uncertainty associated with the estimates generated by the models – attributable to the imputation process, and to uncertainty surrounding benchmark data – is always acknowledged in the analysis based on these estimates. A point estimate is not provided when the level of uncertainty associated with it is very high. Whenever possible, a confidence interval (e.g. for the short-to-medium term unemployment projections) that accounts for the impact of imputations is constructed and presented as a measure of uncertainty*."

Comments

Most of the strengths and limitations regarding the EAP imputations are applicable to the GET model, as the methodology is very similar. The ILO (2010b) methodology paper does not include a specific section on strengths and limitations but discusses the pros and cons in its conclusion:

"There is no doubt that intensive data collection efforts to further expand data coverage and increase the frequency of data collection is necessary. The ILO has a crucial role to play in supporting countries to collect, analyze and disseminate timely labour market information.

In the short run, the ILO will continue to refine and enhance the Trends econometric models. Future work in this regard must include additional sensitivity analysis and testing. Specifically, new equations and model specifications need to be developed and evaluated. For instance, variables representing structural factors that may affect the relationship between growth and employment (e.g. natural resource exports dependence) can be explicitly controlled for in the regressions.

Additional work is also required to further improve the models' flexibility and responsiveness to economic and social shocks resulting in breaks in data series. This could involve the inclusion of variables that capture countries' vulnerability to external shocks, for instance, macroeconomic stability, financial sector development and integration into the global system, dependence on exports (exports as a share of GDP, or exports relative to domestic consumption), dependence on remittances, dependence on foreign aid, among others."

An additional problem not mentioned in the methodology paper on the GET lies in the use of the logistic transformation and its inverse. As mentioned in section I.4.d, this approach results in biased estimates of the conditional 'mean' of the dependent variable.

Finally, it is worth mentioning that in the different background papers (ILO 2010b and ILO 2013b), the details of the various models (specifications, and measure of performance R^2) used to derive regional estimates of employment by occupational status and broad activity sector (agriculture, industry and services) are not provided.

3. ILO Global Employment Trends 2013 –Working Poverty (ILO 2012a)

The GET report also includes regional estimates of working poor[13] (not broken down by age nor sex). The **Working Poverty Model** is one module of the GET models. The methodology has been revised completely recently (see ILO 2012a for a detailed description). The principles of both published data and estimates and the methodology are quite different than in the EAP or GET models, as there are no country level imputations.

In order to compute regional and global aggregates, complete time series of working poverty estimates for 141 countries are generated (but not published). The regional aggregates are published for 8 regions and for the world (based on the sample of 141 countries).

The input file contains 58 national estimates of working poverty (15+) for a total of 53 countries. Overall, an estimate is available for at least one point in time for 78.4 per cent of the developing world's labour force, including highly populated countries such as Brazil, China, India and Indonesia. The years for which national working poverty estimates are available range from 1995 to 2007. In terms of data coverage over time, of the 58 datasets with national estimates, nearly two thirds correspond to the years between 2002 and 2005. The majority of datasets are based from household income and expenditure surveys (HIES) and living standards surveys (LSS), both of which provide details of income and expenditure information together with labour market status.

Imputations

The revised methodology consists of two main steps:

Step 1: Estimate working poverty rates (15+) for countries and years for which no national working poverty estimate is available but for which total poverty rate estimates (0+) are available, using the World Bank database *PovcalNet*.

Step 2: Impute working poverty estimates (15+) for the remaining countries and years, producing a complete time-series of national working poverty estimates.

For Step 1, two regression models are used (for the two thresholds: US$1.25 and the US$2 a day). The (logistic transformed) dependent variable is the Ratio of the working poverty rate (aged 15 years and above) to the total poverty rate (aged 0 years and above). There are 58 records with data on national working poverty estimates (15+) and 422 records on total poverty estimates (0+). Consequently there are 58 computed ratios and model imputations are available for 364 records.

The explanatory variables are:

- employment-to-population ratio (EPR)
- ratio of the working-age (15 years and over) population to the total population (WAP)
- log of labour productivity, measured as output per worker (LP)
- regional dummy variables for five regional groups.

The employment-to-population ratio represents the proportion of a country's working-age population that is employed. The numerator is total employment aged 15 years and over derived from the GET model, while the denominator is the working-age population aged 15 years and over (source: United Nations World Population Prospects, 2010 revision). Labour productivity is represented by output per worker. This is calculated by dividing the annual

[13] The indicators are the shares of employed persons living in a household with an income less than US$1.25 and US$2 (at purchasing power parity) per person, per day. These indicators are also referred to as "International poverty lines".

PPP-adjusted Gross Domestic Product of a country as provided by The World Bank's World Development Indicators database by the total number of employed persons (aged 15 years and over) for the corresponding country and year (source: GET Model).

For Step 2, two regression models are estimated (for the two thresholds: US$1.25 and the US$2 a day). The (logistic transformed) dependent variable is the number of working poor divided by the working-age population (aged 15 years and above). The explanatory variables are:

- the share of agricultural employment in total employment,
- the share of the population aged 25 to 54,
- log labour productivity
- regional dummy variables (which are interacted with the labour productivity) to capture important regional differences in coefficients)
- country dummy variables (country specific fixed effects)

The numbers in the agricultural sector and total employment are both taken from the ILO GET model.

Evaluation of imputation methods

The main evaluation information provided for the working poverty model imputations are indicators of the performance of the applied method. For all regressions (R^2) are provided as well as regression coefficients and their statistical significance. All regressions (from both steps) have substantial explanatory power, with an R^2 of 0.84 and 0.90, respectively, for the US$1.25 and US$2 regressions. However, these measures should be interpreted with care, as the explanatory variables contain a large share of imputations.

The plausibility of imputed values is also analysed on the basis of outlier detection and sensitivity analysis.

Additional accuracy measures of imputed values were computed based the 58 real data points using leaving-one-out cross-validation (simulations). The methodology and main results are presented here (see ILO 2012a, pp. 15-16).

"For both the US$1.25 and US$2 models, in the majority of the cases the values predicted by the model are very close to the actual country-reported values. For the US$1.25 results, for 40 (69 per cent) of the observations, the absolute value of the difference between the actual and predicted rate is less than 3 percentage points, with 21 observations showing a difference of less than 1 percentage point. There are some notable outliers, including Congo (Dem. Rep. of), East Timor, Liberia, Mali, Nepal and South Africa, for which the absolute differences are 6 percentage points or higher."

...

Out of the 58 calculated differences between actual and predicted values, there is an exactly equal split, with predicted values exceeding real values in 29 observations and falling below real values in the remaining 29 observations. The mean absolute error across the 58 observations is 2.5 percentage points for the US$1.25 poverty rate (compared with a simple average US$1.25 poverty rate of 31.3 per cent across the 58 observations)."

Comments

The Working Poverty models use a large share of imputed values, derived from other modules of the GET models. This is the main drawback of that methodology.

On the other hand, the paper provides accuracy measures of imputed values based the 58 real data points using on leaving-one-out cross-validation (and running 58 regressions) and these results indicate reasonable prediction errors (less than 3% percentage points) for around 40% of countries for the 1.25$ model and 69% of countries for the 2$ model.

An additional problem not mentioned in the methodology paper on Working Poverty model lies in the use of the logistic transformation and its inverse. As mentioned in section I.4.d, this approach results in biased estimates of the conditional 'mean' of the dependent variable.

4. World of Work Report (ILO 2013d)

The world of work report 2013 analyses the global employment situation five years after the start of the global financial crisis. It looks at labour market performance and projections both at the global and regional levels.

The options regarding missing data are twofold. Firstly, some the GET estimates and projections, such as the unemployment are used at the global level.

Secondly, for working-age population, total employment and employment rate, covering the 2006–18 period (projections starting in 2013), the sample consists of 65 countries that report quarterly data. So the approach is not to impute and work with a sample of reported countries, producing results that cannot be generalized beyond the countries included in the sample.

5. Global estimates of Child Labour (ILO 2013f and 2010a).

As part of its effort to increase the knowledge base on global child labour developments, the ILO has produced global estimates on child labour every four years since 2000. The 2013 report (see ILO 2013f) is the fourth issue of the ILO's report series: Global Estimates on Child Labour. The 2013 report provides new global and regional estimates on child labour for the year 2012 and compares them with the previous estimates for 2002, 2004 and 2008.

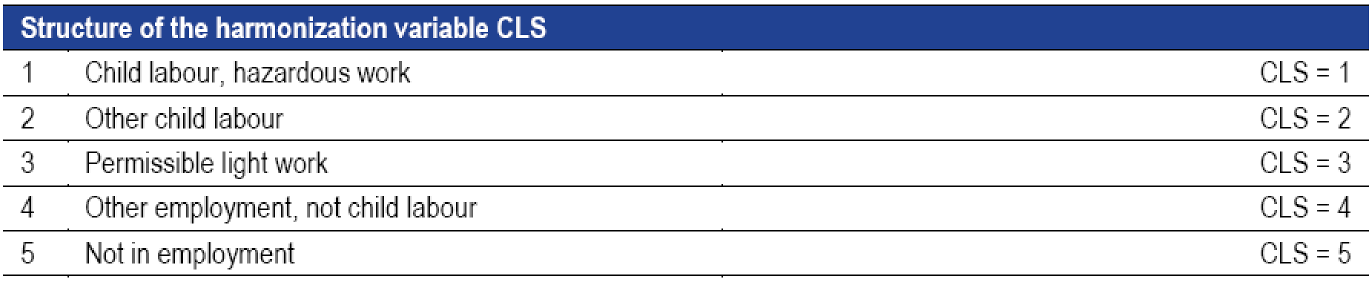
The new child labour estimates are based on refined estimation techniques fully comparable with the ones for 2000, 2004 and 2008 rounds. They also benefited from the international standards on child labour statistics adopted by the 18th International Conference of Labour Statisticians (ICLS) in 2008.

The indicators include children in employment, child labour and hazardous child labour by sex, age group and region. The new estimates are based on child labour data from 75 national household surveys, covering the period from 2008 to 2012. In all, some 75 datasets from 53 countries have been compiled for the 2008-2012 round of the ILO Global estimation of child labour.

The ILO surveys on child labour (SIMPOC) provide the main source of data for the present study. Other data sources are the UNICEF's Multiple Indicator Cluster Surveys (MICS), certain national labour force surveys and other relevant household surveys.

Data harmonisation is undertaken as available national datasets differ from each other with respect to a number of critical elements. These are differences in age groups, types of questions and response categories used in the survey questionnaire. For these reasons, the national datasets need to be harmonized with respect to the key elements before being processed further for global estimation.

The first step in the harmonization process is the construction of a single variable called Child Labour Status (CLS). The variable is composed of five mutually exclusive and exhaustive categories into which each child must be categorized. The five Child Labour Status are the following:



Structure of the harmonization variable CLS		
1	Child labour, hazardous work	CLS = 1
2	Other child labour	CLS = 2
3	Permissible light work	CLS = 3
4	Other employment, not child labour	CLS = 4
5	Not in employment	CLS = 5

Data harmonisation is done on the basis of simple proportional adjustments and logistic regression models.

Methodology used to handle missing data

Imputations are undertaken when missing values are present in the original dataset.

Data on variables used to determine the child labour status (CLS) are missing in some cases from national survey data. The missing data were imputed with the average of the observations in countries of the same region. Then the regional average is combined with the information available from the country to compute the child labour status.

The regional and global estimates of child labour are derived by extrapolations of national data using a composite estimation method. It consists of calculating two initial estimates of child labour in each region and computing a average of the two estimates. One of the estimates uses the **full sample** of countries (75) in each region and the other only the **matched sample** (i.e., those countries which were also part of the sample of the 2008 round of global estimation and those for which child labour data exist for more than one year over the 2008-2012 period).

The rationale of using such a composite method of estimation is "*to improve the accuracy of the estimates both in terms of levels and trends. The full sample estimates have maximum coverage as they include all countries in the sample, while the matched sample has minimum variability as they cover the same countries in their sample and therefore avoid the variability of sample differences. In this sense, the composite estimate is optimum as it uses maximum information with minimum variability*."

The **full-sample estimation** is based on a sampling or **weighting procedure**. There are no imputations. It consists of extrapolating the full sample of harmonized national datasets to regional and global values by weighting each country according to its relative share of children among the total in the region (based on the 2012 revision of UN population estimates and projections). This weighting procedure presumes that the data are missing at random at the country level and there is no non-response bias.

Matched sample estimation involves three main steps:

(a) Standardization of the national survey years to the reference years 2008 and 2012;

(b) Estimation of trends from 2008 to 2012;

(c) Derivation of the matched-sample estimates for 2012.

The main principle of the process implies the use of linear regression fitted over a time trend (after a log-ratio transformation of the dependent variable) estimated for each child labour status across the matched sample. The predicted value from the regression are imputed for each country with missing values in 2008 or 2012 that are included in the matched sample.

The "matched sample estimate of **change** 2008-2012" is estimated using a weighting average of country changes between 2008 and 2012.

Finally, the matched sample estimation for 2012 is derived from a simple addition:

"Matched sample estimate of child labour 2012"
= Estimate in 2008 + "Matched sample estimate of change 2008-2012"

In the final step, the composite estimation calculates a weighted average of the full-sample and matched sample estimates, with weights derived such that they minimize the variance of the final composite estimate on the assumption that the full sample and the matched sample represent both random samples of countries in their regions.

Evaluation of global estimates

The global estimates of child labour are evaluated in terms of their standard errors, calculated as the sampling error variability, assuming that the datasets used for estimating the child labour categories have themselves negligible variability relative to the variability due to differences that would occur had the sample included different countries than the ones used here. As highlighted in the paper, "*The calculation also assumes that the countries covered in the study form a random sample of the countries in the world. Although both of these assumptions are not fully satisfied, the results may still be indicative of the margin of error of the estimates that can be attributed to the selection variability of the countries in the sample.*"

Compared with the results of 2008, the standard errors in 2012 have greatly decreased indicating improvement in the precision of the estimates.

Also the plausibility of the global estimates is undertaken on the basis of four types of comparisons:

(a) comparison with aggregated national survey-based data;
(b) comparison of full-sample and matched sample estimates;
(c) comparison with ILO Labour force estimates and projections for 15-19 year olds;
(d) comparison with UNICEF Child labour estimates for 5-14 year olds published in 2013.

Comments

The methodology uses a variety of techniques, including imputations and weighting procedures and presents several measures of evaluation.

The main concern of the methodology is that it presumes that the data are missing at random at the country level and there is no non-response bias at the regional level. This assumption is both used for the imputations derived from regional averages and for the weighting procedure.

This limitation is acknowledged in the paper (see quote above).

6. Global Wage Report 2012/2013 (ILO 2013c) and ILO (2011)

The report is issued every two years. The last edition (2012/2013) included estimates of real wage growth by region from 2000 to 2011.

For regions with low statistical coverage, figures are published as "Provisional estimates" (based on coverage of 75%) or "Tentative estimates" (based on coverage of 40% to 74%).

The methodology to estimate global and regional wage trends was developed by the ILO's Conditions of Work and Employment Programme (TRAVAIL) in collaboration with the Department of Statistics, following proposals formulated by an ILO consultant and three peer reviews made by four independent experts.

The objective was to collect wage data for a total of 177 countries and territories. Data were finally available for 115 countries and territories. In addition, in some countries with data, the statistical series were incomplete, in the sense that some years were missing ("missing periods").

Methodology used to handle missing data

The methodology includes many steps and is based on several methods.

First, to fill time series gaps (missing periods), a model-based framework is used to impute missing values. This is done in order to hold the set of responding countries constant over time and so avoid the undesired effects associated with an unstable sample.

Depending on the nature of the missing data points, five complementary approaches are used, which are described in the paper and presented in order of preference.

(1) Context: infra-annual (eg. quarterly) are available but not for the whole year. The missing(s) quarters are extrapolated using a combination of a moving average and a linear trend.

(2) Context: a time-series has a short gap between existing data points (gaps of a maximum of three successive years in the time series). The missing years are interpolated using a logarithmic growth function (which forms part of the family of growth curves, such as the linear trend). This approach was used for 17 countries in the 2010/2011 edition.

(3) Context: an alternative (second best) source of wage data for a given country is given for the missing year (e.g. one based on establishment surveys). This additional information is used to fill missing data points in the preferred (first best) time series (variable *a*), by using the growth rate of the second best source (variable b). Formally:

$$\hat{a}_{(t+1)} = a_{(t)} \,.\, b_{(t+1)} / b_{(t)}$$

Note that this imputation can be viewed as an issue of data harmonization. This approach was used for 37 countries in the 2010/2011 edition.

(4) Context: no secondary data source exists and the gap in the series is too long to use the simple interpolation described in (2). In this context, a more complex imputation formula is used at the country-level. It is based on the assumption (economic theory) that – in the long run – wages respond to changes in labour productivity.

This approach was used for 2 countries in the 2010/2011 edition.

(5) Context: none of the four methods described above are feasible. An econometric model is then used to estimate the remaining missing data points. The model is again based on the assumption (drawn on standard economic theory) that suggests that wages respond to

changes in labour productivity. In line with this reasoning, regional elasticities between productivity growth and real wage growth are estimated and used for extrapolation purposes.

The regional elasticities are estimated for each region by regressing wage growth on productivity growth for each region. A weighted regression is actually used, in order to account for the varying sizes of different countries within a region (each observation is weighted by the share of the country in total paid employment in the corresponding region). The robustness of each of the regional estimations was analysed, outliers were excluded and the specifications re-estimated.

This approach was used for 52 countries in the 2010/2011 edition.

The next step (after these five types of imputations) consists in deriving regional and global estimates (defined as a population of 177 countries) on the basis of survey-based data and imputed values for 115 countries.

This is done using a weighting procedure (no imputations are made for non-responding countries) and on the estimation of non-response weights.

Because the missing data mechanism is supposed as MNAR (non-responding countries may have wage characteristics that differ from those of responding countries), a standard approach to reduce the adverse effect of non-response is used. It consists in calculating the propensity of response of different countries and then weighting the data from responding countries by the inverse of their response propensity.

The response propensity was estimated by using an econometric model relating the response or non-response of a given country to its number of employees and its labour productivity (or GDP per person employed in 2005 PPP$). This is based on the observation that wage statistics are more readily available for richer and larger countries than for poorer and smaller countries.

A logistic regression with regional fixed effects is used. Unlike in ILO (2013a) and ILO (2013b), that use panel data, the weights are estimated here on the basis of cross-sectional data for 2008. The logistic regression has 177 observations and produced a pseudo [14]R² of 0.462.

The weights obtained from the logistic regression are then adjusted. This adjustment is called calibration. Calibration adjusts response weights for differences in non-response between regions, ensuring appropriate representation of the different regions in the final global estimate. In the present context, a single variable, number of employees in 2008, was considered for calibration. The calibration factor for each country is calculated simply as the ratio between the **known** number of employees in the region the country belongs to and the **estimated** total number of employees based on the weights obtained from the logistic regression of the same region.

Note that the calibrated response weights are equal to 1 in the regions where wage data are available for all countries (Advanced countries, Central and Eastern Europe, Eastern Europe and Central Asia). They are larger than 1 for small countries and countries with lower labour productivity since these are underrepresented among responding countries.

Finally the regional and global estimates of wage trends are calculated through a complex procedure that includes several steps (see reference ILO 2011 for more details). The need for

[14] When analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squareds have been developed. These are "pseudo" R-squareds because they look like R-squared in the sense that they are on a similar scale, ranging from 0 to 1 (though some pseudo R-squareds never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS R-squared and different pseudo R-squareds can arrive at very different values.

a more complex procedure is due to several reasons. A simple weighted average is actually not applicable for wages, as wages are not variables like employed or unemployed expressed in similar units. Wages are expressed in different monetary units and are also influenced by composition effects that occur when the share of employees shifts between countries. For instance, if the number of paid employees falls in a large high-wage country but expands in a large low-wage country, this would result in a fall of the global average wage (when wage levels stay constant in all countries).

Evaluation of global estimates

The main information provided are indicators of the performance of the applied method (Pseudo R^2 of the logistic regression).

The plausibility of imputed values is also analysed on the basis of outlier detection and sensitivity analysis.

The results are also evaluated in terms of statistical coverage. Regional growth rates are flagged as "provisional estimates" when they are based on coverage of ca. 75 per cent and as "tentative estimates" when the underlying coverage of the database is between 40 and 60 per cent to draw attention to the fact that they might be revised once more data become available. For some regions with too few real observations, the results are not published as data are judged as insufficient to make a reliable estimate. For example, in the 2010-2011 report, the Middle East had data coverage of roughly 13 per cent for 2008 and 2009.

Comments

The methodology includes a wide range of techniques and many steps. It has been validated by different experts.

There is transparency in the methodology as well as transparency regarding the limitations of estimates for some regions (provisional estimates).

What would be very useful in the next editions is to include measures of accuracy of imputed values based on cross-validation simulations (on the sample of countries with data) such as in ILO 2012a.

7. ILO Panorama Laboral 2012. América Latina y el Caribe[15]

The ILO Regional Office for Latin America and the Caribbean publishes every year in Spanish and English an overview of the labour market in Latin America and the Caribbean (see ILO 2012b). The report includes a review on the international outlook, the economic growth in Latin America and the Caribbean in 2012 as well as regional projections of GDP and employment for 2012 and 2013. The performance of the labour market is analysed in details as well. The report also includes data and analyses on employment by status in employment and economic activity, employment and social security and real wages.

Data from national sources are used and most countries report data. The tables include a lot of metadata regarding the specificity of national data (regional coverage, period, break in series, ...). There are no imputations of missing data. Data are actually missing only for a few small countries from the Caribbean.

For most indicators, regional aggregates are presented as sets of selected countries reporting data. For example for employment to-population ratio and labour force participation rate, an aggregate for "Latin America (9 Countries)" is presented in the charts and tables and the list of the 9 selected countries in displayed in the note.

For the unemployment rate (UR), an aggregate for the Latin America and the Caribbean is estimated. It does not cover a few Caribbean countries that do no report data. Also UR are harmonized in terms of concepts for a few countries such as Colombia, Ecuador and Panama that include hidden unemployment in labour force and unemployment.

[15] Also available in English, 2012 Labour Overview. Latin America and the Caribbean. See ILO (2012b).

II. Harmonisation techniques used in cross-country datasets

Harmonisation is the process of comparing two or more data component definitions (eg. different concepts of unemployment) and identifying commonalities among them that warrant their being combined, or harmonized, into a single data component.

Harmonisation of data can be seen as a sub-problem of imputation where the missing data point is accompanied by a data point based on a different definition, concept, geographical coverage or period of reference than the one used for the variable of interest. In other words, some alternative (or proxy) data point is available. So, there is no need to "create" a new figure from scratch.

It is also worth mentioning that in the various ILO reports reviewed in the previous chapter, there is sometimes some confusion between data harmonisation and imputations (eg. some imputations of missing periods are presented as harmonisation).

The point of departure when harmonising data is therefore to assess (judgmentally) by various experts how far is the available (non-harmonised) variable from its theoretical (harmonised) value. This assessment can be made on the basis of data for countries (and periods of time) for which both variables have been observed (harmonised and non-harmonised).

If the gap between the (non-harmonised) data from that of the theoretical (harmonised) variable is judged as small, simple adjustments methods are usually used. These include judgmental adjustments, ad hoc adjustments (such as proportional adjustments) or location-based adjustments.

If data for the (non-harmonised) variable and its theoretical (harmonised) variable are judged to be not comparable at all, the harmonisation techniques may need to be more complex. For example, model-based imputations can be used (on cross-sectional or panel data), where the non-harmonised variable could be used as one explanatory variable (covariate) among other variables in order to predict imputed values.

The task also becomes more time-consuming when there are many different concepts for a same variable (eg. different age limits for the population or different definitions of rural/urban areas). This happens very often for cross-country datasets, as each country may not follow international standards and follow its own standard.

In the various ILO databases and reports reviewed in this paper, all the adjustments are simple adjustments (such as proportional adjustments) or location-based adjustments.

Here are a few **examples**, which would apply to labour market indicators broken down by rural and urban areas.

Harmonising age bands:

In the EAPEP database (ILO 2013a), for the countries with non-standardised age-groups, two types of harmonisation are applied; harmonising the lower and upper age limit and harmonising data from large age bands to the standard ILO 5-year age bands.

A common problem is to harmonise data for the lower age limit (15 years old and onwards being the standard). Countries for which the lower age limit adjustment is required include Iceland, Jamaica, Macau (China), Norway, Puerto Rico, Spain, Sweden, United Kingdom and United States. All the above countries have a lower age limit of 16 years of age, except Jamaica and Macau (China), for which the limit is 14 years of age.

The basic assumption is that the labour force participation rate (LFPR) of 15 year olds for the above-mentioned countries is assumed to be proportional to the labour force participation rate of 16-19 years-old.

For the US, estimates of the LFPR for 16-19 year olds are published for two sub-categories, 16-17 and 18-19. Since LFPRs within this age-group are positively related with age (the LFPR of 18-19 year olds is greater than 16-17 year olds), the assumption is that the same relationship holds for 15 year-olds to 16-17 year olds.

Therefore, the LFPR of the 15 year olds is estimated as using the proportional adjustment:

$$\widehat{LFPR^{US}_{s,y,15}} = \frac{LFPR^{US}_{s,y,16-17}}{LFPR^{US}_{s,y,16-19}} * LFPR^{US}_{s,y,16-17}$$

where s=male, female and y=1980, ..., 2009. For example in 2009, the LFPR of male of 15 year olds was estimated at 17.3%, resulting from:

$$\widehat{LFPR^{US}_{male,2009,15}} = \frac{25.53}{37.63} * 25.53 = 17.3$$

Then, the ratio of the estimated LFPR of the 15 year olds to the LFPR of the 16 to 19 year olds is calculated:

$$USratio_{s,y} = \frac{\widehat{LFPR^{US}_{s,y,15}}}{LFPR^{US}_{s,y,16-19}}$$

In the example above, the ratio for males 15 year olds was estimated at 46% in 2009, resulting from: 0.46 = 17.3 / 37.63

For all the countries, the estimated LFPR for the 15 years-old is calculated by applying the US ratio to the country LFPR of the 16-19 year olds:

$$\widehat{LFPR_{s,y,15}} = USratio_{s,y,} * LFPR_{s,y,16-19}$$

Hence, the estimated economically active persons aged 15 years is:

$$\widehat{LF_{s,y,15}} = \widehat{POP_{s,y,15}} * \widehat{LFPR_{s,y,15}}/100$$

The population aged 15 is derived from population data from the United Nations World Population Prospects' (UN) 15-19 population and the National Statistical Office's (NSO) 16-19 population.

$$\widehat{POP_{s,y,15}} = POP^{UN}_{s,y,15-19} - POP^{NSO}_{s,y,16-19}$$

Then, the above estimated economically active population aged 15 year olds is simply added to the reported economically active population aged 16 to 19:

$$\widehat{LF_{s,y,15-19}} = \widehat{LF_{s,y,15}} + LF^{NSO}_{s,y,16-19}$$

Finally, the estimated LFPR for the lower age group of 15 to 19 year olds is:

$$\widehat{LFPR_{s,y,15-19}} = \frac{\widehat{LF_{s,y,15-19}}}{POP^{UN}_{s,y,15-19}} * 100$$

The LFPR for both sexes is derived from estimates of the labour force and the population for males and females. An example is provided in Table 3 for the UK. The reported male LFPR (16-19) of 51.6% is adjusted downwards to 45.3%.

Table 3. Example of lower age adjustment: United Kingdom, 2009

			Country-reported data			**ILO Estimates**				
	POP (UN, 15-19)	**Ratio (US)**	**POP (NSO, 16-19)**	**LF (NSO, 16-19)**	**LFPR (NSO, 16-19)**	**EST POP 15**	**EST LFPR 15**	**EST LF 15**	**EST LF 15-19**	**EST LFPR 15-19**
Female	1’938	0.460	1’541	766	**49.7**	397	23	91	*857*	***44.2***
Male	2’051	0.421	1’615	834	**51.6**	436	22	95	*928*	***45.3***
Both Sexes	3’989								*1’785*	***44.7***

Note: The country reported data come from OECD, Labour Force Statistics online database.

Harmonising geographical coverage:

In the EAPEP database (ILO 2013a), for several countries, predominantly in Latin America, households (or labour force) surveys do not always cover the entire territory and are restricted to urban areas. Cases in point are Uruguay and Argentina. The surbey-based data are adjusted in order to cover the whole territory.

For those countries, the adjustments are undertaken with a high degree of accuracy provided that two conditions are met. First, household surveys covering the entire territory are available for at least one year and second, the urban areas cover a significant share (e.g., 30%) of the total population.

The principle is to apply a proportional adjustment and to estimate the national participation on the basis on participation rates for urban areas, on the basis of the following relation:

$$LFPR[total]_{a,s,t} = \varpi . LFPR[urban]a,s,t \quad + \ (1-\varpi). LFPR[rural]_{a,s,t}$$

where ω represents the ratio of urban-to-total population. Symbols *a*, *s* and *t* represent respectively the age group, the sex and the year. If for a given year *T*, both urban and rural data are available, then the ratio of urban to rural ratio can be computed:

$$\alpha_{a,s,T} = \frac{LFPR[rural]_{a,s,T}}{LFPR[urban]_{a,s,T}}$$

Usually, this ratio is superior to one. In other words, the LFPR is higher in rural areas than in urban ones. However, this ratio differs considerably across countries, due to differences in agrarian structures and land concentration.

The estimate is given by the following formula:

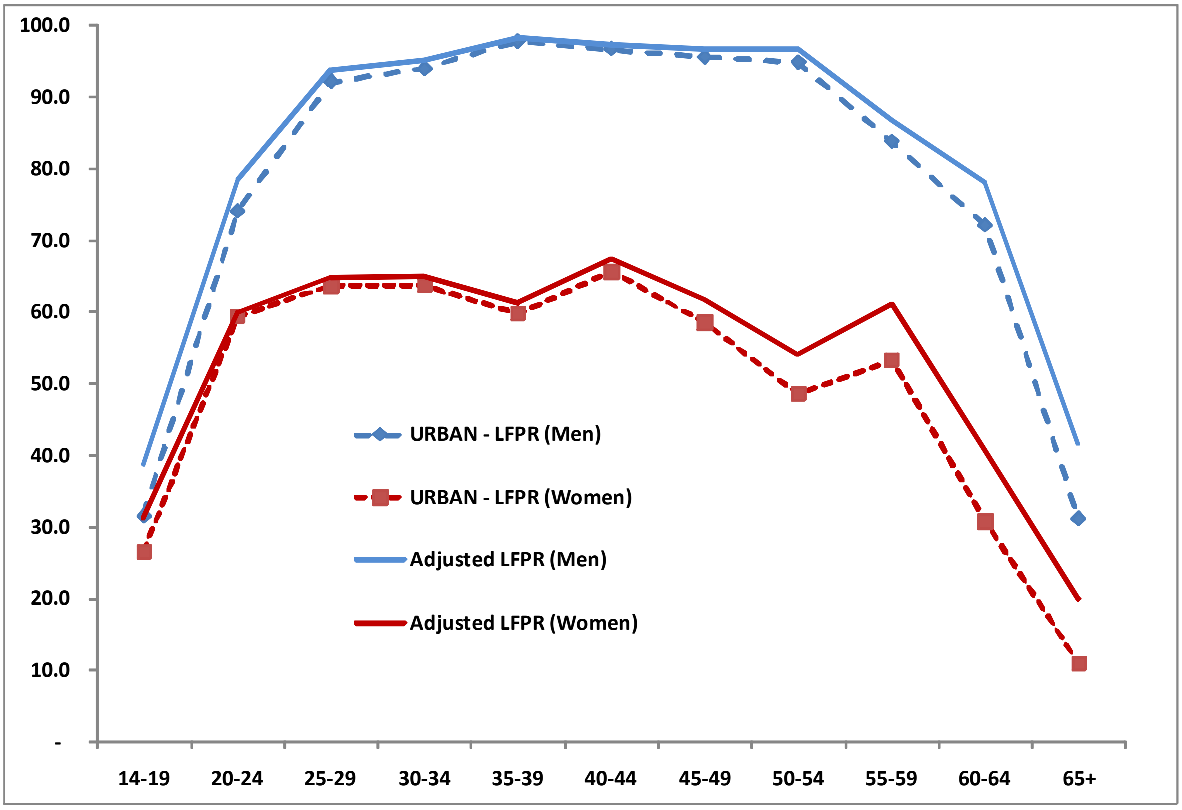
$$PR[total]'_{a,s,t} = PR[urban]_{a,s,t} * (\varpi + (1 - \varpi) * \alpha_{a,s,T})$$

The quality of this estimate depends on the value ω (ratio of urban-to-total population) and the volatility of $\alpha_{a,s,T}$ over time. For example, the value of ω is close to 84% for Uruguay, while it is closer to 30% for Peru (Metropolitan Lima), making the adjustment more accurate for Uruguay.

In the 2013 edition of the EAPEP, the adjustments have been made for Argentina, Bolivia, Peru and Uruguay, on the basis of rural and urban estimates at 5 year intervals (1990, 1995, ..., 2010) published by CELADE in 2006.

Figure 2 highlights the differences between reported urban LFPR and adjusted ones for Peru. The differences between the two curves are notable for both sexes.

Figure 2: Labour Force Participation rates in Peru (2005, LFS).



Note: Urban data covers Metropolitan Lima.

Harmonising data from various sources:

In the global wage report, data are harmonised when an alternative (second best) source of wage data for a given country is given for the missing year (e.g. from establishment surveys). This additional information is used to fill missing data points in the preferred (first best) time series (variable *a*), by using the growth rate of the second best source (variable *b*). Formally: $\hat{a}_{(t+1)} = a_{(t)} . b_{(t+1)} / b_{(t)}$

SUMMARY

The critical issues when choosing an imputation method or an alternative missing data technique for cross-country datasets have been analysed in this document.

These include the amount (**proportion**) of missing data in the total dataset. Easy techniques (judgmental adjustments, mean imputation) can be used when there are only few missing data points.

In setting up the methodology, it is essential to understand the underlying **missing data mechanism**. Random and non-random missing values can be distinguished. In case of data missing completely at random, the probability that *x* is missing doesn't depend on its value (x) or on value of other variables (*y*, *z*,…). Therefore, the missing data process is ignorable in imputation. For example, some survey questions asked of a simple random sample of original sample. In this context, easier imputations or missing data technique can be used such as mean or median imputation.

If values are missing in a non-random, systematic way (Missing Not at Random, MNAR), the distributions of the variable among complete and missing observations cannot be expected to be the same. This effect is also known as **selection bias**. For MNAR datasets the missing data mechanism may not be ignored. An imputation method not taking into account this additional information will typically be biased and over- or under-estimate the variable for the missing time period.

In the case of cross-country datasets, data are most often MNAR. In such cases, there is a selection problem related to unobservable differences in characteristics among reporting and non-reporting countries. Generally speaking, the proportion of missing data is much higher for small and poor countries.

The problem of MNAR missing data mechanism is addressed in various ILO reports by estimating the response probabilities (or weights) on the basis of logit regressions. For example, the explanatory variables used in the EAPEP and GET models include the following country-specific variables: economic growth, population size, per capita GDP and membership in the Heavily Indebted Poor Countries Initiative (HIPC).

The weights are then used in the next steps (weighted regressions or weighting sampling procedure) to correct for non response bias.

Model-based imputations are frequently used in ILO reports and databases. The principle is to use correlations between available explanatory variables or predictors (eg. per capita GDP, gross value added in agriculture) and variables with missing values (eg. rural employment rate) to estimate a model in order to predict the missing values, on the basis of economic theory.

Conceptually, this is a good way to impute values. It is good in the sense that a great deal of information from the individual (for micro-data) or country (for macro-data) is used to predict the missing values. In practice, the higher correlation between the predictors and the missing variable(s), the better the imputation will be.

For micro-data, variables involved in the survey design (eg. variables that define the different strata) can be used as explanatory variables in the model.

For macro-data, the exercise is more difficult because of smaller sample sizes and also it is often the same countries that lack data for all types of macro-economic indicators (see for example the different variables included in the World Development Indicators database maintained by the World Bank). Therefore, it is often the same variables available worldwide (GDP, per capita GDP adjusted for purchased power parity) that are used to compute missing values for various labour market indicators in cross-country datasets. These explanatory variables are then used as second best variables, as proxy of variables that would be determined on the basis of economic theory.

Another key element in the choice of explanatory variables is related to the **reliability of the explanatory variables** in cross-country regressions. Often, explanatory variables that include imputations for a significant number of countries are used to impute missing values of other variables. This aspect is often neglected as databases (including imputed values) published by international organizations are often perceived as real data (for all countries). One can easily see the danger or temptation of "filling holes in a Swiss cheese using whipped Cream", as stated in the title of the paper of Denk and Weber (2011).

Imputation model building is therefore a time-consuming task, as the quest for a good model implies to select useful explanatory variables but also check their source and reliability.

Another crucial aspect to take into account in model-based imputation is that many labour market variables are proportions or ratios (eg. labour market participation rates or unemployment rate), that vary between 0 and 1. When using ordinary least squares, the risk is that predicted values fall outside this range (eg. resulting in a negative unemployment rate). One approach adopted in several ILO models (notably ILO 2013a and IlO 2013b) is to apply a logistic transformation to the ratio variable, then to run the regression on the transformed variable and to compute the missing values on the basis of the inverse transformation applied to the fitted values of the regression. This process guarantees imputed values *y'* within the 0%-100% range. There is however a problem with this approach as it results in biased estimates of the conditional 'mean'. This is a known issue and can be addressed by implementing a non-linear estimate of the expectation (see section I.4.d for more detail).

An alternative to model-based imputations is to use weighting or **sampling techniques**. In this case, no imputations are done but regional averages are computed as weighted averages of available data. What is needed is to specify the weights adequately, on the basis of variables that make sense theoretically and taking into account the missing data mechanism.

In addition to using an appropriate imputation method and document the underlying assumptions, it is also crucial to include in the documentation indicators **assessing** the **quality of the imputation method**(s). There are different types of criteria: indicators of the performance of the applied method (such as the R^2 and the Pseudo R^2); indicators of accuracy of the imputed values (such as the average absolute error); the variability of statistics based on the imputed dataset (such as the sampling variability) and the plausibility of imputed values (based on editing techniques, of which automatic outlier detection).

Out of sample evaluation indicators of accuracy should be encouraged (i.e. cross validation ones) instead of in sample ones. However they are not frequently included in the various ILO reports. The main reason is that this process is time-consuming as indicators of accuracy are based on simulations. Since the true values of the imputed data are unknown, the imputed values cannot be compared to their true counterparts. Hence, accuracy indicators are estimated by treating available values as missing, imputing these artificially missing values, and comparing the imputed values to the ignored true values. This simulation technique of leaving out observations in an estimation procedure to validate estimation results is known as **cross-validation**. A good example of cross-validation can be found in the background paper on ILO global estimates of working poor (ILO 2012a). This exercise has allowed to better understand the magnitude of imputation errors and to identify the countries with very high errors (for which the model is not performing well).

Finally, as in any methodological paper, the background papers on the imputation techniques should explain clearly the methodology (with concrete examples) and should include a final section highlighting its strengths, weaknesses (limitations) and also directions for future work. It is also crucial to publish the regional and global estimates with appropriate metadata and mention possible **limitations** on the interpretation of estimates subject to very high incertitude. A very good practice has been adopted for the Global Wage Report, in which the results are also evaluated in terms of statistical coverage. Regional growth rates are flagged as “provisional estimates” when they are based on coverage of 75 per cent and as “tentative estimates” when the underlying coverage of the database is between 40 and 60 per cent to draw attention to the fact that they might be revised once more data become available. In addition, for some regions with too few survey-based data, the results are not published as data are judged insufficient to make a reliable estimate. For example, in the 2010-2011 report, the Middle East had data coverage of roughly 13 per cent for 2008 and 2009.

In Table 4, the different methodologies used in the ILO are summarised in terms of used imputation method(s) (or alternative missing data techniques) and key evaluation methods.

Table 4: Summary of the most common methodologies involving missing data techniques and evaluation methods used in the ILO for global/regional estimation

Methodology (variables)	**Missing data techniques**	**Key evaluation methods**
ILO EAPEP (Labour Force Participation Rates, by sex and age band, panel data)	Missing periods between two available data points: linear interpolation Other missing data (missing countries or missing periods at the end or beginning or the time series): weighted regional regressions using demographic and macroeconomic covariates and assuming data as MNAR.	In-sample measures (R^2, etc.) Out-of-sample measures for the linear interpolations
ILO GET (Unemployment rates, by sex and age group, panel data)	Missing subcomponents UR (missing variables in sections): proportional adjustments Missing periods between two available data points: country-level (if enough data points) or regional regressions , using GDP growth and time trend as covariates Other missing data (missing countries or missing periods at the end or beginning or the time series): weighted regional regressions using macroeconomic covariates and assuming data as MNAR.	In-sample measures (R^2, etc.) Out-of-sample measures for the 1 to 4 years ahead forecasts of total UR and for regions
ILO Working poverty (Proportion of working poor, unbalanced panel data)	No imputations at the country level. All imputations are model-based. The regressions are based on demographic, macroeconomic, labour market variables (including imputations for some countries) and regional dummies. There is no explicit assumption on the missing data mechanism.	In-sample measures (R^2, etc.) Out-of-sample measures using leaving-one-out cross-validation
World of Work Report (Several labour market indicators)	Complete records analysis (no imputations), based on the sample of countries that report quarterly data.	n.a.
Global estimates of Child Labour (children in employment, child labour and hazardous children work by sex, age)	Missing variables for countries: missing data are imputed with the average of the observations in countries of the same region. Then the regional average is combined with the information available from the country to compute the child labour status. The regional and global estimates are derived by extrapolations of national data based on sampling techniques. The data are presumed missing at random.	Standard errors, calculated as the sampling error variability. The plausibility of the global estimates is undertaken on the basis of several comparisons.
Global Wage Report (real wage growth)	In order to fill time series gaps (missing periods), five complementary approaches of imputation techniques are used, depending on the nature of the missing data points. The five techniques include notably the use of extrapolative methods, growth curves, proportional adjustments and weighted regressions (regressing wage growth on productivity growth). The regional estimates are derived using a weighting procedure (no imputations are made for non-reporting countries) and on the estimation of non-response weights (data are supposed to be MNAR).	In-sample measures (R^2, etc.) The results are evaluated in terms of statistical coverage and published as "provisional estimates" or "tentative estimates" when the regional coverage of the database is low.
ILO Panorama Laboral	Complete records analysis (no imputations), based on the sample of countries that report data. Note that UR are harmonized in terms of concepts for a few countries such that include hidden unemployment in labour force and unemployment.	n.a.

Note: n.a.: not applicable

References

Cameron A., Trivedi P.K. (2010). *Microeconometrics Using Stata, Revised Edition*. Stata Press.

Chambers R. (2000). Evaluation criteria for statistical editing an imputation. *National Statistics Methodological Series No. 28*. ONS, UK.

De Waal T., Pannekoek J., and Scholtus S. (2011). *Handbook of Statistical Data Editing and Imputation.* Wiley, New York.

Dempster and Rubin (1983), Introduction pp-3-10. in: *Incomplete Data in Sample survey (volume 2), Theory and Bibliography*. New York: Academic Press.

Denk M., Weber M. (2011). Avoid Filling Swiss Cheese with Whipped Cream: Imputation Techniques and Evaluation Procedures for Cross-Country Time Series. *IMF Working Paper WP/11/151*. http://www.imf.org/external/pubs/cat/longres.aspx?sk=25007.0

Dobson A.J. (Author), Adrian Barnett A. (2008). *An Introduction to Generalized Linear Models, Third Edition*. Chapman & Hall.

Durbin J., Koopman S. (2004). *Time Series Analysis by State Space Methods*. Oxford Univ. Press.

Efron B., Hastie T., Johnstone I. and R. Tibshirani (2004). "Least Angle Regression". *Annals of Statistics 32* (2): pp. 407–499.

Graham J. (2012). *Missing Data. Analysis and Design*. Springer.

Harvey A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge Univ. Press.

Hasler C. and Tillé Y. (2013). Balanced k-Nearest Neighbor Imputation. Paper presented at the *2013 World Statistics Congress*, Hong Kong. http://www.isi2013.hk/en/scientific_list_Aug_28.php

He Y., Yucel R., Raghunathan T.E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine Early View*, Wiley Online Library http://wileyonlinelibrary.com/

ILO (2010a). Global child labour developments: Measuring trends from 2004 to 2008. http://www.ilo.org/ipecinfo/product/viewProduct.do?productId=13313

ILO (2010b). Trends Econometric Models: A Review of Methodology. EMP/TRENDS Working paper, January 2010. http://www.ilo.org/empelm/pubs/WCMS_120382/lang--en/index.htm

ILO (2011). Global Wage Report 2010/11. Wage policies in times of crisis. http://www.ilo.org/global/publications/books/WCMS_145265/lang--en/index.htm

ILO (2012a). KILM 7the edition. Chapter 1.A: Working poverty in the world. Introducing new estimates using household survey data. http://kilm.ilo.org/2011/download/Chap1AEN.pdf

ILO (2012b). 2012 Labour Overview. Latin America and the Caribbean. ILO Regional Office for Latin America and the Caribbean. http://www.ilo.org/americas/publicaciones/WCMS_213162/lang--en/index.htm

ILO (2013a). ILO Estimates And Projections Of The Economically Active Population: 1990-2030 (2013 Edition). Methodological description. http://www.ilo.org/ilostat/content/conn/ILOSTATContentServer/path/Contribution%20Folders/statistics/web_pages/static_pages/EAPEP/EAPEP%20Methodological%20paper%202013.pdf

ILO (2013b). Global Employment Trends 2013. Recovering from a second jobs dip. http://www.ilo.org/global/research/global-reports/global-employment-trends/2013/lang--en/index.htm

ILO (2013c). Global Wage Report 2012/13. Wages and equitable growth. http://www.ilo.org/global/research/global-reports/global-wage-report/2012/lang--en/index.htm

ILO (2013d). World of Work Report. Repairing the economic and social fabric. http://www.ilo.org/global/research/global-reports/world-of-work/2013/WCMS_214476/lang--en/index.htm

ILO (2013e). A Post-mortem analysis on the Global Employment Trends' Unemployment Rate Forecasts (by Bourmpoula E. and Wieser C.). *Forthcoming*.

ILO (2013f). Global child labour trends 2008 to 2012. http://www.ilo.org/ipecinfo/product/viewProduct.do?productId=13313

Kim J.K. Wayne Fuller W. (2013). Hot Deck imputation for multivariate missing data. Paper presented at the *2013 World Statistics Congress*, Hong Kong. (http://www.isi2013.hk/en/scientific_list_Aug_28.php).

Little R. and Rubin D. (2002). *Statistical analysis with missing data*, 2nd ed. Wiley, New York.

Makridakis S., Wheelwright S. and Hyndman R. (1998). *Forecasting: Methods and Applications* (Wiley, 1998, 3rd edition).

McKnight P., McKnight K., Sidani S., Figueredo A. (2007). *Missing Data: A Gentle Introduction*. Guilford Press.

Mélard G. (2008*). Méthodes de prévision à court terme.* Paris, éditions ELLIPSES.

Pearl J. (2009). *Causality: Models, Reasoning, and Inference, 2nd edition*. Cambridge University Press, New York.

Raessler S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Springer, New York.

Rosenbaum P., Rubin D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1), 41-55.

Rubin D. (1987). *Multiple imputation for nonresponse in surveys*. Wiley, New York.

Shao J. and Sitter R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association 91 (435)*, 1278-1288.

Stata (2013). Stata Multiple-Imputation Reference Manual. Release 13. http://www.stata.com/manuals13/mi.pdf.

Tarsitano A., Falcone M. (2010). Missing-Values Adjustment For Mixed-Type Data. *Working Paper WP15-2010, Department of Economics and Statistics*, University of Calabria.

Wooldridge J.M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68(1), pp. 115 – 132.