ILO modelled estimates – methodological overview

The ILO has designed and actively maintains a series of econometric models, which are used to produce estimates of labour market indicators in the countries and years for which country-reported data are unavailable. The purpose of estimating labour market indicators for countries with missing data is to obtain a balanced panel dataset in order to compute, every year, regional and global aggregates with consistent country coverage. These allow the ILO to analyse global and regional estimates of key labour market indicators and related trends. Moreover, the resulting country-level data combining both reported and imputed observations offer a unique, internationally comparable dataset of labour indicators.

Relevant references

- International Conference of Labour Statisticians (ICLS). The ILO modelled estimates use and promote the use of the recommendations of the ICLS. For instance, the labour underutilisation model estimates the indicators introduced by the 19th ICLS.
- Detailed documentation is available for the ILO modelled estimates of the labour force, employment by economic class (working poverty), and labour income share and distribution.
- For a better understanding of the underlying data used in the ILO modelled estimates please refer to the quick guides on sources and uses of labour statistics, ILOSTAT microdata processing, and interpretation of the unemployment rate.
- The ILO modelled estimates use as input data third-party databases including: the World Economic Outlook from the International Monetary Fund, the World Development Indicators and PovcalNet from the World Bank, the World Population Prospects and National Accounts Data from the United Nations.

Data collection and evaluation

The ILO modelled estimates are generally derived for 189 countries, disaggregated by sex and age as appropriate. Before running the models to obtain the estimates, labour market information specialists from the ILO Department of Statistics, in cooperation with the Research Department, evaluate existing country-reported data and select only those observations deemed sufficiently comparable across countries. The recent efforts by the ILO to produce harmonized indicators from country-reported microdata has greatly increased the comparability of the observations. Nonetheless, it is still necessary to select the data using the following four criteria: (a) type of data source; (b) geographic coverage; (c) age group coverage; and (d) presence of methodological breaks or outliers.

With regard to the first criterion, in order for labour market data to be included in a particular model, they must be derived from a labour force survey, a household survey or, more rarely, a population census. National labour force surveys are generally similar across countries and present the highest data quality. Hence, the data derived from such surveys are more readily comparable than data obtained from other sources. Strict preference is therefore given to labour force survey-based data in the selection process. However, many developing countries, which lack the resources to carry out a labour force survey, do report labour market information on the basis of other types of household surveys or population censuses. Consequently, because of the need to balance the competing goals of data
comparability and data coverage, some household survey data or, more rarely, population census-based data are included in the models.

The second criterion is that only nationally representative (i.e. not geographically limited) labour market indicators are included. Observations corresponding to only urban or only rural areas are not included, because large differences typically exist between rural and urban labour markets, and using only rural or urban data would not be consistent with benchmark data such as gross domestic product (GDP).

The third criterion is that the age groups covered by the observed data must be sufficiently comparable across countries. Countries report labour market information for a variety of age groups, and the age group selected can influence the observed value of a given labour market indicator.

The last criterion for excluding data from a given model is whether a methodological break is present or if a particular data point is clearly an outlier. In both cases, a balance has to be struck between using as much data as possible and including observations likely to distort the results. During this process, particular attention is paid to the existing metadata and the underlying methodology for obtaining the data point under consideration.

Historical estimates can be revised in cases where previously used input data are discarded because a source that is more accurate according to the above-mentioned criteria has become available.

**Methodology used to estimate labour market indicators**

Labour market indicators are estimated using a series of models, which establish statistical relationships between observed labour market indicators and explanatory variables. These relationships are used to estimate missing observations, as well as to make projections of the labour market indicators.

There are many potential statistical relationships, also called “model specifications”, which could be used to predict labour market indicators. The key to obtaining precise and unbiased estimates is to select the best model specification in each case. The ILO modelled estimates generally rely on a procedure called cross-validation, which is used to identify those models that minimize the expected error and variance of the estimation. This procedure involves repeatedly computing a number of candidate model specifications using random subsets of the data: the missing observations are predicted and the prediction error is calculated for each iteration. Each candidate model is assessed on the basis of the pseudo-out-of-sample root mean squared error, although other metrics such as result stability are also assessed depending on the model. This makes it possible to identify the statistical relationship that provides the best estimate of a given labour market indicator. It is worth noting that the most appropriate statistical relationship for this purpose could differ depending on the country.

The basis for the ILO modelled estimates is the United Nations World Population Prospects (UN WPP), which provides estimates and projections of the total population by five-year age groups. The working age population comprises everyone who is at least 15 years of age. First, a model estimates and projects the labour force participation rates by sex and five-year age groups. These estimated and projected rates are applied to the estimates for the working age population in order to obtain the labour force. Second, a model estimates the unemployment rate by sex and for young people (15–24) and adults (25+). Combining the unemployment rate with the labour force estimates, the number of employed and unemployed is obtained. Third, a model estimates the labour underutilization rates (LU2, LU3 and LU4 rates – see further down), from which the time-related underemployment and the potential labour force can be derived. Fourth, the distribution of employment by four different indicators is estimated using four different models. These indicators are employment status, economic activity (sector), occupation, and economic class (working poverty). Lastly, combining national accounts data with the ILO
Harmonized Microdata collection on labour related earnings, the labour income share and distribution are estimated.

Although the same basic approach is used for the models of all indicators, there are nevertheless differences between the various models because of specific features of the underlying data. Further details are provided below for each model.

**Labour force estimates and projections**

The ILO labour force estimates and projections (LFEP) are part of a broader international effort on demographic estimates and projections to which several United Nations agencies contribute. Estimates and projections of the total population, and of its sex and age composition, are produced by the United Nations Population Division; of the employed, unemployed and related populations by the ILO; of the agricultural population by the Food and Agriculture Organization of the United Nations; and of the school-attending population by the United Nations Educational, Scientific and Cultural Organization.

The basic data used as input for the relevant model are single-year labour force participation rates by sex and age groups, of which ten groups are defined using five-year age intervals (15–19, 20–24, ..., 60–64) and the last age group is defined as 65 years and above. The underlying methodology has been extensively assessed in terms of pseudo-out-of-sample performance; the LFEP and the labour income share model, however, are the only models described in this document that do not automatically carry out model searching.

The estimation process is performed in two different steps, each of which is applied recursively. Linear interpolation is used to fill in the missing data for countries for which such a procedure is possible. The performance of this procedure has been found to be reasonable, which is not surprising, given that the labour force participation rate is a very persistent variable. In all other cases, weighted multivariate estimation is carried out. Countries are divided into nine estimation groups, which were chosen on the combined basis of broad economic similarity and geographic proximity. In terms of model specification, taking into account the data structure and the existence of heterogeneity among the various countries in the input data used, it was decided to use panel data techniques with country-fixed effects. The regressions are weighted by the non-response likelihood. The explanatory variables used include economic and demographic variables. The estimates are produced using the detailed five-year age intervals. The global figures are calculated using the benchmark population from the UN WPP and the detailed rates.

The projections are carried out following a different methodology than that used for the imputation of missing values over the historical period. A logistic trend model is used to extrapolate the data. The main advantage of the logistic curve and other sigmoid or S-shaped curves is that they can capture growth processes that ultimately reach a steady state. These curves are frequently used to model populations and labour force participation rates. Furthermore, on the basis of past behaviour of observed labour force participation rates, upper and lower bounds on cumulative change are imposed to avoid extrapolating changes that would be excessive judging by historical experience.

**Unemployment estimates**

This model estimates a complete panel data set of unemployment rates disaggregated by sex and age (15-24, 25+). Real observations are more likely to exist for the total unemployment rate than for the rate disaggregated by sex and age. In order to maximize the use of real information, the model first estimates the total rate. Next, the rates for male and female employment, and for youth and adult
employment, are estimated separately. These estimates are then rebalanced so that the implied total rate matches the total rate estimated in the first step. A similar procedure is used in the final step for the unemployment rates among male and female young people, and among male and female adults.

The estimation of each indicator is performed in a two-step process. In the first step, a cross-country regression is carried out to identify the level of the unemployment rate in countries with completely missing data. This step uses information on demography, per capita income, economic structure and an employment index from the Gallup World Poll. In the second step, the evolution of the unemployment rate is estimated, using information on the economic cycle and also on economic structure and demographics. The two-step process has the advantage of treating two very different econometric problems using separate approaches.

**Unemployment projections**

These models project the future development of unemployment rates from 2018 onwards. A slightly modified cross-validation procedure is used to identify the best-fitting projection models. For forecasting, a specified number of periods are dropped from the end of the sample, and projections are then made for these periods using the candidate model in order to calculate the forecast error for different forecast horizons. By shifting the point at which periods are dropped, the forecast can be evaluated for different historical periods, and hence a root-mean-squared forecast error can be calculated for each candidate model.

In a first set of projection models, quarterly data are used. The use of such higher frequency information increases the forecast accuracy. For 39 countries with available quarterly economic forecasts, a series of models are run to obtain estimates and projections. Some of these models use quarterly labour flows for forecasting, which are estimated from data on unemployment by duration. The models specified either project the unemployment rate directly or determine both inflow and outflow rates, using ARIMA, VARX and combined forecast techniques.

A second set of projection models is used to estimate the unemployment rate for countries without quarterly data, and to make projections over longer horizons for all countries. These models use the full panel data set of unemployment rates up to the last year with reported information as the base; they also make use of projections of the cyclical component of GDP growth. A series of dynamic models are specified and evaluated using the cross-validation techniques described above. The models in question are as follows:

- country-level error correction models for countries that exhibit a cointegrated relationship between employment growth and labour force growth;
- country-level model projecting the unemployment rate itself;
- country-level model projecting the change in the unemployment rate;
- panel regression model projecting the unemployment rate, where the panel dimensions are (a) geographic regions; (b) income groups; (c) oil exporters;
- multi-level mixed model with random intercepts and coefficients projecting the unemployment rate;
- multi-level mixed model with random intercepts and coefficients projecting the change in the unemployment rate.

Models are weighted on the basis of their forecasting performance over different horizons. This means that a model may receive a higher weighting in the short run, but a lower weighting in the long run. The forecast confidence interval is estimated using the weighted root-mean-squared forecast errors from the
cross-validation, together with the weighted variance of forecasts obtained from the various forecasting models.

Estimates of labour underutilization (LU2, LU3 and LU4 rates)
The target variables of the model are the measures of labour underutilisation defined in the 19th ICLS resolution. These include: the combined rate of time-related underemployment and unemployment (LU2), the combined rate of unemployment and potential labour force (LU3), and the composite measure of labour underutilization (LU4). The measures are defined as:

\[
LU2 = \frac{Unemployed + Time \text{ related underemployment}}{Labour \text{ Force}}
\]

\[
LU3 = \frac{Unemployed + Potential \text{ Labour Force}}{Labour \text{ Force} + Potential \text{ Labour Force}}
\]

\[
LU4 = \frac{Unemployed + Potential \text{ Labour Force} + Time \text{ related underemployment}}{Labour \text{ Force} + Potential \text{ Labour Force}}
\]

Persons in time-related underemployment are defined as all persons in employment who, during a short reference period, wanted to work additional hours, whose working time in all jobs was less than a specified hours threshold, and who were available to work additional hours given an opportunity for more work. The potential labour force is made up of people of working age who were actively seeking employment, were not available to have started work in the reference week, but would become available within a short subsequent period (unavailable jobseekers), or who are not actively seeking work but wanted employment and were available in the reference week (available potential jobseekers).

The model uses the principles of cross-validation and uncertainty estimation to select the regression models with the best pseudo-out-of-sample performance, not unlike the unemployment rate model. The labour underutilisation model, however, has three very specific features. First, all demographic groups are jointly estimated, using the appropriate categorical variable as a control in the regression, because the groups are interdependent (and data availability is roughly uniform across breakdown). Second, the model incorporates the information on unemployment and labour force into the regressions (used alongside other variables to reflect economic and demographic factors). Finally, the LU4 rate is uniquely pinned down by the LU2 and LU3, since it is a composite measure of the two indicators.

The resulting estimates include the LU2, LU3, and LU4 rates and the level of time-related underemployment and of the potential labour force.

Estimates of the distribution of employment by status, occupation and economic activity
The distribution of employment by status, occupation and economic activity (sector) is estimated for the total and also disaggregated by sex. In the first step, a cross-country regression is performed to identify the share of each of the employment-related categories in countries with completely missing data. This step uses information on demography, per capita income, economic structure and a model-
specific indicator with high predictive power for the estimated distribution. The indicators for each category are as follows:

- For status, an index of work for an employer from the Gallup World Poll;
- For occupation, the share of value added of a sector in which persons with a given occupation are most likely to work;
- For sector, the share of value added of the sector.

The next step estimates the evolution of the shares of each category, using information on the economic cycle and also on economic structure and demographics. Lastly, the estimates are rebalanced to ensure that the individual shares add up to 100 per cent.

The estimated sectors are based on an ILO-specific classification that ensures maximum consistency between the third and fourth revision of the United Nations International Standard Industrial Classification of All Economic Activities (ISIC). The sectors A, B, C, F, G, I, K, O, P and Q correspond to the ISIC Rev.4 classification. Furthermore, the following composite sectors are defined:

- “Utilities” is composed of sectors D and E;
- “Transport, storage and communication” is composed of sectors H and J;
- “Real estate, business and administrative activities” is composed of sectors L, M and N;
- “Other services” is composed of sectors R, S, T and U.

The estimated occupations correspond in principle to the major categories of the 1988 and 2008 iterations of the ILO International Standard Classification of Occupations (ISCO-88 and ISCO-08). However, subsistence farming occupations were classified inconsistently across countries, and sometimes even within one country across years. According to ISCO-08, subsistence farmers should be classified in ISCO category 6, namely as skilled agricultural workers. A number of countries with a high incidence of subsistence farming, however, reported a low share of category 6, but a high share of category 9 (elementary occupations). This means that the shares of occupational categories 6 and 9 can differ widely between countries that have a very similar economic structure. It is not feasible to determine the extent of misclassification between categories 6 and 9. Consequently, in order to obtain a consistent and internationally comparable classification, categories 6 and 9 are merged and estimated jointly.

**Estimates of employment by economic class**

The estimates of employment by economic class are produced for a subset of countries. The model uses the data derived from the unemployment, status and economic activity models as inputs in addition to other demographic, social and economic variables.

The methodology involves two steps. In the first step, the various economic classes of workers are estimated using the economic class of the overall population (amongst other explanatory variables). This procedure is based on the fact that the distribution of economic class in the overall population and the distribution in the working population are closely related. The economic class of the overall population is derived from the World Bank’s PovcalNet database. In general, the economic class is defined in terms of consumption, but in particular cases for which no other data exist, income data are used instead.

Once the estimates from this first step have been obtained, a second step estimates the data for those observations for which no data on the economic class of the working population nor estimates from step 1 are available. This second step relies on cross-validation and subsequent model selection to ensure a satisfactory performance.
In the present edition of the model, employment is subdivided into five different economic classes: workers living on US$0–1.9 per day, US$1.9–3.2 per day, US$3.2–5.5 per day, US$5.5–13.0 per day, and above US$13.0 per day, in purchasing power parity terms.

Estimates of the labour income share and the labour income distribution

The model estimates a complete panel dataset of the labour income share and the labour income distribution. To this end, national accounts data from the United Nations Statistics Division and the ILO Harmonized Microdata collection are combined. When national accounts data or microdata are not available, the estimates rely on a regression analysis to impute the necessary data. The imputation is based on similar countries in terms of key economic and labour market variables.

The methodology involves two steps. The first step is to compute the labour income share, adjusted for the labour income of the self-employed. Taking into account the labour income of the self-employed has been recognized in the economic literature as a crucial element for international comparability. In order to achieve this, detailed data on status in employment is used (from the model outlined above), which subdivides self-employment into three different groups: own-account workers, contributing family workers, and employers. Furthermore, the labour income of each group of the self-employed relative to the income of employees is based on a regression analysis of the microdata. The resulting estimate corresponds to the share of the economy that accrues to labour:

\[
\text{Labour Income Share} = \frac{\text{Labour Income}}{\text{Gross Domestic Product}}
\]

The second step, based on the level of labour income estimated in the first step and the microdata, produces a detailed distribution, at the percentile level, of the labour income for each country and year. Thus, one can know the n per cent of labour income that accrues to the bottom 1 per cent, to the 2\(^{nd}\) percentile, and so on. Importantly, given that the benchmark of employment follows the International Conference of Labour Statisticians recommendations, the labour income is estimated on a per worker basis, not on a full-time equivalent basis. Additionally, the distribution of labour income at the global and regional level is computed, at the decile level. Given the cross country differences in prices, the distribution of global and regional labour income deciles is computed in purchasing power parity terms.