ILO modelled estimates – methodological overview

The ILO has designed and actively maintains a series of econometric models that 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 data set so that, every year, regional and global aggregates with consistent country coverage can be computed. 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, constitute a unique, internationally comparable data set on labour market indicators.

Relevant references

- International Conference of Labour Statisticians (ICLS). The ILO modelled estimates use and promote the use of the recommendations of the ICLS[1]. 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 the 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, UIS.Stat from UNESCO, and 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. Additionally, for selected indicators an additional disaggregation by geographical area (urban and rural) is performed. 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 have greatly increased the comparability of the observations. Nonetheless, it is still necessary to select the data on the basis of the following four criteria: (a) type of data source; (b) geographical 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

[1] The employment definition following 19th ICLS is not yet implemented in the ILO modelled estimates for countries in which it would generate a methodological break as there are not enough data points based on the new standard to produce reliable global and regional estimates.

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 (non-labour force survey) household survey data and, 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). Nonetheless, when the data are explicitly to be broken down by urban versus rural location, geographically limited data covering the area of interest are included.

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 (see box B.1 for major revisions implemented for the November 2019 edition of the ILO modelled estimates).

Box B.1. Revisions to historical estimates

As in previous years, the ILO modelled estimates have been updated to take into account new information and revisions to historical data.

The main difference between the ILO modelled estimates of November 2019 and those of November 2018 is the revision of historical unemployment rates for India. There are considerable methodological differences between the recently published Periodic Labour Force Survey (PLFS), covering 2017–18, and the previously used National Sample Survey. Consequently, only the most recent data have been used by the ILO; the rest of the time series has been imputed. The new estimates of unemployment are substantially higher than the previous ones, and given the country’s size, this has a large impact on the global aggregates.

The unemployment rate has been derived directly from the PLFS microdata so as to facilitate international comparison, in particular by applying a definition of unemployment that is as close as possible to the standards set by the International Conference of Labour Statisticians. That being said, there is only one question in the PLFS that can be used to identify employment and unemployment: this is not in line with international best practice, which means that both the comparability and reliability of the results obtained using PLFS data are limited.

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 impute missing observations and to make projections for the indicators.

There are many potential statistical relationships, also called “model specifications” that could be used to predict labour market indicators. The key to obtaining accurate 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 benchmark for the ILO modelled estimates is the 2019 Revision of the United Nations World Population Prospects, which provides estimates and projections of the total population broken down into five-year age groups. The working-age population comprises everyone who is at least 15 years of age. First, a model is used to estimate and project the labour force participation rates disaggregated 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, another model is used to estimate the unemployment rate disaggregated by sex and for young people (15–24) and adults (25+). Combining the unemployment rate with the labour force estimates, the numbers of employed and unemployed are obtained. Third, another set of models is used to estimate 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 as a function of four different indicators is estimated using four different models. These indicators are: employment status, economic activity (sector), occupation, and economic class (working poverty). Fifth, a model is used to estimate the share of the youth population not in employment, education or training. Sixth, for all the aforementioned indicators – except for economic class – a breakdown by geographical area (urban and rural) is produced. Lastly, by 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 followed in the models used to estimate all the indicators, there are 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 campaign to obtain demographic estimates and projections to which several United Nations agencies contribute. Estimates and projections are produced by the United Nations Population Division for the total population, and for its sex and age composition; by the ILO for the employed, unemployed and related populations; by the Food and Agriculture Organization of the United Nations (FAO) for the agricultural population by the United Nations Educational, Scientific and Cultural Organization (UNESCO) for the school-attending population.

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

The estimation 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 geographical proximity. In terms of model specification, after taking into account the data structure and the 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 United Nations World Population Prospects 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 2018 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 2019 onwards. In a first set of projection models, quarterly data are used. The use of such higher-frequency information increases the forecast accuracy. For 44 countries with available quarterly economic forecasts, a series of models are run to obtain estimates for 2019 and projections for 2020. These models are evaluated using the model search routines described above, specifically by splitting the data into training and evaluation samples. Because of the high serial correlation of quarterly unemployment rates, a block of observations around the evaluation sample needs to be excluded from the estimation to ensure the training sample's independence from the observation that is being evaluated. Models are combined using a "jackknife model-averaging" technique described in Hansen and Racine (2012), which essentially finds the linear combination of models that minimizes the variance of the prediction error. For countries with available quarterly labour market information, but for which quarterly macroeconomic forecasts are not available, an ARIMA ("Auto Regresive Integrated Moving Average") model is used to project the remaining quarters of the year of which at least one quarter has been observed.

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 a slightly modified cross-validation procedure to identify the best-fitting projection models. For forecasting, a specified number of periods are dropped from the end of the sample, the parameters of the candidate model are re-estimated, and projections are then made for these periods 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 and each projection horizon. 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;
- a country-level model projecting the unemployment rate itself;
- a country-level model projecting the change in the unemployment rate;
- a panel regression model projecting the unemployment rate, where the panel dimensions are (a) geographical regions; (b) income groups; (c) oil exporters;
- a multi-level mixed model with random intercepts and coefficients projecting the unemployment rate;
- a 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 error bounds of the unemployment rate

When observations in the ILO modelled estimates are not real but derived using econometric techniques, they have a certain degree of uncertainty. In addition, projections of the future are also uncertain. These uncertainties are estimated for the unemployment rate. As stated above, we make use of cross-validation techniques to identify the models that minimize the prediction error. This same error describes the uncertainty due to the model-based approach. However, the unemployment rate displays some serial dependence, meaning that adjacent observations will always be closer together than observations far apart in time. Hence, the uncertainty around an estimate adjacent to a real observation is smaller than when the real observation is farther away in time. This effect is also taken into account in the construction of the error bounds.

The unemployment projection model evaluates the forecast performance over different projection horizons, and hence already provides a measure of the model-based forecast uncertainty. In addition, we also compute a measure of the uncertainty around GDP growth projections by comparing the five-year projections of the various vintages since 1991 of the International Monetary Fund's World Economic Outlook database with the realized values. Using this measure of uncertainty, we simulate 100 random realizations of GDP growth projections, use these to project unemployment 100 times, and then compute the variance due to growth forecast uncertainty of these simulated projections. The total variance of the unemployment projection is the sum of the model-based variance and the growth uncertainty variance.

Estimates of labour underutilization (LU2, LU3 and LU4 rates)

The target variables of the model are the measures of labour underutilization defined in the resolution concerning statistics of work, employment and labour underutilization adopted by the 19th International Conference of Labour Statisticians (ICLS) in October 2013. These measures include the combined rate of time-related underemployment and unemployment (LU2), the combined rate of unemployment and the potential labour force (LU3), and the composite measure of labour underutilization (LU4). The measures are defined as:

$$\mathrm{LU2} = \frac{\text{Unemployed} + \text{Time related underemployment}}{\text{Labour force}}$$

$$\mathrm{LU3} = \frac{\text{Unemployed} + \text{Potential labour force}}{\text{Labour force} + \text{Potential labour force}}$$

$$\mathrm{LU4} = \frac{\text{Unemployed} + \text{Potential labour force} + \text{Time related underemployment}}{\text{Labour force} + \text{Potential labour force}}$$

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 their jobs was below a specified threshold of hours, and who were available to work additional hours if they had been given the opportunity to do so. The potential labour force consists of people of working age who were actively seeking employment, were not available to start work in the reference week, but would become available within a short subsequent period (unavailable jobseekers), or who were not actively seeking employment but wanted to work 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 underutilization 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 rates, since it is a composite measure based on 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 people 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. However, a number of countries with a high incidence of subsistence farming reported a low share of workers in category 6, but a high share for 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 138 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 (among 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 neither 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 selection of the best-performing model 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 data set of the labour income share and the labour income distribution. To this end, national accounts data from the United Nations Statistics Division and labour income data from 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 countries that are similar 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 are used (from the model outlined in the preceding section), 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 estimated on the basis of a regression analysis of the microdata. The resulting estimate corresponds to the share of total income that accrues to labour:

$$\text{Labour income share} = \frac{\text{Labour income}}{\text{Gross domestic product}}$$

The second step, drawing on the level of labour income estimated in the first step and on the microdata, produces a detailed distribution, at the percentile level, of the labour income for each country and year. It is thus possible to determine the percentage of aggregate labour income that accrues to the bottom (first) percentile, to the second percentile, and so on. Importantly, given that the definition of employment follows the ICLS 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. Because of the cross-country differences in prices, the distribution of global and regional labour income deciles is computed in purchasing power parity terms.

Estimates related to youth not in employment, education or training

The target variable of the model is the share of youth not in employment, education or training (NEET):

$$\text{NEET share} = \frac{\text{Youth not in employment, education or training}}{\text{Youth population}}$$

It is worth noting that, by definition, 1 minus the NEET share gives the share of young people who are either in employment or enrolled in some educational or training programme. The NEET share is included as one of the indicators used to measure progress towards the achievement of the Sustainable Development Goals, specifically of Goal 8 (“Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all”).

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 NEET model estimates all demographic groups jointly, using the appropriate categorical variable as a control in the regression, because the groups are interdependent and data availability is roughly uniform across breakdown. The model incorporates the information on unemployment, labour force and enrolment rates into the regressions (used alongside other variables to reflect economic and demographic factors). The resulting estimates include the NEET share and the number of NEET youth.

Estimates of key indicators by geographical area: Urban and rural labour market indicators

Separate estimates for urban and rural areas are produced for the following indicators: labour force, unemployment, LU2, LU3, LU4, youth NEET share and the employment distribution by status, economic activity and occupation.

In order to produce the estimates, the models decompose the variable of interest into two components. The procedure described here is for the labour force model; an analogous procedure is used for the other models. The labour force participation rate (LFPR) by geographical area that the model estimates can be expressed as:

$$\text{Labour force participation rate}_{\text{ij}} = \frac{\text{Labour force}_{\text{ij}}}{\text{Population}_{\text{ij}}}$$

$$\text{i} = \{\text{urban}, \text{rural}\}\,;\ \text{j} = \{\text{gender} \times \text{age}\}$$

One relationship of particular importance between the urban and rural rates and the national rates is that the distance of the former rates to the latter rate determines the respective share of the urban and rural population (the denominator of the LFPR expression). The strategy of the modelling approach is to target, for the estimation, two variables that jointly determine the rural and urban LFPRs. The main variable used to produce the LFPR is the spread between urban and rural LFPR:

$$\text{Spread urban} = \frac{\text{Urban LFPR}}{\text{Rural LFPR}} = \frac{1}{\text{Spread rural}}$$

This variable alone does not pin down both the urban and rural LFPRs. Another variable is necessary to complete the system of equations that can be used to produce the two rates. The other variable is the share of the denominator of the LFPR expression by type of area, which is simply the population:

$$\begin{aligned}\text{Share urban} &= \frac{\text{Urban labour force}/\text{Urban LFPR}}{\text{Rural labour force}/\text{Rural LFPR} + \text{Urban labour force}/\text{Urban LFPR}} \\ &= 1 - \text{Share rural}\end{aligned}$$

Decomposing the two rates into the spread and share variables has two main advantages. First, it makes it possible to model explicitly the dependence between the distances of the two rates to the total rate and the share of the population in urban and rural areas. The second advantage is that this framework is easy to extrapolate to the other variables of interest. Once these two auxiliary variables have been estimated using regression methods, the results can easily be used to compute the urban and rural rates of interest:

$$\text{Urban LFPR} = \frac{\text{LFPR}}{\text{Share urban} + \frac{\text{Share rural}}{\text{Urban spread}}}$$

$$\text{Rural LFPR} = \frac{\text{LFPR} - \text{Share urban} * \text{Urban LFPR}}{\text{Share rural}}$$

As mentioned above, the unemployment, labour underutilization, NEET and employment distribution models follow the same procedure.

In order to estimate the spread and share for all the variables, the models of key indicators by geographical area use 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. However, in this case the targets of the estimation are the spread and share variables instead of the variable of interest directly. In the geographical models, 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. The models use various indicators to reflect economic and social factors as explanatory variables for the imputation. Finally, the modelling procedure ensures the consistency of interdependent variables. For this purpose, labour force estimates are used as a basis for the models of the distribution of unemployment and labour underutilization by geographical area. The population benchmark, derived from the labour force model, is used in the model of the NEET distribution by geographical area. Similarly, estimates of unemployment by rural and urban area are used as the basis for the estimates of labour underutilization by geographical area. Finally, the employment estimates derived jointly from the models of the distribution of the labour force and unemployment by geographical area are used as a basis for estimating the distributions of employment with respect to status, economic activity and occupation by geographical area.

The resulting estimates are of the shares or rates and the corresponding levels. The following estimates are available by rural and urban breakdown: LFPR, number of people in the labour force, unemployment rate, unemployment level, LU2 rate, time-related underemployment, LU3 rate, potential labour force, LU4 rate, composite labour underutilization measure, and the distribution of employment by employment status, economic activity and occupation.

Social unrest index

The social unrest index provides a reflection of “social health” at the national level. The index uses data from the Global Database of Events, Language and Tone (GDELT) project on events around the world classified as “protests” (code 14 in the database). Many different types of protest behaviours are recorded, such as street protests, riots, rallies, boycotts, blocking of roads and strikes. Such protests are not necessarily violent, but they always reflect a certain discontent with the social, political or economic situation in the country in question.

The index ranges from 0 to 100 and is computed from a log-transformation of the share of protest events in the total number of events in a year and country, as reported by the GDELT project. An index of 100 corresponds to protest events making up 15 per cent or more of the total number of events.

Social unrest is a relative concept across countries. An equal value of the social unrest index in two countries does not imply identical conditions of social unrest in both because of the inherent differences in countries’ culture, history and methods of reporting. The social unrest index enables a cross-country comparison which identifies those countries or regions that are experiencing periods of heightened unrest. However, it is conceptually incorrect to state that one country experiences, say, 10 per cent more unrest than another.