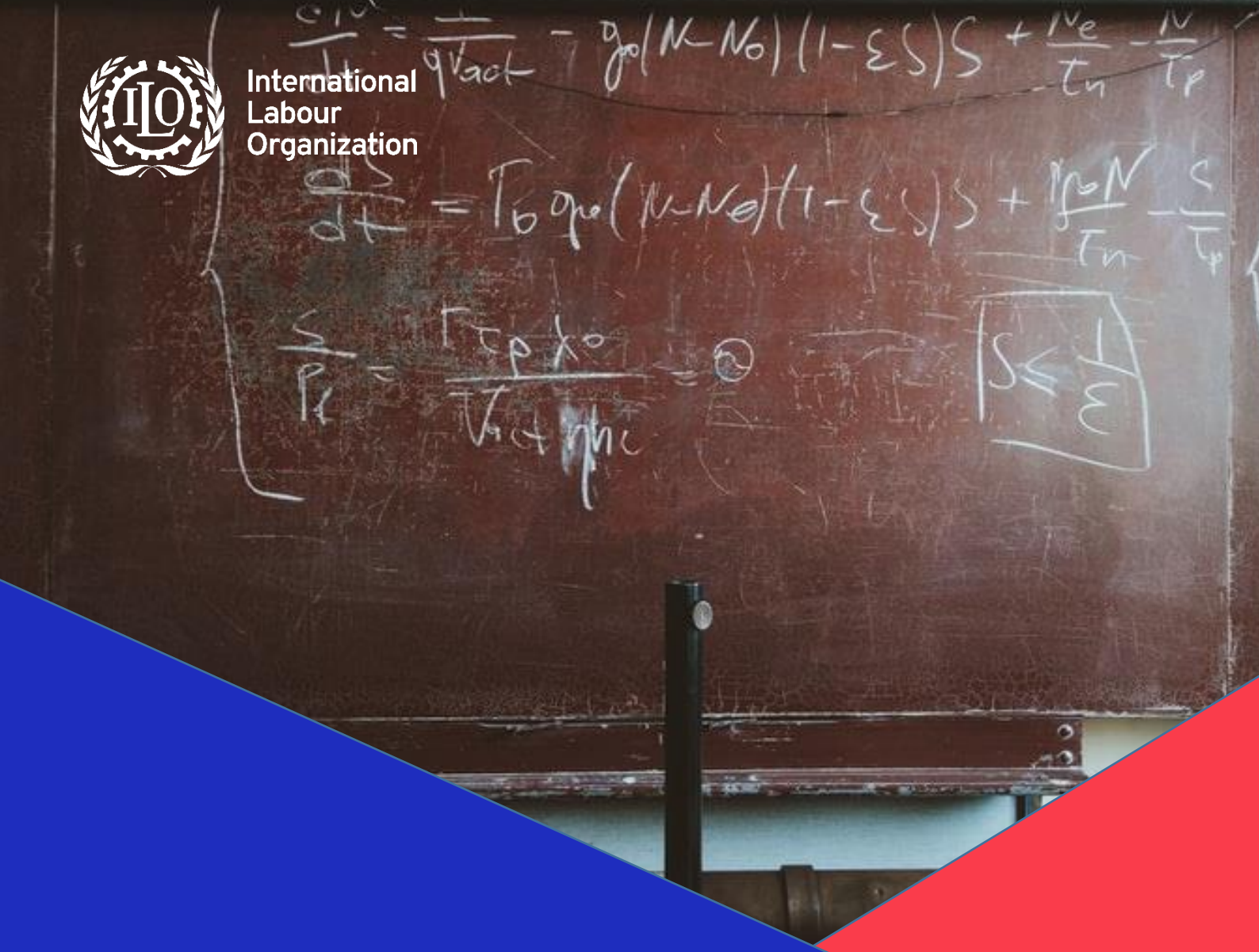

International Labour Organization

ILO modelled estimates Methodological overview

January 2024

Department of Statistics

International Labour Organization

ILO modelled estimates

Methodological overview

Department of Statistics

January, 2024

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, constitutes a unique, internationally comparable dataset of key 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 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 from third-party databases including: the World Economic Outlook from the International Monetary Fund, the World Development Indicators and the Poverty and Inequality Platform (PIP) from the World Bank, UIS.Stat from UNESCO, Women in Parliament Data from the Inter-Parliamentary Union, the World Value Survey Data, Community Mobility Reports from Google, the Oxford Covid-19 Government Response Tracker Database, 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 and territories, disaggregated by sex and age as appropriate. For selected indicators, an additional disaggregation by geographical area (urban and rural) is presented. 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

[1] The employment definition following the 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 standards to produce reliable global and regional estimates.

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 based on 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 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 based on other types of household surveys or population censuses. Consequently, 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 prohibitively 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 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 must 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.

General 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 based on 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. For indicators with subcomponents, such as gender, age or urban/rural, all subcomponent models are estimated separately and, thus, the sum of the subcomponents may be incompatible with the total estimates. The subcomponents are adjusted proportionally to match the total estimates.

The extraordinary disruptions of the global labour market caused by the COVID-19 pandemic rendered the series of models underlying the ILO modelled estimates less suitable for estimating and projecting the evolution of labour market indicators. For this reason, the methodology has been adapted, and explanatory variables that are specific to the COVID-19 pandemic have been introduced into the modelling process.

The benchmark for the ILO modelled estimates is the 2022 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.

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.

Estimates of the labour force

Methodological changes are introduced in the current version of the labour force participation rate (LFPR) model in order to produce more granular age breakdowns. The basic data used as

input for the LFPR model are single-year LFPRs disaggregated by sex and age groups, the latter comprising four intervals (15–24, 25–54, 55–64 and 65+). Compared with earlier years when only two intervals were available (15–24 and 25+), the additional age groups significantly increase the amount of input data. Moreover, estimates for the 25+ age group can still be recovered with the new methodology. The underlying methodology has been extensively assessed in terms of pseudo-out-of-sample performance. However, for certain types of missing data patterns, the LFPR is the only model described in this appendix which does not carry out automatized model selection.

Linear interpolation is used to fill in the missing data for countries for which such a procedure is possible. This procedure produces accurate estimates of low variance, which is not surprising, given that the LFPR is a very persistent variable. In all other cases, weighted multivariate estimation is carried out. Countries are divided into nine estimation groups, chosen on the combined basis of broad economic similarity and geographical proximity. Based on the data structure and the heterogeneity among the countries covered by the input data, the model was specified using panel data with country fixed effects. The regressions are weighted by the inverse of the likelihood of a labour force survey's availability. The explanatory variables used include economic and demographic variables. To produce estimates for 2020, a cross-validation approach is used to select the model that minimizes prediction error in that specific year. The tested models include annual averages of high-frequency indicators related to the evolution of the COVID-19 pandemic. An additional module is used to produce estimates for the year 2021. In addition to the use of cross-validation for model selection, macroeconomic and labour market indicators are utilized to estimate short-run dynamics while accounting for the pre-2020 trend. Finally, for 2022 a separate cross-validation procedure is applied to various models. The model predictors include macroeconomic and demographic factors including country fixed effects, hence leveraging the panel structure and utilizing the entire input data sample. The global figures are calculated using the benchmark population from the United Nations World Population Prospects and the LFPRs.

Rebalancing the estimates ensures that the implied total rate obtained from summing the demographic breakdowns matches the total rate as derived from the labour force surveys or as estimated.

In previous editions of the ILO modelled estimates, detailed age information for the labour force were available.[2] Currently, the model has been discontinued and the associated dataset (except for the indicator of the median age of the labour force) is no longer published to avoid inconsistencies with the LFPR model described above.

[2] Five-year age intervals (15–19, 20–24, and so on until 60–64) and a last age group of 65 years and above.

Estimates of unemployment

This model estimates a complete panel data set of unemployment rates disaggregated by sex and age (15–24, 25+). For countries for which at least one observation is reported,[3] regressions involving country fixed effects are used. Three best models obtained using the cross-validation approach applied over a wide range of models are combined with equal weighting to impute missing values. A separate cross-validation approach is used to select the model that minimizes prediction error in the year 2020. The candidate models include annual averages of high-frequency indicators related to the evolution of the COVID-19 pandemic. An additional procedure is used to produce estimates for 2021 which also uses a cross-validation procedure to select models. These models account for the historical trend and utilize macroeconomic indicators, including the dynamics of the unemployment rate in 2020. The procedure shows unemployment to have displayed a recovery towards that historical trend in 2021. Finally, for 2022, model estimates are calculated using the same approach (including, using the same models) as for the years up to and including 2019. For countries with no reported observations, models are selected based on a separate cross-validation exercise. Rebalancing the estimates ensures that the implied total rate obtained from summing the demographic breakdowns matches the total rate.

Estimates of hours worked

The ratio of weekly hours actually worked by the employed population is the target variable estimated for countries with missing data. Total weekly hours actually worked are derived by multiplying this ratio by the employed population.

The model is estimated in three parts. First, the model addresses countries with at least one observation of the target variable and using distinct models creates estimates up to 2022. Second, estimates for countries without any observation of this indicator are estimated. Finally, projections for 2023 and 2024 are created for all countries.

For countries with at least one observation of the indicator, estimates up to and including 2019, the regression approach uses government consumption as a share of GDP, log GDP per capita, and the rates of unemployed, time-related underemployment, and labour force. For 2020, the growth model approach uses information from the Google Community Mobility Reports, the country's average hours growth trend between 2011 and 2019, and developed and developing country fixed effects. For 2021, the growth model approach uses the indicator's change in 2020 and the average hours growth trends between 2011 and 2019. Finally, for countries with data, the regression approach for 2022 is the same as for the years up to and including 2019. Linear interpolation is then used for the estimates of the indicator to match the observed data.

3 For ease of exposition, we abstract here from the case in which reported observations exist for some demographic groups but not for others in a given country and year.

For countries without any observations of the target variable, the country intercept is estimated as a combination of a regional mean and the country's social protection rate, part-time and self-employment rate of employed people, poverty levels, share of GDP from the agricultural and industrial sectors, log GDP per capita, and government consumption as a share of GDP. Trends are estimated in a second step in a model that includes the following indicators: social protection rate, urbanisation rate, government consumption as a share of GDP, government tax revenues including social contributions as a percentage of GDP, the share of GDP from the agricultural and industrial sectors, unemployment rate, time-related underemployment rate, NEET, part-time employment rate, labour force participation rate, and the interaction of gender, income group, and time fixed effects.

Finally, the target variable projections approach for 2023 and 2024 considers the growth trend from 2011 to 2019 and projections of labour force participation rate, urbanisation rate, and log GDP per capita. Rebalancing the female and male estimates ensures that the implied total rate obtained from summing the sex breakdowns matches the total rate.

Estimates of informal employment

The target variable for this indicator is the share of informal employment in total employment by sex for the population aged 15 and older. The gender-specific country-level data used for the models includes self-employment, part-time employment and social protection rates. The country-level data includes the percentage of people under various poverty lines, percentage of people over 65 and under 15, share of employment in the agriculture and industrial sectors, urbanisation rate, the logarithm of GDP per capita, government consumption as a share of GDP, and categorical variables for geographic region and level of economic development.

The imputations for missing data are produced through five separate econometric models. First, a model produces estimates from 2004 to 2019 for countries with at least one yearly data point of share of informal employment by sex. Second, a model produces estimates from 2004 to 2019 for those countries with no data on share of informality during the entire period. The third and fourth models are growth models used to produce estimates for the 2020 pandemic year and the recovery period of 2021, respectively. The final model estimates the projections for 2022. The five distinct models were chosen from an array of models based on the cross-validation methodology, which selects the models with the highest accuracy in predicting informality rates in pseudo out-of-sample simulations. The predictions from the models are used to estimate the missing observations of the share of informal employment in total employment by sex.

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:

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

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

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

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, similar to 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 breakdowns. 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 jobs gap

The aim of the model is to provide aggregate estimates of the jobs gap rate by sex for the population aged 15 or older. The jobs gap rate is the target variable estimated for countries with missing data and is computed as follows:

$$\text{Jobs gap rate} = \frac{(\text{Unemployed} \ + \ \text{Potential labour force} + \text{Willing non-jobseekers})}{(\text{Labour Force} \ + \text{Potential labour force} + \text{Willing non-jobseekers})}$$

where the potential labour force and willing non-jobseekers include persons who were seeking employment and were not available but would become available in a short time (unavailable jobseekers), persons who were not seeking work but were currently available (available potential jobseekers) and persons who were not seeking work and were not available but were willing to work (willing non-jobseekers). The gender-specific country-level data used for the models includes the unemployment rate, unemployment-to-population ratio, the share of the extended labour force in unemployment or in the potential labour force (LU3), and the economic inactivity rate. The country-level data also includes the percentage of people aged 65 and older, log GDP per capita, and categorical variables for geographic region and levels of economic development.

The imputations for missing country data are produced with the predictions of five separate econometric models. First, a model produces estimates from 2004 to 2019 for countries with at least one observation of the target variable. Second, a model produces estimates from 2004 to 2019 for those countries with no observations of the targe period during the entire period. The third and fourth models are used to produce estimates for the 2020 pandemic year and the recovery period of 2021 and 2022, respectively. The final model generates projections for 2023 using a growth approach that includes the jobs gap growth trend from 2011 to 2019, and the projections for the change in unemployment rate and log GDP per capita.

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 total employment and 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 for which no data are available. 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, the index called “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, as well as on economic structure and demographics. The third step estimates the change in the shares of each category in the years 2020 and 2021. 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 revisions 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 are 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 in 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, 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, along with 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 Poverty and Inequality Platform (PIP) database.[4] 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.

[4] Poverty and Inequality Platform (PIP), version 20230328_2017_01_02_PROD; Mahler, Daniel Gerszon; Yonzan, Nishant; Lakner, Christoph. 2022. The Impact of COVID-19 on Global Inequality and Poverty. Policy Research Working Papers; 10198. © World Bank, Washington, DC.; World Bank (2022).

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 less than US$2.15 per day, on more than US$2.15 and less than US$3.65 per day, on more than US$3.65 and less than US$6.85per day and above US$6.85 per day, in PPP terms.

Estimates of labour force participation rate of couple households with children under age 6

This model estimates a complete panel data set of labour force participation rates disaggregated by sex for prime-age (25-54) couple households with at least one child under the age of 6. The target variable for the model is the difference between labour force participation rates of prime-age persons and the prime-age couple households with at least one child under the age of 6.

All models described below have been chosen based on pseudo-out-of-sample root mean square error (cross-validation). For countries where at least one observation of the target variable is reported from 2004 to 2019, the approach for each gender uses the share of the population aged 14 and below, urbanisation rate, log GDP per capita, and country fixed effects. For countries with no reported observations, the model uses the self-employment rate, social norm perceptions indicators, proportion of seats held by women in national parliaments, informal employment rate, urbanisation rate, the share of the population aged 14 and below, shares of employment in the agriculture, industry, and services sectors, and regional-gender-time fixed effects. A separate growth approach is used for 2020 that includes the interaction between sex, a dummy for country-gender pairs where the target variable is positive in 2019, and the lag of the target variable. An additional procedure is used to produce estimates for 2021 and 2022 using the urbanisation rate, share of population aged 14, log GDP per capita, and country-sex fixed effects. Finally, the labour force participation of prime-age couple households with at least one child under the age of 6 is computed using the existing modelled estimates of the labour force participation of prime-age persons plus the modelled estimates of the target variable.

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 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, as 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 employment by status model), 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. The imputed labour income at the micro level is also used to produce estimates of labour income by gender. 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 of youth not in employment education or training

The target variable of the model is the share of youth (aged 15 to 24) not in employment, education or training (NEET):

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

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 estimation procedure uses cross-validation to select the regression models with the best out-of-sample performance. 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 breakdowns. The model incorporates information on unemployment, labour force participation and enrolment rates into the regressions (used alongside other variables to reflect economic and demographic factors). There is one regression model for countries with at least one data point and a second model for countries with no available data. Note that in contrast to some other indicators, the

same models are used for the entire period, including the COVID-19 pandemic. Since the estimates of the NEET rate rely heavily on the unemployment and labour force participation estimates, the COVID-19-related shock is already sufficiently accounted for. 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:

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

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

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:

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

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.

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, 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 breakdowns. 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 geographic area. Finally, the employment estimates derived jointly from the models of the distribution of the labour force and unemployment by geographic 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 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 status, economic activity, and occupation.

Models used to project labour market indicators

Step 1 - Projections at the quarterly frequency

The quarterly projections for the unemployment rate, the employment-to-population ratio, and the labour force participation rate use high-frequency data such as business confidence indices in addition to economic growth forecasts to test a series of models. The approach is in line with the direct nowcasting approach used to estimate hours worked (Gomis et al. 2022). These models are evaluated using the model search routines described above, including splitting the data into training and evaluation samples. Models are combined using the "jack-knife 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.

The ratios of employment and labour force to the population have been strongly affected by the COVID-19 pandemic. The projection model is based on the assumption that these ratios have a tendency to return to their pre-COVID long-term trend. In technical terms, the projection is based on an error correction model, the correction parameter being estimated using an econometric specification that includes the gap between the actual historical series and the long-term trend.[5]

Step 2 - Projections at the annual frequency

The annual projection pools countries and utilizes vector error correction models. The indicators projected using this method are the employment-to-population ratio, the labour force participation rate, and the unemployment rate. This estimation strategy over-identifies the labour force as a target variable since it can be computed as the sum of unemployment plus employment. However, these redundancies are averaged and the reliance on a single specification is avoided by proceeding this way.

Three different approaches are used to derive projections, which are then combined into a weighted average. In all three approaches the forecast variable of interest is the annual change in the above-mentioned indicators. The first approach contains elements of error correction, while the second and third approaches do not. The first and second approaches pool countries globally, while the third approach pools countries according to geographical and economic similarity.

[5] The long-term trend is estimated using a Hodrick–Prescott filter on data from 1991 to 2019. The smoothing parameter is set at 3,200, which is larger than the parameter of 1,600 usually used in filtering time series at quarterly frequency and hence results in less variability in the trend.