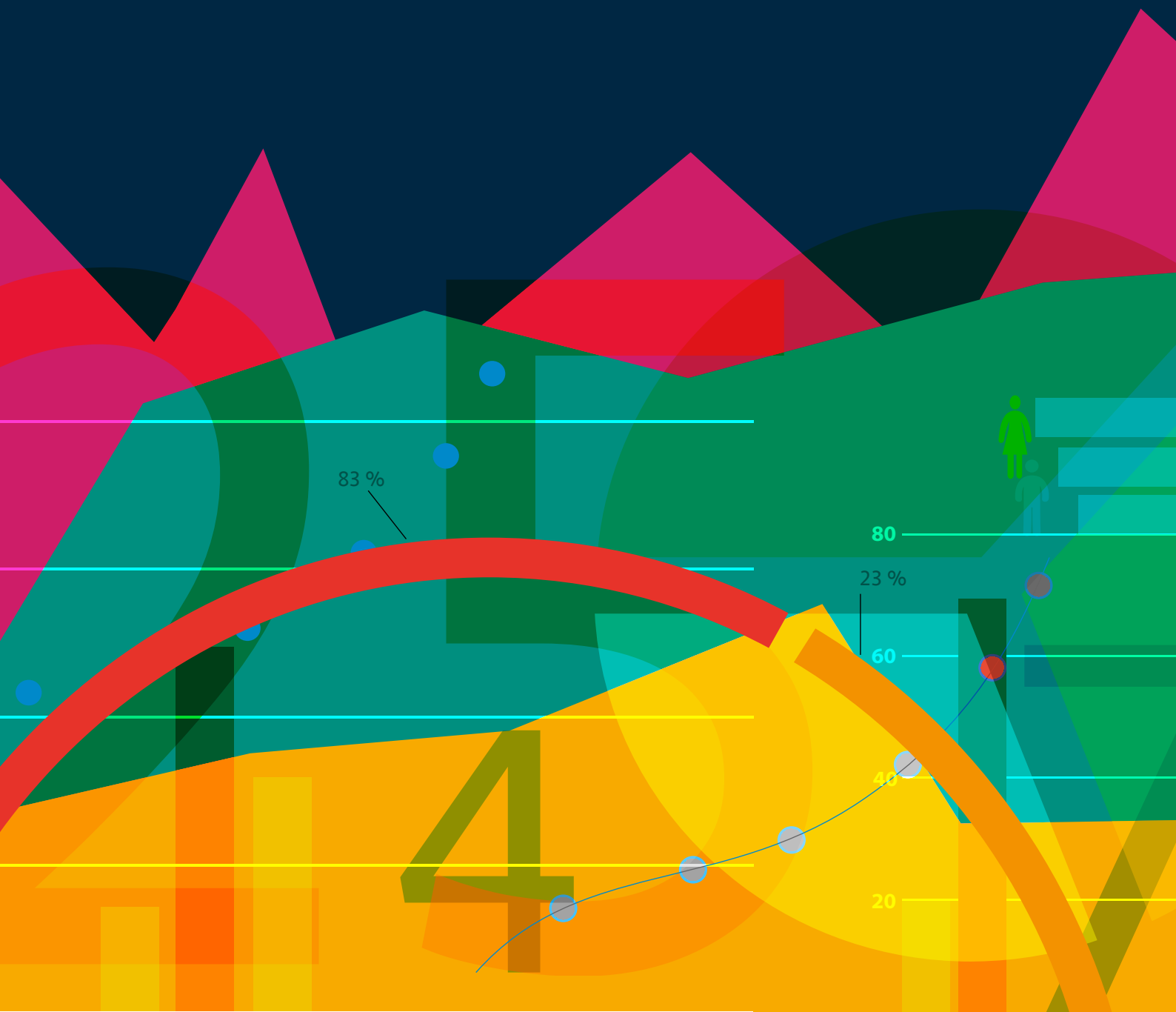
International Labour Office

ILO LABOUR FORCE ESTIMATES AND PROJECTIONS: 1990-2030 (2017 EDITION)

Methodological description

November 2017



ILO LABOUR FORCE ESTIMATES AND PROJECTIONS: 1990-2030 (2017 EDITION)

Methodological description

November 2017

Contents

Preface

The 2017 Edition of the ILO Labour Force Estimates and Projections (LFEP) Database was produced by the ILO Department of Statistics.

There are two important changes in this edition as compared to the previous edition. Firstly, the process for incorporating input data has been substantially modified in order to fully integrate the LFEP model within the ILOSTAT database structure, thus taking advantage of ILOSTAT's quality control system. Secondly, the projections are obtained solely based on econometric methods, becoming more transparent and replicable.

Consistent with the previous edition (2015), the historical estimates (1990-2016) are accompanied by detailed metadata for each data point. The metadata include several fields regarding the source of collected data, the type of adjustments made to harmonise data (when needed) and the type of imputation method used to fill missing data.

This document was prepared by Evangelia Bourmpoula, Roger Gomis (ILO Department of Statistics), Steven Kapsos (ILO Department of Statistics) and Jean-Michel Pasteels. This work has benefited from the valuable comments of Rafael Diez de Medina, Director of the ILO Department of Statistics.

1. Introduction

The ILO programme on labour force estimates and projections is part of a larger international effort on demographic estimates and projections to which several UN agencies contribute. Estimates and projections of the total population and its components by sex and age group are produced by the UN Population Division, the employed, unemployed and related populations by the ILO, the agricultural population by FAO and the school attending population by UNESCO.

The main objective of the ILO programme is to provide member States, international agencies and the public at large with the most comprehensive, detailed and comparable estimates and projections of the labour force for countries and territories, the world as a whole and its main geographical regions. The first edition was published by the ILO Department of Statistics in 1971 (covering 168 countries and territories, with reference period 1950-1985)[1]; the second edition in 1977 (with 154 countries and territories and reference period 1975-2000)[2]; the third edition in 1986 (with 156 countries and territories and reference period 1985-2025)[3]; the fourth edition in 1996 (with 178 countries and territories and reference period 1950-2010)[4]; the fifth edition in 2007 (with 191 countries and reference period 1980-2020) with two subsequent updates (in August 2008 and December 2009)[5]. The sixth edition (2011) covered 191 countries and territories. The reference period for the estimates was 1990-2010 and for the projections was 2011-2020. The 2013 edition covered 191 countries, with a reference period of 1990-2012 for the estimates and 2013-2030 for the projections. The 2015 edition covered 193 countries. The reference period was 1990-2014 for the estimates and 2015-2050 for the projections. The present 2017 edition covers 189 countries. The reference period is 1990-2016 for the estimates and 2017-2030 for the projections.[6]

The basic data are single-year labour force participation rates by sex and age groups, of which ten groups are defined by five-year age intervals (15-19, 20-24, ..., 60-64) and the last age group is defined as 65 years and above. The data are available on ILOSTAT, the ILO's central statistical database: ilostat.ilo.org.

This document describes the main elements of the estimation and projection methodologies adopted for the 2017 edition. This edition continues to use the enhanced methodologies that were developed in order to improve the labour force estimates and projections in the 6th edition in 2011 and continued in the 2013 and 2015 editions.[7] As established in the 6th edition, the historical estimates are accompanied by detailed metadata for each data point. The metadata include several fields regarding the source of collected data, the type of adjustments made to harmonise them (when needed) and the type of imputation method used to fill missing data.

There are, however, some important changes in this edition as compared to the previous editions. Firstly, the process of obtaining and incorporating model input data has been substantially modified in order to fully integrate the LFEP model within the ILOSTAT database structure. Thus the new edition takes advantage of ILOSTAT's quality control system. Secondly, the projections are obtained solely based on econometric methods. Therefore, the projections have become more transparent and replicable.

[1] ILO, Labour force projections, 1965-85 (1st edition, Geneva 1971).

[2] ILO, Labour force projections, 1950-2010 (2nd edition, Geneva 1976).

[3] ILO, Economically Active Population: Estimates and projections, 1950-2025 (3rd edition, Geneva 1986).

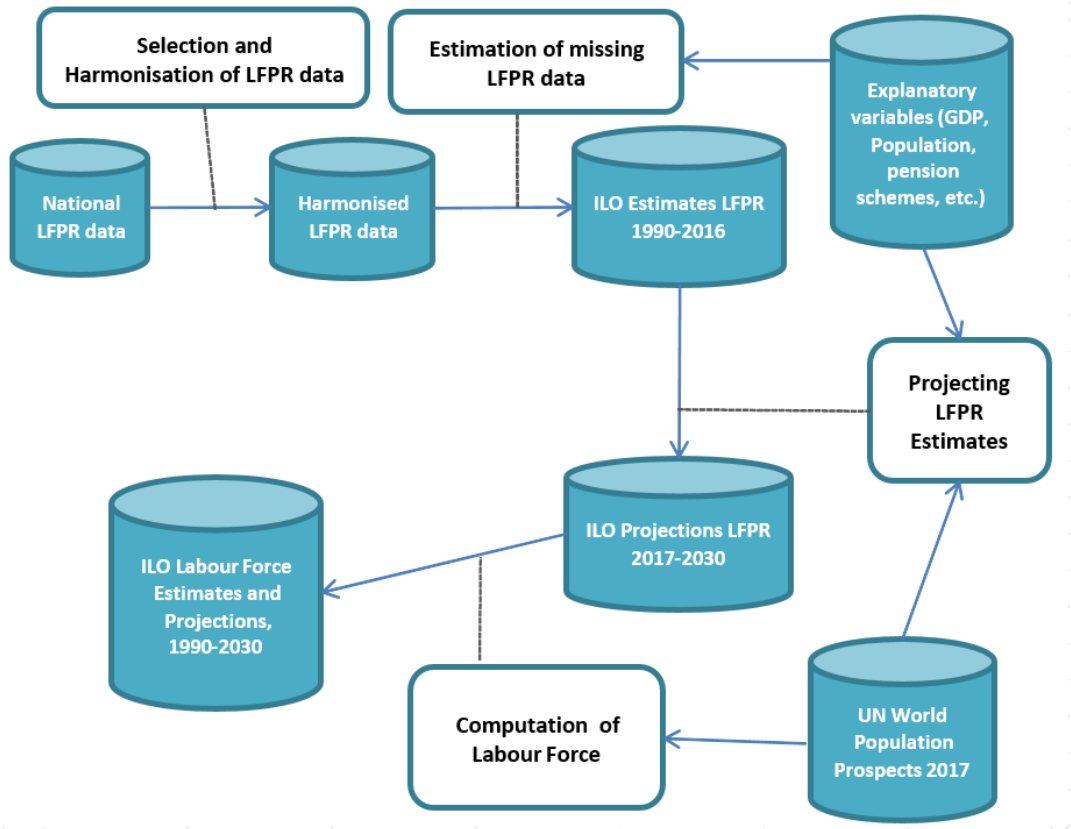
[4] ILO, Economically Active Population Estimates and projections, 1950-2010 (4th edition, Geneva 1996).

[5] ILO, Estimates and Projections of the Economically Active Population, 1980-2020 (5th edition, Geneva 2007, Update August 2008, Update December 2009).

[6] The reduction from 193 to 189 reference areas is due to the inclusion of the data of Dominique, Martinique, French Guyana and Réunion into metropolitan France. The change in horizon from 2050 to 2030 allows to base the projections solely on econometric methods.

[7] ILO, Estimates and Projections of the Economically Active Population, 1990-2030 (2013 edition, Geneva 2013); ILO, Labour Force Estimates and Projections, 1990-2050 (2015 edition, Geneva 2015).

Figure 1: ILO Labour Force Estimates and Projections 1990-2030 (2017 edition)



The determinants of labour force participation are described in section 2. The underlying labour force data used for producing harmonised single-year ILO country estimates of labour force participation rates by sex and standard age groups are described in section 3. This section also includes the description of the statistical treatment of missing values and the estimation models for countries for which no or limited data were available. The projection methodology is described in section 4. The different strengths and limitations of the present methodology are presented in section 5, as well as proposed directions for future work.

2. Concepts, definitions and theoretical background

2.a. Different forms of employment

The labour force is defined as the sum of the employed and the unemployed. For the exact definitions of those two concepts see *Resolution concerning statistics of work, employment and labour underutilization*, adopted by the 19th International Conference of Labour Statisticians (ICLS) in 2013.

According to the Resolution,[8] persons in employment are defined as all those of working age who, during a short reference period, were engaged in any activity to produce goods or provide services for pay or profit. They comprise:

(a) employed persons "at work", i.e. who worked in a job for at least one hour;

[8] Resolution concerning statistics of work, employment and labour underutilization, adopted by the 19th International Conference of Labour Statisticians, Geneva, 2013; http://www.ilo.org/global/statistics-and-databases/standards-and-guidelines/resolutions-adopted-by-international-conferences-of-labour-statisticians/WCMS_230304/lang--en/index.htm.

(b) employed persons "not at work" due to temporary absence from a job, or to working-time arrangements (such as shift work, flexitime and compensatory leave for overtime).

"For pay or profit" refers to work done as part of a transaction in exchange for remuneration payable in the form of wages or salaries for time worked or work done, or in the form of profits derived from the goods and services produced through market transactions, specified in the most recent international statistical standards concerning employment-related income. It includes remuneration in cash or in kind, whether actually received or not, and may also comprise additional components of cash or in-kind income. The remuneration may be payable directly to the person performing the work or indirectly to a household or family member.

Employed persons on "temporary absence" during the short reference period refers to those who, having already worked in their present job, were "not at work" for a short duration but maintained a job attachment during their absence.

Included in employment are: (a) persons who work for pay or profit while on training or skills-enhancement activities required by the job or for another job in the same economic unit, such persons are considered as employed "at work" in accordance with the international statistical standards on working time; (b) apprentices, interns or trainees who work for pay in cash or in kind; (c) persons who work for pay or profit through employment promotion programmes; (d) persons who work in their own economic units to produce goods intended mainly for sale or barter, even if part of the output is consumed by the household or family; (e) persons with seasonal jobs during the off season, if they continue to perform some tasks and duties of the job, excluding, however, fulfilment of legal or administrative obligations (e.g. pay taxes), irrespective of receipt of remuneration; (f) persons who work for pay or profit payable to the household or family; (g) regular members of the armed forces and persons on military or alternative civilian service who perform this work for pay in cash or in kind.

Excluded from employment are: (a) apprentices, interns and trainees who work without pay in cash or in kind; (b) participants in skills training or retraining schemes within employment promotion programmes, when not engaged in the production process of an economic unit; (c) persons who are required to perform work as a condition of continued receipt of a government social benefit such as unemployment insurance; (d) persons receiving transfers, in cash or in kind, not related to employment; (e) persons with seasonal jobs during the off season, if they cease to perform the tasks and duties of the job; (f) persons who retain a right to return to the same economic unit but who were absent; (g) persons on indefinite lay-off who do not have an assurance of return to employment with the same economic unit.

Persons in unemployment are defined as all those of working age who were not in employment, carried out activities to seek employment during a specified recent period and were currently available to take up employment given a job opportunity, where: (a) "not in employment" is assessed with respect to the short reference period for the measurement of employment; (b) to "seek employment" refers to any activity when carried out, during a specified recent period comprising the last four weeks or one month, for the purpose of finding a job or setting up a business or agricultural undertaking. This includes also part-time, informal, temporary, seasonal or casual employment, within the national territory or abroad; (c) "currently available" serves as a test of readiness to start a job in the present, assessed with respect to a short reference period comprising that used to measure employment.

It must be noted that in practice there are large differences in terms of country practices regarding definitions of employment and unemployment (see ILO 2011).

2.b. Labour force participation rates

The labour force projections are obtained by the product of two separate projections: a projection of the population (*POP*) of country *i* at time *t+h* (*t* and *h* are respectively the projection origin and horizon) for the

age group *a* (e.g. those aged [20-24]) and sex *s,* and a projection of the labour force participation rate (LFPR) for the same subgroup of the population.

$$LF_{i,t+h,a,s} = LFPR_{i,t+h,a,s} \cdot POP_{i,t+h,a,s}$$

where: $LFPR_{i,t+h,a,s} = \frac{LF_{i,t+h,a,s}}{POP_{i,t+h,a,s}}$

The decomposition of the projection exercise into two phases has several advantages. Firstly, the determinants of the changes in population and the LFPR are not the same and can be identified.[9] The determinants of the changes in population are primarily due to changes in fertility, mortality and migration flows (United Nations 2011), while the changes in the LFPR can be the result of many factors, including changes in labour demand, as highlighted in the next section. Secondly, the LFPR varies by definition between 0 and 100 per cent, which is convenient, since logistic transformations can be applied to the LFPR in order to ensure that projected values within the 0-100 per cent interval are obtained.

2.c. The determinants of the LFPR

At the macroeconomic level, average aggregated labour force participation rates are observed for the whole population or for population subgroups (male, female, prime age, youth, etc.). These data are typically derived from labour force or other household surveys or from population censuses. The variable "labour force participation rate" is of dichotomous nature: either you participate or you do not. The determinants of the LFPR can be broken down into structural or long-term factors, cyclical factors and accidental factors.

Structural factors include policy and legal determinants (e.g., flexibility of working-time arrangements, taxation, family support, retirement schemes, apprenticeships, work permits, unemployment benefits, minimum wage) as well as other determinants (e.g., demographic and cultural factors, level of education, technological progress, availability of transportation).

Some key findings regarding female labour force participation rates (LFPR)[10]:

- In countries where working-time arrangements are more flexible, there is a higher LFPR of female workers than in other countries.
- Taxation of second earners (relative to single earners) usually has a negative impact on female LFPR.
- Childcare subsidies and paid parental leave usually have a positive impact on female LFPR.
- In countries where the proportion of unmarried women is higher, there is usually a higher female LFPR than in other countries.
- Cultural factors such as strong family ties or religion have a strong impact on LFPR for some subgroups of the population. For example, in many countries, religious or social norms may discourage women from undertaking economic activities.

These types of structural factors are the main drivers of the long-term patterns in the data. Changes in policy and legal determinants (e.g., changes in retirement and pre-retirement schemes) can result in important shifts in participation rates from one year to another.

Cyclical factors refer to the overall economic and labour market conditions that influence the LFPR. In other words, demand for labour has an impact on the labour force. In times of recession, two effects on the

[9] See Armstrong *et al.* (2005) for a presentation on the decomposition of complex time series and its pros and cons.

[10] For more details see Jaumotte (2003).

participation rates, with opposite signs, are referred to in the literature: the “discouraged worker effect” and the "additional worker effect”.

The "discouraged worker effect" applies to persons not working, available for work, but who stopped searching for a job. During times of recession, this effect is very important for younger people, who typically have more problems finding a job than more experienced workers and also may opt to extend the length of studies. As observed by the OECD (2010), in times of adverse labour market conditions, the discouraged worker effect for young people is much higher in places where there is easier access to post-secondary education.

The "additional worker effect" applies more to female or older workers who enter (or re-enter) the labour market in order to compensate for the job losses and decreased earnings of some members of the family or the community. In times of severe downturns, the changes in the LFPR of older persons depend on financial incentives to continue working as compared to taking retirement (OECD 2010).

Lastly, there are **accidental factors** such as wars and natural disasters that also affect LFPR, usually in a temporary manner.

3. Estimation model 1990-2016: data and methodology

3.a. Introduction

The LFEP database is a collection of actual observations and ILO estimated labour force participation rates. The database is a complete cross-sectional time series database with no missing values. A key objective in the construction of the database is to generate a set of comparable labour force participation rates across both countries and time. With this in mind, the first step in the production of the historical portion of the 2017 Edition of the LFEP database is to carefully scrutinize existing labour force participation rates and to select only those observations deemed sufficiently comparable. A subsequent adjustment is done to the national LFPR data in order to increase the statistical basis (in other words, to decrease the proportion of imputed values): the harmonization of LFPR data by age bands (see Annex 3 for a detailed description). In the second step, a linear interpolation model and a weighted least squares model were developed to produce estimates of labour force participation rates for those countries and years in which no country-reported, cross-country comparable data currently exist.

This section contains two main parts. The first part provides an overview of the selection process of the baseline labour force participation rate (LFPR) data that serve as the key input into the 2017 edition LFEP Database. The section includes the description of the obtention process, a discussion of non-comparability issues that exist in the available LFPR data and concludes with a description of the LFPR data coverage, after taking into account the various selection criteria. The second part describes the econometric model developed for the treatment of missing LFPR values, both in countries that report in some but not all of the years in question, as well as for those countries for which no data are currently available.

3.b. The data obtention and selection process

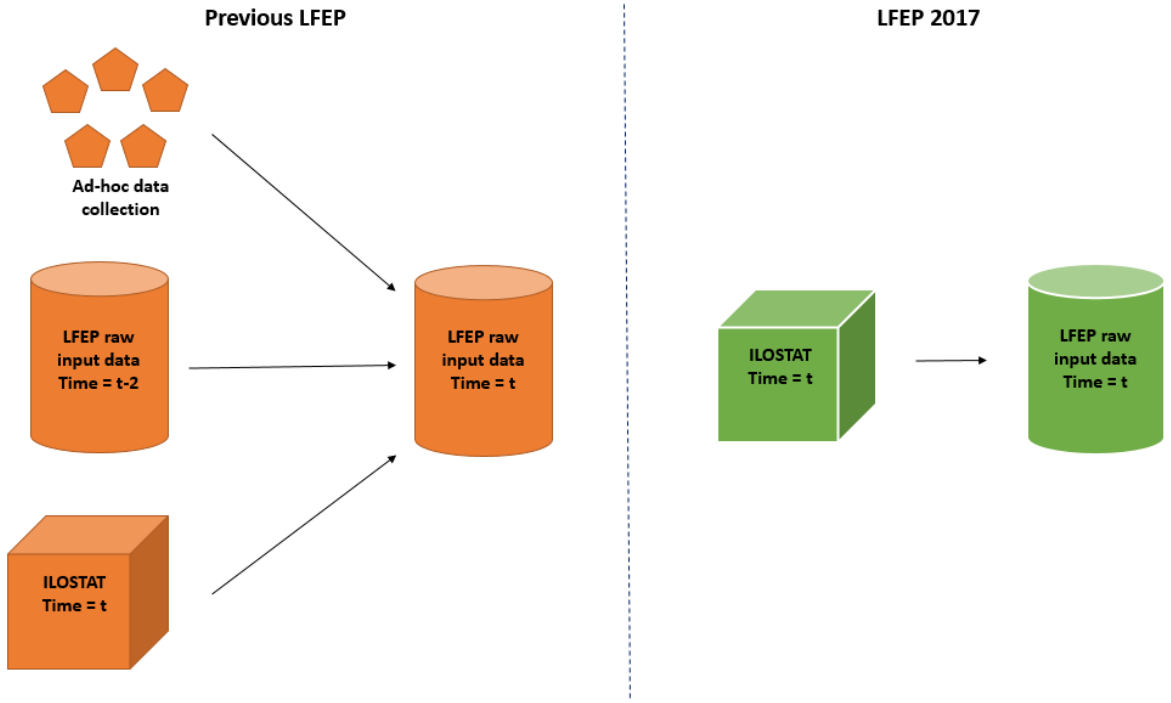
Streamlining the data obtention process

One of the major changes in the 2017 edition of the LFEP model has been streamlining the data collection process. The new set up has one single source of data: the ILOSTAT database. In previous editions, a combination of sources were used: the ILOSTAT database, ad hoc data collection for the LFEP run, and the previous LFEP input data. The procedure used in this edition presents several advantages. First, it is replicable from a methodological standpoint. Downloading the LFPR data from the online database delivers the required data without carrying out any further action. Furthermore, the observations included in ILOSTAT are clearly traced by origin and data collection channel. Second, the labour intensity of the update of the model drastically declines. Not only because the time consuming ad-hoc data collection is avoided, but it eliminates the need to select between up-to-date observations and the data from the previous run. Third, the input data benefits from all the quality control procedures that the ILOSTAT database has in place. This includes the efforts of unreliable observation deletion, error correction, data update, data revision and new data channel creation. Taking advantage of all these procedures significantly increases the quality of the data without any additional burden to carry out the new LFEP run.[11] It is worth highlighting that new ILO micro-data processing efforts as well as new channels for obtaining bulk data has enhanced international comparability of the underlying input dataset. These channels account for almost 50 per cent of the input data in the standard age format. Consequently, the quality gains are substantial. The fourth and final advantage concerns the metadata associated to the LFPR. The metadata system of ILOSTAT is exhaustive and complex. At the same time, the metadata are, as are the data, subject to quality control procedures: error detection, update, revision and

[11] Consider a simple example, the detection of an unreliable data point through one of the appropriate procedures in ILOSTAT that results in the removal from the database of the observation. With the streamlined data obtention this efficiently translates in the removal from the new LFEP edition. In contrast, in the old recursive data collection setting, the deleted observation would appear in the new model version. To detect and correct such inclusion one would have to compare not only all data existing in the previous model run and ILOSTAT, but all previous versions of ILOSTAT.

automatic collection. Taking advantage of this system – as opposed to creating a parallel system within the LFEP process – increases the quality of metadata and hence of decision making regarding the data with little additional effort. The improvements in metadata not only deliver a positive impact during the data obtention step, they facilitate a more replicable and precise data selection step. Figure 2 below illustrates the changes in the data obtention process.

Figure 2: Data obtention process (2017 edition) comparison



Data selection process

In order to generate a set of sufficiently comparable labour force participation rates across both countries and time, it is necessary to identify and address the various sources of non-comparability. Drawing heavily on the labour force participation data comparability discussion in the Key Indicators of the Labour Market (KILM), 9th Edition (Geneva, ILO 2015). The main sources of non-comparability of labour force participation rates are as follows:

Type of source –labour force participation rates are derived from several types of sources including labour force surveys, population censuses, establishment surveys, insurance records or official government estimates. Data taken from different types of sources are often not comparable.

Age group coverage – non-comparability also arises from differences in the age groupings used in measuring the labour force. While the standard age-groupings used in the LFEP database are 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64 and 65+, some countries report data corresponding to other age groupings, which can adversely affect comparability. For example, some countries have adopted a cut-off point at 14 or 16 years for the lower limit and 65 or 70 years for the upper limit.

Geographic coverage – some country-reported labour force participation rates correspond to a specific geographic region, area or territory such as "urban areas". Geographically-limited data are not comparable across countries.

Others – Non-comparability can also arise from the inclusion or non-inclusion of military conscripts; variations in national definitions of the labour force, particularly with regard to the statistical treatment of "contributing family workers" and the "unemployed, not looking for work"; and differences in survey reference periods.

Data selection criteria

Taking these issues into account, a set of criteria was established upon which labour force participation rates would be selected or eliminated from the input file for the LFEP dataset. There are four criteria described hereafter.

Selection criterion 1 (type of source)

Consistent with the discussion concerning the data obtention process, the source selection has been performed based on the best source classification generated by ILOSTAT. This system identifies the optimal type of source for a country and an indicator, in this case, the LFPR. Data is derived from a labour force or household survey, a population census or official government estimates. Labour force surveys are the most comprehensive source for internationally comparable labour force data. National labour force surveys are typically very similar across countries, and the data derived from these surveys are generally much more comparable than data obtained from other sources. Consequently, a strict preference is given to labour force survey data, with population census data included only if no labour force or household survey data exist for a given country. Occasionally a labour force survey coexists with another labour force survey for the same country and period of observation, yet the two surveys do not completely overlap. In the annex a harmonization procedure is described for this case. Due to the need to balance the competing goals of data comparability and data coverage, some population census-based labour force participation rates were included. A similar reasoning, but with more stringent criteria for its use, can be applied to data derived from official government estimates. The methodology for producing official estimates can differ significantly across countries and over time, leading to non-comparability. Nonetheless, they might be used if no other data source is available.

Selection criterion 2 (age coverage)

Age-disaggregated data are included in the initial input file. For example, when the labour force participation rate refers to the total working-age population, this observation is not included. Ideally, the reported rate corresponds to the 11 standardized age-groups (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64 and 65+). For countries with non-standard age-groups, two types of harmonisation have been applied; harmonising the lower- and upper-age limit and harmonising data from large age bands to the above standard 5-year age bands. Detailed descriptions of these harmonisation procedures can be found in Annex 3.

Selection criterion 3 (geographical coverage)

Regarding the geographical selection criteria, data corresponding only to national (i.e. not geographically-limited) labour force participation rates or data adjusted to represent national participation rates are included. Labour force participation rates corresponding to only urban or only rural areas are not comparable across countries. This criterion was necessary due to the large differences that often exist between rural and urban labour markets. Nonetheless, for one country in Latin America (Argentina), the labour force participation rates corresponding to urban areas only were used.

Selection criterion 4 (time series consistency)

The resulting time series for each country and demographic group have been examined for consistency during the whole period 1990-2016. As mentioned above, many types of breaks can occur due to methodological differences. These breaks can generate inexistent trends or excessive volatility. To avoid this, a careful review of both the data and metadata of the whole time series has been carried out, and when necessary the affected data have been deleted.

Resulting input data file

Together, these criteria determined the data content of the final input file, which was utilized in the subsequent econometric estimation process. Table 1 provides response rates and total observations by year. These rates represent the share of total potential (or maximum) observations for which real, cross-country comparable data (after harmonisation adjustments) exist. Compared to the previous editions a natural increase in more recent data can be observed, at the same time, there are substantial decreases in coverage for more distant years. There are two main causes of the decrease. The first is a mechanical loss due to having removed 4 reference areas with a high degree of data coverage.[12] The second is the more selective criteria for data inclusion, both due to ILOSTAT's quality control that affects the data obtention process and more stringent time series consistency checks during the data selection process.

Table 1: Response rates by LFEP version year, both sexes combined

Model:	*2017*		*2015*		*2013*	
Year	Proportion of potential observations (in percentage)	Number of observations	Proportion of potential observations (in percentage)	Number of observations	Proportion of potential observations (in percentage)	Number of observations
1990	23.9	992	27.1	1,149	26.2	1,102
1991	25.4	1,058	28.4	1,206	28.2	1,183
1992	23.4	975	27.0	1,147	26.5	1,112
1993	25.7	1,070	29.0	1,232	28.5	1,198
1994	29.1	1,208	30.3	1,286	29.7	1,248
1995	30.6	1,274	33.8	1,436	33.2	1,393
1996	30.6	1,272	32.6	1,386	31.5	1,323
1997	29.5	1,226	31.7	1,344	31.4	1,321
1998	30.1	1,250	33.6	1,427	33.4	1,403
1999	33.3	1,386	37.0	1,570	36.3	1,526
2000	35.3	1,466	38.5	1,635	38.1	1,600
2001	33.3	1,386	40.1	1,704	39.6	1,664
2002	35.9	1,492	40.0	1,698	39.5	1,658
2003	34.3	1,426	41.9	1,780	41.4	1,738
2004	39.0	1,623	44.9	1,908	44.8	1,884
2005	38.3	1,594	48.0	2,038	47	1,976
2006	40.3	1,674	46.0	1,954	45.5	1,912
2007	40.2	1,670	46.6	1,978	44.7	1,880
2008	40.5	1,682	48.2	2,048	46.2	1,940
2009	50.0	2,078	55.5	2,356	49	2,060
2010	52.0	2,164	55.0	2,336	43.8	1,842
2011	48.4	2,014	50.0	2,122	29.3	1,232
2012	49.8	2,072	50.7	2,153	28.3	1,188
2013	48.5	2,016	50.3	2,136		
2014	49.4	2,056	19.0	806		
2015	47.4	1,969				
2016	38.4	1,595				
Total	37.1	41,688	39.4	41,835	36.6	35,383

In LFEP 2017 the potential number of observations for each year is 4'158 data points (11 age-groups x 189 countries x 2 sexes). Hence, the total potential number of observations during the time period 1990 to 2016 is 112'266 data points.
In LFEP 2015 the potential number of observations for each year is 4'246 data points (11 age-groups x 193 countries x 2 sexes). Hence, the total potential number of observations during the time period 1990 to 2014 is 106'150 data points.
In LFEP (EAPEP) 2013 the potential number of observations for each year is 4'202 data points (11 age-groups x 191 countries x 2 sexes). Hence, the total potential number of observations which covers the time period 1990 to 2012 is 96'646 data points.

[12] The 4 areas have been integrated into the France figures.

In total, comparable data are available for 41,688 out of a possible 112,266 observations, or approximately 37.1 per cent of the total. It is important to note that while the percentage of real observations is rather low, 175 out of 189 countries (92 per cent) reported labour force participation rates in at least one year during the 1990 to 2016 reference period.[13] Thus, some information on LFPR is known about the vast majority of the countries in the sample.

There is very little difference among the 11 age-groups with respect to data availability. This is primarily due to the fact that countries that report LFPR in a given year tend to report for all age groups. On the other hand, there is a clear variation in response by year. In particular, coverage has tended to improve over time, as the lowest coverage occurred in the early 1990s. While the overall response rate is approximately 37 per cent, as will be shown in the next section, response rates vary substantially among the different regions of the world.

3.c. Missing value estimation procedure

Overview

This section describes the basic missing value estimation model developed to produce the historical LFEP database. The methodology contains four steps. First, in order to ensure within-bound estimates of labour force participation rates, a logistic transformation is applied to the input data file. Second, a simple interpolation technique is used to expand the baseline data in countries that report labour force participation rates in some years. Next, the problem of non-response bias (systematic differences between countries that report data in some or all years and countries that do not report data in any year) is addressed and a solution is developed to correct for this bias. Finally, the weighted least squares estimation model, which produces the actual country-level LFPR estimates, is explained in detail. Each of these steps is described below.

Step 1: Logistic transformation

The first step in the estimation process is to transform all labour force participation rates included in the input file. This step is necessary since using simple linear estimation techniques to estimate labour force participation rates can yield implausible results (for instance labour force participation rates of more than 100 per cent). Therefore, in order to avoid out of range predictions, the final input set of labour force participation rates is transformed logistically in the following manner prior to the estimation procedure:

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

where y_{it} is the observed labour force participation rate by sex and age in country i and year t. This transformation ensures within-range predictions, and applying the inverse transformation produces the original labour force participation rates. The specific choice of a logistic function in the present context was chosen following Crespi (2004).

Step 2: Country-level interpolation

The second step in the estimation model is to fill in, through linear interpolation, the set of available information from countries that report in some, but not all of the years in question. In many reporting countries, some gaps in the data do exist. For instance, a country will report labour force participation rates in 1990 and 1995, but not for the years in between. In these cases, a simple linear interpolation routine is applied, in which LFPR estimates are produced using equation 2.

[13] The 14 countries or territories for which no comparable information on labour force participation rate by sex and age is available are: Afghanistan, Central African Republic, Channel Islands, Equatorial Guinea, Eritrea, Guinea-Bissau, Democratic People's Republic of Korea, Libyan Arab Jamahiriya, Solomon Islands, Somalia, Swaziland, Turkmenistan, Uzbekistan, Western Sahara.

$$Y_{it} = \frac{Y_{i1} - Y_{i0}}{t_1 - t_0}(t - t_0) + Y_{i0} \qquad [2]$$

In this equation, Y_{i1} is the logistically transformed labour force participation rate in year t_1, which corresponds to the closest reporting year in country *i* following year *t*. Y_{i0} is the logistically transformed labour force participation rate in year t_0, which is the closest reporting year in country *i* preceding year *t*. Accordingly, Y_{i1} is bounded at the most recent overall reporting year for country *i*, while Y_{i0} is bounded at the earliest reporting year for country *i*.

This procedure increases the number of observations upon which the econometric estimation of labour force participation rates is based. It relies on the assumption that structural factors are predominant as compared to the cyclical and accidental ones.

Table 2: Response rates by estimation group

Estimation group	*Number of observations*	*Number of observations, post-interpolation*	*Proportion of potential observations (%)*	*Proportion of potential observations (%), post-interpolation*
Developed Europe (22 countries)	11,799	11,918	90.3	91.2
Developed Non-Europe (10 countries)	5,478	5,654	92.2	95.2
CEE and CIS (28 countries)	8,240	9,968	49.5	59.9
East and South-East Asia (22 countries)	2,590	4,816	19.8	36.9
South Asia (9 countries)	1,286	3,232	24.1	60.5
Central America and the Caribbean (22 countries)	4,431	7,536	33.9	57.7
South America (9 countries)	3,038	4,608	56.8	86.2
Middle East and North Africa (19 countries)	2,508	5,642	22.2	50.0
Sub-Saharan Africa (48 countries)	2,384	9,136	8.4	32.0
Total (189 countries)	41,754	62,510	37.2	55.7

The potential number of observations for each region is calculated by 11 age-groups x number of countries x 2 sexes x (2016-1990+1).

The increase in observations resulting from the linear interpolation procedure is provided in Table 2. This table also provides a picture of the large variation in data availability among the different geographic/economic estimation groups. In total, the number of observations increased from 41,754 to 62,510 – that is, from 37.2 per cent to 55.7 per cent of the total potential observations. The lowest data coverage is in sub-Saharan Africa, where the post-interpolation coverage is 32.0 per cent. Post-interpolation coverage reaches 91.2 per cent in the Developed Europe region and 95.2 per cent in the Developed Non-Europe region. The resulting database represents the final set of harmonized real and estimated labour force participation rates upon which the multivariate weighted estimation model was carried out as described below.

Step 3: Calculation of response-probabilistic weights

Out of 189 countries in the LFEP database, 14 do not have any reported comparable labour force participation rates over the 1990-2016 period. This raises the potential problem of non-response bias. That is, if labour force participation rates in countries that do not report data tend to differ significantly from participation rates in countries that do report, basic econometric estimation techniques will result in biased estimates of labour

force participation rates for the non-reporting countries, as the sample upon which the estimates are based does not sufficiently represent the underlying heterogeneity of the population.[14]

The identification problem at hand is essentially whether data in the LFEP database are missing completely at random (MCAR), missing at random (MAR) or not missing at random (NMAR).[15] If the data are MCAR, non-response is ignorable and multiple imputation techniques such as those inspired by Heckman (1979) should be sufficient for dealing with missing data. This is the special case in which the probability of reporting depends neither on observed nor unobserved variables – in the present context this would mean that reporting and non-reporting countries are essentially "similar" in both their observable and unobservable characteristics as they relate to labour force participation rates. If the data are MAR, the probability of sample selection depends only on observable characteristics. That is, it is known that reporting countries are different from non-reporting countries, but the factors that determine whether countries report data are identifiable. In this case, econometric methods incorporating a weighting scheme, in which weights are set as the inverse probability of selection (or inverse propensity score), is one common solution for correcting for sample selection bias. Finally, if the data are NMAR, there is a selection problem related to unobservable differences in characteristics among reporters and non-reporters, and methodological options are limited. In cases where data are NMAR, it is desirable to render the MAR assumption plausible by identifying covariates that impact response probability (Little and Hyonggin, 2003).

Given the important methodological implications of the non-response type, it is useful to examine characteristics of reporting and non-reporting countries in order to determine the type of non-response present in the LFEP database. Table 3 confirms significant differences between reporting and non-reporting countries in the sample.

Table 3: Per-capita GDP and population size of reporting and non-reporting countries

	Reporters (175 countries)	***Non-reporters (14 countries)***
Mean per-capita GDP, 2015 (2011 International $)	19504	9024
Median per-capita GDP, 2015 (2011 International $)	13089	2377
Mean population, 2017 (millions)	41.8	9.6
Median population, 2017 (millions)	9.4	4.8

Sources: World Bank, World Development Indicators Database 2017; UN, World Population Prospects 2017 Revision Database.

The table shows that reporting countries have considerably higher per capita GDP and larger populations than non-reporting countries. In the context of the LFEP database, it is important to note that countries with low per-capita GDP also tend to exhibit higher than average labour force participation rates, particularly among women, youth and older individuals. This outcome is borne mainly due to the fact that the poor often have few assets other than their labour upon which to survive. Thus, basic economic necessity often drives the poor to work in higher proportions than the non-poor. As economies develop, many individuals (particularly women) can afford to work less, youth can attend school for longer periods, and consequently, overall participation rates in developing economies moving into the middle stages of development tend to decline.[16]

It appears that factors exist that co-determine the likelihood of countries to report labour force participation rates in the LFEP input dataset and the actual labour force participation rates themselves. The missing data do not appear to be MCAR. Due to the existence of data (such as per-capita GDP and population size) for both responding and non-responding countries and that are related to response likelihood, it should be possible to

[14] For more information, see Crespi (2004) and Horowitz and Manski (1998).

[15] See Little and Hyonggin (2003) and Nicoletti (2002).

[16] See ILO, KILM 7th Edition, (Geneva, ILO, 2011) and Standing, G. *Labour Force Participation and Development* (Geneva, ILO, 1978).

render the MAR assumption plausible and thus to correct for the problem of non-response bias.[17] This correction can be made while using the fixed-effects panel estimation methods described below, by applying "balancing weights" to the sample of reporting countries. The remainder of the present discussion describes this weighting routine in greater detail.

The basic methodology utilized to render the data MAR and to correct for sample selection bias contains two steps. The first step is to estimate each country's probability of reporting labour force participation rates. In the LFEP input dataset, per-capita GDP, population size, year dummy variables and membership in the Highly Indebted Poor Country (HIPC) Initiative represent the set of independent variables used to estimate response probability.[18]

Following Crespi (2004) and Horowitz and Manski (1998), we characterize each country in the LFEP input dataset by a vector (y_{it}, x_{it}, w_{it}, r_{it}), where *y* is the outcome of interest (the logistically transformed labour force participation rate), *x* is a set of covariates that determine the value of the outcome and *w* is a set of covariates that determine the probability of the outcome being reported. Finally, *r* is a binary variable indicating response or non-response as follows:

$$r_{it} = \begin{cases} 1 & reports \\ 0 & does\ not\ report \end{cases} \qquad [3]$$

Equation 4 indicates that there is a linear function whereby the likelihood of reporting labour force participation rates is a function of the set of covariates:

$$r_{it}^{*} = w_{it}^{'}\gamma + \varepsilon_{it} \qquad [4]$$

where a country reports if this index value is positive ($r_{it}^{*} > 0$). γ is the set of regression coefficients and ε_{it} is the error term. Assuming a symmetric cumulative distribution function, the probability of reporting labour force participation rates can be written as in equation 5.

$$P_i = F\left(w_{it}^{'}\gamma\right) \qquad [5]$$

The functional form of *F* depends on the assumption made about the error term ε_{it}. As in Crespi (2004), we assume that the cumulative distribution is logistic, as shown in equation 6:

$$F\left(w_{it}^{'}\gamma\right) = \frac{\exp\left(w_{it}^{'}\gamma\right)}{1+\exp\left(w_{it}^{'}\gamma\right)} \qquad [6]$$

It is necessary to estimate equation 6 through logistic regression, which is carried out by placing each country into one of the 9 estimation groups listed in Table 2. The regressions are carried out for each of the 11 standardized age-groups and for males and females. The results of this procedure provide the predicted response probabilities for each age group within each country in the LFEP dataset.

[17] Indeed, according to Little and Hyonggin (2003), the most useful variables in this process are those that are predictive of both the missing values (in this case labour force participation rates) and of the missing data indicator. Per-capita GDP is therefore a particularly attractive indicator in the present context.

[18] HIPC membership is utilized as an explanatory variable for response probability due to the fact that HIPC member countries are required to report certain statistics needed to measure progress toward national goals related to the program. As a result, taking all else equal, HIPC countries may be more likely to report labour force participation rates.

The second step is to calculate country weights based on these regression results and to use the weights to "balance" the sample during the estimation process. The predicted response probabilities calculated in equation 6 are used to compute weights defined as:

$$s_{it}(w) = \frac{P(r_{it} = 1)}{P(r_{it} = 1 \mid w_{it}, \hat{\gamma})} \qquad [7]$$

The weights given by equation 7 are calculated as the ratio of the proportion of non-missing observations in the sample (for each age group and each year) and the reporting probability estimated in equation 6 of each age group in each country in each year. By calculating the weights in this way, reporting countries that are more similar to the non-reporting countries (based on characteristics including per-capita GDP, population size and HIPC membership) are given greater weight and thus have a greater influence in estimating labour force participation rates in the non-reporting countries, while reporting countries that are less similar to non-reporting countries are given less weight in the estimation process. As a result, the weighted sample looks more similar to the theoretical population framework than does the simple un-weighted sample of reporting countries.

Step 4: Weighted multivariate estimation

The final step is the estimation process itself. Countries are again divided into the 9 estimation groups listed in Table 2, which were chosen on the combined basis of broad economic similarity and geographic proximity.[19] Having generated response-probabilistic weights to correct for sample selection bias, the key issues at hand include 1) the precise model specification and 2) the choice of independent variables for estimating labour force participation.

In terms of model specification, taking into account the database structure and the existence of unobserved heterogeneity among the various countries in the LFEP input database, the choice was made to use panel data techniques with country fixed effects, with the sample of reporting countries weighted using the $s_{it}(w)$ to correct for non-response bias.[20] By using fixed effects in this way, the "level" of known labour force participation rates in each reporting country is taken into account when estimating missing values in the reporting country, while the non-reporting countries borrow the fixed effect of a similar reporter country. The similarity is simply based on economic and social factors, such as per capita GDP and general cultural norms. More formally, the following linear model was constructed (and run on the logistically transformed labour force participation rates):

$$Y_{it} = \alpha_i + x_{it}^{'}\beta + e_{it} \qquad [8]$$

where α_i is a country-specific fixed effect, χ_{it} is a set of explanatory covariates of the labour force participation rate and e_{it} is the error term. The main set of covariates[21] included is listed in Table 4.

[19] Schaible (2000) discusses the use of geographic proximity and socio-economic status to define estimation domains for data estimation including for ILO labour force participation rates. See also Schaible and Mahadevan-Vijaya (2002).

[20] Crespi (2004) provides a test comparing the bias resulting from different missing value estimation models and finds that the weighted least squares model using fixed-effects provides the smallest relative bias when estimating unemployment rates.

[21] Covariate selection was done separately for each of the estimation groups. Full regression results corresponding to the LFEP Version 5 database are published in Kapsos (2007).

Table 4: Independent variables in fixed-effects panel regression

Variable	Source
Per-capita GDP, Per-capita GDP squared Real GDP growth rate, Lagged real GDP growth rate	World Bank, World Development Indicators 2017 and IMF, World Economic Outlook April 2017
Share of population aged 0-14, Share of population aged 15-24, Share of population aged 25-64	United Nations, World Population Prospects 2017 Revision Database

Within the context of the LFEP database, there are two primary considerations in selecting independent variables for estimation purposes. First, the selected variables must be robust correlates of labour force participation, so that the resulting regressions have sufficient explanatory power. Second, in order to maximize the data coverage of the final LFEP database, the selected independent variables must have sufficient data coverage.

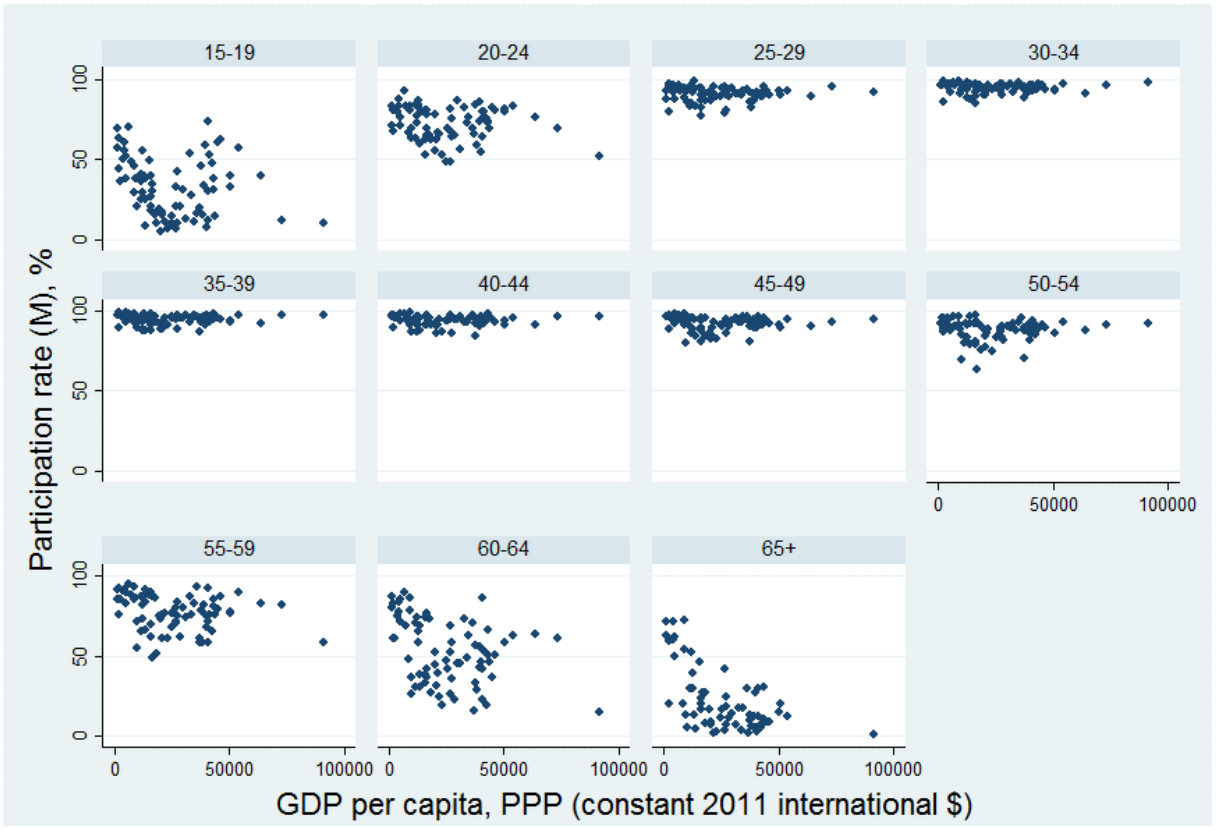
In terms of variables related to economic growth and development, as mentioned above, per-capita GDP is often strongly associated with labour force participation rates.[22] This, together with the substantial coverage of the indicator made it a prime choice for estimation purposes. However, given that the direction of the relationship between economic development and labour force participation can vary depending on a country's stage of development, the square of this term was also utilized to allow for this type of non-linear relationship.[23] Figure 2 depicts clearly that this relationship varies by both sex and age group. For men and women in the prime-age, there is no clear relationship. For both men and women aged 15 to 19 years, and those aged between 55 and 64 years, there is an indication of a U-relationship between per capita GDP and labour force participation rate. Furthermore, annual GDP growth rates were used to incorporate the relationship between participation and the state of the macro-economy.[24] The lag of this term was also included in order to allow for delays between shifts in economic growth and changes in participation.

[22] See also Nagai and Pissarides (2005), Mammen and Paxon (2000) and Clark et al. (1999).

[23] Whereas economic development in the poorest countries is associated with declining labour force participation (particularly among women and youth), in the middle- and upper- income economies, growth in GDP per capita can be associated with rising overall participation rates – often driven by rising participation among newly empowered women. This phenomenon is the so-called "U-shaped" relationship between economic development and participation. See ILO, KILM 4th Edition and Mammen and Paxon (2000).

[24] See Nagai, L. and Pissarides (2005), Fortin and Fortin (1998) and McMahon (1986).

Figure 2: Labour force participation rates by sex and age-group, and per capita GDP, 2006 (78 countries with reported data)



Note: 2006 is selected because it represents a pre-crisis year for which the response rate is the highest.

Changes in the age-structure of populations can also affect labour force participation rates over time. This happens at the country-wide level since different age cohorts tend to have different labour force participation rates, and thus changes in the aggregate age structure of a population can affect the overall participation rate.

Importantly for the present analysis is the potential impact that demographic changes can have on intra age-group participation rates within countries. Changes in population age structure can affect the overall burden of caring for dependents at home, thus affecting individuals' decisions to participate in labour markets. This can have a particularly important effect on women's decisions to enter into work.[25] In order to incorporate these types of demographic effects, the share of the population aged 0-14 (young age-dependent), 15-24 (working-age youth) and 25-64 (prime working age) were incorporated to various degrees in regions in which an important relationship between participation and demographics was found. These variables are by definition correlated and thus increase the presence of multicollinearity in the regressions. However, it was determined that this did not present a prohibitively significant problem in the context of the present estimation procedure.

In all estimation groups, a set of country dummy variables was used in each regression in order to capture country fixed effects. A preliminary examination of the input data revealed that countries in the South Asia estimation group exhibit a particularly large degree of heterogeneity in labour force participation rates, especially with regard to female participation. In order to estimate robust labour force participation rates in non-reporting countries in this estimation group, it was necessary to introduce a dummy variable to further subdivide economies in the region based on observed national labour market characteristics and prevailing cultural norms with regard to male and female labour market participation. This variable was significant in more than 70 per cent of the regressions carried out for the estimation group. Finally, the constant $\alpha_{i,}$ given in equation 8 is country-specific and captures all the persistent idiosyncratic factors determining the labour force participation rate in each country. For 13 out of the 14 countries which do not have any reported comparable labour force participation rates over the 1990-2016 period, the fixed effect of a counterpart economy has been chosen instead of the regional average. The end result of this process is a balanced panel containing real and imputed cross-country comparable labour force participation rates for 189 countries over the period 1990-2016. In the final step, these labour force participation rates are multiplied by the total population figures given in the United Nations World Population Prospects 2017 Revision database, which gives the total labour force in each of the 189 countries, broken down by age group and sex.

Step 5: Data adjustments, blending of model vintages and overfitting

For a few countries, the estimates derived from the weighted panel model (see the previous section) have been adjusted when the estimates were not judged to be realistic by analogy to real data observed in similar countries. This has happened in some countries with a very volatile GDP over time. On occasions, the volatility has its cause on a strong economic dependence on oil. In other cases, it is due to episodes of war. Within this context, the trends derived from the panel model are at times too volatile. In these cases, GDP data has been smoothed and modified to avoid the propagation of this excessive volatility into the estimates of LFPR.

In the previous edition of the LFEP model, a blending of old and new estimates was carried out. This entailed a careful inspection of both sets of results and if deemed necessary, a convex combination of the two observations. This procedure affected 44 countries. The motivation behind this exercise was to avoid large variations in the estimates. In the context of a model with a large number of parameters to be estimated as is the LFEP case, large swings in results can be a clear sign of overfitting. As such, limiting the variability can serve as a sort of regularization procedure to avoid the negative consequences of this phenomenon. It must be noted, however, that the procedure introduces a bias towards previous results and data. Furthermore, there is the risk of incurring the costs of overfitting without some of the alert signs. In the 2017 edition of the LFEP, there has not been any case where such an exercise has been deemed necessary. This is a consequence of the priority given to the time series consistency of the input LFPR data.

[25] Bloom and Canning (2005), Falcão and Soares (2005), O'Higgins (2003), Clark et al. (1999), Fullerton (1999) and McMahon (1986) provide some examples of the relationships between population structure (and demographic change) and labour force participation rates for different groups of the population.

4. Projection model: 2017-2030

4.a. Methodologies used worldwide

In 2011, the ILO Department of Statistics undertook a literature review of all the labour force projection models used and developed by national statistical offices and international organizations.[26] Notably, this document contains a description of the approach adopted by each national or international institution. A two-page template has been used to describe the following aspects: the name of the institution, the frequency of updates and projection horizon, a brief description of the current methodology, the determinants that are captured explicitly, the use (or not) of scenarios, the assessment (or not) of the current methodology, the existence of any previous methodology, reference papers and additional comments. Some major findings of this literature review are presented in this section. To project LFPR, four types of approaches have been identified in this document:

(A) Judgmental (or qualitative) methods based on scenarios or on the targets to be reached.

(B) Time extrapolation models or growth curves. Values for the measured variable can be expressed as a function of time and extrapolated over the projection period. There are many growth curves routinely used in the analysis of growth processes that ultimately reach a steady state. These generally form a class of s-shaped or sigmoid curves, of which the most commonly used is the logistic curve. These sigmoid curves are very useful for modelling populations, labour force participation rates, inflation, productivity growth (not levels) or other processes where, in the long run, it is expected that the variable will not grow any further.

(C) Regression models based on correlations between participation rates and economic, demographic or cultural factors. A regression model with a set of explanatory variables is fitted on observed LFPR. Future scenarios for the explanatory variables are determined and used in the regression model to project LFPR.

(D) Models based on a cohort approach. In this case, LFPRs are not projected by age and sex year after year, but they are projected from the estimated probability of entry or exit of the labour force for each age, sex and cohort (people born in a specific year). More specifically, the probability of entry and exit of the labour force are kept stable at the last observed value or are extrapolated over the projection period for each population cohort.

Table 5 lists the type(s) of methodology used by different national and international institutions as of June 2011. It can be seen that judgmental and extrapolation methods are used most frequently. The main reason is of a practical nature: these methods can be implemented relatively easily. The other approaches are more time-consuming. Regression models are often statistically complex; they can be "heavy users" of historical and projected data. They rely on the accuracy of projected explanatory variables, the choice of which may be subjective and not always optimal. Cohort based-models require historical data over a long period to be implemented. Ideally, it should be pure longitudinal data (the same people surveyed year after year) but in fact, most of the projections are based on annual surveys based on different surveyed households. In addition, statistical procedures for projecting the cohorts' rates of entry or exit of the workforce quickly become complicated.

[26] For more details see Houriet-Segard and Pasteels (2011).

Table 5: Summary of projection methods used worldwide, as of June 2011

Type of projections /Projection models	Judgmental approach (target or scenarios)	Time extrapolation approach	Regression approach	Cohort based approach	Additional modules
ILO (EAPEP edition 5)	X	X			
Algeria	X				
Asian Development Bank			X		
Australia Bureau of Statistics		X			
Australia GPG				X	
Bolivia	X				
Canada		X			
CELADE	X				
European Central Bank			X	X	
European Commission				X	X
EUROSTAT	X				
France		X	X		X
Haiti	X				
Hong Kong		X			
Ireland	X				
Mexico				X	
New Zealand	X				
OECD				X	X
Singapore		X			
Spain			X	X	
Sri Lanka	X				
Switzerland			X		
Tunisia	X				
United Kingdom		X	X		
USA		X			

Source: Houriet-Segard and Pasteels (2011).

Time extrapolation methods are statistically easy to develop and only need a consistent time series. However, they suffer from several drawbacks. Firstly, they only extrapolate past patterns without being able to project changing trends in the future. Fundamentally, these models are meant to implicitly capture all demographic, economic and cultural factors affecting LFPR.

4.b. Methodology used in this edition

In this edition, the projections for the period from 2017 to 2030 are derived exclusively using **econometric methods**. The forecast horizon is 14 years, as opposed to the 36 of the previous LFEP edition. The reduction of the forecast period is due to the exclusive use of econometric techniques and the underlying uncertainty under longer horizons. As defined in the paper by Müller and Watson (2011), forecasting 14 years ahead (2030) based on a sample of 26 years (1990-2016) constitutes a long-horizon problem already. The horizon represents a large proportion of available observations: 53%. In addition, many of the data used are imputed, therefore it is prudent to limit the forecast at the 2030 horizon.

There are several advantages in using exclusively statistical methods. Given the large number of independent time series to produce (11 age groups x 2 sexes x 189 reference areas) the choice of an econometric model to generate the forecasts ensures the following characteristics. First, the transparency of the results; as the model used is the same for all the time series the causes underlying each particular forecast are well justified and clear. Second, the simplicity of the update process; with minor adjustments, the statistical model can be applied to the subsequent updates of the underlying data. Finally, the consistency of both methods and results across editions of the LFEP. The approach used guarantees that a similar set of assumptions and techniques are used across editions.

Given the sensitivity of the data as well as to take advantage of the expert knowledge regarding labour market indicators, a judgmental evaluation of the results is done. This evaluation has been carried out across all independent time series, 4,158, as well as the derived totals, 378, analysing the plausibility of the forecasted LFPR. After the careful consideration of each of the produced series, none of the examined series has been deemed necessary to adjust. In the previous edition judgemental adjustments were widely used, affecting approximately forty per cent of the projections.

Projections

For each time series, three projections are derived from different extrapolative models.

The three projections are based on the following scenarios:

(1) **Constant scenario**. Labour force participation rates at the horizon 2030 are kept at the level that was observed in 2016 for each time series (at the 5 years age-band level).

(2) **Imputation scenario**. Labour force participation rates are obtained for the whole period 2017-2030 using the imputation procedure described in the previous section (at the 5-year age-group level).

(3) **Logistic trend scenario**. Labour force participation rates are extrapolated at the horizon 2030 using different variants of logistic trends for each time series.

The **constant scenario**, which could also be named a “naive projection”, is not a likely outcome but serves the purpose of a reference scenario: the resulting labour force size and composition are purely driven by demographic changes (by the composition of the population, not by changes in participation rates). This scenario is very helpful in assessing what is the total **demographic impact** on participation rates for the total working-age population (15 years and above) when keeping the participation rates constant at the 5-year age band level. This effect can be very important in a context of ageing populations. For example, in Belgium, the male LFPR of the working age population (15 years and above) would drop by more than five percentage points (from 59.0 to 53.8 per cent) between 2016 to 2030 in the constant scenario, as a result of population ageing.

The **imputation scenario** simply follows the methodology described for imputing missing values. The same procedure that is used to fill in gaps in the 1990-2016 can be used to project a LFPR from 2017 to 2030. The demographic data provided by the UN WPP 2017 spans the whole period. Similarly, the GDP data are based on the growth projections from IMF's WEO until 2022 and extrapolated based on the average growth forecast during the 2023-2030 period. The imputation methodology is developed precisely to provide an approximation for missing values, and not as a forecasting tool. However, the resulting data can be of interest if one is interested in the impact of GDP and demographic data on LFPR for the coming years, particularly in the medium term.

In the **logistic trend scenario**, the assumption is that future labour force participation rates will follow past trends and is described in detail below. The three scenarios produced are highly complementary as they are based on completely different assumptions.

Different variants of the logistic trends have been used, consistent with the previous edition of this database (Pasteels, 2012). The parametric form for the basic model is linear but fitted to the logit (logistic transformation) of the proportion participating, scaled to fit between the values y_{min} and y_{max} (the asymptotes) determined for each age-sex group in a separate step. In this model, the participation rate y_t at time t is then given by

$$y_t = y_{min} + \frac{y_{max} - y_{min}}{1 + e^{a+bt}} \quad [1]$$

The transformed variable y'_t , defined as

$$y'_t = \frac{y_t - y_{min}}{y_{max} - y_{min}} \quad [2]$$

is equal to the following expression: $y'_t = \frac{1}{1 + e^{a+bt}}$ [3]

Then, the transformed variable Y'_t, is defined as the logistic transformation of y'_t:

$$Y'_t = \ln(y'_t / (1 - y'_t)) \quad [4]$$

It can also be shown that $Y'_t = -(a + bt)$. Consequently, the parameters a and b can be estimated by running a linear regression on Y'_t.

A special case is when $y_{min} = 0$ and $y_{max} = 1$ (the participation rates can vary between 0 and 100 per cent). In this case, $y_t = y'_t$ and Y'_t corresponds to the same logistic transformation used in the estimation phase (see section 2).

This basic model was used in the previous editions of the projections (see ILO, 2009 and Pasteels 2012). The way to define the asymptotes is the same as in the 2013 edition. It is a very convenient model, which combines the advantages of a logistic curve without suffering from its drawbacks.

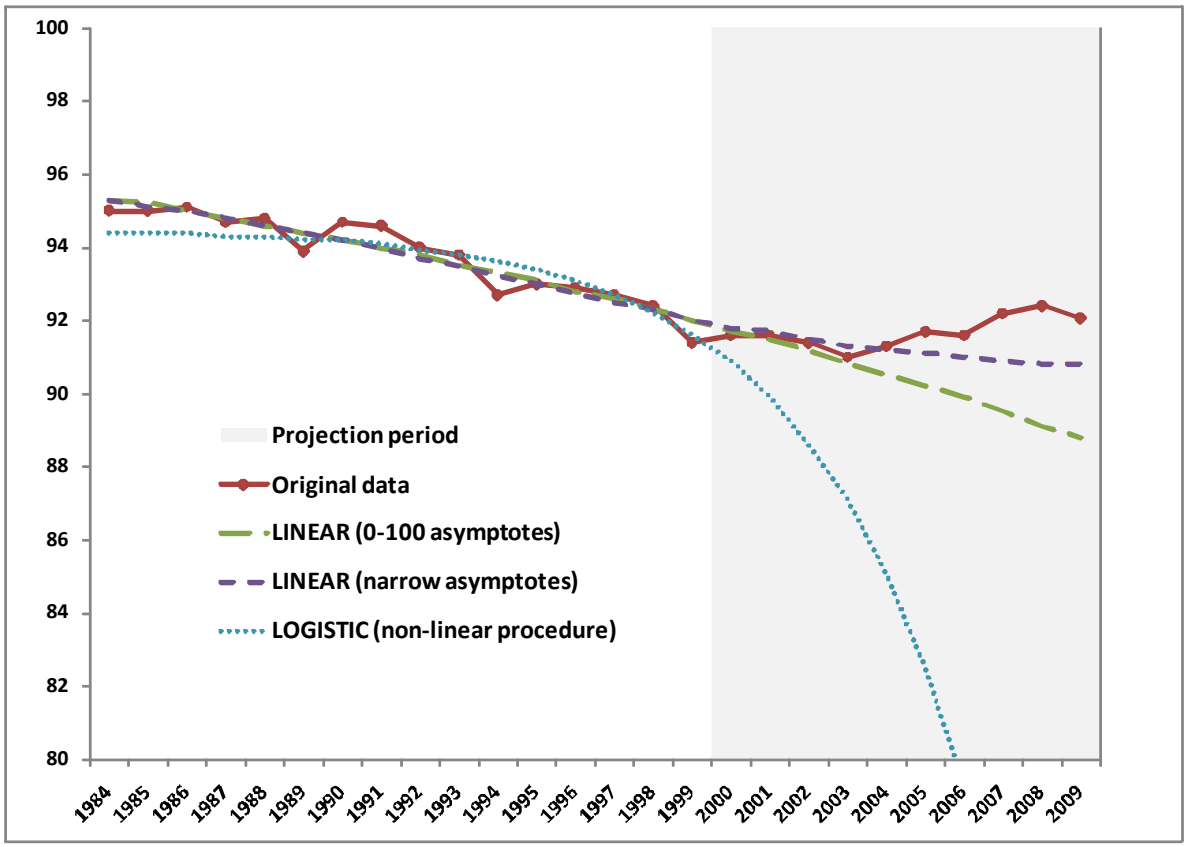
The main advantage of the logistic curve and other sigmoid or S-shaped curves is that they can capture growth processes that ultimately reach a steady state. These curves are frequently used for modelling populations and labour participation rates.

The S-shaped curves, however, are not very easy to estimate. The logistic curve can be estimated using non-linear least squares and maximum likelihood techniques. However, there are often problems in convergence

and sometimes convergence cannot be achieved. In addition, imprecise estimates are obtained if the data do not clearly include an inflexion point, i.e., the time at which the absolute value of the growth rate is maximised (for more details see Kshirsagar & Smith 1995). Estimating a logistic function on such data without imposing an assumption about when the inflexion will occur will sometimes give implausible results.

Figure **3** shows an example of unexpected projections of male LFPR obtained from a logistic curve.

Figure 3: Australia: Male LFPR (age group: 35-39). Projections based on growth curves vs real data



With this basic model, the method for defining the asymptotes is crucial. In the 2009 edition (see ILO 2009) the asymptotes were determined by looking jointly at the patterns of male and female participation rates for the same age group. The main assumption being that the past convergence or divergence between male and female LFPR would continue in the future at the same pace as over the last ten years. This approach has been abandoned as it resulted in too many unexpected results (e.g., strong decreases in male LFPR in the prime age, increased divergence between female and male LFPR). More fundamentally, the assumption of continued joint divergence or convergence of male and female LFPR is not fully justified theoretically, as many of the determinants of male and female LFPR differ and the two rates may often move independently.

In this edition, the asymptotes are estimated statistically in a first phase for each of the 22 subgroups of the population, without imposing any constraint between male and female participation rates (see also Pasteels 2012).

For each time series, the asymptotes y_{min} and y_{max} are defined as:

$$y_{min} = \min(y_t) - \bar{\varepsilon} \quad \text{and} \quad y_{max} = \max(y_t) + \bar{\varepsilon} \qquad [5]$$

Where $min(y_t)$ and $max(y_t)$ are respectively the historical minimum and maximum of the time series y_t, while $\bar{\varepsilon}$ is set as the median volatility of the time series over 14 year time intervals. This period corresponds to the projection horizon, from 2017 to 2030. The volatility is calculated for each time series as follows:

$$\varepsilon = \sum_{t=t_0+14}^{t_1} \left| y_t - y_{t-14} \right| \qquad [6]$$

Where t_0 is the first year and t_1 the last year. Another interpretation of ε is that it represents the average absolute error of the naive method at the horizon of 14 years. The value of ε has been estimated for each of the 22 subgroups of the population and a sample of 88 countries for which consistent historical data exist. Then the median value $\bar{\varepsilon}$ is obtained across country and year.

Table 6 displays the estimated values of $\bar{\varepsilon}$ expressed in percentage points of participation rates. For example, the value of $\bar{\varepsilon}$ is estimated at around 1.2 percentage points for the LFPR of males aged 35 to 39. This means that the LFPR is extremely stable over time for that subgroup of the population and is expected to not deviate more than 1.2 percentage points above its maximum past value and below its minimum past value in the next 14 years. As opposed to males in the prime age, the LFPR for the youngest (15-24) and oldest (55 and above) age cohorts are much more volatile over time. The same pattern concerns female LFPR, for many age groups.

Table 6: Estimated values of $\bar{\varepsilon}$ (measure of volatility over time)

Age group	*Male*	*Female*
15-19	7.8	6.6
20-24	6.2	5.1
25-29	2.2	5.6
30-34	1.4	6.9
35-39	1.2	5.9
40-44	1.3	5.4
45-49	1.3	7.4
50-54	2.0	11.1
55-59	5.0	15.3
60-64	7.8	9.5
65+	3.9	1.7

For the age group 65+ due to excessively tight asymptotes, the value is forced to 5. The volatility of these series is dragged downward by developed economies and earlier years, which is why it is corrected. Female participation rates are further reduced by factors of 0.75 and 0.6, for countries with disparities between 30 and 40, or more than 40 percentage points respectively between male and female participation rates.

The final value of the logistic trend projection is obtained by combining the projections obtained by two variants of equation [p.1].

The two models are:

(i) Basic trend model at the 5-year age-band level (equation [1]),

(ii) Basic trend model at the aggregate age-band level (equation [1])

Model (ii) is similar to model (i) but what changes is simply the time series that is modelled. The principle is to undertake modelling for a larger subgroup of the population (eg. 15-24), to generate projections at that level and to use those intermediate projections to derive the projections at the 5-year age-band level (using simple regression). The underlying assumption of model (ii) is that the future trends are expected to be the same for some age groups within the population (eg. 15-24). Four larger age bands ([15-24], [25-39], [40-54] and [55+]) are used. The prime age-group [25-54] has been subdivided into two groups in order to take into account of the impact of maternity on LFPR.

In order to increase the accuracy of projections, the two projections are combined using a simple average.

Combining forecasts is an approach frequently used by forecast practitioners. Many empirical studies have been undertaken on the subject. Notably, see the reviews undertaken by Clemen (1989) and Armstrong (2001). As stated by Armstrong (2001), "*Combining forecasts is especially useful when you are uncertain about the situation, uncertain about which method is most accurate, and when you want to avoid large errors. Compared with errors of the typical individual forecast, combining reduces errors".*

It is also worth mentioning that there is a debate between practitioners and researchers. As summarised by Armstrong (2001), "*Some researchers object to the use of combining. Statisticians object because combining plays havoc with traditional statistical procedures, such as calculations of statistical significance. Others object because they believe there is one right way to forecast. Another argument against combining is that developing a comprehensive model that incorporates all of the relevant information might be more effective... Despite these objections, combining forecasts is an appealing approach. Instead of trying to choose the single best method, one frames the problem by asking which methods would help to improve accuracy, assuming that each has something to contribute. Many things affect the forecasts and these might be captured by using alternative approaches. Combining can reduce errors arising from faulty assumptions, bias, or mistakes in data.*"

The different empirical experiments also reveal that combining forecasts improves accuracy to the extent that the individual forecasts contain useful and independent information. Ideally, projection errors would be negatively related so that they might cancel each other out. In practice, however, projections or forecasts are almost always positively correlated.

As illustrated in Pasteels (2012), the results from *ex-post* (after the action) simulations indicate non-negligible gains in projection accuracy when combining projections of participation rates.

Computation of intermediate values

Once the final projections at the horizon 2030 are computed, the intermediate values between the projection origin (2016) and the horizon (2030) are computed assuming a linear pattern.

The linear interpolation implicitly assumes that variations of LFPR due to changes in the business cycle are not taken into account. In other words, the ILO projections are designed to minimise the error projections at the horizon 2030. They are determined by labour supply factors and capture essentially structural and demographic determinants. Similar approaches are adopted by regional and national institutes that publish labour force projections which usually do not provide projections for intermediate years between the projection origin and horizon. For example, the US Bureau of Labor Statistics (BLS 2012) built its 2020 projections on the assumption of a full employment economy in 2020 and does not provide intermediate values between 2010 and 2020.

This limitation should be kept in mind when interpreting the data, especially for years that are characterised by turning points and for subgroups of the population, such as youth (15-24), whose participation in the labour market is very sensitive to the business cycle.

5. Strengths, limitations and future work

For the sake of continuous improvement, the strengths and weaknesses of this edition are described hereafter.

5.a. Strengths

(i) The data obtention process has been fully integrated with the ILOSTAT database structure, thus taking advantage of ILOSTAT's quality control system.

(ii) The data have been harmonised and are more comparable across countries. The consistency of participation rates by sex and age group has been checked systematically.

(iii) Detailed metadata for the estimates is provided for each data point.

(iv) The projections are produced solely based on econometric methods.

5.b. Limitations

The limitations specific to the estimates are the following:

(a) The proportional adjustments made in order to harmonise the data have some limitations since, in the absence of complete information over time, some simplifying assumptions regarding constant patterns over time have to be made. An example of this limitation is the following. In the case of harmonisation adjustments of age band 16-19, applying the proportional adjustment based on data from the United States for other countries supposes a priori similar patterns between the United States and those countries. However, the proportion of 15-year-olds that work in a given country as compared to that country's 16-19 year age-group, may differ vis-à-vis the United States, for instance, due to different policies regarding apprenticeships and other work-study programmes.

(b) Some adjustments have not been made in this edition. Cases in point are the exclusion of military forces and prisoners from the surveys, the use of geographically restricted data, the use of different types of sources, and the use of different reference periods (the whole year, a quarter or month in particular, etc.).

(c) The linear interpolation and the weighted panel regressions are not based on the same assumptions. The linear interpolation assumes that the changes in LFPR due to cyclical and accidental factors are negligible compared to structural ones. Furthermore, it assumes that a linear function is an optimal choice.

(d) Concerning the panel regressions, there are many regression models (11 age groups x 2 sexes x 9 regions). It is time-consuming to modify the models and there are not enough data for some regions (eg. Sub-Saharan Africa), reducing confidence in the derived estimates. In particular, there are reasonable concerns about overfitting. In addition, GDP per capita as an explanatory variable is too volatile for a few economies that rely strongly on oil and/or other mineral commodities. In this context, it becomes a poor proxy for what it is meant to capture (wealth per inhabitant, the existence of social security schemes, access to education, etc.).

(e) The regressions are run on the sample of historical data plus the interpolated data.

The limitations specific to the projections are the following:

(a) Extrapolative models have well-known shortcomings as described in the previous section. Additionally, no measure of uncertainty is obtained.

(b) There is a limitation on the intermediate values between the projection origin (2016) and the horizon (2030) that are computed assuming a linear pattern. This linear interpolation implicitly ignores variations of LFPR due to changes in the business cycle. This limitation should be kept in mind when interpreting the data, especially for years that are characterised by turning points and for subgroups of the population, such as youth whose participation in the labour market is very sensitive to the business cycle. On the other hand, given the context of long-horizon forecasting, the linear approximation pins down the average forecast for the projection period which has better known statistical properties than point estimates[27].

5.c. Direction for future work

Future work should address each of the above-mentioned limitations, bearing in mind the costs and benefits of each improvement. For some countries for which no data are available, micro-datasets of other countries could be used as an alternative to econometric models (e.g., using micro-data from some regions in Pakistan for estimating LFPR in Afghanistan).

Regarding the estimation phase, the number of estimation models could be reduced and dummy variables for sub-regions and for countries with similar labour market characteristics (e.g., in terms of share of immigrant workers, the proportion of the female population in the prime age, oil-based economies) could be added. Ideally, the regressions would be estimated only on the basis of historical data. Finally, additional structural variables could be considered (literacy rates, life expectancy, educational variables), provided that complete time series are available.

Regarding the projection phase, it would be convenient to add a measure of uncertainty.

[27] See Müller and Watson (2011).

6. Bibliography

Armstrong J.S. 2001. "Principles of Forecasting: A Handbook for Researchers and Practitioners". Kluwer Academic Publishers, 2001.

Bloom, D. and Canning, C. 2005. “Global demographic change: Dimensions and economic significance”. Harvard Initiative for Global Health, Working Paper No. 1.

BLS 2012. Employment Projections 2010-2020. News Release, February 2012, Bureau of Labor Statistics.

Clark, R., York, A. and Anker, R. 1999. “Economic development and labour force participation of older persons”. Population Research and Policy Review, 18: 411-432.

Clemen, D. R. T. 1989. “Combining forecasts: A review and annotated bibliography”. International Journal of Forecasting, 5, 559-583.

Crespi, G. 2004. “Imputation, estimation and prediction using the Key Indicators of the Labour Market (KILM) data set”. Employment Strategy Paper 2004/16, (Geneva, ILO).

EC 2011. The 2012 Ageing Report: Underlying Assumptions and Projection Methodologies. *European Economy 4|2011*.

EC 2014. The 2015 Ageing Report: Underlying Assumptions and Projection Methodologies. *European Economy 8|2014*.

Falcão, B. and Soares, R. 2005. “The demographic division and the sexual division of labor”. CDDRL, Stanford Institute on International Studies, No. 50, September, 2005.

Fortin, M. and Fortin, P. 1998. “The changing labour force participation of Canadians, 1969-1996: Evidence from a panel of six demographic groups”. Applied Research Branch, Strategic Policy, Human Resources Development Canada, October 1999.

Fullerton, H. 1999. “Labor force participation rates: 75 years of change, 1950-98 and 1998-2025”. Monthly Labor Review, Dec. 1999. US Bureau of Labor Statistics, Washington, DC.

Heckman J. 1979. “Sample selection as a specification error”, Econometrica, vol. 47, 1, 153-161.

Houriet-Segard G. and Pasteels J.M., 2011. Projections of Economically Active Population. A Review of National and International Methodologies. *ILO Department of Statistics, Working Paper* 3, December 2011.

Horowitz J.L., and Manski C.F. 1998. “Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputation”, Journal of Econometrics, vol. 84, 37-58.

Hussmanns R., Mehran F., Varmā V. 1990. *Surveys of economically active population, employment, unemployment, and underemployment: An ILO manual on concepts and methods*. International Labour Organization, Geneva.

Jaumotte J. 2003. Female labour force participation: past trends and main determinants in OECD countries. OECD Economics Department Working Papers, No. 376.

Kapsos S. 2007. “World and regional trends in labour force participation: Methodologies and key results”, Economic and Labour Market Papers 2007/1, Geneva, ILO.

Kshirsagar A. M., Smith W.B. 1995. "Growth curves". Volume 145 of Statistics, textbooks and monographs.

ILO, Key Indicators of the Labour Market (KILM) 6th Edition. CD-ROM. Geneva, ILO. ILO, Laborsta Yearly Statistics 2009. Department of Labor Statistics, Geneva, ILO.

ILO 2009. ILO Estimates and projections of the economically active population: 1980-2020 (Fifth edition). Methodological description.

International Monetary Fund. 2005. World Economic Outlook Database. September 2005.

Little, R. and Hyonggin, A. 2003. “Robust likelihood-based analysis of multivariate data with missing values”. University of Michigan Department of Biostatistics Working Paper Series 2003/5.

Loichinger E. 2015. “Labor force projections up to 2053 for 26 EU countries, by age, sex, and highest level of educational attainment”. Demographic Research, vol. 32, article 15, pp. 443−486.

Madouros V. 2006 "Labour Force Projections 2006-2020", Office for National Statistics (UK), 39 p.

Mammen, K. and Paxon, C. 2000. “Women’s work and economic development”. Journal of Economic Perspectives, Volume 14: Number 4, 141-164.

Mc Morrow, K. and W. Roeger, 1999. "The economic consequences of ageing populations (a comparison of the EU, US and Japan)." European Economy. Economic Papers Nr. 138. European Commission, Economic and Financial Affairs.

McMahon, P. 1986. “An international comparison of labour force participation, 1977-84”Monthly Labor Review, Vol 109, No. 5.

Müller, U. and Watson, M. “Measuring Uncertainty about Long-Run Predictions”. Review of Economic Studies (2011) 83, 1711–1740. Oxford University Press.

Ngai, L. and Pissarides, C. 2005. “Trends in labour supply and economic growth”. Center for Economic Performance, London School of Economics.

Nicoletti, C. 2002. Correcting for sample selection bias: Alternative estimators Compared. ISER, University of Essex.

O’Higgins, N. 2003. “Trends in the youth labour market in developing and transition countries”. Social Protection Discussion Paper Series No. 321. World Bank, 2003.

OECD 2010. The Impact of the Economic Crisis on Potential Output. Working Party No. 1 on Macroeconomic and Structural Policy Analysis ECO/CPE/WP1(2010)3. P.13.

OECD 2013. Pensions at a Glance 2013. OECD and G20 indicators.

Pasteels J.M. 2012. ILO Projections of the Economically Active Population. Revised Methodology of the 2011 Edition. *ILO Department of Statistics Working Paper 4, April 2012*.

Scheible, W. 2000. “Methods for producing world and regional estimates for selected key indicators of the labour market”.

Scheible, W. and Mahadevan-Vijaya, R. 2002. “World and regional estimates for selected key indicators of the labour market”, ILO Employment Paper 2002/36. Geneva, ILO

Standing, G. 1978. Labour Force Participation and Development. Geneva, ILO.

United Nations. World Population Prospects: 2017 Revision Population Database.

World Bank, World Development Indicators Database.

ANNEX 1: Country composition of each sub-regional grouping

AFRICA

Eastern Africa	Burundi ; Comoros ; Djibouti ; Eritrea ; Ethiopia ; Kenya ; Madagascar ; Malawi ; Mauritius ; Mozambique ; Rwanda; Somalia ; Tanzania, United Republic of; Uganda ; Zambia ; Zimbabwe
Middle Africa	Angola ; Cameroon ; Central African Republic ; Chad ; Congo ; Congo, Democratic Republic of ; Equatorial Guinea ; Gabon ; Sao Tome and Principe
Northern Africa	Algeria ; Egypt ; Libyan Arab Jamahiriya ; Morocco ; Sudan ; South Sudan ; Tunisia ; Western Sahara
Southern Africa	Botswana ; Lesotho ; Namibia ; South Africa ; Swaziland
Western Africa	Benin ; Burkina Faso ; Cape Verde ; Côte d'Ivoire ; Gambia ; Ghana ; Guinea ; Guinea-Bissau ; Liberia ; Mali ; Mauritania ; Niger ; Nigeria ; Senegal ; Sierra Leone ; Togo

NORTHERN AMERICA	Canada ; United States

LATIN AMERICA AND THE CARIBBEAN

Caribbean	Aruba ; Bahamas ; Barbados ; Cuba ; Dominican Republic ; Grenada ; Haiti ; Jamaica; Puerto Rico ; Saint Lucia ; Saint Vincent and the Grenadines ; Trinidad and Tobago ; Virgin Islands (US)
Central America	Belize ; Costa Rica ; El Salvador ; Guatemala ; Honduras ; Mexico ; Nicaragua ; Panama
South America	Argentina ; Bolivia ; Brazil ; Chile ; Colombia ; Ecuador; Guyana ; Paraguay ; Peru ; Suriname ; Uruguay ; Venezuela, Bolivarian Rep. of

ASIA

Eastern Asia	China ; Hong Kong, China ; Taiwan, China ; Japan ; Korea, Dem. People's Rep. of ; Korea, Republic of ; Macau, China ; Mongolia
South-Central Asia	Afghanistan ; Bangladesh ; Bhutan ; India ; Iran, Islamic Rep. of ; Kazakhstan ; Kyrgyzstan ; Maldives ; Nepal ; Pakistan ; Sri Lanka ; Tajikistan ; Turkmenistan ; Uzbekistan
South-Eastern Asia	Brunei Darussalam ; Cambodia ; Indonesia ; Lao People's Dem. Rep. ; Malaysia ; Myanmar ; Philippines ; Singapore ; Thailand ; Timor-Leste ; Viet Nam
Western Asia	Armenia ; Azerbaijan ; Bahrain ; Cyprus ; Georgia ; Iraq ; Israel ; Jordan ; Kuwait ; Lebanon ; Oman ; Qatar ; Saudi Arabia ; Syrian Arab Republic ; Turkey ; United Arab Emirates ; West Bank and Gaza Strip ; Yemen, Republic of

EUROPE

Eastern Europe	Belarus ; Bulgaria ; Czech Republic ; Hungary ; Moldova, Republic of ; Poland ; Romania ; Russian Federation ; Slovakia ; Ukraine
Northern Europe	Channel Islands ; Denmark ; Estonia ; Finland ; Iceland ; Ireland ; Latvia ; Lithuania ; Norway ; Sweden ; United Kingdom
Southern Europe	Albania ; Bosnia and Herzegovina ; Croatia ; Greece ; Italy ; Macedonia, The former Yugoslav Rep. of ; Malta ; Montenegro ; Portugal ; Serbia ; Slovenia ; Spain
Western Europe	Austria ; Belgium ; France ; Germany ; Luxembourg ; Netherlands ; Switzerland

OCEANIA

Australia-New Zealand	Australia ; New Zealand
Melanesia	Fiji ; New Caledonia ; Papua New Guinea ; Solomon Islands ; Vanuatu
Micronesia	Guam
Polynesia	French Polynesia ; Samoa ; Tonga

ANNEX 2: Tables of regression specifications by region, sex and age group

Females											
1 Developed Europe											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x	x	x	x	x	x	x	x	x	x	x
Real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Lagged real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.911	0.900	0.878	0.864	0.884	0.909	0.908	0.925	0.913	0.888	0.862
2 Developed Non-Europe											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x	x	x	x	x	x	x	x	x	x	x
Real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Lagged real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.964	0.731	0.836	0.880	0.882	0.918	0.932	0.912	0.852	0.710	0.716
3 CEE and CIS											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x	x	x	x	x	x	x	x	x	x	x
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.749	0.782	0.855	0.861	0.856	0.972	0.968	0.966	0.918	0.967	0.917
4 East & South-East Asia (& the Pacific)											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x	x	x							x	x
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.941	0.971	0.980	0.982	0.985	0.981	0.976	0.960	0.958	0.966	0.913
5 South Asia											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP											
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 15-24	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 25-64	x	x	x	x	x	x	x	x	x	x	x
dummy	x	x	x	x	x	x	x	x	x	x	x
R-squared	0.939	0.953	0.976	0.967	0.975	0.985	0.982	0.960	0.963	0.969	0.959

Females											
6 Central America and the Caribbean											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 15-24	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 25-64	x	x	x	x	x	x	x	x	x	x	x
R-squared	0.913	0.934	0.807	0.857	0.847	0.850	0.865	0.857	0.880	0.899	0.931
7 South America											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 15-24	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 25-64	x	x	x	x	x	x	x	x	x	x	x
R-squared	0.848	0.827	0.957	0.950	0.953	0.972	0.901	0.871	0.814	0.852	0.899
8 Middle East & North Africa											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x	x	x	x	x	x	x	x	x	x	x
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.918	0.909	0.923	0.929	0.908	0.863	0.861	0.822	0.810	0.746	0.825
9 Sub-Saharan Africa											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP			x	x	x	x	x	x	x	x	x
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.95	0.940	0.959	0.964	0.965	0.962	0.964	0.955	0.948	0.972	0.975

Males											
1 Developed Europe											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**,**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x		x	x	x	x	x	x	x	x	x
Real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Lagged real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.900	0.788	0.716	0.656	0.651	0.655	0.688	0.799	0.862	0.901	0.895
2 Developed Non-Europe											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x	x	x	x	x	x	x	x	x	x	x
Real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Lagged real GDP growth rate	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.952	0.890	0.699	0.785	0.855	0.815	0.854	0.872	0.892	0.602	0.724
3 CEE and CIS											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared	x	x	x	x	x	x	x	x	x	x	x
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.772	0.837	0.849	0.868	0.834	0.926	0.901	0.921	0.958	0.976	0.958
4 East & South-East Asia (& the Pacific)											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared			x	x	x	x	x	x	x	x	x
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.970	0.951	0.953	0.863	0.909	0.673	0.785	0.799	0.937	0.903	0.966
5 South Asia											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP											
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 15-24	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 25-64	x	x	x	x	x	x	x	x	x	x	x
dummy	x	x	x	x	x	x	x	x	x	x	x
R-squared	0.905	0.908	0.953	0.921	0.938	0.877	0.790	0.868	0.927	0.970	0.912

Males											
6 Central America and the Caribbean											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.899	0.857	0.642	0.622	0.733	0.797	0.795	0.893	0.841	0.856	0.939
7 South America											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x			x	x
Per-capita GDP squared		x	x	x	x						
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x							x	x		
Share of population aged 15-24	x							x	x		
Share of population aged 25-64	x							x	x		
R-squared	0.847	0.872	0.725	0.838	0.794	0.801	0.936	0.813	0.864	0.876	0.902
8 Middle East & North Africa											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP	x	x	x	x	x	x	x	x	x	x	x
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14											
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.832	0.878	0.809	0.782	0.862	0.852	0.853	0.860	0.864	0.863	0.913
9 Sub-Saharan Africa											
	15-19	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Per-capita GDP			x	x	x	x	x	x	x	x	x
Per-capita GDP squared											
Real GDP growth rate											
Lagged real GDP growth rate											
Share of population aged 0-14	x	x	x	x	x	x	x	x	x	x	x
Share of population aged 15-24											
Share of population aged 25-64											
R-squared	0.910	0.902	0.925	0.892	0.888	0.861	0.851	0.863	0.895	0.937	0.950

Note: The dummy for South Asia equals 1 for the countries: Afghanistan, Bangladesh, Brunei Darussalam, Maldives, Pakistan.

ANNEX 3: Harmonizing LFPR by age bands

Countries often report labour force data for age bands that are not identical to the age bands used by the ILO in the LFEP database. For example, in the United States the labour force is defined for age 16 and over.

Two types of problems arise:

(i) The countries report data based on a different definition than 15 years and above (e.g., 16 and above).

(ii) The data reported by countries are available for age bands **larger** than 5 years (e.g., 35-44 instead of 35-39 and 40-44).

The techniques used to harmonise the data are described below.

a. Harmonising data to 15 years and above

The standard upper and lower age-bands in the LFEP database refer to the age group of 65 years and above and 15 to 19 years of age.

Lower age band adjustments (15-19 age-band)

The countries for which the lower age limit adjustment is required include Azerbaijan, El Salvador, Guam, Iceland, Macau, Mauritius, Mongolia, Nauru, Puerto Rico, United States. All the above countries report data for the 16-19 age-band instead of the 15-19.

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

For the United States, more detailed data is available within the 16-19 years age band. Therefore, these data are used to estimate the proportional coefficient (or ratio) that is applied to countries mentioned above.

Step 1: estimating the ratio for the US

For the United States, estimates of the LFPR for 16-19 year-olds are available for two sub-categories[28], 16-17 and 18-19. Since LFPR within this age-group are positively related with age (the LFPR of 18-19 year-olds is greater than 16-17 year-olds), the assumption is that the same relationship holds for 15 year-olds to 16-17 year-olds.

Therefore, the LFPR of 15 year-olds is estimated as using the proportional adjustment of equation [1]:

$$\widehat{LFPR^{US}_{s,y,15}} = \frac{LFPR^{US}_{s,y,16-17}}{LFPR^{US}_{s,y,16-19}} * LFPR^{US}_{s,y,16-17} \qquad [1]$$

where s=male, female and y=1980, ..., 2016. For example, in 2009, the LFPR of male 15 year-olds was estimated at 17.3 per cent, resulting from:

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

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

[28] The labour force participation rates for the United States come from the U.S. Bureau of Labor Statistics: http://www.bls.gov/data/

$$USratio_{s,y} = \frac{\widehat{LFPR^{US}_{s,y,15}}}{LFPR^{US}_{s,y,16-19}} \quad [2]$$

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

Step 2: estimating LFPR 15-19

The estimated LFPR for the 15 year-olds is calculated by applying the US ratio estimated in equation [2] to the country LFPR of the 16-19 year-olds:

$$\widehat{LFPR_{s,y,15}} = USratio_{s,y,} * LFPR_{s,y,16-19} \quad [3]$$

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

$$\widehat{LFPR_{s,y,15-19}} = 0.2 \cdot \widehat{LFPR_{s,y,15}} + 0.8 \cdot LFPR_{s,y,16-19} \quad [4]$$

The adjustment is done assuming a uniform population: 15 year-olds represent 20 per cent of the 15-19 population and 16 to 19 year-olds represent 80 per cent of the 15-19 population. This assumption is generally a reasonable one, and is much more accurate than combining sources to estimate the share of population of 15 years of age. For the few years for which detailed population census data exists a comparison can be made between the real data and the estimation based on 16-19 NSO data and 15-19 UN WPP data. Figure A1 shows this comparison. The results clearly favour the 20 per cent assumption. Figure A2 shows, using census data, that the approximation used is quite adequate for the countries for which it can be checked.

Figure A1: Comparison of census data and estimation, share of population of 15 years of age in the 15-19 age band

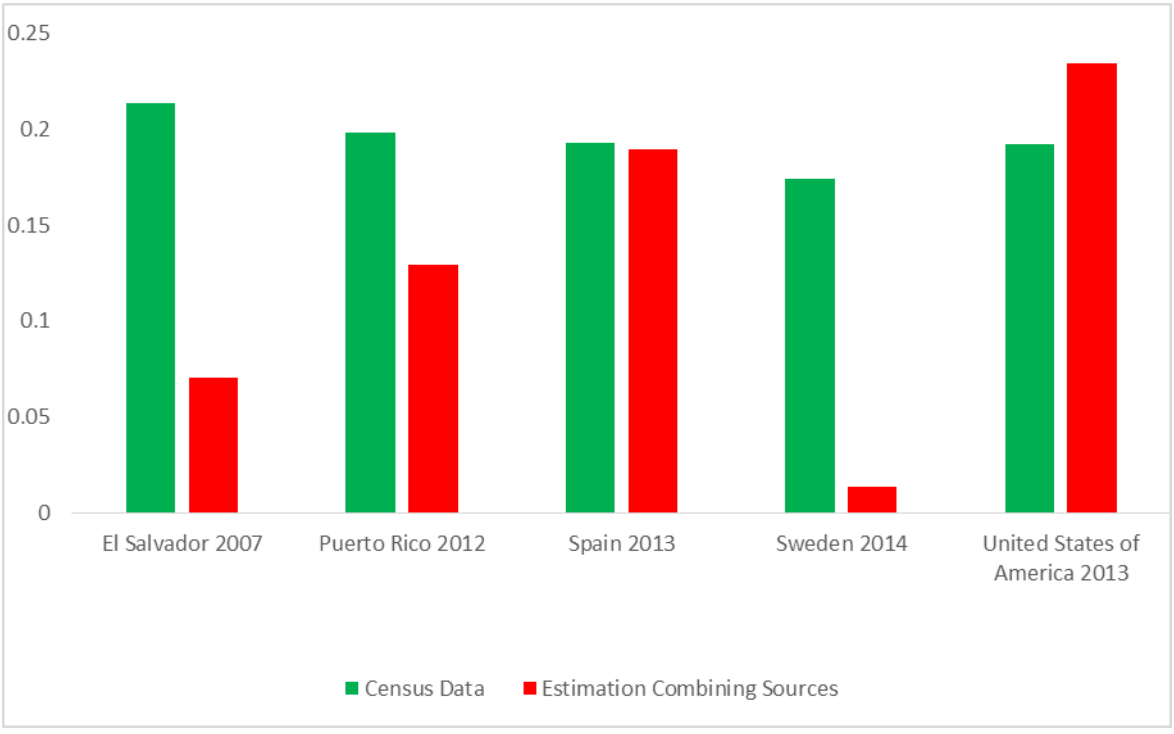
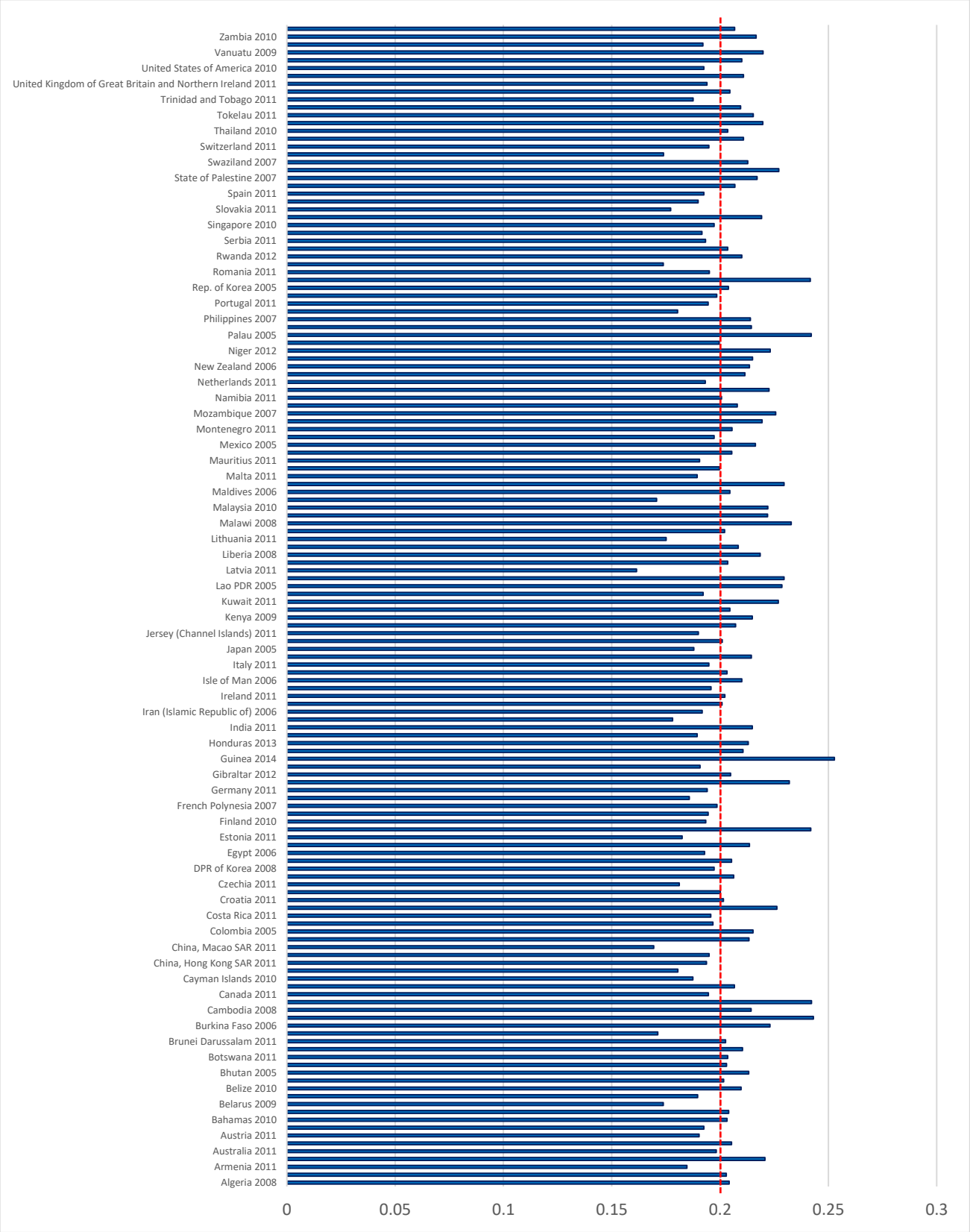


Figure A2: Census data, share of population of 15 years of age in the 15-19 age band



Upper age band adjustments (65+)

After the selection process only one country required an adjustment to the 65+ age-band, Iceland. In particular, the Labour Force Survey presents a cut-off at 74 years of age. An ad-hoc procedure is devised using OECD and UN WPP data. Assuming that the ratio of LFPR of 75+ to that of 70-74 year-olds is the same as the ratio of the LFPR of 70-74 year-olds to that of 65-69 year-olds, the 65+ LFPR is estimated. This result then is compared with the reported 65-74 data for a span of 22 years, which allows obtaining a correction factor 0.66 for the male LFPR and 0.55 for the female LFPR. This factor is then multiplied by the actual data.

Adjusting for change of survey

The data available for Brazil has as source two different labour force surveys: "PNAD" and "PNADC". There are substantial methodological differences between the two surveys that result in a level shift in LFPR for certain age groups. Since 2012 the data reported corresponds to PNADC, whereas PNAD is reported for the period 1990-2011. To avoid introducing a break in the series, as well as introducing an artificial downward trend in the most recent years, the data are adjusted. Taking advantage of the overlap of both surveys in 2012 and 2013 the average deviation between the surveys for each sex and age group is determined. Afterwards, the older data are corrected by the deviation. In this manner data with a reasonable degree of comparability across the two periods is obtained.

b. Harmonising data from large age bands to 5-year age-bands

For many countries, the LFPR reported by the National Statistical Offices are frequently not broken down by the standard 5 year age-groups. For example, in Puerto Rico apart from the tails of the age distribution, the country reports LFPR for the population aged between 25 and 34 years, between 35 and 44 years, between 45 and 55 years, and between 55 and 64 year-olds.

The principle is to break down reported data by large band (eg. 10 years) using a proxy variable, for which data is available by 5-year age bands. This consists of two steps.

Step 1: First estimate based on proxy variable.

Step 2: Adjust the estimate in order to match the original data.

Two types of proxy variables have been used. The best proxy variable to use is LFPR from other years in which data are available by 5-years age bands from national sources. For example, in Puerto Rico complete data by 5-year age bands are available in 2000 and can be used to derive estimates for the years that follow.

The second best proxy variable used here is the regional or sub-regional central values of LFPR by 5-year age bands, calculated using the median. When available, the sub-regional medians are used. However, for some sub-regions such as Central or North Africa there is not always enough information to derive reliable central values. In that case, the regional medians are used.

Formally, the first step consists of computing differences at the large band level *A*:

$$DIFF_{A,t} = LFPR_{A,t} - Proxy_{A,t} \quad [5]$$

where LFPR is a gender-specific LFPR and *t* the corresponding year.

Most of the time the large age band, *A,* includes 10 years. Let us note *a1* and *a2*, the two 5-year age groups that fall into the larger age band *A*.

In identity equation [1], $Proxy_{A,t}$ has been computed as:

$$Proxy_{A,t} = \omega_{a1,t} * Proxy_{a1,t} + (1 - \omega_{a1,t}) * Proxy_{a2,t} \quad [6]$$

where $\omega_{a1,t}$ represents the proportion of the population of the age band *a1* in the age band *A*. The weight $\omega_{a1,t}$ is derived from national data or from UN population estimates when national data are not available.

The first estimates are then computed:

$$LFPR'_{a1,t} = Proxy_{a1,t} + DIFF_{A,t}$$

and

$$LFPR'_{a2,t} = Proxy_{a2,t} + DIFF_{A,t} \quad [7]$$

By construction, these estimates respect the following condition:

$$LFPR'_{A,t} = LFPR_{A,t}$$

Further adjustments can be done if the proxy is not judged to be sufficiently reliable. This can result in out of range estimates (>100 per cent or < 0 per cent), or, more frequently, unexpected results for estimates that belong to different large age-groups, for example LFPR'[60-65] (based on [60+] data) that exceeds LFPR'[55-59] (derived from [50-59] reported data).

The corrective adjustments rely on combining the first estimate and the original data. Formally, the following formula is used:

$$LFPR''_{a1,t} = \lambda * LFPR'_{a1,t} + (1 - \lambda) * LFPR_{A,t} \quad [8]$$

Different values of λ are tested, λ reflecting the weight attributed to the first estimate. First λ=0.5, then if the estimate is still inconsistent, λ=0.25 and if finally λ=0. In the latter case, the final estimate is set to the original data.

The final estimate is then obtained by calculating the discrepancies between these second estimates and the original data and allocating them at the 5-year age-band level, such as:

$$DISCREPANCY_{A,t} = LFPR''_{A,t} - LFPR_{A,t} \quad [9]$$

where

$$LFPR''_{A,t} = \omega_{a1,t} * LFPR''_{a1,t} + (1 - \omega_{a1,t}) * LFPR''_{a2,t}$$

Then the discrepancy is allocated within each 5-year age band, for example for age band *a1*:

$$FINAL\ LFPR_{a1,t} = LFPR''_{a1,t} + DISCREPANCY_{A,t} \quad [10]$$