

► Annex B. Estimation methodology

The estimation methodology used a series of models that establish statistical relationships between observed labour migration indicators reported by countries to the ILO's International Labour Migration Statistics (ILMS) database and selected explanatory variables, detailed in section B2. These relationships were used to impute missing observations to obtain estimates for all 189 countries and territories covered by the estimations.

Once the set of estimates were obtained for each country using the new methodological approach, the resulting values were aggregated to obtain a global figure, and into 5 regions and 11 broad subregions according to the ILO geographical classification, and 4 income group as defined by the World Bank's country classification system (see Annex A). The following sections provide more details on the estimation approach for each specific labour migration indicator estimated.

B1. Input data for estimation

Figure B1 summarizes the main input data used for the estimation of the labour market indicators included in this report:

- Country reported data on the international labour migration indicators to estimate: Data collected at the country level, mainly through household surveys or population censuses, and reported to the ILO. They present some gaps as not all the 189 countries and territories for which the global estimation is performed report those indicators (see Annex C).
- Benchmark data: Harmonized estimates, produced by the United Nations or the ILO, and available for all the 189 countries and territories. Their use aims at ensuring the consistency of the estimates on international migrants in the labour force with the reference estimates of the total working-age population, total international migrant stock and labour market indicators for the total population.

Additional data were used to predict the probability of countries reporting the target labour migration indicators, in order to account for potential non-reporting bias in the series of models, and to estimate specific indicators, such as the proportion of migrants in care employment.

► Figure B1. Benchmark and country reported data serving as input to the estimation

Benchmark data	Country reported data
 Population <i>UNDESA, World Population Prospects, 2022 Revision</i>	 Labour force participation rate by migrant status Employment-to-population ratio by migrant status Employment by economic activity and migrant status Care employment by migrant status <i>ILMS database, latest period</i>
 International Migrant Stock <i>UNDESA, 2020 Revision</i>	
 Labour force participation rate Employment-to-population ratio Employment by economic activity <i>ILO modelled estimates, November 2023</i>	

All input datasets used were the latest publicly available at the time the ILO estimates were finalized in preparation for this report, in June 2024. These input datasets are presented in greater detail in the following sections.

B1.1. Benchmark data

Benchmark population data

The working-age population data are taken from the World Population Prospects (2022 Revision)¹, prepared by the United Nations Department of Economic and Social Affairs (UNDESA). The projection for the reference year 2022 was aggregated into four age groups (15–24, 25–54, 55–64 and 65+) by sex to obtain the benchmark values for working-age population (P).

Benchmark international migration data

The benchmark international migrant data were derived from the International Migrant Stock 2020 (UNDESA, 2020)². The proportion of working-age migrants in the population (m) was computed by sex and age group (15–24, 25–54, 55–64 and 65+) for the 189 countries and territories covered by the estimation³, using the working-age international migrant population (M) and the total population for 2020 from the same dataset. In the absence of more recent dataset on international migrant stock, the assumption was that the proportion of the working-age migrant population in 2020 remained constant in 2022. The decrease in international migration flows observed in some regions during the COVID-19 period, especially in high-income countries, which are the main destinations of migrants (González-Leonardo et al. 2023), could support this assumption of stagnation of international migrant stocks. The proportion of working-age migrants was applied to the working-age population of 2022, from the World Population Prospects dataset, to obtain the number of working-age international migrants for the reference year 2022.

Benchmark labour market data

The benchmark labour market data were taken from the November 2023 revision of the ILO modelled estimates of core labour market indicators (reference year 2022), as disseminated on the ILO Modelled Estimates (ILOEST) database on ILOSTAT, the ILO data portal on labour statistics.⁴ Three datasets were extracted from the ILOEST database, each covering the 189 countries and territories included in the estimation:

- **Labour force participation rate:** This indicator reflects the proportion of persons of working age who are in the labour force; that is, who are in employment or in unemployment.
- **Employment:** This indicator measures the proportion of the working-age population that is currently employed.⁵
- **Employment by economic activity:** This indicator reflects the number of employed persons by broad categories of economic activity; that is, in agriculture, industry or services.

¹ Available at <https://population.un.org/wpp/Download/Archive/Standard/>.

² These estimates take account of refugees, asylum seekers and other related groups. However, they do not provide separate estimates that could be used to support disaggregation. Taiwan, China is not included in the dataset. The proportion of international migrant population in Taiwan, China, according to Integral Human Development was used instead (see <https://migrants-refugees.va/country-profile/taiwan/>).

³ The international migrants living in the 189 countries and territories represent 99.6 per cent of the world's migrant population.

⁴ Available at <https://ilostat.ilo.org/data/>.

⁵ Most country-reported employment data and the employment estimates for total population are based on the 13th International Conference of Labour Statisticians (ICLS) definition of employment, which includes production of goods mainly intended for own final use in employment. A few countries reported employment data based on the 19th ICLS definition of employment, which excludes this groups. See <https://ilostat.ilo.org/data/> for details.

B1.2. Country reported data on international labour migration

Information on the labour market characteristics of international migrants were obtained from data reported by countries from different sources, compiled and disseminated by the ILO through the ILMS database on ILOSTAT.⁶ In recent years, the ILMS database grew significantly from 19 indicators and 27 reporting countries and territories in October 2018 to more than 70 indicators and 170 countries and territories in June 2024 and is now the only source of country reported data on labour migration used for the ILO estimates.

Three sets of indicators, listed below, were extracted from the ILMS database to serve as input data. The ILMS database compiles indicators separately by place of birth and citizenship, which are the two main criteria to define international migrants. Priority was given to indicators based on the criterion of place of birth when available (see Annex C). The three set of indicators are:

- **Labour force participation rate (LFPR):** LFPRs by sex, international migrant status and age group (15–24, 25–54, 55–64 and 65+), available for 142 countries and territories, with 126 based on place of birth and 16 based on citizenship.
- **Employment-to-population ratio (EPR):** EPRs by sex and international migrant status for the population aged 15 years and over, available for the same 142 countries and territories with data on LFPRs.
- **Employment by economic activity (EA):** The number of persons employed by economic activity by sex and international migrant status for the population aged 15 years and over, available for 144 countries and territories.⁷ The proportion of persons employed by economic activity (pEA) was calculated from the ratio of the number of persons employed in each economic activity and the total number of persons employed.
- **Employment in the care economy (CE):** The number of persons employed in the care economy by sex and international migrant status for the population aged 15 years and over, available for 109 countries. The proportion of persons employed in the care economy (pCE) was calculated from the ratio of the number of persons employed in the care economy and the total number of persons employed.

B1.3. Additional data

The estimation methodology required the use of two additional datasets:

- **Gross domestic product per capita (GDPpc):** This variable was used as an explanatory variable to model the probability of a country to report labour market indicators by international migrant status. GDP itself is not a direct cause of a lack of reporting data, but there are several ways in which GDP-related factors might influence or correlate with data reporting issues, such as economic development and infrastructure, administrative capacity, data collection priorities, and data quality and reliability. The most recent available GDPpc in constant 2015 US dollars⁸ for the 189 countries and territories reported by the ILO was obtained from the data repositories of the World Bank⁹, the International Monetary Fund¹⁰ and the United Nations¹¹. Of the 189 countries and territories, GDPpc in 2022 was used for 176, in 2021 for 9 and in 2015 or earlier for 4.
- **Healthy life expectancy (HALE) at age 60:** HALE at age 60 is the average number of years in full health a person at age 60 can expect to live based on current rates of ill health and mortality. HALE at age 60 is

⁶ Available at <https://ilostat.ilo.org/data/>.

⁷ Compared to the 142 countries and territories with LFPRs, data on economic activity were not available for Liberia and Saudi Arabia, but were available for Azerbaijan, Japan, the Occupied Palestinian Territory and the Republic of Korea.

⁸ GDP per capita in current US dollars was used for: the Channel Islands; the Democratic People's Republic of Korea; Taiwan, China; the Bolivarian Republic of Venezuela; and Western Sahara.

⁹ See <https://data.worldbank.org/indicator/NY.GDP.PCAP.KD>.

¹⁰ See <https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD>.

¹¹ See <https://data.un.org/>.

used as an explanatory variable to model pCE, as it is expected that people living longer without good health will demand more services from the care economy. HALE at age 60 for the 109 countries and territories reporting pCE was obtained from the data repository of the World Health Organization.¹² The latest HALE at age 60 available corresponds to 2021.

B2. Main components and steps

B2.1. Estimation of international migrants in the labour force

A linear mixed model was used to estimate the LFPR of the population by sex, international migrant status and age group. Model estimates are used to impute LFPR values in countries with missing data to derive the number of international migrants in the labour force. A linear relationship is assumed between LFPR of migrants and LFPR of the total population. The model specification is given in equation 1 (i refers to country and t refers to year):

$$L_{it} = \ln \left(\frac{\text{LFPR}_{it}}{(1 - \text{LFPR}_{it})} \right) = x_{it}\beta + z_{it}\varphi + p_i + y_t + \alpha_i + \varepsilon_{it} \quad (1)$$

L_{it} is the logit transformation of reported LFPR by sex, international migrant status and age group.¹³ The vector x_{it} is a set of explanatory covariates that included the LFPR for the total working-age population, the proportion of international migrants in the population and the square of this proportion. The vector z_{it} is a set of dummy variables including sex, international migrant status, age group, broad subregion and income group. In addition, a dummy variable for year (y_t), a country-specific random effect (α_i)¹⁴ and an error term (ε_{it}) were included.

The model also includes the probability of having reported LFPR by international migrant status (p_i) to account for the non-response bias.¹⁵ This is important as only 142 out of the 189 countries and territories reported LFPR by international migrant status to the ILO, and ignoring or mishandling missing data can lead to biased results, reduced statistical power and incorrect conclusions in data analysis or modelling.

The fitted values of LFPR ($\widehat{\text{LFPR}}_{it}$) were obtained from the estimated $L_{it}(\widehat{L}_{it})$, as follows:

$$\widehat{\text{LFPR}}_{it} = \frac{\exp(\widehat{L}_{it})}{(1 + \exp(\widehat{L}_{it}))} \quad (2)$$

The international migrant labour force (MW) was computed based on the benchmark estimate of population (P) and benchmark proportion of international migrants (m) for the reference year, to harmonize the estimates with the international estimates of population and of international migrants (see section B1.1 for details on benchmark data). For a bottom-up approach aiming at producing aggregate results by geographical location and income group, the computation of MW was performed separately for eight categories or subgroups of the population, derived from the combination of sex and the four age categories (15–24, 25–54, 55–64 and 65 years and above). Estimates were obtained using the following equation

¹² See <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy>.

¹³ LFPRs were included in the model as a logit transformation to normalize skewed data in order to make them more suitable for statistical analyses and to reduce the impact of outliers for increased reliability.

¹⁴ While both random effects and fixed effects models provide identical coefficient estimates, the random effects specification offers several advantages that make it more suitable for the current analysis. These benefits include the ability to account for heterogeneity among countries, incorporate time-invariant country-specific covariates and potentially apply the model to countries not present in the sample. Given our focus on imputation, the random effects approach is particularly appropriate. To validate this choice, a Hausman-type test was used, showing that the random effects model should be preferred.

¹⁵ Unreported LFPRs are missing at random, meaning that reporting countries are known to be different from non-reporting countries and the factors that determine whether a country reports data are identifiable. To account for missing data, a logistic regression was run to estimate, conditional on identifiable country characteristics such as GDP per capita (constant 2015 US dollars), national working-age population, share of migrant working-age population at the national level, income group and broad subregion. The dependent variable took the value of 1 if the LFPR by sex, international migrant status and age group was reported, and 0 otherwise. The estimated was then included in the main model as an additional variable (Heckman 1979).

for each category of the population in a given country, where MLFPR is the migrant labour force participation rate:

$$MW = MLFPR \times P \times m \quad (3)$$

The MW values are computed differently according to the availability of observed data on LFPRs by international migrant status. For countries with observed MLFPR in the relevant sex–age category, MLFPR corresponds to the reported national data for the reference year or the most recent year if 2022 data were not available (see equation 4a). For countries with missing MLFPR in the relevant sex–age categories, MW was computed using MLFPR derived from the model estimates (see equation 4b). Details on the data availability of each country can be found in annex C, table C1.

$$MW = MLFPR_{obs} \times P \times m \quad (4a)$$

$$MW = \widehat{MLFPR} \times P \times m \quad (4b)$$

In equations 4a and 4b, $MLFPR_{obs}$ corresponds to the reported values of LFPR for migrants from national data, and $\widehat{MLFPR}$ to the fitted values of LFPR for migrants from the model.

The non-migrant labour force was calculated as the difference between the ILO modelled estimates of the labour force and the MW estimates from equations 4a or 4b. This approach ensures internal consistency of the results while preserving the levels of MW derived from the model.

B2.2. Estimating employed international migrants

A similar approach was used to estimate the number of international migrants in employment. The outcome variable of the estimation (R_i) is the logit-transformed EPR by sex and international migrant status for the population aged 15 years and over. A linear model was used,¹⁶ and the explanatory variables are the benchmark EPR, the proportion of migrants and its square (vector x_i), and the probability of having reported EPR, estimated similarly to p_i . Dummy variables for sex, international migrant status, broad subregion, and income group were also included (vector z_i).

$$R_i = \ln \left(\frac{EPR_i}{(1 - EPR_i)} \right) = x_i \beta + z_i \varphi + p_i + \varepsilon_i \quad (5)$$

Migrant employment (ME) by sex was calculated using different equations depending on the availability of observed EPR, similar to the estimation of LFPR. Equation 6a was used where reported data were available and equation 6b, based on fitted values of EPR from the model, was used where EPR was not reported.

$$ME = MEPR_{obs} \times P \times m \quad (6a)$$

$$ME = \widehat{MEPR} \times P \times m \quad (6b)$$

In equations 6a and 6b, $MEPR_{obs}$ corresponds to the reported values of migrant employment-to-population ratio from national data and $\widehat{MEPR}$ to the fitted values of migrant employment-to-population ratio from the model. Non-migrant employment was calculated as the difference between the employment of the total population from ILO modelled estimates and ME.

¹⁶ In this case, only the most recent data on employment was used, since employment fluctuates more over time compared to LFPR. Therefore, a linear model was used, since the dataset did not present the characteristics of an unbalanced panel.

B2.3. Estimating the distribution by economic activity of employed international migrants

This step builds on the estimated employment by international migrant status and sex for the population aged 15 years and over. The output dataset described in section B2.2 serves to derive the distribution of the employed population by three broad categories of economic activity: agriculture, industry and services. The distribution by economic activity (pEA) can be modelled as a set of dependent fractional variables, each bounded between 0 and 1 and whose sum is equal to 1. Therefore, a fractional multinomial logit model¹⁷ was used to estimate the proportion of persons employed in each sector by sex and international migrant status.

Explanatory variables, included in vector x_i , were the proportion of persons employed in agriculture, the proportion of persons employed in industry and the benchmark EPR, and the probability of having reported pEA, estimated similarly to p_i . As before, dummy variables for sex, international migrant status, broad subregion and income group were included (vector z_i), and i refers to country by sex and international migrant status and j refers to broad category of economic activity by sex and international migrant status:

$$pEA_{ij} = x_i \beta + z_i \varphi + p_i + \varepsilon_i \quad (7)$$

The number of persons employed in each broad category of EA by international migrant status and sex was calculated using different equations depending on the availability of observed pEA, similar to the estimation of LFPRs and EPRs. Equation 8a was used where reported data of pEA for all three categories of economic activity were available for a given sex and international migrant status (for example, available for all migrant women). Equation 8b was used in all three categories of economic activity for a given sex and international migrant status when at least one category of pEA was missing (for example, agriculture data missing for migrant women).

$$EA = pEA_{obs} \times E \quad (8a)$$

$$EA = p\hat{EA} \times E \quad (8b)$$

In equations 8a and 8b, pEA_{obs} corresponds to the reported values of pEA from national data, $p\hat{EA}$ to the fitted values of pEA from the model, and E to the estimates of employment obtained as described in section B2.2.

B2.4. Deriving unemployment estimates

The number of persons in the labour force equals the number of persons employed plus the number of persons unemployed (see definitions in box 2, in the Introduction). As the labour force and employment by international migrant status were already estimated using the approaches described above, the number of unemployed persons by international migrant status can be derived as the difference between the labour force and employment. Unemployment rates can then be calculated as the ratio of the number of unemployed persons to the labour force, for each sex and migrant status category.

B2.5. Estimating the proportion of employed international migrants in care employment

This estimation focuses only on a subset of 109 countries and territories given the data limitations on care employment. There are no harmonized benchmark estimates for total population in care employment and reported national data for total population by international migrant status is limited (reported data for 86 countries and territories only, see Annex D).

¹⁷ A fractional multinomial logit model is a statistical model used in the analysis of categorical outcomes where the response variable represents fractions or proportions that sum to one within each observation (Papke and Wooldridge 1996).

Similar to EPR, a linear model was used to estimate the number of international migrants in care employment (CE). The outcome variable C_i is the logit-transformed proportion of persons in care employment (pCE) by sex and international migrant status for the population aged 15 years and over. The explanatory variables, included in vector x_i , are the benchmark EPR for total working-age population, the proportion of migrants and the proportion of dependent population (for example, population aged 0–14 and 65+). HALE at age 60 by sex and the probability of having reported CE, estimated similarly to p_i , were included. Dummy variables for sex, international migrant status, broad subregion, income group (vector z_i), and year (y_i) were also included, where i refers to country:

$$C_i = \ln \left(\frac{pCE_i}{(1 - pCE_i)} \right) = x_i \beta + z_i \varphi + HALE + p_i + y_i + \varepsilon_i \quad (8a)$$

The number of migrants employed in the care economy (MCE) by sex was calculated using different equations depending on the availability of observed pCE, similar to the estimation of LFPRs, EPRs and pEAs. Equation 10a was used where reported data of pCE were available and equation 10b was used when pCE was missing.

$$MCE = pCE_{obs} \times E \quad (10a)$$

$$MCE = \widehat{pCE} \times E \quad (10b)$$

In equations 10a and 10b, pCE_{obs} corresponds to the reported values of migrant pCE from national data, $\widehat{pCE}$ to the fitted values of migrant pCE from the model, and E to the estimates of employment from section B2.2. MCE is then aggregated at the global level and the proportion of migrants employed in the care economy (pMCE) by sex is derived using equation 10c:

$$pMCE = \frac{MCE}{E} \quad (10c)$$

B3. Quality of estimates

The approach used in the current edition allows the production of high-quality estimates of international migrants in the labour force given improvements on several aspects. First, the new methodology employs a robust regression-based imputation approach using specific parameters based on the relationships between migrants and population characteristics at the regional level. Second, the use of country-level LFPR as the variable for imputation makes the estimation less sensitive to data availability at the regional level. Third, the imputation at the country level with a bottom-up approach allows estimates to be produced with more detailed disaggregation. The quality of the ILO global estimates can be assessed in terms of coverage of the underlying country reported data, the consistency of the estimates and the accuracy of the results.

B3.1. Coverage of country reported data

Labour market indicators by international migrant status were available for at least 142 countries and territories, representing 75 per cent of the 189 countries and territories covered by the estimation presented in this report. However, the availability of the national data was somewhat unevenly distributed over the ILO geographical regions. As shown in table B1, Northern America, Northern, Southern and Western Europe, and Eastern Europe have a complete, or almost complete coverage of country reported data, while data coverage remains low in the Arab States, Central and Western Asia, and Eastern Asia.

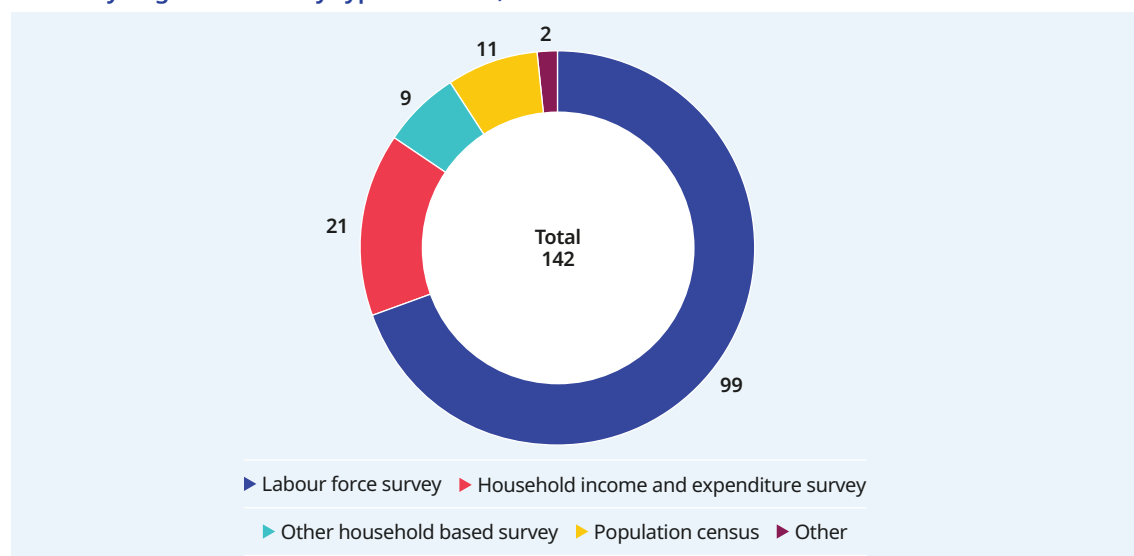
► **Table B1. Coverage of country reported data by international migrant status by broad subregion**

Region	Broad subregion	No. of countries reporting data			Share of countries reporting data (%)		
		LFPR	EPR	EA	LFPR	EPR	EA
Africa	Northern Africa (7)	3	3	3	43	43	43
	Sub-Saharan Africa (47)	39	39	38	83	83	81
Americas	Latin America and the Caribbean (31)	23	23	23	74	74	74
	Northern America (2)	2	2	2	100	100	100
Arab States	Arab States (12)	5	5	5	42	42	42
Asia and the Pacific	Eastern Asia (8)	1	1	3	13	13	38
	South-Eastern Asia and the Pacific (22)	17	17	17	77	77	77
	Southern Asia (9)	7	7	7	78	78	78
Europe and Central Asia	Central and Western Asia (11)	6	6	7	55	55	64
	Eastern Europe (10)	10	10	10	100	100	100
	Northern, Southern and Western Europe (30)	29	29	29	97	97	97
Total (189)		142	142	144	75	75	76

Note: EA – employment by economic activity; EPR – employment-to-population ratio; LFPR – labour force participation rate.

Source: ILO estimates.

The labour market indicators by international migrant status from the ILMS database come from different sources. Figure B2 shows the distribution of the 142 countries and territories with available data on LFPs by international migrant status. As observed, most of the data by migrant status come from labour force surveys (70 per cent) and household income and expenditure surveys (15 per cent) (see Annex C for details by country). The reference year of country reported data coincided with the 2022 reference year of this report for around half of the countries (67). Only around one out of five countries (30) had less recent data, dating from 2018 or earlier.

► **Figure B2. Number of countries and territories that reported data on labour force participation rate by migrant status by type of source, 2022**

Source: ILO estimates.

An increased country coverage improves the quality of estimates by incorporating a wider range of data, enhancing representativeness and reducing biases. This leads to more accurate and comprehensive statistical models and predictions.

B3.2. Consistency

As part of the assessment of data quality, internal consistency of the estimates was examined. The methodology used in this report ensured that the results respect inherent relationships among the key indicators of population and labour market. For example, the number of international migrants in the labour force (MW) and the number of employed international migrants must always be lower than the international migrant working-age population and MW, respectively. The same requirements apply to the estimated labour market indicators for non-migrants, which were calculated as the difference between benchmark labour market data and estimated migrant labour market data to preserve the values for migrants derived from the different models.

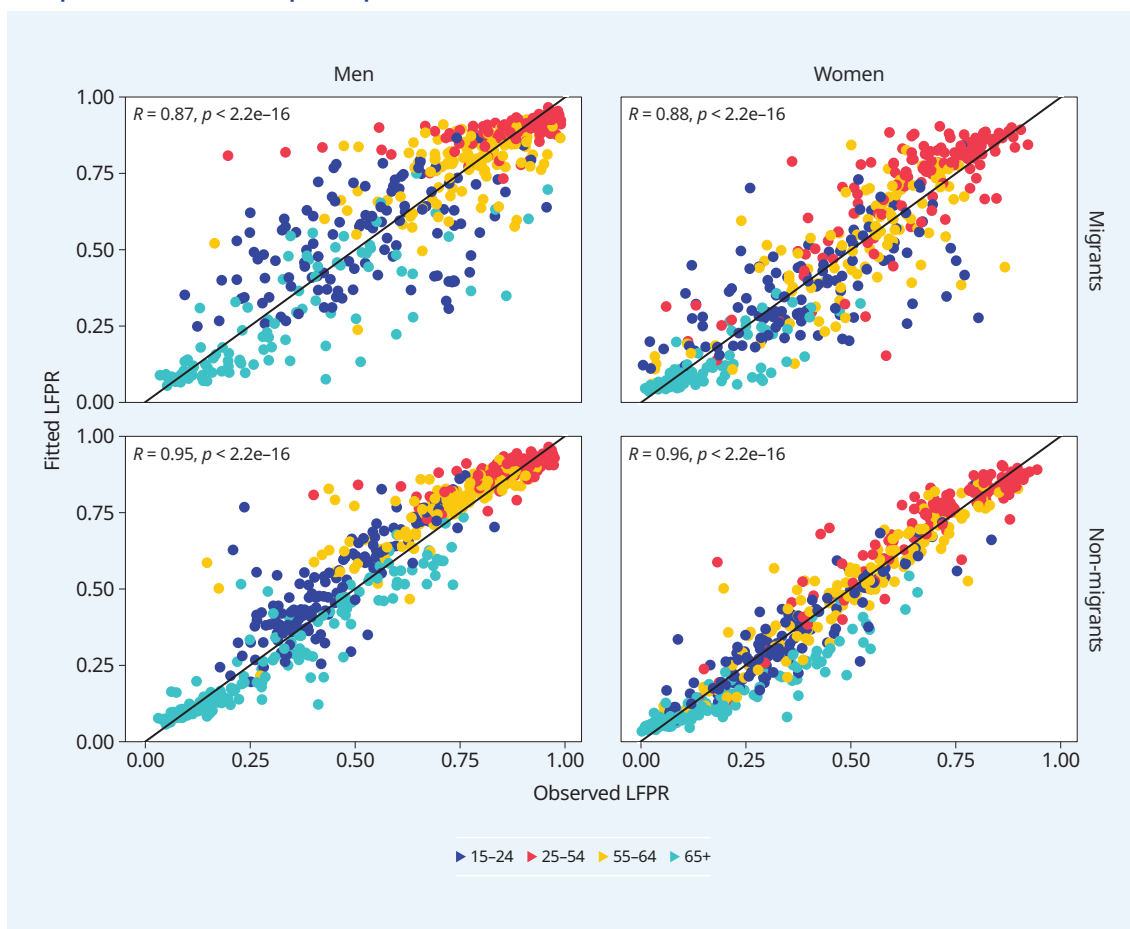
Overall, this approach performed well, and there was internal consistency except for 0.9 per cent of the cases (26 of the 3,024 categories), all of which were non-migrant categories. For those few cases, another approach was used to adjust the estimated labour market indicators to benchmark data. The ratio of the migrant labour force (MW) to the labour force by sex and age group estimated by the model was calculated, and this ratio was multiplied by the benchmark labour force. The non-migrant labour force was then calculated as the difference between this adjusted MW and the benchmark labour force.

B.3.3. Accuracy

The accuracy of the method was evaluated in two ways. A first simple assessment involved the comparison of the fitted estimates with the observed corresponding indicators reported by countries. Figure B3 plots the relationship between fitted LFPR by age group and country reported LFPR by age group for each sex and migrant status category. The Pearson coefficient shows a very strong correlation in all cases ($R > 0.8$). These results suggest that the model is able to predict accurate estimates.

The accuracy of the method was also assessed using cross-validation, which is a statistical technique used to evaluate the performance and generalizability of predictive models. Cross-validation aims to assess how well a model will perform on unseen data by partitioning the available data into subsets, training the model on some subsets and validating it on the remaining subsets. In this case, 50, 60 and 70 per cent of the original dataset were used as training dataset. The metrics used to assess the accuracy of the method were the out-of-sample R -squared and root mean square error (RMSE). The out-of-sample R -squared measures how well the model generalizes to new, unseen data. Out-of-sample R -squared usually tends to be lower than the in-sample R -squared. In this case, the metric is equally high for all four versions of the training dataset used for the assessment (table B2). The out-of-sample RMSE measures the average magnitude of the error between fitted and observed values on new, unseen data, and should be close to the in-sample RMSE, indicating that the model generalizes well without overfitting. The out-of-sample RMSE is low in all four versions of the assessment. Considering these metrics, the model has a very good predictive performance for new, previously unseen data.

► **Figure B3. Comparisons between model fitted labour force participation rate and country reported labour force participation rate**



Source: ILO estimates.

► **Table B2. Evaluation of the accuracy of the model from cross-validation assessment**

Cross-validation	Out-of-sample R-squared	Out-of-sample root mean square error
100% training dataset (original dataset)	0.92 ¹	0.505 ¹
70% training dataset	0.91	0.502
60% training dataset	0.91	0.505
50% training dataset	0.91	0.504

¹ This value corresponds to an in-sample metric.

Source: ILO estimates.