Policy Research Working Paper 10034

# When Is There Enough Data to Create a Global Statistic?

*Daniel Gerszon Mahler*
*Umar Serajuddin*
*Hiroko Maeda*

WORLD BANK GROUP
Development Economics
Development Data Group
May 2022

Policy Research Working Paper 10034


## Abstract

To monitor progress toward global goals such as the Sustainable Development Goals, global statistics are needed. Yet cross-country data sets are rarely truly global, creating a trade-off for producers of global statistics: the lower is the data coverage threshold for disseminating global statistics, the more statistics can be made available, but the lower is the accuracy of these statistics. This paper quantifies the availability-accuracy trade-off by running more than 10 million simulations on the World Development Indicators. It shows that if the fraction of the world's population for which data are lacking is x, then the global value will on expectation be off by 0.37*$x$ standard deviation, and it could be off by as much as x standard deviations. The paper shows the robustness of this result to various assumptions and provides recommendations on when there is enough data to create global statistics. Although the decision will be context specific, in a baseline scenario, it is suggested not to create global statistics when there are data for less than half of the world's population.


This paper is a product of the Development Data Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dmahler@worldbank.org, userajuddin@worldbank.org, and hmaeda@worldbank.org.

# When Is There Enough Data to Create a Global Statistic?

*Daniel Gerszon Mahler, Umar Serajuddin, and Hiroko Maeda*[1]

*JEL codes: C15, C18, C53, C82, Y10.*
*Keywords:* Aggregation, data, simulation, SDGs, statistics, missing data, monitoring, India.

[1] All authors are with the Development Data Group of the World Bank. Corresponding author: Daniel Gerszon Mahler (dmahler@worldbank.org). We are grateful for comments received from Aart Kraay, R. Andrés Castañeda Aguilar, Bob Rijkers, Christoph Lakner, and Dean Jolliffe. We are also grateful for discussions on the topic from participants of the Interagency Task Team on Aggregation coordinated by the United Nations. The code for the paper is available at https://github.com/danielmahler/Aggregation.

## Introduction

Open a newspaper and chances are you will find some statistic referring to how the world is fairing: "Global growth is projected to recover", "the number of refugees worldwide is set to increase for the third straight year", "global CO2 emissions are reaching an all-time high." The demand for global statistics is perhaps best embodied in the Sustainable Development Goals (SDGs), whose 231 indicators for the most part can be and are aggregated to the global level.

In reality, there is rarely complete global data behind such statistics. Due to lack of resources, capacity, and political will, some countries do not produce information on the indicators of interest (Dang et al. 2021; Serajuddin et al. 2015). When creating global statistics, estimates for these countries are either imputed or simply ignored. This inevitably creates a trade-off between the *availability* of global statistics, and the *accuracy* of these statistics. If global statistics are only published when data are universally or near-universally available, there will be many important topics that cannot be illuminated. If global statistics are published even when the data coverage is weak, the accuracy of the statistics may be doubtful in the sense that they are likely to deviate from the figure had all data been available.

In this paper, we quantify this trade-off between data accuracy and data availability using the World Bank's World Development Indicators. We select 165 indicators spanning a wide range of topics where, for a given year, data are available for at least 99% of the world's population. We randomly remove data from these indicators and calculate the expected difference in the global statistic as a function of the share of the world's population without data. We show that if the fraction of the world's population on which one lacks data is $x$, then one should expect to be $0.37*x$ standard deviations away from the true mean, and as much as $x$ standard deviations from the mean at times. Here the standard deviation is based on the distribution of country-level estimates. As data producers might not be used to thinking in standard deviations from the mean, we provide examples of what such deviations imply.

In further results we show how these errors change (i) if one is interested in regional statistics, (ii) if data are imputed, (iii) if the probability of data missing is correlated with the indicator of interest, (iv) if one uses the share of countries rather than the share of population as a coverage threshold, and (v) if one has specific coverage requirements for populous countries, such as India. We end with recommendations on when to produce global statistics. We hope these recommendations can be used to ensure that global statistics are only made available when the accuracy is deemed sufficiently high. This has implications for international organizations and researchers producing cross-country data sets that are aggregated to create a global statistic. By consequence, the recommendations have implications for any users of such global statistics, including academia, the media, and policy makers.

To our knowledge, we are the first to study when there is enough data to create a global statistic. Yet, our paper relates to several streams of literature, such as the challenges of measuring the SDGs (MacFeely 2018; Sachs 2012), missing data on SDG reporting (Dang & Serajuddin 2020; World Bank 2021), and making inference with missing data (Dang et al. 2019).

## Method

To quantify the impact of data availability on the precision of global statistics, we rely on the World Bank's World Development Indicators (WDI). The WDI is arguably the world's largest database of relevant country-year indicators spanning a wide range of topics. The WDI contains information on around 1,400 indicators covering topics such as poverty, health, agriculture, education, climate change, infrastructure and more. We select 165 different indicators that for a given year have near universal coverage (>99% of the world's population). We diversified the indicators to cover as many different topics as possible. We focus on

indicators where one is interested in the population-weighted mean of an indicator, such as global growth, the global unemployment rate, global electricity access, and so on. The indicators chosen are listed in Table A.1.

For these 165 indicators, we abstract from the small degree of missingness and consider the statistic they produce as the ground truth. Next, we randomly delete a subset of the data for each indicator, calculate the new global mean, and compare it to the ground truth. This gives us an estimate of the error when only a fraction of the global population has data. By repeating this exercise more than 10 million times using different indicators and different probabilities of missingness, we can calculate the expected error as a function of population coverage. To compare indicators in different units we first standardize all variables to have mean 0 and variance 1. This allows us to express the error as standard deviations from the mean and average these errors across all indicators.

Most data producers may not be used to thinking of their indicator in terms of standard deviations from the mean. To foster some intuition, Table 1 shows what one standard deviation implies for five selected indicators. If one is a standard deviation away from the true mean when creating a global statistic, one could get life expectancy off by 7 years, global growth off by 3 percentage points, and the share using at least basic sanitation services off by 24 percentage points. Even if these errors are cut by four, and one is 0.25 of a standard deviation off the truth, they still represent large errors.

*Table 1: Examples of one standard deviation for selected indicators*

| Indicator | Global mean | 1 standard deviation |
|---|---|---|
| Life expectancy at birth, total (years) | 71.1 | 6.9 |
| GDP growth (annual %) | 5.0 | 3.0 |
| People using at least basic sanitation services (% of population) | 71.5 | 24.2 |
| Agricultural land (% of land area) | 47.9 | 17.5 |
| CO2 emissions (metric tons per capita) | 4.4 | 4.2 |

*Note*: The table shows what one standard deviation (using the distribution of country estimates) implies for five different indicators for a particular year.

Another way of interpreting standard deviations from the mean is by looking at how much global statistics change from one year to the next. For 144 of the 165 indicators we have chosen, we have at least 99% coverage two years after each other. This means that we can calculate how much the global mean changed expressed in standard deviations from the mean in the first year. Half of all indicators change by 0.03 standard deviation or less from one year to the next, and no indicator changes by more than 0.33 standard deviation.

On the one hand, this means that if one is 0.03 standard deviation from the true mean in a single year because of missing data, then for half of all indicators, one would not be able to tell apart true changes in the statistic from changes driven by inaccuracy. On the other hand, to the extent that countries with missing data remain the same from one year to the next, missingness is less likely to impact year-to-year changes and more likely to cause a systematic and consistent bias. Across WDI, 93% of instances with missing data in one year also have missing data in the next year, suggesting the latter channel may dominate in many cases.

One may wonder why we use simulations to quantify the availability-accuracy trade-off rather than derive an analytical solution. We know from the central limit theorem that the mean value of independent random draws from some distribution tends towards a normal distribution, even if the original distribution is not normally distributed. Hence, the expected deviation off the true mean measured in standard deviations has an analytical solution. In our set-up, where the independent random draws are countries

and the original distribution is the distribution of the indicators we use, we draw from a finite distribution (all countries of the world). Yet it is possible to adjust the analytical solution for this using a finite population correction factor. What prevents us from deriving an analytical solution is that when we make random independent draws, we also randomly draw a weight for each observation (the population of each country). To our knowledge, analytically determining the distribution of the mean in this setting is not possible.

## Main results

In Figure 1 we plot the results from our simulations. We plot the expected error in the global statistic as a function of the share of the global population without data. The expected error increases linearly with the share of population without data. The linear fit suggests that if the share of the world's population on which one lacks data is *x*, then one should expect to be 0.37**x* standard deviation off the true mean, with the upper bound of this estimate being about *x* standard deviations off the true mean. Put reversely, if one is willing to tolerate being *y* standard deviations away from the true mean, then one can tolerate missingness on *y**2.7 (=*y**1/0.37) of the global population.

As an example, if one has data for half of the world's population, the global statistic will on expectation be 0.185 (0.37*0.5) standard deviation off the truth, and it could be as much as 0.5 standard deviation off the truth. The wide confidence interval reflects that when one only has data for some of the population, one might be lucky and get the mean right, or unlucky and be far off. Note that this is unrelated to the uncertainty surrounding the expected deviation – 0.37 -- for which the 95$^{th}$ percentile confidence interval is 0.36-0.39.

*Figure 1: Relationship between global data coverage and accuracy*

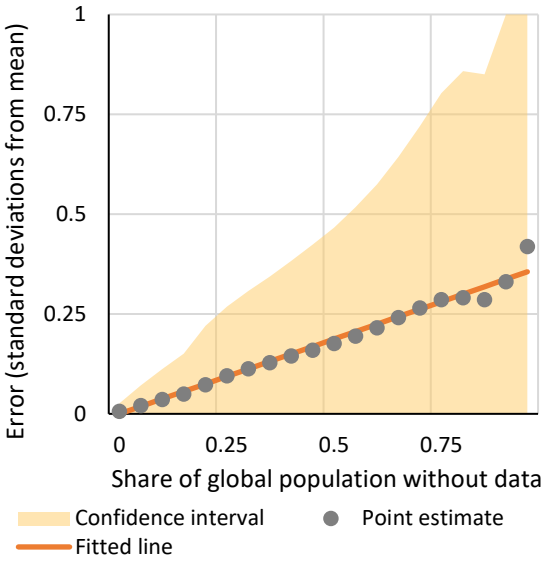
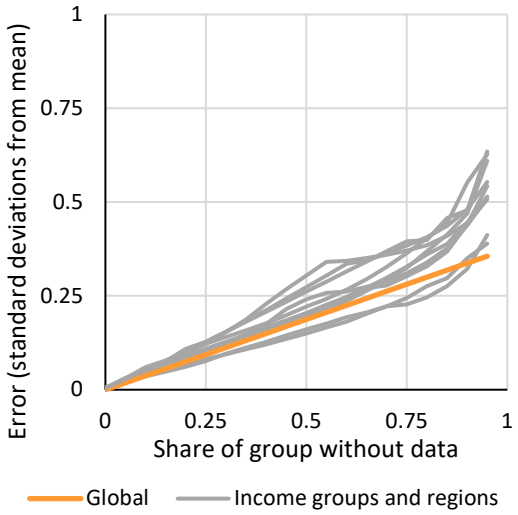



*Figure 2: Relationship between data coverage and accuracy by subgroups*




*Note*: Each grey line represents a World Bank region or a World Bank income group.

If one is interested in producing regional statistics (or any other sub-global statistic), the errors could be smaller or larger. On the one hand, to the extent that countries within regions are relatively alike, just having data on a few countries in the region might be sufficient to get the mean relatively right. This pushes the errors down relative to the global statistics. On the other hand, aggregating to a smaller number of countries means that for a fixed population share, there are fewer estimates to average over, which pushes

the uncertainty and expected error up. Figure 2 shows the estimated errors across the World Bank's geographical regions and four income groups. Evidently, the errors tend to be higher for regions and income groups than for the world as a whole. The only subgroups with lower errors are Europe & Central Asia and High-Income Countries.

### Alternative missingness assumptions

Our analysis so far may provide too optimistic errors if the data are systematically missing in the sense that the correlation between the probability of missingness and the value of the indicator is not zero. This is the case with SDG 1.1.1---the share living below the international poverty line---where less data is associated with higher poverty. Countries with at most five poverty estimates since 1980 have an average poverty rate of 32%, countries with 6-10 poverty estimates have an average poverty rate of 22%, and countries with more than 10 estimates have an average poverty rate of 4%. On the other hand, the errors we have presented so far might be too pessimistic if imputations are used to get proxy values for the countries with missing data.

In this section, we try to address these two issues. First, we assume that all missing values are imputed using regional averages. This is a common way of dealing with missing values in applied work. Second, we order countries by their share of missingness in an indicator in the years without full coverage and delete observations using this order rather than at random. The purpose is to only retain the data for the countries that are most likely to have data in any other year. The results from these two exercises are shown in Figure 3.

Using the empirical missingness from WDI does not systematically make the errors greater. This suggests that the probability of missingness in WDI is not systematically correlated with the indicator values and that our main results are not too optimistic. Yet, if for specific indicators the probability that data is missing is correlated to the values of the indicator, as with SDG 1.1.1, then we would be underestimating the error. Imputing with regional averages reduces the error by about 20%. If true imputations are better than using regional averages, then the expected error will be even lower.

*Figure 3: Relationship between data coverage and accuracy with alternative assumptions*

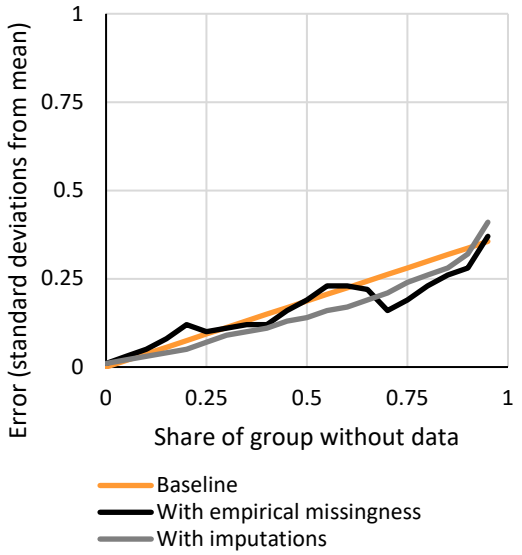
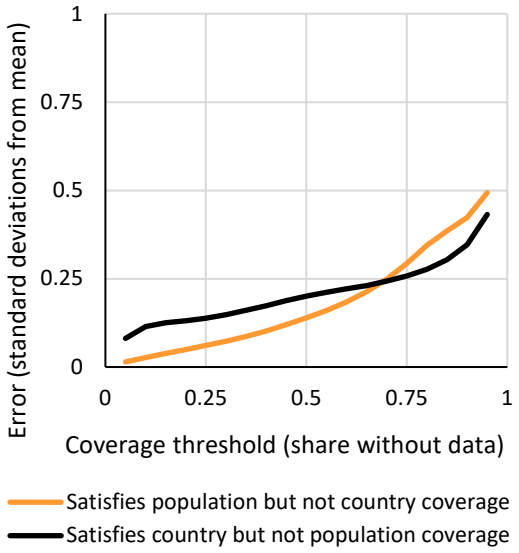



*Figure 4: Comparing error with population coverage and country coverage*



## Alternative coverage weights

Data producers sometimes use coverage rules based on the share of countries covered rather than the share of the global population covered. Suppose one is in doubt between which rule to use. A way to determine this would be for a given coverage requirement, say 50%, to compare the average error of the statistics that satisfy the country criterion but not the population criterion, and vice versa. Conditional on the same number of statistics passing the two coverage thresholds, ideally the coverage rule should minimize the error of those that pass.

Figure 4 tests this as a function of the threshold. We find that for any missing data tolerance less than 0.7, population weights work better than country weights. The intuition for this is as follows. If one has data on a large fraction of the world's countries, one might still get the statistic far off if one is lacking data on some of the most populous countries of the world. To the contrary, if one is willing to tolerate a large share of missingness, one might pass the bar by only having data on, say, India. If India is very different from the rest of the world, one can risk being quite off, and it might be better to average over more countries even if they account for a smaller population share. To see the latter argument more clearly, suppose one can choose between having data for one country of 40 million people or 4 countries of 10 million people. The latter would probably be better given that it would average out idiosyncrasies, outliers, and possible measurement error.

Does this mean that one ought to use country weights when tolerating high degrees of missingness? Actually not---intermediate options might be preferable. Notice that population weighting is equivalent to giving each country a weight of their population^1 while country weighting is equivalent to giving each country a weight of their population^0. By altering the exponent, we can get intermediate options. For example, using the square-root of the population size as weights would give the same weight to having data from two countries of 10 million and one country of 40 million (rather than half the weight, as population-weighting would do, or twice the weight, as country-weighting would do).

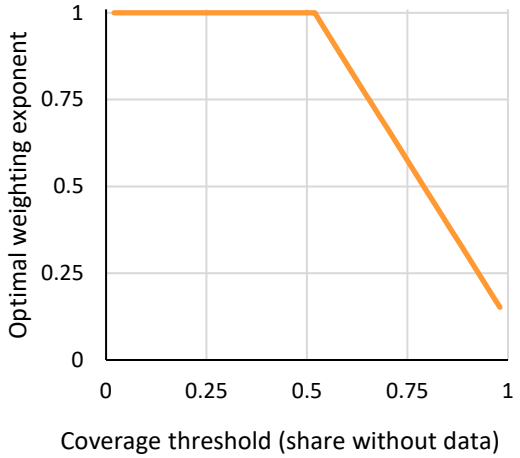
By making pairwise comparisons between such intermediate weighting schemes, we can see which weighting scheme for a given coverage threshold minimizes the expected error. The results are presented in Figure 5. The ideal weighting scheme is to use population weights for any missingness tolerance less than around 50%. After that, the optimal exponent declines. Using square root weights is optimal if one is willing to tolerate around 80% missingness. Using country weights (exponent = 0) is not optimal at any relevant missingness tolerance.

An alternative intermediate approach to taking an exponent of the population size is to condition coverage rules on data for the most populous countries being available. Comparing the performance of such coverage rules is a bit more challenging, given that for a certain population or country coverage threshold, the coverage rule is stricter. For example, the global statistics that cover at least half of the world's population *and* India on average cover a larger population share and are thus bound to be more precise than the global statistics that cover at least half of the world's population regardless of whether India is covered. Instead, we can compare a rule that conditions on data in India being present to a rule that does not condition on India being present, but has a slightly higher coverage threshold, such that they are equally stringent, meaning that equally many global statistics pass the rule.

Figure 6 makes such a comparison by replacing the x-axis with the share of all simulations that pass a given rule as the rule is made more lenient. Hence, for a given x-value, the various rules are equally stringent. The y-axis shows the average error of rules that pass. We continue to use India as an example country whose presence the global statistic is conditional on, though it could be replaced by other or multiple populous countries. The orange dotted line shows the average error of the global statistics that

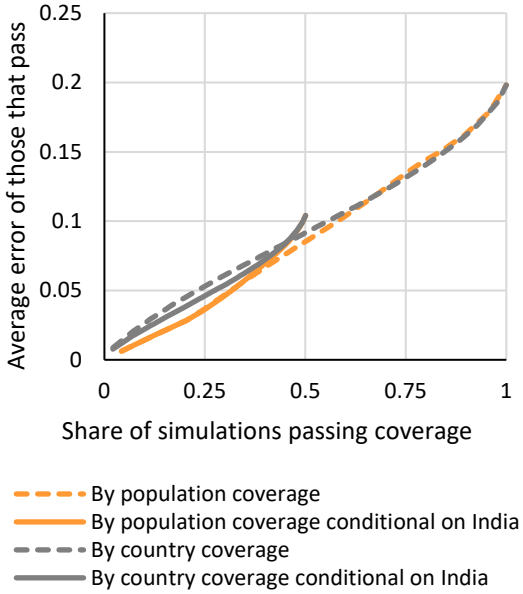
pass the regular population coverage rule as the population coverage threshold is decreased from 100% to 0%. The solid orange line shows the same for the subset of simulations that have data for India. Notice that when conditioning on India, about half of all simulations never pass the coverage threshold, as they do not have data for India.

*Figure 5: Optimal population weight as a function of coverage threshold*




*Note*: The figure shows the optimal country weight for a given coverage threshold. A y-axis value of *y* means an ideal country weight of population^*y*.

*Figure 6: Coverage rules that condition on data for the largest country being present*




For about the 20% of simulations that first pass the bar, the error is the same whether conditioning on India or not: if a simulation has at least 80% of the world's population covered, by necessity it must have India covered. After that, they are nearly identical until around 40% of the simulations pass the threshold. At that point, conditioning on India being present gives higher errors. This means that for a given stringency level, there is little evidence in favor of conditioning on data for the most populous country being present. If one is worried about having too imprecise global statistics, it would be better instead to increase the coverage threshold.

Though conditioning on India being present does not help when using population coverage, it more obviously might help when using country coverage. It only increases the country coverage strictness by one country but can substantially reduce the error. The grey lines in Figure 6 show that if using country coverage, conditioning on data being present for the most populous country helps. Yet since the solid grey line is above the orange lines, it is still preferable not to use country coverage rules at all – even when conditioning on data being present for the most populous country.

## Conclusion

In conclusion we offer some advice for how to decide when there is sufficient data to create global statistics. The most important to note is that there is no single threshold that can guide when to publish global statistics or not. The decision will depend on the context. In particular, we think the data producer should ask her- or himself the following questions:

- How large errors am I willing to tolerate?
- How pervasive is missing data in my indicators of interest?
- Is the probability of a country not having data likely correlated with the indicator of interest?
- [If producing time series] How much do the global statistics change from year to year and do the same countries consistently have missing values?
- [If missing data is imputed] How confident am I in the precision of the imputations?
- [If producing sub-global statistics] How large are the groups and how much of the variation happens between subgroups rather than within subgroups?

Jointly answering these questions should afford an approximate slope of the error as a function of the population coverage, as well as a ceiling on how large an error one is willing to tolerate. Judging from the table comparing standard deviations with original units, our (admittedly, subjective) take is that errors should never on expectation exceed 0.25 standard deviation. Even in the less optimistic cases we presented, this roughly corresponds to not publishing statistics when data for less than half of the relevant population is available. For certain purposes, such as comparing statistics over time, it is likely that much lower errors are needed. A corollary of these recommendations is that it is always optimal to use the share of the global population covered rather than the share of countries covered as the coverage threshold. A corollary of this is that rather than conditioning on data being present for populous countries, it would be better to increase the coverage threshold.

## References


Dang, Hai-Anh H., Dean Jolliffe, and Calogero Carletto (2019): "Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments." *Journal of Economic Surveys* 33 (3): 757-797. https://doi.org/10.1111/joes.12307

Dang, Hai-Anh H., John Pullinger, Umar Serajuddin, and Brian Stacy (2021): "Statistical Performance Indicators and Index." World Bank Policy Research Working Paper 9570. Washington, D.C.: World Bank Group. https://openknowledge.worldbank.org/handle/10986/35301

Dang, Hai-Anh H., and Umar Serajuddin (2020): "Tracking the Sustainable Development Goals: Emerging measurement challenges and further reflections." *World Development* 127: 104570. https://doi.org/10.1016/j.worlddev.2019.05.024

MacFeely, Steve (2018): "The 2030 Agenda: An Unprecedented Statistical Challenge." Friedrich-Ebert-Stiftung, Global Policy and Development. http://library.fes.de/pdf-files/iez/14796.pdf

Sachs, Jeffrey D. (2012): "From millennium development goals to sustainable development goals." *The Lancet* 379 (9832): 2206-2211. https://doi.org/10.1016/S0140-6736(12)60685-0

Serajuddin, Umar, Hiroki Uematsu, Christina Wieser, Nobuo Yoshida, and Andrew Dabalen (2015): "Data Deprivation: Another Deprivation to End." World Bank Policy Research Working Paper 7252. Washington, D.C: World Bank Group. https://openknowledge.worldbank.org/handle/10986/21867

World Bank (2021): "World Development Report 2021: Data for Better Lives." Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/35218.

# Appendix

### Table A.1: Indicators used for the analysis

| WDI code | Description |
|---|---|
| AG.LND.AGRI.ZS | Agricultural land (% of land area) |
| AG.LND.ARBL.HA.PC | Arable land (hectares per person) |
| AG.LND.ARBL.ZS | Arable land (% of land area) |
| AG.LND.CROP.ZS | Permanent cropland (% of land area) |
| AG.LND.FRST.ZS | Forest area (% of land area) |
| AG.PRD.CROP.XD | Crop production index (2014-2016 = 100) |
| AG.PRD.FOOD.XD | Food production index (2014-2016 = 100) |
| AG.PRD.LVSK.XD | Livestock production index (2014-2016 = 100) |
| AG.YLD.CREL.KG | Cereal yield (kg per hectare) |
| BX.KLT.DINV.WD.GD. | Foreign direct investment, net inflows (% of GDP) |
| BX.TRF.PWKR.DT.GD. | Personal remittances, received (% of GDP) |
| EG.EGY.PRIM.PP.KD | Energy intensity level of primary energy (MJ/$2011 PPP GDP) |
| EG.ELC.ACCS.RU.ZS | Access to electricity, rural (% of rural population) |
| EG.ELC.ACCS.UR.ZS | Access to electricity, urban (% of urban population) |
| EG.ELC.ACCS.ZS | Access to electricity (% of population) |
| EG.ELC.RNEW.ZS | Renewable electricity output (% of total electricity output) |
| EG.FEC.RNEW.ZS | Renewable energy consumption (% of total final energy consumption) |
| EN.ATM.CO2E.GF.ZS | CO2 emissions from gaseous fuel consumption (% of total) |
| EN.ATM.CO2E.LF.ZS | CO2 emissions from liquid fuel consumption (% of total) |
| EN.ATM.CO2E.PC | CO2 emissions (metric tons per capita) |
| EN.ATM.CO2E.SF.ZS | CO2 emissions from solid fuel consumption (% of total) |
| EN.ATM.GHGT.ZG | Total greenhouse gas emissions (% change from 1990) |
| EN.ATM.METH.AG.ZS | Agricultural methane emissions (% of total) |
| EN.ATM.METH.EG.ZS | Energy related methane emissions (% of total) |
| EN.ATM.NOXE.AG.ZS | Agricultural nitrous oxide emissions (% of total) |
| EN.ATM.NOXE.EG.ZS | Nitrous oxide emissions in energy sector (% of total) |
| EN.ATM.PM25.MC.ZS | PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total) |
| EN.POP.DNST | Population density (people per sq. km of land area) |
| EN.URB.LCTY.UR.ZS | Population in the largest city (% of urban population) |
| EP.PMP.DESL.CD | Pump price for diesel fuel (US$ per liter) |
| EP.PMP.SGAS.CD | Pump price for gasoline (US$ per liter) |
| ER.H2O.FWAG.ZS | Annual freshwater withdrawals, agriculture (% of total freshwater withdrawal) |
| ER.H2O.FWDM.ZS | Annual freshwater withdrawals, domestic (% of total freshwater withdrawal) |
| ER.H2O.FWIN.ZS | Annual freshwater withdrawals, industry (% of total freshwater withdrawal) |
| ER.H2O.FWTL.ZS | Annual freshwater withdrawals, total (% of internal resources) |
| ER.LND.PTLD.ZS | Terrestrial protected areas (% of total land area) |
| ER.PTD.TOTL.ZS | Terrestrial and marine protected areas (% of total territorial area) |
| FB.CBK.BRCH.P5 | Commercial bank branches (per 100,000 adults) |
| IC.BUS.DFRN.XQ | Ease of doing business score (0 = lowest performance to 100 = best performance) |
| IC.BUS.DISC.XQ | Business extent of disclosure index (0=less disclosure to 10=more disclosure) |
| IC.CRD.INFO.XQ | Depth of credit information index (0=low to 8=high) |
| IC.CRD.PRVT.ZS | Private credit bureau coverage (% of adults) |
| IC.CRD.PUBL.ZS | Public credit registry coverage (% of adults) |
| IC.ELC.TIME | Time required to get electricity (days) |
| IC.EXP.CSBC.CD | Cost to export, border compliance (US$) |
| IC.EXP.CSDC.CD | Cost to export, documentary compliance (US$) |
| IC.EXP.TMBC | Time to export, border compliance (hours) |
| IC.EXP.TMDC | Time to export, documentary compliance (hours) |
| IC.IMP.CSBC.CD | Cost to import, border compliance (US$) |
| IC.IMP.CSDC.CD | Cost to import, documentary compliance (US$) |
| IC.IMP.TMBC | Time to import, border compliance (hours) |
| IC.IMP.TMDC | Time to import, documentary compliance (hours) |
| IC.LGL.CRED.XQ | Strength of legal rights index (0=weak to 12=strong) |
| IC.LGL.DURS | Time required to enforce a contract (days) |

| | |
|---|---|
| IC.PRP.DURS | Time required to register property (days) |
| IC.PRP.PROC | Procedures to register property (number) |
| IC.REG.COST.PC.FE.ZS | Cost of business start-up procedures, female (% of GNI per capita) |
| IC.REG.COST.PC.MA.Z | Cost of business start-up procedures, male (% of GNI per capita) |
| IC.REG.COST.PC.ZS | Cost of business start-up procedures (% of GNI per capita) |
| IC.REG.DURS | Time required to start a business (days) |
| IC.REG.DURS.FE | Time required to start a business, female (days) |
| IC.REG.DURS.MA | Time required to start a business, male (days) |
| IC.REG.PROC | Start-up procedures to register a business (number) |
| IC.REG.PROC.FE | Start-up procedures to register a business, female (number) |
| IC.REG.PROC.MA | Start-up procedures to register a business, male (number) |
| IC.TAX.DURS | Time to prepare and pay taxes (hours) |
| IC.TAX.LABR.CP.ZS | Labor tax and contributions (% of commercial profits) |
| IC.TAX.OTHR.CP.ZS | Other taxes payable by businesses (% of commercial profits) |
| IC.TAX.PAYM | Tax payments (number) |
| IC.TAX.PRFT.CP.ZS | Profit tax (% of commercial profits) |
| IC.TAX.TOTL.CP.ZS | Total tax and contribution rate (% of profit) |
| IT.CEL.SETS.P2 | Mobile cellular subscriptions (per 100 people) |
| IT.MLT.MAIN.P2 | Fixed telephone subscriptions (per 100 people) |
| IT.NET.SECR.P6 | Secure Internet servers (per 1 million people) |
| IT.NET.USER.ZS | Individuals using the Internet (% of population) |
| MS.MIL.TOTL.TF.ZS | Armed forces personnel (% of total labor force) |
| NV.AGR.TOTL.ZS | Agriculture, forestry, and fishing, value added (% of GDP) |
| NV.IND.TOTL.ZS | Industry (including construction), value added (% of GDP) |
| NV.SRV.TOTL.ZS | Services, value added (% of GDP) |
| NY.ADJ.AEDU.GN.ZS | Adjusted savings: education expenditure (% of GNI) |
| NY.ADJ.DCO2.GN.ZS | Adjusted savings: carbon dioxide damage (% of GNI) |
| NY.ADJ.DKAP.GN.ZS | Adjusted savings: consumption of fixed capital (% of GNI) |
| NY.ADJ.DMIN.GN.ZS | Adjusted savings: mineral depletion (% of GNI) |
| NY.ADJ.DNGY.GN.ZS | Adjusted savings: energy depletion (% of GNI) |
| NY.ADJ.DPEM.GN.ZS | Adjusted savings: particulate emission damage (% of GNI) |
| NY.GDP.COAL.RT.ZS | Coal rents (% of GDP) |
| NY.GDP.DEFL.KD.ZG | Inflation, GDP deflator (annual %) |
| NY.GDP.FRST.RT.ZS | Forest rents (% of GDP) |
| NY.GDP.MINR.RT.ZS | Mineral rents (% of GDP) |
| NY.GDP.MKTP.KD.ZG | GDP growth (annual %) |
| NY.GDP.NGAS.RT.ZS | Natural gas rents (% of GDP) |
| NY.GDP.PCAP.CD | GDP per capita (current US$) |
| NY.GDP.PCAP.KD | GDP per capita (constant 2010 US$) |
| NY.GDP.PCAP.KD.ZG | GDP per capita growth (annual %) |
| NY.GDP.PETR.RT.ZS | Oil rents (% of GDP) |
| NY.GDP.TOTL.RT.ZS | Total natural resources rents (% of GDP) |
| NY.GNP.PCAP.CD | GNI per capita, Atlas method (current US$) |
| SE.PRM.DURS | Primary education, duration (years) |
| SE.SEC.DURS | Secondary education, duration (years) |
| SG.GEN.PARL.ZS | Proportion of seats held by women in national parliaments (%) |
| SH.ALC.PCAP.LI | Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age) |
| SH.DTH.COMM.ZS | Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of |
| SH.DTH.INJR.ZS | Cause of death, by injury (% of total) |
| SH.DTH.NCOM.ZS | Cause of death, by non-communicable diseases (% of total) |
| SH.DYN.0509 | Probability of dying among children ages 5-9 years (per 1,000) |
| SH.DYN.1014 | Probability of dying among adolescents ages 10-14 years (per 1,000) |
| SH.DYN.1519 | Probability of dying among adolescents ages 15-19 years (per 1,000) |
| SH.DYN.2024 | Probability of dying among youth ages 20-24 years (per 1,000) |
| SH.DYN.MORT | Mortality rate, under-5 (per 1,000 live births) |
| SH.DYN.NCOM.ZS | Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70 (%) |
| SH.DYN.NMRT | Mortality rate, neonatal (per 1,000 live births) |
| SH.H2O.BASW.ZS | People using at least basic drinking water services (% of population) |
| SH.IMM.HEPB | Immunization, HepB3 (% of one-year-old children) |

| | |
|---|---|
| SH.IMM.IDPT | Immunization, DPT (% of children ages 12-23 months) |
| SH.IMM.MEAS | Immunization, measles (% of children ages 12-23 months) |
| SH.MMR.RISK | Lifetime risk of maternal death (1 in: rate varies by country) |
| SH.MMR.RISK.ZS | Lifetime risk of maternal death (%) |
| SH.STA.AIRP.P5 | Mortality rate attributed to household and ambient air pollution, age-standardized (per 100,000 |
| SH.STA.BASS.ZS | People using at least basic sanitation services (% of population) |
| SH.STA.DIAB.ZS | Diabetes prevalence (% of population ages 20 to 79) |
| SH.STA.MMRT | Maternal mortality ratio (modeled estimate, per 100,000 live births) |
| SH.STA.ODFC.ZS | People practicing open defecation (% of population) |
| SH.STA.POIS.P5 | Mortality rate attributed to unintentional poisoning (per 100,000 population) |
| SH.STA.SUIC.P5 | Suicide mortality rate (per 100,000 population) |
| SH.STA.TRAF.P5 | Mortality caused by road traffic injury (per 100,000 population) |
| SH.STA.WASH.P5 | Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 |
| SH.TBS.DTEC.ZS | Tuberculosis case detection rate (%, all forms) |
| SH.TBS.INCD | Incidence of tuberculosis (per 100,000 people) |
| SH.UHC.SRVS.CV.XD | UHC service coverage index |
| SH.XPD.CHEX.GD.ZS | Current health expenditure (% of GDP) |
| SH.XPD.CHEX.PP.CD | Current health expenditure per capita, PPP (current international $) |
| SH.XPD.EHEX.CH.ZS | External health expenditure (% of current health expenditure) |
| SH.XPD.EHEX.PP.CD | External health expenditure per capita, PPP (current international $) |
| SH.XPD.GHED.CH.ZS | Domestic general government health expenditure (% of current health expenditure) |
| SH.XPD.GHED.GD.ZS | Domestic general government health expenditure (% of GDP) |
| SH.XPD.GHED.PP.CD | Domestic general government health expenditure per capita, PPP (current international $) |
| SH.XPD.PVTD.CH.ZS | Domestic private health expenditure (% of current health expenditure) |
| SH.XPD.PVTD.PP.CD | Domestic private health expenditure per capita, PPP (current international $) |
| SL.AGR.EMPL.ZS | Employment in agriculture (% of total employment) (modeled ILO estimate) |
| SL.EMP.1524.SP.ZS | Employment to population ratio, ages 15-24, total (%) (modeled ILO estimate) |
| SL.EMP.MPYR.ZS | Employers, total (% of total employment) (modeled ILO estimate) |
| SL.EMP.SELF.ZS | Self-employed, total (% of total employment) (modeled ILO estimate) |
| SL.EMP.TOTL.SP.ZS | Employment to population ratio, 15+, total (%) (modeled ILO estimate) |
| SL.EMP.VULN.ZS | Vulnerable employment, total (% of total employment) (modeled ILO estimate) |
| SL.EMP.WORK.ZS | Wage and salaried workers, total (% of total employment) (modeled ILO estimate) |
| SL.FAM.WORK.ZS | Contributing family workers, total (% of total employment) (modeled ILO estimate) |
| SL.IND.EMPL.ZS | Employment in industry (% of total employment) (modeled ILO estimate) |
| SL.SRV.EMPL.ZS | Employment in services (% of total employment) (modeled ILO estimate) |
| SL.TLF.ACTI.1524.ZS | Labor force participation rate for ages 15-24, total (%) (modeled ILO estimate) |
| SL.TLF.ACTI.ZS | Labor force participation rate, total (% of total population ages 15-64) (modeled ILO estimate) |
| SL.TLF.CACT.FM.ZS | Ratio of female to male labor force participation rate (%) (modeled ILO estimate) |
| SL.TLF.CACT.ZS | Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate) |
| SL.UEM.1524.ZS | Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate) |
| SL.UEM.TOTL.ZS | Unemployment, total (% of total labor force) (modeled ILO estimate) |
| SM.POP.TOTL.ZS | International migrant stock (% of population) |
| SP.ADO.TFRT | Adolescent fertility rate (births per 1,000 women ages 15-19) |
| SP.DYN.CBRT.IN | Birth rate, crude (per 1,000 people) |
| SP.DYN.CDRT.IN | Death rate, crude (per 1,000 people) |
| SP.DYN.IMRT.IN | Mortality rate, infant (per 1,000 live births) |
| SP.DYN.LE00.IN | Life expectancy at birth, total (years) |
| SP.DYN.TFRT.IN | Fertility rate, total (births per woman) |
| SP.RUR.TOTL.ZS | Rural population (% of total population) |
| SP.URB.GROW | Urban population growth (annual %) |
| SP.URB.TOTL.IN.ZS | Urban population (% of total population) |
| TG.VAL.TOTL.GD.ZS | Merchandise trade (% of GDP) |