United Nations ESCAP
Economic and Social Commission for Asia and the Pacific

SD/WP/04/March 2017 STATISTICS DIVISION

Working Paper Series

A weighted extrapolation method for measuring the SDGs progress

Author:

Arman Bidarbakht Nia
Statistician, ESCAP-Statistics Division

Acknowledgements:

The author is grateful to the following reviewers for their valuable contributions to the paper: Asian Development Bank (Arturo Martinez Jr), United Nations Development Programme (Bishwa Nath Tiwari), and ESCAP (Sudip Ranjan Basu, Daniel Clark, Margarita Guerrero, Wenyun Qian, Yanhong Zhang). The author also would like to thank staff of the ESCAP- Statistics Division for their support and useful discussions.

For more information, please contact:
Statistics Division
Economic and Social Commission for Asia and the Pacific (ESCAP)
United Nations Building, Rajadamnern Nok Avenue, Bangkok 10200, Thailand
Email: stat.unescap@un.org

Abstract

In the absence of timely data on the development indicators, the UN system, including various regional and global organizations, often apply statistical and mathematical methods to extrapolate existing data series to fill the data gaps. This paper is proposing a time-related weighting system that increases accuracy of extrapolation methods, as used by ESCAP and its partner organisations in recent years to monitor the regional progress towards internationally agreed development goals. The results from a simulation study on a set of indicators across all Sustainable Development Goals (SDGs) prove substantive gain in accuracy by applying the time-related weighting factors.

Contents

Previous Working Papers:

1. Working Paper Series (SD/WP/01/May 2015): Measuring Trade in Value Added: Concepts, Estimation and Analysis *(Published 28 May 2015)*

2. Working Paper Series (SD/WP/02/April 2016): Asymmetries in International Merchandise Trade Statistics: A case study of selected countries in Asia-Pacific *(Published 5 April 2016)*

3. Working Paper Series (SD/WP/03/September 2016): Business Demography Statistics: A case study of selected countries in Asia-Pacific *(Published 20 September 2016)*

1. Introduction

In order to plan for achieving the desirable goal, it is essential not only to set high standards (targets) but also to acknowledge the departure point and current position (baseline). At the dawn of the Sustainable Development Goals (SDGs) that established aspirational goals and targets for the international community, it is of paramount importance to understand where the Asia-Pacific region stands in relation to the SDGs targets and where it will be by 2030. Lack of timely data on development indicators was a challenge for monitoring the Millennium Development Goals (MDGs) and continues to be so in any regional review of progress towards achieving the SDGs that would be based on the evolving, challenging and hard-to-measure global indicator framework[1]. In the face of insufficient and infrequent national data on development indicators, regional and global development partners have resorted to extrapolating national indicator values for the years for which data is missing and have used varying methods to do so. .

A diverse range of methods-- from simple mathematical techniques to model-based statistical methods-- are available for extrapolating socio-economic indicators. Although metadata on methods used by agencies is not always available, existing information suggest that practices vary widely across topics and agencies. For instance, World Bank interpolates income values for missing years by projecting incomes for that year based on GDP growth[2]. Another example is the use of time series models by the International Monetary Fund (IMF) for nowcasting quarterly national accounts indicators[3]. These methods are well established within specialized agencies that deal with a lesser number of indicators (compared to those required by regional or global reviews) and developed to address specific topics. However, for the purpose of regional monitoring, the ESCAP regional database[4] does not lend itself to most sophisticated statistical models. Some of the unique characteristics of the regional database and MDG/SDG report requirements that hinder application of more accurate extrapolation methods are low frequency of national data and lack of auxiliary variables. Given the large number of indicators in the database for which national data is missing this study aims to develop a low-cost (computational) and easy-to-understand method that can be applied across a diverse range of indicators.

The two extrapolation methods used by ESCAP in the past are based on geometric growth rates and log-transformed regression models. Although both methods are statistically sound and widely practiced, the underlying assumptions are not necessarily realistic. The geometric mean assumes that changes between two periods differ by a constant ratio, which is not necessarily the case for most (if not all) of development issues. Among log-transformed regression models, only an S-shaped logistic function is based on the more realistic assumption that indicator values increase exponentially at early stage and the growth rate then reduces as indicator values increase/decrease. Nonetheless, when indicator values are not proportions, logarithmic transformation assumes an exponential growth rate throughout the distribution.

This paper proposes that the importance of the indicator values should be proportional to their *recency*, i.e. how recent the data are, when incorporated in the model. The proposed extrapolation method is making optimum use of time, as the only auxiliary variable available, in nowcasting and predicting future values of development indicators by introducing time-related weights that improves the predictive power of the model.

[1] Available: http://unstats.un.org/sdgs/indicators/indicators-list/ [Accessed 23 March 2017]
[2] Available: https://openknowledge.worldbank.org/handle/10986/25078 [Accessed 23 March 2017]
[3] Available: http://www.imf.org/external/ns/search.aspx?hdCountrypage=&NewQuery=extrapolation&search=Search&filter_val=N&col=SITENG&collection=SITENG&lan=eng&iso=&requestfrom=&countryname=&f= [Accessed 23 march2017]
[4] Available: http://data.unescap.org/escap_stat/ [Accessed 23 March 2017]

The same technical issues remain relevant to the proposed model. The first issue is the error term in the model. Statistical models require a sufficient number of data points to provide reasonably accurate estimates while in many cases we have only two data points for one country. Therefore, one has to choose between a model-based estimate with a large error term or a smaller sample of national data for regional aggregates. The second challenge is that one single method cannot provide the most accurate estimate for all indicators and there is also lack of comparative study to inform our decision on the choice of appropriate method. A flexible approach has been taken in this paper that allows for identifying the best performing method for each indicator.

The main objectives of this working paper are to: (a) review the existing practices for extrapolating development indicators for regional monitoring in the Asia-Pacific region; (b) propose a new weighting system that enhances the predictive power of the log-transformed models; and (c) identify the most accurate and robust method for different development indicators. The rest of the paper is organized as follows: Section II provides a review of the geometric and log-transformed regression methods. Section III proposes a time-related weighting system and finally section IV presents findings from a simulation study that compares various extrapolation methods and identifies the best method for each indicator.

2. Selected Methods for Extrapolation

Suppose that n data points are available on indicator I for a given country/region over a period of T years and we are interested in extrapolating the indicator value to the year t_{n+a} *(a=1,2,..).*

$T = t_n - t_1$ where t_n and t_1 are the latest and the earliest years, respectively, for which data on indicator I are available respectively.

2.1 Geometric growth:

Geometric growth method assumes that the observed value of the indicator changes by a constant ratio from one year to another and estimates average annual growth rate as:

$$r_{GM} = \left(\frac{I_n}{I_1}\right)^{\frac{1}{T}} - 1$$

Where I_1 and I_n are indicator values for the first and the last years, respectively. And the indicator value for the year t_{n+a} is estimated as

$$\hat{I}_{GM} = I_n \times \left(\frac{I_n}{I_1}\right)^{\frac{a}{T}}$$

2.2 Log-transformed Regression Method

In this method, the average annual growth rate (r_1) is estimated by fitting a linear regression model of transformed indicator values over normalized time values:

$$L_i = r_0 + r_1 t_i^* + \varepsilon_i \ \ (i = 1,2,..,n)$$

Where L_i is the transformed value of the indicator I for the year t_i. The transformation is done in two steps: in the first step the indicator I is converted to Y by dividing it by an appropriate scale to standardize it to a scale of 0 to 1. For example, indicators shown in percentage are divided by 100 and indicators expressed in other rates such as "per 1000", "per 100,000" divided accordingly by 1000 and 100,000. For some indicators that cannot be expressed in the form of probability or rate, no transformation is applied. In the second step a natural log transformation is applied to indicators that needed transformation in the first step:

$$L = \ln\left(\frac{Y}{1-Y}\right) \; if\ I\ is\ probability\ or\ rate$$

$$= ln(I)\ if\ I\ is\ ratio\ of\ proportions\ \ (odds\ ratio)$$

$$= \mathrm{I}\ \ Otherwise$$

And t is normalized by subtracting the mean year $\bar{t}$:

$t_i^* = t_i - \bar{t} \quad (i = 1,2,..,n)$

Applying estimated parameters from the model, $(\widehat{r_0}, \widehat{r_1})$, the extrapolated values of indicator I for year t_{n+a} is obtained as follows:

$$\hat{I}_{Reg} = scale \times \frac{EXP}{1+EXP} \quad if\ I\ is\ rate\ or\ probability$$

$$= scale \times EXP \quad if\ I\ is\ odds\ ratio$$

$$= \widehat{r_0} + \widehat{r_1} \times (t_{n+a} - \bar{t}) \quad if\ I\ is\ odds\ ratio$$

where $EXP = exp(\widehat{r_0} + \widehat{r_1} \times (t_{n+a} - \bar{t}))$ and "*scale*" is the appropriate scaling factor used in the first step transformation of the indicator (e.g., 100, 1000, 100,000). For further details on the regression method, readers may refer to the Asia-Pacific Regional MDG Report 2011/12[5].

2.3 Composite Estimation Method:

Since the purpose of regression models here is only for predicting the missing values, it would seem theoretically correct to assume that models with greater R-square provide more accurate estimates. If this assumption is true, a composite estimator that combines estimators from two methods, with R-square being used as the weight of model-based estimates, is expected to provide optimum more accurate estimator; i.e.:

$$\hat{I}_{Comp} = (1 - R^2) \times \hat{I}_{Mean} + R^2 \times \hat{I}_{Model}$$

In this equation, $\hat{I}_{Model}$ can be any of the model-based estimation methods that we discuss in this paper and $\hat{I}_{\mathrm{Mean}}$ any of the other mathematical methods. It is obvious that as the R-square approaches 1, the composite estimate approaches the model-based estimate and for the models with insignificant R-squares, the composite estimate is equal to the mathematical estimate. For the purpose of this study, weighted and unweighted regression models were combined with the weighted geometric mean as the best mathematical estimator to construct two types of composite estimators, as discussed in the next section.

[5] Available: http://www.unescap.org/sites/default/files/MDG-Report2011-12_0.pdf [Accessed 23 March2017]

3. Time-related Weights

In the aforementioned methods, observed rates of change in the indicator values in all periods for which data is available are given equal importance (weight) in predicting the future values. For instance, if we aim to predict poverty rate in 2015 and data from 2000 to 2014 are available, the annual rate of change in poverty observed in 2001 and 2014 are given equal importance in predicting poverty in 2015.

When extrapolation of indicator values is done in the context of rates of progress towards achieving a fixed target value, an important consideration is that as we approach the target value, it is expected that a different rate of change (often decreasing) will be observed. In other words, as we get closer to the target value, it takes more effort to make the same progress as we made in previous periods. Therefore, it is more plausible to produce estimates which account for the *recency* of data. The time-related weights are proposed to work as a multiplier that inflates the rate of change in each period proportional to its temporal distance to the target year (t_{n+a}). The time-related weight for the i^{th} observation for given country/region is proposed as:

$$w_i = \frac{(t_{n+a} - t_1)}{(t_{n+a} - t_i)} \quad (i = 1,2,..n)$$

With this weighting factor, more recent values are given greater weight in the estimation. The resulting estimators relevant to this study are:

<u>Weighted geometric mean:</u>

$$\hat{I}_{WGM} = I_n \times \left(\prod_{i=2}^{n} \left[\frac{I_i}{I_{i-1}} \right]^{w_i} \right)^{\frac{a}{W}}$$

Where $W = \sum_i w_i$.

<u>Regression model estimator:</u>

$$\widehat{r_1} = \frac{\sum_i \mathrm{w_i t_i^* L_i} - \frac{\sum_i \mathrm{w_i t_i^*} \sum_i \mathrm{w_i L_i}}{\mathrm{W}}}{\sum_i \mathrm{w_i}\, \mathrm{t_i^{*2}} - \frac{(\sum_i \mathrm{w_i t_i^*})^2}{\mathrm{W}}}$$

Note that with $w_i = 1$ and $W = n$, the two weighted estimators will reduce to their unweighted forms; $\hat{I}_{GM}$ for the geometric mean and ordinary least square (OLS) estimates for the regression model.

4. Results

The methods were applied to selected SDG indicators, for which data were available to compare the predictive power of each method and to assess the effect of the proposed weighting system on accuracy of extrapolation. Seventy-seven (77) indicators shown in Table 1 were selected for the simulation study. The criteria for selecting these indicators are: (1) there are at least three observations in the past 15 years in order to be able to fit a non-linear growth model and (2) data are available for at least 10 countries in the Asia-Pacific region to allow for calculation of Mean Absolute Errors (MAE).

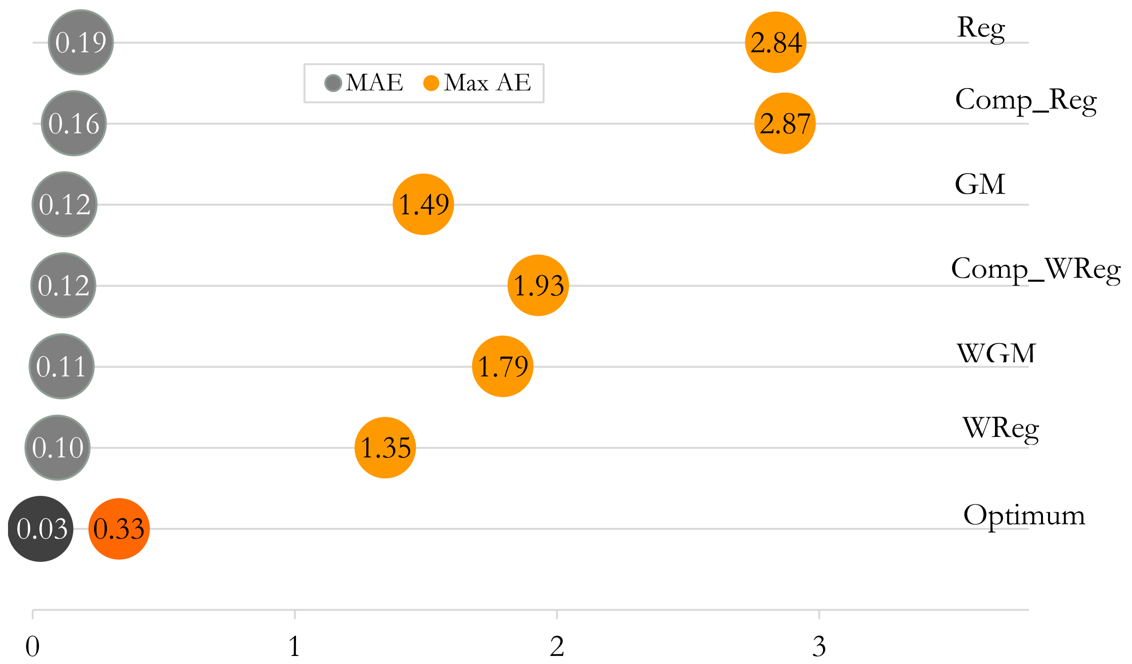
To compare accuracy of predictions across different methods, the latest observation is dropped from each data series and remaining observations (at least two) are used to predict the omitted observation. The deviation of the predictions from real values of the latest year is used as a measure of accuracy for each method. Assume that for a given country there are n observations on indicator I. For a given method, if we estimate the latest observation (I_n) by $\hat{I}_n$, then the Absolute Error (AE) can be measured as:

$$AE = \left| I_n - \hat{I}_n \right|$$

MAE, the average of the AE's over all countries for which data on indicator I is available, is used as a measure of accuracy of the given method for indicator I.

Figure 1 shows the results from the simulation comparing MAEs and maximum value of AEs for each extrapolation method. In general, time-related weighting has improved accuracy of the estimates across all methods. The weighted geometric mean, weighted regression and composite of the two estimators yield the three lowest MAEs. This proves that using time variable (as the only auxiliary variable available for extrapolation) as both explanatory and temporal distance information can significantly improve accuracy of extrapolation methods. The time as temporal distance is incorporated in the model as weighting factor to borrow strength from the most recent changes for predicting the future. The weighting performed best when regression models were used. While the unweighted regression estimator is ranked as the least accurate and least robust (the highest MAE and the widest gap between MAE and maximum AE), the weighted regression in contrast is the most accurate and robust estimator with 0.1 MAE and the largest possible AE of 1.35.

Figure 1 - The mean and maximum absolute error for various extrapolation methods[6]



[6] GM: geometric mean; WGM: weighted geometric mean; Reg: regression model; WReg: weighted regression model; Comp_Reg: composite estimation combining WGM and Reg; Comp_WReg: Composite estimation combining WGM and WReg. Optimum: applied the most accurate estimator to each indicator

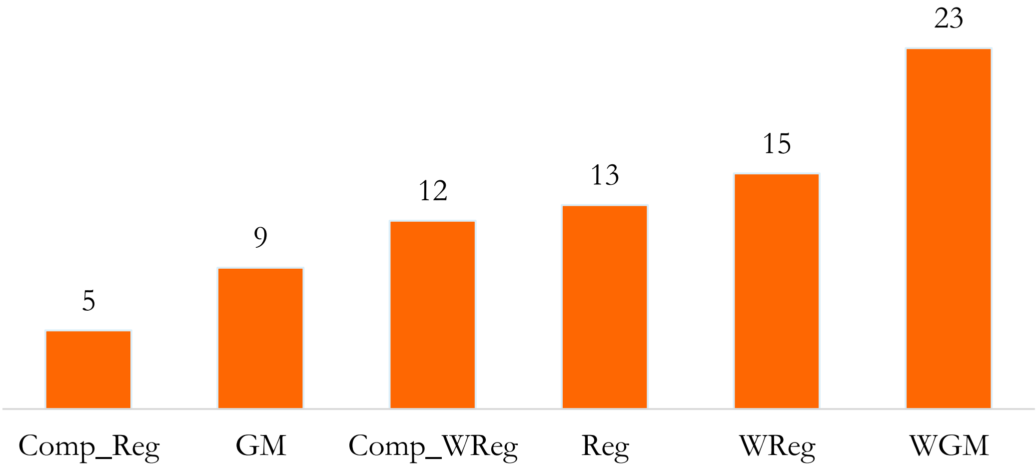
The results in Figure 1 also suggest that we did not gain accuracy by applying the proposed composite estimator. In other words, R-square is not necessarily a reliable criterion for selecting among the possible models.

This could be due to the lack of sufficient number of observations in most of the cases. As is expected, no one single method can be identified as the most accurate across all indicators. Even methods such as unweighted regression model, ranked as the least accurate on average, can perform the best for some indicators. Figure 2 shows the distribution of the 77 selected indicators by type of the best extrapolation method. The two most accurate predictors (WGM and WReg) were the best method for over 60% of the indicators. However, alternative methods such as unweighted regression and composite method with weighted regression performed the best for a significant number of indicators. The study did not find any common markers that would characterize indicators that performed well under a given method.

In order to achieve the highest accuracy for the Asia-Pacific 2016 SDG baseline report[7], the most accurate extrapolation method was selected at the indicator level (optimum method[8]). With this flexible approach, the overall MAE went down to only 3% and, more importantly, the maximum AE was only 0.33-- meaning no estimates at the country level was expected to have an absolute error larger than 0.33 which is significantly lower than maximum AE for any of the methods (Figure 1).

As the computational cost of the extrapolation method to be adopted may be a concern for automating the reporting process, and given the fact that over 60% of the indicators lend themselves to one of the two weighted methods, a practical alternative was also evaluated in this study. By replacing the Comp_Reg and GM methods with WGM and the Comp_WReg and Reg methods with WReg the resulting MAE and maximum AE are 7% and 1.79, respectively. This approach produces results with an average error of lower than any of the single methods but the maximum error committed for one indicator may go beyond those of weighted regression.

Figure 2 - Distribution of indicators by type of the best extrapolation method



[7] Extrapolations for the baseline and target years (2015 and 2030) were done at the national level to make the best use of existing data. The national indicator values were then used to calculate regional estimates and only regional estimates are used for the reporting.

[8] Geometric mean method was only applied to indicators with positive values. Therefore, the best method for indicators with negative values was selected only among the two regression estimators.

5. Conclusion and Way Forward

In conclusion, the proposed time-related weighting factors facilitate optimum contribution of time (as the only auxiliary variable available) in extrapolating development indicators for the purpose of regional/global monitoring and review. The results from simulation study on 77 indicators from the evolving global SDG indicator framework[9] showed that weighted methods outperform all other methods that have been used in the past for regional monitoring. Nonetheless, no one method can be identified as the most accurate across all indicators. A flexible approach that chooses the best performing method for each indicator can lead to a substantial gain in overall accuracy and robustness of extrapolation results. Finally, a practical approach that limits the choices to only two weighted methods, namely weighted geometric mean and weighted regression model, may be adopted to decrease the number of computations required.

Two possibilities may be considered for further improvements in the methods proposed in this paper. The first is exploring different functional forms for the time-related weights. The proposed weights are constructed by an exponential function, meaning that relative importance of the values increase exponentially over time. This type of weighting system can be very sensitive to structural changes that occur in more recent years. Other functional forms, such as linear functions, may be examined for achieving more robust estimates. The other area for further investigation is the time period for which predictions are produced and optimum sample size required for providing accurate predictions for given durations.

Table 1- List of indicators selected for the simulation and the most acurate estimator identified for each indicator

Indicator	Method
Tuberculosis incidence rate / Per 100,000 population	Comp_WRe
Material Footprint - Biomass / Kg per 1 US dollar (2005 GDP)	Comp_WRe
Neonatal mortality rate / Deaths per 1,000 live births	Comp_WReg
Air transport passengers carried / Millions	Comp_WReg
Minimum organized teacher training, lower secondary education, total / Percentage	Comp_WReg
Prevalence of undernourishment / Percentage	Comp_WReg
Researchers, full-time equivalents / Per million inhabitants	Comp_WReg
Gender parity index for participation rate in organized learning (one year before the official primary entry age) / Female-to-male ratio	Comp_WReg
Share of persons in informal employment in total non-agricultural employment, total / Percentage	Comp_WReg
Social insurance / % of population	Comp_WReg
Manufacturing employment / % of total employment	Comp_WReg
Malaria incidence rate / Per 100,000 population	Comp_WReg
Minimum organized teacher training, upper secondary education, total / Percentage	Comp_Reg

[9] In a few cases, a different variant of an indicator was selected.

Indicator	Method
Population living below the national poverty line / % of population	Comp_Reg
Population living in poverty at $1.90 a day in 2011 PPP / % of population	Comp_Reg
Urban slum population / % of urban population	Comp_Reg
Unemployment rate, total / % of labour force	Comp_Reg
Number of nursing and midwifery personnel / Per 10,000 population	GM
Total primary energy supply (TPES) / Kg of oil equivalent per 1,000 dollars GDP (2011 PPP)	GM
Share of extremely poor living on less than US$1.90 a day in total employment, total / % of total employment	GM
Mortality rate attributed to chronic respiratory disease / Per 100,000 population	GM
Debt service / % of exports of goods, services and income from abroad	GM
Gross domestic expenditure on research and development / % of GDP	GM
Gross domestic expenditure on research and development / Current PPP dollars per capita	GM
Domestic material consumption intensity / Kg per 1 US dollar (2005 GDP)	GM
Total official flows for the agriculture sector / Million 2014 US dollars	GM
Demand for family planning satisfied with modern methods / % of women of reproductive age	WGM
GDP by activity: Manufacturing / % of value added	WGM
Maternal mortality / Deaths per 100,000 live births	WGM
Infant mortality rate / Deaths per 1,000 live births	WGM
Under-five mortality rate / Deaths per 1,000 live births	WGM
Health worker density and distribution / Per 10,000 population	WGM
Access to improved water sources / % of population	WGM
Renewable energy production, total / % of TPES	WGM
New HIV infections (all ages) / Per 100,000 population	WGM
NEET rates, total / % of population aged 15-24	WGM
Weighted tariff-average: Agricultural / % weighted average	WGM
Mortality rate attributed to cardiovascular disease / Per 100,000 population	WGM
Personal remittances received / % of GDP	WGM
Access to improved sanitation / % of population	WGM
Adolescent fertility rate/ Live births per 1,000 women (aged 15-19)	WGM
Material Footprint - Metal Ores / Kg per 1 US dollar (2005 GDP)	WGM
GDP per employed person / Constant 2005 US dollar	WGM
Seats held by women in national parliament / % of seats	WGM
Material Footprint - Non-metal ores/Minerals / Tons per capita	WGM

Indicator	Method
Material Footprint - Fossil Fuels / Tons per capita	WGM
Material Footprint total by type / Tons per capita	WGM
Children under 5 stunting / % of children under 5	WGM
Material Footprint - Fossil Fuels / Kg per 1 US dollar (2005 GDP)	WGM
Exports of merchandise / Million US dollars	Reg
Material Footprint - Non-metal ores/Minerals / Kg per 1 US dollar (2005 GDP)	Reg
Number of automated teller machines (ATMs) / Per 100000 adults	Reg
Fixed-broadband equal to or above 10 Mbit/s subscriptions / Per 100 population	Reg
Mortality rate attributed to diabetes / Per 100,000 population	Reg
Public expenditure on education / % of GDP	Reg
Children under 5 overweight / % of children under 5	Reg
Population covered by a mobile-cellular network / % of population	Reg
Public expenditure on education / % of total government expenditure	Reg
Number of physicians / Per 10,000 population	Reg
Weighted tariff-average: Industrial / % weighted average	Reg
Labour share of GDP / % of GDP	Reg
Participation rate in organized learning (one year before the official primary entry age), total / Percentage	Reg
Internet users / % of population	Wreg
Domestic material consumption per capita / Tons per capita	Wreg
Material Footprint - Metal Ores / Tons per capita	Wreg
Children under 5 wasting / % of children under 5	Wreg
Total international support to infrastructure / Million 2014 US dollars	Wreg
Total revenue / % of GDP	Wreg
General government health expenditure / % of government expenditure	Wreg
Intentional homicide / Per 100,000 population	Wreg
Material Footprint - Biomass / Tons per capita	Wreg
Material Footprint total by type / Kg per 1 US dollar (2005 GDP)	Wreg
Weighted tariff-average: Petroleum / % weighted average	Wreg
Social assistance / % of population	Wreg
Agriculture orientation index / Index	Wreg
Dollar value of financial and technical assistance / Million 2014 US dollars	Wreg
Mortality rate attributed to cancer / Per 100,000 population	Wreg