Data Futures Exchange Indicator Aggregation

Methodology Note

I. INTRODUCTION

To support the identification of patterns, trends and disparities across UNDP's five geographical regions of work, the Data Futures Platform sheds light on regional dynamics by reporting aggregated data values for a growing number of indicators.

By building on the daily refresh rates for 350+ indicators across UNDP's six focus areas[1], the regional and subregional aggregates aim to build a nexus where up-to-date data on poverty and inequality, governance, resilience, environment, energy and gender equality can be explored simultaneously in order to provide a holistic representation of the development trajectories in UNDP's programme countries.

This methodological note presents an overview of the analytical procedures used for the calculation of the regional aggregates presented through the Data Futures Platform. Briefly, this involves:

1) Determining if all required data for the aggregation of an indicator is available.
2) Filling-in missing data at the national level using cubic spline imputation
3) Calculating the aggregate value using population-weighted values for each country with reported or imputed data.
4) For each year, assessing the share of the population in the region or subregion that there is reported data for. Aggregates where the share of the population living in non-reporting countries exceeds the share of the population living in reporting countries are excluded.

Gaps in data availability remain a persistent challenge to the effective use of data for development. Thus, due to the prevalence of both point gaps and countries with missing timeseries, regional and subregional aggregates should be interpreted as approximations of the real but unknown indicator values.

II. METHODS

1. Identifying the Required Information for Aggregating an Indicator

The country-level indicators reported through the Data Futures Platform can be categorized into one of the following categories:

a) *Totals* (e.g. *Total Population*)
b) *Ratios* (*e.g. Gross Domestic Product per Capita*), where a total is scaled by the value of another variable
c) *Indices and Normalised Scores* (e.g. *Rule of Law Index),* where a calculation is performed on one or more indicators to derive a single score that references a given standard

[1] https://strategicplan.undp.org

d) *Ranks* (e.g. *Climate Change Readiness Ranking),* where countries' values for an indicator are ranked in ascending or descending order based on a total, a ratio or an index or normalized score

At launch, the redesigned Data Futures Platform provides aggregations for total- and ratio-type indicators and the development of robust methods for aggregating ranks and indices and normalized scores is ongoing.

While, depending on the frequency of missing data, *totals* can be aggregated readily, *ratio*-type indicators require the availability of data for the scaling factor, that will be needed to calculate the final population-weighted aggregates. For most indicators, the scaling factor is total population, gross domestic product or land area. Scaling factors not currently available through the Data Futures Platform are retrieved, where available.

2. Imputing Missing Data

Missing data was either interpolated or extrapolated using natural cubic spline imputations. Briefly, cubic spline functions are fit in regions with missing data points so as to join the ends between reported timeseries using a smooth curve that is flexible between missing regions but dependent on the closest non-missing data points. Imputed missing data is only used for the calculation of aggregates— they are not used to fill in data gaps in the national data available from other modules of the Data Futures Platform.

- For total-type indicators, cubic spline imputations were performed
- For ratio-type indicators, cubic spline imputations were first performed for the scaling factor indicator, following which the cubic spline imputations were performed for the product of the ratio-type indicator and the scaling factor

    - ***Example:*** For the 'number of physicians per 100,000 population' indicator, cubic spline imputation was performed for the scaling factor (population), following which spline imputations were performed for the expression

$$\frac{(Number\ of\ Physicians\ per\ 100{,}000\ population)\ \times\ (Total\ Population)}{Population\ Scaling\ Factor},$$

Where the population scaling factor represents the factor by which the indicator being aggregated is being scaled.

To restrict the uncertainty associated with data modelled significantly beyond the range of the observed data, for extrapolations, missing data was not imputed beyond the five years preceding the first non-missing data point or for beyond the five years after the last non-missing data point.

3. Calculating Regional Aggregates

a) ***Total-type*** indicators: the regional aggregate for each year was calculated by a single summation for each indicator for all reporting countries:

$$\overline{X_{sy}} = \sum_{i=1}^{n} X_{iy}$$

Where X_{sy} is the regional or subregional aggregate for year *y* and X_{iy} is the value of the indicator for country *i* in year *y*.

b) ***Ratio-type indicators***: the regional aggregate for each year was calculated by summing the values of the rescaled indicator and divided by the totals of the scaling factor for each region:

$$\overline{X_{sy}} = \frac{\sum_{i=1}^{n} r_{iy} * w_{iy}}{\sum_{i=1}^{n} w_{iy}}$$

where r_{iy} represents the unscaled value of the indicator (e.g. GDP per capita) in year *y* and country *i* and w_{iy} represents the value of the weighting factor (national population) in year *y* and country *i*

4. Calculating The Share Of The Population Represented By The Observed Data

For each indicator, the proportion of the population with observed data is calculated for each year. To ensure that the estimated regional aggregates approximate the real but unknown aggregate value, for the cases where the share of the population represented by the imputed data exceeded the share of the population represented by the reported data, regional aggregates were set to missing.