United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
FOR
STATISTICS

TECHNICAL
COOPERATION
GROUP

SUSTAINABLE
DEVELOPMENT
GOALS

7th meeting of the TCG
27-29 October 2020

TCG-7/I/2

UIS METHODOLOGY FOR AGGREGATION OF NATIONAL EDUCATION DATA TO REGIONAL AND GLOBAL ESTIMATES

Table of contents

List of Figures

List of Tables

Acronyms and abbreviations

AD	Auxiliary data or information
ANIR	Adjusted net intake rate
CFVB	Carry first value backward
CI	Coverage index
CLVF	Carry last value forward
E	Enrolment
EFA	Education for All
GED	Global Education Digest
GER	Gross enrolment ratio
GIR	Gross intake ratio
GMI	Group mean imputation
ImpTimeGap	Imputation time gap
IMF	International Monetary Fund
ISCED	International Standard Classification of Education
NE	New entrants
NER	Net enrolment rate
NYIAD	Nearest-year imputation with auxiliary data
PFE	Percentage of female enrolment
Pr	Private
Pub	Public
RI	Related indicator
SAP	School age population
SDG	Sustainable Development Goal
SDG 4	Sustainable Development Goal 4 on education
SIM	Single imputation method
SNYI	Simple nearest-year imputation
T	Teachers
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization
UWGMI	Unweighted group mean imputation
WGMI	Weighted group mean imputation

Introduction

Regional and global indicator estimates produced by the UIS are widely used to monitor progress towards international education goals and as benchmarks for national statistics. A sound and transparent methodology for calculating these statistics over time is a prerequisite for the credibility and the trustworthiness of international statistics and helps users of UIS statistics to adequately interpret and analyse regional and national data.

While formulas for calculation of regional and global statistics are typically straightforward, attention is needed in developing appropriate methodologies for the imputation of missing data. Starting in 2010, the UIS implemented an automated imputation methodology, as part of the revised tools developed by the UIS for indicator calculation.

This document presents the key aspects of the UIS methodology for imputation of missing data and the calculation of regional and global statistics. The document is divided into four sections. Section 1 introduces the terminology used in this document. Section 2 outlines the requirements and constraints for imputation methodologies at the UIS. The imputation process and methods used in the UIS data context to impute missing data are detailed in Section 3 and illustrated with examples. Section 4 describes the calculation of regional and global aggregates, as well as the quality rating approach adopted by the UIS to assess the quality of the aggregated figures.

1. Terminology

To aid the reader, **Table 1** presents the definitions for the terminology used in this document.

Table 1. Terms and definitions

Term	***Definition***	***Level to which the term applies***
Raw data	Data point stored in the UIS database and collected through the UIS survey or from other data sources (e.g. population or economic data). Raw data are almost exclusively absolute data (e.g. enrolment in primary education), typically reported by countries or edited by the UIS. They are inputs for further calculations (e.g. estimation or indicator calculation).	Country
Absolute data	A quantity of a statistical unit expressed in absolute numbers, such as number of students or teachers. The scale of absolute data is usually correlated with the size of a country. Absolute data are mostly raw data, but can be also derived data, calculated using raw data, such as sums of raw data.	Country
Derived data	Absolute data which are not raw data. Derived data are typically the sum of raw data. They are calculated using the same calculation tools as indicators. They are just different because they use an absolute scale. Examples: sum of enrolment at ISCED 1 and 2, school age population, school age enrolment.	Country, region, world

Term	***Definition***	***Level to which the term applies***
Indicator	A function of raw data that produces a statistic that is cross-nationally comparable. Examples: enrolment ratios and rates. In the framework of SDG 4, the data used to produce certain indicators are collected using country-specific definitions, which may limit comparability across countries. Examples: proportion of teachers with the minimum required qualification to teach a given level of education, pupil-qualified teacher ratio.	Country, region, world
Missing data	Country-level data points not available and that need to be imputed in order to produce regional/global estimates.	Country
Auxiliary information	Country level absolute data or indicators used for estimation or imputation of other missing data.	Country
UIS estimate	Publishable[1] estimate to replace missing raw data. UIS estimates are permanently stored as raw data in the database and can be published as such for the country. Indicators and derived data based on UIS estimates are marked as "UIS estimates" when published for a country. However, UIS estimates are exclusively generated for raw data. Indicators that are the result of calculations using UIS estimates are also marked as UIS estimates.	Country, region, world
Substitute	A substitute is an imputation of missing raw data. It is produced when the data reported by the country fail UIS quality standards and are deemed unpublishable by the UIS. Substitutes are produced after background research. Substitutes are stored permanently in the database like UIS estimates, but cannot be published as a value for the given country, meaning they are exclusively to be used as a basis for calculation of regional and global statistics. Substitutes are typically used for countries having a large demographic weight within their region, when an informed estimate leads to more accurate results than automatic imputations.	Country
Valid data	Any data that are either reported by a country, produced as a publishable UIS estimate or a country-level substitute for missing data.	Country
Imputation	Replacement of a missing data point by a value. Imputations are generated at the country level, based on standard algorithms. They are not intended to be published at country level and are exclusively used to calculate regional statistics. Technically, even though UIS estimates and UIS substitutes are special types of imputations, the term imputation in this document refers exclusively to imputations used in the calculation of regional and global estimates	Country

[1] The definition of publishable and unpublishable estimates is presented in section 2.

Term	***Definition***	***Level to which the term applies***
Regional aggregate	A regional estimate that is calculated as an aggregation of country level data.	Region
Regional sum	Refers to absolute data for a region. Typically, a regional aggregate (the sum of absolute data for countries within a region), but can also be a derived regional estimate. Examples: Sum of enrolment at ISCED 1, sum of school age population.	Region
Regional derived estimates	A regional statistic, indicator or absolute number that is not calculated by aggregating country data, but by making calculations using other regional aggregates or statistics.	Region
Regional average	A regional aggregate calculated as average of country indicators. Not all regional indicators are regional averages but they also can be regional derived estimates.	Region

Source: UNESCO Institute for Statistics (UIS)

2. Designing a methodology for UIS regional and global estimates: requirements and constraints

The choice of methodology for imputation of missing data depends on the requirements and constraints of statistics to be produced with the imputed data. The methodology should fit the needs of statistics produced by the UIS. Yet, the limits of the resulting imputed data should be clearly identified and considered when the database is to be used for other analytical purposes. Limiting the imputation system to a pre-set purpose helps to avoid unnecessary complexity (principle of parsimony) and avoids serving imputation issues, not relevant to the system. The purpose of imputation at the UIS is not to produce a more complete database but to prepare the production of statistical output, i.e. regional and global aggregates.

This section starts with an introduction to types of missing data and a discussion of the special nature of missing data in the UIS education database. It further aims to clarify the UIS requirements for regional aggregates. It concludes by discussing imputation techniques and a brief justification for the approach chosen for the production of regional aggregates by the UIS.

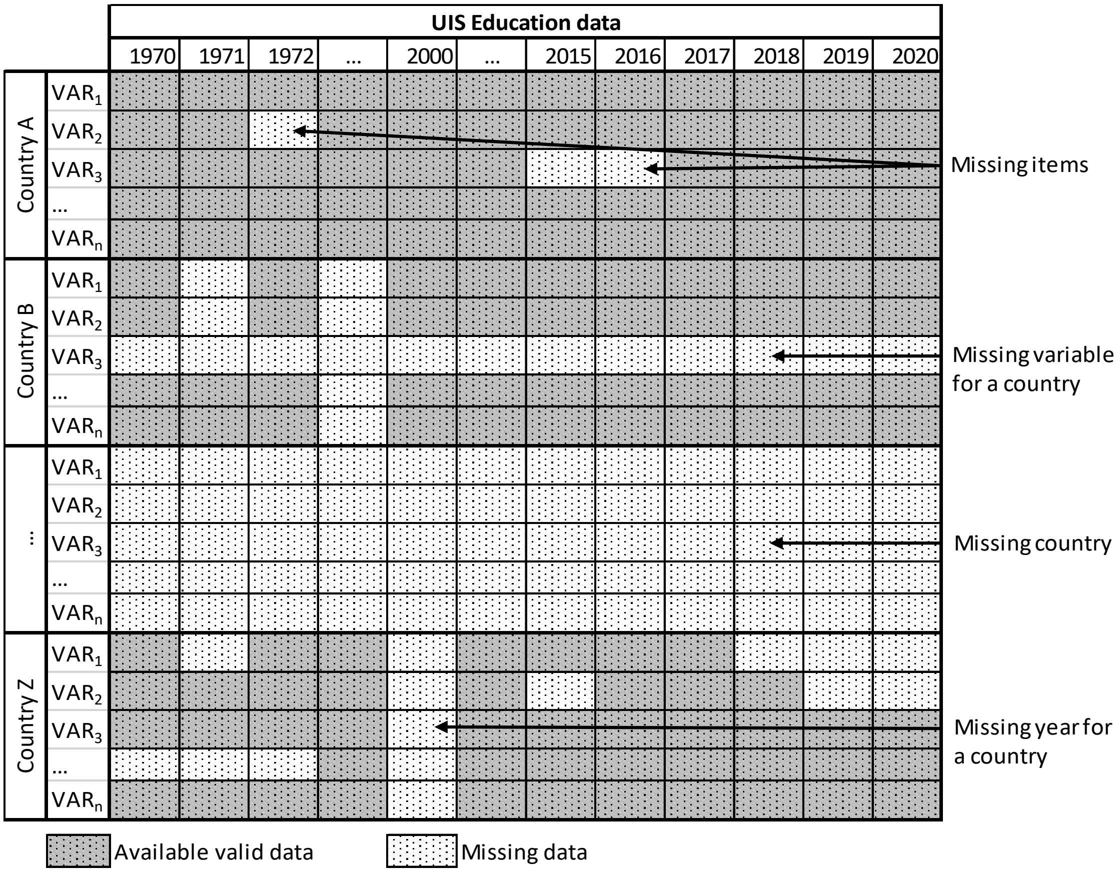
2.1 Types of missing data

UIS data are national aggregates (macro data). They do not contain detailed data for individual observation. As such, the statistical units are countries.

UIS data are missing if they have not been reported by a country or have been reported but are judged by the UIS to be inaccurate. The UIS education database holds cross-sectional multivariate time series: multiple variables/indicators observed for multiple countries over time. In this case the data structure is complex as none of the three dimensions (country, variable, time) are held constant. Consequently, all patterns of missing data need to be considered (see **Figure 1**):

- Missing items: Single data point missing, but the data point exists for other years and other data points exist for the same year;
- Missing years: No data exist for the country in the given year, but data for other years are reported. Missing periods can occur between other years with data, or at the beginning and/or end of the time series.
- Missing variables for individual countries: Data series that are missing for all years, even though other data series are reported.
- Missing country: Countries without any data at all. In the UIS database, which covers a period of more than four decades, this occurs rarely, for example in the case of countries that only recently became UNESCO Member States. It can occur for a subset of data, such as education finance data.

Figure 1. Patterns of missing data in time series data*



* Adapted from Denk and Weber, 2011.

2.2 UIS requirements

2.2.1. Data consistency requirement

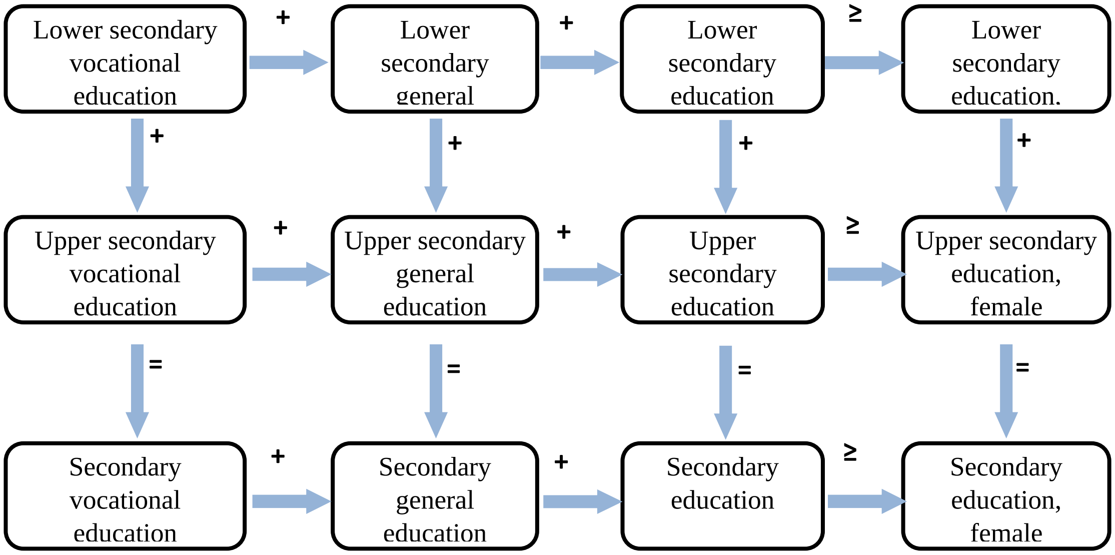
Regional and global statistics produced by the UIS present a system of interlinked indicators and absolute data. Statistics are grouped in families that are related and interdependent. For UIS reporting, it is essential that the data are consistent and free of contradictions. The requirements include:

- Consistency between indicators and absolute data: When regional indicators are reported together with absolute data that are the source of the indicators, figures must be consistent.
- Consistency between related indicators: When indicators are related, their trends should be consistent.
- Consistency between components and their sum: When three or more absolute figures are reported together, and one figure represents the sum of the others, the figures must sum up correctly.
- Natural boundaries for indicators: Indicators such as Net Enrolment Rate (NER) must not exceed 100% with their imputation and the NER must be less than or equal to the Gross Enrolment Ratio (GER).

Understanding the requirement of consistency is necessary to apprehend the complexity of the UIS imputation model. For example, if indicators would be imputed independently from absolute data or other indicators, the use of regression techniques would be a method of choice. Yet, it seems impossible to reach consistency within indicator families with such statistical imputation techniques. Therefore, the UIS system is based on *deterministic calculation methods* producing one fixed number for each imputation based on a system of logical algorithms rather than statistical or stochastic estimates (see section 2.3).

To illustrate the importance of consistency, **Figure 2** shows the logical relationships that must prevail between enrolment data for the secondary level of education.

Figure 2. Example of data dependency: Enrolment in secondary education


Lower secondary vocational education
+
Lower secondary general
+
Lower secondary education
≥
Lower secondary education.
+
+
+
+
Upper secondary vocational education
+
Upper secondary general education
+
Upper secondary education
≥
Upper secondary education, female
=
=
=
=
Secondary vocational education
+
Secondary general education
+
Secondary education
≥
Secondary education, female

Source: UNESCO Institute for Statistics (UIS)

The dependencies reach beyond the data shown in Figure 2 and can be extended to indicators that are based on those absolute data, including GER at lower and upper secondary education, percentage of enrolment by type of programme (vocational or general) at different levels of secondary education, percentage of female enrolment, and parity indices (e.g. gender, location, wealth) for various indicators.

2.2.2. Regional consistency requirement

The UIS serves multiple clients, providing regional statistics in accordance with their regional (and sub-regional) classification system (based on groups of countries), with most clients having a global category ("world").

For reporting, it is essential that absolute data (after imputation of missing data) such as enrolment or the number of teachers in sub-regions sum up to the correct figures for the region. Regional figures must sum up to identical world totals for different systems of regions which comprise in total the same group of countries.

2.2.3. Reporting of absolute data

A key feature of UIS regional statistics is the production of regional and global totals of absolute numbers, such as out-of-school children. Therefore, missing data must not be ignored because doing so would lead to incorrect regional and global figures.

In the case of absolute data, ignoring missing data is equivalent to an implicit imputation of "nil" at the national level, hence, reporting the regional sum of absolute data requires imputation. For indicators (e.g. ratios), ignoring missing data is identical to an implicit imputation of missing national data by the regional average of the given indicator.

2.2.4. Optimal use of available information

Missing data are not only the result of country non-response, but can also be observed when countries respond to the UIS survey but single data points are missing. In the latter case, it is desirable to make best use of the data available for a given year. Unfortunately, there are many potential situations of available and missing data. Therefore, the automated imputation system must provide alternative imputation formulas to cover different scenarios of missing data. A one-fits-all method would simply impute data based on other years and ignore available data from that given year which might help improve the quality of the imputed values. The UIS methodology therefore provides a set of alternative imputation formulas that are designed for different scenarios of missing data. Following a predefined imputation sequence, the first imputation approach is applied, which requires the most complete data. It then escalates to less data demanding formulas until resorting to a formula that does not require any auxiliary information.

2.2.5. Consistency with data updates

UIS imputations methodology strongly draws on auxiliary information, such as data on the school-age population or education data that are correlated with the missing data, such as changes in enrolment figures, which provide useful information for the imputation of teacher numbers. Yet, the basis for imputation may change often: Auxiliary information, especially UN population data, are subject to frequent revisions and updates; countries provide updates or corrections to earlier data submissions.

UIS implicit imputations[2] are therefore regularly, i.e. for every data release, completely revised to reflect changing auxiliary information.

2.2.6. Transparency and efficiency

To be efficient during the regular updates, implicit imputations by the UIS are fully automated. Techniques that require human judgement and calculations (i.e. manual substitution) are not only inefficient but also more prone to error, lack of documentation and deviation from standard UIS procedures. Hence, the automated system produces imputations that are fully documented and can be reproduced since the underlying algorithms are documented. The methodology is standardized in its basic approach in order to ensure consistency and transparency.

The use of deterministic imputation methods makes it possible to explain each single imputed data point and demonstrate its calculation. In order to allow reproduction of imputations, the methods applied must not have any random elements. Techniques using stochastic processes for imputation, would lead to fluctuations between data releases that could not be explained in detail by the UIS. The ability to explain each data point is needed to answer questions from data users concerning changes between data releases.

2.2.7. Time series reporting

UIS education data are central for monitoring progress towards international goals, especially the SDGs. Therefore, good quality of time series data is essential and it is a responsibility of the UIS to be able to explain changes over time. Any imputation technique must therefore ensure consistency within time series and between time series of related data and indicators. Substantial change in a statistic over time (or absence of such change) should never be the result of the imputation methodology but always be explainable by real change (or constancy) in observed data, either in the statistic itself or in auxiliary information used for the imputation process.

Imputations to fill a gap between the last year with data and the present should not build on any assumptions that past trends continue. Progress reported by the UIS should be based on progress actually observed in countries. This conservative approach may under- or overestimate current regional statistics since unobserved change is not considered by imputation techniques. Similarly, when missing data for previous years are imputed on the basis of data for later years, statistics for previous years can potentially be over- or underestimated.

As a consequence, UIS regional figures for a given reference period represent the best information available on the situation in the given year. Yet, change between two reference years represents only changes in those countries of the region for which data updates were provided.

[2] Implicit imputation refers to the imputation that is generated for a given missing data to generate regional and global figures only. Implicit imputation is not published for the particular country. As described by Holt (2003), the implicit imputation may be appropriate because of the risk of publishing imputed value for indicator of a high political importance, but adequate to feed into the aggregation process to generate regional and global estimates.

2.3. Imputation techniques

Numerous imputation techniques are available for survey data, but most of them are not applicable to UIS data when considering the requirements discussed above.

Due to the need to document and explain imputations, the range of techniques the UIS can apply is limited to deterministic techniques, in which case the imputed value is determined by the data available for the country or, in the case of countries without data, for other countries in the region.

The UIS relies on **deterministic imputation techniques**. They are designed to impute one and only one value based on a fixed algorithm (single imputation method) leading to the creation of a complete dataset. The resulting complete dataset is utilized to calculate regional and global averages. Deterministic simple imputation approaches are used due to their straightforward applicability and to the common availability of the auxiliary information. They permit estimation of imputation errors.

The UIS applies deterministic models, which are automated and build on algorithms reflecting the relation within time series and/or between time series of related data or indicators.

Manual deterministic imputation is the simplest imputation approach using SIM. It replaces missing values by values that are specified ad-hoc manually by statistical staff in the data editing process. Each missing value may be treated differently in a manual procedure, or a few rules of thumb may be formulated based on experience and by subject matter experts. In the presence of sufficient auxiliary information and/or additional research, manual deterministic estimates can provide a more realistic picture of the reality and thus be superior to automated imputations. Well researched, they can lead to imputations that can be published at individual country level, while automated imputations should not.

The UIS system applies a system of **automated deterministic imputation**. A set of formulas describes the relations between data series for a given country and allows calculation of missing values based on non-missing data combined with linear interpolation building on time series. In exceptions, model formulas rely as a last resort on mean imputation using data from other countries in the same reference region. The advantage of automated deterministic imputation is that a fixed and transparent set of formulas exists, making imputations transparent. Imputations can be justified by referring to publishable data[3] . In addition, when related data series change, imputations are automatically updated based on the new auxiliary information.

2.4. What kinds of data are imputed?

The UIS publishes regional time series data with many dependencies between them. The gross enrolment ratio (GER), for example, is related to the absolute numbers for enrolment and school age population. The male and female absolute numbers and respective indicators are also interrelated.

[3] By default, imputations are based on data published by the UIS. Yet, in a small number of exceptions, the UIS carries out manual substitutions for missing country data before imputation, for example with data from household surveys as substitutes for missing administrative data. These exceptions are documented in a table in the UIS database (Afghanistan, China, Ethiopia, Haiti, India, Nigeria for very few years) but the data substituted in this manner are not published themselves to avoid the impression that data are available for these countries when they are in fact missing.

Wherever both regional statistics, absolute numbers and their dependent indicators, are published, consistency needs to be ensured between series. To this end, regional indicators are typically not calculated as aggregates from country indicators, but calculated based on regional aggregates of the absolute data underlying the indicator. This means that imputation is needed mostly for country level absolute data (e.g. enrolment) that are aggregated to the regional level. Finally, regional indicators (e.g. GER) are calculated using regional absolute data.

The preference to impute absolute data instead of the indicator is a choice made by the UIS. The inverse approach (impute indicators, calculate weighted regional averages from them and finally calculate regional absolute data out of regional indicators) could also work.

Imputation is carried out for absolute data and not for indicators, as long as absolute data are to be published as regional statistics. For absolute data that are dependent, such as data on male, female and total, only some of the data are imputed. The remaining data are derived from the imputed data. For instance, male enrolment at country level is not imputed and the regional sum of male enrolment is calculated as the difference between the regional sum of total and female enrolment.

Example 1. Calculation of regional gross intake ratio (GIR) into primary education (ISCED 1)

The gross intake ratio into primary education (ISCED 1) GIR_1 is never imputed, because the regional number of new entrants in the first grade $G1$ of primary education NE_1 and the regional value of the official school entrance age population $SAP_{1,G1}$ are calculated and published. To ensure consistent regional results, $GIR_{1,region}$ for a given sex (female, male or both sexes) is not calculated as a weighted mean of country level results but as:

$$GIR_{1,region,sex}^{y} = \frac{NE_{1,region,sex}^{y}}{SAP_{1,G1,region,sex}^{y}} 100$$

where y is the reference year.

Therefore, country level imputation is needed for NE_1 and $SAP_{1,G1}$ only if missing. Additionally, because the total number of new entrants is the sum of male and female figures, only female and total new entrants are imputed, which serve to calculate the regional number of new entrants for female, male and both sexes.

$$NE_{region,total}^{y} = \sum_{c}^{C} NE_{c,total,imputed}^{y} + \sum_{c\prime}^{C\prime} NE_{c',total,not\ imputed}^{y}$$

and

$$NE_{region,female}^{y} = \sum_{c}^{C} NE_{c,female,imputed}^{y} + \sum_{c\prime}^{C\prime} NE_{c',female,not\ imputed}^{y}$$

Where:

- *c* represents a country where *NE* is imputed, and *c′* a country where *NE* is available and no imputation is needed.

- *C* is the number of countries in the region for which *NE* is imputed, and *C′* the number of countries in the region for which *NE* is available and no imputation is needed.

C + *C′* is the total number of countries in the region.

Male regional values are then derived from the regional total and female values:

$$NE^{y}_{region,male} = NE^{y}_{region,total} - NE^{y}_{region,\,female}$$

3. Treatment of missing education data

The UIS addresses the issue of missing data at different phases of its data production and quality assurance process.

Step 1: UIS estimation - Publishable country estimates

Deterministic imputation of missing country level raw data generates UIS estimates that are sufficiently good[4] to be used to produce publishable country statistics. UIS estimates are marked as such when published. The estimates are stored permanently in the UIS data base together with country responses to the UIS survey.

Step 2: Deterministic substitution

For the most populous countries, additional research is carried out to produce imputations to replace missing data. These proxies are not published but they provide better information than automated imputation for countries that influence regional results the most. The imputations are stored permanently in the data base and identified as unpublishable. The imputations are revised each year.

Step 3: Automated deterministic imputation

The final step in the preparation for the calculation of regional statistics is the automated imputation process. It aims to produce a complete data set to be used for the calculation of regional statistics.

The production of the complete data set is automated to avoid the application of ad-hoc methods and to ensure that only one set of transparent standard formulas and rules is applied for all countries and all years. The results of the automated imputation are not stored in the same data base with the publishable data. A copy of imputations is kept for documentation purposes. Automated imputations are recalculated each time regional statistics are prepared.

[4] UIS estimates are based on credible secondary data sources, e.g. the website of a national ministry of education or a national household surveys.

3.1. Publishable UIS estimates

During the initial validation of country level data and the subsequent production of country statistics, the ultimate goal is to produce the maximum amount of publishable raw data. Therefore, the UIS attempts to produce publishable country estimates for raw data as a substitute for missing data. These estimates of good quality are stored permanently in the UIS database and used for calculation of country level indicators and derived data. Subsequently, the UIS estimates are used for the production of regional statistics.

Before the estimation of missing raw data at the UIS, countries are asked to estimate the missing values themselves. This step is crucial because countries may have ad-hoc statistics derived from sample-based surveys or other sources that can be used to reliably estimate missing data. If the country does not provide estimates for missing data, the UIS team responsible for the country searches for information from other national data sources such as the country's own publications or websites.

UIS estimations are exclusively used for countries that provide at least partial data for a given reference year. Any estimate draws on those newly reported data. The UIS does not produce publishable estimates for a given reference year for countries that reported no data at all for that year. UIS estimates are clearly identified by footnotes or symbols in UIS publications and databases. They are validated with the concerned country.

UIS estimates are generated according to the UIS estimation guidelines.[5] Most common are estimates produced with deductive methods, which consist of inferring missing data from values for preceding year (*y-1)* using auxiliary information for the reference year *y*.

For instance, if total enrolment in primary education is missing for a country for the year *y* because enrolment in private institutions is not known, total enrolment is estimated using the reported public enrolment in year *y* and an estimate of the percentage of enrolment in private institutions based on information from previous years.

The estimates for total enrolment are publishable and classified as "**UIS estimates**" if these conditions are fulfilled:

- Primary private education in the preceding year (*y*-1) represents a small percentage (less than 20%) of the entire education system; and
- The enrolment in primary public education has not changed by more than 5% compared to the previous year.

Missing data for private enrolment are not estimated since the resulting indicator, the percentage of private enrolment, would be simply a reproduction of the previous year's indicator. By contrast, the estimate for total enrolment is informed by data for the year *y* and provides therefore new information compared to what was presented the year before, e.g. the resulting gross enrolment ratio (GER) in year *y* reflects a change that is observed in public enrolment.

[5] The UIS guidelines are an internal document that summarizes manual estimation standards.

In exceptional cases[6], estimation is based on information from other sources than the UIS survey. The estimates are validated with the concerned countries. This situation can occur, for instance, when the age distribution of enrolment is unknown from administrative sources but can be estimated from household survey data. The disadvantage of referring to household surveys is that they are not conducted on an annual basis, which means the UIS estimate for the age distribution of pupils can be based on a survey distribution that may be out of date.

At the end of the validation and estimation process, the UIS data base contains only publishable raw data and, in exceptions, unpublishable data that remain in the data base to generate on aggregates that will be publishable. Only publishable raw data are used for subsequent calculation of regional statistics, with the exceptions of the few country substitutes discussed below.

3.2. Unpublishable UIS substitutes

As preparation for the calculation of regional statistics, the UIS conducts further research on countries with a large weight within their region. For these countries, the UIS identifies the need to impute missing data to reflect known trends, which would not be reflected by automated standard imputation ignoring external information. Even though the substituted raw data are considered unpublishable, they are stored permanently in the data base. Production of substitute raw data, typically based on non-standardized methodologies, is an exception at the UIS because it would mean that the UIS publishes regional figures without being transparent about underlying raw data. It is desirable that the UIS makes the underlying research leading to the substitutes available and avoids application of undocumented ad-hoc methods.

As an example, for the UIS September 2019 data release, substitution of missing data by unpublishable raw data occurred for 19 countries from the set of 214 countries and territories. Among the 19 countries, only 10 required substitutions for more than two reference years when looking at a long time series dating back to 1970: Bangladesh, Brazil, China, Egypt, Ethiopia, Haiti, India, Nigeria, the Russian Federation, and Uganda. It is important to underline that only data for key years that are necessary to generate better imputations are substituted. Data for other years are subsequently imputed using the automated system.

3.3. The UIS automated imputation system

After explicit imputation[7] is completed, country level indicators and derived data are produced. However, explicit imputation does not lead to complete data series and some gaps due to missing data remain. Before production of regional statistics another level of imputation (implicit imputation) is needed to produce complete data series, free of missing values, for data series that are required to calculate regional statistics. Missing data points that are not needed for regional and global estimates and for which no publishable estimates can be produced are left missing.

The impact of the quality of imputations on the quality of regional and global estimates is assessed based on UIS standards (see section 4.4).

[6] Exceptions are determined based on the size of the country and the availability of household survey data.

[7] Explicit imputation refers to the publishable imputed value for the particular country. The explicit imputation is based on clearly established method and standards.

This third round of imputation is implicitly and automatically carried out with software developed by the UIS. Imputed national data are neither published nor stored in the UIS data base with the publishable values, but stored separately in an imputation data base with the exclusive purpose of production of UIS regional statistics.

Implicit imputations are replaced by new calculations every time regional statistics are produced. This ensures that imputations are always based on the most updated auxiliary information and weights.

The underpinning imputation formulas are clearly defined and established. No ad-hoc imputation deviating from the standardized formulas is allowed at this stage. Formulas depend on the statistical characteristics of each indicator and its dependencies on other related statistics. The automatic imputation is done in the following steps.

3.3.1. Nearest-year imputation

The UIS imputation algorithm takes into account time series information. Imputation can be performed as linear interpolation of the time series that is to be imputed. This process is referred to as Simple Nearest-Year Imputation (SNYI). Often, nearest-year imputation is carried out by combining an auxiliary time series with further auxiliary information from the same year for which imputation is required. This is referred to as Nearest-Year Imputation with Auxiliary Data (NYIAD).

3.3.1.1. Simple Nearest-Year Imputation (SNYI)

Simple Nearest-Year Imputation of missing data for a given statistics in a given year relies on linear interpolation of available valid data of the closest year(s). Valid data are defined as data reported by a country, publishable UIS estimates or substitutes. Assume year y of the statistics X is missing and needs to be imputed. Depending on the available valid data, there are three scenarios for imputation: valid data exist (i) only for years before year y, (ii) only for years after y or (iii) for years before and after year y.

For all nearest-year imputations, the UIS algorithm determines an imputation value and, for quality assessment, an imputation distance in years.

a. Data exist only before year y

If valid values of the statistics X are only available for years earlier than the reference year y, the "carry last value forward" (CLVF) strategy is applied: the value var_b for year b, the most recent year before year y, is used as the imputed value for the reference year: $var_y = var_b$. In this case, the imputation distance (in years) is $d_b=y-b$.

The "carry last value forward" strategy may lead to over-estimation but most likely under-estimation of the true value of X in year y, especially in the case of expanding education systems.

b. Data exist only after year y

If valid values of the statistic X are only available for years after the reference year y, the "carry first value backward" (CFVB) strategy is applied: the value for the first year after the reference year y, year a, is used as the imputed value for the reference year: $var_y = var_a$. In this case, the imputation distance (in years) is $d_a=a-y$.

The “carry first value backward” strategy may lead to under-estimation, but most likely over-estimation of the true value of X in year y, especially in the case of expanding education systems.

c. Data exist for years before and after year *y*

If valid values of the statistic *X* are available for years before and after the reference year *y*, data are imputed assuming a linear trend between the closest valid data in the years before and after the reference year, year *b* (before) and year *a* (after).

The missing value *var^{y}* for the year *y* is calculated based on the valid data points *var$_{b}$* and *var$_{a}$* as

$$\mathrm{X_y} = \mathrm{X_b} + \left[\left(\frac{\mathrm{X_a} - \mathrm{X_b}}{a-b}\right)(y-b)\right]$$

with $\left(\frac{\mathrm{X_a} - \mathrm{X_b}}{a-b}\right)$ expressing the average annual change in the period between years *b* and *a* and *(y-b)* being the number of years between the years *b* and *y.*

In the case of imputation by linear interpolation, the imputation distance (in years) is calculated as the minimum of the differences between *y* and *a* and *y* and *b,* respectively. To account for the fact that linear interpolation draws on more information than a situation with only data before or after the year with missing data, the imputation distance is reduced by 1:

$$d_m = \min\big((a-y),(y-b)\big) - 1$$

For example, if data are imputed for the year 2016 and valid data exist for the years 2013 and 2018, the imputation distance is

$$d_m = \min\big((2018-2016),(2016-2013)\big) - 1 = (2018-2016) - 1 = 1\text{ year}$$

The actual imputation distance (2 years) is reduced by 1 to indicate the higher quality of the imputed value compared to cases where valid data are only available for a year before or after the year of imputation *y*.

An advantage of this imputation method is that the imputed value always falls in the range of two observed values. However, this simple computational approach assumes that education indicators evolve smoothly over time, which may not always be correct.

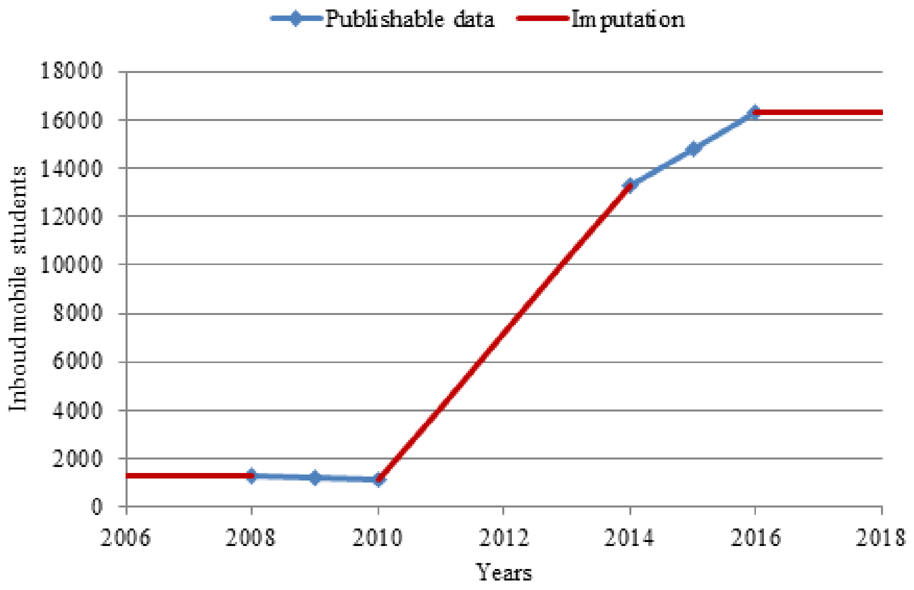
If for a given country data are missing for all years, no data can be imputed with SNYI and a different imputation method is required.

Example 2. Simple Nearest-Year Imputation

SNYI is used for key data series that cannot be informed by other trends. Examples are the total population and the school-age population for small island countries that have gaps in time series. Another example is the number of inbound mobile students.

The chart below shows SNYI for the number of inbound mobile students in tertiary education with missing data imputed before, between and after valid data.

Figure 3. Inbound mobile students of a given country for the period 2006 to 2018



As shown in the chart, missing data for 2011, 2012 and 2013 are imputed assuming a linear trend between the closest years with valid data (2010 and 2014). In the case of missing data for 2017 and 2018, publishable data for 2016 are used (CLVF strategy). Similarly, the number of inbound mobile students in 2008 is used to impute missing data in the years 2007 and before (CFVB strategy).

3.3.1.2. Nearest-Year Imputation with Auxiliary Data (NYIAD)

The most common algorithm in the UIS imputation methodology allows imputing a missing value of the statistic *X* for the year *y* based on the time series of a related indicator *RI* and auxiliary information *AD,* which describes a relation between *X*, which is missing for some years, and *RI,* which exists for those years*.* This imputation technique, based on an imputed time series of a *RI* and an additional calculation for the reference year using *AD*, is identified as Nearest-Year Imputation with Auxiliary Data (NYIAD).

A standard situation is the use of NYIAD to impute missing absolute data based on a time series of related indicators. This is done because imputation based on SNYI for absolute data is less satisfactory than SNYI for indicators. Absolute data, such as enrolments, are therefore not directly imputed using SNYI only but as a combination of SNYI for the related indicator *RI* and auxiliary information for the reference year *y*.

Similar to SNYI, NYIAD returns an imputed value and, for quality assessment, an imputation distance. The imputation distance is calculated in the same manner as for SNYI, but with the nearest years before and/or after the reference year with valid data for the related indicator *RI,* for which the nearest year imputation is applied.

When a country never reported data for the indicator *RI,* SNYI for *RI* returns missing data and as a consequence NYIAD fails. For almost all imputation formulas, NYIAD is therefore complemented by unweighted group mean substitution to produce estimates of *RI* for non-reporting countries (see section 3.3.2).

Example 3. Nearest-Year Imputation with Auxiliary Data: Imputation of total primary enrolment

Change in enrolment in primary education E_1^y is interrelated with the change in the gross enrolment ratio for primary education GER_1^y and on the change in population size. In this example, the GER_1^y is the related indicator *RI*, and the school-age population for primary education SAP_1^y is the auxillary information *AD.*

If for a given reference year *y* total enrolment in primary education E_1^y is missing and needs to be imputed, the corresponding GER_1^y is naturally also missing. GER_1^y can be imputed using the SNYI algorithm. E_1^y is calculated as the product of the imputed GER_1^y and the primary school age population (SAP_1) for the year *y*.

$$E_1^y = \frac{SNYI(GER_1)}{100} SAP_{1,imputed}^y$$

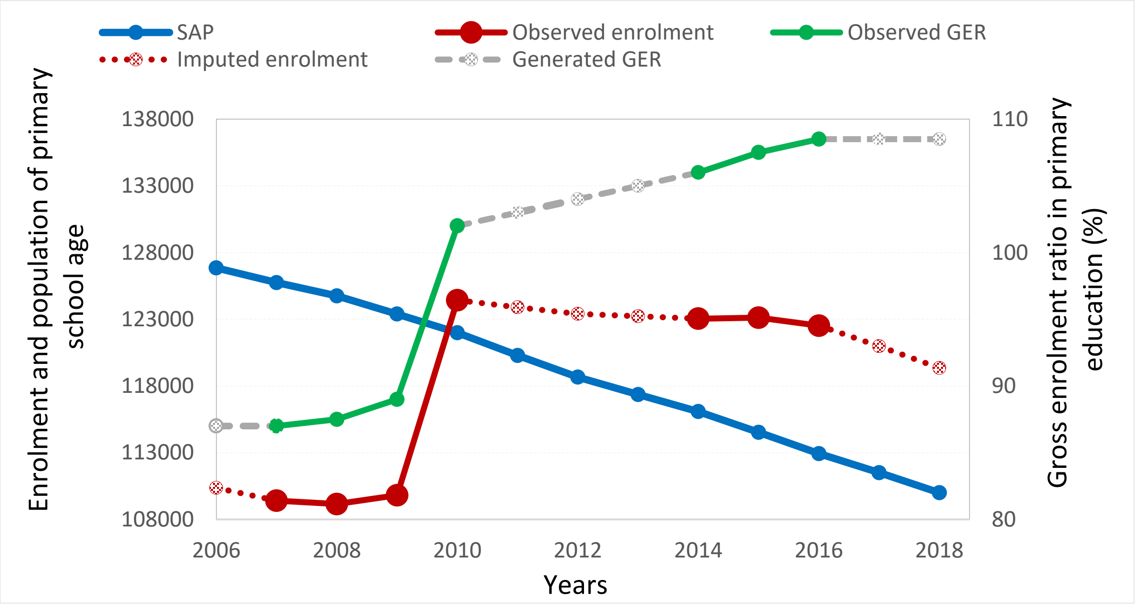
First, SNYI imputes a value for GER_1 in the year *y* based on valid GER_1 data of the nearest years without storing it in the UIS database (GER_1 is temporarily generated to impute enrolment). In a second step, the imputed GER in percent is then divided by 100 and multiplied by SAP_1^y (which is observed or imputed before applying the imputation algorithm for enrolment in primary education). If there is no value for the related indicator GER_1 for all years, SNYI returns missing data and thereby NYIAD fails to return a value for E_1^y. At this point, unweighted group mean imputation becomes necessary (see section 3.3.2).

Table 2 and the **Figure 4** illustrate the imputation of enrolement in primary education for a hypothetical country during the period 2006 to 2018. Observed data of GER_1 for 2007, 2008, 2009, 2010 and 2014 show an increase with a drastic change in 2010. The school-age population is decreasing over time and enrolment in 2014 is lower than in 2010, despite the GER being higher. After imputation with SNYI, the GER_1 has a smooth trend between 2010 and 2014 and no change in 2006 and after 2016. Yet, because the imputation of enrolment uses information from the series with school-age population data, the trend for E_1 after 2016 is not flat but represents the expected decrease in enrolment following the known decrease in school-age population since the imputated GER_1 remains constant after 2016.

Table 1. Imputed values of enrolment in primary education for a hypothetical country

Year	School age population *SAP_1^y*	Gross enrolment ratio *GER_1^y*			Enrolment *E_1^y*	
		Value	Data status	Nearest year(s) with publishable data	Value	Data status
2006	126,850	87.0	SNYI	2007	110,360	Imputation
2007	125,750	87.0	Publishable	-	109,403	Publishable
2008	124,750	87.5	Publishable	-	109,156	Publishable
2009	123,390	89.0	Publishable	-	109,817	Publishable
2010	121,990	102.0	Publishable	-	124,430	Publishable
2011	120,290	103.0	SNYI	2010, 2014	123,899	Imputation
2012	118,660	104.0	SNYI	2011, 2014	123,406	Imputation
2013	117,360	105.0	SNYI	2012, 2014	123,228	Imputation
2014	116,080	106.0	Publishable	-	123,045	Publishable
2015	114,530	107.5	Publishable	-	123,120	Publishable
2016	112,930	108.5	Publishable	-	122,529	Publishable
2017	111,500	108.5	SNYI	2016	120,978	Imputation
2018	110,000	108.5	SNYI	2016	119,350	Imputation

Figure 4. Trend of imputed values of enrolment in primary education for a hypothetical country (GER1 is temporarily generated for the purpose of enrolment imputation)



3.3.2. Unweighted group mean imputation

Where no information is available for a given country, imputation draws on data from other countries. Since the population can substantially differ between countries in the same region and in order to reduce the influence of countries with a large population, countries are weighted equally regardless of their population. In this case, the unweighted group mean of a given indicator is used as the imputed value. This is referred to as Unweighted Group Mean Imputation (UWGMI), typically carried out in combination with the NYIAD algorithm. For illustration, consider a statistic *X* with a missing value for the year *y* that should be imputed based on the time series of a related indicator *RI* and auxiliary information *AD.* If the time series of *RI* are missing, *RI* cannot be determined for the given country, not even by SNYI, and is therefore replaced by the unweighted group mean. The unweighted group mean is then combined with the auxiliary information *AD* based on a given algorithm to generate an imputed value of *X*.

UWGMI can only be used to replace missing data which are not absolute. In imputation of absolute data, UWGMI can only be used to impute a ratio or indicator, which serves as auxiliary information to impute absolute data, such as enrolment, using additional information representing the absolute size of the country.

UWGMI considers each country in a region as one entity regardless of its population and treats all countries equally. It implicitly assumes that the education indicator in a given country is uniformly influenced by the values of the indicator in all other countries in the region regardless of the population. For countries with substantial relative weights in their region (for example, China in East Asia and the Pacific), the UIS avoids the application of UWGMI. In such cases, manual imputation of missing values is required, even if it generates unpublishable data (see section 2.2).

Because the imputation method is sensitive to regional groupings, the UIS imputes missing values for a given country only once a year, based on the principal UIS regional classification, (see Annex). For instance, data for a Caribbean country are imputed based on the group mean of the Caribbean region. Since the group mean of the Caribbean is different from the mean for Latin America and the Caribbean or for the world, repeating UWGMI for each group would lead to different imputations for the same country. This is avoided by reusing the values calculated with UWGMI for smaller regions within a larger region. For example, when calculating regional statistics for Latin America and the Caribbean or for the world, UWGMI is not repeated, but the result from the UWGMI for the Caribbean region is reused.

An alternative approach would be the Weighted Group Mean Imputation (WGMI). WGMI implicitly assumes that education indicators in a given country are similar to the values for the population of the region as a whole. However, this approach would bias imputations for countries of all sizes towards data from big countries (i.e. countries that have a large population (weight) have a greater influence on the results).

In the case of unweighted group mean imputation, the determination of the imputation distance (in years) is difficult and can be arbitrary because it depends on the quality of the imputation values, which in turn depends on how close the missing information for a given country is to the unweighted regional averages.

UWGMI uses external information to impute missing data for a country, which might not reflect the real situation in the country. However, the UWGMI best reflects the situation in the region, given no information is available on the missing country.

Since the UWGMI is not based on data from another year, no imputation distance is defined and available for quality assessment. To allow for quality assessment of the regional figures that the UWGMI contributes to, the imputation distance needs to be substituted by fixed value that reflects the low quality of the imputation without overstating the impact of a single UWGMI. The imputation distance for UWGMI is defined as 10 years.

Example 4. Unweighted Group Mean Imputation

To illustrate UWGMI, consider a hypothetical region established based on the principal UIS regional classification. The region contains four countries, A, B, C and D. The net enrolment rate in primary education NER_1 for country B is missing and no time series information is available to use SNYI. In addition, there is no available auxiliary information (such as GER_1) of country B in order to explicitly impute NER_1.

Table 2. Hypothetic net enrolment rate in primary education and the corresponding school age population for given countries

Country	Net enrolment rate in primary education (NER_1, %)	School age population
A	95	1,000
B	missing	2,000
C	89	3,000
D	98	2,000

In this case, the unweighted group mean technique is applied to impute the net enrolment rate in primary education ($NER_{1,UWGMI}$). UWGMI for NER_1 for the given region yields:

$$NER_{1,UWGMI} = \frac{95 + 89 + 98}{3} = 94\%$$

This imputed value of NER_1 for country B is used to calculate all regional and global estimates of this indicator for all regional groupings containing country B.

As illustrated, UWGMI is a simple method but allows reducing the bias that can occur when data are missing for some countries with a reasonably small weight within a region using other methods such as ignoring countries with missing values. For countries with dominating weight in a given region, the UIS substitutes data based on research. Producing regional averages without imputation of missing data is equivalent to implicit imputation of Weighted Group Mean Imputation (WGMI) to missing values.

3.3.3. Sequential imputation

In numerous cases, automated imputation requires that some auxiliary data are imputed in a previous step. In other words, most UIS imputations are a sequence of several imputation steps and could be described as a chain of applications of SNYI, NYIAD and UWGMI to yield imputed values. The automated system must therefore ensure that a pre-determined execution sequence is respected to ensure that imputations needed in subsequent imputations are conducted first.

Example 5. Sequential imputation

If, for a given country, female enrolment in secondary education (ISCED 2 and 3) $E_{23,y,female}$ is missing for year *y*, the imputation of $E_{23,y,female}$ needs to draw on the enrolment in secondary education (total enrolment for both sexes) $E_{23,y,imputed}$. It is worth noting that in formulas data with the subscript *imputed* never contain missing values, since the symbol refers to data that were imputed in a separate preceding step. *In this example,* $E_{23,y,imputed}$ is the imputed version of *E23,*.

To impute female enrolment the system tries first to use $E_{23,y,imputed}$ in combination with SNYI of the percentage of female enrolment in secondary education PFE_{23}:

$$E^{y}_{23,female,imputed} = \frac{SNYI(PFE_{23})}{100} E^{y}_{23,imputed}$$

If SNYI fails to generate a value the unweighted group mean imputation of PFE_{23} is used.

$$E^{y}_{23,female,imputed} = \frac{UWGMI(PFE_{23})}{100} E^{y}_{23,imputed}$$

The imputation of $E_{23,y}$ is defined in a separate set of formulas to be applied in a preceding step. The imputation of $E_{23,y}$ itself uses SNYI imputation of the gross enrolment ratio of secondary education GER_{23}, which can be imputed by UWGMI if SNYI fails to return a value. In turn, the imputation of $E_{23,y}$ requires the school-age population in secondary education SAP_{23}, which in exceptions, need to be also imputed in a preceding step using SNYI. The imputation algorithm formula for $E_{23,y}$ to create $E_{23,yimputed}$ is:

$$E^{y}_{23,imputed} = \frac{SNYI(GER_{23})}{100} SAP^{y}_{23,imputed}$$

If the time series for GER_{23} is completely missing, $SNYI(GER_{23})$ returns a missing value. The sequential imputation escalates in this case to use UWGMI for the year *y* to generate an imputation:

$$E^{y}_{23,imputed} = \frac{UWGMI(GER_{23})}{100} SAP^{y}_{23,imputed}$$

3.3.4. Multi-option imputation model

As discussed in section 1.1, missing data in time series follow different patterns in different countries. The UIS automated system therefore provides a pre-determined sequence of formulas for deterministic imputation. The first alternative (the first imputation formula in the pre-determined imputation sequence) requires more data and is perceived as better, but fails in a situation when the pattern of missing data is less favourable, meaning auxiliary information are also missing. The system cascades to the following alternatives that are considered less favourable. The first formula in the pre-determined sequence that generates a value is used.

To ensure completeness of data, the UIS imputation model must be designed in such a way that the last option, requiring the least amount of data, produces an imputation result in all cases, without exception.

This means that the last option draws on unweighted group means and auxiliary information that must be available, such as previously imputed data.

For a given statistic, this process is carried out for all years (year by year), which generates time series with the best available imputation for each year.

Example 6. Multi-option imputation model

Assume statistics on teachers in private institutions (*T_{pr}*) are missing and should be imputed using auxiliary information on public and private enrolment *(E_{pub}* and *E_{pr})* and public teachers (*T_{pub}*). Assume further the pattern of missing data in different countries indicated in **Figure 5**.

Figure 5. Hypothetical pattern of enrolments and teachers in public and private institutions of four countries

		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Country A	E_{pub}	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
	E_{pr}	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
	Tpub	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
	Tpr	▨	▨	▨	▨	▨	▨	▨	░	▨	▨	▨	▨	░
Country B	E_{pub}	▨	▨	▨	▨	▨	▨	▨	░	▨	▨	▨	▨	▨
	E_{pr}	▨	▨	▨	▨	▨	▨	▨	░	▨	░	▨	▨	▨
	Tpub	▨	▨	▨	▨	▨	▨	▨	░	▨	░	▨	░	▨
	Tpr	▨	▨	▨	▨	▨	▨	░	░	░	░	░	░	▨
Country C	E_{pub}	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
	E_{pr}	▨	▨	▨	▨	▨	▨	▨	░	░	░	▨	▨	▨
	Tpub	▨	▨	▨	▨	▨	▨	▨	▨	░	░	▨	▨	▨
	Tpr	░	░	░	░	░	░	░	░	░	░	░	░	░
Country D	E_{pub}	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
	E_{pr}	░	░	░	░	░	░	░	░	░	░	░	░	░
	Tpub	░	░	░	░	░	░	░	░	░	░	░	░	░
	Tpr	░	░	░	░	░	░	░	░	░	░	░	░	░

▨ Available valid data ░ Missing data

Different options are applied to make the best use of *E_{pub} and E_{pr}* and *T_{pub}* in the given situations, using information on *E_{pub} and E_{pr}* to reflect the change in the size of the system and *T_{pub}* and *T_{pr}* to reflect trends in pupil teacher ratios. It is important to note that the process starts by looking for publishable data first before applying the imputation algorithm.

No imputation: *T_{pr}* is available.

In this trivial case, the publishable value is used and thereby none of the imputation options need to be applied. This situation applies to countries A and B for years with publishable values (observed or UIS estimates).

Imputation option 1: *T_{pr}* is missing for the given year and available for other years and *E_{pr}* is available for the given year.

Option 1 applies to country A, years 2013 and 2018, and country B, years 2012, 2014, 2016 and 2017. In this case, the private pupil-teacher ratio $PTR_{pr} = E_{pr}/T_{pr}$ can be imputed using SNYI because the ratio E_{pr}/T_{pr} has valid results in other years. The missing T_{pr} is imputed as:

$$T_{pr} = \frac{E_{pr}}{SNYI(PTR_{pr})} = \frac{E_{pr}}{SNYI(E_{pr}/T_{pr})}$$

This option assumes that the number of teachers in the private system is changing at the same rate as enrolment.

Imputation option 2: T_{pr} and E_{pr} are missing and E_{pub} is available.

For country B, year 2015, the private pupil-teacher ratio $PTR_{pr}=E_{pr}/T_{pr}$ can be imputed with SNYI drawing on publishable data in 2011 and 2018. E_{pr} can be imputed using SNYI *(E_{pr}/E_{pub}),* the ratio of private to public enrolment in 2014 and 2016. Based on these assumptions, T_{pr} is imputed as:

$$T_{pr} = \frac{E_{pub} * SNYI(E_{pr}/T_{pr})}{SNYI(PTR_{pr})}$$

This option reflects changes in the size of total enrolment, assuming private enrolment changes at the same rate as public enrolment.

Imputation option 3: T_{pr} is missing for all years and E_{pub}, T_{pub} and E_{pr} are available.

This option applies to country C, years 2006 to 2012 and 2016 to 2018. In this case, the private pupil-teacher ratio ($PTR_{pr}=E_{pr}/T_{pr}$) cannot be imputed using SNYI. T_{pr} can be imputed assuming that the pupil-teacher ratio in private education (PTR_{pr}) is identical to the pupil-teacher ratio in public education:

$$T_{pr} = \frac{E_{pr}}{E_{pub}} T_{pub}$$

Further options build on option 3 but use SNYI for different combinations of E_{pub}, T_{pub} and E_{pr} being missing.

When T_{pr} and T_{pub} are missing for all years, no PTR can be calculated for public or private education. In this case, the unweighted group mean of the PTR in private education is used to impute T_{pr} based on E_{pr}. This applies to country D for all years.

Final imputation option:

Finally, in the worst possible case, E_{pub} is also missing. As a last resort, a final option must be designed to provide an imputation purely based on several unweighted group mean imputations to estimate $T_{pr,}$ E_{pr} and E_{pub} and population data.

Sometimes alternative options are implicit and hidden by the fact that the applied algorithm uses values that are previously imputed for which SNYI or NYIAD were already performed separately. In other words, the complexity of the imputation model is not shown in the formulas but hidden in the execution sequence of imputations, which the system automatically determines and optimizes.

3.3.5. Imputation based calculations

Finally, missing country level data needed for the complete data set are not imputed but calculated based on imputed values of variables from which they can be derived, even though they could be imputed. This is the case for raw data that should sum up. For example, if total enrolment in secondary education, enrolment in secondary general education and enrolment in secondary vocational education would be imputed separately, this could cause inconsistencies in the sums. Therefore, imputation is only done for two of the three data points. The third data point is calculated on the basis of the other imputations to ensure numerical consistency between imputed values.

As a rule of thumb, the data point representing typically the smallest fraction of the total sum is selected to be calculated based on the other two imputations. The previous imputations of the other data points must ensure that the imputations are consistent, so that calculations cannot turn out negative results.

Example 7. Imputation of teachers in upper secondary education

Assume the total number of teachers in upper secondary education (ISCED 3) is missing. Because the number of teachers in lower secondary education (ISCED 2) and the number of teachers in upper secondary education must sum up to total teachers in secondary education (ISCED 2 and 3), the imputation algorithm must insure consistency between all numbers.

In this case, missing data on teachers in upper secondary education T_3 are imputed using the previously imputed number of teachers in total secondary education $T_{23,imputed}$ and the previously imputed number of teachers in lower secondary education $T_{2,imputed}$. No other algorithm is applied. Since $T_{2,imputed}$ and $T_{23,imputed}$ contain by definition no missing values, the calculation always leads to non-missing results.

$$T_3 = T_{23,imputed} - T_{2,imputed}$$

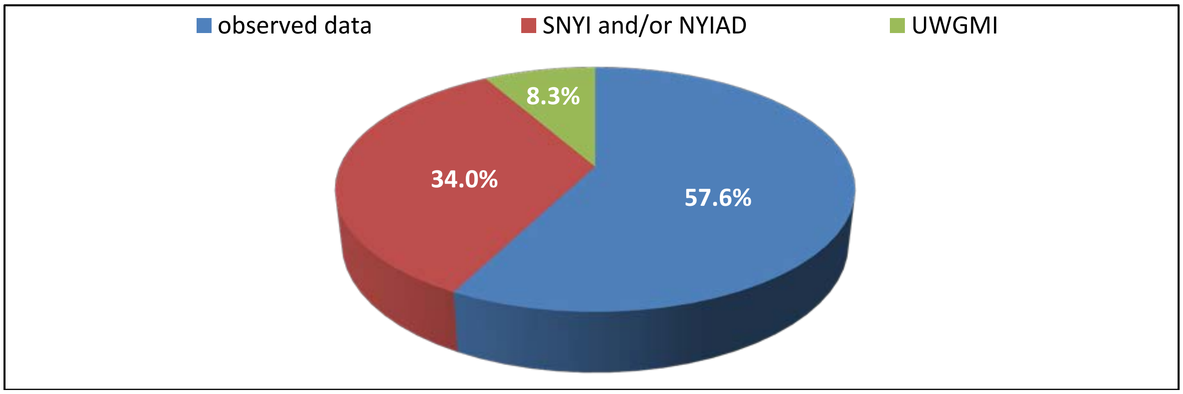
3.3.6. Judging the impact of imputation: The proportion of missing data

One aspect to describe the quality of regional aggregates is to calculate the proportion of missing data at the country level. The complete UIS dataset utilized to calculate regional or global education figures in a given year is, as of 2019, comprised of 631 raw data points and indicators (including disaggregation by gender, ISCED level, type of programmes, etc.) for 214 countries and territories.

Figure 6 illustrates that in September 2019, 34 per cent of the complete UIS dataset used to calculate regional averages for the period 2000 to 2018 were imputed directly or indirectly with SNYI and/or NYIAD and 8.3 per cent were imputed with UWGMI.

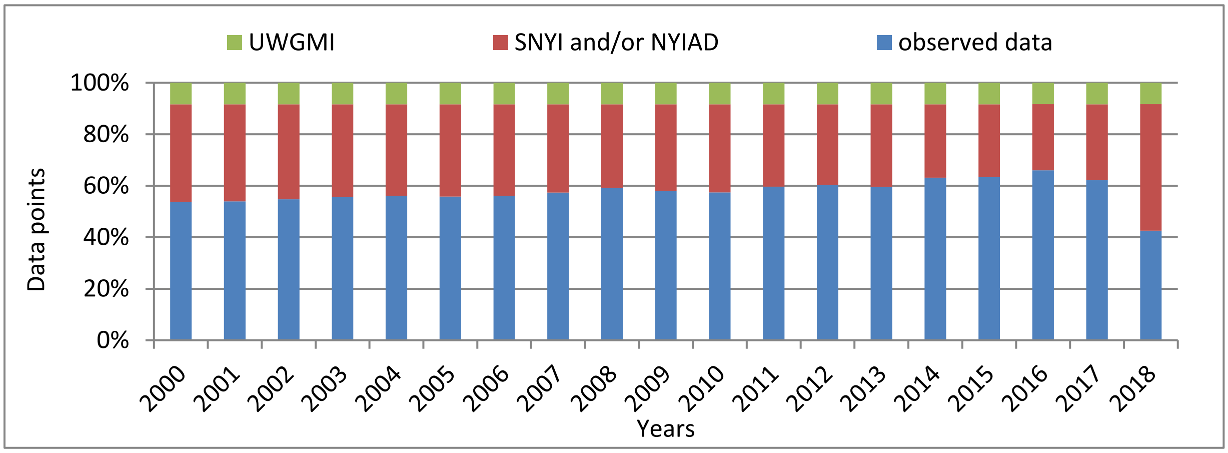
When we look at the composition of the complete data set by single year, the situation is improving over time with a lower need for imputations, except for 2018 (the most recent year with regional averages in the education data release by the UIS in September 2019), as shown in **Figure 7.** There is a high proportion of imputations in the most recent year (2018) contrary to the previous years. In fact, 57.6% of the required data to calculate regional and global figures for 2018 were imputed. This situation for the most recent year in the UIS database is due to the fact that some countries have not yet reported their data. The situation improves each time when the following UIS data collection round is completed and countries reporting always one year late are also added to the database.

Figure 6. Distribution of regional aggregates by type of imputation method for 2000-2017



Source: UIS database, September 2019.

Figure 7. Distribution of regional aggregates by type of imputation method and by year

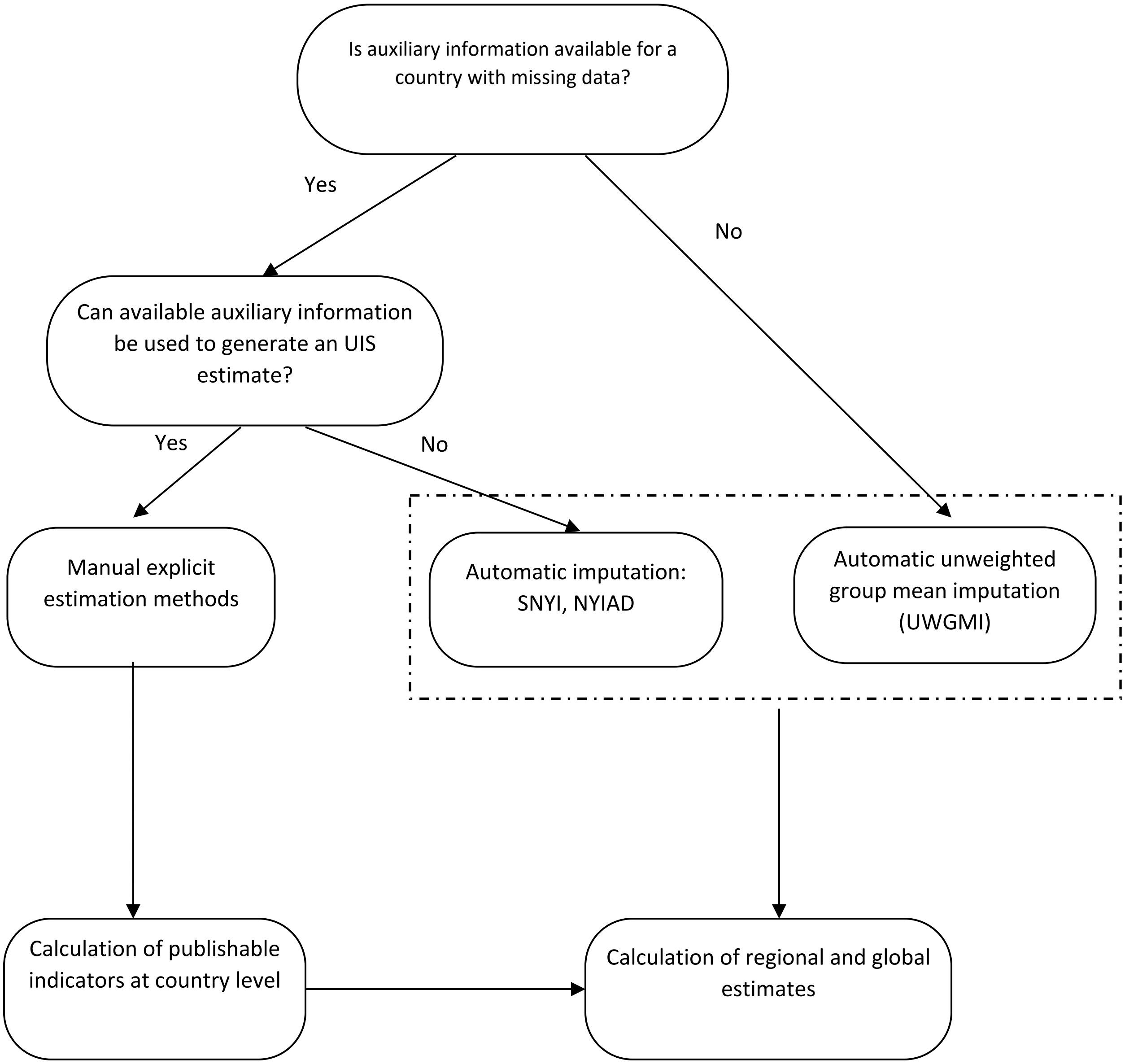


Source: UIS database, September 2019.

3.3.7. Summary: Steps of handling of missing data

Figure 8 summarizes the UIS methodology to deal with missing data and to generate UIS publishable data at country level, as well as the imputation methods for calculation of regional figures.

Figure 8. Summary of UIS methodology for imputation of missing data


Is auxiliary information available for a country with missing data?
Yes
No
Can available auxiliary information be used to generate an UIS estimate?
Yes
No
Manual explicit estimation methods
Automatic imputation: SNYI, NYIAD
Automatic unweighted group mean imputation (UWGMI)
Calculation of publishable indicators at country level
Calculation of regional and global estimates

Source: UNESCO Institute for Statistics (UIS)

4. Regional and global estimates

The previous sections introduced the methodology for data imputation at the country level. The sole purpose of the imputation process is to prepare a complete database, meaning without any missing data points, for the computation of regional aggregates. The following section now describes how country level data are aggregated to the regional level.

Section 4 describes the aggregation of absolute figures to regional sums. It presents the aggregation of ratios and percentages into regional estimates and the calculation of derived regional statistics. Finally, it discusses the UIS approach to determine the quality of the regional estimate with respect to the fraction of missing data and thus imputation that contributed to the aggregation. It concludes by describing the rules used to decide which regional figures cannot be published due to too high fractions of missing information.

Methodologies to produce regional averages for dependent indicators for which the variability of one affects the other (e.g. Gross graduation ratio and Net enrolment rate) are chosen to insure their consistency. However, because of the difference between independent indicators[8], their imputation methods and regional or global estimates are based on methodologies that are not necessarily similar.

4.1. Regional and global sums

The regional sum of a given variable *X*, such as enrolment, is computed as follows:

$$X_{region_i}^{y} = \sum_{country} X_{country}^{y}$$

Where:

$X_{region_i}^{y}$ is the regional sum of the variable *X* for all countries in region *i* in year *y*.

$X_{country}^{y}$ is the value of the variable *X* for a given country in region *i* in year *y*.

[8] In this context, indicators are independent if the change of one does not affect the other indicator.

Example 8. Calculation of regional sums

Consider a hypothetical region that contains 7 countries: A, B, C, D, E, F and G. Total enrolment in primary education (ISCED 1) E_1 in 2016 is missing for countries C, E and F (**Figure 9**).

Figure 9. Enrolment and school-age population of primary education for countries C, E and F

		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Country C	E_1	▨	▨	▨	▨	▨	▨	▨	▨	▨	░	░	▨	▨
	SAP_1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Country E	E_1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	░	░	░
	SAP_1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Country F	E_1	░	░	░	░	░	░	░	░	░	░	░	░	░
	SAP_1	▨	▨	▨	░	░	▨	▨	▨	▨	░	░	▨	▨

▨ Available valid data

░ Missing data

To calculate the regional sum of total enrolment in primary education in the reference year 2016, all missing country values must be imputed. This is basically discussed above, but it is shown here again to demonstrate the complete process of regional aggregation. The deterministic imputation method for E_1 is based on the following algorithm, where the sequence is important:

No imputation: Publishable values for a given country are available.

Imputation option 1: Total enrolment E_1 and the school age population SAP_1 are available for other years. The school-age population should be separately and previously imputed for the reference year (i.e. the subscript "imputed" in $SAP_{1,imputed}$ means that the value of SAP_1 exists or has already been imputed).

$$E_1 \ = SNYI\left(\frac{E_1}{SAP_1}\right) * SAP_{1,imputed}$$

Imputation option 2: Total enrolment E_1 is missing for all years. In this case, the unweighted group mean imputation of the ratio E_1/SAP_1 is multiplied with the imputed value of SAP$_1$ for the reference year.

$$E_1 \ = UWGMI\left(\frac{E_1}{SAP_1}\right) * SAP_{1,imputed}$$

Depending on the available auxiliary information, automatic imputation generates estimates based on different imputation options. Because publishable data are available for countries A, B, D and G, imputation is not applied. In the case of country C, option 1 is applied. The ratio E_1/SAP_1 can be imputed using SNYI, which means that the value of this ratio in 2016 is imputed assuming a linear trend between 2014 and 2017.

Because publishable data of the ratio data E_1/SAP_1 for country E exist only for years before 2016, the value of this indicator in 2015 (the most recent year with publishable data) is carried forward and used as the imputed value.

For country F, SNYI fails to return a value for the ratio E_1/SAP_1 because there are no data for E_1 in any year. In this case, UWGMI is applied to generate a value for the ratio, which in turn is used with the pre-imputed value of SAP_1 to impute E_1[9]

Table 3. Imputation of enrolment in primary education for a hypothetic region.

Country	Enrolment in ISCED 1 (before automatic imputation)	Enrolment in ISCED 1 (after automatic imputation)	Imputation method
A	72,000	72,000	No imputation
B	4,565	4,565	No imputation
C	missing	33,800	Option 1
D	21,025	21,025	No imputation
E	missing	41,020	Option 1
F	missing	2,130	Option 2
G	35,138	35,138	No imputation

After imputation of all missing values, the regional sum is simply:

$$E_{region}^{2016} = \sum\nolimits_{country} E_{country,imputed}^{2016}$$

$$E_{region}^{2016} = 72{,}000 + 4{,}565 + 33{,}800 + 21{,}025 + 41{,}020 + 2{,}130 + 35{,}138$$

$$E_{region}^{2016} = 209{,}678$$

Whether this regional sum and the related regional averages are considered publishable depends on the quality of the imputations (see section 3.4).

4.2. Regional and global averages of ratios

For ratios, the regional average is calculated as a weighted average of the given indicator using the values in the denominator as weights. For example, the school age population is used as the weight in the case of enrolment rates, calculated as the ratio of enrolment over school-age population.

After imputation of the missing values for a given indicator *V* and its weight *W* at country level (*V* is the ratio of the variable *X* and the weight *W*), the regional average is computed based on the following formula:

[9] Because the imputation method is sensitive to regional groupings, the UWGMI is based on the UIS regional groupings (see Annex).

$$V_{region}^{y} = \frac{X_{region}^{y}}{W_{region}^{y}}$$

Which is equivalent to:

$$V_{region}^{y} = \frac{\sum_{country} V_{country}^{y} \times W_{country}^{y}}{\sum_{country} W_{country}^{y}}$$

Where:

$V_{y,region}$: regional average of indicator *V* in year *y*

$X_{y,region}$: regional sum of variable *X* for all countries in the region in year *y*

$W_{y,region}$: regional sum of weight *W* in year *y*

$V_{y,country}$: value of indicator *V* for a given country in year *y*

$W_{y,country}$: weight of indicator *V* for a given country in year *y*

It is worth noting that the weighted regional averages methodology consists of calculating the value of a given indicator for a region as one unit. The results are consequently more influenced by the indicator values of populous countries in the region.

Example 9. Calculation of regional average of a ratio

Consider again Example 4 and assume that countries A, B and C constitute a sub-region *N*. The regional average of the net enrolment rate (NER) is computed based on the NER of each country in sub-region *N*. The values of the NER of primary education and the weights (the primary school-age population) for each country are given in **Table 5**.

Table 4. Example of imputed net enrolment rates in primary education and the corresponding school age population

Country	Imputed net enrolment rate in primary education (NER_1, %)	Imputed school-age population
A	95	1,000
B	94	2,000
C	89	3,000

The regional average of the NER of primary education for sub-region *N* is:

$$NER_{region\ N} = \frac{\sum_{all\ countries} NER_{country} \times SAP_{country}}{\sum_{all\ countries} SAP_{country}}$$

Hence, the regional average of the primary NER of sub-region *N* is:

$$NER_{region\ N} = \frac{95 \times 1{,}000 + 94 \times 2{,}000 + 89 \times 3{,}000}{1{,}000 + 2{,}000 + 3{,}000} = 91.7\%$$

4.2.1. Weight determination

As a rule of thumb, the weight for a given indicator (ratio) is its denominator. **Table 6** gives some examples of indicators and theirs corresponding weights.

Most UIS regional and global estimates for ratios are calculated using this rule. However, there are some exceptions. For instance, the regional and global figures of the school life expectancy for children of age 5 use the population of 5-year-olds as weight and the gender parity index is calculated from the regional figures of its numerator and denominator.

Table 5. Examples of indicators and their corresponding weights

Indicator	Weight
Gross enrolment ratio, net enrolment rate, out-of-school rate	School age population
Net intake rate into primary education	Population of theoretical primary entrance age
Percentage of enrolment in private education	Total enrolment (public and private)
Youth/adult literacy rate	Population of youth/adults (excluding persons with unknown literacy status)
Proportion of schools offering basic services (at a specific level of education)	Total number of schools (at the specific level of education)
Proportion of teachers with the minimum required qualification	Total number of teachers
Percentage of female graduates	Total number of graduates
School life expectancy for children aged 5 years	Population aged 5 years

Source: UNESCO Institute for Statistics (UIS)

4.3. Regional derived estimates

Finally, for reasons of efficiency and consistency many regional statistics are not created as aggregate from country level data but calculated based on other regional aggregates. This is the case for regional estimates that should sum up with other regional statistics to equal a third regional statistics or for indicators for which the needed underlying absolute data are already aggregated. For example, if total enrolment in secondary education, enrolment in secondary general education and enrolment in secondary vocational education would be aggregated separately, this could cause minor inconsistencies due to rounding in the sums. Therefore, aggregation is only done for two of the three data points. The third data point is calculated on the basis of the other imputations to ensure numerical consistency between imputed values. Similarly, GER or NER are calculated from the aggregates of the needed enrolment and population data.

4.4. Quality rating

Regional and global figures are derived from both publishable and imputed country data. Where some countries data are missing and thus imputed, the regional figure is an approximation of its unknown real value. The quality of the resulting estimate depends on the (population) weighted coverage of publishable data and the quality of information from related indicators and auxiliary information. This section discusses how the UIS calculates indices of coverage for regional aggregates to determine whether aggregates can be published.

4.4.1. Coverage indexes

There is no straightforward rule to determine the quality threshold (in terms of the percentage of weighted publishable data) at which regional and global aggregates can be considered valid and publishable. Even if the threshold is defined, the quality of regional estimates will vary depending on the weighted coverage of publishable data used to calculate these figures. For instance, regional estimates based on data with 55 per cent coverage of publishable country data is not as accurate as the estimate for another region with a coverage rate of 95 per cent.

Instead of having only one threshold to decide whether regional estimates are publishable, all estimates are rated to distinguish between regional estimates of high and low weighted coverage.

The UIS assigns quality ratings to regional averages based on two factors: (a) the weight of the imputed values and (b) the time lag (where applicable) between the publishable value on which an imputation is based and the reference year of the imputed value. When calculating regional aggregates, the UIS monitors the imputation process and evaluates its impact. Four coverage indexes are calculated:

1) CI1: The weighted coverage of publishable data in the region in the reference year, defined as the percentage of the population (the weight) covered by publishable country data in the reference year.

2) CI2: The weighted coverage of publishable data in the region in the reference year, the following year, and the preceding year. It is calculated as the percentage of the population (the weight) covered by publishable country data in the reference year, one year before and one year after.

3) CI3: The weighted coverage of both publishable data in the reference year and imputations based on publishable data from years before and after the reference period. It is defined as the percentage of the population (the weight) covered by publishable country data in the reference year and imputations established using publishable data from years before and after the reference year.

4) CI4: Imputation time gap**,** defined as the weighted average distance (in years) between the nearest year with publishable data and the reference year.

4.4.2. Publishable regional averages

Regional and global estimates are considered publishable if one of the following three conditions is met:

1) The weighted coverage of publishable data in the region for the reference year (CI1) is at least 33% and the imputation time gap (CI4) is less than or equal to 4 years;

2) The weighted coverage of publishable data in the region for the reference year, the preceding year or the following year (CI2) is at least 33% and the imputation time gap (CI4) is less than or equal to 4 years;

3) The weighted coverage of both publishable data for the reference year and imputations established based on publishable data from the nearest years before and after the reference year (CI3) is at least 33% and the imputation time gap (CI4) is less than or equal to 4 years.

Publishable regional averages are further classified into two categories depending on the value of the weighted coverage of publishable data: (1) regional averages without qualifier and (2) regional averages with qualifier.

4.4.2.1. Publishable regional averages without qualifier

Publishable regional averages without qualifier refer to regional averages that are calculated based on at least 60% of weighted publishable data in the reference year for countries in the region.

Example 10. Calculation of the coverage index: Case of publishable regional average without qualifier

This example focuses on the calculation of the regional adjusted net intake rate into primary education (ISCED 1) ($ANIR_1$). The weight in this case is the population of school entrance age for primary education.

Suppose that a given region *N* contains ten countries in a given year (2016 is used as example in **Table 7**). Imputation is needed for countries A, E, I and J. For country A no data are available and imputation is based on the unweighted group mean. For countries E and I, imputations are based on data from a previous year (2008) and a following year (2017), respectively. For country J, imputation is based on data from years before and after the reference year (2013 and 2017).

For each imputation the gap between the nearest year with publishable data and the year of imputation is calculated. For country J, the gap is calculated as the absolute value of the difference between the reference year and the nearest year minus 1. The subtraction of 1 is justified by the fact that the imputation with linear interpolation draws on more information than in a situation with only data before or after the reference year (see section 3.3.1). An imputation time gap of 10 years is assigned to imputations based on the unweighted group mean (see section 2.3.2).

After imputation of missing values, the complete set of the $ANIR_1$ for region *N* in 2016 is reported in Table 7.

Table 6. Example of data to calculate a publishable regional average without qualifier

Country	ANIR$_1$ (%)	Data type	Nearest year(s) with publishable data	Imputation time gap (years)	Population of primary school entrance age (weight)
A	89.5	Group mean imputation	--	10	109,985
B	87.6	Publishable	2016	0	515,672
C	78.2	Publishable	2016	0	269,008
D	99.3	Publishable	2016	0	851,456
E	94.6	Nearest year	2008	8	421,300
G	91	Publishable	2016	0	399,400
H	81.3	Publishable	2016	0	235,246
I	88.2	Nearest year	2017	1	673,213
J	94.9	Nearest year	2013 - 2018	1	278,100
K	90.3	Publishable	2016	0	2,138,947
Region *N*	90.8				5,892,327

The main quality index, the coverage index CI1 of publishable values of $ANIR_1$ in region *N* is:

$$CI1 = \frac{515{,}672 + 269{,}008 + 851{,}456 + 399{,}400 + 235{,}246 + 2{,}138{,}947}{5{,}892{,}327} = 74.8\%$$

Because the value of the coverage index CI1 (74.8%) exceeds 60% of the population of official primary school entrance age (the weight), the adjusted net intake rate for region *N* (90.8%) is published without qualifier and without calculation of other quality indicators (coverage indexes).

4.4.2.2. Publishable regional averages with qualifier

If the coverage of the publishable values used in the calculation of a regional average is less than 60%, the regional average is published and marked with a qualifier (two asterisks **) if one of the following conditions is fulfilled:

1. The regional average is calculated based on at least 33% but less than 60% of weighted publishable data (observed or UIS estimates) for countries in the region in the reference year, and the imputation time gap (CI4) is less than or equal to 4 years;

2. At least 33% of weighted national data in the given region are publishable for the reference year or imputed from publishable data for the reference year plus or minus one year, and the imputation time gap (CI4) is less than or equal to 4 years.

3. The regional average is based on publishable data for the reference year or on imputations from publishable data from years before and after the reference year, and the imputation time gap (CI4) between the nearest publishable data and the reference year is less than or equal to 4 years.

Example 11. Calculation of the coverage index: Case of publishable regional average with qualifier

This example is an extension to Example 10 and focuses on the calculation of the regional average of the Adjusted Net Intake Rate (ANIR) for a region where the coverage of publishable data for the reference period (2016) is less than 60% (the weight is the population of school entrance age into primary education). Suppose that the hypothetical region *Z* contains nine countries in 2016.

As illustrated in the **Table 8**, imputation is needed for countries L, N, O, S and T. For countries L, N, O and T imputations are based on the simple nearest year imputation method since some auxiliary information is available. For country S no data are available and imputation is based on the unweighted group mean.

For each imputation the time gap between the nearest year with observed data and the year of imputation is calculated. The imputation time gap for imputation based on the unweighted group mean is set to 10 years (see section 3.3.2).

After imputation of the missing values, the complete set of the ANIR_1 for region Z in 2016 is reported as in Table 8.

Table 7. Example of data to calculate publishable regional average with qualifier

Country	ANIR_1 (%)	Data type	Nearest year(s) with publishable data	Imputation time gap (years)	Population of primary school entrance age
L	86.3	SNYI	2014	2	61,885
M	85.9	Publishable	2016	0	175,569
N	80.2	SNYI	2014	2	77,204
O	91.5	SNYI	2017	1	315,159
P	90.0	Publishable	2016	0	121,874
Q	92.8	Publishable	2016	0	64,600
R	89.3	Publishable	2016	0	181,246
S	88.1	UWGMI	--	10	119,317
T	89.0	SNYI	2008	8	662,906
Region Z	88.9				1,779,760

To assess the quality of the regional estimate of $ANIR_1$ and determine the qualifier to be assigned to its value, coverage indexes are calculated.

1) Coverage index CI1: The coverage rate of publishable values of $ANIR_1$ in region *Z* is:

$$CI1 = \frac{175{,}569 + 121{,}874 + 64{,}600 + 181{,}246}{1{,}779{,}760} = 30.5\%$$

Because the coverage rate is less than 60% of the population of school entrance age for primary education in region Z, the regional average of $ANIR_1$ for that region cannot be published without qualifier. The CI1 value is less than 33%, which indicates that the regional average should not be published. To decide whether the value can nevertheless be published with a qualifier (**), other quality indexes are calculated and evaluated:

2) Coverage index CI2: The weighted coverage rate of publishable data of *ANIR*$_1$ in region Z in 2016, one year before the reference year (2015) or one year after (2017):

$$CI2 = \frac{175{,}569 + 315{,}159 + 121{,}874 + 64{,}600 + 181{,}246}{1{,}779{,}760} = 48.2\%$$

The value of the coverage index CI2 is greater than 33%. In this case, the quality rating is determined by the average imputation time gap.

3) Imputation time gap (CI4)**:** The CI4 for region *Z* is calculated with 0 as the time gap for countries with observed data:

$$impTimeGap = \frac{2 \times 61{,}885 + 2 \times 77{,}204 + 1 \times 315{,}159 + 10 \times 119{,}317 + 8 \times 662{,}906}{1{,}779{,}760} = 3.9 \text{ years}$$

Because the average (weighted) time gap is 3.9 years and below the threshold of 4 years, the indicator can be published and marked with two asterisks (**). Note: 3.9 years can be interpreted as the average distance (in years) that separates 2016 (the reference year) and the theoretical year from which auxiliary information for region Z as unit (e.g. a country) is used to estimate the regional average of *ANIR*$_1$. To explain this statement, consider that data for all countries in the former example are missing and are imputed based on data of 2015 using SNYI. In this case, the imputation time gap for each country is 1 and thereby the *impTimeGap* is 1 too. In this case, the regional average of *ANIR*$_1$ in 20169 for region *Z* is estimated based on auxiliary information of region *Z* as one unit (e.g. one country) in 2015.

4.4.3. Unpublishable regional estimate

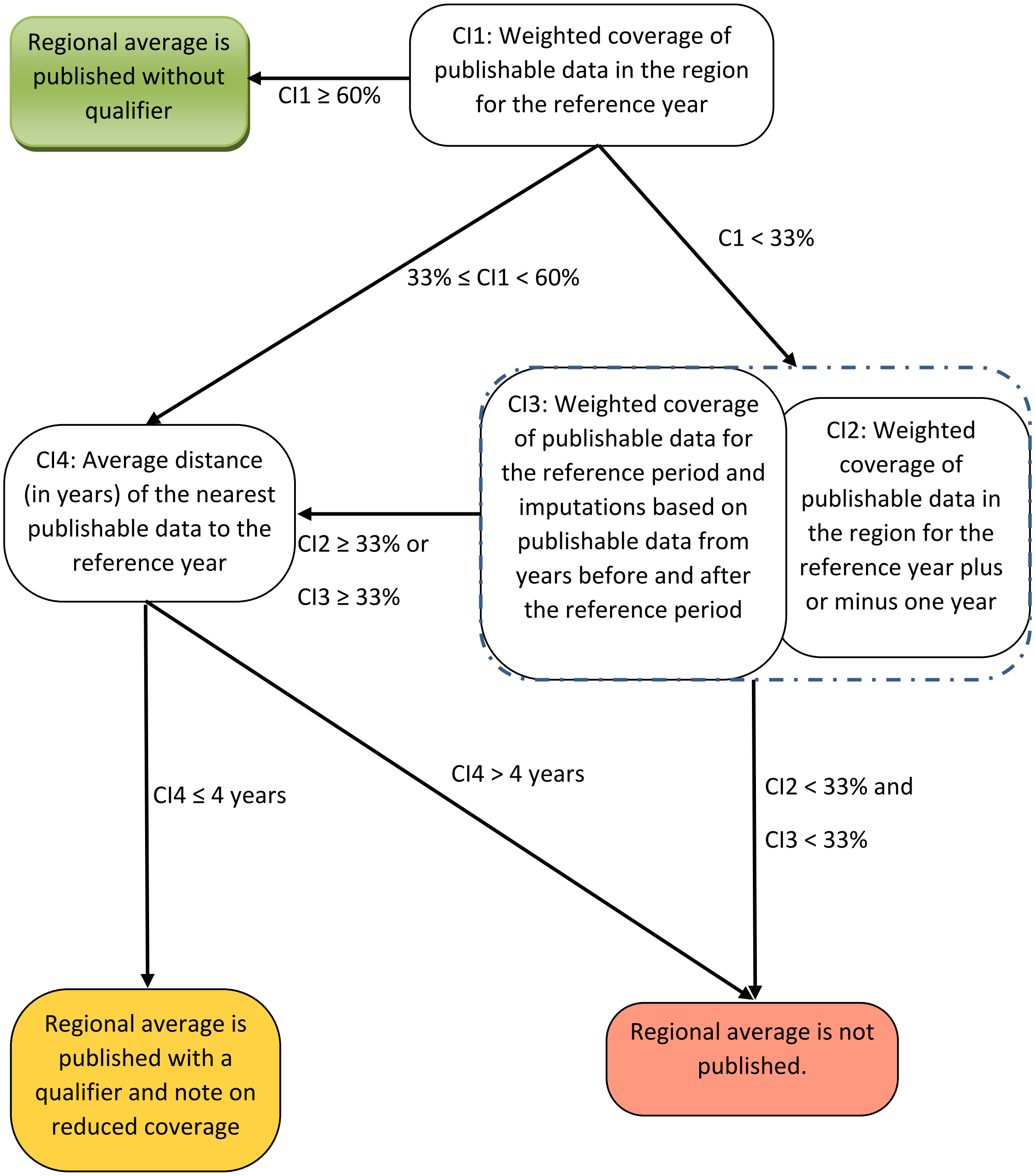
Regional averages that do not meet the minimum quality standards described in section 4.4.2 are considered unpublishable. They are substantially based on imputation, meaning that weighted publishable data for the countries in the region represent less than 33% of the regional population and either:

1) The weighted coverage of publishable data in the region for the reference year plus or minus one year (CI2) is less than 33%, and the weighted coverage of both publishable data for the given reference period and imputations based on publishable data from years before and after the reference period (CI3) is also less than 33%, or

2) One or both coverage indexes (CI2 and CI3) are greater than 33% but the average imputation time gap (CI4) is greater than 4 years.

Figure 10 summarizes the UIS standards used to assess the quality of regional and global figures.

Figure 10. UIS quality standards for regional averages


Regional average is published without qualifier
CI1 ≥ 60%
CI1: Weighted coverage of publishable data in the region for the reference year
C1 < 33%
33% ≤ CI1 < 60%
CI3: Weighted coverage of publishable data for the reference period and imputations based on publishable data from years before and after the reference period
CI2: Weighted coverage of publishable data in the region for the reference year plus or minus one year
CI4: Average distance (in years) of the nearest publishable data to the reference year
CI2 ≥ 33% or
CI3 ≥ 33%
CI4 ≤ 4 years
CI4 > 4 years
CI2 < 33% and
CI3 < 33%
Regional average is published with a qualifier and note on reduced coverage
Regional average is not published.

Source: UNESCO Institute for Statistics (UIS)

Conclusion

This document describes the UIS methodology to generate regional and global figures and the imputation methodology in the presence of missing data. It explains the nature of missing data in the UIS education database, and the requirements and constraints of the UIS education statistics to be produced. These constraints guide the choice of imputation techniques.

The UIS imputation techniques allow constructing complete datasets to calculate regional and global figures for absolute values and ratios. The quality of regional and global figures that are calculated based on publishable data, as well as on imputations, depends on the quality of imputations, which in turn strongly depends on the quality of the auxiliary information used to generate them and on the underlying assumptions.

Missing data for countries with a large population pose a serious challenge for the imputation methodology. This issue is currently addressed with manual imputation even if it generates unpublishable country level data.

Finally, it is worth noting that availability of reported data is of paramount importance to calculate accurate regional and global statistics. Therefore, collection of observed data should remain the essential element in the process of calculating regional figures. Imputation should be seen as the last solution to overcome the problem of missing data and it must never be considered as objective in the calculation of global statistics and replace any effort to collect observed data.

References

Denk, Michaela and Weber, Michael (2011). “*Avoiding Filling Swiss Cheese with Whipped Cream: Imputation Techniques and Evaluation Procedures for Cross-Country Time Series”*. IMF Working Paper WP/11/151, June.

Holt, Tim (2003) “Aggregation of National Data to Regional and Global Estimates”. Report prepared for the Committee for the Coordination of Statistical Activities, Geneva, 8-10 September.

World Bank (2012). “Aggregation Rules”. Available at: http://data.worldbank.org/about/data-overview/methodologies.

Annex: List of principal UIS regional groupings used to impute missing data

Arab States (21 countries or territories)

Algeria, Bahrain, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, , Oman, Palestine, Qatar, Saudi Arabia, Sudan, Sudan (pre-secession). Syrian Arab Republic, Tunisia, United Arab Emirates, Yemen.

Central and Eastern Europe (21 countries or territories)

Albania, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Montenegro, Poland, Republic of Moldova, Romania, Russian Federation, Serbia, Slovakia, Slovenia, The former Yugoslav Republic of Macedonia, Turkey, Ukraine.

Central Asia (9 countries or territories)

Armenia, Azerbaijan, Georgia, Kazakhstan, Kyrgyzstan, Mongolia, Tajikistan, Turkmenistan, Uzbekistan.

East Asia and the Pacific (34 countries or territories)

East Asia (17 countries or territories)

Brunei Darussalam; Cambodia; China; China, Hong Kong Special Administrative Region; China, Macao Special Administrative Region; Democratic People's Republic of Korea; Indonesia; Japan; Lao People's Democratic Republic; Malaysia; Myanmar; Philippines; Republic of Korea; Singapore; Thailand; Timor-Leste; Viet Nam.

Pacific (17 countries or territories)

Australia, Cook Islands, Fiji, Kiribati, Marshall Islands, Micronesia (Federated States of), Nauru, New Zealand, Niue, Palau, Papua New Guinea, Samoa, Solomon Islands, Tokelau, Tonga, Tuvalu, Vanuatu.

Latin America and the Caribbean (42 countries or territories)

Latin America (19 countries or territories)

Argentina, Bolivia (Plurinational State of), Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, Venezuela (Bolivarian Republic of).

Caribbean (24 countries or territories)

Anguilla, Antigua and Barbuda, Aruba, Bahamas, Barbados, Belize, Bermuda, British Virgin Islands, Cayman Islands, Dominica, Grenada, Guyana, Haiti, Jamaica, Montserrat, Curaçao, Sint Maarten (Dutch part), Puerto Rico, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, Turks and Caicos Islands.

North America and Western Europe (31 countries or territories)

Andorra, Austria, Belgium, Canada, Cyprus, Denmark, Faroe Islands, Finland, France, Germany, Gibraltar, Greece, Greenland, Holy See, Iceland, Ireland, Israel, Italy, Liechtenstein, Luxembourg, Malta, Monaco, Netherlands, Norway, Portugal, San Marino, Spain, Sweden, Switzerland, United Kingdom of Great Britain and Northern Ireland, United States of America.

South and West Asia (9 countries or territories)

Afghanistan, Bangladesh, Bhutan, India, Iran (Islamic Republic of), Maldives, Nepal, Pakistan, Sri Lanka.

Sub-Saharan Africa (46 countries or territories)

Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Congo, Côte d'Ivoire, Democratic Republic of the Congo, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritius, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, South Sudan, Togo, Uganda, United Republic of Tanzania,