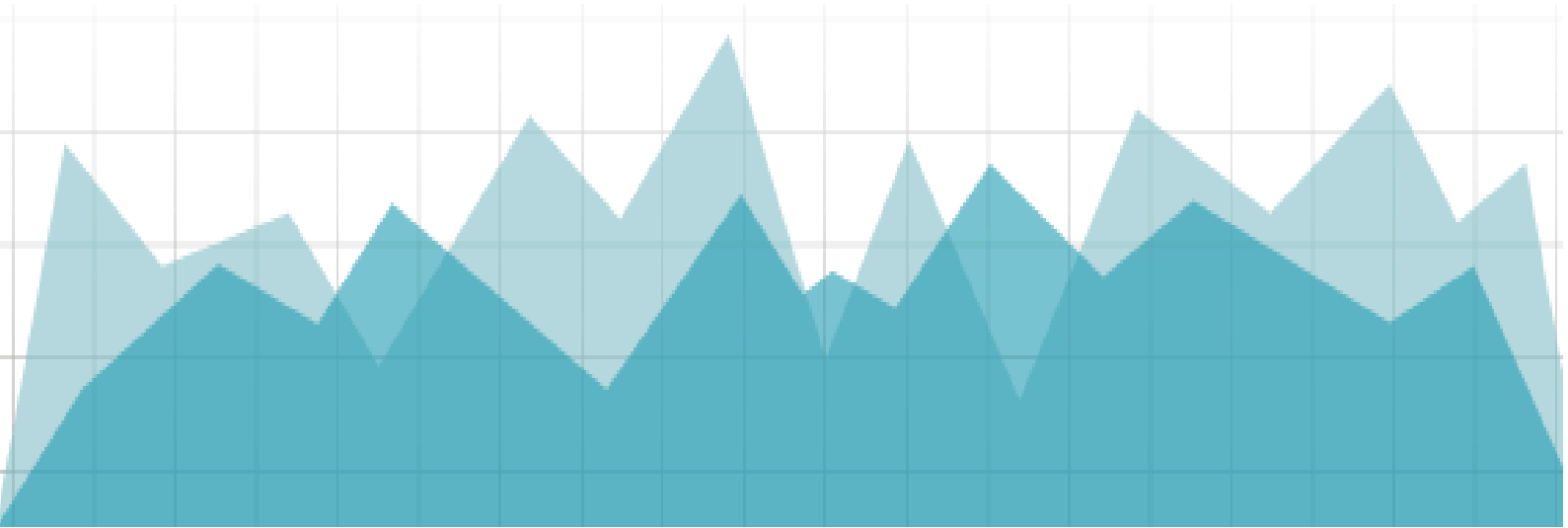
Food and Agriculture Organization of the United Nations

Statistical Standard Series

Data aggregation

Endorsed by the Inter-Departmental Working Group
Technical Task Force on Statistics

19 December 2019



The FAO Statistical standard on data aggregation provides guidance and recommendations on the production of statistical aggregates (e.g. regional aggregates using country estimates).

It includes key definitions as well as general and technical recommendations on the compilation of aggregated sums, averages, proportions, ratios, growth rates and more complex indicators.

Contents

BACKGROUND

The production of FAO statistical outputs generally involves a data processing phase whose final step consists in the calculation of aggregates (Generic Statistical Business Process Model (GSBPM) sub-phase 5.7). In particular, collected data, after suitable treatment (to recode, convert unit of measurement, validate, etc.), are processed to derive statistical outputs referred to the whole target population being investigated and to specific sub- populations (**see Annex 1 for definition**). For instance, FAO's statistical processes that collect data at country level usually end up producing statistical outputs (estimates) referring to established country groupings (e.g. regions) and to the whole world (so called "global" estimates). This standard series document addresses issues related to the calculation of some well-known aggregates that can be generated using sum, mean (average) and ratio functions.

DEFINITIONS

Most of the aggregates calculated at the end of a statistical process are, in fact, sums, means or ratios.

- **Sum** is used to count the number of occurrences of an event (frequencies) or derive the total amount of a given continuous variable (e.g. arable land).

- **Mean (average)** is the sum of values divided by the number of units (or items) contributing to it.

- **Ratio** is obtained by dividing the values of two related variables, e.g. the yield for a particular crop in agriculture corresponds to the total production (usually in tons) divided by the corresponding area planted (in hectares).
  - **Proportion** is a special type of ratio in which the denominator includes the numerator, e.g. area with a specific crop divided by total area; **number of occurrences** of an event divided by the total number of cases (denoted as **relative frequency of occurrence**; percentage if multiplied by 100). The relative frequency of an event can be seen as an average (variable recoded in 1 if the events occur and 0 otherwise).
  - **Growth rate** (or **rate of change**) from one period to the next. It is a particular ratio that compares the total change in a specified time reference period to the value at the beginning of the period or at a specified earlier time reference, e.g. (present-past)/past (sometimes multiplied by 100; **see Annex 2 for formulas**).

For major details on the definitions, please see Annex 1.

GENERAL RECOMMENDATIONS

Calculating aggregates can be straightforward in the case of sums, while in the case of ratios, different strategies can be chosen to address the difficulties raised by the presence of non-responses (data gaps or missing values), i.e. statistical units that do not report a value for a part or all the requested variables[1].

When the data gaps are filled-in using imputation (see corresponding Statistical Standard Series), then the aggregates are calculated by treating the imputed values as if they were observed. In certain situations, imputation is performed just for obtaining the required aggregates, while the single imputed value is not disseminated externally. The accuracy of the estimates calculated using observed and imputed values will also depend on the imputation procedure, which is a source of additional variance (wider confidence intervals than expected) and can introduce also bias (i.e. underestimation or overestimation of the target aggregate). Calculating regional or global level estimates by excluding the non-reporting units implicitly involves making some assumptions about the contribution of non-reporting units to the target aggregates.

Sums

In the absence of missing values (i.e. when all data items are available), the total amount for a group of units (or all the units) is obtained by summing up the observed values. The same applies when data gaps have been filled-in by imputation. In that latter case, it is important to assess the contribution of imputed values to the final total. If the non-reporting units are expected to have a negligible contribution to the total amount[2], then the sums of the observed values (discarding the missing values) may still provide an accurate estimate of the total (affected by negligible bias, i.e. a slight underestimation). When missing values are expected to have a non-negligible contribution on the totals, summing only the respondents' values will provide estimates affected by bias (i.e. a significant underestimation of the target total). To avoid this problem, some Agencies have established threshold-based rules. For instance, the World Bank calculates sums without imputing missing values only if the proportion of missing values does not exceed one third of the observations[3].

Means

Their calculation is straightforward in the absence of missing values or when missing values have been imputed. The same rationale as for sums applies.

Care should however be exercised when discarding missing values during the compilation of aggregates using means. The practice of considering just the observed units in the calculation – using

[1] Missing value here refers to the case in which a statistical unit did not provide a value for "existing" phenomena. A missing value should be distinguished from an "impossible" value, i.e. case in which a phenomenon cannot exist for a specific unit.

[2] The rough quantification of bias requires additional information (e.g. historical data) or a deep understanding of phenomenon being studied, that usually permits to state that relative contribution of non-reporting units is below a given threshold. Commonly used threshold are 2%, 5% or 10%.

[3] https://datahelpdesk.worldbank.org/knowledgebase/articles/198549-what-methods-are-used-to-calculate-aggregates-for

the total estimated on the reporting units divided by the number of reporting units – would correspond to a mean imputation (i.e. replacing each missing value with the mean calculated on the reporting units). This assumption is very strong and may not be valid.

When it is assumed that the non-reporting units have a negligible contribution to the total amount of a given variable (see footnote 2), then the average should be calculated as the ratio between the sum over the responding and total number of units (both reporting and not).

Proportions

When proportions are calculated discarding missing values special consideration should be given to whether and how missing values affect both the numerator and the denominator. For categorical variables, where proportions correspond to relative frequencies of occurrence, the estimation strategy in the presence of missing values are similar to the case of the mean. Estimating the relative frequency based on the observed cases (discarding missing values) corresponds to assuming that the same relative frequency would be observed on the subset of units presenting missing values. With relative frequencies, understanding how missing values can affect the final estimate is straightforward, by calculating the interval of the admissible estimates assuming that all the missing values present the given characteristic or not (See Annex 2 for major details).

Aggregating ratios (and proportions)

Aggregation of ratios is needed when the ratios calculated at country level should be processed to obtain global or regional estimated ratios (i.e. yield at regional level). Aggregation can be done in different manners; typically, average functions are considered and, in particular, *weighted averages* are preferred, where the weight is often the value at the denominator of each single ratio; the advantage of this is that the final aggregate corresponds to the ratio between the sums (i.e. ratio between the sum of values of the variable at the numerator and the sum of values of the variables at the denominator of the ratio; **see Annex 2** for major details). In some circumstances, an alternative weighting system can be applied to better represent the "importance" of each unit in the statistical domain being considered. For instance, in economic statistics, the GDP is a commonly used weight when calculating the regional aggregates, while in social statistics it is considered the country population and, in environmental statistics, the land area.

The aggregation of ratios can also simply consist in the *unweighted average* of the observed ratios. This choice gives equal importance to all units, but then the aggregate will not correspond to the ratio between the sums. In addition, simple averages are sensitive to extreme values.

Aggregation of ratios in the presence of missing values may be challenging since all units where one or both the values involved in the ratio are missing must be discarded; this way of working may dramatically reduce the number of units usable for calculations (i.e. the ones having both the values observed). As in the previous case, aggregation of available ratios can be done using an average function (weighted or unweighted); the accuracy of the resulting estimated aggregates will depend on many factors: (i) the aggregation method; (ii) the expected impact of missing values on the single variables' aggregates; and (iii) the relationship between variables involved in the ratios at unit level.

As mentioned before, a simple average of ratios based on reporting units only, corresponds to assuming that non-reporting units share the same average ratio (i.e. imputation of the average ratio to all non-reporting units).

Summarizing growth rates

The calculation of growth rates requires some care when averaging over time series or calculating regional or global level aggregates.

The *average growth ratio* of a time series can be calculated in different ways, depending on the growth model followed by the phenomenon being studied. Four common methods can be used: (i) arithmetic, (ii) geometric endpoint, (iii) exponential endpoint, and (iv) least squares (**see Annex 2 for formulas**).

The arithmetic averaging assumes that the variable of interest increases by a constant amount in each period; this is a very simplistic assumption which is rarely true. Geometric formulas assume a *compound growth* over discrete periods, while exponential formulas assume continuous growth compounded over time. The geometric growth formula is commonly adopted for indicators related to economic phenomena, while the exponential growth formula is considered for indicators related to the population. The least squares method does not assume a specific growth pattern; rather it takes into account all the observations in the time series and can be used for both demographic and economic phenomena.

Aggregation of growth rates when grouping data (e.g. countries forming a region) can be done by calculating the weighted average of the growth rates contributing to it or by considering the time series of the group totals.

More complex indicators

Calculation of complex indicators for group of units may require specific methods, to be decided on a case-by-case basis, according to the phenomena analysed and the objective of the analysis. In some circumstances, the aggregated indicators are derived by substituting the various quantities with the corresponding aggregates (sums); in other cases, a weighted average (or a simple average) can be a good approximate solution. Ad hoc complex indicators may require the development of specific aggregation strategies which also determine the rules for dealing with eventual missing values.

TECHNICAL RECOMMENDATIONS

- Both statisticians and subject matter experts should be involved in developing the strategies for aggregating data in line with sound methodologies. The objectives, methods and underlying assumptions should be clearly stated and documented.

- Implementation of the aggregation procedure in the chosen software package should be carefully controlled to avoid processing errors. The procedure should be tested and should provide reproducible results.

- The disseminated aggregates should be accompanied by all the relevant information to enable their correct interpretation (unit of measure, reference periods, etc.). In particular, when disseminating ratios, the quantities being compared (numerator and denominator) and the constant by which the ratio has been multiplied (if any) should be specified. Similarly, for growth rates it is important to clarify the reference periods, if they are expressed in percentage terms or not, etc.

- Coherence in the aggregation procedure should be addressed by explaining whether it is possible to obtain the overall aggregate (global) by using only the sub-aggregates (regional estimates) by applying the same method that was used to calculate the sub-aggregates.

- When the aggregates are disseminated jointly with the elementary data contributing to them, the users should be informed about the coherence of the aggregation process; in addition, it should be clearly stated whether they could perform the aggregation by themselves (achieving the same result).

- Users should be informed when the aggregates are calculated considering only the reporting units (i.e. discarding missing values) and they should be made aware about the possible consequences of this choice on the accuracy of final statistical outputs (aggregates) being disseminated.

- When imputation is used to fill-in the data gaps (sometimes just for aggregation purposes), the assumptions underlying the imputation procedure should be clearly stated and users should be informed about them. The fraction of missing values and expected impact of imputed values on the final aggregates should be assessed and communicated to the users.

GOVERNANCE PROCEDURES

- Technical units are responsible to identify and apply the most appropriate aggregation methods. The Office of the Chief Statistician can provide support if necessary.
- Technical units should apply consistent aggregation methods overtime. Any changes in the aggregation procedure should be clearly motivated and, when applicable, the impact on the time series of published statistical results should be carefully assessed. Users should be informed accordingly.

ANNEX 1: ADDITIONAL DEFINITIONS

Statistical unit

An object of statistical survey and the bearer of statistical characteristics. The statistical unit is the basic unit of statistical observation within a statistical survey. Statistical units are the entities for which information is sought and for which statistics are ultimately compiled.

http://www.unece.org/fileadmin/DAM/stats/publications/53metadaterminology.pdf

https://stats.oecd.org/glossary/detail.asp?ID=6157

Target population

The set of elements about which information is wanted and estimates are required.

https://stats.oecd.org/glossary/detail.asp?ID=2645

Aggregation

The combination of related categories, usually within a common branch of a hierarchy, to provide information at a broader level to that at which detailed observations are taken.

With standard hierarchical classifications, statistics for related categories can be grouped or collated (aggregated) to provide a broader picture, or categories can be split (disaggregated) when finer details are required and made possible by the codes given to primary observations.

https://stats.oecd.org/glossary/detail.asp?ID=68

Average value

A familiar but elusive concept. Generally, an "average" value purports to represent or to summarise the relevant features of a set of values; and in this sense the term would include the median and the mode. In a more limited sense an average compounds all the values of the set, e.g. in the case of the arithmetic or geometric means. In ordinary usage "the average" is often understood to refer to the arithmetic mean.

https://stats.oecd.org/glossary/detail.asp?ID=3601

Ratio

A ratio is a number that expresses the relative size of two other numbers. The result of dividing a number X by another number Y is the ratio of X to Y.

https://stats.oecd.org/glossary/detail.asp?ID=6688

Proportion

A proportion is a special type of ratio in which the denominator includes the numerator. An example is the proportion of deaths that occurred to males which would be deaths to males divided by deaths to males plus deaths to females (i.e. the total population).

https://stats.oecd.org/glossary/search.asp

Rate

A rate refers to the occurrence of events over a specific interval in time. Similarly, a rate refers to the measure of the frequency of some phenomenon of interest.

Caution must be used with the term “rate” as it is sometimes applied to ordinary percentage changes such as a “literacy rate” which is the percentage of a population that is literate. Different constants (commonly 100, 1 000, 100 000) are used in the presentation of different rates (e.g. crude death rates and crude birth rates are usually expressed per 1 000).
https://stats.oecd.org/glossary/detail.asp?ID=6691

Growth rate (or rate of change)
Is a ratio of total change in a specified time reference period to values at the beginning of the period or at a specified earlier time reference.
https://stats.oecd.org/glossary/detail.asp?ID=2236

Index number
A quantity which shows by its variations the changes of a magnitude over time or space.
Index type refers to any of the various indices (e.g., Laspeyres, modified Laspeyres, Paasche, Value-Added, Fisher, Törnqvist, etc.) used in the Index type refers to any of the various indices (e.g., Laspeyres, modified Laspeyres, Paasche, Value-Added, Fisher, Törnqvist, etc.) used in the statistical production process. Important features in the construction of an index number are its coverage, base period, weighting system and method of averaging observations.
https://stats.oecd.org/glossary/detail.asp?ID=3750

Price index
Reflects an average of the proportionate changes in the prices of a specified set of goods and services between two periods of time.
https://stats.oecd.org/glossary/detail.asp?ID=2110
https://www.imf.org/external/np/sta/tegppi/gloss.pdf

Indicator (statistical indicator)
Data element that represents statistical data for a set of characteristics, one of which allows for meaningful comparisons of the data.
An aggregation such as the number of accidents, total income or female members of Parliament, are not in themselves indicators for comparison across countries, as they are not comparable between populations. However, if a transformation is applied to make the data comparable, e.g. number of accidents per thousand of population, average income, or female members of Parliament as a percentage of the total, the result meets the criteria for an indicator.
Indicators can be used to reveal relative positions and/or show positive or negative change.
https://sdmx.org/wp-content/uploads/SDMX_Glossary_Version_2_0_October_2018.docx

Composite indicator

A composite indicator is formed when individual indicators are compiled into a single index, on the basis of an underlying model of the multi-dimensional concept that is being measured. A composite indicator measures multi-dimensional concepts (e.g. competitiveness, e-trade or environmental quality) which cannot be captured by a single indicator. Ideally, a composite indicator should be based on a theoretical framework / definition, which allows individual indicators / variablesto be selected, combined and weighted in a manner which reflects the dimensions or structure of thephenomena being measured.

https://stats.oecd.org/glossary/detail.asp?ID=6278

ANNEX 2: TECHNICAL DETAILS AND FORMULAS

Notations:

j: identifier of a statistical unit;
n: total number of statistical units (or total number of units belonging to a group);
m: number of reporting statistical units ($n - m$ are the non-reporting units or non-respondents);
x_j: value of variable *X* observed on unit *j*;
y_j: value of variable *Y* observed on unit *j*;

Number of occurrences of the event "A":

$$n_A = \sum_{j=1}^{n} I(y_j = 'A')$$

where $I(y_j = 'A') = 1$ if the condition within parenthesis (unit *j* presents a value of *Y* equal to 'A') is met and 0 otherwise.

Relative frequency of occurrence of the event "A":

$$p_A = \frac{n_A}{n}$$

Sometimes multiplied by 100, to express the change in percentage terms.

Total amount of *Y* (continuous variables):

$$s_y = \sum_{j=1}^{n} y_j$$

Average of *Y* is:

$$\bar{y} = \frac{s_y}{n} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

Ratio between the values of *X* and *Y* observed on the unit j:

$$r_j = x_j / y_j$$

Growth rate for variable *Y* between time *t* and time *t+1*:

$$g = \frac{y_{t+1} - y_t}{y_t} = \frac{y_{t+1}}{y_t} - 1$$

Sometimes multiplied by 100, to express the change in percentage terms. g can also be computed by considering two non-consecutive points in a time series of a given variable (e.g. after 10 years).

Aggregation of ratios

Simple average of ratios:

$$\bar{r} = \frac{1}{n}\sum_{j=1}^{n} r_j = \frac{1}{n}\sum_{j=1}^{n} \frac{x_j}{y_j}$$

Weighted average of ratios:

$$\bar{r}_w = \frac{\sum_{j=1}^{n} r_j w_j}{\sum_{j=1}^{n} w_j}$$

If the weight is set equal to the variable at the denominator of the ratio, i.e. $w_j = y_j$, then:

$$\bar{r}_{w=y} = \frac{\sum_{j=1}^{n} r_j w_j}{\sum_{j=1}^{n} w_j} = \frac{\sum_{j=1}^{n} r_j y_j}{\sum_{j=1}^{n} y_j} = \frac{\sum_{j=1}^{n} x_j}{\sum_{j=1}^{n} y_j} = \frac{s_x}{s_y}$$

In other words, in this case, the weighted average of the ratios is equal to the ratio of the totals.

Sums and averages with non-reporting units

In presence of missing values, i.e. when only m units ($m < n$) are observed, discarding the missing values means having:

$$s_{ym} = \sum_{j=1}^{m} y_j \leq s_y$$

and

$$\bar{y}_m = \frac{s_{ym}}{m} = \frac{1}{m}\sum_{j=1}^{m} y_j$$

As said earlier, if s_{ym} is close to s_y because of negligible contribution of missing units, then an estimate of the whole sum would be provided by summing over reporting units:

$$\hat{s}_y = s_{ym}$$

and consequently, an estimate of the mean is obtaining by dividing it by all the units (both responding and nonresponding):

$$\hat{\bar{y}} = \frac{s_{ym}}{n}$$

Proportions with non-reporting units

If m denoted the number of reporting units and m_A is the occurrence of event "A" in the subset of the reporting units, then the estimated relative occurrence of "A" over the responding is:

$$p_{Am} = \frac{m_A}{m}$$

It easy to assess how non-reporting units can affect the estimation of the relative frequency, since:

$$\frac{m_A}{n} \leq p_A \leq \frac{m_A + (n-m)}{n}$$

In practice a range of possible values of the relative frequency is obtained by imputing all the non-reporting units with respectively 0 (event "A" does not occur) or 1 (occurrence of "A").

Aggregation of ratios with non-reporting units

The simple average of ratios on responding units is:

$$r_m = \frac{1}{m}\sum_{j=1}^{m} r_j = \frac{1}{m}\sum_{j=1}^{m} \frac{x_j}{y_j}$$

If it is considered a valid estimate of $\bar{r}$, then the underlying assumption is that the $n - m$ non-reporting units show the same ratio.

The weighted version is:

$$\bar{r}_{mw} = \frac{\sum_{j=1}^{m} r_j w_j}{\sum_{j=1}^{m} w_j},$$

That, when $w_j = y_j$, becomes

$$\bar{r}_{m,w=y} = \frac{s_{xm}}{s_{ym}}$$

If this latter estimate is considered as a valid estimate of $\bar{r}$, the underlying assumption is that both s_{xm}

and s_{ym} are both valid estimates of s_x and s_y, respectively (i.e. nonreporting units have a negligible contribution to both the sums of the target variables).

Summary of growth rates over time

In the presence of a regular time series, as for yearly data, achieving a summary measure of the growth rate over time is not straightforward.

Let T be the observations in the time series, say one observation per year, the arithmetic growth rate is obtained as:

$$\bar{g} = \frac{1}{T}(y_T - y_1)$$

The method uses only first (y_1) and last observation (y_T) in the series and assumes that the variable *Y* increases yearly by a fixed amount ($\bar{g}$). This assumption is seldom valid in the real world. In fact, usually, time series show a compound growth over time.

When a yearly compound growth is assumed (compound growth over discrete periods) the geometric average should be considered:

$$\bar{g}_{GEO} = \left(\frac{y_T}{y_1}\right)^{(1/T)} - 1$$

In practice, the changes between two periods differ by a constant ratio:

$$y_T = y_1(1 + \bar{g}_{GEO})^T$$

This growth rate is a special case of exponential growth, and it is used for indicators related to economic phenomena (trade, GDP, etc.).

Finally, when the compound growth takes places continuously over time, the exponential growth rate should be applied:

$$g_{EXP} = \frac{1}{T} \times ln\left(\frac{y_T}{y_1}\right)$$

Whereas the underlying growth model is:

$$y_T = y_0 \times exp(T \times \bar{g}_{EXP})$$

This assumption is the one being usually considered for growth of demographic indicators (population, etc.).

Another possibility of summarizing compound growth relies on fitting a linear regression trend line on the log-transformed values, i.e.

$$ln(y_t) = ln(y_0) + ln(1 + g_{OLS}) \times t$$
$$= \alpha + \beta \times t$$

Estimates of both α and β are derived by applying the ordinary least squares (OLS) method using all the observations in the time series (not just the first and the last), and finally:

$$\hat{g}_{OLS} = exp(\hat{\beta}) - 1$$

This method is suited for almost all growth patterns, but a sufficiently long time series is required achieve a reliable estimate of the parameters of the model[4].

Fitting a linear trend works with different types of growth patterns, but requires the availability of sufficiently long time series, without missing values. Whereas the other "averaging" methods rely uniquely on the extremes of the whole series (first and last observations) and therefore require the availability of just these two values (the other ones may be also missing).

Growth rates for group of units

Aggregation of growth rates is required to estimate, for instance, the growth rate at regional or global level. Two aggregation methods are usually suggested: (a) weighted average of single growth rates or (b) calculation of aggregated growth rate from the series of group totals.

In the absence of missing values, both the methods are straightforward. As usual, the weighted average requires selecting a weighting system (w_j) that should reflect the importance of each single unit with the respect of the phenomenon being studied:

[4] There is not a general rule that identifies the minimum length of as time series to fit a linear trend model with two parameters to be estimated. A popular rule of thumb claims that at least $n = 20$ observations are needed.

$$g_W = \frac{\sum_{j=1}^{n} w_j g_j}{\sum_{j=1}^{n} w_j}$$

Estimating the aggregated growth rate from the time series of group totals is simpler, since it just requires the sum of the elements in the time series. In particular, if y_{tj} is the value of *Y* observed for country *j* at time *t*, then the total for the group of countries at time *t* will simply correspond to the sum of values:

$$s_{y,t} = \sum_{j=1}^{n} y_{tj}$$

And consequently the aggregated growth rate at $t + 1$ compared to t is:

$$g_A = \frac{s_{y,t+1} - s_{y,t}}{s_{y,t}} = \frac{s_{y,t+1}}{s_{y,t}} - 1$$

Similarly, “average” growth rates can be calculated according to the assumed growth pattern by considering group totals for the reference time periods.

Aggregation can pose a number of issues when one or more time series are incomplete (because of some missing values) or completely missing (i.e. it is not available the whole time series of a country within a given region). In such cases, different strategies can be adopted ranging from discarding of missing values/time series (only available data are considered) to imputing them and using observed and imputed to calculate the required aggregates. There is no clear-cut distinction between the two opposite approaches since, as already mentioned, the simple average of growth rates for available countries would correspond to assuming that growth rates for non-reporting countries are equal to the average value estimated on reporting. The Word Bank suggests not computing aggregated growth rates when a half of the observations for a period are missing.

REFERENCES

Economic and Social Commission for Asia and the Pacific (ESCAP). 2015. *Average growth rate: computation methods*. Stats Brief, Issue No. 07. Bangkok, ESCAP. https://www.unescap.org/sites/default/files/Stats_Brief_Apr2015_Issue_07_Average-growth-rate.pdf

World Bank (WB). 2018. Methodologies. Data Compilation Methodology. In *The World Bank*. Washington, DC. Cited 10 December 2019. https://datahelpdesk.worldbank.org/knowledgebase/articles/906531-methodologies

WB. 2018. What methods are used to calculate aggregates for groups of countries? Data Compilation Methodology. In *The World Bank*. Washington, DC. Cited 10 December 2019. https://datahelpdesk.worldbank.org/knowledgebase/articles/198549-what-methods-are-used-to-calculate-aggregates-for

WB. 2018. How are aggregate growth rates computed for National Accounts series? Data Compilation Methodology. In *The World Bank*. Washington, DC. Cited 10 December 2019. https://datahelpdesk.worldbank.org/knowledgebase/articles/114952-how-are-aggregate-growth-rates-computed-for-nation