

Standardized description of index numbers for application in generic index computation software modules

Photis Stavropoulos¹, Georges Pongas², Spyros Liapis¹, George Petrakos¹, Tonia Ieromnimon¹

¹Agilis S.A. Statistics and Informatics, e-mail: Photis.Stavropoulos@agilis-sa.gr,
Spyros.Liapis@agilis-sa.gr, George.Petrakos@agilis-sa.gr,
Tonia.Ieromnimon@agilis-sa.gr

²EUROSTAT, e-mail: Georges.Pongas@ec.europa.eu

Abstract

The aim of this paper is to present a scheme for the description of index numbers in a standardized manner and to show how this scheme was realized in a generic software module for the computation of index numbers. Index number theory provides statisticians with a palette of formulae and computational operations (chaining, aggregation, etc.) to build index calculation since there is no unique way of measuring an index number. The proposed scheme has been developed in such a way so as to allow the user to define the characteristics/parameters of the index: index type, base period, reference period of the weights, periodicity of input data and of the index, dissemination dimensions, mode of aggregation, etc. These characteristics may be viewed as ‘process metadata’, which can guide the computation of an index. The tool allows the different mix of index types and computational operations. It covers computational tasks common in several domains where indices are used. Moreover it can be expanded to accommodate new operations, with little extra programming, that apply to particular domains and are needed for the computational of additional indices. Further work is envisaged in making the tool even more generic to accommodate more computations and be applicable to different domains.

Keywords: Index numbers, Process metadata, Building block

1. Introduction

Index numbers are widely used in Official statistics to convey information about the relative size of a variable (price, quantity, etc.) between different points in time or between different geographical locations. Consumer price indices and purchasing power parities are examples of the former and latter type of use respectively.

A large number of index number formulae are available to the Official statistician who wishes to select the most appropriate one for each application. The statistician can also choose whether or not to use a chained index. Moreover, the statistician needs to specify the periodicity of the index, which may differ from the periodicity of the raw data (e.g. monthly data may be used to produce monthly, quarterly or annual indices), the dissemination ‘dimensions’ to break the index down by (e.g. a consumer

price index may be disseminated broken down by items, regions, income classes, etc) and the particular classes of each dimension.

There are alternative ways to compute an index at different periodicities or different levels of dissemination breakdowns, in other words to aggregate the index. A quarterly index for example, may be a weighted average of the respective monthly indices or may be computed from quarterly aggregates of the monthly data that produce the monthly indices; similarly, the price index of a dissemination class may be a weighted average of the indices of the component sub-classes, may be a ratio of aggregated numerators and denominators of the component sub-indices or may be computed on aggregates of the input data that produced the sub-classes' indices. Quite often each alternative produces a slightly different value for the aggregated index.

2. Index formulae and operations applied to index numbers

2.1 Index formulae

Several alternative kinds of mathematical formulae are in common use for the calculation of index numbers. Depending on the objective the index is calculated for, the statistician needs primarily to decide on the index type (Lowe, Laspeyres, Paasche, Fisher, etc.) and the series of weights. For instance, CPIs intend to measure either price inflation (price index) between two time periods for a given set of quantities ('basket') or changes in the amount of products that a household consumes for a given budget (quantity index). The choice between Laspeyres or Paasche or Lowe type index depends primarily on the weight reference period, which could be either the base period or the current period or any period different from the previous two. In that case, timeliness and appropriateness of weights are therefore vital.

Commonly used indices are the following: Laspeyres, Paasche, Fisher, Bowley, Lowe, Edgeworth (also known as the Marshall-Edgeworth index), Walsh, Törnqvist-Theil (also known as Törnqvist index), Arithmetic mean – aggregation and two PPP-specific types: Fisher type binary PPP at Basic Heading level and Fisher type binary PPP above Basic Heading.

Examples on the computation of six of them are based on the set of the mathematical formulae presented in Table 1.

Let

${}_{T_0}P_{T,D}^{(G)}$	denote an index number of type G ,
$w_{T,i}$	denote index weights,
$v_{T,i}, V_{T,i}$, etc	denote variables participating in the index formulae. Sometimes they might be indices.

where T is the 'current' point and T_0 is the base point of the reference dimension (time, country, etc.) and D a set of indicators denoting the value of other dimensions.

Table 1. Examples of common index numbers and their formulae

Index	Formula	Index	Formula
Laspeyres	${}_{T_0}P_{T,D} = \frac{\sum_{i \in D} w_{T_0,i} \cdot v_{T,i}}{\sum_{i \in D} w_{T_0,i} \cdot v_{T_0,i}}$	Walsh	${}_{T_0}P_{T,D} = \frac{\sum_{i \in D} \sqrt{w_{T,i} \cdot w_{T_0,i}} \cdot v_{T,i}}{\sum_{i \in D} \sqrt{w_{T,i} \cdot w_{T_0,i}} \cdot v_{T_0,i}}$
Paasche	${}_{T_0}P_{T,D} = \frac{\sum_{i \in D} w_{T,i} \cdot v_{T,i}}{\sum_{i \in D} w_{T,i} \cdot v_{T_0,i}}$	Arithmetic mean - aggregation	${}_{T_0}P_{T,D} = \frac{\sum_{i \in D} w_{T,i} \cdot {}_{T_0}P_{T,i}}{\sum_{i \in D} w_{T,i}}$
Lowe ¹	${}_{T_0}P_{T,D} = \frac{\sum_{i \in D} w_{t,i} \cdot v_{T,i}}{\sum_{i \in D} w_{t,i} \cdot v_{T_0,i}}$	Fisher type binary PPP above Basic Heading level	$F_{T T_0} = \sqrt{\frac{\sum w_{T_0,i} \cdot \frac{v_{T,i}}{v_{T_0,i}}}{\sum w_{T,i} \cdot \frac{v_{T_0,i}}{v_{T,i}}} \cdot \frac{\sum w_{T,i}}{\sum w_{T_0,i}}}$

Exceptionally, for the computation of PPP a single formula does not apply. For the computation of PPP indices PPP formulae are applied in a repetitive way in all possible combinations of country pairs. PPP calculation is a streamlined process that involves individual steps that must be executed in the correct order:

- EKS direct Fisher type PPP (below Basic Heading)

For each ordered pair of countries T_0 and T^2 a Fisher type PPP is computed with country c as base country, using the formula 'Fisher type binary PPP at Basic Heading level'. The products range across all items of the basic heading which are priced in both countries.

- EKS direct Fisher type PPP (above Basic Heading)

For each ordered pair of countries T_0 and T a Fisher type PPP is computed with country c as base country, using the formula 'Fisher type binary PPP above the Basic Heading level'. The sums range across all Basic Headings of the desired aggregate.

- EKS transitivity adjustment

This procedure (a) fills gaps in a list of binary direct Fisher-type PPPs and (b) adjusts the binary Fisher-type PPPs to make them transitive.

(a) The ordered pairs of countries T_0 and T for which no Fisher type PPP was computed (due to missing data) are estimated with the help of indirect PPPs, as follows

$$F_{T|T_0} = \left[\prod_{T^* \neq T_0, T} F_{T|T_0} \right]^{\frac{1}{n}} = \left[\prod_{T^* \neq T_0, T} \frac{F_{T|T^*}}{F_{T_0|T^*}} \right]^{\frac{1}{n}}.$$

In this formula n is the number of countries which are different from T_0 and T and for which both PPPs of the ratio have been computed previously

(b) A new PPP is computed for each ordered pair of countries T_0 and T , as follows

¹ Point t is the weight reference dimension different from both the current and base point (T_0 or T)

² This means that the same procedure is followed for countries T_0 and T .

$$EKS_{T|T_0} = \left[F_{T|T_0} \cdot \prod_{T^* \neq T_0, T} \frac{F_{T|T^*}}{F_{T_0|T^*}} \right]^{1/(n+2)}$$

In this formula n is the number of countries which are different from T_0 and T and for which both PPPs of the ratio have already been computed

- PPP standardization

The PPPs are standardized and one PPP is calculated per country (implicitly taking the average of all countries as base), as follows

$$PPP_{T_0} = \frac{EKS_{T_0|T}}{\left[\prod_{T^*} EKS_{T^*|T} \right]^{1/m}}$$

In this formula m is the number of countries. The choice of country T^* is not important. Any country would give the same result.

2.2 Chaining

An index can either be fixed-based or chained. Fixed-based indices are calculated directly by producing the comparisons between two different periods, y and y_0 . A chain index is a special index type obtained by linking sub-indices (links) in terms of time (multiplication); those refer to the previous period, which means that they have a weighting pattern that changes every period. In that case, any two-periods comparison between y and y_0 is performed indirectly, that is by multiplying the sub-indices/links. Any index number formula can be used for the individual links in the chain index.

Examples of chained indices are the Unit Value Index and Volume Index calculated by Eurostat in the External Trade domain. The chained index for the current month m and current year y with the fixed base year y_0 as reference for block b (combination of reporting country / partner country / flow / product class) is given by the following formula ${}_{y_0}P_{m,y,b} = {}_{y-1}P_{m,y,b} \cdot {}_{y-2}P_{y-1,b} \cdot \dots \cdot {}_{y_0}P_{y_0+1,b} \cdot 100$.

2.3 Aggregation

Aggregation implies the computation of an index at different periodicity or different breakdown. By aggregation we mean the computation of an index number, at a given 'current point' of the reference dimension, for a level of the other dimensions' nomenclatures, using as input the values and weights of the index number at lower levels of the nomenclatures. Different ways of aggregation are envisaged for weights and index numbers depending on the index type.

The table below summarizes the procedures needed to aggregate an index number, by type of index number.

Table 2. Aggregation of weights and index numbers

Type of Index number	Aggregation of Weights	Aggregation of Index number
Weighted means, Fisher type PPP indices	Summation of lower level weights	Weighted mean of lower level indices
	-	Summation of numerator and denominator
Laspeyres, Lowe	-	Summation of numerator and denominator
Chain index, Index ratio	-	Aggregation of constituent indices and application of chain or ratio formula on the aggregates
All types	Summation of lower level weights	Aggregation of data and application of index number formula at the required level ³

3. Proposed scheme for standardized description of index numbers

The combination of index number formula, use of chaining or not and way of aggregation creates a multitude of options which are at statisticians' disposal. In order to make this selection standardized, a scheme for describing computational parameters (index formula, way of aggregation, base period, etc.) has been developed. The scheme comprises a set of distinct items with a small number of possible values for each item. The scheme described below is free of domain-specific semantics so as to achieve re-usability across domains. The scheme does not cover any pre-processing that may be required, like weights calculation, monetary transformations, etc.

The list of parameters which comprise the scheme are grouped into the following three categories:

Basic Information

1. *Index formula for computation.* Formula selection (or PPP-specific procedure)
2. *Reference dimension.* Reference time or geographical location (PPP only) and the respective unit of time or geographical unit.
3. *Output time format.* [Only relevant for Time reference dimension] Frequency of the index number.
4. *Index reference point.* The base point of the index number to be computed. It is possible that this is at a higher level of aggregation than the data (e.g. in foreign trade where data are monthly the base point is the year). In the case of chained indices, this is the starting point for the computation of the indices. In the case of PPPs base point is not relevant. The computation will take into

³ This mode of aggregation is not used in the domains of study. It is however conceivable that such a requirement may arise in the future.

account all countries (or other geographical domains present) and use their ‘average’ as base point.

5. *Index type*. There are two possibilities: A) Fixed, B) Chained [Only relevant for Time reference dimension]

Variables and weights

6. *Indexed variable*. The variable to be indexed (the values of $v_{T,i}$ if one refers to Table 1).
7. *Indexed variable at index reference point*. [Only relevant for Time reference dimension and where the reference period is defined at a more aggregated level of time than the data]. The data of the variable to be indexed referring to the reference point.
8. *Weights*. Data of the variable serving as weight.
9. *Weights at index reference point*. [Only relevant for Time reference dimension and where the reference period is defined at a more aggregated level of time than the data]. Weights referring to the index reference point.
10. *Weight reference point*. [Only relevant for Time reference dimension and the Lowe index] Base point of the weights if different from base point of the index.
11. *Weights at weight reference point*. [Only relevant for Time reference dimension, the Lowe index and only if the weights’ reference period is defined at a more aggregated level of time than the data]. Weights referring to the weights’ reference period.

Constraints

12. *Current points*. [Only relevant for Time reference dimension – in PPPs all points will serve as current points] All points by level of the reference dimension where the index must be computed.
 - a. *Current points. First*. This is the first point in time (the smallest value of T if one refers to Table 1) at which to compute the index.
 - b. *Current points. Last*. This is the last point in time (the largest value of T if one refers to Table 1) at which to compute the index.
13. *Incremental Aggregation*. [Only relevant for Time reference dimension] Aggregation mode in respect to the time dimension.
14. *Other dimensions*.
 - a. *Order of aggregation*.⁴ Order in which each dimension will be aggregated. Order 1 means the relevant dimension will be aggregated before all other dimensions (the reference dimension is not taken into account).
 - b. *Start level of computation*. [Not applicable to EKS direct Fisher type PPP (below BH)] The most detailed level where the index will be

⁴ The scheme supports a standard way of aggregation: one formula for aggregating numerator and denominator. Otherwise (use sum of component details) pre-processing by user is required

computed [Separate for each of the other dimensions – default: most detailed level of the nomenclature]

- c. *End level of computation*. [Not applicable to EKS direct Fisher type PPP (below BH)] The most aggregated level where the index will be computed [Separate for each of the other dimensions – default: most aggregated level of the nomenclature]

4. Software tool

4.1 Typical workflow

The scheme described in Section 3 has been integrated into the software module developed for calculating index numbers. Index computation requires three entities: numerical data, nomenclatures (metadata) and computation parameters. Pre-processing is however to convert data and metadata into a format recognizable by the application. In summary the steps that comprise index computation are:

1. Collecting and preparing data and metadata
2. Entering data into the application
3. Entering metadata/nomenclatures in the application
4. Associating data with metadata
5. Defining computation parameters
6. Starting the computation process
7. Viewing results
8. Saving results for further analysis.

Since the user is expected to compute indices on similar input datasets using the same parameters over and over again the application also offers the ability to save and load parameter sets. This greatly simplifies step 5 as shown above.

It is also worth noting that a result from one computation can be the input dataset for the next. So after step 8 the user might just well go back to 1 (or 2) in order to continue with the computation process.

The application has also support for PPP indices. As already described in Section 2.1, the structure of the input files (and nomenclatures) for PPP are more strict than other indices and time is no longer a reference dimension for the index. PPP calculation is split in several steps. The application performs one step each time so the user is expected to feed the next step with the result of the previous as already explained.

4.2 Technical description of the tool

The system developed for the needs of index number computation is based on two basic components, the Graphical User Interface (GUI) and the Computational Engine.

The interface of the Computational Engine is based on SOAP since it has been implemented as a Web Service to conform to Eurostat's guidelines for Building Blocks. The index Computation Engine can be called by any application that requires index estimation. The Web Service interface consists of 3 basic parts: the input dataset (CSV format), the related nomenclatures (XML format) and the computational parameters (XML format).

The index Computation Engine is designed in a generic way that is possible to glue any new index estimation with little extra programming. The Engine has adopted framework that consists of two basic artefacts, data formulator and data estimator. These are abstract classes that are extended for satisfying specific estimation needs. E.g. there are concrete formulators for each type of data needed. The formulators are responsible for formulating data needed for the estimation by querying the database. For each elementary estimation and index there are concrete estimators implemented. Depending on the computational parameters the computation engine selects the appropriate formulator and estimator implementations. The types of the formulators and estimators have been based on the analysis of the proposed standardized description of index numbers presented above. This scheme facilitates the extension to new indices in the future. Moreover it utilizes a database for storing temporarily data for processing so as to enable processing of very large datasets.

The GUI of the system offers a user-friendly interface that facilitates the user to make formulation estimation requests to the computational engine. The user interface provides the means to the user to import the input data in CSV format and preview them in the application in excel like sheet. It also supports importing nomenclatures based on two CSV formats that can be extracted from RAMON. Finally it offers a very efficient interface to specify computational parameters for estimations. The parameters can be saved to a local repository so as to be reused in next computations. These can be exported and reused in another installation of the GUI.

Both GUI and Computational Engine have been implemented in Java. Due to this choice the system is portable and can be deployed in almost all platforms since most commonly used operating systems have the Java Runtime Environment preinstalled. The Computational Engine has some limitations. It needs a J2EE application server to be installed and a RDBMS so as to store temporary data for the estimations.

At the time being there isn't a standalone deployment of the application. However it's feasible to be ported in the future, by eliminating the use of the database for the processing and hardwiring the computational logic directly to the GUI, if there is such a need.

5. Index computation - Example

In the example below we demonstrate how the scheme can be applied for the computation of a commonly used index, the Unit Value index from External Trade. Data pre-processing required for the series of weights or/and the variable to be indexed is not in the scope of this section.

We assume that monthly data are available for years 2004 to 2006 and a Laspeyres chained index is to be calculated for the months between February 2004 and

December 2006. Base point will be January 2004. Suppose that data comes in the format: YEAR, MONTH, COUNTRY, PRODUCT, QUANTITY, UNITVALUE and we wish to compute the index for the following combinations:

- Country (Country level) – Product (2-digit level)
- Country (Country level) – Product (Total)
- Country (EU, Non-EU) – Product (Total)

We thus assume that

- Countries have a 3-level nomenclature with individual countries at level 3, country groups (such as EU) at level 2 and ‘Total’ (i.e. all countries) at level 1;
- Products have a 3-level nomenclature with 3-digit codes at level 3, 2-digit codes at level 2 and ‘Total’ (i.e. all products) at level 1;

Based on the proposed scheme, the following parameters need to be specified:

- *Index number formula:* LASPAYRES
- *Reference dimension:* TIME
 - *Level 1:* YEAR
 - *Level 2:* MONTH
- *Output time format:* MONTH
- *Base point:* 200401
- *Index type:* Chained
- *Indexed variable:* UNITVALUE
- *Indexed variable at base point:* Not applicable (Base point is at the same level as the data)
- *Weights:* QUANTITY
- *Weights at reference point:* Not applicable (Base point is at the same level as the data)
- *Reference point of the weights:* Not applicable (applies to Lowe type indices only)
- *Weights at weight reference point:* Not applicable (applies to Lowe type indices only)
- *Current points. First:* 200402
- *Current points. Last:* 200612
- *Incremental Aggregation:* Not-applicable (index computed at monthly level)
- *Other dimensions, Order of aggregation, Start level of computation, End level of computation*
 - COUNTRY, 2, 3 (country-level), 2 (EU and Non EU-level)
 - PRODUCT, 1, 2 (2-digit level), 1 (top level of product Nomenclature)

6. Conclusions and Future work

The primary aim of this exercise, i.e. to compute index numbers in a standard and generic way, has been achieved to a very satisfactory extent. The proposed scheme and the respective software have been tested and applied to a number of indices from different domains that required a different mix of computations. The results have very promising in calculating index numbers in a concrete way through a generic and standard process. The flexibility of the tool to accommodate additional computations adds value to its usability in view of expanding it to more specific computations/operations that are less frequently used.

Work has already been initiated towards that direction, in identifying a number of functionalities that need to be considered for future developments:

- un-chaining of chained indices,
- price-updating of weights,
- re-basing indices to a new index reference period,
- changing the item or country composition of an index
- selection of aggregation procedure
- rounding of computed indices
- computation of derived statistics
- currency conversions
- output of aggregate weights (they are used but are not included in the output on screen or in a file)
- chaining monthly with annual links

The development and integration of the extra functionalities will result in a revised scheme and a new version of the software tool that could be widely usable to a broader variety of users.

References

- Eurostat - OECD, 2006. Methodological manual on purchasing power parities.
- Fisher, I. (1967) *Making of Index Numbers*, Augustus M Kelley Pubs, 3rd Edition.
- Forsyth, F.G., R.F. Fowler (1981). The Theory and Practice of Chain Price Index Numbers, *Journal of the Royal Statistical Society*, Series A, Vol. 144, No. 1, pp. 224–46.
- International Monetary Fund, 2001. Quarterly National Accounts Manual.
- International Labour Organisation, 2004. Consumer Price Index Manual: Theory and Practice.