A do-it-yourself guide in Excel for composite indicator development

MICHAELA SAISANA
michaela.saisana@jrc.ec.europa.eu

Version: October 2012

Aim of this Excel Spreadsheet

This Excel spreadsheet aims to provide a practical guide to the construction of composite indicators, for European Commission officials and experts from EU institutions and Member States, central agencies and authorities, Institutes, NGOs and researchers involved in the development of composite indicators and rankings.

In particular, it contains a set of technical guidelines that can help builders of composite indicators to improve the quality of their outputs. The guide is meant to serve as a tool to people less familiar with programming software.

The "do-it-yourself" guide starts from the premise that the developers of a composite indicator have already conducted a thorough literature review on the topic of interest, namely: definition(s) of the phenomenon, relevant studies, conceptual frameworks, methodological concerns.

The features included in this "do-it-yourself" guide are the following:

- calculating descriptive statistics of the data,
- spotting and treating potentially problematic indicators,
- analyzing the data correlation structure,
- estimating missing data,
- normalizing indicators (z-scores, min-max, ranks),
- aggregating indicators using (weighted) arithmetic averages, geometric averages, trimmed mean, median rank, summation of ranks, Borda rule, Copeland rule.

This guide was developed by Michaela Saisana (JRC, Italy) in the context of the JRC Annual Seminar on Composite Indicators. This is the October 2012 version. All presentations accompanying this practical guide are available at: http://composite-indicators.jrc.ec.europa.eu

Suggested citation:

How the Excel spreadsheets are organised

"1-Data"

Things to do:

1) Organise the data in countries (rows) x indicators (columns), grouping the indicators according to the conceptual structure (here: 26 countries, 27 indicators, 4 dimensions). Report desired direction for the indicators (here: good = 1, bad = -1)

2) Calculate descriptive statistics, such as missing values, min, max, mean, std, skewness, kurtosis for each indicator.

Practical rules:

a) Require at least 60-65 percent data coverage per country and dimension (requirement can be relaxed or be more conservative depending on the degree of correlation between indicators within a dimension).

b) Treat potentially problematic indicators (i.e. those with skewness >2 AND kurtosis > 3.5) either by winsorisation (if few outlier values, roughly 5 percent of countries) or by a box-cox transformation (e.g. log) (here only B4 seems to be problematic and that is due to a single value 34.16 for country W; simple solution to winsorise i.e. assign the next best value to country W that is 17.35, see next sheets).

3) Calculate correlations between indicators. However, if the indicators do not have the same direction, it will take more time to reflect on the signs of correlations; in that case better look at correlations after having normalised (so that desired direction has been taken into account). (here: A1 and A3 are expected to have a negative association as their directions are "opposite")

Practical rules:

Highly collinear indicators (r > 0.92 roughly) within a given dimension need to be treated (either by eliminating one of the two, or counting them as a single indicator) otherwise they will influence the results of principal component analysis and will dominate the country scores in the respective dimension.

4) Estimate missing data by regression: Use relatively high correlations (roughly r > 0.80) between indicators to estimate missing data. (here: high correlation between indicators C6 and D8, hence we can estimate missing data in C6 by using the linear relation (or second order, etc) between C6 and D8). Note that this type of imputation does not change significantly the degree of correlation (before: 0.87, after imputation: 0.91)

"2-Data-hotdeck"

Things to do:

5) In case correlations between indicators are not high enough, you can use hot-deck imputation (here: missing values were estimated with one nearest neighbour and Manhattan distance; example: the nearest neighbour to country B is country H, hence all missing values for B are replaced by those of H). Use cross-validation to identify the most suitable number of nearest neighbours and the most suitable distance measure. Remember to normalise data prior to calculating distances between countries. Note that this type of imputation may have an impact on the degree of correlation between indicators (example here C6 and D8 before: 0.87, after imputation: 0.71).
Remarks: Note that after having "corrected" for the indicator value B4 for country W (new value: 17.35, old value: 34.16), the indicator B4 is no longer flagged as problematic (skewness = .79, kurtosis = -.83)

"3-Dataminmax"

Things to do:

6) Normalise indicators. Most commonly used methods are: min-max or z-scores (here: min-max). Remember to take into account the direction of the indicators.

General formula:

newvalue=(oldvalue-min)/(max-min)*direction+0.5*(1-direction)

7) Calculate correlations between indicators. At this point it will be easier (than in point 4), as all correlations are expected to be positive. If there are negative correlations, it means that either the desired direction of the indicator is wrong or that there are trade-offs between indicators (nothing to do about it, it is a feature of the phenomenon you are trying to measure). Generally desirable: not to have negative correlations at least within the same dimension.

(Here only in B dimension there are negative correlations, but given the small sample size they may be spurious, hence no further action on them).

Practical rules:

a) Highly collinear indicators (r > 0.92 roughly) within a given dimension need to be treated (either by eliminating one of the two, or counting them as a single indicator) otherwise they will influence all principal component analysis and will dominate the country scores in the respective dimension. (here highly collinear indicators were assigned half weight, see B4 and B5).
b) You may decide to eliminate indicators that are randomly associated to any of the remaining indicators in the dimension (e.g. C5 gets a weight of 0).

Remarks: Note that after having "corrected" for the indicator value B4 for country W (new value: 17.35, old value: 34.16), both country W and H get a maximum normalised score of 1.

"4-Dataz"

Things to do:

8) Normalise indicators. Most commonly used methods are: min-max or z-scores (here: z-scores). Remember to take into account the direction of the indicators.

General formula: new value = (old value-mean)/std * direction

9) Calculate correlations between indicators. At this point it will be easier (than in point 4), as all correlations are expected to be positive. If there are negative correlations, it means that either the desired direction of the indicator is wrong or that there are trade-offs between indicators (nothing to do about it, it is a feature of the phenomenon you are trying to measure). Generally desirable: not to have negative correlations at least within the same dimension.

(Here only in B dimension there are negative correlations, but given the small sample size they may be spurious, hence no further action on them).

Practical rules:

a) Highly collinear indicators (r > 0.92 roughly) within a given dimension need to be treated (either by eliminating one of the two, or counting them as a single indicator) otherwise they will influence all principal component analysis and will dominate the country scores in the respective dimension. (here highly collinear indicators were assigned half weight, see B4 and B5).

b) You may decide to eliminate indicators that are randomly associated to any of the remaining indicators in the dimension (e.g. C5 gets a weight of 0).

Remarks: Note that after having "corrected" for the indicator value B4 for country W (new value: 17.35, old value: 34.16), both country W and H get a maximum normalised score of 1.75.

Note that the correlations are the same as those with min-max normalisation (or those with raw data, when all data were available, otherwise depending on the imputation method the degree of correlations might be slightly affected).
"5-Dataranks"

Things to do:

10) Calculate ranks on individual indicators. They will help in the interpretation of results when trying to argue why one country is doing better than another within a dimension.

11) Correlations between ranks are equivalent to the Spearman rank correlation coefficient (correlation coefficients calculated in the previous sheets were Pearson coefficients).

**Remark**: Note that all correlations in dimension B are positive (the negative ones are not statistically significant. Hence our prior decision not to take any action on them is justified at this point).

"6-PCA"

Things to do:

12) Apply Principal Components Analysis within each dimension (use other software such as SPSS, Statistica, or the XLSTAT for Excel, or other). Identify how many latent dimensions exist within each conceptual dimension (e.g. components with eigenvalue > .9). Select only those components and then rotate (e.g. varimax rotation). After rotation, hopefully each indicator is loaded mostly on a single component, and all loadings in the same component have the same sign.

13) Ideally, if the sample size is adequate, one should also apply Factor Analysis across all indicators in order to assess whether the conceptual structure is statistically confirmed.

"7-DEA"

How to work with Data Envelopment Analysis (DEA)

DEA works with non-negative values. Use another program (e.g. R or Matlab) to calculate the weights of DEA. (here: only C5 was
excluded, decision due to low correlation with other indicators in dimension C).

Multiply the weights by the respective normalised indicator scores and for the classical DEA consider only the values in the main diagonal. These scores can only be used when "efficiency" comparisons need to be made but NO ranking.

If you want to rank countries, better use the cross-efficiency version of DEA and preferably the average cross-efficiency as even the median version can result in many ties between countries.

"8-Aggregation Methods"

How to apply different normalisation, sets of weights, aggregation methods (part 1/2)

Description of the case study:

For illustration purposes, an example based on 20 countries and 8 indicators is used. There are no missing data and none of the indicators is highly skewed. The correlation matrix is calculated, showing that indicators #5 and #6 are negatively associated to the other six indicators. The desired direction of the indicators is given. In fact, the desired direction of indicators #5 and #6 is the opposite (-1) namely the lower the better, whilst for all other indicators the desired direction is the higher the better (1).

Notation: N is the total number of countries (here N=20), M is the total number of indicators (here M=8).

What calculations are done here:

Normalisation

1) Raw data are normalised using z-scores with a mean=0 and a standard deviation=1. The direction of the indicators is taken into account during the normalisation. The value 5 is added to all scores, so as to allow for the use of geometric averaging that necessitates strictly positive values. Hence, the normalised data have mean=5 and standard deviation=1. In general, if you want the data to have a mean=MEANdesired and SD=SDdesired, the formula is:

Formula:  
\[ \text{newvalue} = \frac{\text{oldvalue} - \text{mean}}{\text{std}} \times \text{direction} \times \text{SDdesired} + \text{MEANdesired} \]

2) Raw data are normalised using min-max within the 0-1 range. The direction of the indicators is taken into account during the normalisation. The minimum value is set at 0.1, so as to allow for the use of geometric averaging that necessitates strictly positive values. Hence, the normalised data range between 0.1-1.0. In general, if you want the data to have min=MINdesired and max=MAXdesired, then

Formula:  
\[ \text{newvalue} = \frac{\text{oldvalue} - \text{MIN}}{\text{MAX} - \text{MIN}} \times \text{direction} + 0.5 \times (1 - \text{direction}) \times (\text{MAXdesired} - \text{MINdesired}) + \text{MINdesired} \]

3) Raw data are normalised using ranks. The direction of the indicators is taken into account during the normalisation.

Formula:  
\[ \text{newvalue} = \text{RANK(olvalue, range of values ranked)} + 0.5 \times (1 - \text{direction}) \]
4) Raw data are normalised using Borda points. The direction of the indicators is taken into account during the normalisation. For \( N \) countries in a sample, the top ranked country gets \( N-1 \) points, the second ranked country gets \( N-2 \) points and so on. The last ranked country gets 0 points.

**Formula:** newvalue = N - country rank

[remark: use the calculation in (3) for the country rank]

**Weights**

5) The weights are randomly selected using a uniform distribution in 0-1 (X1:AE1). The weights are then re-scaled to unity sum (X2:AE2). By pressing F9, the weights are automatically changed, and so are the calculations of the country scores in the overall aggregate (see aggregation below). In general, if you want the weights to have a certain range, \( \text{MINdesired, MAXdesired} \), the formula is:

**Formula:** weight = randbetween(MINdesired, MAXdesired)

[remark: the MINdesired and MAXdesired have to be greater than 1 in this Excel formula. Hence for a range of weights in 0.2 to 0.5 simply write: weight = randbetween(2, 5)/10]

**Aggregation**

The aggregation functions used here are: Arithmetic mean (equal weights, unequal weights), Geometric mean (equal weights, unequal weights), trimmed mean (equal weights), summation of ranks, median rank, trimmed median rank, Borda (equal weights, unequal weights).

6) Normalised data (using either z-scores or min-max) are aggregated using arithmetic averaging.

**Formula- equal weights:** score = average(normalised values)

**Formula-unequal weights:** score = sumproduct(weights*normalised values)

7) Normalised data (using either z-scores or min-max) are aggregated using geometric averaging.

**Formula-equal weights:** score = product(normalised values)^{(1/M)}

**Formula-unequal weights:**

score = EXP(SUMPRODUCT(weights, LN(normalised values)))

8) Normalised data (using either z-scores or min-max) are aggregated using trimmed mean. In this example, the best and worst values of the eight indicators are discarded.

**Formula- equal weights:** score = (SUM(normalised values) - LARGE(normalised values,1) - LARGE(normalised values,8))/(M-2),

Besides country scores in the aggregations (6) to (8) above, also the corresponding country ranks are calculated.

9) Country ranks for each indicator are aggregated using median rank. **Formula-equal weights:** rank = MEDIAN(ranks for the same country across all indicators)

10) Country ranks for each indicator are aggregated using Borda rule.

**Formula-equal weights:** rank = N+1-RANK(sum(ranks for the same country across all indicators))
Formula-unequal weights:

\[
\text{rank} = N+1-\text{RANK}(\text{sumproduct}(\text{weights} \times \text{ranks for the same country across all indicators}))
\]

The country scores and ranks in the light blue columns are updated every time the weights are calculated (by pressing F9).

"9-Copeland rule"

How to apply different normalisation, sets of weights, aggregation methods (part 2/2)

Description of the case study:

Same raw data as those in the sheet "8-Aggregation Methods".

The direction of the indicators and the weights are updated automatically from the sheet "8-Aggregation Methods".

What calculations are done here:

This sheet calculates the outranking matrix that is needed as input information in the Copeland rule. You need to copy the raw data in N11:U30 and also in transposed form in W2:AP9.

Outranking matrix:

Countries are compared pairwise. For each comparison, all the weights corresponding to the indicators that favour country A versus B are added up as evidence in favour of “A better than B” (abbreviated as AB). For N countries, there are N*(N-1) comparisons to be made. The diagonal elements are set at 0 by definition. In practical terms, for each pairwise country comparison the following formula is used:

Formula: Summation((weight for indicator \(i\)) *(1+direction of indicator \(i\)) *SIGN(raw value of country A on indicator \(i\) - raw value of country B on indicator \(i\)))/2, where the summation is done across all indicators.
All the values from the pairwise comparisons are entered in the so-called outranking matrix. The outranking matrix is calculated here automatically (see cells W11:AP30). Note that above/below diagonal entries add up to one (since the sum of weights is one).

**Copeland rule:** Score +1 if a majority of indicators prefers country A to country B, score -1 if a majority of indicators prefers B to A, and 0 if it is a tie. Hence in the outranking matrix, replace all values greater than 0.5 with +1, all values lower than 0.5 with -1 and all ties (=0.50) with 0. The diagonal elements are set at 0 by definition. The Copeland score for each country is the sum of the values in a given row. A final ranking is then calculated.

**Remark:** In general, some compensability/substitutability is desired at the lower aggregation level. However, at a higher aggregation level (that is dimensions/objectives/pillars), compensability is less desirable. To this end, one may use one of the other aggregation methods calculated in sheet "8-Aggregation Methods" at the lower level of aggregation and then use the Copeland rule to aggregate the dimensions.

"10-Scenaria"

**Scenario in developing a composite indicator**

**Description of the case study:**

Same as in sheets "8-Aggregation Methods" and "9-Copeland rule".

**What calculations are done here:**

This sheet simply gathers all the composite indicator rankings calculated in sheets "8-Aggregation Methods" and "9-Copeland rule". Overall, 15 rankings are produced combining different normalisation methods and different aggregation methods. The rankings presented in blue are fixed as they are based on equal weights among all indicators. Instead, the rankings presented in
light blue are updated every time the weights are updated. Simply press F9 to update the weights (ranks are also updated).

A median rank across all the fifteen scenarios together with the interval (minimum and maximum rank) is also calculated for each country.

ENJOY!