Data Analysis and Visualization

CA1 Specification Index Generation and Visualization

BSc (Honours) in Computing in Software Development

Individual Project

Weighting: 50%

Due: End of Week 12, week 28-4-24

DAV Objectives

To practice the following:

- Finding data to support the index
- Selecting variables
- Multivariate analysis
- Cluster analysis
- Normalisation of data
- Weighting and aggregation
- Visualisation of the individual and composite indicators
- Use version control to track development of the project

Learning outcomes addressed

- Compare and apply common regression techniques in multivariate analysis
- Clustering
- Visualization

Project Brief

For this project you are required to develop a composite indicator which is a single number that can be used to compare countries, people or cities.

An indicator is a quantitative or qualitative measure derived from a series of observed facts that can reveal relative positions (e.g. of a country) in a given area. When evaluated at regular intervals, an indicator can point out the direction of change across different units and through time.

A composite indicator is formed when individual indicators are compiled into a single index on the basis of an underlying model. The composite indicator should ideally measure multidimensional concepts which cannot be captured with a single indicator.

(Handbook on Constructing Composite Indicators: Methodology and UserGuide).

Table 1. Checklist for building a composite indicator

Step	Why it is needed
1. Theoretical framework Provides the basis for the selection and combination of variables into a meaningful composite indicator under a fitness-for-purpose principle (involvement of experts and stakeholders is envisaged at this step).	• To get a clear understanding and definition of the multidimensional phenomenon to be measured. • To structure the various sub-groups of the phenomenon (if needed). • To compile a list of selection criteria for the underlying variables, *e.g.*, input, output, process.
2. Data selection Should be based on the analytical soundness, measurability, country coverage, and relevance of the indicators to the phenomenon being measured and relationship to each other. The use of proxy variables should be considered when data are scarce (involvement of experts and stakeholders is envisaged at this step).	• To check the quality of the available indicators. • To discuss the strengths and weaknesses of each selected indicator. • To create a summary table on data characteristics, *e.g.*, availability (across country, time), source, type (hard, soft or input, output, process).
3. Imputation of missing data Is needed in order to provide a complete dataset (*e.g.* by means of single or multiple imputation).	• To estimate missing values. • To provide a measure of the reliability of each imputed value, so as to assess the impact of the imputation on the composite indicator results. • To discuss the presence of outliers in the dataset.
4. Multivariate analysis Should be used to study the overall structure of the dataset, assess its suitability, and guide subsequent methodological choices (*e.g.*, weighting, aggregation).	• To check the underlying structure of the data along the two main dimensions, namely individual indicators and countries (by means of suitable multivariate methods, *e.g.*, principal components analysis, cluster analysis). • To identify groups of indicators or groups of countries that are statistically "similar" and provide an interpretation of the results. • To compare the statistically-determined structure of the data set to the theoretical framework and discuss possible differences.
5. Normalisation Should be carried out to render the variables comparable.	• To select suitable normalisation procedure(s) that respect both the theoretical framework and the data properties. • To discuss the presence of outliers in the dataset as they may become unintended benchmarks. • To make scale adjustments, if necessary. • To transform highly skewed indicators, if necessary.

Step	Why it is needed
6. Weighting and aggregation Should be done along the lines of the underlying theoretical framework.	• To select appropriate weighting and aggregation procedure(s) that respect both the theoretical framework and the data properties. • To discuss whether correlation issues among indicators should be accounted for. • To discuss whether compensability among indicators should be allowed.
7. Uncertainty and sensitivity analysis Should be undertaken to assess the robustness of the composite indicator in terms of e.g., the mechanism for including or excluding an indicator, the normalisation scheme, the imputation of missing data, the choice of weights, the aggregation method.	• To consider a multi-modelling approach to build the composite indicator, and if available, alternative conceptual scenarios for the selection of the underlying indicators. • To identify all possible sources of uncertainty in the development of the composite indicator and accompany the composite scores and ranks with uncertainty bounds. • To conduct sensitivity analysis of the inference (assumptions) and determine what sources of uncertainty are more influential in the scores and/or ranks.
8. Back to the data Is needed to reveal the main drivers for an overall good or bad performance. Transparency is primordial to good analysis and policymaking.	• To profile country performance at the indicator level so as to reveal what is driving the composite indicator results. • To check for correlation and causality (if possible). • to identify if the composite indicator results are overly dominated by few indicators and to explain the relative importance of the sub-components of the composite indicator.
9. Links to other indicators Should be made to correlate the composite indicator (or its dimensions) with existing (simple or composite) indicators as well as to identify linkages through regressions.	• To correlate the composite indicator with other relevant measures, taking into consideration the results of sensitivity analysis. • To develop data-driven narratives based on the results.
10. Visualisation of the results Should receive proper attention, given that the visualisation can influence (or help to enhance) interpretability	• To identify a coherent set of presentational tools for the targeted audience. • To select the visualisation technique which communicates the most information. • To present the composite indicator results in a clear and accurate manner.

Above is a checklist for developing a composite indicator taken from Handbook on Constructing Composite Indicators: Methodology and UserGuide. You are not required to carry out steps 7 and 8 but should consider all other steps.

Step 1 will also be problematic for you as you probably won't have access to domain experts. However, you should research the index you are creating and justify your choice of variables and data sources.

Some examples of existing indices are:

Siemens Green City Index - https://assets.new.siemens.com/siemens/assets/api/uuid:fddc99e7-5907-49aa-92c4-610c0801659e/european-green-city-index.pdf

UN Human Development Index: http://hdr.undp.org/en/statistics/hdi/

Movies: https://mdblist.com/

Footballers: https://football-observatory.com/IMG/sites/instatindex/

Theoretical Framework

Set the scene for your composite index. Justify why you have chosen this problem and why the data you are using to solve it is appropriate. Normally this would be done with expert opinion. If this is not available to you, you can research using Google or survey classmates.

The index would normally be made up of sub-indices which are combined together create the final index.

(10 Marks)

Data Selection

You need to get the data that you will use for the index. This could be done by using publicly available data sets or survey.

(10 Marks)

Imputation of Missing Data

If your data is not complete you will need to infer values to complete the dataset.

(10 Marks)

Multivariate Analysis

Analyse and report on the structure of the data. Decide what are the most important variables and what should be excluded from the index.

(10 Marks)

Normalisation

In order to compare variables you will need to normalise them. Choose an appropriate normalisation.

(10 Marks)

Weighting and Aggregation

Combine indicators into sub-indicators and a final composite index.

(10 Marks)

Link to other Indicators

Try to find existing indicators similar to yours and compare the results.

(10 Marks)

Visualisation of Results

Draw graphs to visualise the composite index and sub-indices.

(10 Marks)

Version Control

All work should be committed to version control.. You should follow the rules set out here https://chris.beams.io/posts/git-commit/ when writing commit messages.

(10 Marks)

Deliverable

You should write a document detailing the development of the index including outputs from the visualisations that you have produced.

(10 Marks)

Submission Requirements

1. Please submit your documentation, any supporting code or demos and a code repo url in one.zip through Moodle.
2. Each member of the team has to make a clear contribution to the CA and have a well-defined role. You must use version control when writing the documentation and each member of the team should make a clear contribution to the documentation. This will be assessed by your commits and interviews if necessary.
3. **Plagiarised assignments will have their mark withheld**. The HOD will be notified and you will go through DKIT's process for suspected plagiarism. This also applies to the individual allowing their work to be plagiarised.

4. Any plagiarism will be reported to the Head of Department and a report will be added to your permanent academic record.
5. Late assignments will only be accepted if accompanied by the appropriate medical note. This documentation must be received within 10 working days of the project deadline. The penalty for late submission is as follows:
   - Marked out of 80% if up to 24 hours late.
   - Marked out of 60% if 24-48 hours late.
   - Marked out of 40% if 48-72 hours late.
   - Marked out of 20% if 72-96 hours late.
   - Marked out of 0%, if over 96 hours late.

Generative AI

Generative artificial intelligence (AI) tools are not restricted for this assessment task. In this assessment, you can use generative artificial intelligence (AI) to assist you in any way. Any use of generative AI must be appropriately acknowledged (in accordance with DkIT Academic Integrity Policy and Procedures, https://www.dkit.ie/about-dkit/policies-andguidelines/academic-policies.html).