PATHWAY 2: DEVELOPING COMPOSITE INDICATORS

Christopher Gandrud

SG1022, City University London

AIMS

- What are composite indicators and who uses them?
- Making Composite Indicators
- Pros and cons

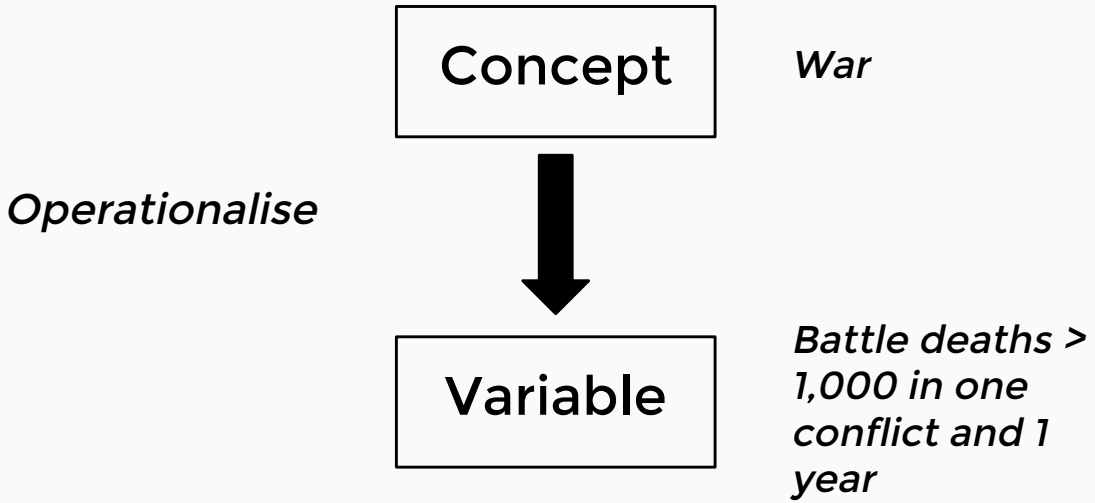
WHAT ARE COMPOSITE INDICATORS

VARIABLES

Last semester, we learned about concepts and variables.

- *Concept*: A phenomenon (e.g. poverty, democracy, human development, trust in institutions) we are interested in studying.

- *Variable*: Observable characteristic of a unit (e.g. person, city, country) that operationalises the concept.


Concept
War
Operationalise
Variable
Battle deaths > 1,000 in one conflict and 1 year

WHEN ONE VARIABLE IS NOT ENOUGH

Frequently in social science we are interested in complex concepts that cannot be operationalised with one variable.

Instead, they likely involve some combination of variables.

WHAT IS IT?

For example, what is democracy?

WHAT IS DEMOCRACY?

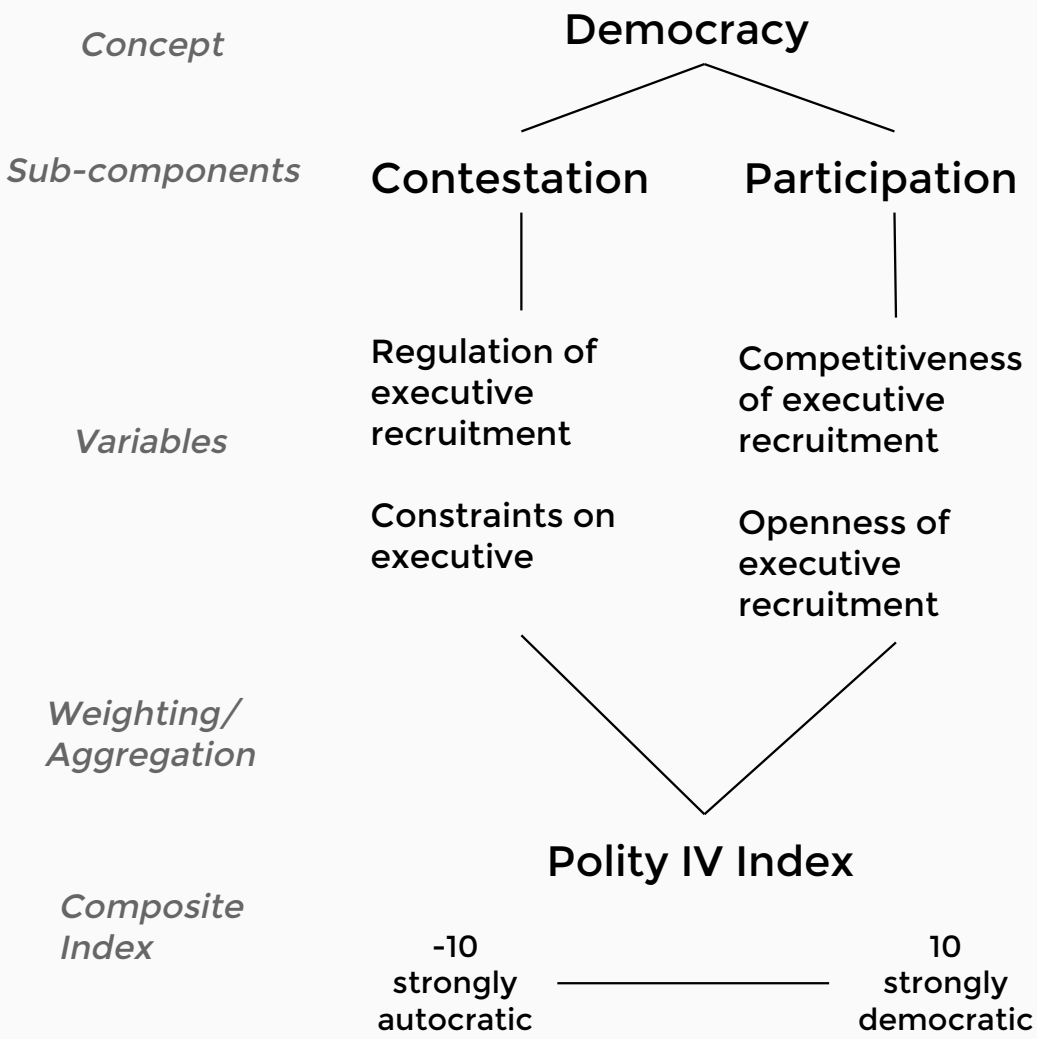
Conceptualisation: Dahl (1971) argued that democracy has two core attributes:

- *contestation*: competitive elections to choose leaders,
- *participation*: inclusive rules for and rates of participation.

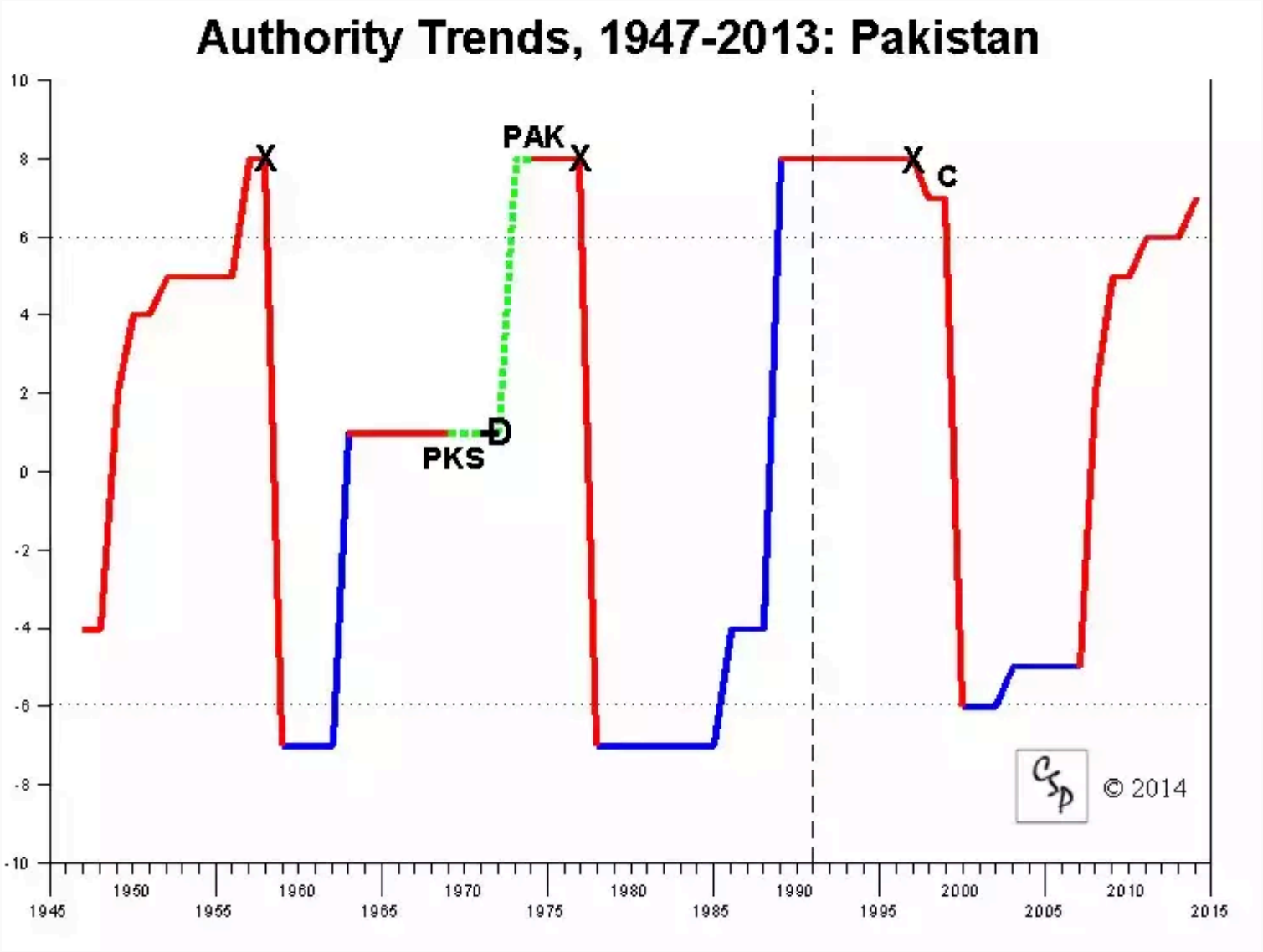
Note: not just elections/no elections or universal suffrage/limited suffrage. So ...

...we need a composite indicator to operationalise Dahl's democracy.

CREATING THE POLITY IV DEMOCRACY INDEX



PAKISTAN POLITY IV SCORE



WHY CARE?

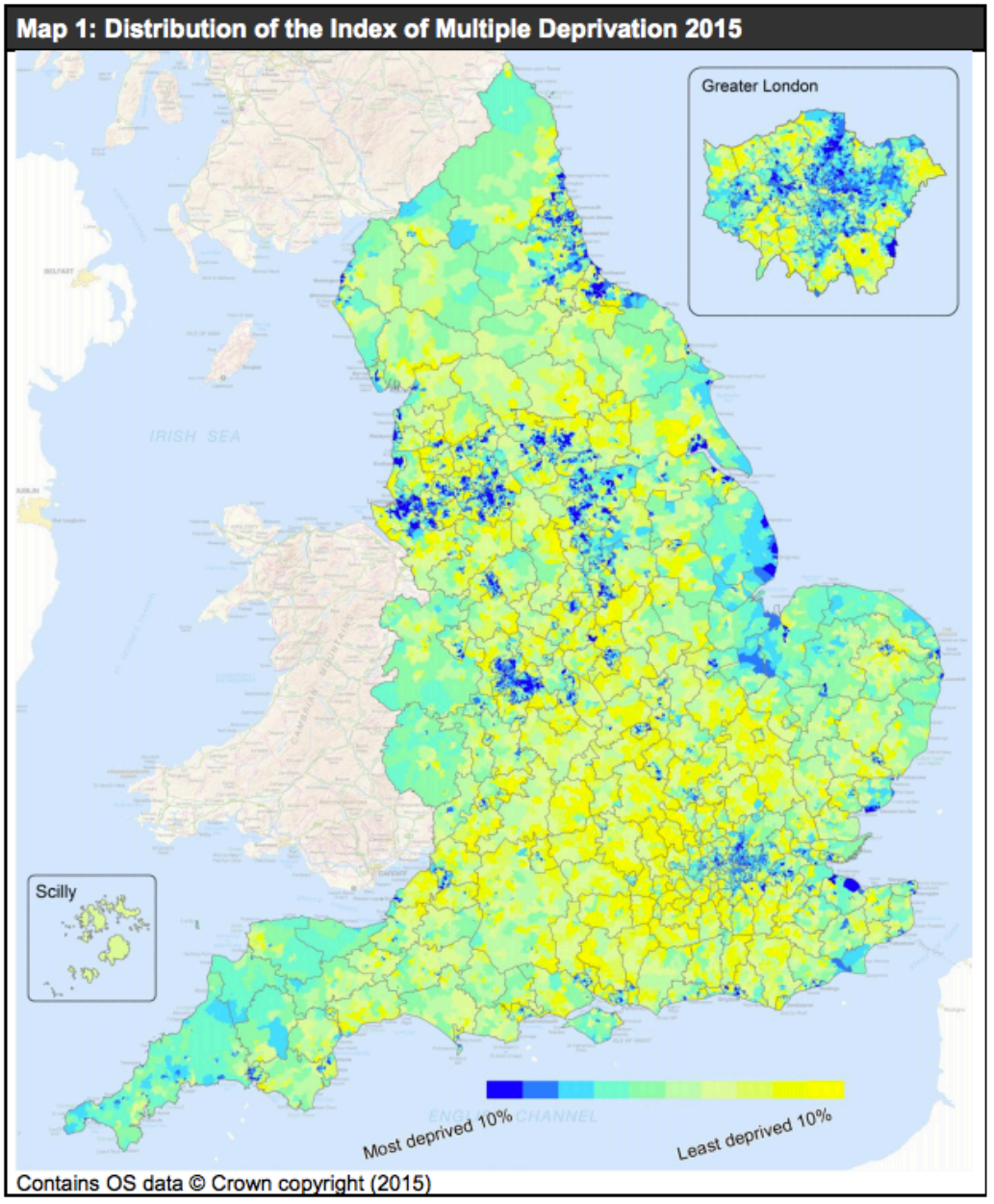
Why should we care about measuring concepts well?

WHY CARE?

If we don't have good measures of our concepts, we can't know how one thing effects another (relationships *between* concepts), how characteristics change over time, and how to improve the social world.

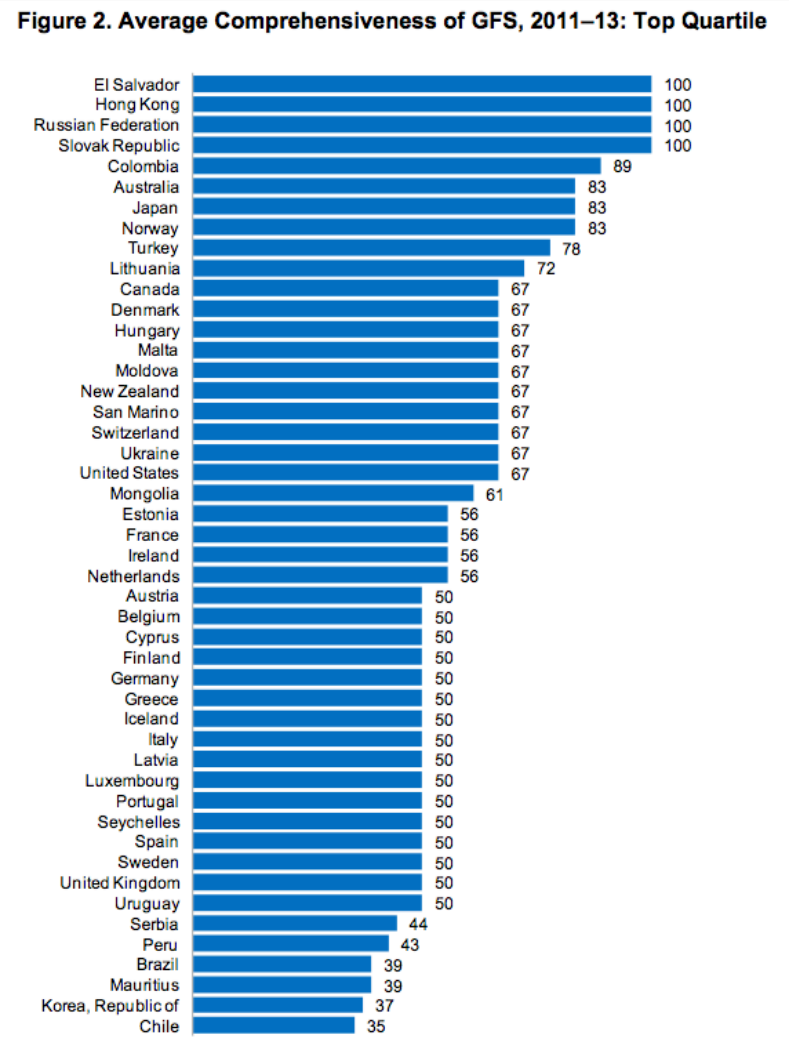
Composite indicators and policy-making

Composite indicators are a popular tool used by governments, international organisations, NGOs, and public policy advocacy groups to both understand the world and advocate change.



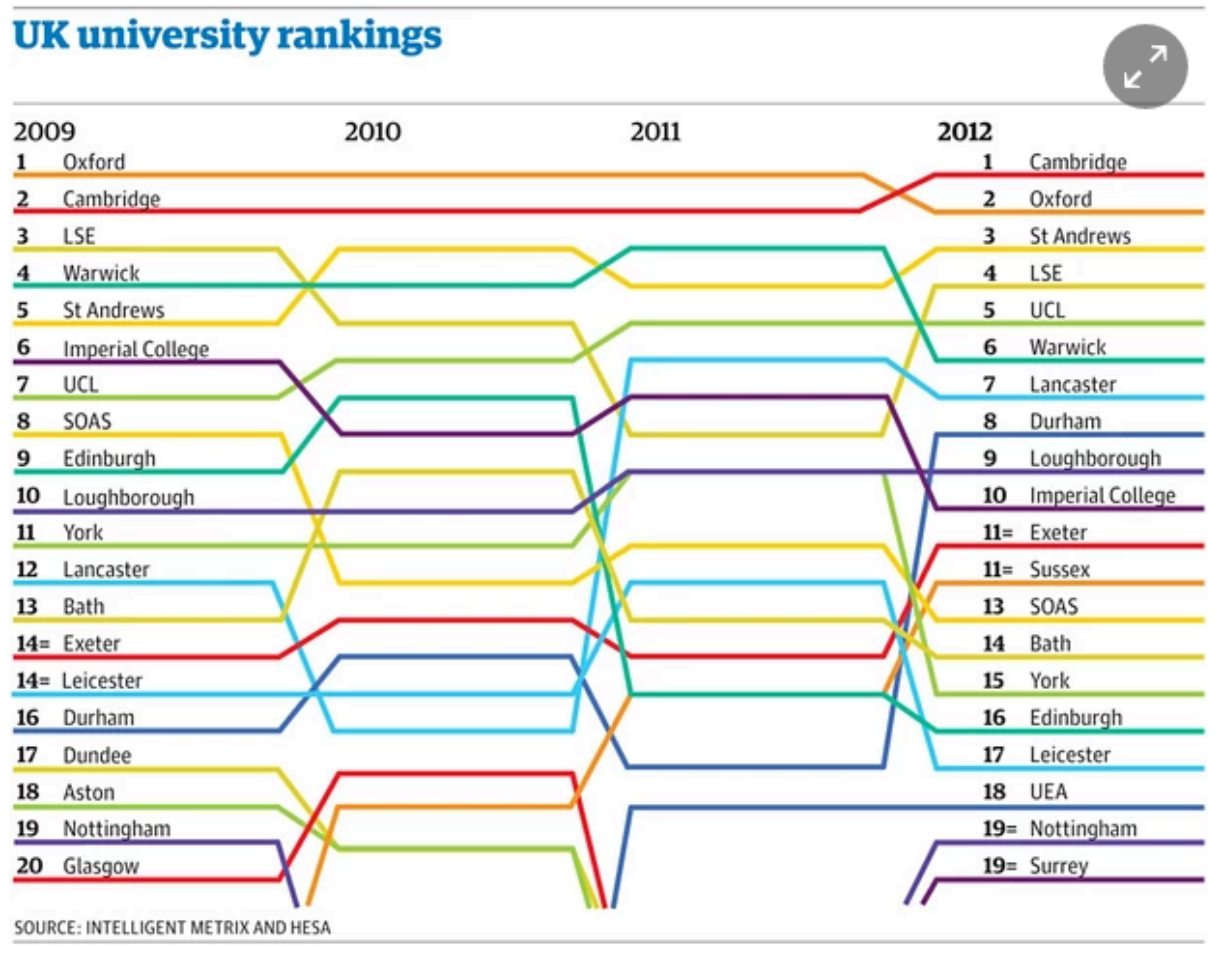
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/465791/English_Indices_of_Deprivation_2015_—_Statistical_Release.pdf

IMF FISCAL TRANSPARENCY (GFS) INDEX



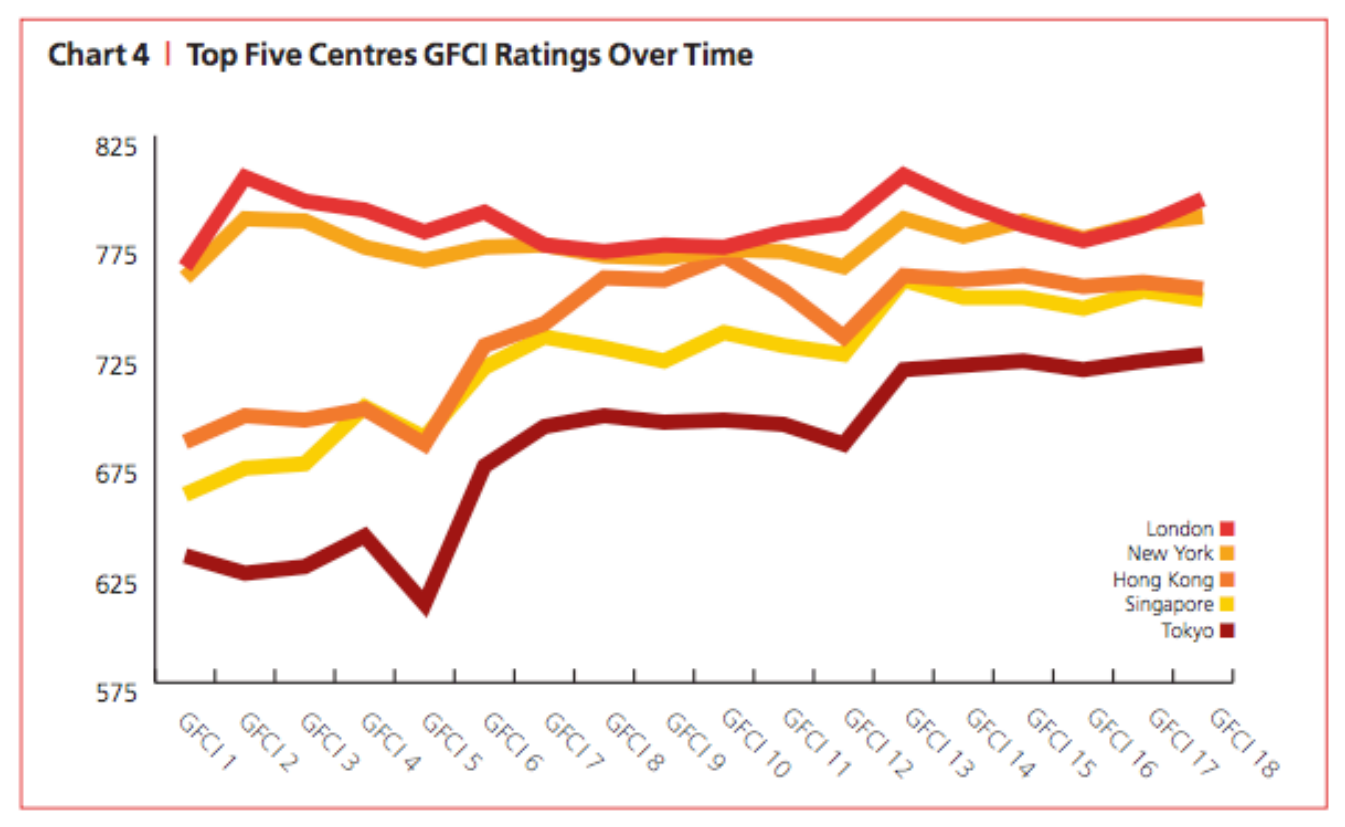
Wang et al. (2015, 10)

UNIVERSITY RANKINGS



(http://www.theguardian.com/news/datablog/2011/may/17/university-guide-2012-data-guardian)

GLOBAL FINANCIAL CENTRES INDEX (2015)



Z/Yen Group (2015, 8)

IMPACTS OUR UNDERSTANDING OF THE WORLD AND POLICYMAKING

New York and London vie for crown of world's top financial centre

Financial centres face competition from Asia and increased regulation

Financial Times (1 Oct 2014)

STEPS FOR CREATING COMPOSITE INDICATORS

STEPS FOR BUILDING COMPOSITE INDICATORS

1. Theoretical framework
2. Data selection
3. Address missing data
4. Multivariate analysis (note: we don't really much this in this course)
5. Normalisation
6. Weighting and aggregation
7. Validation

Modified from OECD (2008, 20-21)

All of these steps should be clearly documented in detail.

THEORETICAL FRAMEWORK

"What is badly defined is likely to be badly measured." (OECD 2008, 22)

Concept definitions should:

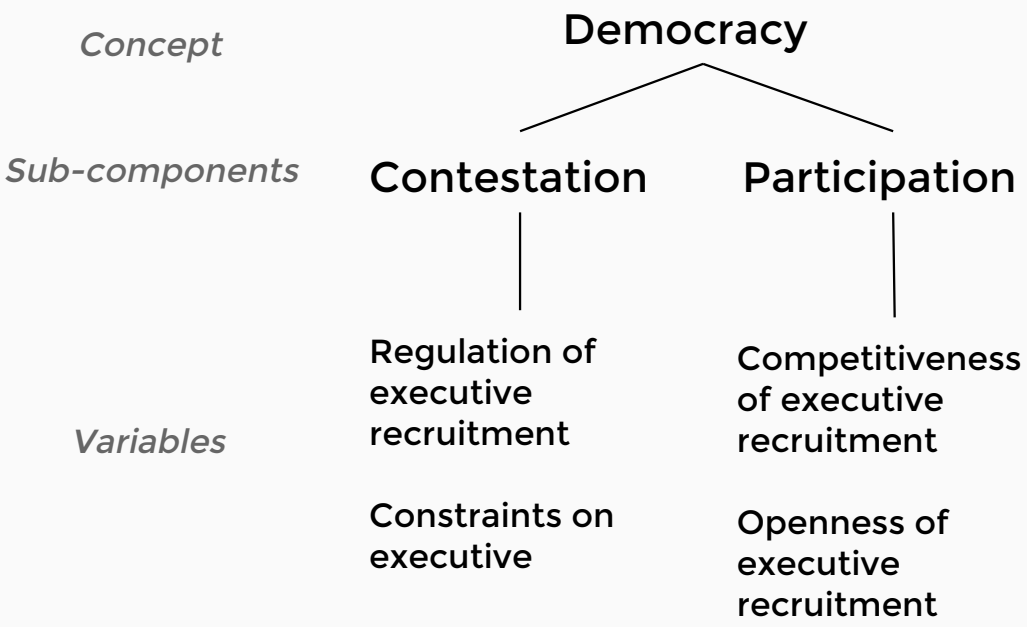
- be clear and parsimonious,
- refer to a theoretical framework,
- select a relevant unit of analysis,
- link various sub-components and the underlying indicator.

Sub-components:

- should be *statistically independent* of each other (want as few components as possible, don't want to double up. Though can compensate in the weighting stage),

- existing *linkages* between them should be *described theoretically and empirically* as much as possible.

CONCEPTUALISING DEMOCRACY



DATA SELECTION

DATA SELECTION

A composite indicator is only as good as its parts.

Data on the indicators' component should be selected based on at least their:

- relevance to the theoretical framework,
- analytical soundness,
- timeliness,
- accessibility.

GATHERING, CLEANING, MERGING

Gathering, cleaning, and merging the component variables is not trivial.

Has important substantive implications (e.g. error can bias your indicators) and can be time consuming.

So, needs to be well documented and reproducible.

MISSING DATA

MISSING VALUES

To aggregate your components into a composite indicator you need a complete data set, with no missing values.

For example, imagine we are going to make a component indicator with variable_1, variable_2, and variable_3. A complete data set would look like:

country	variable_1	variable_2	variable_3
Algeria	1	10034	0
Cambodia	0	30020	0
Zambia	0	50302	1

MISSING VALUES

However, we often (especially when working with country-level data) have missing data on at least some of our variables.

For example:

country	variable_1	variable_2	variable_3
Algeria	NA	10034	0
Cambodia	0	30020	0
Zambia	0	NA	1

Note: NA in R means missing value.

MISSING VALUES

You should always assess how much missing data you have and consider why the data is missing.

WHY MISSING

Data can be:

- *Missing at random*: why the variable value is missing has nothing to do with the variable–random.

- *Not missing at random*: the variable value is missing because of some reason related to the variable.
  - E.g.: Low income countries don't have the money to pay for enough staff to gather GDP data.

Not missing at random is complex to deal with and requires explicit modeling of the missing process.

We do not cover this modeling process in this course.

However, if you suspect your data is not missing at random, you should discuss this, including the reasons why you believe the data is not missing at random.

ADDRESSING MISSING DATA

If you have missing data, to get complete cases you can:

- drop incomplete cases,
- single impute data (e.g. replacing the missing value with the variable's mean/median/mode),
- multiple impute data (not covered in this course).

ANALYSIS OF THE SUB-COMPONENT STRUCTURE

UNDERLYING STRUCTURE

"Analysing the underlying structure of the data is still an art."

(OECD 2008, 25)

Understanding the underlying structure is important for informing weighting and aggregation decisions, whether or not to include a variable.

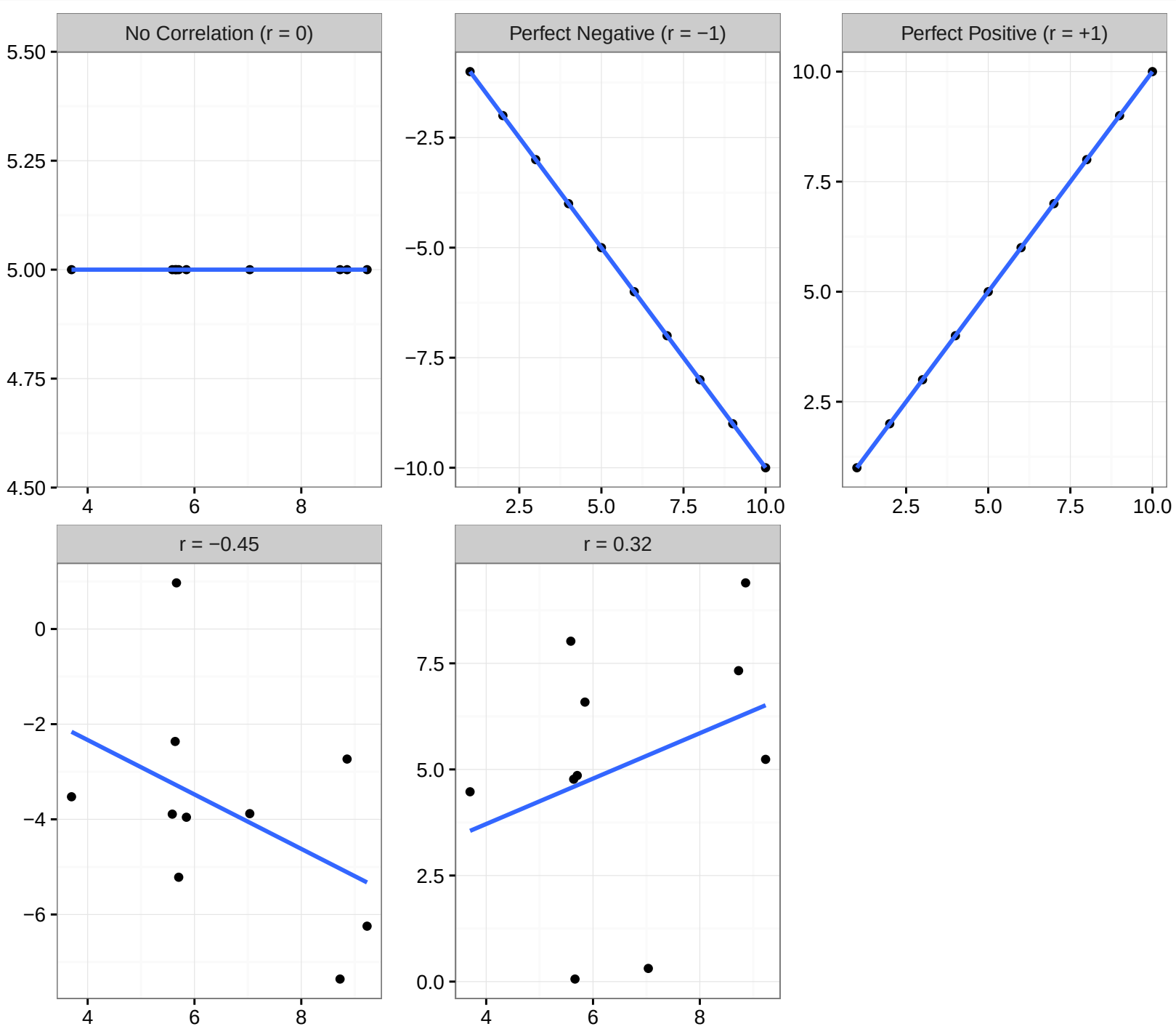
Note: important multivariate analysis tools–e.g. principal component analysis, cluster analysis–are beyond the scope of this class.

SIMPLE CORRELATION ANALYSIS

A simple way to examine the relationships between your sub-components is to create a correlation matrix.

Correlation: a mutual relationship between two variables.

Correlation coefficient: a number ranging from -1 to 1 calculated to represent the linear correlation between two variables. Usually denoted *r*.


No Correlation (r = 0)
Perfect Negative (r = −1)
Perfect Positive (r = +1)
r = −0.45
r = 0.32

SIMPLE CORRELATION ANALYSIS

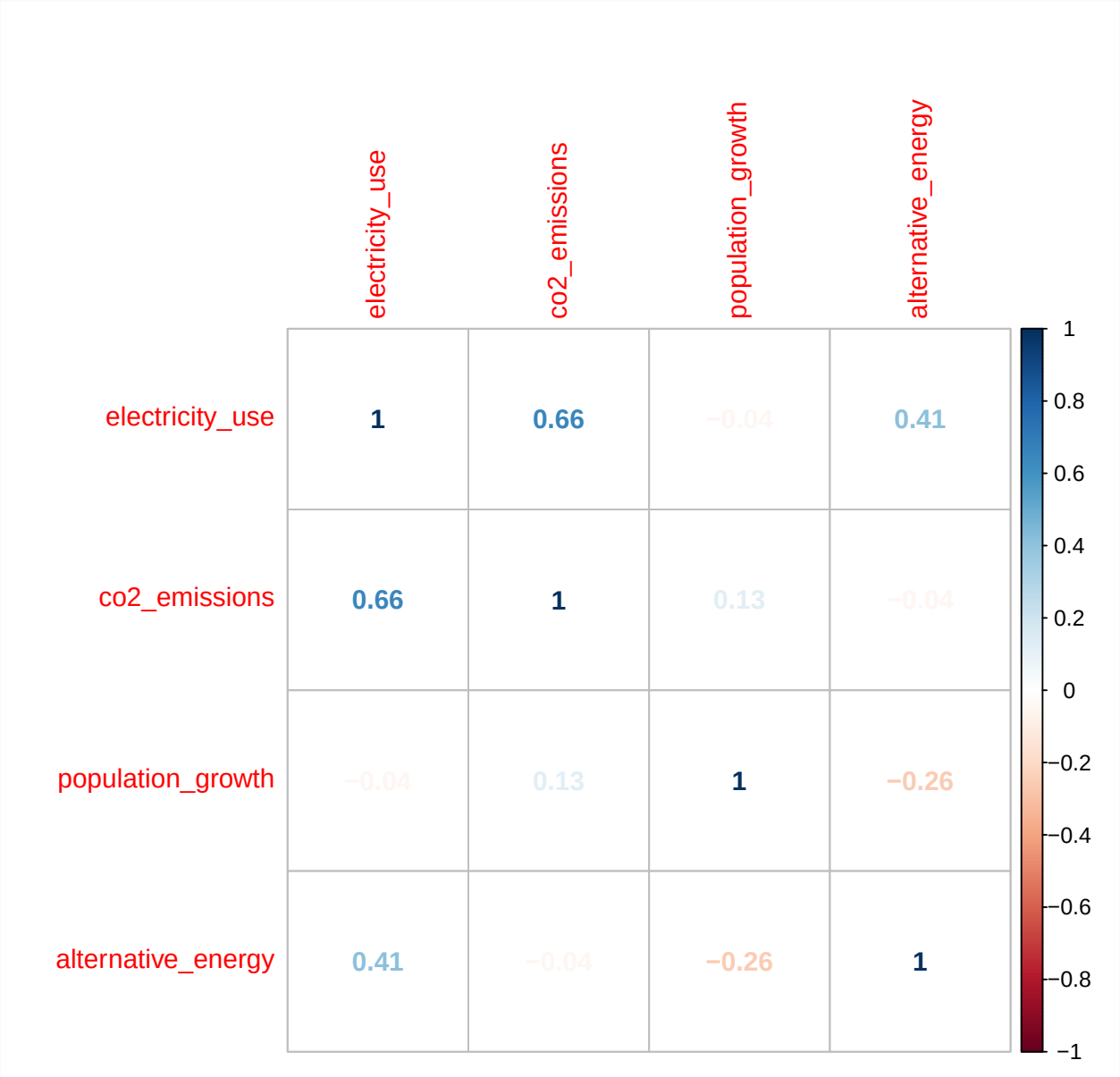
A simple way to examine the relationships between your sub-components is to create a correlation matrix.

Correlation: a mutual relationship between two variables.

Correlation coefficient: a number ranging from -1 to 1 calculated to represent the linear correlation between two variables. Usually denoted *r*.

Correlation matrix: A matrix showing correlations between many variables at once.

CORRELATION MATRIX



ROUGH GOAL

The overall goal is to theoretically understand how the variables are related to each other theoretically and empirically.

Roughly balance:

- not wanting very highly correlated variables, as these indicate duplication

and

- not wanting very lowly correlated variables, as this indicates they are related to different concepts.

Note: this is a very rough guide. More sophisticated tools await you in later courses.

NORMALISE VARIABLES

ON A LEVEL PLAYING FIELD

Observable variables are often on different scales as they can have different measurement units.

For example, *life expectancy at birth* is in years ranging from 0 to > 100 and *GNI per capita* is in US dollars starting from > 300.

Obviously, adding these two variables together would weight GNI more than life expectancy.

SOME NOTATION

Some notation:

$x_{u,t}$ observed variable value for unit u (e.g. country) at time t (e.g. year)

$I_{u,t}$ normalised value of the variable value for unit u at time t

NORMALISING APPROACHES

There are many approaches to normalising variables (see OECD 2008, 27-29).

Here are two for interval/ratio data ...

Min-Max rescales a variable to be between 0 and 1 based on the observed minimum and maximum values.

$$I_{u,t} = \frac{x_{u,t} - \min(\mathrm{X})}{\max(\mathrm{X}) - \min(\mathrm{X})} \tag{1}$$

where X are all of the observed values of the variable. (For R students: it is a vector of the values.)

Note: extreme values can distort the measure and can widen the range of closely spaced values.

Z-SCORE STANDARDISATION

Z-Scores create a common scale with a mean of 0 and a standard deviation of 1.

$$I_{u,t} = \frac{x_{u,t} - \mu_{\mathrm{X}}}{\sigma_{\mathrm{X}}} \tag{2}$$

where μ_{X} is the mean of the variable and σ_{X} is the standard deviation.

Note: extreme values can have a greater effect on the composite indicator.

DIRECTION

All of the variables that you think have a “positive” influence on the underlying concept need to be positive.

All of the variables that you theoretically think have a “negative” influence on the underlying concept should be negative.

NEED TO CONSIDER VARIABLE DIRECTION, EXAMPLE

For example, if we want to create a measure of environmental un-sustainability with the components: $C0^2$ Emissions, Population Growth, and Alternative Energy Use.

We would expect:

- More $C0^2$ Emissions $\rightarrow$ less sustainable.
- More Population Growth $\rightarrow$ less sustainable.
- More Alternative Energy Use $\rightarrow$ more sustainable

So, we need to reverse the scale of our Alternative Energy Use variable

REVERSE THE DIRECTION OF A VARIABLE

To reverse the direction of a variable simply subtract each value by the maximum observed value, i.e.:

$$I_{u,t} = \max(\mathrm{X}) - x_{u,t} \tag{3}$$

WEIGHTING AND AGGREGATION

WHY WEIGHT?

Once you have your normalised variables ($I_{c,t}$), then you need to consider how to aggregate them.

Specifically consider:

- Weighting: how important are each individual variables to the composite?

“No uniformly agreed methodology exists to weight individual indicators before aggregating them into a composite indicator.”

But generally, “what matters more …weighs more.”

https://composite-indicators.jrc.ec.europa.eu/?q=content/step-6-weighting

EQUAL WEIGHTING

If you simply add all of the variables together, you are implicitly assuming that they have an equal weight of 1. E.g.:

$$CI_{u,t} = \sum I_{u,t} * 1 \tag{4}$$

Sometimes this makes sense: e.g. economic activity is often measured in the same currency, or if you have no prior knowledge about how important each component is.

UNEQUAL WEIGHTING

There are many advanced techniques to determine variable weights (e.g. factor analysis, data envelopment analysis and unobserved components models).

In this course we are going to use "expert judgement" – you are the experts and will determine how much each component contributes to the composite indicator.

Imagine that you have four component variables X_1, X_2, X_3 and X_4.

Based on your expert knowledge, you believe that X_1 is much more important than the other three combined. So you give it a weight of 0.7 and the others 0.1:

$$CI_{u,t} = 0.7x_{1u,t} + 0.1x_{2u,t} + 0.1x_{3u,t} + 0.1x_{4u,t} \tag{5}$$

THRESHOLD AGGREGATION

Sometimes the concept we are measuring might be discrete. E.g. you are in a financial crisis or not in a financial crisis.

So, you might use your expert judgement to set a threshold, a point past which a unit goes from having the characteristic to not having the characteristics.

THRESHOLD EXAMPLE

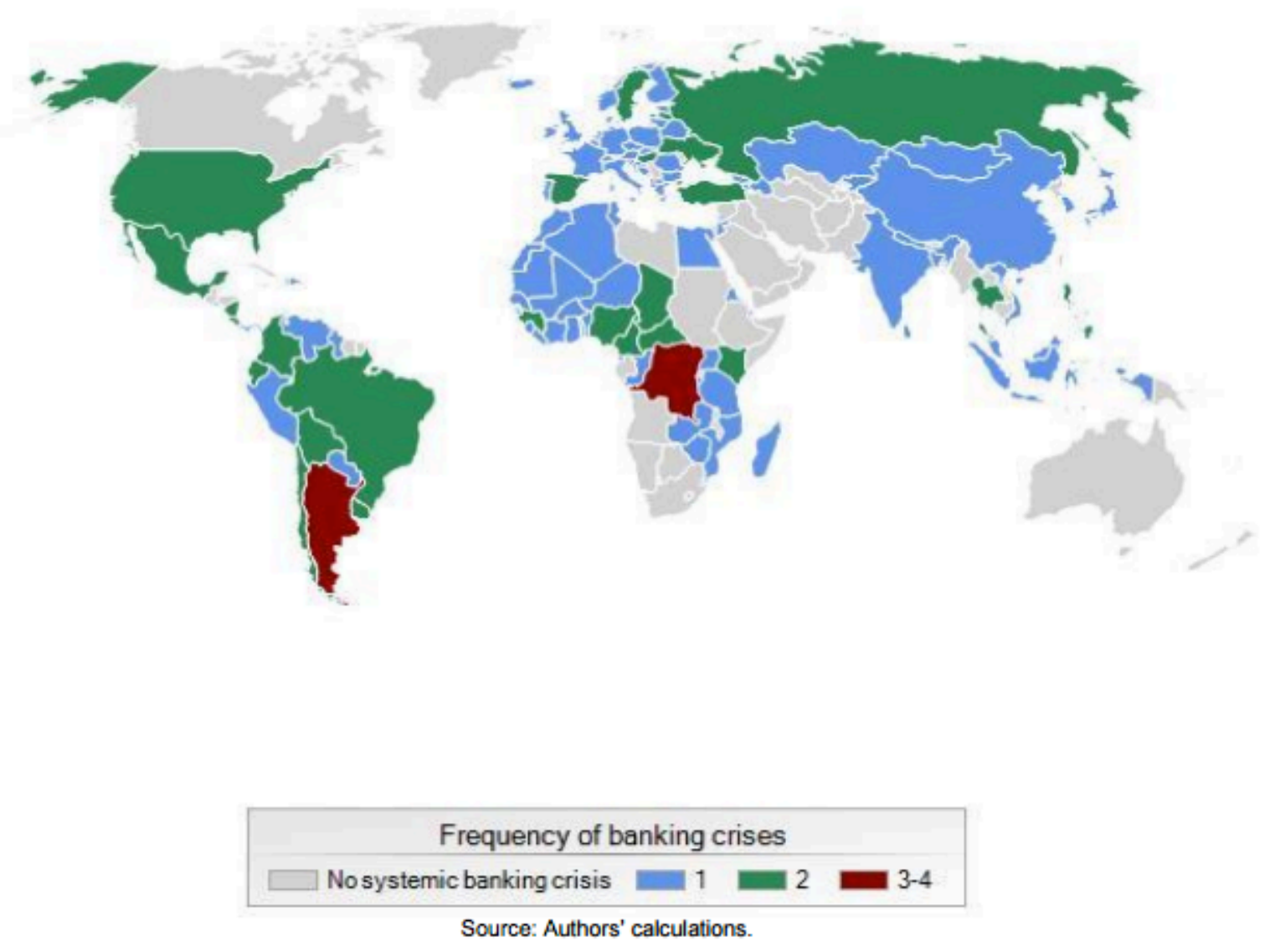
Laeven and Valencia (2013) determine a country has crossed a financial crisis threshold if:

- There is 'signficant distress' in a country's financial system.

and

- At least three of six policy responses are used (e.g. bank holidays, bank nationalisations).

Laeven & Valencia banking crises (1970-2011)



Source: Authors' calculations.

WHAT ARE WE MEASURING?

VALIDITY

Once you have a composite indicator, your work is far from done.

You need to conduct numerous tests to determine if your indicator is a valid measure of the concept you are trying to measure.

We will discuss this more in Week 10.

There is likely no perfect composite indicator for any given concept.

Transparency–conceptualisation, data gathering/cleaning, construction, validation–is crucial for others to be able to understand and evaluate your indicators.

PROS AND CONS

UNCERTAINTY

We need to remember that our indicators are estimates of what we want to measure.

We are uncertain about how well our indicators capture reality.

Uncertainty can be caused by at least:

- Error in our construct (i.e. including or omitting important variables)
- Measurement error in our raw data $x_{c,t}$
- Error in our weighting/aggregation.

Composite indicators are popularly used to rank units (e.g. cities, countries).

However, this can be highly misleading.

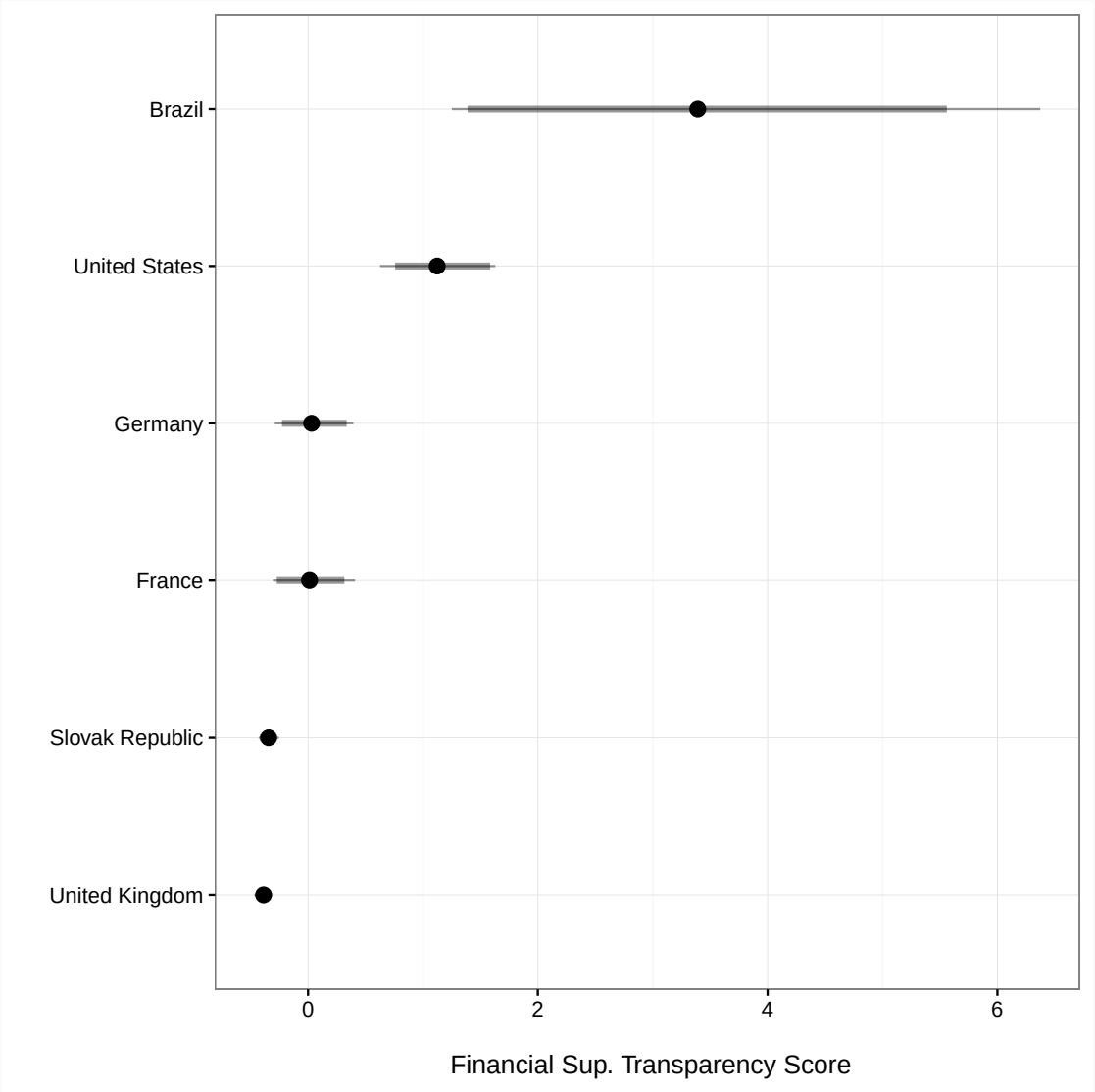
RANKING EXAMPLE

For example, Copelovitch, Gandrud, and Hallerberg (2015) created a measure of financial supervisory transparency.

Table: Financial Regulatory Transparency Index Ranking (2011)

Country	Rank
Brazil	1
United States	7
Germany	23
France	25
Slovak Republic	48
UK	61

INDICATOR WITH ABSOLUTE SCORE & UNCERTAINTY



In this course, we won't be covering more advanced ways to quantify your uncertainty about your composite indicators.

However, you should be careful and honest when you compare units by ranking their composite scores.

Be honest about what you don't know.

PROS AND CONS OF COMPOSITE INDICATORS

Pros	Cons
Summarise complex multi-dimensional concepts so that we can understand relationships between them	May be misleading if poorly constructed & non-transparent
Support policy-making	May lead to simplistic policy conclusions or missed to support particular predefined policy goal
Facilitate communication with the public.	May disguise problems in one dimension if construction is not transparent.
	May exaggerate differences between units if uncertainty is not acknowledged.

Partially from OECD (2008, 13-14)