THE ATTRACTIVENESS INDEX OF GLOBAL ECONOMIES

Thesis submitted by

DMITRIJ PETROV

ON 15th of June 2015

in partial fulfilment of the requirements
for the degree of B.Sc. Business Informatics.

Department of Computer Science
Reutlingen University

Supervisor: Prof. Dr. rer. nat. Bernhard Mößner, Reutlingen University
Co-reader: Dr. habil. Konstantin Kholodilin, German Institute for Economic Research

Statutory Declaration

I assure that this thesis is a result of my personal work and that no other than the indicated aids have been used for its completion. Furthermore, I assure that all quotations and statements that have been inferred literally or in a general manner from published or unpublished writings are marked as such. Beyond this I assure that the work has not been used, neither completely nor in parts, to pass any previous examination.

Reutlingen, 15 June 2015 Dmitrij Petrov _____

Abstract

In the following bachelor's thesis, I have developed countries' attractiveness index based on weights of the factor analysis. The aim was to create a ranking of 30 global economies build upon two dimensions – *economy & business* and *education*. Due to targeting a group of young adults being between 20 and 40 years old, six indicators from aforementioned fields were chosen in a way, which might relate to how appealing they perceive each country.

This work is divided into four chapters starting with the first one, which is an introduction to the topic. Chapter two (p. 13) describes the main part where I laid out my theoretical foundation, gathered necessary data and built my index. Because of data being unavailable for several countries, I needed to impute them first. Subsequently I have performed feature scaling using min-max technique and continued further with the multivariate analysis. In the final steps, I have constructed weights and aggregated them together with the scaled data. To expand the scope of my uncertainty and sensitivity analysis, I have performed comparison of two other weighting schemes as well.

In the third chapter (p. 45), I describe some technical aspects of my work and detail two code snippets of R programming language that were used during the construction of my composite indicator. In the last chapter (p. 50), my results are presented and summarized. Indeed, not surprisingly the United States of America is a leading nation in my index (score of 79.9) with a substantial distance from the second United Kingdom (73.9) and third Germany (73.5). On the other end of the ranking, Ghana, Kenya and Nigeria achieve last positions showing clearly that these countries are the least attractive to the young adults.

Acknowledgements

I wish to acknowledge my deep appreciation to my parents. I hope that this bachelor thesis will warm their heart.

Even more importantly, without my supervisor Dr. Bernhard Mößner, I would not be able to complete this work. Not only am I grateful to him that I could write this thesis under his supervision but also because of his openness, support and advices he gave me. Furthermore, I would like to thank Dr. Konstantin Kholodilin for accepting a position of being my co-reader of this thesis.

Lastly, I want to express recognition of people working at Quandl Inc.

Table of Contents

List of figures

List of tables

List of abbreviations

BOD	**B**enefit **O**f the **D**oubt
CI	**C**omposite **I**ndicator
CNBC	**C**onsumer **N**etwork and **B**usiness **C**hanel
CSV	**C**omma-**s**eparated **v**alues
DEA	**D**ata **E**nvelopment **A**nalysis
DFM	**D**ynamic **F**actor **M**odel
ETL	**E**xtraction, **T**ransformation and **L**oading process
EW	**E**qual **W**eighting
FA	**F**actor **A**nalysis
Fed	**Fed**eral Reserve System
FFR	**F**ed **F**unds **R**ate
GCI	**G**lobal **C**ompetiveness **I**ndex
GCR	**G**lobal **C**ompetiveness **R**eport
GDP	**G**ross **D**omestic **P**roduct
IDE	**I**ntegrated **D**evelopment **E**nvironment
ILO	**I**nternational **L**abour **O**rganisation (UN)
IMF	**I**nternational **M**onetary **F**und
LMCI	**L**abour **M**arket **C**onditions **I**ndex
MA	**M**ultivariate **A**nalysis
OALD	**O**xford **A**dvanced **L**earner's **D**ictionary
OECD	**O**rganisation for **E**conomic **C**o-operation and **D**evelopment
PCA	**P**rincipal **C**omponent **A**nalysis
PISA	The **P**rogramme for **I**nternational **S**tudent **A**ssessment (OECD)
PRC	**P**eople's **R**epublic of **C**hina
R&D	**R**esearch and **D**evelopment
SA	**S**ensitivity **A**nalysis
STEM	**S**cience, **T**echnology, **E**ngineering and **M**athematics
UA	**U**ncertainly **A**nalysis
UN	**U**nited **N**ations
UNESCO	**U**nited **N**ations **E**ducational, **S**cientific, and **C**ultural **O**rganization
WEF	**W**orld **E**conomic **F**orum

This page is intentionally blank.

1 Introduction

In the following part I am going to introduce my reader to this thesis, firstly by providing a reason why I choose to write about *composite indicators* (thereafter abbreviated as CI or an index) and secondly to explain my goals. In addition, I define aforementioned term and present two distinct examples of indices. Finally, I also discuss advantages and disadvantaged of using CIs as a tool that summarizes information.

1.1 The mission

During one early morning in September 2014 in New Jersey (USA), the headquarters of CNBC, Steve Liesman[1] reported on the story that Federal Reserve Bank has introduced a new index, which assess "*overall labour market conditions*" – hence the name *Labour Market Conditions Index* (LMCI; described in its own chapter 1.3.2). As he reported on the story, I become curious of how are indices designed and created, in particular ones that try to compare different countries' performances in many aspects of citizens' everyday life. Due to that, I want to develop my own composite indicator, which will be able to compare countries' attractiveness based on data from fields of economy, business and education. Through that, I want to investigate how nations stand against each other, in eyes of young adults who can choose to live in any country worldwide.

To have a representative sample, 30 countries[2] are chosen from six continents – with seven countries each from Asian and European and six from American region, see figure 1. Additionally, there are four African, four Middle-Eastern and two nations from Oceania. Together, they represent almost 17% of 193 officially recognized member countries of the United Nations (UN) with fourteen states being also a part of G20 club.

By using – now very popular[3] – R programming language (hereafter R), I am going to work with raw datasets to clean and transform data in order to build my own ranking. My methodology and further steps will closely follow *Handbook on constructing composite indicators: methodology and user guide* published jointly by the OECD and European Research Centre in 2005.

---

[1] "Steve Liesman – CNBC Senior Economics Reporter" CNBC, Web. 14 July 2015 http://www.cnbc.com/id/15838058

[2] An assumption is made that the sample represents entire population of reference.

[3] Measuring popularity of a programming language is hard and is a research topic for itself. Any results cause disputes among debaters and communities. Nevertheless, [Muen12], [Tippm14] [Tiob04], [Zapp14] and [Jaxen14] all conclude that R-language had seen a major growth in popularity in 2014.

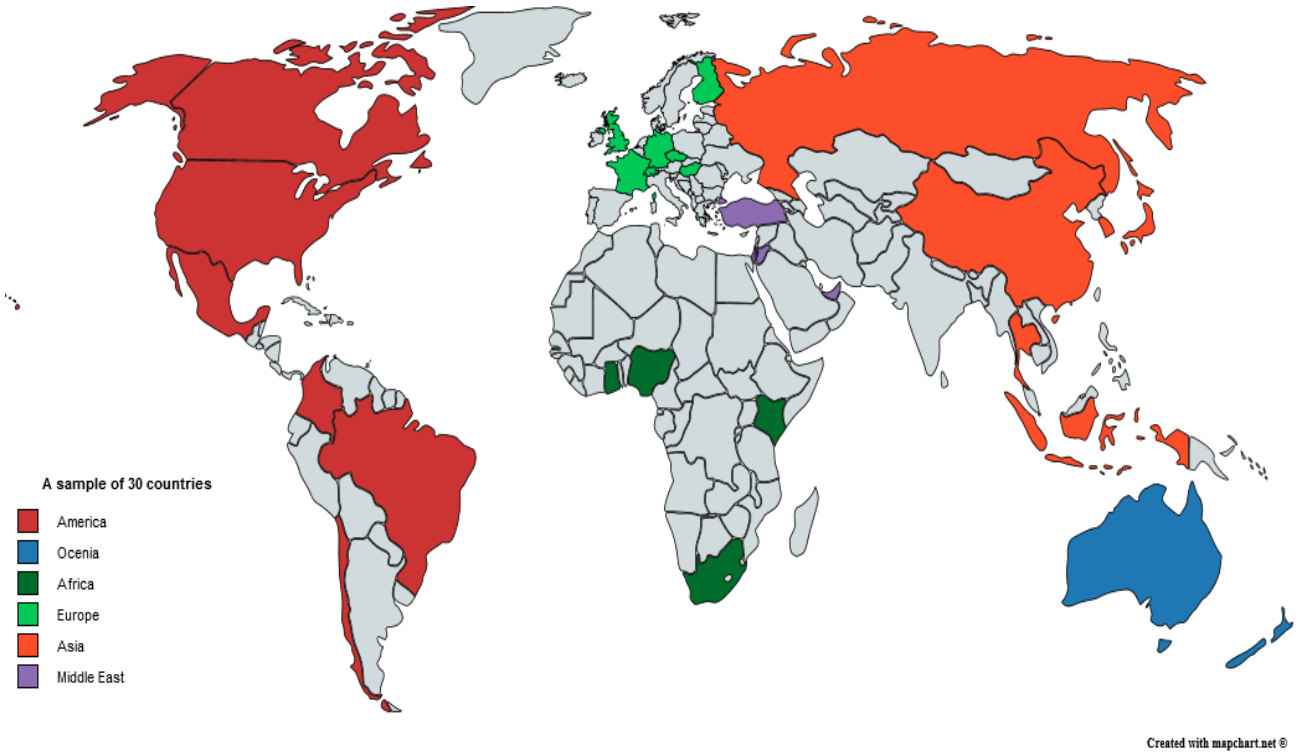

Figure 1 presents 30 countries sampled. For a named list, see table 10.

1.2 Defining composite indicator

Previously I have mentioned the word *composite indicator* and thus let me first begin by providing the reader a short summary of the word *index*.

In fact, it summarizes many individual indicators into the one numeric value. This leads to the creation of a ranking allowing measuring a predefined concept that otherwise would be hard or even impossible to evaluate. Therefore, for example, when comparing countries, such a composite indicator is able to show their relative performance to other peers. To continue further, I would like to split this phrase into two separate words and explain both of them.

An *indicator*, as Heink and Kowarik (2010) conclude in their article, is "*a profoundly ambiguous term that has different meanings in different contexts*". Being a noun, it can be used as a standalone expression in a sentence and, for example, the OALD defines the word as "*a sign that shows (…) what something is like or how a situation is changing*" [Indic14]. Thus, an *economic indicator* describes a level of economic activity. It may be also used to predict future level of something allowing reacting promptly to a problem and applying necessary policy, if needed. Concrete examples are the unemployment rate with the GDP,

which are all (lagging[4]) indicators describing the state of economy [Smith11].
In mathematics, Dodge (2008) describes the indicator as "*a statistic whose objective is to give an indication about state, behaviour, and changing nature during some period of an economic or political phaenomenon*". Although differences of opinion exist, there appears to be agreement that indicators refer to observed or unobserved values of variables, which "*describe[s] the state of a system*" [Walz20].
On the other hand, the term *composite* can be a noun, a verb and even an adjective. Being the last one, it is used with another noun to describe something that is "*made of different parts*", e.g. the concrete, which is a composite material consisting of sand, cement and water [Compo14].
Even though the reader now understands both words separately, defining the whole term *composite indicators*, per Saisana (2005) "*often called indices*", is not an easy task. She provides two types of definitions [Saitar02].
The first one is a *technical* definition which Saisana (2004, p. 3) describes CIs as "*mathematical combinations (or aggregations) of a set of indicators*". This goes later hand in hand with definition of Nardo et al. (2005, p. 15), who write that CI "*is formed when individual indicators are compiled [aggregated] into a single index on the basis of an underlying model*". Saisana (2004) further specifies that there is also a *conceptual* definition, which says that CIs are based "*on sub-indicators that have no common meaningful unit of measurement and there is no obvious way of weighting these sub-indicators*".
From a construction point of view, a common basis for indices is described by Freudenberg (2003) and Chernyak and Shumayeva (2014) who write that CIs contain of three levels of hierarchy; see figure 2. The foundation of pyramid carries individual indicators – "*quantitative or qualitative assessment of certain factors obtained through series of observations*" [Cheshu14, p. 77]. Then there are different dimensions, called *thematic indicators*. These are grouped together on a predefined topic or a theme, e.g. the quality of life dimension consisting of many socio-economic variables. Finally, both levels are necessary for the index itself that is formed "*when thematic indicators are compiled into a synthetic index and presented as a single composite measure*" [Jrcth15]. An equation (1), outlined by Freudenberg (2003, p. 7), is frequently used for calculating "*the performance*

[4] A lagging indicator is a "*measurable economic factor that changes after the economy has already begun to follow a particular pattern or trend*" [Lagg03].

of countries on different dimensions" and here the reader can see that w_i plays a major role in determining the final score of the index I for each country.

$$I = \sum_{i=1}^{n} w_i X_i \ \ where$$

X_i *is value of indicator* i (1)
w_i *is weight of indicator* i
$i = 1 \ldots n$ *indicators*
with $\sum_{i=1}^{n} w_i = 1$ *and* $0 \leq \ w_i \leq 1$

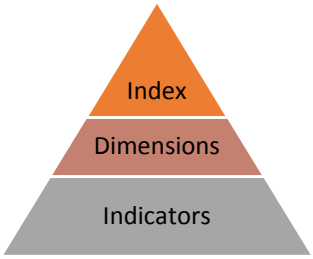

Figure 2 gives an overview of three levels of hierarchy, based on [Freud03].

1.3 Domains of application

Now that I have explained the term, we can look further at the domain of application of such indices.

In fact, they can be used in many environments not only to compare countries, as one can later see in the LMCI. However, in this thesis I am going to write about composite indicators that, which as Munda and Nardo (2003) report, "*stem from the need to rank countries and benchmarking their performance*" [p. 1]. Thus, indices are being developed by many think tanks, foundations and other organisation to compare nations each having a specific environment for example in the health care, ecology or an infrastructure. Moreover, not only are CIs common in areas above, but are "*used in a variety of policy domains such as industrial competitiveness, sustainable development, quality of life assessment, globalization and innovation*" too [Muna03, p. 1]. These domains are ideal "*multidimensional concepts which cannot be captured by a single indicator, e.g. (…) knowledge-based society*", and therefore their evaluation by composite indicators can be well suited [Hcci05, p. 15].

Indices are further "*useful in identifying trends and drawing attention to particular issues*" [p. 15]. This follows Freudenberg (2003, p. 3) who writes that CIs "*are valued as a communication and political tool*" as well, because they help in setting policy priorities to

benchmark and monitor countries' accomplishments. This, in turn, should assist decision-makers to create and implement the right policies.

Some examples of indices that rank countries and are often quoted in the media include:

- *Human Development Index* published by *United Nations*
- *World Press Freedom Index* compiled by *Reporters Without Borders*
- *Global Competitiveness Index* by *World Economic Forum*
- *Corruption Perceptions Index* by *Transparency International*

In the next several chapters, I am going to present two different examples of indices. One that ranks countries – the *Global Competitiveness Index* (WEF) – and another one, which measures conditions in the US labour market – the *Labour Market Conditions Index* – developed by the US Federal Reserve Bank.

1.3.1 WEF's Global Competiveness Index

The World Economic Forum has created *Global Competiveness Report (GCR)* "*to build a shared understanding of the main strengths and weaknesses of each of the economies covered, so that stakeholders can work together on shaping economic agendas that can address challenges and create enhanced opportunities*" [Wef14, p. 15]. One of outcomes of this report is the *Global Competitiveness Index (GCI)*, which ranks nations on scale of 1 to 7 – from the least to the most competitive. For six years in a row, the first place has been held by Switzerland (score 5.7), followed closely by Singapore (5.65).

The index is based on over 100 indicators and with 144 economies benchmarked in the current 2014/2015 edition it claims to represent today the most extensive study of its kind globally [p. 15]. GCI uses data from many international organisation such as the IMF, WHO and World Bank. However, most of the data the *Forum* needs for its report comes from their own *Executive Opinion Survey*, which "*captures the opinions over 14,000 business leaders in 148 economies on a broad range of topics"*, including *"skills gap, the level of corruption, or the intensity of market competition*" [p. 101].

1.3.1.1 The pillars of competitiveness

In the report, the *Forum* writes that competitiveness represents "*set of institutions, policies, and factors that determine the level of productivity of a country. The level of productivity, in turn, sets the level of prosperity that can be reached by an economy. (…) In other words, a more competitive economy is one that is likely to grow faster over time*" [Wef14, p. 20].

To achieve high level of competitiveness WEF identifies 12 major pillars, which a country needs to master. Each of these pillars below consists of several subcategories and these again contain of many individual indicators.

1. Institutions
2. Infrastructure
3. Macroeconomic environment
4. Health and primary education
5. Higher education and training
6. Goods market efficiency
7. Labour market efficiency
8. Financial market development
9. Technological readiness
10. Market size
11. Business sophistication
12. Innovation

The pillars above are further grouped in three different subindices – '*basic requirements*', '*efficiency enhancers*' and '*innovation and sophistication factors*' (see figure 4). The first four pillars are key for factor-driven economies. Pillars five to ten are for efficiency-driven economies, while last two pillars form a third stage – innovation-driven economies, where the most of the modern world resides [p. 27].

The framework uses two criteria to allocate a country into specific stage of economic development. The first criterion is "*the level of GDP per capita measured at market exchange rates [in US$]*" and its threshold can been seen in figure 3 [p. 26]. The second criterion is used only for an adjustment when the prosperity "*is based on the extraction of [mineral] resources*" [p. 26].

In reality, under the first criterion, countries such as Sierra Leone would be beyond the first stage of development. However, because their export depends on mineral resources and exceeds 70%, WEF considers them "*to a large extent factor driven*", thus in the first stage of development [p. 26]. It further captures states such as India or Ghana too. The second stage now includes countries such as Indonesia and South Africa, whereas the last one incorporates all developed economies e.g. Israel and the UK. On top of these three stages, there are additionally two transitional ones as well [p. 27].

To create an index – a score for each of 144 economies – "*the GCI takes the stages of development into account by attributing higher relative weights to those pillars that are more relevant for an economy given its particular stage of development*" [p. 25f.]. That means that although all 12 pillars do matter somewhat to all countries, some pillars are more relevant to nation's particular stage of development. For example, for Cambodia being very innovative is less important than having superior infrastructure and accessible health care. Consequently, "*the weights attributed to each subindex in every stage of development are shown*" in figure 3 [p. 26].

	STAGE OF DEVELOPMENT				
	Stage 1: Factor-driven	Transition from stage 1 to stage 2	Stage 2: Efficiency-driven	Transition from stage 2 to stage 3	Stage 3: Innovation-driven
GDP per capita (US$) thresholds*	<2,000	2,000–2,999	3,000–8,999	9,000–17,000	>17,000
Weight for basic requirements	60%	40–60%	40%	20–40%	20%
Weight for efficiency enhancers	35%	35–50%	50%	50%	50%
Weight for innovation and sophistication factors	5%	5–10%	10%	10–30%	30%

Note: See individual country/economy profiles for the exact applied weights.
* For economies with a high dependency on mineral resources, GDP per capita is not the sole criterion for the determination of the stage of development. See text for details.

Figure 3 displays "subindex weights and income thresholds for stages of development" [Wef14, p. 10].

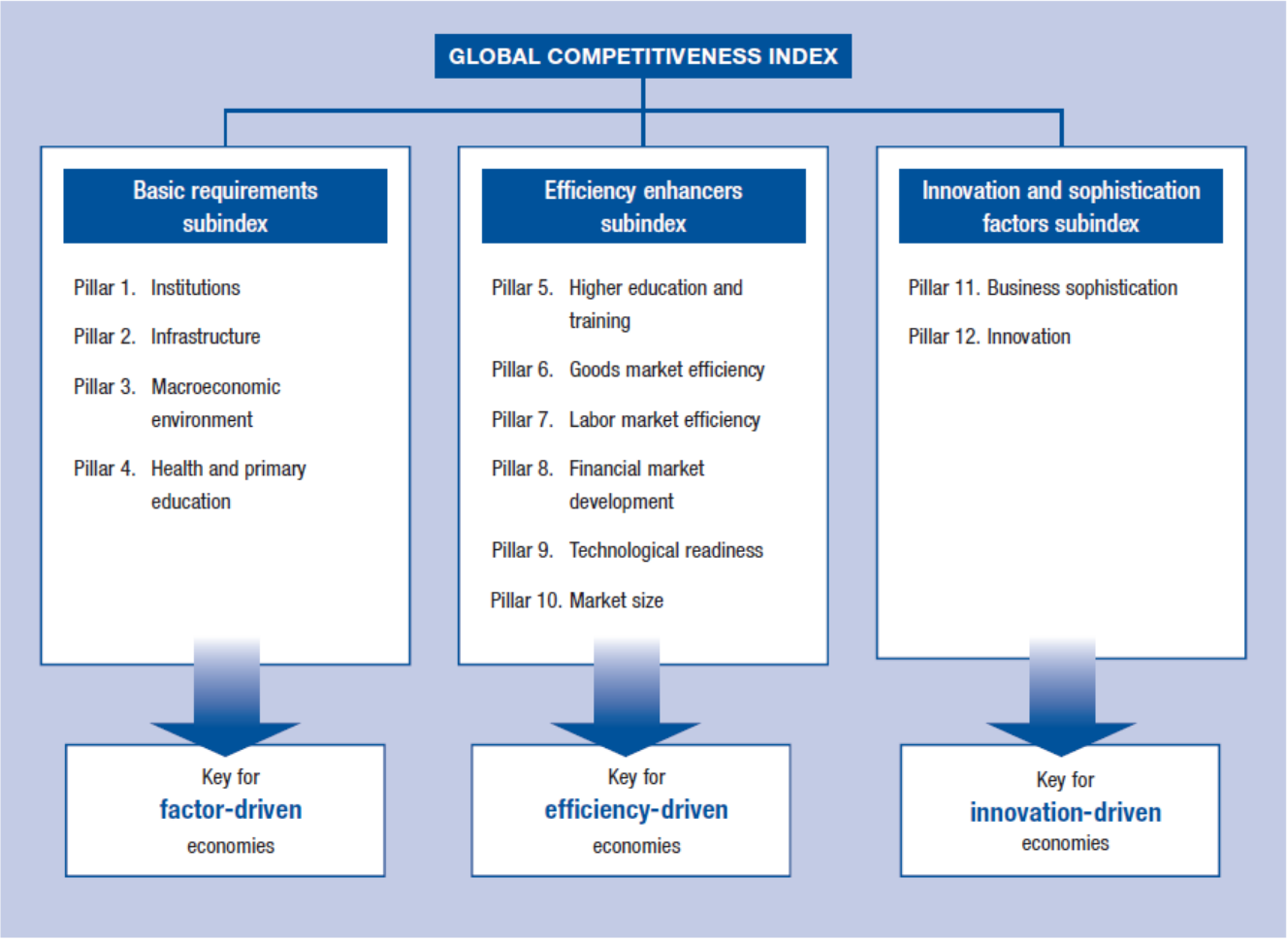

Figure 4 details how pillars are organized into three subindices, which are later weighted relative to each stage of economic development of countries [Wef14, p. 25].

1.3.1.2 The case of Singapore

This Southeast Asian country is for the fourth consecutive year on the second place in the overall ranking through possessing "*world-class infrastructure*" with "*strong focus on education*" [p. 28]. The methodology of how WEF calculates its competitive score is extensively described in the GCR. Therefore, here I want to show final steps that reflect aforementioned process described by Freudenberg (2003) too.

Being in the third stage of economic development, Singapore's score consists of weighting three subindices. Whereas '*basic requirements*' weighs only for 20%, pillars five to ten already account for a half of its index. Lastly, '*innovation and business sophistication factors*' justify 30% of country's competitiveness score, see figure 3 and table 1 [p. 350]. Once individual values for pillars had been calculated, a table such as one below can be derived.

Weights	20 %				50 %						30 %	
	Pillar 1	Pillar 2	Pillar 3	Pillar 4	Pillar 5	Pillar 6	Pillar 7	Pillar 8	Pillar 9	Pillar 10	Pillar 11	Pillar 12
Score	6.0	6.5	6.1	6.7	6.1	5.6	5.7	5.8	6.1	4.7	5.1	5.2

Table 1 provides an overview of Singapore's scores for each pillar and its corresponding weights [Wef14, p. 350].

To calculate Singapore's GCI, first I average scores within a given subindex:

$$\bar{x}_{basic\ requirenment} = \frac{6.0+6.5+6.1+6.7}{4} = 6.33 \quad (2)$$

$$\bar{x}_{efficiency\ enhancers} = \frac{6.1+5.6+5.7+5.8+6.1+4.7}{6} = 5.67 \quad (3)$$

$$\bar{x}_{inn.\ and\ soph.\ factors} = \frac{5.1+5.2}{2} = 5.15 \quad (4)$$

Then I use predefined weights for each of subindices and multiply them with the average scores above. Consequently (2), (3) and (4) with appropriate weights gives me a score of 5.64 (5), which corresponds to the 5.65 in the report where only exact figures have been used.

$$GCI\ score_{Singapore} = (6.33 * 0.2) + (5.67 * 0.5) + (5.15 * 0.3) = 5.64 \quad (5)$$

By using such a *linear aggregation*, see later chapter 2.6.2, this whole process follows the notion described by Tangian (2007) as well, who writes that CIs "*measure (…) particular properties and to summarize them, eventually with weights which reflect their importance*" – and for Singapore the *efficiency* and *innovation factors* are clearly the most important ones to stay globally competitive.

1.3.2 Fed's Labour Market Conditions Index

As briefly noted in the chapter 1.1, the decision to write about CIs was reporting about *Labour Market Conditions Index* (see figure 5). The LMCI is, as Chung et al. (2014) write, a "*dynamic factor model of nineteen labour market indicators [time series]*" that was first introduced in May 2014 to gauge US labour market [Lmci14]. This composite indicator includes data from "*broad categories of [un]employment, (…), workweeks, wages, vacancies, hiring, layoffs, quits, and surveys of consumers' and businesses' perceptions*" and the most strongly correlated data with the index are unemployment rate and payroll employment [Lmcidoc14, p. 2 and 18].

There had been one important reason for developing such an index. As authors write "*often-cited indicators (…) measure a particular dimension of labour market activity, (…) [and it is common] for different indicators to send conflicting signals about labour market conditions. (…) analysts typically look at many indicators when attempting to gauge labour market improvement. However, it is often difficult to know how to weigh signals from various indicators*" [Lmci14]. In consequence, furthermore, LMCI can assist Fed officials in a decision when to begin raising US interest rates (FFR) [Fedfunds, Flmi14, Fomc11]. Because of that, Chung et al. (2014) have developed a model summarizing different labour market indicators.

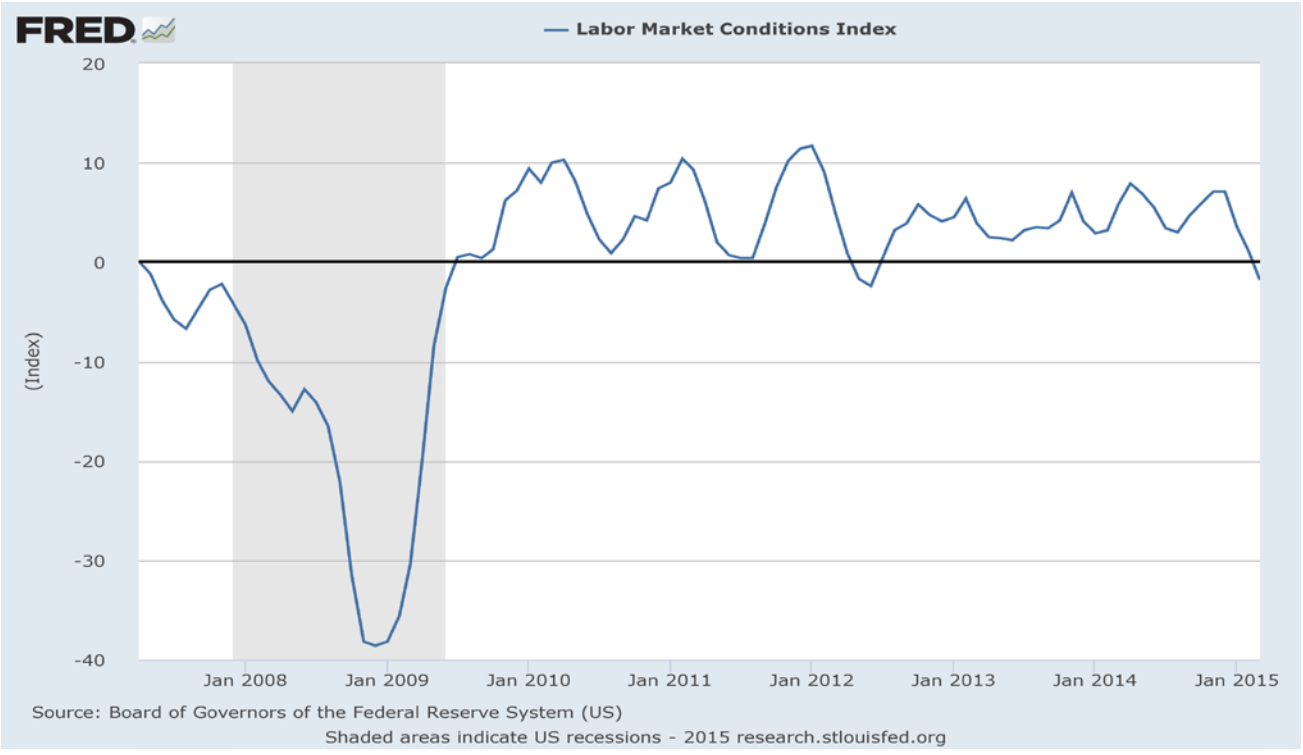

Figure 5 presents the Labour Market Conditions Index, monthly & seasonally adjusted between April 2007 and April 2015 [Frblm14].

1.3.2.1 The labour market model

The level of the index "*has no obvious economically-meaningful interpretation*" [Lmci14]. While growing curve means improvement of conditions in the labour market, a declining curve means just the opposite. Here, researches have focused on building the index where users are interested in its change, rather than its level. This is a strong contrast to the previous example with the *Forum's* GCI.

As can be seen in figure 5 and authors describing their work, the index "*generally declines during recessions (the grey shaded areas) and typically rises during expansions*" [Lmci14]. Therefore, it saw a huge decline between 2008 and second quarter of 2009 – as conditions in the labour market had been the worst (e.g. layoffs etc.). Since then, however, the market has improved and currently we "*see rather the uneven pace of the ongoing labour market recovery*" [Lmci14]. In part due to this improvement, the US Fed has already announced that it may start its *normalisation process* already in the second half of 2015, i.e. the first increase of interest rates since July 2006 [Fedfunds].

In recent years, much research in the econometrics has been attributed to creating statistical models summarizing information from numerous indicators, especially time series. Increasingly common are dynamic factor models (DFM), which extract "*a small number of unobserved factors that summarize the comovement among a larger set of correlated time series*" [Lmci14]. These models can be used to forecast key economic indicators or, for example, to develop labour market indices similar to the one described in this chapter, see Hakkio and Willis (2013). Alternatively, they can be also used to predict business cycles of the economy, see [Negotr08] or [Chazey12].

The essential feature of the dynamic factor model used by Chung et al. (2014) is that "*its inference about labour market conditions places greater weight on indicators whose movements are highly correlated with each other*" [Lmci14; Geweke77]. Once indicators provide contrary signals, the model "*reflects primarily those indicators that are in broad agreement*" [Lmci14]. Also very importantly is that "*the entire history of the LMCI may revise each month*". This is due to the model as well as data, which may be unavailable for the monthly release [Lmci14].

In their conclusion, authors write that their LMCI "*appears to be a useful tool for assessing the change in labour market conditions*" and providing "*a direct and transparent way to asses (…) questions arising when the model's [many individual] indicators appear to send different signals about labour market conditions*" [Lmci14, Lmcidoc14, p. 25].

In my conclusion, however, I want to stress that Fed researchers do not release raw datasets

with precise steps and formulas used for calculation, through which others could easily verify the methods and results[5]. In addition, some elements of this index have been criticised, notably by Carola Binder or Timothy Duy, who published an opinion on this index, concluding that one may use the index "*with caution. Extreme caution.*" [Duy14; Cbinder14]. He wrote that "*the Fed appears to want you to believe the LMCI is important, but they really don't give you reason to believe it should be important*". Hence, he questions the assumption that it may have any policy relevance, e.g. on a decision of raising interest rates.

1.4 Advantages and disadvantages of indices

Up until this chapter, I have explained what composite indicators mean, how and where they can be used. Therefore, let me conclude the first part of this thesis with a discussion of the pros and cons of their application.

Generally, CIs are used to abstract complex issues into a perspective that can support decisions-making processes, e.g. as the reader could just see in the timing of raising interest rates. Nardo et al. (2005) write, when done correctly, CIs enable easier interpretation of many separate indicators by providing a 'bigger picture'. This creates a discussion among parties involved and finally it should lead to a better decision and policy. Alas, if the index is poorly constructed, it may "*send misleading policy messages*" [Hcci05, p. 15]. This goes even beyond that as when policymakers implement wrong policies, based on the knowledge of such index.

Furthermore, indices simplify interpretation and conclusion. However, this needs to be seen both positively as well as negatively. On the one hand, this may lead to improper decisions because some facts are not included and hence not providing a holistic overview of the subject. From a political point of view, they can be misused "*to support a desired policy*" too, same authors continue [p. 15]. One the other hand, as already mentioned, indices comprehend many individual data, which for example in case of LMCI may send diverse signals. Thus, they may provide easier and clearer look on many individual facts. From the users' point of view, if transparency of construction of the model is missing, it may be subjected to (political) disputes and more importantly, it may hide "*serious failings in some dimensions*", lack "*sound statistical or conceptual principles*" and as a result

[5] Yet, the model is shortly discussed in their *Working Paper* [Lmcidoc14].

"*increase the difficulty of identifying proper remedial action*" [p. 15f.].

Based on Nardo et al. (2005), Tarantola et al. (2005) and others I have comprehended some arguments in table 2. From studied literature on this subject there appears to be no conclusion in favour of either one. Nevertheless, it is always noted that the modeller needs to follow steps, which can provide the reader a clear use case for the index. Moreover, the modeller should be also required in the first place to provide maximum transparency of how the index has been developed.

Pros	Cons
Summarises complex dimensions into 'bigger picture' by reducing size of indicators, resulting in one data output. This allows easier interpretation (at least, it should).	Comprehends many indicators, which may have nothing in common. Thus, it may send false policy messages, invite simplistic conclusions and negatively influence taken actions.
In the case of popular indices such as WEF's GCI, it attracts public interest and places country's performance at the centre of political discussion, forcing the government to (re-)act.	Having a construction process not transparent (enough), it will be subjected to mistrust, misuse and disputes. All that resulting in supporting (I) wrong policies and (II) having 'zero effect' on the current ones.

Table 2 summarizes the pros and cons of using composite indicators based on [Toci05] and [Hcci05].

2 Countries' Attractiveness Index

In the second part of my thesis, I am going to build my index. Therefore, first I begin with a description of my framework, which needs to be sound as only then users can understand and interpret my outcomes. At later stages, I advance with a selection of variables, handling of missing data and their scale normalisation. After performing the multivariate analysis, I aggregate normalised scores with a newly calculated weights based on the factor analysis. This results into the *Attractiveness Index* for a sample of 30 global economies, which additionally I am going to compare with outcomes of two other weighting methods. Before going further, I want to go ahead and mention that in every step I make unavoidable subjective choices, which will result in bringing a source of uncertainty on the model and outcomes. To compensate my subjectivity, in each step I attempt to explain my reasons and provide a broader perspective on the matter.

2.1 Theoretical framework

> *What is badly defined is likely to be badly measured.*
> *[Nardo et al. 2005, p. 24]*

According to Nardo et al. (2005, p. 51), the most important quality properties for any theoretical framework are the relevance, credibility and interpretability. The first two are even more important at later stages because through "*careful evaluation and selection of basic data*" the modeller ensures "*that the right range of domains is covered in a balanced way*" to fulfil the goals of the composite [p. 48]. The interpretability on the other hand "*reflects the ease with which the user may understand and properly use and analyse the data*" [p. 49].

Developing a framework consists of defining a concept, determining subgroups and identifying selection criteria for my indicators [p. 24]. Therefore, let me first begin with the concept, because this is essential to understand how I define what countries' attractiveness is in the regard to my target group.

Based on people's education, environment they are in or their social status, each person has its own definition and understanding of attractiveness. For example, for high-income people the country may be attractive only when it has a stable political and business atmosphere, high quality of life and low taxation. On the contrary, for a younger generation (e.g. students), it may go into the direction where living in a specific country is interesting to their goals, desires and future prosperity. In addition to the abovementioned

quality of life, now it might be also its affordability, climate or just the 'cool' factor that can play equally important role. For a group of people between 40 and 60 years, attractiveness can be something entirely different too. They might be more interested in the social welfare and having future for their children.

In my index, I want to concentrate on a definition of what attractiveness means for the younger adults, who are between 20 and 40 years old as my composite indicator targets that group. It is founded on the idea that the country is attractive to them when it has good education system (e.g. students achieve high results in international assessments) and appealing business & economic climate (e.g. low level of bureaucracy). This doesn't mean that if a country is ranked last in my ranking, it will be automatically unattractive. To assess such a concept extensively, it is necessary to include far more dimensions, not just two. Nevertheless, in this work I want to investigate what are the outcomes of having just these two elements in my ranking.

The index is divided into 2 subgroups, as described above. The first one is the educational dimension, which is important because it plays a major role for country's innovations, R&D and broadly speaking its future in the world. Even if people have jobs, but education is suffering from many problems, the country – in the longer term – cannot be as innovative and attract the best talent. Therefore, its potential for succeeding and being recognized worldwide is substantially lower. Take for example nations such as South Korea or Japan that are known for their high quality of education and concentration on STEM fields[6], resulting into major accomplishments in many – not only scientific – domains.

The second dimension is the economic and business perspective of the country. This dimension can play an important part in decision of young adults in which nation they want to live and stay. For instance, nowadays young adults often consider building their own company rather than working for one. Having said that, the country needs to attract businesses and stimulate their creation in order to set up a whole environment where people want to live, work and raise their children. Thus further advancing the nation. However, if the country has complicated tax code, slow government with unnecessary bureaucracy and weak economy combined with the high youth unemployment[7], the probability of people coming into this country and founding there new businesses is rather low. In today's world, some of abovementioned elements can be seen in Greece, where e.g.

[6] STEM is an acronym for subjects in *science*, *technology*, *engineering*, and *mathematics*.

[7] The International Labour Organization (ILO), a UN agency, defines it between 15-24 years old [Ilo11].

youth unemployment, according to data of Eurostat, is between 50% and 60% [Eurostat]. Undeniably, young adults from Greece are trying to relocate to other countries worldwide to gain better career perspectives [Eurostat15].

To conclude my section of the theoretical framework, lastly I need criteria for my indicators. Here it is important to identify them in a way that enables the index to have a meaningful value and guide me "*to whether an indicator should be included or not in the overall composite index*" [Hcci05, p. 24]. Indeed, below mentioned ones "*strongly affect(s) [indicator's] accuracy and credibility*" too [p. 50].

I am measuring an attractiveness of countries and hence only ***output*** indicators should be considered [p. 24]. Meaning that it would be a mistake to include country's spending on the education, because this input indicator doesn't provide any meaningful interpretation of how good the education is, based solely on this variable. In fact, high spending on education doesn't mean that students will achieve good results in international assessments and vice-versa. Other criteria, which are equally important for me, are the overall quality and timeliness of data. Therefore, I am going to prefer data from organisations such as UN instead of collecting them individually through nations' departments of statistics. Even though, if some values will be missing or lag (a bit) behind the most current ones.

Finally I want to say that while the choice of all indicators will be guided by my theoretical framework, the "*process [of data selection itself] will be quite subjective as there may be no single definitive set of indicators*" [p. 25].

2.2 Data selection

> *Poor data will produce poor results in a 'garbage-in, garbage-out logic'.*
> *[Nardo et al. 2005, p. 25]*

In this chapter, I am going to present all my chosen indicators and describe how they relate to my index, see table 9 and table 11. Similarly, now, the accuracy, credibility and timeliness are very valuable quality characteristics of the data selection process. According to Nardo et al. (2005, p. 49), the accuracy is "*extremely important*" dimension, because together with credibility they reflect the "*confidence that users place in (…) products based (…) on their image of the data producer [e.g. the International Monetary Fund]*". Users, consequently, trust such data due to being "*in accordance with appropriate statistical standards and policies (…)*" [p. 49].

As the reader is going to observe, twice 'hard' data are unavailable for specific countries.

Therefore, in the first case, a proxy measure will be applied. However, in the second case, where there is no such proxy possible, an imputation will be conducted, see later chapter 2.3.

2.2.1 Index of Economic Freedom

Let me first begin with a ranking published annually jointly by *The Heritage Foundation* and *The Wall Street Journal*.

The foundation defines *economic freedom* broadly as each person controlling his own destiny in a free society, where everybody is "*free to (…) produce, consume, and invest in any way they please*" and the government "*allow[ing] (…) capital, and goods to move freely, and refrain from (…) constraint of liberty beyond the extent necessary to protect (…) liberty itself*" [Frein15]. These are three "*fundamental principles of economic freedom (…) that underpin every measurement and policy idea presented*" in the index, consisting of four elements – *rule of law*, *limited government*, *regulatory efficiency* and finally the *market openness* [Frein15]. Due to all of that, such a ranking is important for my composite, because it shows which countries can provide best conditions for development and empowerment of each individual and business. Thus potentially contributing to the success of the whole nation.

Through measuring those four sub-components in 186 states, the foundation has documented over the years a clear and strong "*link between economic freedom and long-term development*" in education, health care and many other fields as well, all of which significantly affect nations' prosperity [Frein15].

A remarkable fact is that "*for 21 consecutive years*" the Hong-Kong SAR (89.6) has been sitting at the top of the ranking [Frein15]. Yet the gap with the second Singapore (89.4) "*has almost vanished*" in recent months and years. Additionally, because of the Hong-Kong's protests in 2014/2015 and the fact that the current edition of the ranking mostly covers data between "*the second half of 2013 through the first half of 2014*", Singapore can already overtake the territory in the next 2016 version [HKprotests, Frein15].

The indicator for my index is the final score of each country where having highest score on scale from 0 to 100 implies being economically the freest nation. The access to the data has been provided by using *xlsx* R library to read the Excel worksheet available on their website [xlsx14, Frein15]. The reader can see the corresponding R script in *RawData / FreedomIndex.R*.

2.2.2 Global Competiveness Index

This ranking has been already described in chapter 1.3.1 and thus here I want to highlight that the index is one of its kind that assesses countries' competiveness on the global level. By being highly competitive, any nation is able to overcome economic transitions and adapt itself quickly to new trends and needs. To quote WEF, the report "*shed light on key factors (…) that determine economic growth and the level of present and future prosperity in a country*", each being in its own stage of economic development and unique environment [Wef14, p. 15].

To access data, the *Forum* provides an Excel worksheet, which together with the R library mentioned in the previous chapter I have used to extract data [Wef14]. The reader can see the R script in *RawData / WEF.R*.

2.2.3 Human Development's Education Index

Since 1990, the United Nations Development Program (UNDP) has published annual *Human Development Report* highlighting progress in areas such as (un)employment, economic growth, gender inequality, poverty or human mobility [Undr14].

One of report's outcomes is the *Human Development Index* (HDI), which emphasizes "*that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone*" [Undr14]. According to authors, the index can be used for questioning how nations – with a similar level of gross national income per capita – "*can end up with different human development outcomes*" [Undr14]. Thus stimulating the debate about government's policy priorities.

The HDI, which in the current 2014 edition benchmarks 187 countries, is a measure of achievement in three dimensions of human development – (I) a long and healthy life, (II) being knowledgeable and (III) have a decent standard of living [Hditech14]. The second dimension, which I am also going to use for my composite, is expressed in the *Educational Index* ranging from 0 to 1 where a value close to one means that the nation is being very knowledgeable [p. 1]. It is calculated by establishing two other results – the *mean* and *expected years of schooling*, which had been added up and divided by two [p. 2].

Consequently, such a measure is important for my index, because the education is a major source of human progress, which is further easily seen e.g. in the R&D when comparing countries and their scientific achievements. Additionally, the knowledge contributes back in the capacity of being more competitive with others nations, having more skilful workers

(e.g. for STEM fields) and solving many problems of today's world such as an income inequality or the environmental protection. Therefore, not only the education has a measurable impact on our economic growth (not limited to, of course) but also allows us to distinguish between advanced, developing and underdeveloped nations. Hence, by improving citizens' knowledge moving the whole economy from one stage to another.
The data have been scraped from the UNDP's website using the *rvest* R library [Eduin13, rvest]. The reader can see the corresponding R script in *RawData / EUI.R*.

2.2.4 Pearson's Learning Curve Index

First compiled in 2012, *Pearson* and the *Economist Intelligence Unit* (EIU) published an updated report in 2014 with the major objective "*to collate and compare international data on national school systems' outputs in a comprehensive and accessible way (…)*" [Leacu14, p. 25]. For construction of the index, the EIU has selected several indicators that "*measure countries' output performance in education*" [p. 25]. These indicators were grouped into two dimensions – *cognitive skills* and *educational attainment*. For the first one, the index "*uses (…) reading, maths and science scores from PISA (…), TIMSS (…) and PIRLS (…) [educational assessments]*" [p. 25]. For the next one, "*literacy rate and graduation rates at the upper secondary and tertiary level*" were used [p. 25]. Through applying a z-score, a ranking of 40 countries has been created and one of key findings was not surprisingly that "*East Asian countries – always strong performers – now have a monopoly at the top of most education measurements*" [p. 10].
Authors of the index have further concluded that reliable data for many countries in the world simply do not exist, thus they have measured only 40 of them. This, for example, has caused a situation with the mainland People's Republic of China, where data are not available "*due to the lack of test results at a national level*" [p. 25]. Therefore, *The Learning Curve Index* uses outcomes of the Hong-Kong SAR as a proxy for China. Indeed, due to such a proxy, not only it has skewed results for their ranking but now it may effect mine too. The reason is widely acknowledged fact that at least in two provinces – the Hong-Kong SAR and Shanghai – the level and quality of education is far better than in the mainland China.
Additionally, the reader should be acquainted with the fact that for countries such as Nigeria and the United Arab Emirates values are not available at all. Specifically this happens only in my two regions (Africa and the Middle East), where out of eight countries

only two – Israel and Turkey – could have been measured. At this stage, I want to note that a solution to the missing data in *The Learning Curve Index* will be presented in the next chapter 2.3. Nevertheless, to extract existing data *Pearson* provides an Excel worksheet, which I also use with *xlsx* R library again. The reader can see the corresponding R script in *RawData / LearningCurve.R*.

2.2.5 Countries' H-Index

Based on an article published in 2012, Guerrero-Bote and Moya-Anegón proposed a new indicator called SJR2 (SCImago Journal Rank), which measures scientific journals' "*impact, influence, and prestige*" [Sjr12, Scim15]. To construct their indicator, the described methodology has been applied to all published publications between 1996 and 2013 in the *Scopus.com* bibliographic database. In addition to the ranking of journals, an analysis allowed them to create other rankings where for example users can look up aggregated (self-)citations or a h-index for each country.

For my purpose, I take countries' h-index named after its founder Jorge Hirsch [hindex]. Such an index is defined as "*the number of papers with citation number higher or equal to h*" and although it is usually used in reference to the researcher, here it reflects both number of written articles and citations per country as well [Calchin14]. From all six indicators, this one is also the most disputed and criticised one. Although, it is considered sometimes as *the* measure of scientific output, it can be (easily) manipulated too. Yet, the resulting number is important for my composite, because it shows what an impact and productivity researches can achieve in a particular country reflecting local circumstances and environment, e.g. shown through governmental spending.

SCImago provides its index for countries, ranging from 1 to 1518, both on its website as well as in the Excel worksheet. I use the later one, again using same aforementioned R library, to extract data for my index. The reader can see the R script in *RawData / HIndex.R*.

2.2.6 Youth unemployment

As already briefly noted in the business & economic dimension in the chapter 2.1, the youth unemployment rate might also play a certain role in decision to which economy young adults may consider moving into. The rate is defined as "*a measure of the inability of an economy to generate employment for those persons who are not employed but are*

available and actively seeking work" [Ilo11]. In my situation, data for each of 30 countries estimate a total labour force aged between 15 and 24 years, in percent, without having a job as defined by the ILO, a United Nations agency [Wbyu15].

This indicator is an exception from others as well, namely high value for a nation has to be interpreted as 'bad', not 'good' as opposed to other five variables in my dataset. "*Since only indicators with the same direction [in the index] can be aggregated*", I will need to transform this variable during the normalisation process, see later chapter 2.4 [Luxguc15].

For this indicator, data have been accessed through *Quandl* R library and the original source is the abovementioned ILO [Wbyu15, Quandl14]. The corresponding R script can be found in *RawData / Unemplo.R*.

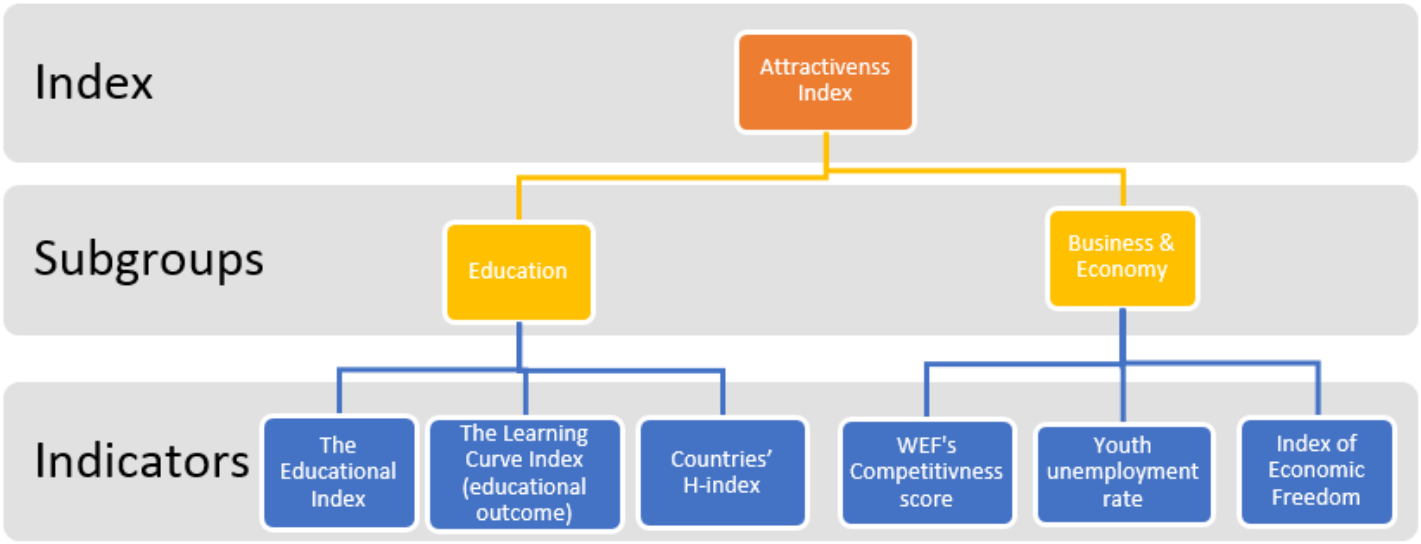

Figure 6 gives a graphical representation of my composite.

2.3 Imputation of missing data

When data are missing, Nardo et al. (2005, p. 26) write that it can significantly make it harder to develop a "*robust composite indicator*". As already mentioned in ch. 2.2.4, six values (which are just under 4% of total) simply do not exist in *The Learning Curve Index* for countries such as the UAE or Jordan. The *imputation* now deals with the question how should this missing information be handled, e.g. replaced or re-created. As a matter of fact, for my index it is absolutely necessary to have a complete dataset and non-existent values thus require some kind of a solution.

To identify why data are missing, it needs to be said that neither *Pearson* nor EIU could have been able to 'observe' them, at all. For instance, the index draws data from PISA exam. However, countries such as Kenya and South Africa are not a part of OECD as only

its member countries are automatically assessed [Pisa2000]. Due to this, countries must specifically opt-in for the assessment, which they have not done. Therefore, only a small subset of countries in *The Learning Curve Index* could have been measured.
Indeed, one solution can be using a proxy measure (a similar representation of missing phenomenon; e.g. proxy results of the Hong-Kong SAR for the PRC, see 2.2.4), which I would use as a replacement of my missing information. However, finding such a proxy is here extremely challenging because the abovementioned index comprehends many individual educational assessments. Thus, it would be a mistake to pick just one because results from other assessments would not be included. Hence, other methods need to be considered and will be further discussed in the following two chapters.

2.3.1 Possible (non-proxy) treatments

Starting with the easiest resolution, I would be simply omitting rows with NAs – deleting them altogether. This method is called *Complete-Case Analysis (CCA)* and it is not a right approach to my situation, because in such case my data frame has just 24 rows (i.e. countries with all six variables present). Such a loss of information is unwanted, in addition to requiring other special assumptions [Pigott01].
An option would be to use *Available-Case Analysis (ACA)*, where I take all present observations. This, however, introduces a bias making it hard to compare different analyses due to the different sample size of *The Learning Curve Index*. In addition to the requirement of having complete data, see previous page [Igl04].
Then, there are methods of simple imputation, such as *hot* or *cold imputation*. Moreover, there is also a *mean imputation* where by taking an average of available observations I would replace all six missing values with that mean [Howel12]. However, (I) the average adds no new information but at the same time decreases variance and (II) equally important it would be assumed that e.g. South Africa's quality of education is on the same level with Jordanian's one. For the context, the mean suggests 0.003[8], which would put them both at the level of France and Chile – and this is clearly not representative of their real situation. Not only is a bias introduced again but the method "*distorts covariances and intercorrelations between variables*" too [Schgram02, p. 13]. Pigott (2001, p. 13) writes that "*mean imputation cannot be recommended under any circumstances*".
Another technique is to use a *regression imputation* and through utilizing other indicators

[8] Calculated as the mean of 24 countries' z-score. For more, see *RawData / LearningCurve.R*

in my dataset, I could predict the score for each of six countries. A downside of this approach is that imputed data would be lying in the regression line, thus having "*the variability of the imputations (…) too small, so the estimated precision of regression coefficients will be wrong and inferences will be misleading*" [Lshtm15].
To summarize, "*the uncertainty in the imputed data should be reflected by variance estimates*" [Hcci05, p. 27]. Yet, simple imputation procedures "*are known to underestimate*" it making themselves inappropriate for handling missing data in certain cases such as mine [p. 27].

2.3.1.1 Model-based methods and my approach

Finally, there are so-called *model-based methods* where the "*researcher must make assumptions about the joint distribution of all variables in the model*", e.g. the *multiple imputation* [Pigott01, p. 14; Humph13]. Furthermore, he is also required to identify one of three *missingness mechanisms* too, e.g. *missing data at random* (MAR) as described first by Rubin in 1976 [Rub76; Rubin04]. Yet correctly identifying and applying such methods is not only outside of the scope of this thesis but even with these techniques, data cannot be seen as 'true representation' of country's accomplishment or failure. Rather they are intended "*to create an imputed dataset which maintains the overall variability in the population while preserving relationships with other variables*" [Wayman03, p. 4]. In addition to all of that, applying *multiple imputation* further requires deep understanding of the whole *missingness* concept. Not coincidently Schafer and Graham (2002) believe that "*researcher's time and effort are probably better spent building an intelligent model for the data rather than building a good model for the missingness*" [p. 25].
Due to that, I am not going to use any of the abovementioned mechanisms for handling missing data, but will return to a much simpler method. Namely, given my knowledge, I will choose and assign six values for Nigeria, Kenya, Jordan, Ghana, South Africa and the UAE. On the one hand, this is not a scientifically good approach as it brings a tangible source of uncertainty on my results. In the case of large dataset and/or very high rate of *missingness* it may be even impossible doing so. On the other hand, if data are not available and the reason is not related to other variables in my dataset – as it is the case here – it is very hard to impute them in a preferable ('desired') way even with the most advanced statistical models, simply because data do not exist.
As result, I decide to assign z-score of -2.1 to Nigeria, -1.9 to South Africa, -1.5 to Kenya,

-1 to Ghana, -0.5 to Jordan, and finally -0.2 to the UAE. To conclude the whole chapter, I would like to point out that the best solution to the problem of missing data is not to have a problem of missing data. However, this is often not possible and therefore in this chapter I showed several available techniques and finally assigned values to those countries considering my best (yet also limited) knowledge of their real situation.

2.4 Normalisation

One of problems researches often need to deal with are different scales of variables. In my case, I have units in percent (the youth unemployment), in the z-score (*The Learning Curve Index*) and four different score ranges (1-7 in GCI, 1-1518 in h-index, 0-100 in the economic freedom index and 0-1 in the Education Index). To avoid comparing apples with oranges, it is necessary to bring them to the common 'denominator' by using one of several normalization techniques [Hcci05, p. 85]. Naturally, different methods will provide different results. However, as Bostoc (2014) describes, all of them result into transformation of the original input domain into the new output range.

One of very common transformations is a standardised score, which converts "*indicators to a common scale with a mean of zero and standard deviation of one*" [Hcci05, p. 29]. It is calculated by applying an equation (6) where μ is the mean of population, σ is its standard deviation and x represents the actual value of a variable.

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

Resulting z-score gives a distance from the mean, where if it is positive then the country is above the mean, otherwise below the mean. The reader should note that outliers would have a "*greater effect on the composite indicator*" [p. 29]. Furthermore, as Nardo et al. (2005, p. 30) write, "*if the intention is to reward exceptional [good] behaviour*" z-score alone may not be the most appropriate choice for the index. This can be corrected by assigning different weights to indicators or for example by exclusion of "*the best and worst individual indicator scores from inclusion in the index*" [p. 29]. An example of using this technique for construction of composite indicators is many times aforementioned *The Learning Curve Index*.

Another commonly used method for building an index is a distance to the reference country. This approach benchmarks nations based on a distance to the frontier country, e.g. the leader or an average [p. 87]. It is used for example by the *Ease of Doing Business Index* developed by the World Bank [Wbg14]. Additionally, there is also a way to rank

countries given their overall position in the indicator. On the one hand, it is very simple and independent to outliers; on the contrary, there is a "*loss of information on absolute levels (…) [leading to] impossibility to draw any conclusion about difference[s] in performance[s]*" of countries [Toci05, p. 46].

The last normalisation technique, which I want to mention here (also used in the construction of my index), is called min-max normalisation. Using the equation (7) normalised values lie in the range between 0 (the lowest measure) and 1 (the highest one) [Hcci05, p. 87]. Same authors note once again that "*extreme values (…) could distort the transformed indicator*" [p. 30]. Moreover, it can "*widen the range of indicators lying within a small interval, increasing the effect on the composite indicator more than the z-score transformation*" [p. 30]. Conversely, an advantage is that the equation can be easily adapted "*that all the values are annealed within certain range*", not necessarily 0 and 1 or commonly in the case of z-score ± 3.5 [Sarman13]. (7)

$$I_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Due to that, I am going to use a range between 0 and 100 and as briefly mentioned in the chapter about the youth unemployment rate, it will be required to transform its polarity, i.e. from having the highest number being the worst to having the lowest number being the worst. For that matter, the equation for normalising all five variables is (8) where lower and upper range are 0 and 100 and *max(x)* and *min(x)* are minimal and maximal value of the range x_i[9] [Irritate11; Ttns13; Luxguc15]. The reader is now also encouraged to see the *Normalization / Scale.R* script producing newly scaled data. (8)

$$I_i = \frac{(upper\ range - lower\ range) * (x_i - min(x))}{max(x) - min(x)} + lower\ range$$

2.5 Multivariate analysis

> *It requires a very unusual mind to undertake the analysis of the obvious.*
> *– Alfred North Whitehead, 1925*

Multivariate analysis (MA) stem from the need to analyse data with more than just one variable, at once [Camo15, Ma09]. One of several goals of this analysis is to determine the structure of the data, e.g. in regard to their mutual relationships [Balem15]. As Nardo et al. (2005, p. 27) write this step is very "*helpful in assessing the suitability of the dataset and*

[9] For the unemployment rate, the equation (8) needs to swap functions of *min()* and *max()*. Consequently, former *min()* becomes new *max()* and former *max()* becomes new *min()*. See *Normalization / Scale.R.*

will provide an understanding of the implications of the methodological choices, e.g. weighting and aggregation". By applying the MA, same authors continue, I "*can increase both the accuracy and the interpretability of final results*" in addition to identifying "*redundancies among selected phenomena and (…) evaluat[ing] possible gaps in basic data*" [p. 50].

Three commonly used techniques for the MA – in the context of building composite indicators – are the *Principal Component Analysis* (PC/PCA), *Factor Analysis* (FA) and *Cluster Analysis* (CA). These methods will be further used in this thesis for analysing "*two dimensions of dataset: individual indicators [variables – PCA & FA] and countries [observations – CA]*" [p. 28]. Not only are such procedures useful for indices but all of them have very wide range of applications in other fields too, e.g. in the finance, biology, psychology, marketing and even urbanism.

Unfortunately, the PCA and FA are often confused in the literature. Both of them "*aim to reduce the dimensionality of a set of data, but the approaches taken to do so are different*" [Jolliffe02, p. 181]. Due to that, I will give an overview of the pair, first beginning with the PCA. Then, I am going to continue further with the factor and cluster analysis and finally conclude the whole chapter of the MA with a summary.

2.5.1 Principal component analysis

In the literature, one can find the term *exploratory analysis* concerning PCA, because as the first step in my data analysis I am going to explore my dataset [Butler14].

The aforementioned technique is able to find and identify "*patterns to reduce the dimensions of the dataset with minimal loss of information*" by preserving data's overall variability, i.e. meaningful information [Raschka14, Pcafa12]. In fact, sometimes "*much of the data's variation can often be accounted for by a small number of variables*" – principal components that I am interested in [Hcci05, p. 65; Weispca06]. These may reduce my original 6-dimnesional dataset by projecting new variables on n-dimensional subspace [Raschka14]. To sum up, I will carry out PCA to understand if I can use a smaller number of components to capture "*most of the variation between data*" [Coghlan15; Jolliffe02, p. 142].

However, before I can proceed further, it is usually suggested doing two steps. The first one is actually a requirement "*to prevent one variable having an undue influence*" on my PCs [Hcci05, p. 66]. Due to all variables having very different scales (for more see table 11),

I am going to use here my min-max normalised values that have been previously described toward the end of the chapter 2.4 [p. 66].

The second suggestion is to inspect how, and if at all, are my variables correlated. Nardo et al. (2005, p. 65) write that "*if the original variables are uncorrelated, then the [such] analysis is of no value*", as "*it is unlikely that they [indicators] share common*" components [Toci05, p. 56]. By producing the correlation matrix, see figure 7, I can demonstrate that e.g. the competitiveness and economic freedom have moderate to strong correlations with other variables in the dataset [Correla15]. Yet, this applies far less for the youth unemployment rate, where there is either a weak relationship with other indicators or even none at all. Due to that, I could consider omitting youth unemployment rate from other analyses in this section.

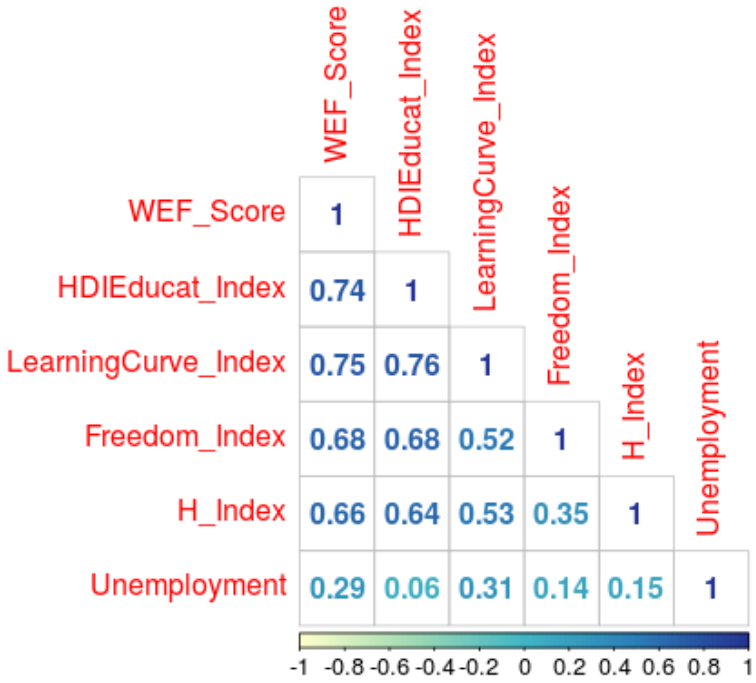

Figure 7 details correlation coef. between variables. For R script, see *MultivariateAnalysis / Correlation.R*.

In order to obtain a number of principal components[10] to retain, one may use many techniques. See Jolliffe (2002) discussing seven of them [p. 143ff.]. One, which is very common, is to plot a scree plot. This displays "*the successive eigenvalues, which drop sharply and then level off*" and it can suggest how many components could be retained, see figure 8 [Schwabj15]. Because of the elbow that occurs most significantly at the second

[10] Mathematically, principal components are "*found by calculating the eigenvectors and eigenvalues of the data covariance matrix*". In fact, PCs are "*eigenvector[s] with the highest eigenvalue*" in the whole dataset, i.e. variables that explain highest variation [Gillies15, p. 1; Smith02, p. 17].

component, I could retain two PCs [Pcatut12].

Another common methodology is to use *Kaiser's criterion* developed in the 1960 (suggesting two PCs as well) or just to apply "*some minimum amount of the total variance*" that needs to be kept [Kaiser60; Coghlan15; Hcci05, p. 72]. For example, I can decide to keep a number of PCs that explain cumulatively at least 90% of the variance. In that case, it leads me retaining 3 eigenvectors, instead of 2. From the table 3, the reader can see that the largest proportion of variability in the original data is captured indeed by the first three components. Together, they contain 90% of the total variation and due to low variance in the fourth, fifth and sixth component (each less than 6%), all three "*can be omitted from any further analysis*" [Nazim11, p. 3].

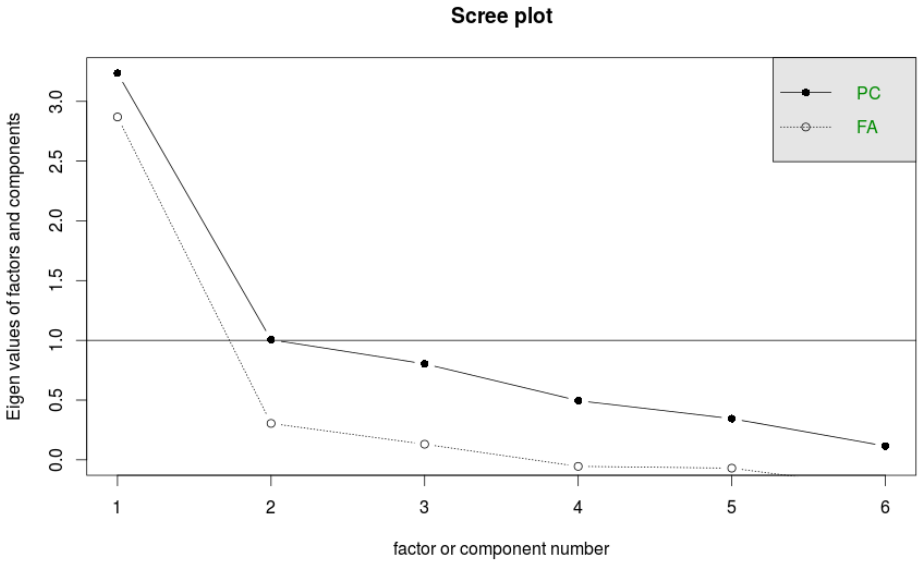

Figure 8 shows for how much variance accounts each component (black line) or a factor (dotted line) in the dataset. PCs "*with an eigenvalue of less than 1 account for less variance than did the original variable (which had a variance of 1), and so are of little use*" [Fultz12]. For R script, see *MultivariateAnalysis / PCAandFA.R*.

Additionally, I want to examine relationships between original variables and my principal components – specifically "*how different variables change in relation to each other and how they are associated*" [Hcci05, p. 28]. Table 4 allows me to interpret component loadings "*based on finding which variables are most strongly correlated*" with each PC [Weispca06]. Although the determination of strong correlation is largely subjective process, I will consider values above 0.6 threshold as a significant.

To consider first two principal components, the reader should now observe that they are strongly correlated with only one indicator, namely the h-index. Furthermore, with an exception of the abovementioned h-index and *The Learning Curve Index*, which is indeed

very close to 0.6 in fourth PC, all other variables are loaded only in one PC alone. This is also the case with the youth unemployment rate, which – due to its rather lower correlation with all five indicators – is now almost solely being responsible for the third component.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     27.107 13.118 9.94791 7.66408 5.28765 4.18017
Proportion of Variance  0.662  0.155 0.08916 0.05292 0.02519 0.01574
Cumulative Proportion   0.662  0.817 0.90615 0.95907 0.98426 1.00000
```

Table 3 shows a summary of PCA by providing information on standard deviation and (cumulative) proportion of variance.

```
Loadings:
                        PC1    PC2    PC3    PC4    PC5    PC6
Unemployment          0.087 -0.161 -0.868 -0.300 -0.348  0.044
Freedom_Index         0.216 -0.316  0.235 -0.699  0.151 -0.535
WEF_Score             0.319 -0.178 -0.032 -0.254  0.550  0.706
LearningCurve_Index   0.446 -0.518 -0.173  0.597  0.231 -0.305
HDIEducatIndex        0.421 -0.271  0.393 -0.004 -0.707  0.309
H_Index               0.684  0.707 -0.078 -0.009  0.025 -0.157
```

Table 4 presents principal component loadings for individual indicators, i.e. correlation coefficients between components and variables [Hcci05, p. 70]. Absolute values above 0.6 are in bold and "*the sign of its [PC] (…) loadings is arbitrary and meaningless*" [Signpca14, Jolliffe02].

2.5.2 Factor analysis

The goal of factor analysis is little bit different. Whereas PCA is "*purely mathematical transformation*" of data, the purpose of FA is to identify and "*confirm the latent factor*[11] *structure for a group of measured variables*" [Dontas10; Pcafa12]. These factors are unobserved, unmeasurable and assumed to "*cause the scores we observe*" [Pcafa12]. Namely, it is speculated "*that the data is based on the underlying factors of the model*" and by analysing such factors, it may help me to understand a pattern of variability in the dataset [Hcci05, p. 71].

Nardo et al. (2005, p. 71) write that the most common method to extract number of factors is to use PCA, which extracts "*first m principal components and (…) consider them as factors, neglecting those remaining*". This also suggested by Jolliffe (2002, p. 191) who writes that "*the use of PCs to find initial factor loadings (…) will often not be misleading in practice*". He continues by saying that "*for a given set of data, the number of factors*

[11] Latent factor is "*a variable that cannot be measured directly*", so that it may be "*inferred indirectly using other variables that are observed*" [Root10].

required for an adequate factor model is no larger – and may be strictly smaller – than the number of PCs required to account for most of the variation in the data" [p. 190]. This so-called *principal (components) factor analysis* (PFA) is "*most preferred in the development of composite indicators (…) as it has the virtue of simplicity and allows [later] for the construction of weights representing the information content of individual indicators*" [Hcci05, p. 71]. Given that, I decide to retain 2 factors, which also initially proposed the scree plot of PCA/FA in figure 8.

Next, the "*standard practice*" is "*to perform [a factor] rotation (…) to enhance the interpretability of the results*", i.e. I "*want to spread variability more evenly among factors*" [Hcci05, p. 72; Cososb05; Garrett06]. By doing that, the rotation of eigenvectors will attempt to "*simplify and clarify the data structure*" [Cososb05, Brown09]. As a result, each factor "*should have a few high loadings with the rest of the loadings being zero or close to zero*" [Brown09, p. 4f.].

To perform factor rotation there are again several possible functions, however "*the most common rotation method is the [orthogonal] varimax rotation*" [Hcci05, p. 72; Jolliffe02, p. 193]. Such a rotation results into factor loadings, which are "*correlations between each variable and the factor*" and they are shown in table 5 [Torrey10]. These values will be later used in the chapter 2.6 for construction of weights for my *Attractiveness Index*. From the table 5, the reader can now observe that two factors account for about 63% of cumulative variance. In contrast, the total variance explained by first two principal components was higher, namely about 82%, 90% with three PCs respectively.

The first factor is now dominated by the HDI's educational index, having highest loading of 0.997 across both of them [Weisner06; Jolliffe02, p. 195]. It is then followed by competitiveness and educational outcome. Additionally, in the two cases – for the youth unemployment rate in the first factor and HDI's educational index in the second one – R has omitted their loadings due to being very low (close to 0) and thus meaningless. However, despite the factor rotation with the objective to "*minimize the number of significant loadings on each row of the factor matrix*", the reader can still see that the competitiveness is again loaded significantly in the second factor, so that "*a meaningful interpretation (…) [of the variable] is not straightforward*" [Fatow15, p. 13f.; Hcci05, p. 72]. Very importantly, the reader should note that "*this is just the pattern that exists in the data and no causal inferences should be made from this interpretation. It does not tell us why this pattern exists [and] it could be very well that there are other essential factors that are not seen at work here*" [Weisner06].

To "*assess the adequacy of a factor model*", *communality* can be used. Namely, it tells "*how well the model is working for the individual variables*" and it "*can be interpreted as the proportion of variation (…) [in the indicator] explained*" by two factors [Weisner06]. In my particular case, this means that the youth unemployment has the smallest communality from all. Thus it can be implied that it doesn't "*move [well] with the other individual indicators in the dataset*" and hence it is not well represented by two common factors [Weisner06; Hcci05, p. 73]. On the other hand, "*the results suggest that the factor analysis does the best job of explaining variation in*" the competitiveness, educational output and HDI's educational index with *good* explanations in the economic freedom and h-index [Weisner06, Brown09].

Loadings:	Factor1	Factor2	Communality h^2
Unemployment		0.363	0.135
Freedom_Index	0.673	0.278	0.529
WEF_Score	0.738	**0.671**	0.995
LearningCurve_Index	0.761	0.281	0.658
HDIEducatIndex	**0.997**		0.995
Ranking_HIndex	0.634	0.282	0.482

	Factor1	Factor2
SS loadings	2.976	0.818
Proportion Var	0.496	0.136
Cumulative Var	0.496	0.632

Table 5 provides "*rotated factor loadings for individual indicators*" using varimax rotation (maximum-likelihood FA) [Hcci05, p. 73]. 'SS loadings' means sum of squared factor loadings. For the corresponding R script, see *MultivariateAnalysis / PCAandFA.R*.

2.5.3 Cluster analysis

In the literature, one can find that cluster analysis has not one single definition [Castro02]. This is in part due to many different types of clustering (e.g. (non-) hierarchical, fuzzy), clusters (e.g. graph-based) and algorithms used [Panmiv05, p. 4ff.]. In my situation, having no a prior knowledge of group memberships of 30 countries, I will be exploring their classification "*based on their similarity on different individual indicators*" [Hcci05, p. 28]. By doing that, I can discover some insights on their common characteristics [p. 75]. To summarize, "*clustering is an [unsupervised] learning technique which groups the similar objects into appropriate*" sets [Sarman13, p. 4].

As mentioned earlier, CA is divided into several clustering methods. However, here I want present only one clustering method, namely the hierarchical clustering by plotting its dendogram [Spenc11]. This doesn't require me to specify number of clusters in advance because it takes all observations and merges them into appropriate sets giving me their possible range, see figure 9 [Techni09]. In the picture, I highlight two biggest clusters in red rectangles, which show two groups of countries – labelled as 'developing' and

'advanced' economies. Nevertheless, the reader may also observe that there can be more groups available, depending on where one cuts the tree.

Finally, I want to present a line chart, see figure 10, which displays means of indicators for two abovementioned clusters [Hcci05, p. 79f.]. It shows that the biggest difference between 'developing' and 'advanced' economies, as one could assume, is indeed in their h-index and educational outcome. On the other hand, in the unemployment and economic freedom developing countries are at closest to their advanced peers.

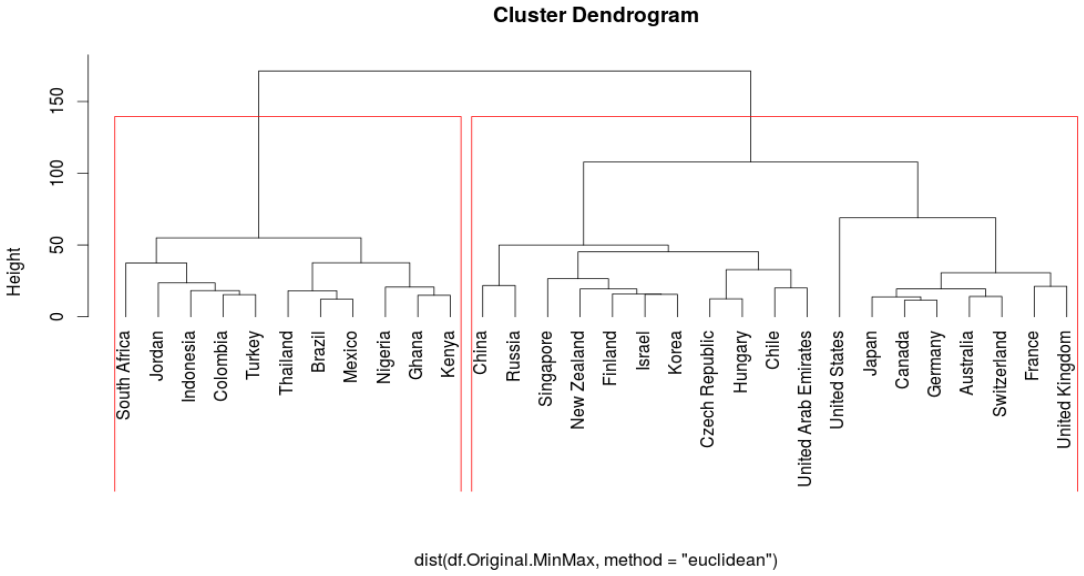

Figure 9 displays all possible clusters of countries. "*The height (…) indicates the distance between the objects*", i.e. it is "*a measure of closeness*" [Matlab15; Flom14]. Two rectangles have been drawn, highlighting a cluster group of South Africa – Kenya (11 'developing') and China – the UK (19 'advanced' nations). See *MultivariateAnalysis / ClusterAnalysis.R*.

2.5.4 The outcomes

At the beginning, I set out to explore six indicators, which are being used for the construction of my index. My goal was to inspect the structure of my data. Therefore, in this chapter I want to conclude the multivariate analysis with a summary of principal component, factor and cluster analyses. For the PCA, I may (re-)consider removing the youth unemployment rate from further analyses as its correlation is three times below or equal to ± 0.2, see figure 7. Having such a low correlation means that it will make its own principal component, see PC3 in table 4, which goes against the notion of reducing data dimensionality [Fultz12].

Moreover, my decision was to retain first three PCs as they contain 90% of meaningful information. Thus, the conclusion would be to reduce my original 6-dimenstional dataset

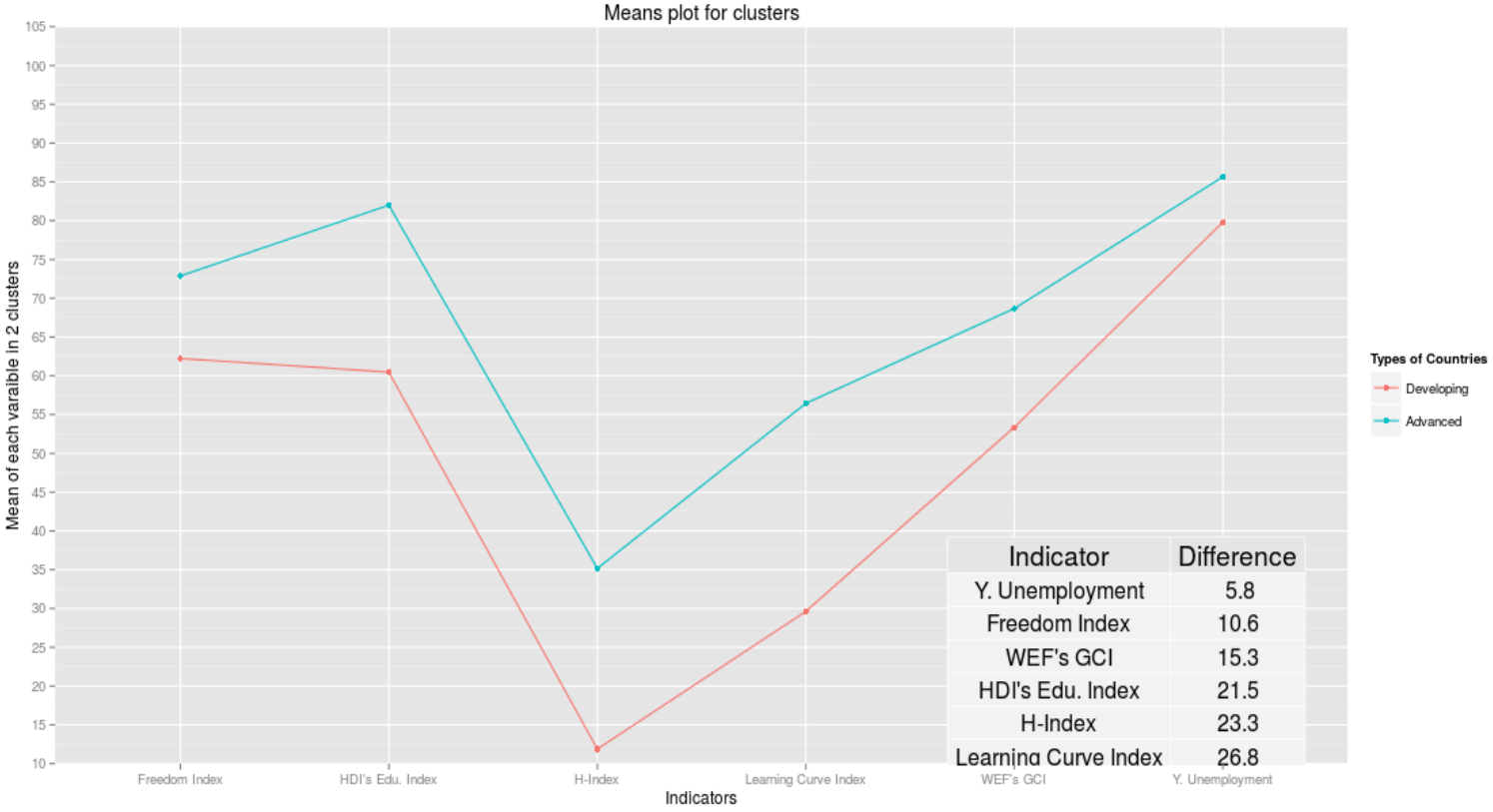

Figure 10 presents a mean of each individual indicator for both cluster (red and blue points), i.e. types of countries. The table further shows the difference between advanced and developing nations in each variable.

on a 3-dimensional subset. Even though, the reduction will be applied, almost 90% of total variability in data will be preserved. Nevertheless, it needs to be mentioned that due to having smaller sample size, "*caution is needed in making any inference about the PCs in the population*" of all 193 countries [Jolliffe02, p. 99].

In the case of factor analysis, I followed Jolliffe (2002, p. 190), namely given my prior retention of 3 PCs, I retained only two of them for the FA in order to have even a smaller number of factors to work with. This is followed by their *varimax* rotation for having easier interpretation of those factors and allowing me to answer a question of what constructs are being measured by them [Saspca07]. As already described, the first one is about nation's knowledge, competitiveness and educational outcomes whereas the second one is again about competitiveness but now also to a much smaller extend about unemployment as well. Lastly, I attempt to assign names "*to the pattern of factor loadings*" and hence call the first one 'Educational Performance' while the second one 'Monetary Success of Individuals' [Fatow15, p. 14].

Both PCA as well as FA are often confused because they "*summarise a set of individual*

indicators while preserving the maximum possible proportion of the total variation in the original dataset" [Hcci05, p. 28]. The biggest difference between two is that whereas PCA uses "*no explicit model*" of achieving dimensionality reduction, FA attempts to achieve that using a model relating variables to factors [Jolliffe02, p. 182]. The caveat is that they "*cannot be measured or observed, [thus] the existence of these hypothetical variables is open to question*" [Brihor11, p. 175]. Moreover, another difference is that by increasing number of principal components (e.g. from 2 to 3) values of those components remain unchanged. In that way, I simply increase the total variance explained by them. However, if the number of factors is increased in the same fashion, then consequently all factors "*can [and will] be substantial[ly] change[d]*" [Brihor11, p. 173]. On the contrary, one similarity of both techniques is that they are sensitive to the outliers, modifications and having sufficiently large sample size [Hcci05, p. 28].

Cluster analysis aims "*to reduce the dimensionality of a dataset by exploiting the [dis]similarities between*" countries and grouping them into several sets [Hcci05, p. 75; Ma09]. In the case of very popular hierarchical clustering, I highlighted 2 clusters that might be appropriate [Rodrigues14]. Furthermore, by calculating their *agglomerative coefficient* value of which is 0.89, I prove that there is a strong clustering structure available [Storti15]. Let me now also say that a disadvantage of CA is that it's only a descriptive technique, which always produces clusters [Hcci05, p. 29; Gupta02, p. 74].

Additionally, I want to note that the imputation described in the chapter 2.3 could have been done in a different way too. The missing value for the United Arab Emirates – being the only nation in the larger set – could correspond to the average of countries in that cluster[12]. The resulting measure would be 0.48[13] as opposed to the current -0.2, i.e. either being above the average in the educational output or slightly below. Comparable method using cluster information for the imputation has been suggested by Ishioka (2013) too [Cimi14].

To sum up the whole chapter, the "*application of multivariate statistics, including FA [and] CA, is something of an art, and it is certainly not as objective as most statistical methods*" [Toci05, p. 34]. However "*if it is thought of [MA] as a purely descriptive tool, with limitations that are understood, then it must take its place as one of the important steps*

[12] Two assumptions are made here: (I) I would begin with the CA first without using any values of the UAE (i.e. having only 29 rows) and (II) later I would assign this nation the average of the educational output from the larger cluster.

[13] For more see *MultivariateAnalysis / ClusterAnalysis.R*.

during the development of composite indicators" [p. 34]. In fact, the multivariate analysis has allowed me to look at potential relationships of my variables and as a matter of fact only results of the factor analysis will further used in this thesis.

2.6 Weighting and aggregation

In this chapter, I am going to conclude my construction phase of the attractiveness ranking for 30 nations through assigning weights to indicators and later their aggregation with normalised scores as described in the chapter 2.4.

This stage is also one of the most critical ones, because outcomes are heavily influenced here. The whole chapter is divided into two sections. In the first one, I am going to present several choices for weighting whereas in the second one, I will aggregate my normalized scores with newly calculated weights. Indeed, there are many weighting schemes as well as several aggregation methods possible, however in both cases only some of them will be discussed.

By comparing results of different normalisation, weighting and aggregation techniques, the modeller can get an interesting insight into methodological differences of applied choices and thus the range of possible positions for a given country [Luxguc15]. Yet for my index, I will analyse only three weighting procedures that are combined through a linear aggregation.

Similarly to previous chapters too, I want to stress that there is not the 'best' and "*'objective' way to determine weights and aggregation methods*" for the composite indicator [Hcci05, p. 35]. In fact, all presented methods have their own advantages and disadvantages, which need be taken into an account e.g. due to the theoretical foundation, when defining and building the composite [p. 103].

2.6.1 Weighting methods

Some weighting techniques really on empirical analyses (e.g. PCA/FA and data envelopment analysis), while others use participatory processes (e.g. budget allocation and public opinion) [Hcci05, p. 33]. Therefore, here I introduce the reader to a few methods, which only three of them will be further used in my index.

The simplest and the most transparent weighting method is the *equal weighting* (EW) [Sharand12, p. 10; Hcci05, p. 33]. In this, all variables are assigned same weight, hence implying that they "*are 'worth' the same in the composite*" [Hcci05, p. 33]. A disadvantage

of this method is that it may camouflage an absence of sound statistical basis, e.g. when "*there is insufficient knowledge of causal relationships*" or no consensus on the choice of weighting method has been reached [p. 33]. In my particular case, the EW means that each indicator weighs $\overline{0.166}$. However, as Nardo et al. (2005, p. 33) write, it can happen that two highly correlated indicators (e.g. the educational output and competitiveness value of which is 0.75) would be double counted in my index[14]. To correct it, I could either give a less weight to the pair or adjust number of indicators accordingly.

Another common technique is to apply a budget allocation process or a public opinion – both of which represent different participatory methods. In two cases, various shareholders – i.e. experts or a society – are asked to assign weights to indicators. In the BAP, "*experts are given a 'budget' of N points, to be distributed over a number of individual indicators, 'paying' more for those indicators whose importance they want to stress*" [p. 34]. On the one hand, it "*increase[s] the legitimacy of the composite*" due to opinion of experts, in addition to the ease of communication to the broader public and general marketability [p. 103]. On the other hand, their opinion "*could [also] reflect specific local conditions*", which do "*not measure the importance of each individual indicator but rather the urgency (…) for political intervention (…) of the (…) indicator concerned*" [p. 103]. Although the costs and complexity required to conduct such a survey can be high, it leads to have a very representative and objective evaluation of all taken indicators [Sharand12, p. 50].

Finally, there are several statistical approaches, which one of them is the *benefit of the doubt* (BOD) method [Hcci05, p. 34]. It stem from the application of data envelopment analysis (DEA) on composite indicators and in basic terms it takes "*best performing observations in each indicator to create a 'boundary' [the efficiency frontier] of feasible performance which is then used to measure the score of each observation*" [Sharand12, p. 19]. In fact, "*the weighting of each component will be determined uniquely for each observation in the data*" [p. 19]. Using that, the index is then "*defined as the ratio of a country's actual performance to its benchmark performance (…)*" and it will range "*between zero (worst possible performance) and 1 (the benchmark)*" [Hcci05, p. 94f.].

[14] Meaning that two indicators will be measuring a dimension that has weights $w_1 + w_2$ in the composite [Hcci05, p. 34].

2.6.1.1 Weighting based on factor analysis and my own preference

Another very common (statistical) weighting technique is to use outcomes of the principal component and factor analysis [Hcci05, p. 91f.; Nicscbo0, p. 15]. Here weighting is used to "*correct for overlapping information (…) between correlated indicators*" and is not a measure of their importance [Hcci05, p. 91]. Hence, such indicators' weights follow "*patterns of variance within data*" and can have "*a bias against possibly important variable that vary little*" [Sharand12, p. 49]. Nicoletti et al. (2000, p. 18) summarizes such approach, when they write "*each component (…) is weighted according to its contribution [proportion] to the overall variance in the data*".

The reader may remember that "*each factor reveals the set of indicators with which it has the strongest association*" [Hcci05, p. 91]. Due to the use of smallest number of factors that explain most of variability between data, "*the composite may no longer depend upon dimensionality of the dataset*" (i.e. all variables), "*but rather is based on the 'statistical' dimensions of the data*" [p. 91]. This process of assignment of weights is also the *least* transparent among four presented options, because it is "*statistically heavy*" and, as I show, it depends on many calculation steps. As a result, it is "*not being easily communicated to the public*" too [Sharand12, p. 49].

An advantage of this process is that the significance of indicators is data-based. This ensures that the largest weights are assigned to indicators that explain most variability in the data, independently of their (assumed or believed) importance [p. 18]. The same authors continue that this is a desirable property, because the focus is set on those indicators "*that are potential useful for explaining the cross-country variation*" in nation's attractiveness [p. 18]. Furthermore, it also solves 'double counting' problem because weights will differ for each indicator based on their variance [Toci05, p. 58].

On the other hand, a disadvantage of this method is that such weights are sensitive to data revisions and "*the presence of outliers, which may introduce a spurious variability in the data*" [see ch. 2.5.4; Nicscbo0, p. 18f.]. Moreover, a small sample size such as mine may influence results too; therefore, the caution is needed in extending any interpretation to the whole population of 193 countries. To sum up, factor analysis is going to be "*used to group individual indicators according to their degree of correlation*" [Hcci05, p. 34].

As described in the chapter 2.5.2, the first step was to decide about a number of factors to retain. In fact, the analysis identifies "*indicators which are most associated with different underlying (unobserved) factors*" [Nicscbo0, p. 18]. Here PCA is "*usually used to extract [a subset of those] factors*", as they "*account for the largest amount of the variance*" [Hcci05,

p. 91]. As written, I decided to retain two of them and further it was necessary to perform their (*varimax*) rotation in order to have each indicator loaded ideally "*on one of the retained factors*" [p. 92]. Thus, "*minimis[ing] the number of individual indicators*" that have high correlations on several factors[15] [p. 92]. Although the rotation has been performed, the reader could still observe that the competitiveness is loaded significantly in both factors, which doesn't makes its interpretation any clearer.

After the rotation, the next step now "*deals with construction of the weights from the matrix of [rotated] factor loadings*" [Hcci05, p. 92]. This is described by Nicoletti et al. (2000, p. 19), who applied weighting of individual indicators "*according to the proportion of [their total unit] variance that is explained by the factor [they are] associated to*". Hence, it is necessary to compute normalised squared factor loadings – ('Scaled SFL)', see table 6 [p. 18f.]. Then, highest ones from each factor are grouped "*into 'intermediate' composite indicators*" (highlighted in red) [Hcci05, p. 92]. As observed, the first composite contains economic freedom, educational outcome, h-index and HDI's educational index while the second one contains the youth unemployment rate and competitiveness.

The last step is to compute weights for both composites. Indeed, they have been weighted "*according to [their] relative contribution to the explanation of the overall variance of the two factors*", i.e. I take each value of 'SS loadings' from table 5 and divide it by the sum of both [Nicscbo0, p. 22]. Table 7 shows not only normalised squared factor loadings (column 'Domain Weight') but also corresponding factor weights. These two have been later multiplied ('Weight Score') and scaled to unity sum of 1 ('Norm. Weight Score'). As a result, values of the last column can now be used for correction of "*overlapping information*" between indicators [Hcci05, p. 91]. More importantly, however, they still should not be seen as a measure of indicator's importance [Toci05, p. 56].

The reader can now also compare FA weights with those equal ones, see table 8. In fact, only in two indicators – the youth unemployment rate and HDI's educational index – weights differ significantly. Hence, one can conclude that due to low/large correlations with most variables, these FA weights will attempt to correct it having potentially a powerful impact on the overall ranking. On the other side, in the educational output, h-index, economic freedom and competitiveness both weighting methods result into having similar results. Indeed, the above FA results also imply that the educational dimensions will be more significant in the index than the business & economic one.

[15] This was for example the case with the PCA, see table 4, where h-index has sizable loadings in two PCs.

The last weighting method I want to write about here is my own choice of weights, namely how I think indicators could have been weighted given my theoretical framework. Because in the FA weighting, only moderate weights have been calculated for the economic & business dimension, now I consider all indicators – with an exception of two educational ones – relatively more important, see table 8.

At this stage, variability of weights and their influence at the index will be further discussed in the chapter 2.7 Uncertainty and sensitivity analysis.

Indicator	Factor 1		Factor 2	
	Squared FL	**Scaled SFL**	**Squared FL**	**Scaled SQL**
Youth Unemployment	0.004	0.001	0.131	0.161
Economic Freedom Index	0.452	0.152	0.077	0.094
WEF's competitiveness	0.544	0.183	0.451	0.551
Learning Curve Index	0.579	0.194	0.079	0.097
HDI's Education Index	0.995	0.334	0	0
H-Index Ranking	0.402	0.135	0.080	0.097
Sum	**2.976**	**1**	**0.818**	**1**

Table 6 presents squared factor loadings (SFL), which have been scaled to unity sum of 1 (i.e. "*weights of variables in factor*") [Nicscbo0, p. 22]. Scaled SFL values, which are in red, correspond to the maximum from both columns. For detailed calculation, see *WeightingAggregation / WeAg.R*.

Indicator	Domain Weight	Factor Weight	Weight Score	Norm. Weight Score
Youth Unemployment	0.161	0.216	0.035	0.044
Economic Freedom Index	0.152	0.784	0.119	0.150
WEF's competitiveness	0.551	0.216	0.119	0.150
Learning Curve Index	0.194	0.784	0.153	0.192
HDI's Education Index	0.334	0.784	0.262	0.330
H-Index Ranking	0.135	0.784	0.106	0.134

Table 7 displays weights assigned to each indicator and factor in the index. These have been later multiplied and scaled to unity sum of 1. Adapted based on table from [Sharand12, p. 16].

Indicator	Equal weighting	Factor analysis	My choice
Youth Unemployment	0.166	0.044	0.140
Economic Freedom Index	0.166	0.150	0.170
WEF's competitiveness	0.166	0.150	0.230
Learning Curve Index	0.166	0.192	0.220
HDI's Education Index	0.166	0.330	0.130
H-Index Ranking	0.166	0.134	0.110

Table 8 details equal weighting, weighting based on FA and my personal consideration of indicators' significance. All columns sum up to 1.

2.6.2 Aggregation methods

Once weights have be established, it is necessary to aggregate them with the normalized scores. Therefore, the first aggregation technique, which I had already mentioned, is the *linear aggregation* where weights are simply multiplied with corresponding scores. This leads to the full (constant) compensability, "*which means that poor performances in one indicator can be fully compensated by good ones in another indicator*" [Luxguc15]. It rewards "*indicators proportionally to the weights*" and an equation (1) described by Freudenberg (2003) is used here to calculate each country's index, see page 4 [Hcci05, p. 35]. An example for Germany using the default weighting method would be 73.45[16]. (9)

$$Attractiveness\ Index_{Germany} = (92.2 * 0.044) + (73.78 * 0.150) + (74.80 * 0.150) + (55.91 * 0.192) + (88.4 * 0.330) + (53.66 * 0.134) = 73.45$$

Alternatively, there is also a *geometric aggregation*, which "*reward[s] (…) countries with higher scores*" in individual variables [p. 35]. Thus, the nation that has low value in one indicator will need much higher value in another indicator in order to improve its situation in the ranking. In that way, "*the geometric aggregation limits the compensability*", while at the same time 'punishes' the country for the unbalanced performance [Luxguc15]. This means that "*a country would have a greater incentive to address those sectors (…) with low scores (…) , as this would give it a better chance of improving its position in the ranking*" [Hcci05, p. 106]. This method is for example used during the construction of the *Human Development Index*, see chapter 2.2.3.

As stressed by Nardo et al. (2005, p. 35), neither with the linear nor with the geometric method weights do not represent indicators' 'importance'. "*This implies an inconsistency between how weights are conceived (usually measuring the importance of the associated variable) and the actual meaning when geometric or linear aggregations are used*" [p. 35]. Being able to interpret weights as 'importance' coefficients, I would be required to apply some *non-compensatory multi-criteria approaches* (MCA) such as one described by same authors (see p. 114ff.).

However, as already mentioned, for my composite I use only a linear aggregation, which doesn't 'penalize' nations for diverse performances in all indicators. Due to the choice of min-max normalisation technique with weights based on the factor analysis as my primary method, I can now compare my results with two other weighting schemes in table 13 too.

[16] During calculation only exact figures have been used, see *WeightingAggregation / WeAg.R*.

At this stage, I have finished building my *Attractiveness Index* and therefore, in the next chapter I want to analyse the outcomes and look back at my decisions.

2.7 Uncertainty and sensitivity analysis

At the beginning of the second part, I have mentioned that in each step I was going to take many subjective choices that could result into sending "*misleading, non-robust policy messages*" if the index is poorly constructed or misinterpreted [Toci05, p. 85]. Given that "*good modelling practice requires that the modeller provide[s] an evaluation of the confidence in the model*", now I want to assess my taken choices [Hcci05, p. 119]. In particular, the uncertainty of my results may arise from all five major stages of construction process of the composite. These stages were:

i. Selection of indicators
ii. Treatment of missing data
iii. Data normalisation process
iv. Conducting multivariate analysis
v. Choice of diverse weighting and aggregation schemes

For an investigation of possible sources of concern, there are two suggested techniques [Toci05, p. 85]. Both of them are usually conducted together in order to "*help to gauge the robustness of the composite indicator ranking, [and] to increase its transparency, [and] to identify which countries are favoured or weakened under certain assumptions and to help frame a debate around the index*" [Hcci05, p. 119]. The first one is an *uncertainty analysis* (UA), which analyses how uncertainty "*propagates through the structure of the CI and affects (…) [its] values*" [Toci05, p. 85]. It "*aims to quantify the overall uncertainty in country rankings as a result of*" mistrust in the input data and procedures [Hcci05, p. 119]. The second one is a *sensitivity analysis* (SA) that examines "*how much each individual source of uncertainty contributes to the output variance*", i.e. "*how 'sensitive' a model is to changes in the value of the parameters of the model and to changes in the structure of the model*" [Toci05, p. 85; Brlcmch01, p. 47].

Therefore, I first begin with a selection of basic data. Namely, I have selected only 6 variables and together with 30 observations, these numbers can be viewed as a very low (not even one-fifth of 193 UN recognized nations). For example, in order to perform PCA or FA there needs to be a sufficient number of cases. And although opinions differ widely – from 3:1 ratio up to having at least 200 measurements – with rather smaller numbers the

inference may not be very clear so that such interpretation cannot be easily extended to all economies worldwide [Hcci05, p. 68]. Besides, it is worth mentioning that five indicators are also indices with only one being a 'true raw' measurement.

Additionally, the user might question how indicators have been selected in the first place and if they (possibly) relate in any regard to my target group. Even though in the chapter 2.1 I have defined criteria for my variables that I believe have been met (e.g. only output and current ones being included in the composite), there is not a complete list of suitable indicators available due to the choice that is (largely) a subjective process. Yet on the other hand, indicators can be poorly chosen and therefore it is necessary that they clearly follow modeller's theoretical framework, which I believe they have.

In the next chapter, I dealt with the imputation of missing observations in *The Learning Curve Index*. After an introduction of several possible approaches, I have decided to assign values through my (limited) knowledge of countries' real situations. Although I didn't conduct any model-based methods such as *multiple imputation*, one of their clear advantages – over the other methods introduced – is that they can provide unbiased estimates by using all available data and thus preserving their overall variability. Indeed, "*the multiple imputation (…) provides several values for each missing value, [hence] it can more effectively represent the uncertainty*" [p. 27]. However, these techniques are also complex and further require fulfilling specific *missingness* conditions for my data, see Azur et al. (2011, p. 41) discussing some of them.

During feature scaling, I used a very common min-max transformation, which has an additional advantage of being compatible with all weighting and aggregation schemes. This applies far less e.g. for the standardized score that can be used – according to Nardo et al. (2005) – with only some weighting and some aggregation methods [p. 33 & 120]. Such normalized values have been used throughout the rest of my thesis.

Then, I continued with the multivariate analysis as correctly conducting it and interpreting its outcomes brings me further ahead toward the composite. During the PCA, I have retained three principal components, therefore reducing my original 6-dimnesional dataset only on 3-dimensional subspace. For the factor analysis, I have followed Jolliffe (2002, p. 190f.) and further decreased number of factors to two, labelling them as 'Educational Performance' and 'Monetary Success of Individuals'. In the third section, in the cluster analysis, I performed only the hierarchical clustering by omitting the popular *k-means* one. The resulting two largest sets have suited well too due to very simple distinction of advanced and developing economies. Not surprisingly, two-thirds of countries are

advanced, thus being in the larger cluster.
Lastly, there are several weighting and at least three aggregation techniques, yet, only some have been described in this thesis. This part has been also crucial for my composite due to the necessity of having correct calculation of FA weights, which are later aggregated with normalised scores. My choice of applying three weighting schemes (other two do not require any calculation) was driven by the fact that I wanted to compare what an impact have distinct weights on results that use same normalisation (the min-max) and aggregation (the linear) method.

2.7.1 Analysis of final results

Additionally in this chapter, I want to interpret figure 12 and figure 13, starting with the first one that compares three different outcomes.
On the x-axis, the reader sees all 30 countries whereas on the y-axis those nations are now assigned a position in the ranking. The chart shows three coloured lines that go through white scatter points. Not only lines interfere with each other but also symbolize nations' ranks using distinct weighting methods – the red colour for the equal weighting (EW), the blue one is about my personal preference (MC) and the straight green line for the factor analysis weights (FA), the default one.
As one can observe, only eight states have no matter which weighting scheme is chosen always the same positions in the ranking. The same number of nations has also different positions in all three indices, namely they capture their places up to five distances away from the green line. On the contrary, the vast majority of countries differ in only two weighting schemes. Therefore, due to having a moderate variability of results (especially at the beginning and toward the end of ranking), I can conclude that the choice of weighting scheme – in my exercise – has only modest impact and thus the influence.
As a matter of fact, the blue (MC) and red (EW) line move often very close to each other, see additionally figure 15 too. At the beginning, the reader rather sees both of them being at different positions for each country while already starting in the middle two lines converge and continue together toward the end (with small exceptions of the UAE, Brazil and Mexico). This is also expected, as equal weights with 'my choice' are closest to each other, see table 8. What, however, has caught my attention are distances of some countries based on FA weighting, e.g. Canada (71.95), Switzerland (71.55) and Japan (71.51) or South Africa (49.51) and Colombia (49.33) all of which are within the interval of 0.5.

Similarly, this applies to other two weighting schemes as well.

Let me now specifically mention several countries starting first with China. While achieving its 18th position – 15th in the EW and MC respectively – the reader should embrace the fact that its success (or failure) can be disputed. Namely, some statistics are not available for the whole nation (e.g. in *The Learning Curve Index* a proxy of the Hong-Kong's results has been taken), while others must be viewed very carefully as China doesn't "*use the internationally accepted metrics*" (e.g. in the case of youth unemployment rate) [Esptse14]. Even though I have considered data only for the mainland PRC, one could also argue that its position should be rather seen for a specific region or a part of China, instead of for the whole country. Not coincidently, territories such as the Hong-Kong SAR, Taiwan and Shanghai are very often presented separately of mainland China, e.g. seen in the *WEF's Competitiveness Report* or in *The Heritage's Economic Freedom*.

A similar situation concerning distinctly strong and weak parts of one country can apply to others as well, e.g. to Germany. Although it has reliable statistics, there are still differences in the economic (and educational) situation between the West and the East part of the nation [Hock14, Kmk13, p. 11ff.]. Furthermore, concerning Singapore, South Korea and Finland, I am a little bit surprised that they have not achieved better overall results given their quality of education. In fact, all of three are known to have exceptionally good schooling systems in Asia and Europe, respectively.

Finally, who are the most attractive economies? The top three are the USA (79.93), the UK (73.99) and Germany (73.45). Additionally, in the first top ten, countries are almost equally represented across four regions – six coming from the European and Asian continent with four nations from the (North) America and Oceania as well. From my sample of 30 nations, these results do not come as a surprise to me. In fact, the United States together with the United Kingdom – e.g. according to various university rankings – have the best universities in the world. By having them exceptionally well-funded, they are able to showcase high-quality results in most scientific fields (and thus attracting the best talent from oversees). Additionally, due to the high quality of life (even though it was not included in my index) and being one of the most advanced economies in the world, these countries confirm the usual perception of being the most attractive to young adults[17].

The last picture I want to present here it the decomposition of two dimensions and their

[17] Which in the case of the USA, at least for the last several years, is further proved by exceeding annual limit of 65 000 H-1B visas for skilled foreign workers [Wsjm15].

impact on each nation's position in the ranking, see figure 13. Such an analysis further extends interpretation of my results as each sub-component contributes differently to nations' attractiveness index [Hcci05, p. 37]. In fact, given my outcomes of the factor analysis weighting in the chapter 2.6.1.1, the educational dimension clearly prevails over the business & economic one.

For example, in the case of Russia and the United States, the reader observes that their educational score is responsible for 2/3 of their overall index; see *UnserSensi / BackToDetails.R* for detailed percentages. On the other side of the ranking, for Nigeria its score is the most equally distributed on both dimensions, having the highest representation of business & economic measure across all other nations (50.77%). This is due to its educational dimension, which although has the lowest results in individual indicators, it is still valued very significantly in the overall index. On the contrary, for the second subgroup, which achieves far better results, it is valued also far less. Hence, the balanced distribution between both of them.

Furthermore, as already briefly noted at page 37f., when comparing results based on the equal weighting, they would be just the opposite, see figure 16. Namely, the business & economic subgroup now prevails with the USA being the most equally distributed on both dimensions while for Nigeria, out of 72.5%, its score is instantly (and naturally) attributed to the newly prevailing measure. Even though distribution percentages of two subgroups change radically, they have only moderate effect on countries that are being in the first top ten as they just slightly adjust their ranks and consequently index values too. Similarly, this applies to my preference of weights as well.

Lastly, I want to put my results into the relationship with nations' GDP and investigate if they relate to each other. By plotting IMF's projections for this year's GDP per capita based at the purchasing-power-parity with my results, I can examine its correlation, see figure 14. Undeniably, the value of 0.79 shows that there is a strong association between both of them, yet also with several notable outliers. The United Arab Emirates and Singapore have very high level of GDP, however, for them it doesn't automatically translates into being attractive as they 'only' achieve 19th, 10th place respectively. On the contrary, the USA – only achieving fourth highest GDP in my sample – it is the most attractive nation. The conclusion could be that the higher GDP per capita economies have, the more attractive there are. However, here it is also very important to mention that *correlation does not imply causation*, i.e. that there could be other factors affecting this relationship as well.

3 Technical notes

In this chapter, I want to describe some technical aspects of my work. As noted earlier, for all statistical analysis and development of my index I was going to use R statistical language [Rcore08]. Below I provide my version of the operating system, R and an excerpt of loaded packages[18] that together with the table 12 present further a non-extensive list of libraries, which may be required to reproduce my source code. The reader is strongly encouraged to run *Util / Install_packages.R* script for their installation, instead of installing them later individually. In the case of unsuccessful installation, the user is also encouraged to load a binary *Util / Index.RData* file to see my data. My preferred choice of IDE has been RStudio Desktop, developed by RStudio Inc[19].

During my work, not only I have profoundly improved my R programming skills, but also saw its drawbacks, which two of them I want to highlight here. First, in my personal opinion, one of disadvantages of the language is that it requires installing huge amount of packages being able to manipulate and further work with data. At the end of chapter 2.7, I

```
> sessionInfo()
R version 3.2.0 (2015-04-16)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.10

locale:
 [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C                  LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8        LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8          LC_NAME=en_US.UTF-8           LC_ADDRESS=en_US.UTF-8
[10] LC_TELEPHONE=en_US.UTF-8      LC_MEASUREMENT=en_US.UTF-8    LC_IDENTIFICATION=en_US.UTF-8

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gridExtra_0.9.1 ggthemes_2.1.2  clustrd_0.1.2   irlba_1.0.3     e1071_1.6-4
 [6] corpcor_1.6.7   NbClust_3.0     cluster_2.0.1   psych_1.5.1     scales_0.2.4


(many other packages have been omitted)


loaded via a namespace (and not attached):
 [1] Rcpp_0.11.6         bitops_1.0-6        class_7.3-12        tools_3.2.0
 [5] digest_0.6.8        gtable_0.1.2        lattice_0.20-31     Matrix_1.2-0
 [9] DBI_0.3.1           parallel_3.2.0      proto_0.3-10        XLConnectJars_0.2-9

(…)
```

[18] "*Packages are collections of R functions, data, and compiled code in a well-defined format. The directory where packages are stored is called the library*" [Kabac14]. The goal of packages is to extend functionality of base the language.

[19] http://www.rstudio.com/products/RStudio/

had over 100 different libraries installed on my system (not counting those, which come with the *base* the installation of R). In fact, the *base* language is very small, and therefore it relies on over 6000 packages from CRAN – *The Comprehensive R Archive Network*[20]. On the one hand, this is clearly an advantage, on the other hand R needs to be significantly extended in order of having 'more-or-less' basic functions such as plotting *nice* cluster plot or easily merging two tables.

The second disadvantage of the language I want to write about is the insufficient official documentation, where it is very often necessary to rely on third-party resources[21]. Having a second choice, most probably I would be using another popular language called Python[22] because of its better documentation and the fact that it is a multipurpose programming language, i.e. designed not only for statisticians or for programmers.

Furthermore, in this chapter I want to present two code snippets, which have been used in my thesis. The goal of that should be providing the reader an insight into manipulation and visualization of data in R. Indeed, the first snippet deals with the question of how a *data frame* [23] has been created so that it contains 180 (non-normalised) values in 30 rows and 6 columns, for more see *RawData / DataFrame.R*.

At first, the R script on the next page loads two libraries – *plyr* and *dplyr* – in the first two lines making their functions available to use [plyr11, dplyr14]. Then, starting at line #3, the script runs another six R files using *source* function, which reads, parses and executes them. As the output, each of them produces one data frame containing values of the corresponding indicator. Such R script, which is executed using *source* function, is unique as it needs to process data in different ways. This has been already briefly described in section 2.2 where some indicators use *Quandl* R library while others read Excel worksheet or a text file. Not only resulting data frames have all different number of rows and columns but also each data source they stem from had to be further manipulated in its own way, e.g. in regard to the data cleansing (the ETL process).

Once all those six data frames have been executed, the former script continues processing the rest of the file. The idea is now to merge (join) all those six data frames in a new one. Therefore, in the line #9, first I merged two data frames containing data of youth unemployment rate ('unemplo') and economic freedom ('freedom'), using dplyr::left_join

[20] http://cran.r-project.org/web/packages/index.html
[21] E.g. https://stackoverflow.com/questions/tagged/r; https://google.com etc.
[22] https://www.python.org/
[23] *Data frame* is a fundamental data structure in R, in a form of a table [Rdatafr].

```
#1      library("plyr")
        library("dplyr")

#3      source("1_RawData/Unemplo.R")
        source("1_RawData/FreedomIndex.R")
        source("1_RawData/WEF.R")
        source("1_RawData/LearningCurve.R")
        source("1_RawData/EUI.R")
        source("1_RawData/HIndex.R")

#9      df.Original <- dplyr::left_join(unemplo, freedom, by = "Country")
#10     df.Original <- dplyr::left_join(df.Original, wef, by = "Country")
#11     df.Original <- dplyr::left_join(df.Original, hdi, by = "Country")
#12     df.Original <- dplyr::left_join(df.Original, completionRate, by = "Country")
#13     df.Original <- dplyr::left_join(df.Original, hindex, by = "Country")

#14     df.Original <- subset(df.Original, select = c(Country, Unemployment_NonScaled,
       Freedom_Index_NonScaled, WEF_Score_NonScaled, LearningCurve_Index, HDIEducatIndex,
       H_Index_NonScaled))

#15     df.Original <- plyr::arrange(df.Original, df.Original$Country)

#16     sapply(df.Original, class)
```

function of *dplyr* package. This has merged them in a new data frame called 'df.Original' in a way that returns all rows and columns matched by the country. Now, the abovementioned data frame stores 30 countries with all columns from both dataset. Although I could also only use left_join without explicit reference to its namespace (dplyr::), here I haven't done so because of potential problems that could happen due to similar functions being in both packages[24] (and sometimes better readability too).

This process continues until line #13 as I always merged the prior 'df.Original' with the one containing data of each of four indicators. Due to joining many unnecessary columns, in the line #14, I used a subset function to select only those that are relevant. In the last two steps, I sorted my countries lexicographically using plyr::arrange function, while in the second step I check if all variables are type of numerical (continuous) in order to perform mathematical operations with them. Of course, except those for the country, which is a character type.

The second code snippet now deals with the visualisation of results as seen in the figure 12, see *UnserSensi / US_Graphs.R*. Therefore, first, I needed to load two libraries again – this time *reshape2* and *ggplot2* [reshape2, Ggplot11]. Starting at line #2, it was necessary to make sure that data from all five R scripts have been executed using *source* function.

[24] For example, both libraries contain function arrange, yet their implementations are slightly different. Thus, without explicit reference to its namespace, R could be confused from which package to use.

However, due to having those data already stored in R's environment, in RAM, I didn't load them again and thus commented appropriate lines out.
In lines #3-5, I had to prepare my datasets in order of plotting a line chart with them. Therefore, I first extracted row names using rownames function and stored them in their own column, see figure 11. Without doing this step – in this case – it would be an undesirable property to have data in such format and because of that, I wouldn't be able to work with such tables later.

	Value	RankMM.FA	Country
United States	79.93082	1	United States
United Kingdom	73.98922	2	United Kingdom
Germany	73.45203	3	Germany

Figure 11 presents an example, where countries are stored in rows (far left column without the title) and the result of lines #3-5 (red rectangle).

```
#1      library("reshape2"); library("ggplot2")

#2      #source("1_RawData/DataFrame.R");          #source("2_Imputation/Imputation.R");
        #source("4_Normalization/Scale.R");        #source("3_MultivariateAnalysis/PCAandFA.R");
        #source("5_WeightingAggregation/WeAg.R")

#3      df.Original.MM.FA$Country <- rownames(df.Original.MM.FA)
#4      df.Original.MM.EW$Country <- rownames(df.Original.MM.EW)
#5      df.Original.MM.MyChoice$Country <- rownames(df.Original.MM.MyChoice)

#6      meltingOriginal.MM.FA.Subset <- melt(df.Original.MM.FA[, c("Country", "RankMM.FA")], id =
       "Country")
#7      meltingOriginal.MM.EW.Subset <- melt(df.Original.MM.EW[, c("Country", "RankMM.EW")], id =
       "Country")
#8      meltingOriginal.MM.MC.Subset <- melt(df.Original.MM.MyChoice[, c("Country", "RankMM.MC")], id =
       "Country")

#9      me1 <- ggplot()
#10     me1 <- me1 + geom_line(data=meltingOriginal.MM.FA.Subset, aes(reorder(Country, value), value,
       colour=variable, group = variable))
#11     me1 <- me1 + geom_point(data=meltingOriginal.MM.FA.Subset, aes(reorder(Country, value), value,
       colour=variable, group = variable), size = 4, shape=21, fill="white")
#12     me1 <- me1 + geom_line(data=meltingOriginal.MM.EW.Subset, aes(reorder(Country, value), value,
       colour=variable, group = variable))
#13     me1 <- me1 + geom_point(data=meltingOriginal.MM.EW.Subset, aes(reorder(Country, value), value,
       colour=variable, group = variable), size = 4, shape=21, fill="white")
#14     me1 <- me1 + geom_line(data=meltingOriginal.MM.MC.Subset, aes(reorder(Country, value), value,
       colour=variable, group = variable))
#15     me1 <- me1 + geom_point(data=meltingOriginal.MM.MC.Subset, aes(reorder(Country, value), value,
       colour=variable, group = variable), size = 4, shape=21, fill="white")

#16     me1 <- me1 + coord_cartesian(ylim = c(0, 35)) + scale_y_continuous(breaks = seq(0, 35, 1))
#17     me1 <- me1 + ggtitle("Small title") + ylab("Positions") + xlab("Countries") + labs(color =
       "We./No. methods")
#18     me1
```

Once this step has been accomplished, with lines #6-8, I proceeded to the most important part of plotting my desired graphic. Namely, I needed to reshape all three data frames into the long format[25]. This is also the reason why I loaded *reshape2* library, as in fact its function melt and the right parameter that accomplishes my goal. As a result, I get three separate tables containing each three columns with the first one for names of countries, second one for the type of rankings (column called 'variable'; e.g. EW) and last one for positions of those nations in rankings (column called 'value'; e.g. 6th place).

Next, to create figure 12, I used the *ggplot2* library, which has been developed as a *grammar of graphics* with the intention of making all kinds of different plots in R easier and nicer. The idea is now to construct a line chart having three lines, which one of them is a strait one. In basic terms, I needed to create three separate lines and combine them in one picture.

For that, in the line #9, first I initialized and stored *ggplot* object using ggplot function in new variable – 'me1'. Then, for each of three datasets I used two *ggplot2* functions – the geom_line and geom_point. Both of them take one data frame (e.g. *meltingOriginal.MM.FA.Subset*) and order its nations by their value in the ranking, from smallest to highest. All of that by creating a connected line and applying a colour to it. Moreover, also forming white scatter points of size '4', which is done by using the second function taking same arguments plus additional ones. This repeats for two other datasets as well (#11-15). In the line #16 for a better resolution, I set limits of the y-axis and its breaks, ranging from 0 to 35 moving up by 1, and finally give a description of my chart (#17) too. Once the code snippet on the previous page has been executed, line #18 makes sure that the graphic will be plotted automatically.

[25] See Grace-Martin, Karen. "The Wide and Long Data Format for Repeated Measures Data." *The Analysis Factor*. n.d. Web. 14 July 2015. http://www.theanalysisfactor.com/wide-and-long-data/

4 Conclusion

In this thesis, my quest was to create an attractiveness index for a sample of 30 global economies. In the first part of my thesis I have introduced the reader to the topic of composite indicators first by defining the term, later presenting two examples from distinct fields and finally discussing advantages and disadvantages of indices.

In the main part, I have defined my framework with a target group and gathered necessary data about 6 indicators forming two dimensions. By performing different analyses on normalized values I was able to create my ranking using linear aggregation in combination with three weighting schemes, having factor analysis weights as default ones. This has allowed me to compare various outcomes, concluding that even though there are clear methodological differences between various weighting methods, specifically in my exercise, they in fact only show to have a moderate impact on the attractiveness of each country. The reason is their relatively small variation coming from table 8. Yet on the other hand, the reader should note that there *are* differences and it is proven e.g. by the majority of countries having distinct positions in at least two rankings.

Least, but not the last, I have showed two R code snippets which I have used during the construction process of my index.

Finally, concerning my attractiveness – based on min-max normalization with weights of the factor analysis – not surprisingly, the United States of America has achieved its number one position, followed by countries such the UK, Germany or Australia. Indeed, in the top five, countries from the European and American region are equally represented. Furthermore, an interesting aspect is that nations' performance inside of regions can widely differ. Specifically this can be seen in Asia, where e.g. China and Russia (with Indonesia being fourth from last) are in the second half of the rankings whereas their (geographical) neighbors such as Japan or South Korea clearly gain far better results. Similarly in the American region too, where the USA and Canada achieve front positions, yet Mexico with Colombia struggle and are in the last fourth.

This compares strongly for group of countries in Africa and the Middle-East, as they do not seem to differ so widely. With an exception of Israel, all of them are ranked in the second half of my index, with most of African nations being also the least attractive. Thus only confirming that they are still developing itself and have a lot of to improve in years and decades ahead.

Overall, my ranking doesn't present any new 'groundbreaking' information, which

however was also not its mission. Rather it confirms that nations from America, Europe and Asia are best positioned to attract young talent. All of that, however, also needs to be taken carefully as the attractiveness is a very broad term, which not only is tough to measure and describe but the term itself depends on each person's definition. As I have only targeted a group of young adults, being between 20 and 40 years old, I made a big assumption that for them it is *all* about the knowledge economy and business environment that can (yet, not necessarily) make the country very attractive [Ee06].

To conclude my thesis, I would like to mention that any multidimensional concept of reality (which the *attractiveness* is part of) when measured by a single number, can be seen problematically. This is due to many different input aspects, both practical (e.g. calculating weights) as well as theoretical ones (e.g. selection of my variables) that influence each outcome. Hence, the user of such index is required to be aware of composite's drawbacks and understand – quoting Luzzati and Gucciardi (2015) – that "*composites respond to our need of simplifying the representation of reality (…) [and] they also require accepting the idea that reality can be condensed into a single figure*" [Luxguc15].

With that I would like to conclude my quest of constructing my *Attractiveness Index* for a sample of 30 global economies.

Dimension	Dataset (Indicator)	Unit measured / Definition	Year of avail. data, frequency	Source
Business and Economy	Index of Economic Freedom (IEF)	Country's score of economic freedom	2015 edition, annual	The Heritage Foundation & The Wall Street Journal
	Youth Unemployment Rate	Percentage of 15-24 years old adults without having a job (of the total labour force)	2012-2013, annual	The International Labour Organisation
	Global Competiveness Index	Country's competitiveness score	2014/2015 edition, annual	World Economic Forum
Education	Learning Curve Index	Country's index (z-score)	2014, differs	Pearson & Economist Intelligence Unit
	HDI's Educational Index	Country's index of being knowledgeable	2014, annual	The United Nations Development Program
	H-Index	Country's number of articles (h) receiving h citations	1993-2013, differs	SCImago

Table 9 describes data for my index.

Region	Asia	America	Europe	Africa	Middle East	Oceania
	Singapore	USA	Germany	Kenya	Israel	New Zealand
	Russia	Canada	Czech Republic	Nigeria	Jordan	Australia
	(Mainland) China	Mexico	UK	South Africa	United Arab Emirates	
	South Korea	Brazil	Switzerland	Ghana	Turkey	
	Japan	Chile	Finland			
	Thailand	Colombia	France			
	Indonesia		Hungary			

Table 10 shows countries divided into geographic regions. For their role in the G20 & OECD, the reader can see the corresponding Excel file.

Variables	Possible Values	Sample Mean	Standard Deviation	Missing Values (NAs)
Ind. of Economic Freedom	0-100	68.97	9.56	-
Youth Unemployment	0-100%	16.50%	9.66%	-
Glob. Competitiveness Ind.	1-7	4.784	0.603	-
Learning Curve Ind.	[-∞, ∞] (limited to ±3.5)	0.003	0.961	6 (see ch. 2.3)
Education Index (HDI)	0-1	0.741	0.131	-
H-index	1-1518	404.9	315.04	-

Table 11 details variables used for my index. The calculation of an arithmetic mean and sd for *The Learning Curve Index* is based on my complete sample of 24 countries.

Building Data Frame	Multivariate Analysis	General Purpose Libraries
Quandl [Quandl14]	fpc [fpc14]	rJava [rJava13]
rvest [rvest]	mclust [mclust12]	ggplot2 [Ggplot11]
dplyr [dplyr14]	cluster [cluster]	reshape2 [reshape2]
xlsx [xlsx14]	NbClust [NbClust]	GPArotation [GPArot]
plyr [plyr11]	clustrd [clustrd]	scales [scale]
stringr [stringr12]	ellipse [ellipse13]	psych [psych15]
	corrplot [corrplot]	rgl [rgl]

Table 12 presents a list of libraries, which can be needed in order to reproduce (some of) my results, see *Util / Install_packages.R*.

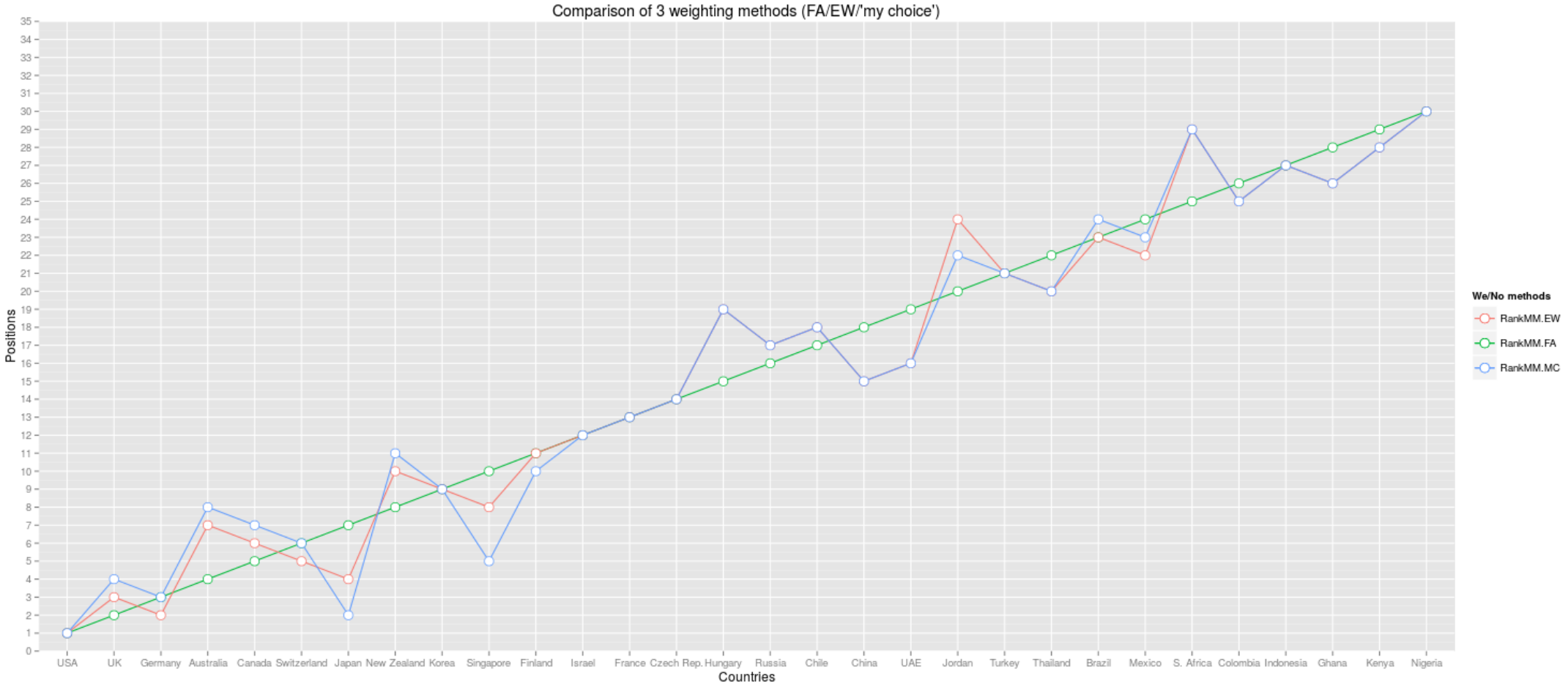

Figure 12 shows countries' positions in different rankings, for more see *UnserSensi / US_Graphs.R*.

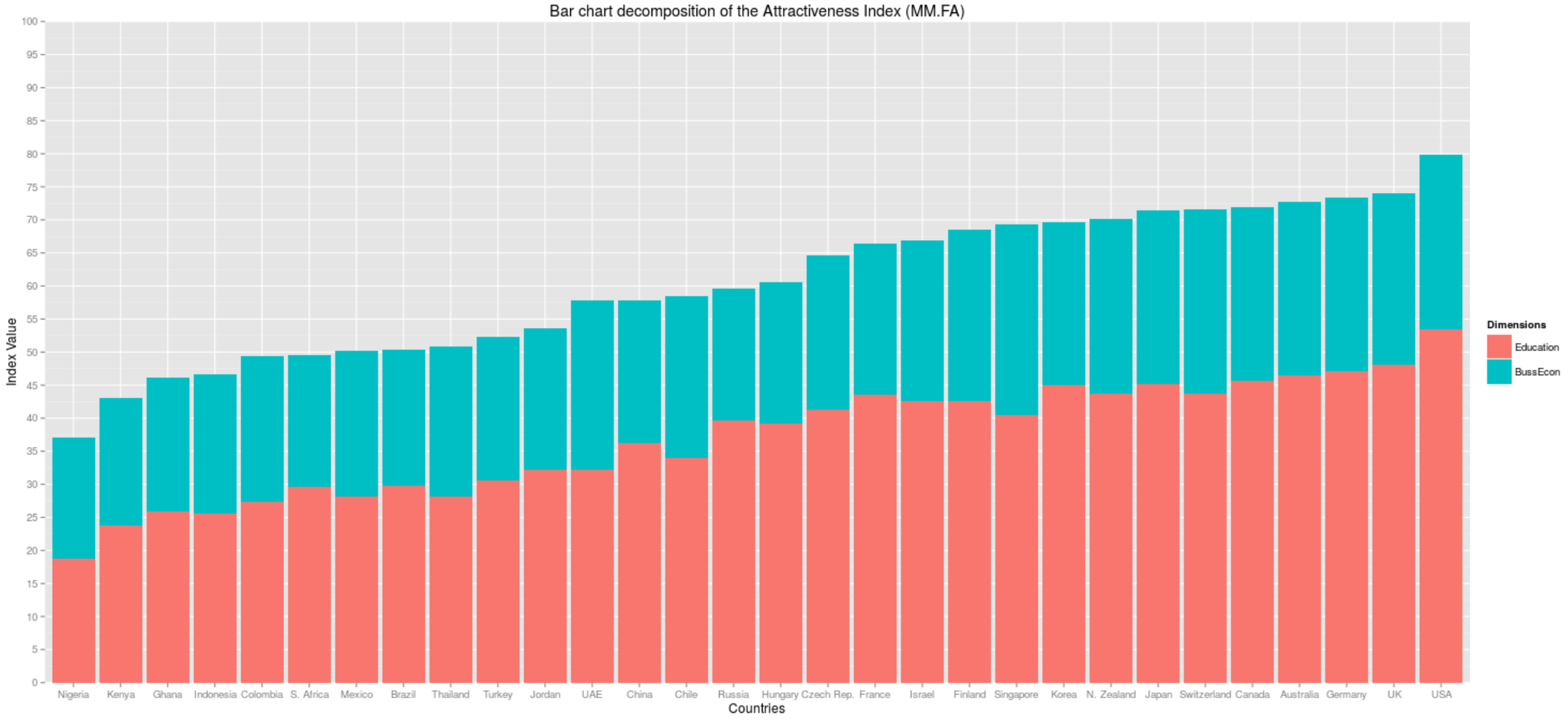

Figure 13 displays a bar chart decomposition of two dimensions of my composite, see *UnserSensi / BackToDetails.R* [Hcci05, p. 38].

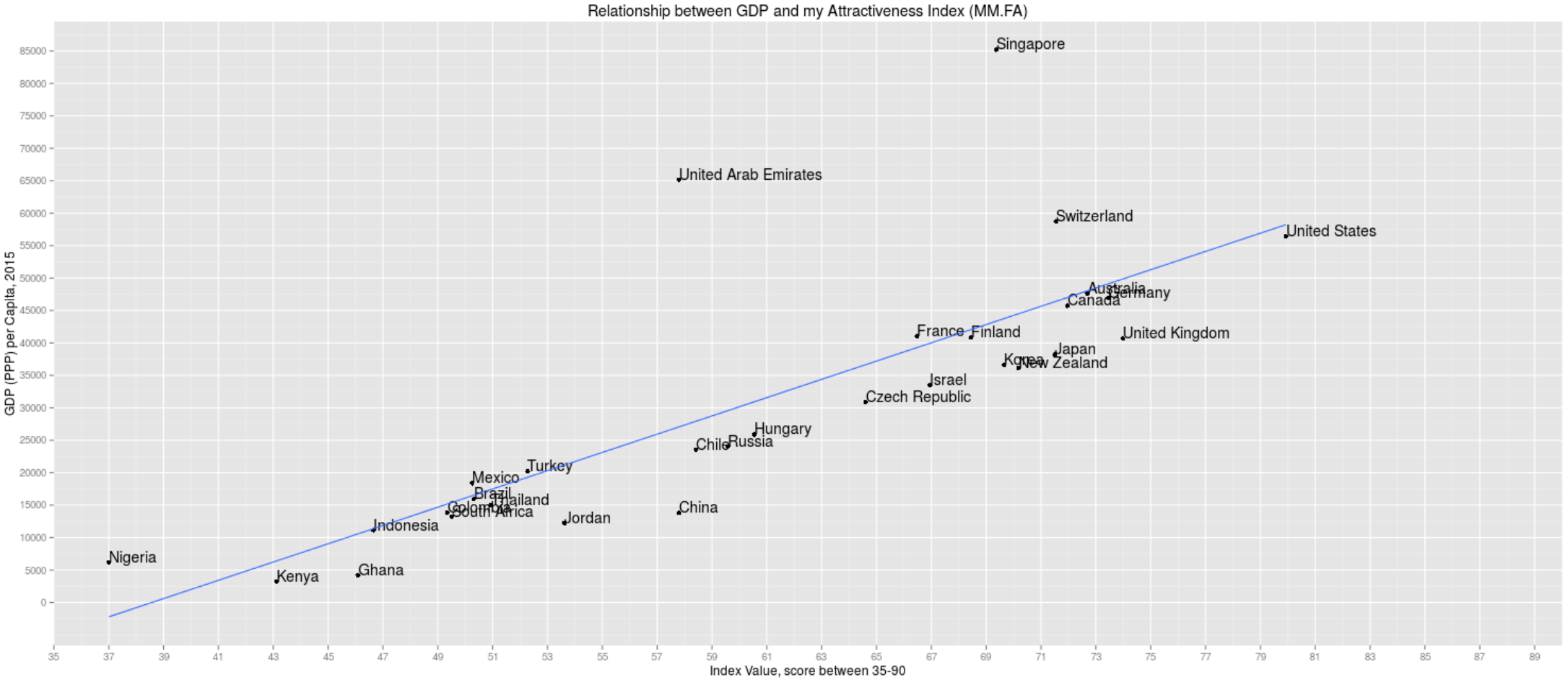

Figure 14 shows the relationship between two variables – IMF's projections of GDP at purchasing-power-parity (PPP) per capita in USD for 2015 (y-axis) and nations' Attractiveness Index. Correlation rounds up to 0.79, R-squared to 0.63, see *UnserSensi / IndexVsGDP.R.*

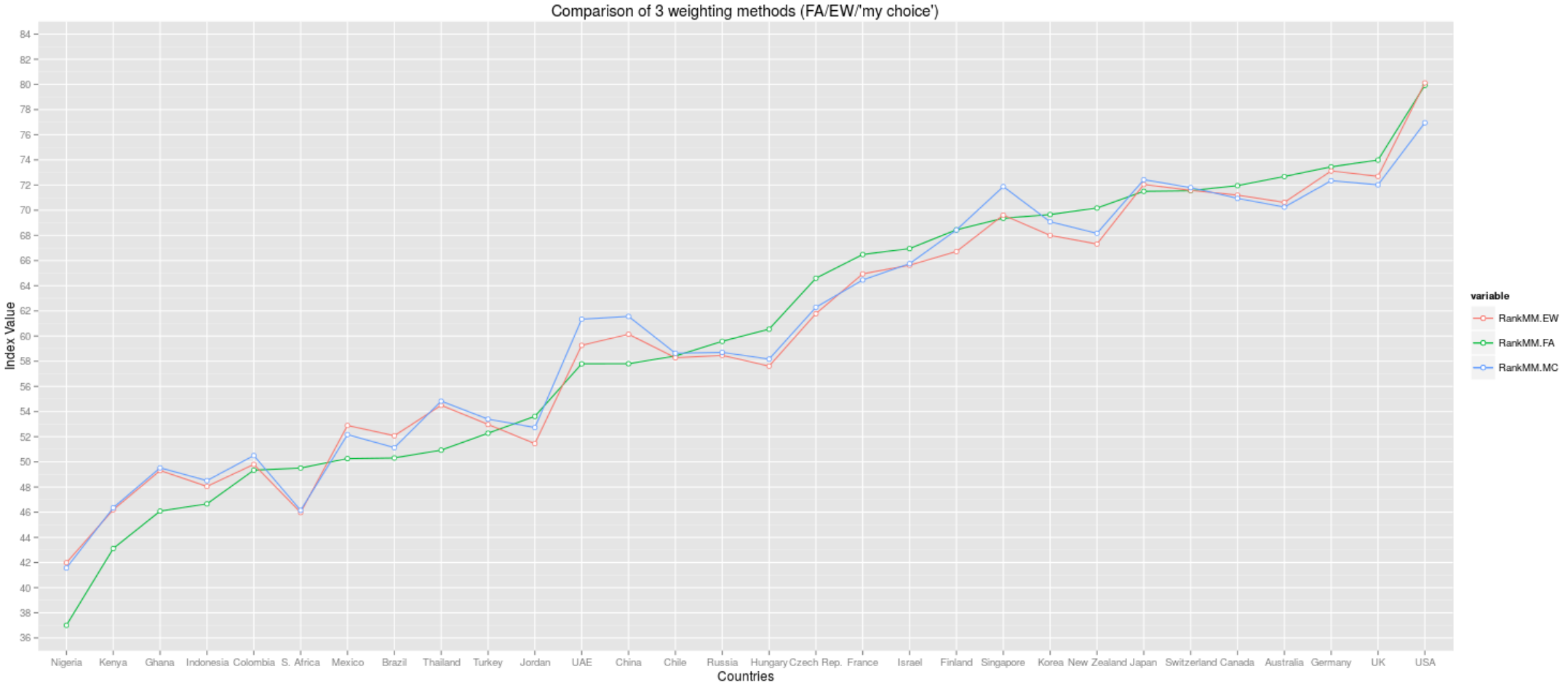

Figure 15 is analogous to the figure 12. However now with the difference that the y-axis shows the value of country's attractiveness, see *UnserSensi / US_Graphs.R*.

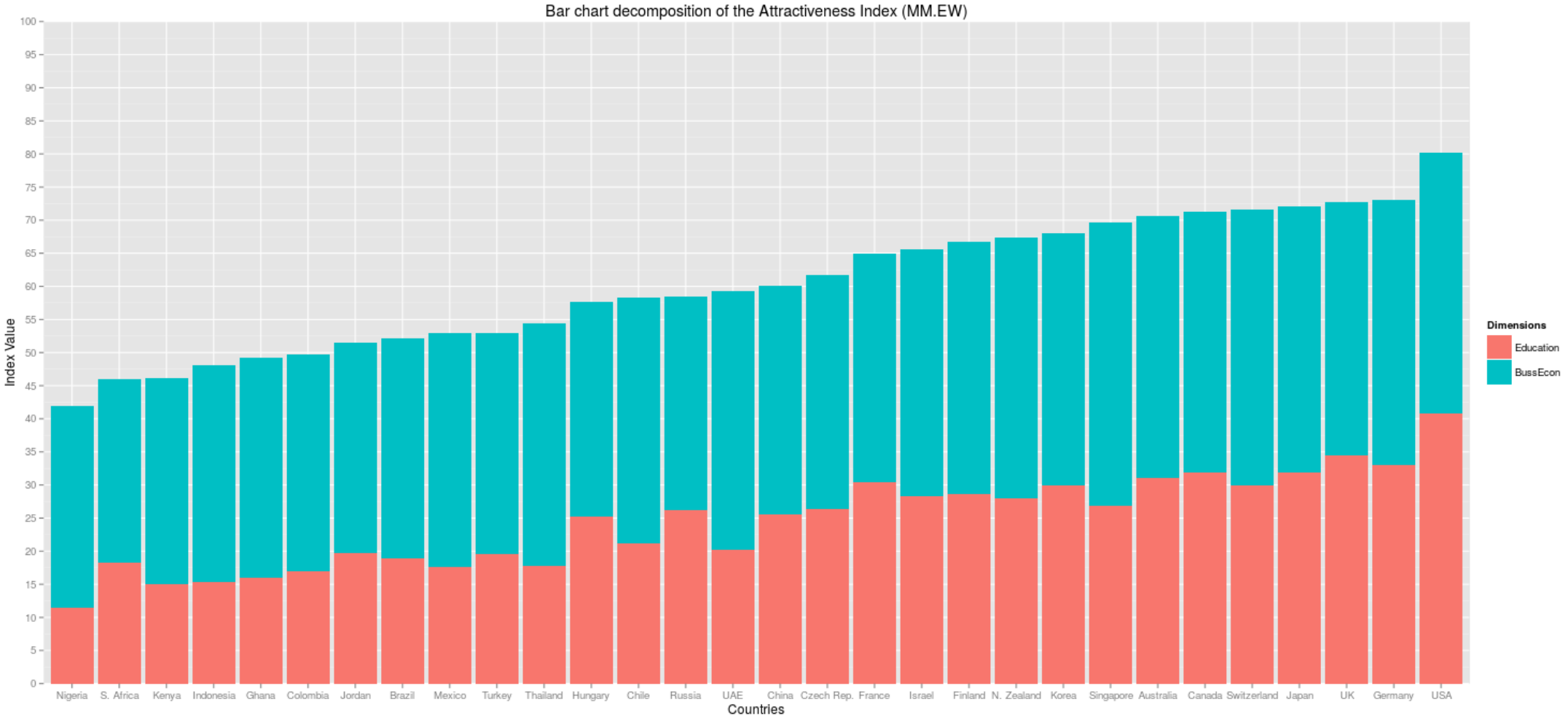

Figure 16 is analogous to the figure 13, now with the difference that the bar chart decomposition is based on the results of equal weighting; see *UnserSensi / BackToDetails.R*.

Country	RankMM.FA	RankMM.EW	RankMM.MC
United States	1	1	1
United Kingdom	2	3	4
Germany	3	2	3
Australia	4	7	8
Canada	5	6	7
Switzerland	6	5	6
Japan	7	4	2
New Zealand	8	10	11
Korea	9	9	9
Singapore	10	8	5
Finland	11	11	10
Israel	12	12	12
France	13	13	13
Czech Republic	14	14	14
Hungary	15	19	19
Russia	16	17	17
Chile	17	18	18
China	18	15	15
United Arab Emir.	19	16	16
Jordan	20	24	22
Turkey	21	21	21
Thailand	22	20	20
Brazil	23	23	24
Mexico	24	22	23
South Africa	25	29	29
Colombia	26	25	25
Indonesia	27	27	27
Ghana	28	26	26
Kenya	29	28	28
Nigeria	30	30	30

Table 13 shows three different results of my weighting techniques. MM = min-max norm.; FA = weights based on factor analysis; EW = equal weights; MC = 'my choice'. Names of countries are labelled according to the colours of continents in the first figure.

Bibliography

[Azur11] Azur, Melissa J., et al. "Multiple imputation by chained equations: what is it and how does it work?" *International journal of methods in psychiatric research* 20.1 (2011): 40-49. http://www.ncbi.nlm.nih.gov/pubmed/21499542

[Balem15] Balemi, Andrew. "Multivariate Analysis." *Statistics in Market Research*. University of Auckland, n.d. Web. 14 July 2015 https://www.stat.auckland.ac.nz/~balemi/Multivariate%20Analysis.ppt

[Bostoc14] Bostock, Mike. "mbostock/d3 - Wiki: Quantitative Scales." *GitHub*. Ed. Last Editor: dwtkns. 30 Oct. 2014. Web. 14 July 2015 https://github.com/mbostock/d3/wiki/Quantitative-Scales

[Brihor11] Everitt, Brian, and Torsten Hothorn. *An Introduction to Applied Multivariate Analysis with R*. New York: Springer, 2011. Print. http://link.springer.com/book/10.1007%2F978-1-4419-9650-3

[Brlcmch01] Breierova, Lucia, and Mark Choudhari. "An Introduction to Sensitivity Analysis." MIT System Dynamics in Education Project, Oct. 2001. Web. 14 July 2015 http://clexchange.org/ftp/documents/Roadmaps/RM8/D-4526-2.pdf

[Brown09] Brown, James Dean. "Choosing the Right Type of Rotation in PCA and EFA." *Shiken: JALT Testing & Evaluation SIG Newsletter. 13 (3) November 2009 (p. 20 - 25)* (n.d.): n. pag. University of Hawaii at Manoa, Dec. 2009. Web. 14 July 2015 http://jalt.org/test/PDF/Brown31.pdf

[Butler14] Butler Scientifics. "Exploratory vs Confirmatory Research." N.p., 8 Oct. 2014. Web. 14 July 2015 http://www.butlerscientifics.com/#!Exploratory-vs-Confirmatory-Research/ciyl/CA5AF661-7462-4AE8-815C-04B6F70BCC76

[Calchin14] Stirling, Peter, and Kathy MacDonald. "Calculate Your Academic Footprint." *Calculate Your H-index*. Uwaterloo.ca, 11 Sept. 2014. Web. 14 July 2015 http://subjectguides.uwaterloo.ca/content.php?pid=84805&sid=1885850

[Camo15] "Multivariate Data Analysis." *CAMO Software*. n.d. Web. 14 July 2015 http://www.camo.com/multivariate_analysis.html

[Castro02] Vladimir Estivill-Castro. 2002. Why so many clustering algorithms: a position paper. *SIGKDD Explor. Newsl.* 4, 1 (June 2002), 65-75. http://doi.acm.org/10.1145/568574.568575

[Cbinder14] Binder, Carola. "Thoughts on the Fed's New Labour Market Conditions Index". Quantitative Ease, 17 July 2014. Web. 14 July 2015 http://carolabinder.blogspot.de/2014/07/thoughts-on-feds-new-labor-market.html

[Chazey12] Chauvet, Marcelle, and Zeynep Senyuz. "A Dynamic Factor Model of the Yield Curve as a Predictor of the Economy." Federal Reserve Board, FederalReserve.gov, March 2012, Web. 14 July 2015 https://www.federalreserve.gov/pubs/feds/2012/201232/201232pap.pdf

[Cheshu14] Oleksandr Chernyak & Marina Shumayeva, 2014. "The Analysis Of Methods For Constructing Composite Indicators." Revista Economica, Lucian Blaga University of Sibiu, Faculty of Economic Sciences, vol. 66(2), pages 75-82.

[Cimi14] User: John. "R - Using Cluster Information in Multiple Imputation." N.p., 14 July 2014. Web. 14 July 2015. http://stats.stackexchange.com/q/107530 ; see *Imputation / notUsed/UnsuperRF.R* script

[cluster] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2015). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.1. http://cran.r-project.org/package=cluster

[clustrd] Angelos Markos, Alfonso Iodice D' Enza and Michel Van de Velden (2013). clustrd: Methods for joint dimension reduction and clustering. R package version 0.1.2. http://CRAN.R-project.org/package=clustrd

[Coghlan15] Coghlan, Avril. "Using R for Multivariate Analysis." *Using R for Multivariate Analysis — Multivariate Analysis 0.1 Documentation*". n.d. Web. 14 July 2015 http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html

[Compo14] "Composite (adj.)", Turnbull, Joanna, ed. *Oxford Advanced Learner's Dictionary*. 8th Edition ed. Oxford: Oxford UP, 2010. Print.

[Correla15] "Correlations: Direction and Strength Based on Dancey and Reidy's (2004) Categorisation." *Dancey, C., & Reidy, J. (2004). Statistics without Maths for Psychology: Using SPSS for Windows, London: Prentice Hall.* Web. 14 July 2015 https://www.strath.ac.uk/aer/materials/4dataanalysisineducationalresearch/unit4/correlationsdirectionandstrength/

[corrplot] Taiyun Wei (2013). corrplot: Visualization of a correlation matrix. R package version 0.73. http://CRAN.R-project.org/package=corrplot

[Cososb05] Anna B. Costello, Jason W. Osborne, *Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis*, *Practical Assessment, Research & Evaluation*, Vol. 10 (2005), pp. 173-178

[Dodg08] Dodge, Yadolah. *The Concise Encyclopedia of Statistics*. Berlin: Springer, 2008. 259-63. Print. http://link.springer.com/referencework/10.1007%2F978-0-387-32833-1

[Dontas10] Dontas, George. "What Are the Differences between Factor Analysis and Principal Component Analysis?" Ed. Nick Cox. N.p., 12 Aug. 2010. Web. 14 July 2015 http://stats.stackexchange.com/a/1584

[Duy14] Duy, Tim A. "The Labour Market Conditions Index: Use With Care." *Tim Duy's Fed Watch*. N.p., 06 Oct. 2014. Web. 14 July 2015 http://economistsview.typepad.com/timduy/2014/10/the-labor-market-conditions-index-use-with-care.html

[Eduin13] UNDR. "Human Development Reports - Education Index." UN, 15 Nov. 2013. Web. 14 July 2015. http://hdr.undp.org/en/content/education-index

[Ee06] Dr. C. George Boeree. *Erik Erikson*. N.p., 2006. Web. 14 July 2015 http://webspace.ship.edu/cgboer/erikson.html

[ellipse13] Duncan Murdoch and E. D. Chow (2013). ellipse: Functions for drawing ellipses and ellipse-like confidence regions. R package version 0.3-8. http://CRAN.R-project.org/package=ellipse

[Esptse14] Esposito, Mark, and Terence Tse. "Jobless Youth in China: Crisis in the Making?" CNBC, 20 Feb. 2014. Web. 14 July 2015. http://www.cnbc.com/id/101433696

[Eurostat] Eurostat. "Harmonised Unemployment Rate by Sex - Age Group 15-24." Web. 14 July 2015. http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=teilm021&plugin=1

[Eurostat15] "Migration and Migrant Population Statistics". Eurostat - Statistics Explained, 25 Mar. 2015. Web. 14 July 2015 http://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics

[Fatow15] User: Towriss. "Factor Analysis (Adapted from Hare Et Al 1998)." N.p., Jan. 2007. Web. 14 July 2015 https://www.networkedcranfield.com/cell/Knowledgebase/Quants%20Material/Factor%20Analysis.pdf

[Fedfunds] Board of Governors of the Federal Reserve System (US), *Effective Federal Funds Rate* [FEDFUNDS], retrieved from FRED, Federal Reserve Bank of St. Louis, 14 July 2015 https://research.stlouisfed.org/fred2/series/FEDFUNDS/

[Flmi14] Leubsdorf, Ben. "Fed's New Labor-Market Index Saw Conditions Improve in September." Real Time Economics at WSJ.com, 06 Oct. 2014. Web. 14 July 2015 http://on.wsj.com/1uQZRQB

[Flom14] Flom, Peter. "How to Interpret the Dendrogram of a Hierarchical Cluster Analysis", 15 Jan. 2014. Web. 14 July 2015. http://stats.stackexchange.com/a/82332

[Fomc11] "The Structure of the Federal Reserve System. - The Federal Open Market Committee." Fed, 14 Jan. 2011. Web. 14 July 2015 http://www.federalreserve.gov/pubs/frseries/frseri2.htm

[fpc14] Christian Hennig (2014). fpc: Flexible procedures for clustering. R package version 2.1-9. http://CRAN.R-project.org/package=fpc

[Frblm14] Board of Governors of the Federal Reserve System (US), *Labor Market Conditions Index* [FRBLMCI], retrieved from FRED, Federal Reserve Bank of St. Louis, 14 July 2015 https://research.stlouisfed.org/fred2/series/FRBLMCI/

[Frein15] "2015 Index of Economic Freedom | The Heritage Foundation." *Promoting Economic Opportunity and Prosperity by Country*. The Heritage Foundation & The Wall Street Journal, n.d. Web. 14 July 2015. http://www.heritage.org/index/about; http://www.heritage.org/index/book/executive-highlights ; http://www.heritage.org/index/download

[Freud03] Freudenberg, Michael. *Composite indicators of country performance: a critical assessment*. No. 2003/16. OECD Publishing, 2003. https://ideas.repec.org/p/oec/stiaaa/2003-16-en.html

[Fultz12] Fultz, Neal. "Annotated SPSS Output: Principal Components Analysis." Institute for Digital Research and Education, 5 Nov. 2012. Web. 14 July 2015 http://statistics.ats.ucla.edu/stat/spss/output/principal_components.htm

[Garrett06] Garrett-Mayer, Elizabeth. "Lecture 8: Factor Analysis I." JHSPH Department of Biostatistics. Fall 2006. *Statistics in Psychosocial Research: Measurement*. Web. 14 July 2015 http://ocw.jhsph.edu/courses/statisticspsychosocialresearch/pdfs/lecture8.pdf

[Geweke77] Geweke, John. 1977. "The Dynamic Factor Analysis of Economic Time Series Models" Latent Variables in Socio-Economic Models. eds. Dennis J. Aigner and Arthur S. Goldberger, Ch. 19. North-Holland Publishing Co.

[Ggplot11] H. Wickham (2009). ggplot2: elegant graphics for data analysis. Springer New York, http://CRAN.R-project.org/package=ggplot2

[Gillies15] Gillies, Duncan Fyfe. "Lecture 15: Principal Component Analysis." 21.9.2011. Web. 14 July 2015 http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture15.pdf

[GPArot] Bernaards, Coen A. and Jennrich, Robert I. (2005) Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis, Journal of Educational and Psychological Measurement: 65, 676-696. http://www.stat.ucla.edu/research/gpa

[Graham09] Graham, John W. "Missing data analysis: Making it work in the real world." *Annual review of psychology* (2009): 549-576. http://www.stats.ox.ac.uk/~snijders/Graham2009.pdf

[Gupta02] Gupta, Barun K. Sen. "Modern Foraminifera" *Barnes & Noble*. Springer Science & Business Media, 31 May 2002. Web. 14 July 2015 http://www.barnesandnoble.com/w/modern-foraminifera-sen-gupta/1100747404?ean=9781402005985

[Hakwil13] Hakkio, Craig S., and Jonathan L. Willis. 2013. "Assessing Labour Market Conditions: The Level of Activity and the Speed of Improvement." Federal Reserve Bank of Kansas City Macro Bulletin (July 18). http://www.frbkc.org/publicat/research/macrobulletins/mb13Hakkio-Willis0718.pdf

[Hcci05] Nardo, Michela, et al. *Handbook on constructing composite indicators: methodology and user guide*. No. 2005/3. OECD publishing, 2005. ISBN 978-92-64-04345-9. 14 July 2015 http://www.oecd.org/std/42495745.pdf

[Hditech14] UNDR. "2014 Report Technical Notes", UN. June 2014. Web. 14 July 2015 http://hdr.undp.org/sites/default/files/hdr14_technical_notes.pdf

[Heko10] Ulrich Heink, Ingo Kowarik, *What are indicators? On the definition of indicators in ecology and environmental planning, Ecological Indicators*, Volume 10, Issue 3, May 2010, Pages 584-593, http://dx.doi.org/10.1016/j.ecolind.2009.09.009

[hindex] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America,* 102(46), 16569-16572. http://arxiv.org/abs/physics/0508025

[HKprotests] User: STSC. "2014 Hong Kong Protests." *Wikipedia*. 21 May 2015. Web. 14 July 2015. https://en.wikipedia.org/wiki/2014_Hong_Kong_protests (only for the reference to the event)

[Hock14] Hockenos, Paul. "OPINION: Differences Persist between Eastern and Western Germany" Al Jazeera America, 9 Nov. 2014. Web. 14 July 2015. http://alj.am/10Kq5Y8

[Howel12] Howell, David C. "Treatment of Missing Data--Part 1." *Treatment of Missing Data*. N.p., 9 Dec. 2012. Web. 14 July 2015 http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

[Humph13] Humphries, Melissa. "Missing Data & How to Deal: An overview of missing data." Population Research Center (2013). https://www.utexas.edu/cola/centers/prc/_files/cs/Missing-Data.pdf

[Igl04] Igl, Wilmar. "Behandlung Fehlender Werte." Rehabilitationswissenschaftlicher Forschungsverbund Bayern. 8 Jun. 2004. Web. 14 July 2015 http://www.rehawissenschaft.uni-wuerzburg.de/methodenberatung/Igl_040604_Halle_Fehlende_Werte.pdf

[Ilo11] "Indicator 2: Youth Unemployment Rate." ILO, 2 Aug. 2011. Web. 14 July 2015 http://www.ilo.org/public/english/employment/yen/whatwedo/projects/indicators/2.htm

[Indic14] "Indicator", Turnbull, Joanna, ed. *Oxford Advanced Learner's Dictionary*. 8th Edition ed. Oxford: Oxford UP, 2010. Print.

[Irritate11] User: Irritate. "How to Scale down a Range of Numbers with a Known Min and Max Value." N.p., 14 Mar. 2011. Web. 14 July 2015 http://stackoverflow.com/a/5295202

[Ishioka13] Ishioka, Tsunenori. "Imputation of missing values for unsupervised data using the proximity in random forests." *eLmL 2013, The Fifth International Conference on Mobile, Hybrid, and On-line Learning*. 2013. http://www.rd.dnc.ac.jp/~tunenori/doc/elml_2013_3_20_50062.pdf

[Jaxen14] Ed. JAX Editorial Team. "The Search for the Best Programming Language of 2014 - JAXenter." JAXenter, 12 Dec. 2014. Web. 14 July 2015 http://jaxenter.com/best-programming-language-2014-113110.html

[Jolliffe02] Jolliffe, I. T. *Principal Component Analysis*. New York: Springer, 2002. Print. http://wpage.unina.it/cafiero/books/pc.pdf

[Jrcth15] "Step 1: Theoretical Framework - JRC Science Hub - European Commission". n.d. Web. 14 July 2015. https://ec.europa.eu/jrc/en/coin/10-step-guide/step-1

[Kabac14] Kabacoff, Robert I. "Quick-R: R Packages." Statmethods.net, n.d. Web. 14 July 2015 http://www.statmethods.net/interface/packages.html

[Kaiser60] Henry F. Kaiser, *The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement* April 1960 20: 141-151, doi:10.1177/001316446002000116

[Kmk13] KMK. "The Education System in the Federal Republic of Germany." http://www.kmk.org/fileadmin/doc/Dokumentation/Bildungswesen_en_pdfs/dossier_en_ebook.pdf , May 2013. Web. 14 July 2015. http://www.kmk.org/information-in-english/the-education-system-in-the-federal-republic-of-germany.html

[Lagg03] "Lagging Indicator Definition". Investopedia, 23 Nov. 2003. Web. 14 July 2015 http://www.investopedia.com/terms/l/laggingindicator.asp

[Leacu14] "The Learning Curve." Ed. The Economist Intelligence Unit & Pearson. May 2014. Web. 14 July 2015 http://thelearningcurve.pearson.com/ http://thelearningcurve.pearson.com/content/download/bankname/components/filename/The_Learning_Curve_2014-Final_1.pdf http://thelearningcurve.pearson.com/content/download/bankname/components/filename/2014Indexfordownload_FINAL-v2.xlsx.xls

[Lmci14] Chung, Hess T., Bruce Fallick, Christopher J. Nekarda, and David D. Ratner. "FEDS Note: Assessing the Change in Labour Market Conditions." FederalReserve.gov, 22 May 2014. Web. 14 July 2015. http://www.federalreserve.gov/econresdata/notes/feds-notes/2014/assessing-the-change-in-labor-market-conditions-20140522.html and subsequent update http://www.federalreserve.gov/econresdata/notes/feds-notes/2014/updating-the-labor-market-conditions-index-20141001.html

[Lmcidoc14] Chung, Hess T., Christopher J. Nekarda, Bruce Fallick, and David D. Ratner. "Assessing the Change in Labour Market Conditions." Federal Reserve Board, 17 Dec. 2014. Web. 14 July 2015 http://www.federalreserve.gov/econresdata/feds/2014/files/2014109pap.pdf

[Lshtm15] Bartlett, Jonathan. "Regression Mean Imputation." London School of Hygiene and Tropical Medicine, n.d. Web. 14 July 2015 http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=69:regression-mean-imputation&catid=39:simple-ad-hoc-methods-for-coping-with-missing-data&Itemid=96

[Luxguc15] T. Luzzati, G. Gucciardi, *A non-simplistic approach to composite indicators and rankings: an illustration by comparing the sustainability of the EU Countries*, Ecological Economics, Volume 113, May 2015, Pages 25-38, ISSN 0921-8009, 14 July 2015 http://www.sciencedirect.com/science/article/pii/S0921800915000658

[Ma09] Clark, Mike. "Multivariate Analysis - Overview". University of North Texas. Mar. 2009. Web. 14 July 2015. Class of Advanced Techniques in the Science of Human Nature http://www.unt.edu/rss/class/mike/6810/IntrotoMV.pdf

[Marr15] Marr, Paul. "Principal Component Analysis." *Http://webspace.ship.edu/pgmarr/Geo441/Lectures/Lec%2017%20-%20Principal%20Component%20Analysis.pdf* Web. 14 July 2015 http://webspace.ship.edu/pgmarr/Geo441/Geo441.htm

[Matlab15] Matlab. "Documentation - Hierarchical Clustering." The MathWorks, Inc., n.d. Web. 14 July 2015. http://de.mathworks.com/help/stats/hierarchical-clustering.html

[mclust12] Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca (2012), *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, Technical Report No. 597, Department of Statistics, University of Washington http://CRAN.R-project.org/package=mclust

[Muen12] Muenchen, Robert. "The Popularity of Data Analysis Software." r4stats.com, 12 Apr. 2012. Last update: 25 Nov. 2014. Web 14 July 2015 http://r4stats.com/articles/popularity/

[Muna03] Munda, Giuseppe, and Michela Nardo. "On the methodological foundations of composite indicators used for ranking countries." Ispra, Italy: Joint Research Centre of the European Communities (2003). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.201.2232&rep=rep1&type=pdf

[Nazim11] Khan, R. Nazim. "Principal Component Analysis." The University of Western Australia. 4 Nov. 2011. *STAT3366 3S6 Applied Statistical Methods - Semester One 2011*. Web. 14 July 2015. http://staffhome.ecm.uwa.edu.au/~00013289/3S6/Lectures/PCA.pdf

[NbClust] Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36. URL http://www.jstatsoft.org/v61/i06/.

[Negotr08] Marco Del Negro & Christopher Otrok, 2008. "Dynamic factor models with time-varying parameters: measuring changes in international business cycles". Staff Reports 326, Federal Reserve Bank of New York. https://ideas.repec.org/p/fip/fednsr/326.html

[Nicscbo0] Giuseppe Nicoletti & Stefano Scarpetta & Olivier Boylaud, 2000. "Summary Indicators of Product Market Regulation with an Extension to Employment Protection Legislation" OECD Economics Department Working Papers 226, OECD Publishing, https://ideas.repec.org/p/oec/ecoaaa/226-en.html

[Panmiv05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining, (First Edition)*. Chapter 8 – Cluster Analysis: Basic Concepts and Algorithms Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

[Pcafa12] "Principal Components Analysis & Factor Analysis." *Research and Statistical Support*. University of North Texas, 12 June 2012. Web. 14 July 2015 http://www.unt.edu/rss/class/Jon/R_SC/Module7/M7_PCAandFA.R

[Pcatut12] User: Echo. "Tutorial: Principal Component Analysis."17 Oct. 2012. Web. 14 July 2015 http://wiki.originlab.com/~originla/howto/index.php?title=Tutorial%3APrincipal_Component_Analysis

[Pigott01] Pigott, Therese D. "A review of methods for missing data". *Educational research and evaluation* 7.4 (2001): 353-383. http://galton.uchicago.edu/~eichler/stat24600/Admin/MissingDataReview.pdf

[Pisa2000] *About PISA*. OECD, n.d. Web. 14 July 2015. http://www.oecd.org/pisa/aboutpisa/

[plyr11] Hadley Wickham (2011). *The Split-Apply-Combine Strategy for Data Analysis*. Journal of Statistical Software, 40(1), 1-29. http://www.jstatsoft.org/v40/i01/

[psych15] Revelle, W. (2015) psych: Procedures for Personality and Psychological Research, Northwestern University, Illinois, USA, http://CRAN.R-project.org/package=psych

[Quandl14] Raymond McTaggart and Gergely Daroczi (2014). Quandl: Quandl Data Connection. R package version 2.3.2. http://CRAN.R-project.org/package=Quandl

[Raschka14] Raschka, Sebastian. "Implementing a Principal Component Analysis (PCA) in Python Step by Step." *Principal Component Analysis Step by Step*. N.p., 13 Apr. 2014. Web. 14 July 2015 http://sebastianraschka.com/Articles/2014_pca_step_by_step.html

[Rcore08] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org

[Rdatafr] "R: Data Frames." *R-devel Version of Documentation*. N.p., n.d. Web. 14 July 2015 https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html

[reshape2] Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. http://www.jstatsoft.org/v21/i12/

[rgl] Daniel Adler, Duncan Murdoch and others (2014). rgl: 3D visualization device system (OpenGL). R package version 0.95.1201. http://CRAN.R-project.org/package=rgl

[rJava13] Simon Urbanek (2013). rJava: Low-level R to Java interface. R package version 0.9-6. http://CRAN.R-project.org/package=rJava

[Rodrigues14] Rodrigues, Isabel M. "Cluster Analysis." Lisbon, Instituto Superior Tecnico. Mai 2014. Web. 14 July 2015 https://fenix.tecnico.ulisboa.pt/downloadFile/3779580670953/Clusters.pdf

[Root10] Root, Elisabeth D. "Principal Components Analysis." 26 Apr. 2010. *University of Colorado*. Web. 14 July 2015 http://www.colorado.edu/geography/class_homepages/geog_4023_s11/Lecture18_PCA.pdf

[Rub76] Rubin, Donald B. "Inference and missing data." *Biometrika* 63.3 (1976): 581-592. http://qwone.com/~jason/trg/papers/rubin-missing-76.pdf

[Rubin04] Rubin, Donald B. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004. http://books.google.de/books?id=bQBtw6rx_mUC

[rvest] Hadley Wickham (2015). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.2.0. http://CRAN.R-project.org/package=rvest

[Sais04] Saisana, Michaela. "Composite Indicators – A Review." Second Workshop on Composite Indicators of Country Performance. OECD, Paris. 26 Feb. 2004. Lecture. 14 July 2015 www.oecd.org/sti/ind/29398640.pdf

[Saisana05] Saisana, Michaela, Andrea Saltelli, and Stefano Tarantola. "Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.2 (2005): 307-323.

[Saitar02] Saisana, M. and Tarantola, S., (2002). "State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development", European Commission, Joint Research Centre, Ispra, Italy, EUR 20408 EN http://bookshop.europa.eu/en/state-of-the-art-report-on-current-methodologies-and-practices-for-composite-indicator-development-pbEUNA20408/

[Sarman13] C. Saranya, G. Manikandan. "A Study on Normalization Techniques for Privacy Preserving Data Mining." *International Journal of Engineering and Technology* 5.3 (2013): 2701-704. June 2013. Web. 14 July 2015 http://www.enggjournals.com/ijet/docs/IJET13-05-03-273.pdf

[Saspca07] "Principal Components Analysis - Chapter 1." SAS, 19 Oct. 2007. Web. 14 July 2015 http://support.sas.com/publishing/pubcat/chaps/55129.pdf

[scale] Hadley Wickham (2014). scales: Scale functions for graphics.. R package version 0.2.4. http://CRAN.R-project.org/package=scales

[Schgram02] Schafer, Joseph L., and John W. Graham. "Missing data: our view of the state of the art." *Psychological methods* 7.2 (2002): 147. http://www.nyu.edu/classes/shrout/G89-2247/Schafer&Graham2002.pdf

[Schwabj15] Schwab, James. "Principal Components Factor Analysis - Stage 4: Deriving Factors and Assessing Overall Fit". The University of Texas at Austin, n.d. Web. 14 July 2015 https://www.utexas.edu/courses/schwab/sw388r7/Tutorials/Principal_Components_Analysis_doc_html/024_Stage_4___Deriving_Factors_and_Assessing_Overall_Fit.html

[Scim15] SCImago. (2007). SJR – SCImago Journal & Country Rank. 14 July 2015 http://www.scimagojr.com/

[Sharand12] Sharpe, Andrew, and Brendon Andrews. "An Assessment of Weighting Methodologies for Composite Indicators: The Case of the Index of Economic Well-being." *CSLS Research Report No. 2012 - 10*. Centre for the Study of Living Standards, 11 Dec. 2012. Web. 14 July 2015 http://www.csls.ca/reports/csls2012-10.pdf

[Signpca14] "Does the Sign of PCA or FA Components Have a Meaning?" Ed. User: Amoeba, 5 Mar. 2014. Web. 14 July 2015 http://stats.stackexchange.com/q/88880

[Sjr12] Guerrero-Bote, Vicente P., and Félix Moya-Anegón. "A further step forward in measuring journals' scientific prestige: The SJR2 indicator." *Journal of Informetrics* 6.4 (2012): 674-688. http://www.sciencedirect.com/science/article/pii/S1751157712000521

[Smith02] Smith, Lindsay I. "A tutorial on principal components analysis." *Cornell University, USA* 51 (2002): 52. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[Smith11] Smith, Kalen. "List of 16 Major Leading & Lagging Economic Indicators." Money Crashers, 06 Sept. 2011. Web. 14 July 2015 http://www.moneycrashers.com/leading-lagging-economic-indicators/

[Spenc11] Spector, Phil. "Cluster Analysis." *Cluster Analysis*. N.p., 16 Mar. 2011. Web. 17 Feb. 2015. http://www.stat.berkeley.edu/~s133/Cluster2a.html

[Storti15] Storti, Davide. "Cluster Analysis." UNESCO and Information Processing Tools - IDAMS Statistical Software, n.d. Web. 14 July 2015 http://www.unesco.org/webworld/idams/Doc/ManualHtml/E2clusfi.htm

[stringr12] Hadley Wickham (2012). stringr: Make it easier to work with strings. R package version 0.6.2. http://CRAN.R-project.org/package=stringr

[Tang07] Andranik Tangian, *Analysis of the third European survey on working conditions with composite indicators*, European Journal of Operational Research, Volume 181, Issue 1, 16 August 2007, Pages 468-499 http://dx.doi.org/10.1016/j.ejor.2006.05.038

[Techni09] "Data Mining Methods - Clustering Methods." Israel - Israel Institute of Technology (Technion). Spring 2009. Web. 14 July 2015 http://moodle.technion.ac.il/pluginfile.php/150699/mod_resource/content/0/Lectures/chap11_Clustering4pp.pdf

[Tiob04] "TIOBE Index for November 2014: November Headline: Statistical Language R on Its Way to the Top 10." TIOBE Software: The Coding Standards Company, Apr. 2004. Web. 14 July 2015 http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html http://blog.revolutionanalytics.com/2014/11/r-tiobe-12.html

[Tippm14] Tippmann, Sylvia. "Programming Tools: Adventures with R." *Nature.com*. Nature Publishing Group, 29 Dec. 2014. Web. 14 July 2015 http://www.nature.com/news/programming-tools-adventures-with-r-1.16609

[Toci05] Michela Nardo, Michaela Saisana, Andrea Saltelli, Stefano Tarantola. "Tools for Composite Indicators Building" 2005 — European Commission, EUR 21682 EN, Institute for the Protection and Security of the Citizen, JRC Ispra, Italy

[Torrey10] Torres - Reyna, Oscar. "Getting Started in Factor Analysis (using Stata 10)" Data and Statistical Services, Sept. 2010. Web. 14 July 2015 http://dss.princeton.edu/training/Factor.pdf

[Ttns13] User: ttnphns. "How to Normalize Data to 0-1 Range?" N.p., 23 Sept. 2013. Web. 14 July 2015 http://stats.stackexchange.com/a/70808

[Undr14] UNDR. "Human Development Reports." *2014 Human Development Report - Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience*. UN. http://hdr.undp.org/en/content/human-development-index-hdi http://hdr.undp.org/en/2014-report/download, Web. 14 July 2015

[Walz20] Walz, Rainer. "Development of environmental indicator systems: experiences from Germany." *Environmental Management* 25.6 (2000): 613-623. http://link.springer.com/article/10.1007%2Fs002670010048

[Wayman03] Wayman, Jeffrey C. "Multiple imputation for missing data: What is it and how can I use it." *Annual Meeting of the American Educational Research Association, Chicago, IL*. 2003. http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf

[Wbg14] World Bank Group. "Easy of Doing Business Index". *Ranking of Economies*. Http://www.doingbusiness.org/methodology, June 2014. Web. 14 July 2015 http://www.doingbusiness.org/rankings

[Wbyu15] "Unemployment, Youth Total (% of Total Labour Force Ages 15-24) (modelled ILO Estimate)." ILO, N.d., Web. 14 July 2015 http://data.worldbank.org/indicator/SL.UEM.1524.ZS

[Wef14] Schwab, Klaus, ed. *The Global Competitiveness Report 2014–2015*. Geneva: WEF, 2014. Web. 14 July 2015. (pages in this thesis refer to pages in the PDF document, not the pages in the original document) http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2014-15.pdf http://reports.weforum.org/global-competitiveness-report-2014-2015/downloads/

[Weisner06] Wiesner, Andrew. "Factor Analysis." Pennsylvania State University, 2006. Web. 14 July 2015 http://sites.stat.psu.edu/~ajw13/stat505/fa06/17_factor/

[Weispca06] Wiesner, Andrew. "Principal Components Analysis (PCA)." Pennsylvania State University 2006. Web. 14 July 2015 http://sites.stat.psu.edu/~ajw13/stat505/fa06/16_princomp/ ; http://sites.stat.psu.edu/~ajw13/stat505/fa06/16_princomp/06_princomp_interpret.html

[Wsjm15] Jordan, Miriam. "Demand for Skilled-Worker Visas Exceeds Annual Supply." *WSJ*. 7 Apr. 2015. Web. 14 July 2015. http://www.wsj.com/articles/u-s-demand-for-skilled-worker-visas-exceeds-annual-supply-1428431798

[xlsx14] Adrian A. Dragulescu (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7. http://CRAN.R-project.org/package=xlsx

[Zapp14] Zapponi, Carlo. "GitHut - Programming Languages and GitHub." N.p., 2014. Web. 14 July 2015 http://githut.info/