Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Department of Business Administration

Creation of a composite indicator to assess students' academic satisfaction at Engineering Schools

Bachelor Thesis

Supervisor: Professor Lucas Van Wunnik

Author: Javier Morales Bellsolà

Barcelona, January 2016

Creation of a composite indicator to assess students' academic satisfaction at Engineering Schools

Abstract

Universities have traditionally responded to two main functions: teaching and research. Universities provide training for high-level jobs and increase the body of theoretical knowledge as well as their possible applications through research. Competitiveness in the academic world has led to the upcoming of many different University tables that order Universities' according to their academic performance using various combinations of different factors. However, such tables do not put emphasis on teaching quality and are mostly based in objective (though arguable) indicators of Universities' research performance. In the opinion of the author current league tables can be misleading and confusing for students who use tables to choose a University for their Studies and even for Universities themselves who might be compelled to adopt policies to contempt leagues tables instead of focusing on long term policies aimed to improve both their research and their teaching quality. This document presents the process of creation of an alternative composite indicator which will serve to assess academic performance at Higher Education Institutions in Engineering-related fields. A composite indicator is formed when individual indicators are compiled into a single index, on the basis of an underlying model of the multi-dimensional concept that is being measured. Such multi-dimensional concept to be measured is in this case academic performance understood as the ability of a university to contenting the expectations, enhancing the capacities and providing the tools students consider to be relevant in their process of being transformed from high school students to competent and ready professionals to enter in the labor force as engineers. As a result of such considerations this composite indicator will be focused rather on University as a training and learning institution and not so much as an engine of knowledge creation.

The composite indicator created will be put into practice to assess academic performance at four Engineering Schools (Escola Tècnica Superior d'Enginyeria Industrial de Barcelona (Spain), Facoltà d'Ingegneria di l'Università di Roma Tor Vergata (Italy), Facoltà d'Ingegneria di Roma La Sapienza (Italy) and Delft Technical University (Netherlands) to guarantee its applicability in practice.

.

Content

Figures

Tables

Equations

Definitions

Composite Indicator: A composite indicator is formed when individual indicators are compiled into a single index, on the basis of an underlying model of the multi-dimensional concept that is being measured. **Index** can be used as a synonym of **Composite Indicator.** In Rankings parlance a quite accurate concept to refer to Composite indicators is the concept of **Ranking methodology** which is the set of indicators which are taken into account, normalized, weighted and aggregated to appraise Universities that will subsequently be ranked. Those three concepts will be used indistinctly as synonyms throughout the text.

Normalization: is the process prior to any data aggregation required as indicators in a data set often have different measurement units.

Weighting and aggregating: These two concepts are defined together because of their interdependence. **Weights** are fundamentally value judgements assigned to certain categories when conforming a composite indicator. **Aggregating** is the way such value judgements are put together quantitatively.

Analytic Hierarchy Process definitions (Each one of the following concept will be defined in a more comprehensive and broad way in section 4.5)

Analytic Hierarchy Process (AHP) is a methodology of multi-criteria analysis developed at the end of the 70's decade at the Wharton School of Business by Ph. D. in mathematics Thomas L. Saaty. Such methodology provides support to decision-making in situations of complex structuration of the problem, under the assumption of bounded rationality and where a plurality of decisional criteria (probably contraposed one to another) are involved.

Element (in AHP parlance): criterion included in any hierarchic level of an AHP hierarchy.

Matrix of Pairwise Comparisons $\{A\}$ **(MPC)** is a square matrix used in the **AHP** constituted by a number of rows and columns equal to the number of nodes or criteria included in a given set. The row i and the column j represent the criteria i and j that need to be confronted whereas the element $a_{ij} \in \{A\}$ represents the quotient between the absolute priority of each criterion (or the relative priority between criteria i and j) that the decision-maker estimates according to his opinions.

Local Weight (LW) in AHP parlance expresses the degree in which a given **element** 'collaborates' into achieving the objective represented by its 'father' element in an absolute way. That is, how important is that element for the 'father' criterion.

Local Weights Vector (LWV): set of 'son' local weights that 'collaborate' into achieving the objective expressed by a father element in an absolute way.

Normalized Local Weights Vector (NLWV): set of 'son' local weights that 'collaborate' into achieving the objective expressed by a father element in a relative way. The sum of all the elements inside this vector needs to be equal to 1.

Consistency Index (CI): Parameter that measures the degree of consistency of an **MPC.**

Random Index (RI): Index associated to matrices of dimension n which is obtained by calculating the **CI** of a sufficiently big number of matrices (i.e. 50000) equal in dimension that have been filled at random. Once those (i.e. 50000). **CI**'s have been obtained they are averaged and as a result the **RI** is defined.

Consistency Ratio (CR): Quotient defined as CI/RI**.**

1. Introduction

1.1. Motivation

The motivation for writing this thesis was born after taking an undergraduate course called 'Project I : Creating an Index of Well-Being' at ETSEIB (Escola Tècnica Superior d'Enginyeria Industrial de Barcelona) a faculty inside the Polytechnic University of Catalonia (UPC). Such course was intended to provide students with a theoretical framework and set of tools so that they could build their own composite indicator of well-being throughout the course.

During the course, Professor Van Wunnik motivated us to question whether neoclassical economics metrics were correct in evaluating societies' welfare through a series of different references (often through the reading of scholars' articles).

Some of the references presented in class included the Easterlin's Paradox (which defends the non-positive correlation between income and happiness after a certain level of income) or the Maslow Pyramid of needs.

The course was divided into two main blocks. In the first place there was the analysis of GDP as an indicator of well-being and the realization of the many factors that make of GDP a quite limited indicator to address the problem of quantifying well-being. In the second place realizing that GDP was indeed a poor way measure of well-being, we were urged to build a composite indicator that could transcend such limitations. The creation of such composite indicators should act as an element to evaluate performance in different countries that could be used both for decision makers and International Organizations.

I would say that in the end, however, the main purpose of the course (to put it in common words the 'big picture') was that of analyzing why some metrics that have widely been used in the past might not be useful anymore and how to find the right metrics to evaluate concrete issues. What's more, how to find metrics that are suitable to measure variables that are themselves rather ambiguous and open to discussion (such as well-being or happiness).

In this way I feel the problem of metrics not only has a prominent relevance when it comes to evaluating societies' well-being but also when it comes to evaluating other kinds of well-being (i.e. well-being at work or at university).

Indexes of Well-Being aimed at evaluating a country/society's performance have been a major topic in literature (see Index of Economic Well-being[1], Happy Planet Index [2] or OECD Better life Index[3]) in precedent years and I have felt it would not be interesting to approach a field that has already been explored so many times by many eminent scholars (i.e. Sen, Stiglitz or Fitoussi (2009)).

Thus, I have tried to carry out a research that would contribute to help a collective of people with which I can identify myself. As a student I have noticed that the problem of addressing satisfaction and well-being at Higher Education Institutions (HEIs) has not been approached in a profound manner. Though several studies on academic performance have been made, they have mostly not considered subjective parameters that might indeed have an effect on academic performance.

---

[1] In the paper '*The measurement of Economic Well-Being'* (1985) Osberg set the theoretical framework that would allow researchers at the Centre for the Study of Living Standards in Canada (CSLS) develop the Index in 1998. Such index seeks to measure Economic Well-Being through a set of categories such are Effective per capita consumption, Net societal accumulation of stocks of productive resources, Economic Inequality or Economic Security.

[2] The New Economics Foundation introduced this alternative Index in 2006. HPI (Happy Planet Index) seeks to 'measure what matters: the extent to which countries deliver long, happy, sustainable lives for the people that live in them. The Index uses global data on life expectancy, experienced well-being and Ecological Footprint to calculate this' extracted from HPI website *http://www.happyplanetindex.org/about/*.

[3] In the paper '*Report by the Comission on the Measurement of Economic Performance and Social Progress (Executive summary) (2009)'* Fitoussi, Sen and Stiglitz set out a series of guidelines aimed at assessing Well-being both in its economic and social array. The OECD used such document to launch the OCDE Better Life Index in 2011 which allows users create a composite indicator weighting different variables according to their own preferences. The Index takes into account different dimensions such as Housing, Income, Jobs, Community, Education, Environment, Governance, Health, Life Satisfaction, Safety and Work-Life balance to set a ranking of performance of all the OCDE countries.

1.2. Objective

The objective of the present thesis is that of creating a composite indicator to evaluate performance at universities taking into account subjective parameters. In particular, given that evaluating performance as a whole would require a much broader study I am particularly interested in measuring how Engineering schools prepare their students as perceived by students themselves.

The project as a whole, then, should serve to create and apply a composite indicator to quantify **academic performance** with a strong focus in **quality of teaching** at Engineering Schools. This composite indicator could be considered together with other indicators of different nature to evaluate overall university performance.

The composite indicator should serve to help Engineering Higher Education Institutions (HEIs) improve the perception students have of their teaching institution.

1.3. Structure

The thesis will be divided in 5 main sections.

The first section serves as an introduction to the topic. The motivation, the objective and the structure are presented here.

The second section is aimed at highlighting that once a metric or model gains a certain popularity (even if it's not completely accurate) it is very difficult to be critical of it. This might lead to inaccurate data evaluation and might end up in poor decision-making processes. Two examples will be given to exemplify this phenomenon: the supremacy of GDP as macroeconomic indicator and the great attention University Rankings are receiving when assessing HEIs academic quality. This section also comprises an analysis of some University Rankings and some criticism to the way they are articulated and used.

In the third section the theoretical framework that articulates the thesis is set. In that section it is defined what academic performance is and which factors contribute to accomplish it. In other words, a formal definition is given regarding the aspects that are being measured in the present study.

In the fourth section the composite indicator to evaluate 'academic performance' is created. Firstly some background into multi-criteria decision making is presented. In particular a long and wide explanation of the Analytic Hierarchy Process (AHP) is developed given that such methodology has provided the theoretical basis under which the index has been created. Once the theoretical knowledge has been presented it is applied into the practical case of evaluating academic performance at Engineering Schools. The model is applied to 4 different Engineering Schools and each university obtains a score according to the variables, weighting and aggregation methods that will be presented throughout the section. A comparison between the 4 Universities under study will be carried out once the scores for each University are obtained.

In the fifth section some conclusions help to evaluate whether the overall accomplishment of the objective has been reached and if not which further actions could be taken in the future.

2. Overview:

In this section the theoretical framework that articulates and gives meaning to the present thesis will be set. This section serves as an introduction to the problem of metrics and is intended to highlight some cases in which inappropriate metrics have conditioned decision making often leading to decisions that are not based on a critical analysis but rather in the application of a non-necessarily correct model.

The following section illustrates that decision-making based on an absolute or 'monopolistic' indicator often leads to incomplete analyses. The model becomes the main focus for the analyst and the real phenomenon is set aside as if the model was indeed the actual scope of study.

For a long time, economic indicators such as GDP have deserved all the attention in contrast with other indicators that have become ignored or set aside. Similarly, in recent years we have seen the coming of a set of University Rankings that have driven mass opinion into the belief that rankings are the correct way to evaluate how well a university is doing and that only by looking at how high on a ranking a university is we can obtain an accurate idea of the overall quality of such institution.

Much in the way GDP has been considered the most relevant indicator of macroeconomic performance University Rankings are starting to be seen as the new 'absolute truth' when it comes to evaluating University Performance.

GDP as a country performance indicator is presented as an example where metrics fail and some parallelisms are drawn to University Rankings as quality of teaching of HEIs indicators.

2.1. GDP and its Inability to assess Well- Being

Throughout the 20th century, an eminently materialistic approach to economic science has created an 'absolute preponderance' of certain indicators (such as GDP) that have been key to determining which course to take when decision makers have been compelled to make data-driven decisions. Though data-driven decisions do never rely uniquely in one indicator, GDP has broadly been used as the synthesis indicator to evaluate economic country performance though actually GDP is only meant to measure the economic value of final goods and services produced by a country (OECD, 2008).

The supremacy of GDP has become of such importance that it has been used in a quite vague manner by politicians to evaluate whether a country was 'doing well'. In other words, as Gertner (2010) states governments have long held the view that only one statistic, the measure of gross domestic product can really show whether things seem to be getting better or getting worse.

This rather materialistic approach finds its origin in the *Principles* of Marshall (1890) (regarded as one of the founders of economics) where he stated that Economics should not deal with 'well-being' but with the 'material requisites' of it.' The heritage of this conception is still very present in most of the indicators that are being used extensively nowadays. This conception has motivated that often indicators constituted from variables that allow a fast conversion into quantitative outputs have been regarded as more reliable than other indicators based on rather intangible parameters or subjective data.

2.1.1. Redefining the scope of measure

The new approach to Economics claims that 'happiness' and well-being should be main concerns of Economic Science contrasts with the traditional approach that would only take into account its material requisites.

As Dixon (1997) has outlined economists from different backgrounds...all believe that happiness must play a more central role in economic science once again. In the same way, Oswald (1997) has highlighted 'The importance of the economic performance is that is that it can be a means for an end. The economics matters interest only as far as they make people happier'.

Sen (1985) talks about the concept of utility in two different dimensions: desire fulfillment and happiness. Under his assumptions, there is no discussion that economics if aimed at maximizing utility should also be responsible for assuring 'well-being' and 'happiness' to citizens.

2.1.2. Creating appropriate measuring methodologies

Therefore, if the aim has changed, new metrics that adjust correctly to the actual beliefs should be developed. In the field of Economics of Development, Composite indicators have appeared as a more holistic way to approach the new problems of Economics. Composite Indicators that compare country performance are increasingly recognized as a useful tool in policy analysis and public communication. (OECD,2008)

The development of this alternative Composite Indicator to evaluate University performance as a contraposition to current University Rankings in greater use (as ARWU or QS) is analogous to the motivation that critics with GDP (i.e. Sen) had when creating Well-Being Indexes such as the Human Development Index or the Happy Planet Index (both of which have been referred in the footnotes of *subsection 1.1*).

2.2. University Performance Rankings: Background and Criticism

University Rankings, are increasing in popularity year after year mainly because they aim to respond to a 'public demand for transparency and information that institutions and government have not been able to meet on their own.' (Usher & Savino, 2006, p 38). Rankings compare Higher Education Institutions (HEI) through a ponderation of different items or criteria that the creator/s of the list deems as relevant to assess/reflect academic quality or performance.

It has been highlighted in literature (see Hazelkorn (2009), (2014), (2015)) that the way they are configured is sometimes too simplistic and inaccurate and has been criticized as being conditioned by the judgements of the compilers. However, Rankings have acquired a power they could never imagine and now affect not only students' perception but also model the opinion of governments, governments' agencies, businesses and employers. Its power has been increased as a product of its 'appearance of scientific objectivity' (Ehrenberg, (2001)).

Given this new vision of Higher Education Institutions (HEI) it seems obvious that evaluating the academic quality of a university by only looking at its position on a certain ranking is as reductionist as looking at a single macroeconomic indicator (i.e. GDP) to determine the performance of a country's economy.

University Rankings have been set up mainly by the ponderation and addition extrinsic, objective, tangible and quantitative parameters. The creators of the Shanghai Ranking (one of the most accepted University Rankings of the world) have argued Liu and Cheng (2005, p. 135) that the Shanghai ranking uses "carefully selected objective criteria", is "based on internationally comparable data that everyone can check", and is such that "no subjective measures were taken".

If university performance is modelled as the output of a function that operates with an array of inputs or variables, it is clear that those inputs have to include both objective and subjective data.

It is difficult to believe that no subjective measure is to be considered in order to evaluate University Performance. If rankings are aimed at assessing which higher institutions perform better, how can no subjective data be considered in doing so?

Though probably actual rankings are representative of how well universities perform, the most influential ranking (ARWU) has evaded the responsibility to properly measure and take into account inputs of rather subjective nature that have an undeniable relevance when it comes to evaluating the quality of a university.

Other rankings such as the Times Higher Education or the QS have tried to consider include subjective criteria in their ranking methodology but the current trend is that of avoiding subjective evaluation. The Times Higher Education Ranking has appealed to the need for more reliable world rankings that could be used for everything from government policy to shaping institutions' reputations made the previous QS methodology – largely based on surveys (subjective evaluation) –too volatile.

Many have argued that most rankings fail at evaluating the quality of teaching and consider only variables related with faculty members expertise, awards or recognition[4].

The US News & World Report ranking (the most influential HEI Ranking in the US) has also received critics. Researchers, professors and students have questioned the variables that serve as a basis for the final ranking. In 2007, presidents of more than 60 liberal arts colleges refused to participate in a key component of the U.S. News & World Report rankings. The rankings, they wrote, "imply a false precision and authority" and "say nothing or very little about whether students are actually learning at particular colleges or universities."(Enserink, 2007).

One of the pieces of criticism that has been referred to repeatedly in literature is that of former vice-president of ASSU (Associated Students of Stanford University) Nick Thompson (1996) in which he argued that 'Rankings are subjective, but are taken as dogma by college applicants and employers and encourage college administrations to mold their policies towards boosting their school's ranking, rather than focusing on legitimate educational goals'.

Globalization and the marketization of higher education are responsible for Rankings to have gained major attention in recent years because they appear to (i) gauge world-class status; (ii) provide accountability; and

[4] The Shanghai ranking has been one of the most questioned rankings in recent times. 'It is a remarkably stable list, relying on long-term factors such as the number of Nobel Prize-winners a university has produced, and number of articles published in Nature and Science journals. But with this narrow focus comes drawbacks. China's priority was for its universities to 'catch up' on hard scientific research. So if you're looking for raw research power, it's the list for you. If you're a humanities student, or more interested in teaching quality? Not so much. *"University rankings: which world university rankings should we trust?". The Telegraph. 2015. Retrieved 14th Nov. 2015*

(iii) measure national competitiveness (Hazelkorn 2009). Assuming that higher education is the engine to achieve economic growth, governments have been induced in recent times to create policies to position their HEIs in the rankings instead of adopting long-term HE strategies.
It is clear that rankings clearly influence decision making and as Hazelkorn has underlined 'the history of rankings shows that measuring the wrong things can produce distortions'.

The following section should be able to explain how the major international rankings are built. A brief introduction to the criteria and the aggregation method on each Ranking methodology should help the reader to understand the criticism that such methodologies generate in opinion of the writer of the present study.

2.3. Existing Rankings: An Introduction

Marginson, former professor of higher education in the Centre for the Study of Higher Education at the University of Melbourne and actual Professor of International Higher Education at UCL divides rankings into two[5] main categories: those that rely integrally on objective research metrics (such as the Shanghai ARWU and probably the Leiden Ranking) and multi-indicator rankings systems like those produced by *Times Higher Education (THE)* and *U.S. News & World Report*, which assign weights to various objective and subjective indicators (reputation surveys, faculty-student ratio, selectivity in admissions…).
Here an example and further discussion of two existing, well consolidated rankings will be presented as a particular case of each category. The Shanghai Ranking will be studied as representative as a quality research ranking and the Times Higher Education ranking as an example of multi-indicator ranking.
The ARWU Ranking has been chosen because it is probably the most influential Ranking worldwide whereas the Times Higher Education (THE) Ranking considers a broad span of factors and is one of the only rankings that take into account quality of teaching (and therefore better represents what is to be measured an evaluated in the present study).
The Leiden University Ranking (focused on impact of papers and focused in scientific performance) is not worth examining in the present paper because of its eminently research approach (similar in scope to that of the ARWU but with a lower degree of recognition). The US News &World Report Ranking will also be discarded of profound study because it is focused in US Universities whereas the present Study is centered to evaluate a subset of European HEIs. The QS Ranking will not be studied deeply either because of the similarities in core with the THE[6].

[5]Marginson has talked about three categories arguing that the third category "is uniquely occupied by QS because Q.S. simply doesn't do as good a job as the other rankers that are using multiple indicators." See Bibliography [44]

[6]THE had previously collaborated with Quacquarelli Symonds (QS) to publish a *THE–QS World University Rankings* from 2004 to 2009 but the partnership finished and both started to publish their own Rankings. QS continued using the same methodology that was used in the collaborative Ranking whereas the THE developed new criteria to be considered and new aggregation weights. It has been considered after analyzing the two methodologies that the one developed by the THE is a revision (probably an improved one) of the previous version of the Ranking and thus, will be presented in a higher degree of detail. However, both rankings are not sharing the same audience which is something that has also changed since they separated. QS claims that they have always been clear they are aimed at prospective international students (Danny Byrne, QS) and adds that their rankings are easy to understand and of direct relevance to students arguing they are unique in asking potential employers what they think of the universities, for example. However, THE highlights how

An interesting comparative chart of main composite indicators Rankings is attached below. In such chart three main aspects of League Tables are reflected: (i) where/how the data is obtained, (ii) which are the criteria that have been selected to build the Ranking and (iii) the assignation of weight to each criterion.

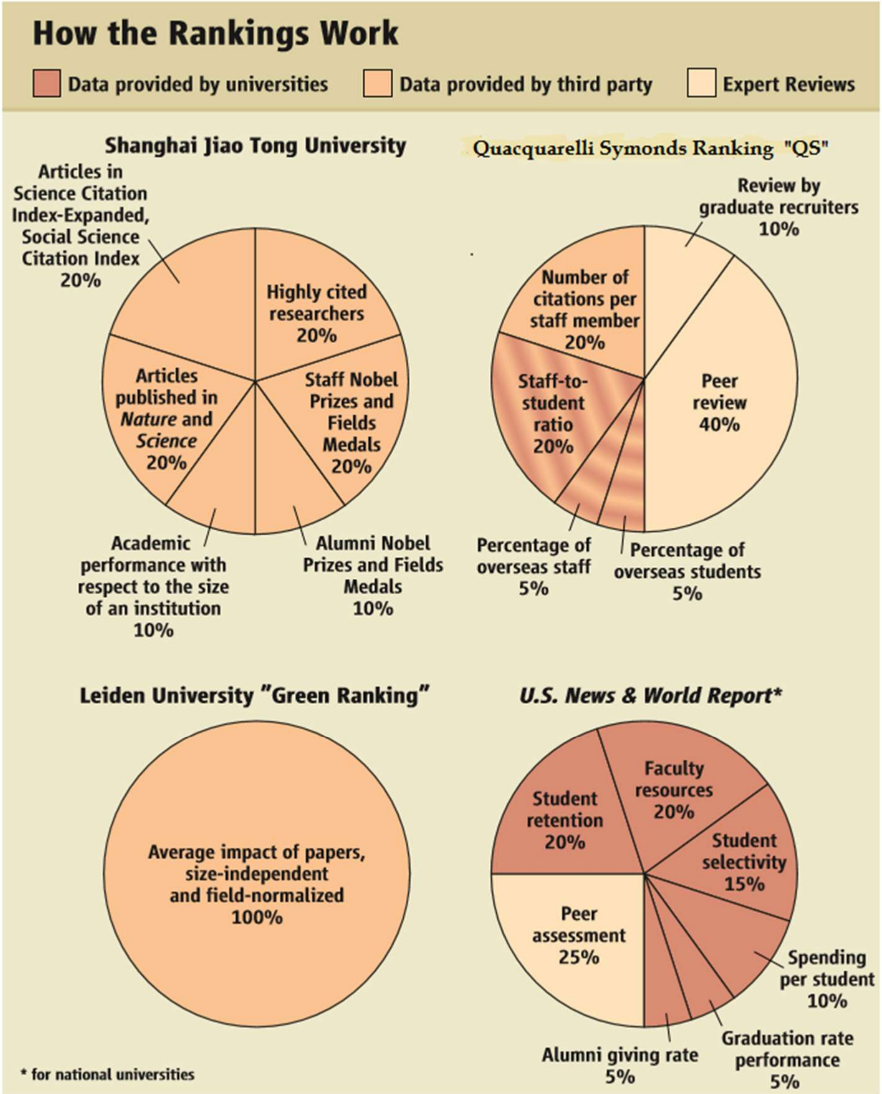


Figure 1: Comparison between major Rankings criteria selection and assigned weights and provenance of the data used. Source: Science Magazine (2007)

their rankings stand up to more academic scrutiny (Phil Baty, THE editor) and adds "We produce high-end rankings which are used by governments around the world. And we're the only global rankings that take teaching seriously."

2.3.1. The Times Higher Education Ranking (THE) (based on 2015/2016 Report)

Times Higher Education World University Rankings are global performance tables that judge research-intensive universities across all their core missions: teaching, research, knowledge transfer and international outlook.

Times Higher Education Rankings has defined both a general ranking and four different subject Rankings (Arts & Humanities, Life Sciences & Physical Sciences, Engineering & Technology and Social Sciences). In this study the Engineering & Technology Ranking methodology will be analyzed given the nature of the study.

2.3.1.1. University selection

The 2015/2016 version of the Rankings includes 800 Universities Worldwide which is a significant change in comparison to the set of universities which had been included in the period 2004-2010 (200) or in the period 2010-2015 (400). The Ranking includes institutions from 70 countries and has collected data from 1128 institutions from 88 countries. Institutions are included on the list if they teach at undergraduate level and achieve to publish at least 200 research papers a year in journals indexed by Elsevier in the 5 precedent years (i.e. 2010-2014 in the case of the 2015 Ranking). To be included in the Engineering & Technology Ranking institutions need even higher rates of publications (minimum 500 research papers a year according to Elsevier).

2.3.1.2. Data collection

Institutions provide their institutional data to be used in the rankings. If a particular data point is missing a low estimate between the average value of the indicators and the lowest value reported is entered (the 25th percentile of the other indicators). This allows not to penalize an institution too harshly with 0 values for not-provided/measured data.

For the reputation surveys that account for a significant share of the overall ranking result invitation-only questionnaires are sent by the THE to a limited number of institutions around the world. UN data is used to ensure a representative coverage of the scholar world. The target of such questionnaires are experienced scholars who give their opinion about teaching and research of other institutions. Scholars are asked about the level of other universities in their field. They are asked to name 10 institutions they believe are the best.

In 2015 reputation rankings are based on a survey based on 10507 responses from 142 countries (from which 9794 were deemed valid).

2.3.1.3. Normalization and aggregation

In order to make it possible to integrate specific data included in independent indicator to a total score the THE uses a standardization approach for each indicator and combine all the indicators using the proportions added in the following subsection. The standardization technique is based on the data distribution within a particular indicator where a cumulative probability function is calculated and an evaluation of a particular institution's indicator lies within such function. A cumulative probability score of an indicator X reveals that a university with random values for that indicator would fall below that score X per cent of the time. Mathematically this can be expressed as:

$$Z = \frac{X - \mu}{\sigma}$$

Equation 1: Standardized normalization

Where X is the score obtained by a certain university on a given indicator, μ is the expected value and σ the standard deviation of the population.

For all the indicators except from the Academic Reputation Survey the cumulative probability function is calculated using a version (not specified) of Z-scoring. The distribution of data in the Academic Reputation Survey requires the addition of an exponential component.

2.3.1.4. Ranking Criteria

Figure 2 provides a clear picture of the methodology and the criteria used by THE to set up its Engineering and Technology University Ranking.

ENGINEERING & TECHNOLOGY

Group weight	30					30			27.5	5	7.5		
Indicator weight	3	1.5	4.5	19.5	1.5	4.5	4.5	21	27.5	5	2.5	2.5	2.5

SUBJECT RANKINGS METHODOLOGY

Indicator	Total students/ academic staff	PhD awards/ bachelor	PhD/Academic staff	Reputation Survey (teaching)	Institutional income/ Academic staff	Scholarly papers/ Academic Staff	Research income/ Academic Staff	Reputation Survey (research)	Citations: Research impact	Income from industry/ Academic Staff	Ratio of international to domestic staff	International co-authorship	Ratio of international to domestic students
	Teaching: The learning environment					Research: volume, income and reputation			Citations per paper	Industry income: innovation	International outlook		

Figure 2: Percentage distribution of different dimensions that constitute the THE Engineering and Technology Ranking methodology

THE makes use of 13 performance indicators to set up its rankings. Such indicators can be grouped into five areas:

- **Teaching**
- **Research**
- **Citations**
- **International Outlook**
- **Industry Income**

2.3.1.4.1. Teaching (the learning environment) (30%)

- Staff/Student Ratio (3%)
- Doctorate/Bachelor Ratio (1.5%)
- Doctorates awarded/academic staff Ratio (4.5%)
  A high proportion of postgraduate research students suggests the provision of teaching at the highest level that is attractive to graduates and effective at developing them.
- Reputation survey (teaching) (19.5%)
  The Academic Reputation Survey examines the perceived prestige of institutions in teaching.
- Institutional Income (1.5%)
  Gives a sense of facilities and infrastructure available both to students and staff.

2.3.1.4.2. Research (volume, income and reputation) (30%)

- Reputation survey (21%)
  This category looks at a university's reputation for research excellence among its fellows, based on the responses scholars give to an annual Academic Reputation Survey.
- Research income (4.5%)
  Research income (indispensable for research) is scaled against staff numbers and adjusted for purchasing-power parity (PPP). This indicator takes into account that often research grants in science subjects are bigger than those awarded for the highest-quality social science, arts or humanities research.
- Research productivity (4.5%)
  This indicator counts the number of papers published in the academic journals indexed by Elsevier's Scopus database per scholar, scaled for institutional size and normalized for subject. This gives a sense of the university's ability to get papers published in quality peer-reviewed journals.

2.3.1.4.3. Citations (research influence) (27.5%)

Research influence is examined by capturing the number of times a university's published work is cited by scholars globally, compared with the number of citations a publication of similar type and subject is expected to have. The data supplier Elsevier examines millions of citations (51 million) belonging to journal articles (11.3 million in 23000 different journals) published in the last 5 years. Publications taken into account include journal articles, conference proceedings and reviews.

Citations give an idea of how much a university contributes to the sum of human knowledge outlining influence researchers and the impact of their researches. Data is treated to take into account the distribution of publications in different fields not to be detrimental to institutions that excel in low publications' volumes fields. Also measures are taken to give equal opportunities to universities which publish in languages other to English.

2.3.1.4.4. International outlook (staff, students, research) (7.5%)

- International-to-domestic-student ratio (2.5%)
- International-to-domestic-staff ratio (2.5%)
  The ability of a university to attract undergraduates, postgraduates and faculty from all over the planet is key to its success on the world stage.
- International collaboration (2.5%)
  This indicator calculates the proportion of a university's total research journal publications in the last 5 years that have at least one international co-author and reward higher volumes.

2.3.1.4.5. Industry income (knowledge transfer) (5%)

This category seeks to capture such knowledge-transfer activity by looking at how much research income an institution earns from industry (adjusted for PPP), scaled against the number of academic staff it employs. A university's ability to help industry with innovations, inventions and consultancy has become a core mission of the contemporary global academy and that is what the indicator is trying to evaluate.

2.3.1.5. Criticism

The THE Ranking is a rather well reputed ranking which has not created as much controversy as other major rankings (such as the ARWU presented below). The fact that it is a composite multi-dimensional index makes it quite robust and though some might differ in their specific weights 'arbitrary' assignation, it is generally considered quite complete. However, high degrees of disagreement have been created around the way Reputation Surveys (which account for the 34.5% of the overall result) are carried out. Expert opinion based on reputation are conditioned by three phenomena (i) halo-effect: a particular department's reputation an expert is familiar with can influence rating of a whole institution (ii) Experts may be uninformed about the institutions they rate (experts are most likely to succeed when rating their own universities but probably not so much when rating others) (iii) the seriousness with which respondents treat opinion polls.

Other complaints have come from the way Scholars are chosen and selected though throughout the years the ranking has improved the transparency of their methodologies.

The same has happened with former criticism about the citations indicator under which English-publishing Universities had a significant advantage in relationship to others. In the actual version this biases have been taken into account.

2.3.2. The Shanghai Ranking. Academic Ranking of World Universities (ARWU)

University selection

Liu and Cheng (2005, p. 127–128) assure to have analyzed around 2000 institutions and argue their selection includes the major universities in every country. The authors sustain to have included all universities having Nobel prize and Fields medal laureates, a significant number of papers published in either Nature or Science, highly cited by Thomson Scientific (ISI) researchers and a significant amount of publications indexed in the Thomson Scientific databases and claim their inclusion criteria achieve to take into consideration all the major universities in each country. The published ranking, however, only includes 500 institutions. From the 1st to the 100th are ranked ordered whereas from the 100th till the 201th position are ranked in groups of 50 and from that one on institutions are ranked in groups of 100.

Ranking Criteria

The authors use six criteria that belong to four different domains:

Quality of Education

This domain is based on a single criterion: the number of alumni of the institution having received a Nobel Prize (Peace and Literature are excluded, the Bank of Sweden Prize in Economics included) or a Fields medal. An alumni is defined as a person having obtained a Bachelor, a Master or a Doctorate in the institution. If a laureate has obtained a degree from several institutions, each one of them receives a share. Not all prizes or medals have the same weight (older awards are less valuable than recent ones). Their value is weighted following a descending linear scheme that diminishes with time (an award received after 1991 counts a 100 %, an award received between 1981 and 1990 counts a 90 % and each decade diminishes the value of the award by a 10%). When several people are awarded the prize or the medal, each institution receives a share. This defines the first criterion labeled ALU.

Quality of Faculty

This domain has two criteria. The first one counts the total number of staff members of an institution who have won Nobel prizes in Physics, Chemistry, Medicine and Economics, and Fields Medals in Mathematics. By staff is meant those who worked at the institution in question at the time of winning the prize. Again, different weights are set, according to period: the weight is 100 percent for winners between 2001 and 2003, 90 percent for winners between 1991 and 2000, 80 percent for winners between 1981 and 1990, 70 percent for winners between 1971 and 1980, and so on; finally, 10 percent for winners between 1911 and 1920. If a winner is affiliated to more than one institution each one receives a share. This criterion is labeled as AWA

The second criterion in this domain is the number of highly cited researchers in each of the 21 areas of Science identified by Thomson Scientific in Life Sciences, Medicine, Physical Sciences, Engineering, and Social Sciences. Researchers among the most highly cited researchers within each category acquire 'points' for their institution. The definition of categories, and detailed procedural information, can be found on the website of the ISI Web of Knowledge4 of Thomson Corporation. These highly cited researchers, in each of the 21 domains, are obtained through a list of 250 people who have received the largest number of citations in their domain according to the Thomson Scientific databases (see http://www.isihighlycited.com). This is computed over a period of 20 years and defines the criterion HiCi.

Research output

This domain includes two criteria. The first one Nature and Science (N&S) indicates the number of articles published in Nature and Science in the past five years. To distinguish the order of author affiliation, a weight of 100 percent is assigned to corresponding author affiliation, 50 percent to first author affiliation (second author affiliation if the first author affiliation is the same as corresponding author affiliation), 25 percent to the next author affiliation, and 1 percent to all other author affiliations. Only publications of article type are considered. Since this criterion is not significant for institutions that specialize in Social and Human Sciences, it is deemed irrelevant (thus not considered) for such institutions.

Science Citation Index (SCI) indicates the amount of papers published by members of the academic staff of an institution in the Science Citation Index Expanded and/or the Social Science Citation Index in the previous year. Only publications of article type are considered. Since Thomson Scientific indexes detriment research in Social and Human Sciences, each publication indexed in the Social Science Citation Index is assigned a coefficient of 2. This two criterion are labeled under the tag PUB.

Productivity

This domain is called size (PCP) by the authors indicates 'the total scores of the above five indicators divided by the number of full-time equivalent (FTE) academic staff' (Liu and Cheng 2005, p. 129) which is clearly a measurement of productivity. The criterion is deemed irrelevant if the number of FTE cannot be obtained.

Data collection

Data are mostly collected online. Data Sources used by the authors include the official site of the Nobel Prizes (http: //nobelprize. org/nobel_prizes/), the official site of the International Mathematical Union (http: //www.mathunion. org/general/prizes, and different Thomson-related sites (http://www.isihighlycited.com and http: //www.isiknowledge.com) though they do not openly share the provenance of some data (i.e. number of faculty staff in each institution).

Normalization and aggregation

Once each criterion has been measured they are normalized in comparison to the institution that receives the best score in each dimension. This methodology concedes a score between (0,100) to each university in each criterion. The six normalized criteria are aggregated using a weighted sum (for more information see section 4.4.1) that provides the global 'performance' of an institution.

The weights awarded to the six criteria are the following: ALU (10%), AWA (20%), N&S (20%), HiCi (20%), PUB (20%) and PCP (10%). The aggregation analytical formula (equation (1)) is attached below:

$$(1) \qquad ARWU = 0{,}1.\mathrm{ALU} + 0{,}2.\mathrm{AWA} + 0{,}2.\mathrm{N\&S} + 0{,}2.\mathrm{HiCi} + 0{,}2.\mathrm{PUB} + 0{,}1.\mathrm{PCP}$$

Equation 2: ARWU aggregation formula

The global results are normalized once again at the end to assure that the highest performing institution is awarded a score of 100.

Criticism

A lot of authors have dedicated long analytic pieces of criticism to the configuration of the ranking. In academia (see Dill and Soo 2005; van Raan 2006) and in position papers (see Saisana and D'Hombres 2008).

Here the criticism will be tackled from two different perspectives. On the one hand, the criteria selected will be critically commented. On the other hand normalization and aggregation methods will be discussed and questioned.

Criticism to criteria selection

ALU and AWA

The alumni criterion (ALU) fails to consider the ***temporal leap*** *between the time when a graduate of a certain institution leaves his/her 'alma mater' and the moment when he/she obtains the award*. Most Nobel-winners are awarded the prize long after their graduation (the average age of Nobel laureates is around 60 years old) so it is difficult to assume a causal relationship between the instructing institution and the award recognition. It is complex to assess in which degree the alma mater contributes to the award achievement.

The AWA dimension might cause confusion in the case that faculty members are awarded the Prize in a different stage of their life than that in which they are teaching/researching at an Institution. As with ALU it is difficult to evaluate the degree in which institutions are involved in the award concession.

HiCi

Van Raan (2005a) has outlined how the tables of Thomson Scientific classifying highly cited authors (accounting for the HiCi criterion) are biased in favor of certain domains (Medicine and Biology). The 21 categories are not representative of the amount of researchers involved in each domain which leads to better performance of Universities that focus on certain fields. As an example, the Thomson Scientific category of 'Immunology' is based on 57 main Journals whereas the category of 'Engineering' considers 997 Journals. Nonetheless, both categories are weighted equally despite the difference in absolute paper production in each category.

N&S

Another considerable piece of criticism can be added for the N&S criterion. Such criterion is subject to opportunistic behavior from the institutions. Assuming the impact of the ARWU, Ranking HEIs could encourage their faculty staff to produce co-authored papers (written by several researchers of the same institution) in order to 'boost' its 'Payback' from a publication. Note how a multiple author publication scores substantially better than a single author publication. As highlighted previously, adopting a different Education Policy to satisfy a certain Ranking methodology has often undesired consequences (Hazelkorn 2009).

PUB

The most highly criticism to be added in this domain is the fact that publications are counted only in terms of quantity and not in terms of impact.

Other general considerations include: (i) the mixing of different time periods for each criterion (100 years for ALU and AWA, 20 years for HiCi, 5 years for N&S and a year for PUB) without a justified explanation, (ii) the strong link/correlation between performance criteria and institution size that favors big HEIs ('big becomes beautiful' (Zitt and Filliatreau (2006)) or (iii) the poor importance given to performance of those institutions that are not ranked among the best ones (the fact that authors rank top 100 in a cardinal manner but set 'performance' groups that increase in size as one strolls down the ranking table without meticulously ordering HEIs inside such groups).

Criticism of normalization and aggregation methods

The discussion will be developed from a MCDM (Multiple Criteria Decision Making) perspective since a rigorous and consistent methodology is required to support the technical arguments presented below. (Further insight about MCDM methods is given in sections 4.3 to 4.5).

The first aspect to comment on is (i) the aggregation methodology adopted by the authors. MCDM proves that when aggregating multiple criteria using a weighted sum (see section 4.4.1) the weights assigned to each criterion are not to be interpreted as the degree of 'importance' of the criteria. Such weights, strange as it may seem, are determined as a consequence of the normalization methodology that has been used. An example might help to understand this point:

If one is trying to aggregate a criterion 'Weight' with other criteria, its performance could either be measured in grams or in kilograms. If an arbitrary weight (or scaling constant in MCDM literature) is set for the criterion as measured in grams (i.e. 0,2) this weight should be multiplied by 1000 if the analyst decides to measure the criterion in kilograms. As a consequence it is clear that the comparison of such scaling constant with scaling constants referred to other criteria is not reflecting a comparison of importance (criterion weight as measured in grams could be less relevant than another criterion whereas if measured in kilograms the opposite thing could occur). Weight attribution in weighted sums is a complex problem and the authors of the Ranking should have explained the reasons to set the scaling constants they have chosen in a justified manner. Without a proper explanation one might assume that such weights have been assigned arbitrarily which is in any case a wrong strategy because it will probably lead to inconsistencies and contradictory results.

A second aspect that deserves to be highlighted derives from the previous observation and relates to (ii) the relationship between weights assigned (aggregation) and the criteria normalization methodology. It seems clear that a change in the normalization methodology intrinsically implies a weights modification. If this is not taken into account most likely absurd result might arise. Note that the normalization method adopted by the authors is based on giving a score of a 100 to the best scoring institution in each criterion. As the reader might notice, non-normalized scores in a specific criteria are likely to change every year and this might lead into what has been called 'most common critical mistake' (Keeney (1992)) or alternatively 'the normalization trap'.

An example (Boyssou, Billaut, Vincke (2010)) can help to illustrate this point:

Example 1

Alternatives	P_1	P_2	P_1^n	P_2^n	Score	Rank

H[7]	2000	500	100	100	100,0	1
A	160	435	8	87	47,5	2
B	400	370	20	74	47,0	3
C	640	305	32	61	46,5	4
D	880	240	44	48	46,0	5
E	1120	175	56	35	45,5	6
F	1360	110	68	22	45	7
G	1600	45	80	9	44,5	8

Table 1: Weighted Sum with equal weights (expressed as a percentage)

Hypothetic data (*Tables 1 and 2*) will be used to rank the K HEIs (A,B C ,D,E,F,G,H) according to M criteria (1,2). To clarify notation, each performance will be represented as P_{ij} where i represents each HEI (i=A,B,C,D,E,F,G,H) and j the evaluation criterion considered (j=1,2) or simply P_j as it appears in the heading of *tables 1 and 2*. The criteria selected (to maintain the criteria used by ARWU) will be (i) PUB and (ii) 10 x Hi-Ci of each institution. The normalization methodology applied is defined as a ratio between the HEI of study in a particular criterion and the score of the best scoring HEI in such criterion (see equation (2)). This operation of normalization needs to be applied iteratively in order to build columns $P_1{}^n$ and $P_2{}^n$ of *tables* 1 and 2.

$$(3) \qquad {P_{ij}}^n = \frac{P_{ij}}{Max\left(P_{Aj} \dots P_{Hj}\right)} \quad for\ j = 1,2$$

Equation 3: Normalization methodology applied by the ARWU

For example, ${P_{C1}}^n = \frac{P_{H1}}{P_{C1}} = 0{,}32 = 32\%$ analogously ${P_{H1}}^n = \frac{P_{H1}}{P_{H1}} = 1 = 100\%$

Let's assume the aggregation is made using a weighted sum with an assignation of equal weights to both criteria. Operating like this the 'Score' column will be defined (note there is no need to re-normalize the global score, since the score of H is already 100). The global score will be used to rank the alternatives in order (see Rank column of *Table 1*) that expresses the overall ordered performance of HEIs:

$$H > A > B > C > D > E > F > G$$

Consider now a situation (i.e. the data after a year) in which everything remains unchanged except the performance of H in terms of the first criterion ($P'_{H2} = 700\ in\ contrast\ to\ P_{H2} = 500\ from\ table\ 1$)

Alternatives	P_1	P_2	$P_1{}^n$	$P_2{}^n$	Score	Rank
H	2000	700	100	100	100,0	1
A	160	435	8	62,14	35,07	8
B	400	370	20	52,86	36,43	7
C	640	305	32	43,57	37,79	6
D	880	240	44	34,29	39,14	5
E	1120	175	56	25	40,5	4
F	1360	110	68	15,71	41,86	3
G	1600	45	80	6,43	43,21	2

Table 2: Weighted sum with equal weights (expressed as a percentage): h increases on g

[7] Note how H is ranked first in reference to Harvard (known to be one of the highest reputed Universities Worldwide)

Data presented in *Table 2* presents the results of the new situation. Normalization is carried out as explained previously (according to equation (3)) but given that $P_{H2} \xrightarrow[changes\ to]{} P'_{H2}$ the normalized scores also vary.

If we proceed to aggregate the two normalized criteria using the same weights as before, the ranking order is modified (see column 'Rank' of *Table 2*) with the following result:

$$H > G > F > E > D > C > B > A$$

It is clear that changing P'_{H2} has inverted the ranking of all other alternatives. However, this results are obvious as the normalization of the second criterion has changed. In order to avoid this problem weights should be changed according to changes in the normalization methodology. The consequence of not changing weights when normalization methodology changes provokes a diminishment in the importance of certain criteria as it has been presented. If a HEI is weak on a certain criterion and close in the overall rank to another one, this HEI will be interested in increasing (and not diminishing!) the distance between its performance in that criterion and the one obtained by the best scoring to eventually become ranked (mechanically) before his precedent competitor. This contradiction (the fact that lower criterion performance can in certain cases lead to higher rank positions) clearly diminishes the confidence that can be given to the Shanghai Ranking.

NOTE: A lot of attention has been put into developing a critical evaluation of the Shanghai Ranking mostly because of the great influence around it. Given that the motivation of the present thesis is that of creating an academic quality performance indicator (and potentially a ranking based on such performance) it is important to critically analyze what has been done beforehand, taking advantage of what has been done correctly but most importantly, improving those aspects that could have been done better. This critical observations are not aimed at disqualifying the overall ranking (it goes without saying the great amount of work behind a Ranking of such characteristics) but in this paper the negative aspects have been highlighted as 'raw material' for further development.

3. What needs to be measured and how does it have to be interpreted?

This section is aimed at defining the core features that determine what this ranking methodology is trying to measure and also serves as a discussion of how the results should be interpreted. In other words this can be seen as the philosophical framework which gives meaning to the present ranking and is aimed at making clear the change of point-of-view in relationship to previous rankings. Assuming that every single ranking methodology is anchored to a certain degree of criticism or discussion and that there is no objectively correct ranking methodology, it is also obvious that actual major rankings are far from perfect and they mostly fail to accurately evaluate academic performance. The present model will offer an alternative of evaluation to traditional rankings.

3.1. The re-definition of academic performance

Though the scope of the present thesis is not that of undermining the validity of widely renowned rankings, it seems indispensable to cover what major rankings have not covered, that is, including quality of teaching as measured by a series of subjective parameters as a fundamental aspect of University evaluation.

As stated in the introduction, the objective of this thesis is to create a model with which 'teaching quality' can be assessed in Engineering Schools. Therefore the approach of this study is substantially different than that of major Rankings.

As observed in van Raan (2006), "academic performance" in University Rankings parlance can mean two very different things: the prestige of an institution based on its past performances and its present capacity to attract excellent researchers. In this study emphasis will be put on the fact that research performance of faculty staff is not representative of the **quality of teaching** of a HEI and that new parameters of evaluation will have to be created to correctly address this issue.

The definitions of **academic performance** and **quality of teaching** need to be clarified because the use given to such concepts in the present thesis could differ from how these concepts might have been used in different contexts.

In the case of this study **academic performance** has a much more specific meaning than that of Van Raan and its scope is mainly oriented towards (i) contenting the expectations, (ii) enhancing the capacities and (iii) providing the tools students consider to be relevant in their process of being transformed from high school students to competent and ready professionals to enter in the labor force as engineers. In this sense, and assuming that Universities are a lot more than teaching institutions, the present study will consider that 'academic performance' relates exclusively to the ability of the University to prepare their students for their future occupations and to the degree in which they make their students learn. **Quality of teaching** will also be a term which will be used throughout the thesis and refers exclusively to the ability faculty staff have at presenting concepts, fostering critical thinking, making students intervene, motivating students…Quality of teaching, thus has a narrower meaning than **academic performance**.

3.2. The democratic component

Students and faculty staff should be involved in the evaluation of their own institutions by sharing their opinions and their judgements. However, it is true that faculty staff have ways to be active into changing things they consider to not be working well whereas students are more likely to not have so many opportunities.

In the present project, more than having external organizations that rate a university performance from the outside using the criteria they consider to be relevant, 'evaluation power' will be given to students. Students though lacking the knowledge and data that these organizations presume to have play a much more important role into the evaluation process because they are the ones who better 'know' their institution and are probably more able to critically speak (constructively) about it.

This subsection is particularly interesting because it highlights two different important aspects. In the first place the fact that people participation is indispensable. From this it is derived that no accurate definition of the study object can be delivered a priori (that is without having asked what quality education really constitutes for students). In this sense the process of defining of what needs to be measured cannot be dissociated from the fact that the nature of the study is participative. Guidelines will be given to collaborators (expressed in section

4.6.2) and from their ideas a clear vision of the object being measured (academic performance) will be better understood.

3.3. The interpretation of the results

It is fundamental to understand the change of point-of-view that this new Ranking methodology introduces as far as its implementation and interpretation is concerned.

While Rankings serve for institutions to acknowledge their position in relation to one another, the present ranking is more focused onto highlighting if a university is doing well enough and can look for ways to improve its performance in specific dimensions. Therefore, the objective of the composite index is that of seeing how good a university is doing in comparison to how well it could be doing. The eminently subjective approach of the index also alters the validity that can be given to the obtained results mostly as far as comparisons are concerned. Diener has outlined (2000) how when evaluating subjective well-being individuals living in different countries respond in different ways according to a cultural influence. On this behalf Diener has added that satisfaction scales might have a somewhat different meaning between cultures related to a global positivity response tendency. For example American people claim to be very happy (probably because being happy is valued as a positive feature in the US) whereas Japanese people claim to be unhappier (probably because the role humbleness plays in their culture). The results of Diener experiments have been brought up to outline that self-reports often provoke an intrinsic bias (not only of cultural kind but also in regard to expectations and beliefs). When confronting a student with a subjective evaluation he/she will unavoidably be conditioned by the set of convictions and past-experiences which will most likely influence his/her report. In this sense given this paradigm it is **extremely difficult to effectuate comparisons across universities in different countries**. Grading systems are another example that might illustrate this reality. In Italy, there is a policy under which exams can be retaken easily so that makes students aim for very high marks. Probably a student often receiving almost-perfect marks will tend to evaluate his university better than students in other countries just because he assumes that low marks represent very poor performance. In contrast, a student in France (or in a Spanish technical University) where professors are generally less solidary when grading might be a lot more severe when grading their institutions. This fact is certainly a limitation to the potential of the present Composite Indicator and obviously diminishes the objective reliability that can be given to the results (‘ranking’) obtained. However no other method other than self-reports has been found as a means to assure the objective of creating a ‘participative’ Index.

3.4. The teaching quality in Engineering Schools. Literature Review.

In order to know what has previously been said about Engineering Schools teaching quality there is a referent who has developed a series of observations and is currently regarded as an expert in the engineering education field. Richard Felder has extensively explored the Engineering learning processes and has published extensively on the knowledge he has acquired throughout his teaching career. Here some of his ideas and principles will be shared. It is worth outlining that his approach is not broad enough to be applied in the present study because he only explores techniques focused on how engineering professors should teach (associated with teaching quality) and does not consider other university-related parameters which could be relevant for students learning processes (associated academic performance). However, his work serves as a good starting point in absence of a proper and accurate definition of what does **academic performance** in the broadest sense mean as far as engineering learning is concerned (this will be defined in section 4.6.5).

Dr. Felder has articulated engineering teaching quality into a series of items which should be taken into account in order to prepare students properly. These are: instructional objectives, learning styles, active learning, collaborative learning and teaching with technology.

Instructional objectives

Instructional objectives are the explicit statements of what students should be able to do if they have learned something. The idea of Felder is that the augmenting the degree of specificity of the teaching objectives (especially those which involve higher levels of critical or creative thinking) students will be more likely to achieve them. He stresses how in configuring such objectives using vague verbs like know, learn or understand is not a good strategy given that such actions cannot be observed directly. He encourages to use verbs like listing, identifying, implementing or calculating. He shares his instructional objectives with his students at the beginning of the course so that they know what they should be able to do after the course. He argues that giving a very long list of competences to accomplish help students feel more challenged and motivated.

Learning styles

Learning styles are the different ways students have to process information. Among them some prefer to operate with concrete information (experimental data or facts) while others prefer to work with abstract concepts or models. Some are the visual kind and prefer to look at figures, pictures and diagrams instead of using verbal information and some are just the opposite way. The objective of the professor should be trying to use different learning styles to 'be fair' with all the different students in the classroom. Felder argues that most classes are based in lectures that involve words and formulas (and not that many visual info). Moreover some students are active learners and need to 'do things' (problem solving, ideas discussing) in order to learn. In this sense he adds laboratory sessions might help to reach that objective. He also points out how some students are sensing learners and need to work slowly and carefully with facts and real objects and see direct connections between what they are learning and the real world. Most engineering education is though focused towards intuitive learners which are comfortable working with theoretical material. Students are expected to write long tests that only fastest-working students might finish and those who understand things very well and work more carefully might end up failing. He demonstrates how students taught with mismatched learning styles learn significantly less as if taught in their preferred styles (Felder, 1996). However he says that making students learn in their favorite styles cannot be achieved and is even undesirable because professionals have to function effectively in all learning styles to carry out their works and duties. The key for him is balance which means increasing visual content and putting more emphasis on observable phenomena and less on theory and experimental data and providing opportunities in class for student activity instead of requiring student to spend the class watching and listening.

Active learning

Active learning can be defined as a method that engages students in course related activities. The existence of feedback and recurrent practice in contrast with passive observation has appeared to be key to make students get better at something. Felder adds the example of giving brief exercises to students in class and give them direct feedback on their results. He usually asks short open questions in class and give students a short amount of time (less than 3 minutes) to come up with a solution. Then recurrently asks to different random students to share their results or evaluations till their answers are satisfactory. He outlines how the factor time is indispensable to avoid that students that finish early get bored or those who can't find a solution experience long unproductive periods. Challenging students by asking them questions they have spent time thinking about results into a higher degree of engagement and commitment with the class and a higher motivation. This provokes students' performance augments significantly.

Cooperative learning

Cooperative learning is a type of collaborative learning that makes students work in teams. In cooperative learning, the team activities are aimed at accomplishing five criteria (Felder (1994)):

- Positive Interdependence (if all team members fail to do their parts correctly each one is penalized)
- Individual Accountability (each team member is held accountable for all the learning that was supposed to take place in the assignment)
- Face-to-Face Interaction
- Appropriate use of interpersonal skills
- Regular self-assessment of team functioning

In this items probably the one that deserves further explanation is the individual accountability one. It refers to the fact that not because a project or assignment has been done in group students can rely on their 'expert' peers to answer questions they would not know to how to answer. That means that when working in a team students not only need to know what they have done but also what others have done. An example of application of this 'principle' that can be found in an oral defence of a group project. Normally students will organize themselves so that the best candidate on each part of the project explains it. Assigning a certain part of a project to a student in a random way is fostering this item.

Some studies (Springer, Stanne & Donovan (1999)) have carried out a meta-analysis of cooperative learning research in science, mathematics, engineering, and technology. They put the results of 39 rigorous studies on a common basis, and showed that on average cooperative learning significantly improved academic performance, lowered dropout rates, and increased student self-confidence.

Teaching with technology

The apparition of 'digital books' and the internet is gradually changing students' learning processes. Textbooks are not only presenting information but most of them come along with courseware interactive resources that can ask questions or pose problems and provide instant feedback to the student. Courseware also allows operations lecturers cannot do (i.e. interrupting a lesson to get a detailed explanation of a certain term; add illustrative images, diagrams, animations, movies…; solve equations…).

Technology also provides new ways to simulate physical, chemical or biological systems and actively explore the effects of parameter changes on system behavior. Experimenting with a system through simulation and learning from it cannot be achieved by watching and listening to a lecturer. Other interactive resources such as video-conference or lecture recording are practices that are being increasingly adopted in HEIs.

There are some things that professors will always be able to do better than virtual ones: motivate and inspire students; promote a sense of community among them while maintaining individual accountability for learning or helping them develop desirable professional, social, and ethical values. Nonetheless, universities that do not take advantage of innovative, technological ways to educate their students will probably be left behind.

After this introduction about which factors deserve attention in order to define what constitutes an engineering school that performs good at teaching, the following section is aimed at defining a methodology to measure this set of factors and some others (which will be presented in section 4.6.2).

4. The creation of the Ranking Methodology. An AHP approach

The main and most general question to ask in order to create an appropriate methodology of measurement is knowing how to find variables that are relevant for a given problem and that accurately turn inputs (variables) into outputs (performance synthesis) in a reliable manner.

Other relevant questions that derive directly from the previous one are stated below:

How to choose significant and relevant variables?

How can variables of different characteristics and backgrounds be added up in order to build indicators?

How can subjective data be transformed into quantitative data?

How to assign weights to different dimensions to simultaneously take into account outputs based on rather objective criteria with variables of subjective nature?

Most of this problems will be solved (or partially solved) in the following subsections.

4.1. A subjective approach

Scientific methodology has traditionally considered rigor and objectivity as key factors to obtain reliable results. However, the apparition of increasingly complex problems of in has made it impossible to find models that deterministically explain the functioning of certain systems especially in cases where human behavior is involved.

The pursue of objectivity is fundamental to approach any subject that has to do with the physical world though models or metrics that try to evaluate human factors exclusively through the quantification of objective parameters are likely to fail. In the present study, reducing academic performance to an objective evaluation of objective performance criteria has seemed to be a wrong strategy. Some might consider the methodology used to be un-scientific and probably they will be right. However, taking into account subjectivity does not mean working without a rigorous method. (This is justified properly in the following 4.2 section).

Subjectivity has been taken into account from two different perspectives:

In the first place, subjectivity has been used to evaluate parameters that do not have a settled unit of measure as most of the variables with which this study will work with. The role of students' participation has been deemed crucial to elaborate this project and the value that has been given to students' reports, opinions and subjective evaluations has been the only way to quantitatively evaluate such variables.

Second, subjectivity has allowed observation to be carried out collectively. It is obvious that when analyzing a particular phenomenon or trying to understand how a system works different points of view normally provide a broader and more holistic view of the system making it easier to analyze.

The following section sets the theoretical paradigm which articulates the way measurements will be carried out and metrics will be defined.

4.2. An alternative conception of metrics

Metrics are defined in mathematics as functions that define distances or relationships between two concrete elements in a set. Traditional metrics have articulated its meaning around the mathematical concept of metric topology, that is, the amount a certain element has of a certain property in comparison to a scale defined through an arbitrary unit. Measuring the weight of a person is an example of metric topology given that there is a property to be measured (Weight) a unit (for example the kilogram) and a scale of reference (a set of all the possible ratios measured property: unit).

Metric properties have extensively been used in the measurement of physical properties and have been used mostly in disciplines that relate to the material world (engineering, natural sciences…).

In contrast, order topology refers to the measurement of dominance of one element over the other with respect to a common attribute (Saaty, 2008). Order properties are related to human values and personal preferences and therefore operate in the opposite direction that metrics do. In this new paradigm, human judgement occurs previously to its measurement and this fact allows the analyst to operate with intangible factors (subjective) and tangible (objective) factors simultaneously in a reasonably justified manner. An example of order topology is found when an expert delivers an evaluation on whether the economic feasibility of a project should be measured in terms of the rate of investment (ROI)[8] or in terms of Net Present Value (NPV)[9] and which weight should be assigned to each indicator in case of a 'multi-indicator' evaluation. In this example, the expert, based on his previous experience or knowledge should deliver a relation of dominance of one element (ROI) over the other (NPV) with respect to a common attribute (the performance each indicator obtains in opinion of the expert in evaluating the economic feasibility evaluation of the project).

As stated before, priorities are set before the actual measurement of properties which can remain as an unknown quantity in the judgement process. Note how in the example, there is no need to know the numeric value of the ROI nor of the NPV of the project to decide which relationship of importance exists between them.

This example serves to illustrate how measurement according to the laws of order topology is inherently different than that of metric topology. This switch of Point-of- view allows a consistent and meaningful combination of diverse and unrelated parameters to make complex decisions in absence of a general construct of a relationship between given quantities often referred to as formulas (typically found in science). In

[8] **Return on investment** (**ROI**) is the benefit to the investor resulting from an investment of some resource. Mathematically it can be expressed as $ROI = \frac{gain\ from\ investment - co\quad of\ investment}{cost\ og\ investment}$ .A high ROI means the investment gains compare favorably to the investment cost. As a performance measure, ROI is used to evaluate the efficiency of an investment or to compare the efficiency of a number of different investments.

[9] **Net Present Value (NPV)** is the difference between the present value of cash inflows and the present value of cash outflows. NPV is used to analyze the profitability of a projected investment or project. The formula used is the following $\sum_{t=1}^{n} \frac{\mathrm{Ct}}{(1+\mathrm{r})^{t}} - \mathrm{Co}$ where C_t = net cash inflow during the period t; C_o = total initial investment costs; r = discount rate and t = number of time periods.

particular, this paradigm has been used in different models of Multi criteria decision analysis (MCDA) or Multi criteria decision making (MCDM) which will be the methodology adopted to select variables and add them to create the composite indicator.

4.3. Multi criteria decision making

Why is it necessary to adopt a multi criteria decision-making methodology?

As previously stated it is necessary to operate with a methodology that is able to combine multiple dimensions, objectives, stakeholders and scales of reference that are involved in a standard process of decision-making without renouncing to the accuracy, quality or the degree of quorum in the results. One of the strengths of the multi criteria methodology is the great amount of factors that can be integrated in the decision-making evaluation process. This characteristic is the reason why an approach Multi Criteria Decision Making (MCDM) approach will be used to conceptualize and develop the ranking methodology.

MCDM is a sub discipline of Operations Research that creates models to deal with decision problems under the presence of a number of decision criteria.

MCDM has been used extensively in industrial engineering and some of their applications are aimed at assessing decision making in many different areas: Integrated Manufacturing, technology investment decision evaluation, layout design among others.

The MCDM problem is an optimization with a set of simultaneous objective functions and a unique decision agent. In other words, an MCDM problem is aimed at evaluating a set of alternatives in terms of a set of decision criteria.
Mathematically an MCDM problem can be formulated as it follows:

$$\max F(x) \ \ where\ x\ \in X$$

$x\ is\ the\ vector\ (x_1, x_2 \dots x_n)$ of the decision variables, criteria or attributes. The analyst should try to choose the 'best' (or more appropriate) ones.
X is known as the feasible region this is the set of possible values that the decision variables may take. $F(x)$ is the vector $(f_1(x), f_2(x) \dots f_p(x)$ of the p objective functions that define the simultaneous objectives (or goals) of the problem.

The problem of the present thesis can be regarded as a particular case of MCDM and consequently will be treated as such. In this subsection, general theory of MCDM will repetitively be linked to the particular case that concerns the present thesis.
In particular, MCDM studies two different types of problems (Zimmerman, 1991): Multi-Objective Decision Making (MODM) is concerned with decision making when the space of choices/alternatives (X) is continuous whereas Multi-Attribute Decision Making (MADM) (the discipline that will be applied in this study) addresses the problem of decision-making when the decision space (X) is discrete.

Chen and Hwang (1992) identified a series of aspects that all MADM methods had in common and defined them. Their contribution will be helpful to a better understanding of the MADM methodology and will help to clarify the 'technical' previous definition of the MCDM problem.

- **Alternative**: Choice of action available to the analyst. In the case of study, each university to be evaluated constitutes an 'alternative'. Note that alternatives are a finite set and they are supposed to be analyzed and prioritized and most likely ranked.

- **Attributes/decision criteria/variables:** Attributes are the different dimensions from which the alternatives can be evaluated. The number of attributes to be considered in a decision process can range from a few to hundreds depending on the case, so it is a common practice to order them in a hierarchy. In the present study the decision criteria will be presented in *section 4.6.2.* given that further detail needs to be added to justify the analyst's attributes selection.

- **Incommensurable units:** Each attribute can be measured in a different unit. Being forced to consider together dimensions that are measured in different units might add a high degree of complexity to the MADM problem. If a company is trying to rent a new office and two decision criteria are to be considered (price and dimension) it is clear that each attribute will be measured in a different unit (euros and square meters respectively). The existence of mechanisms aimed at combining criteria of different nature (and measured in different units) is a particular strength of the MADM methodology. In the case of study there will be many criteria to be contemplated and the MADM will help to add them coherently.

- **Decision weights**: MADM methods often require the analyst to assign weights of importance to each decision criteria. A common methodology is that of **normalization** (see definitions) so that the sum of weights assigned to each attribute equals to one.

- **Decision matrix:** A matrix representation of an MADM problem can be a very useful way to represent a decision-making problem. A decision matrix A is an (M x N) matrix where each component of the matrix a_{ij} represents the performance of an alternative *Ai* in relationship to a certain decision criteria *Cj (for i=1,2...M and j=1,2…N).* M is the number of possible alternatives considered whereas N is the number of decision criteria involved in the decision making process.

In the particular case under study the set of alternatives are the different Engineering Schools under study and the decision criteria are the set of variables that are aimed at determining what makes a quality HEI as perceived by its own students (this will be explained with detail in *section 4.6.2.).*

4.4. Some MCDM methods:

MCDM methods are characterized for operating in three steps in any decision-making technique in which a numerical analysis of alternatives is involved:

1. Determining the set of relevant criteria and the possible alternatives.
2. Defining numerical values to the relative importance between criteria and to the performance of each alternative in relationship to such criteria.
3. Treating the numerical values obtained previously to define a ranking of each alternative.

In this section three of the main methodologies used to develop *step 3* will be presented to give a certain background to the reader. However, an alternative methodology will be used in the present study to provide a higher degree of consistency and coherence in the results.

The following techniques presented are known as methodologies to solve the aggregation problem.

4.4.1. The Weighted Sum Model

The Weighted Sum Model (WSM) is a methodology that has been used extensively in particular in single dimensional problems. With M alternatives and N attributes, the best alternative (the first in a ranking) is the one that satisfies the following equation (3) (Fishburn (1967)):

$$(4) \qquad A^{*} = \max_i \sum_{j=1}^{N} a_{ij}.W_j \quad for\ i = 1,2 \dots M$$

Equation 4: Weighted Sum Model aggregation formula

where: A* is the Weighted Sum Model score of the best alternative, N is the number of attributes, a_{ij} is the present value of the i-th alternative as measured by the j-th criterion, and W_j is the weight of importance of the j-th criterion. Under the WSM the total value of each alternative is obtained through the sum of products given by (3). In single-dimensional cases that is those in which all units are the same (i.e. meters, kilograms, marks…) the WSM results in a very effective technique.

An example might serve to illustrate a practical application of the method:

Example 2

Suppose that an MCDM problem involves three criteria, which are expressed in the same unit, and three alternatives. The relative weights assigned to the three criteria are determined to be: W_1 = 0.30, W_2 = 0.45 and W_3 = 0.25. The corresponding a_{ij} values are assumed to be the following:

$$A = \begin{pmatrix} 25 & 20 & 15 \\ 11 & 30 & 20 \\ 30 & 15 & 24 \end{pmatrix}$$

When equation (3) is applied on the previous data, the scores of the three alternatives are obtained: A_1 = 25×0.30 + 20×0.45 + 15×0.25 = 20.25. Similarly, A_2= 21.80 and A_3= 21.75 Thus, the best alternative (in a maximization case) is alternative A1 (achieves the highest WSM score). The ranking in this case would result into $A_2 > A_3 > A_1$ where '>' represents 'better than').

The reader should identify how the WSM is used in many different cases and areas, for example when evaluating the average score of a student at university given a series of grades that have a different ponderation or impact in the overall mark.

Difficulties in this model arise when trying to combine different dimensions and consequently working in different units. In this case, the additive utility assumption[10] *is violated and the result is equivalent to adding pears and apples.*

After this introduction to the WSM it seems clear that it is not a perfect way to aggregate criteria in multi-criteria decision processes given that there is no methodology that explains or defines how to solve the adding pears with apples problem in a coherent manner.

Moreover, weighted sums provoke a problem that has been referred as the **unsupported efficient alternatives** in the MCDM literature. Let's see what defines an efficient alternative. An alternative is **dominated** by another one if this other one has better evaluations on all criteria and a strictly better in any of such criteria.

The concept of **efficient alternative** is defined, then, as an alternative that cannot be dominated by another one. In this sense it seems that given a set of weights W any efficient alternative could potentially reach the top of a ranking if weights were chosen in an audacious manner.

However it has been proven that weighted sums are not able to provoke that any efficient alternative is set in the first place of a particular ranking. A simple example should serve to illustrate this idea. (Bouyssou et al. 2000).

Example 3

Alternatives	*Criterion 1*	*Criterion 2*
A	*5*	*19*
B	*20*	*4*
C	*11*	*11*
D	*3*	*3*

Table 3: Data that represents a theoretical MCDM problem under 2 evaluation criteria

In this problem a combination of 2 criterion needs to be maximized in order to set up a ranking between the 4 (A,B,C,D) alternatives. Note how three of the alternatives (A,B,C) are efficient (not dominated). However, the degree in which alternatives contribute to the maximization of the objective is different. Whereas A and B 'compensate' a bad performance in a certain criterion with an excellent performance in the other criterion C acquires a fair performance (though far from excellent) in each evaluation criteria. When aggregating the criteria using a weighted sum it is impossible to find a combination of weights that result into C heading the ranking.

Let's see the conditions that would make C stand at the top of the ranking:

1. *C ranked before A:* $11\alpha + 11(1-\alpha) > 5\alpha + 19(1-\alpha)\ which\ results\ into\ \alpha > 8/15 \approx 0{,}53$
2. *C ranked before B:* $11\alpha + 11(1-\alpha) > 20\alpha + 4(1-\alpha)\ which\ results\ into\ \alpha < \frac{7}{15} \approx 0{,}44$

See how conditions 1. And 2. Cannot be accomplished simultaneously and thus, C (an efficient alternative) will not be ranked first whatever the weights assigned to criteria 1 and 2. However, it is important to highlight that decision-making processes do not necessarily lead to an optimal solution because often a lack of information

[10] In plain words describes the assumption under which the utility of the whole is equal to the sum of the utilities of the parts.

conditions the choice taken and therefore forces the decision-maker to choose with imperfect information. This is called the assumption of bounded rationality treated with a higher degree of detail in footnote 12. However, in this example, given that it is an 'academic' problem an efficient (or optimal) choice should be pursued by the decision-maker (i.e. the ranking compiler).

Despite the flaws of WSM it works beautifully when aggregating elements that belong to the same dimension and thus it will be used as an aggregation method in the present study when the conditions are appropriate.

4.4.2. The Weighted Product Model

The Weighted Product Model (WPM) is another aggregation method that differs from the WSM because addition is substituted by multiplication. Alternatives are compared with each other by multiplying a series of ratios (equal to the number of criteria involved in the decision process). The ratios are raised to the power equivalent to the relative weight of the corresponding criterion. In general, in order to compare the alternatives A_r and A_l through this methodology the following product (Miller and Starr (1969)) has to be calculated:

$$(5) \qquad R(A_R/A_L) = \prod_{j=1}^{N} (a_{Rj}/a_{Lj})^{W_j}$$

Equation 5: Weighted Product Model aggregation formula

N is the number of criteria, a_{ij} is the actual value of the i-th alternative in terms of the j-th criterion, and W_j is the weight of importance of the j-th criterion. If the relationship $R(A_R/A_L)$ is greater than one, then alternative A_r is more desirable than alternative A_l (in the maximization case) and if the ratio is smaller than 1 that indicates that alternative A_l is more desirable than A_r. The best alternative is the one that is better than or at least equal to all the other alternatives. The WPM has an advantage in comparison to the WSM because its structure deletes the unit of measure and it is therefore called dimensionless analysis. In consequence, the WPM can be used both in single- and multi-dimensional decision-making problems. Another advantage of such method is that it allows the analyst to work with relative values between criteria performance instead of the real ones.

First let's define how to find a relative value. Note that a relative value defines a relationship between the actual performance of an alternative (R in this case) in relationship to a criterion j and the summation of all the alternatives that are to be evaluated according to such criterion (equation (5)).

$$(6) \qquad a'_{Rj}(relative\ value) = a_{Rj}(actual\ value) / \sum_{i=1}^{N} a_{Ri}$$

Equation 6: Expression to obtain a relative value

The analytic equation that demonstrates that operating with relative values between criteria is a consistent strategy can be found below (equation (6)):

$$(7) \quad \left(a_{Rj}/a_{Lj}\right) = \frac{a_{Rj}/\sum_{i=1}^{N} a_{Ri}}{a_{Lj}/\sum_{i=1}^{N} a_{Li}} = \quad a'_{Rj}/a'_{Lj}$$

Equation 7: Demonstration that operating with relative values is consistent.

Example 4:

Retake the problem presented in the previous example 2 (note that now the restriction to express all criteria in terms of the same unit is not needed). The WPM is applied (equation (5)) to the same decision matrix values and the following results are obtained: *R (A_1 /A_2) = $(25/11)^{0.30}$ × $(20/30)^{0.45}$ × $(15/20)^{0.25}$ = 0,992< 1, R (A_1/A_3) = 0.958 < 1, and R(A_2/A_3) = 0.966 < 1. Therefore, the best alternative is A_3, since it is superior to all the other alternatives. In this case, the ranking of these alternatives is the following: A_3 > A_2 > A_1.*

Note how the methodology used may change the order in which alternatives are ranked. If the WSM Ranking is used A_2 is the most desirable alternative whereas in the case of using the WPM the alternative that performs better is the alternative A_3. It is extremely important to take into account the effect of the methodology used in the results in order to choose a methodology that fits with the objectives of the study.

4.5. The Analytic Network Process

Note: The analytic network process will deserve a lot of attention given that has become the framework in which the ranking methodology has been articulated. After trying to develop a procedure that would suit the necessities of the present study, that is creating appropriate metrics to assess Academic Performance in Engineering Universities I realized that it was out of my scope trying to develop myself a mathematical methodology to aggregate different dimensions in a justified, logical and consistent way. Moreover I lacked knowledge and experience to carry out such a task and was afraid my intention to confront such a large and extremely difficult problem would result into me becoming frustrated and unable to 'get it done' .I would have liked to pursue the accomplishment of such task but this process has also served as an exercise of humility given that I have acknowledged my limitations and I have realized that though I would probably be able to develop this task, I was certainly not able to do it in the amount of time given/assigned. I have realized that an undergraduate thesis does not allow such a deep research project and I understood at the end that my legitimate will to address this problem in a correct manner would require at least a Master Thesis if not a Ph. D. Therefore, in the present thesis, the Analytic Hierarchy Process will be used to evaluate the quality of teaching in different Engineering Universities. Note how the approach of AHP keeps modeling the problem as a 'particular case' of an MCDM problem.

The following subsection is aimed at explaining what the AHP is, where it has been used, the particularities of the methodology and obviously how it has been applied to the present study to obtain results.

The Analytic Network process is a methodology of multi-criteria analysis developed at the end of the 70's decade at the Wharton School of Business by Ph. D. in mathematics Thomas L. Saaty. Such methodology provides support to decision-making in situations of complex structuration of the problem, under the assumption of bounded rationality[11] and where a plurality of decisional criteria (probably contraposed one to another) are involved. As time has gone by it has become one of the most-used multi-criteria methodologies.

The AHP has been used extensively in many different fields and by many different institutions from governments (US) in strategic decisions to companies Xerox, Ford, IBM, British Airways…) where confronting difficult decision-making processes (i.e. IBM used it in their quality improvement strategy to design its AS/400 computer (1988)). The AHP has also been used in student admissions, military personnel promotions or hiring decisions.

The AHP tries to cover all the steps involved in the process of decision-making:

1. Models the problem through a hierarchic structure (hierarchic decomposition)
2. Uses a priorities' scale based in the preferences of certain elements over the others with the objective of 'converting' the multiplicity of different reference scales.
3. Synthesizes the judgements emitted in order to yield a ranking of alternatives based on the priorities that have been expressed previously.

This methodology presents a way to order the analytic process of thinking into three basic principles:

1. The principle of hierarchy construction
2. The principle of compared judgements
3. The principle of logic consistency

4.5.1. The hierarchic decomposition

Complex systems can be better understood through its decomposition in its constituent elements.

This technique allows to convert a general complex problem into more and more simple problems in which to drift the attention. In doing that the analyst needs to maintain the interdependence between the subsystems so that at the end the correct assembly of the solutions of smaller problems can give a solution to the general problem. The hierarchic decomposition of the AHP allows to divide the system into simpler parts without losing the interactions between parts that conform the overall system.

The structuration of such elements/parts in a hierarchic structure and the posterior composition or synthesis of such judgements according to the relative importance of the elements of every level of the hierarchy is a fundamental requisite to use the AHP.

[11] **Bounded rationality** refers to the idea that when individuals are confronted with decision-making processes, their rationality is limited by the available information, their own cognitive limitations or the time available to make a choice. The approach of decision-makers under the 'bounded-rationality' assumption is normally the satisfactory one, the one that seeks to find a satisfactory solution more than an optimal one. In 1957 Simon presented bounded rationality as an alternative basis for the mathematical modeling of decision-making. Bounded rationality contrasts with the paradigm of 'rationality-as-optimization', which views decision-making as a fully rational process of finding an optimal choice given the information available.

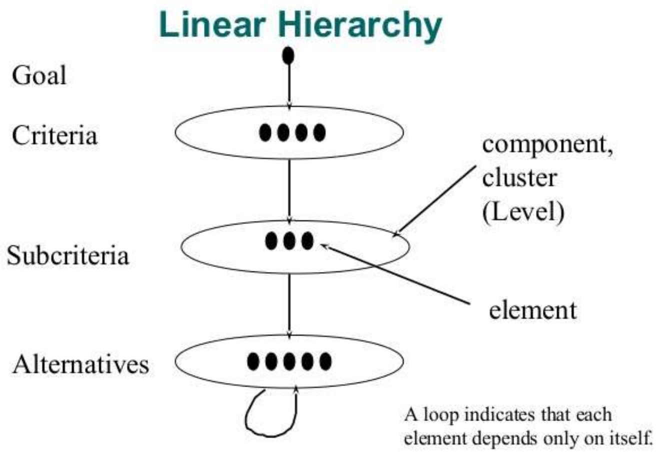


Figure 3: Linear Hierarchy architecture. Source: Saaty (2008)

The so-called functional/lineal hierarchy (figure 3) represents a complex problem using a structure of several levels. The first level represents the goal to achieve. Afterwards factors, decision criteria and decision sub-criteria are found in successive levels and at the lowest level the alternatives to solve the problem are found.

As it may be read in figure 3, each level of the hierarchy has a name (criteria, sub-criteria and so on). Similarly, we name elements to each criterion included in a given hierarchic level. Moreover In this explanation, and as a means of clarifying concepts, criteria in higher hierarchic levels will be referred to as 'father' nodes whereas those of inferior levels will be considered 'son' nodes. In this way each criterion 'father' will have a series of dependent criteria 'sons' that at the same time might have other 'sons'.

The objective of hierarchic decomposition is that of quantifying how much elements that belong to lower levels of the hierarchy contribute to the realization of their fathers' objectives in order to be able to quantify, at the end, how much each element at the lower level of the hierarchy contributes into the satisfaction of the general objective expressed by the goal. (Figure 3 can help to understand the concept).

The hierarchies that operate in the AHP are those that lead the system towards the common goal which can either be either the solution of a conflict, an efficient assignment or an output maximization. Each set of elements in a goal-oriented hierarchy occupies a certain hierarchic level in the general hierarchy. Then, each criterion in the first level (right below the goal level) may have a certain number of son elements that are hierarchically related to it. Similarly, elements in the second level might have sons in the following level and so on. The number of 'son' elements of a given element is not fixed but is regularly a 'small' number between 3 and up to 9. The AHP model assumes in general that elements in the same level must be of the same order-of-magnitude.[12]

When decomposing a problem into a hierarchy it is necessary to take into account some conditions:

---

[12] Posterior revisions of the model by the same author (Saaty) and other scholars have permitted to operate with the AHP taking into account criteria which do not share the same order-of-magnitude. However, partly because of the complexity that would be derived from it but mostly because it is not necessary in the present study, the case of comparing criteria that do not belong to the same order of magnitude will be omitted.

1. Each level needs to have intern independence. This means that criteria contained in the same level cannot be dependent to one another. In other words (in the graph-theory parlance) there are not arches that connect elements that belong to the same subgroup.
2. Every level needs to have external dependence of the precedent level. In other words, every node or element needs to at least be dependent of another element in a superior level of the hierarchy.

The decomposition of the problem using the AHP is very useful because reproduces the process of our mind when confronting complex problems. Often our mind selects macro-elements of the system and orders them under a decrescendo of importance; then, our mind assigns the various elements to the several groups of the same level in importance according to an importance criterion.

Note how in the AHP, hierarchic levels act as the physical quantities (i.e. length, volume or weight) of metric topology. In the same way that a meter cannot be added to a kilogram, elements of a certain level will not be compared, added or operated to one another. The principle of homogeneity of dimension in metric topology has an analogous 'principle' of homogeneity in the AHP methodology in which only criteria that depend of the same 'father' node can be compared.

4.5.2. The compared judgements

The second principle that needs to be highlighted regarding the AHP methodology is the judgment setting between the elements that configure a hierarchy. A scale of priorities is proposed as a means to 'get-rid-of' the multiple scales that exist among their components, elements or criteria.

In the paradigm of **metric topology** (revise section *4.2.*) in order to carry out any measurement there are two main things required: the existence of a **reference scale** and the existence of a **measuring instrument**.

However, how can measures be carried out in absence of one or both of those elements?

It is known that human beings can use their intuition or perception to assign levels of importance to distinct elements if such levels can be measured under an objective scale: for example, it is possible for us to determinate the weight of an object and given a set of n objects we will probably be able to rank them in a descending order. Before measurement scales were invented, people had no way to measure because there were no scales and the only way to do it was by comparing things with each other or against as a standardized method to determine their relative order. Probably this fact has made us able to order objects or elements in absence of both a **reference scale** and a **measuring instrument.** This can be made by evaluating elements in a relative way, in other words, confronting one element to another to establish which one of them 'performs' better. For example we can say that one element weights four times another one in absence of a scale of reference.[13]

[13] (i) Note how working with relative values does not require absolute evaluations. A relative value is a quotient (q) between two numbers $q = \frac{a}{b}$ but if the result of the quotient (relative value) can be obtained through an approximate

However this problem becomes a lot more difficult to solve when many criteria are involved and therefore it is not so easy to combine the orders obtained with respect to the different criteria to obtain a total order unless there is associated with each partial order a set of numbers that are in some sense commensurate so they can be combined using the numbers (weights or priorities?) associated with the criteria (Saaty, 2008). The AHP takes advantage of this forced order to solve the problem of comparing different units in a decision-making process as we will see later.

Another advantage of operating using the AHP methodology is that it makes it possible to confront not only objectively measurable elements but also elements which apparently are not. In the case of measurable elements the absence of a scale of reference might provoke that human judgements are not accurate enough (in absence of a measuring instrument) but anyway, the problem of comparison (which is at the end what is under study) will have been solved. In the case of not measurable elements we are equally familiar with delivering a judgement that regards a certain degree of importance, for example when we say that independence at work is three times more important than income when evaluating job satisfaction.

To summarize, the idea behind the principle of compared judgements is taking advantage of the fact that human beings generally perceive relationships between elements that describe a situation and are able to conduct pairwise comparisons between them in order to express their preferences of one to another. The synthesis of the group of such judgements delivers a scale of *preference intensity* between the total of compared elements. Therefore it is possible to integrate rational thinking with feeling, intuition and experience in a general/overall scale that allows comparison across dimensions.

Arrived at this point, the following step is stablishing a methodology to carry out pairwise comparisons between the elements that belong to the same hierarchic level.

In order to solve this problem, comparison matrices are probably the most convenient and efficient way to do it. Pairwise comparison matrixes need to be referred to a particular sub-tree of criteria or level. Once a level has been selected, the father node will represent the **control criterion** or in plain words the principle under which 'son' criteria need to be evaluated in priority. Thus, for every pair of sons criteria inside the subset considered the pairwise comparison will serve to determine the relative priority that every node (criterion) takes in order to 'satisfy' the criterion represented by the father node.

4.5.2.1. The matrix of pairwise comparisons

The **Matrix of Pairwise Comparisons** $\{A\}$ **(MPC)** is a square matrix constituted by a number of rows and columns equal to the number of nodes or criteria included in a given subset. The row i and the column j represent the criteria i and j that need to be confronted whereas the element $a_{ij} \in \{A\}$ represents the quotient between the absolute priority of each criterion (or the relative priority between criteria i and j) that the decision-maker needs to estimate according to his opinions.

---

judgement and therefore there is no need to know the value of either a or b; (ii) See how this new way of thinking based on human judgements is related to the concept of **order topology** presented in section *4.2.*

Estimating which Food has more Protein

Food Consumption in the U. S.	A	B	C	D	E	F	G
A: Steak	1	9	9	6	4	5	1
B: Potatoes	1/9	1	1	1/2	1/4	1/3	1/4
C: Apples	1/9	1	1	1/3	1/3	1/5	1/9
D: Soybean	1/6	2	3	1	1/2	1	1/6
E: Whole Wheat Bread	1/4	4	3	2	1	3	1/3
F: Tasty Cake	1/5	3	5	1	1/3	1	1/5
G: Fish	1	4	9	6	3	5	1

Figure 4: Example of MPC that delivers judgements on protein values of different aliments. Source: Saaty (2008)

As highlighted before, if the decision-maker had at disposal an instrument to measure absolute priority of criterion i and j, he/she could be able to determine the element a_{ij} as a quotient between $w_i \; and \; w_j$ assuming they are respectively the absolute priorities of i and j (equation (8)):

$$(8) \qquad a_{ij} = \frac{v_i}{v_j}$$

Equation 8: Determination of a matrix element aij when a scale of reference and a measuring instrument can be used

Note how in the case of figure 4, the estimations could indeed be objectively quantified with a measuring instrument though in this case they have been obtained through the preferences of the decision-maker. The decision maker has given a_{ij} values directly without trying to assign values to v_i or to v_j.

In situations in which the criteria and sub-criteria adopted for the evaluation of alternatives cannot be measured under an objective basis the decision-maker will not be able to quantify v_i nor v_j so the only possibility will be to give a direct value for a_{ij}.

To report this estimation the decision-maker will consider its own subjective judgements obtained through a critical analysis of the collected data and taking into account its own experience as well. It is also a common practice that decision-makers consult experts in a certain discipline in order to define their judgements accurately.

The evaluation supplied by the decision-maker for an element a_{ij} will be expressed through a numerical score that will represent the intensity of priority that element i assumes in relationship to j (which acts as the unit of reference). The semantic scale defined by Saaty is the table in which judgements are based:

Scale	Intensity of Relevance	Description
1	Same relevance	Both tasks contribute equally to the purpose.
3	A little more relevant	Experience and judging favor one task in relation to another.
5	Considerably more relevant	Experience and judging strongly favor one task in relation to another.
7	A lot more relevant	A task is a lot more strongly favored in relation to another.
9	Extremely relevant	An evidence favors a task in relation to another with the highest level of reliability.

Table 4 : Fundamental Scale of Saaty. Source: Saaty (2006) rev.

In table 4 the numerical values that a_{ij} may take according to the degree in which i is more or less important than j as considered by the decision-maker are stated. One of the advantages of using this scale is that it allows 'natural' language and allows to obtain experts' opinions and judgements without having to introduce the AHP methodology in a detailed way.

The fundamental scale of Saaty serves as a 'translator' from vague and undefined preferences or opinions into quantitative judgements. Having defined Saaty's scale it is easier to interpret the matrix in figure 4. If we analyze judgements delivered by the decision maker we in the first row we see that he has considered steaks to have extremely more proteins than potatoes and apples. Similarly, the decision maker has considered steaks to have considerably more proteins than cakes and so on.

A particular characteristic that needs to be commented is the fact that Saaty's scale only arrives till 9. This maximum intensity of relevance is used to represent that criterion i is extremely more important than criterion j. Note how this is coherent with the previous rule of same order-of-magnitude condition defined at 4.5.1.

The range of values that each a_{ij} element of the matrix may take is bounded from 0 to 9. a_{ij} takes a 0 value in the case that no priority can be set between i and j (we will see later how this avoids that the AHP can be used). In any other case a_{ij} takes strictly positive and most likely (but not necessarily) integer values.

Once the pairwise comparison matrix of a given subset has been obtained it needs to be transformed into a synthesis vector that defines an estimation of priority between all elements included inside such subset. In other words, a method should allow to transform the relationships between pairs of criteria into a single criterion that represents the priority of a given element inside the set of elements to which it belongs. This general priority called **local weight** expresses the degree in which a given element 'collaborates' into achieving the objective represented by its 'father' element in an absolute way. That is, how important is that element for the 'father' criterion.

Arrived at this point and in order to provide a clearer understanding of **local weights** it may be interesting to compare the **MPC** to the **matrix of adjacency** defined in graph-theory.

A little introduction to graph-theory basic concepts needs to be given in order to continue with the explanation.

Given an **MPC** such as $\{A\}$ there is a graph which represents this matrix.

$$\{A\} = \begin{pmatrix} 1 & 2 & 6 \\ 1/2 & 1 & 3 \\ 1/6 & 1/3 & 1 \end{pmatrix}$$

The dimension of the matrix (equal to the number of decision-criteria involved) defines the number of nodes that the graph needs to include. In the particular case of $\{A\}$ of dimension 3, the graph attached to $\{A\}$ will have 3 nodes.

A representation of the graph that represents the $\{A\}$ matrix is attached below:

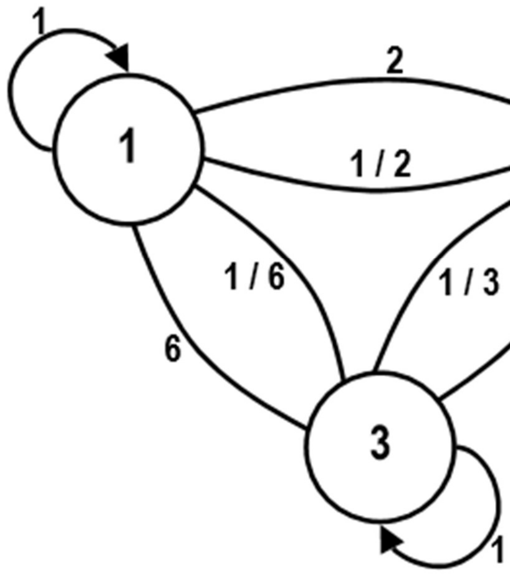


Figure 5: Graph that represents an MPC $\{A\}$.

In figure 5 'circles' represent nodes that account for the set of different criteria aimed at achieving an 'objective' which is defined by the 'father' node which is not included in the representation.

The graph that represents the $\{A\}$ matrix contains arches between each pair of nodes (i and j) and each arch is accompanied by a numeric value that comes to define the intensity with which a certain node (criterion) 'dominates' another node. Note how 'outside' arches define the domain relationships of i over j whereas 'inside' arches define the relationships of dominion of j over i. Relating those values to the matrix, note how outside bonds are defined by the upper-triangular part of the matrix whereas the lower-triangular part defines inside bonds. Each one of this numerical values expressing priority is called an **arch intensity (**$\boldsymbol{i} \underset{\boldsymbol{to}}{\rightarrow} \boldsymbol{j}$ in the case of outside arches and $\boldsymbol{j} \underset{\boldsymbol{to}}{\rightarrow} \boldsymbol{i}$ in the case of inside arches).

In a graph of this kind the j node is reachable from the i node not only through a direct pathway constituted by the arch $\boldsymbol{i} \underset{\boldsymbol{to}}{\rightarrow} \boldsymbol{j}$ but also through pathways that starting from i go through other nodes and finish in j. That means that priority is not only evaluated through direct pathways but rather by all the possible pathways that link an i node to a j node. Therefore, it seems that in order to determine the **local weight** associated to each one of the nodes involved it is necessary that for every pair of criteria i,j exists a pathway of length at least one between i and j**.**

Traducing this condition to the matrix approach determines that in order to be able to assign **local weights** to elements considered in an **MPC** all the pairwise comparisons need to be defined. That means that (a_{ij} needs to be strictly bigger than 0). This, however is not the only property that **MPC**s need to take into account.

The set of properties that **MPC**s need to have is presented below:

1. The confrontation of an element i against itself has always a relative weight (equation (8)) equal to 1.

$$a_{ii} = 1 \; \forall \; i = 1 \ldots n;$$

2. All the elements of the matrix define a confrontation of priority and thus, the matrix has only strictly positive elements.

$$a_{ij} > 0, \forall \quad i, j = 1 \ldots n;$$

3. Given an element a_{ij} the element a_{ji} is automatically defined.

$$a_{ij} = \frac{1}{a_{ji}}, \qquad \forall \; i, j = 1 \ldots n;$$

$$where\ n\ defines\ the\ number\ of\ criteria\ involved$$

As a consequence of the second property it is derived that graphs that are defined by **MPC**s need to be **strongly connected.**[14]

As a consequence of the third property the number of preferences that need to de defined to configure an **MPC** is equal to $c = \frac{n-1}{2}$

4.5.2.2. *Defining local weights.*

Let's continue now with the previous arguments of finding a method to define **local weights** aimed at synthesizing the information contained in the pairwise comparison matrix.

The **normalized local weight** that is defined through the concept of **local weight** is defined as the percentage in which a certain criterion *i* included in a certain set *(S)* contributes to achieve the objective expressed by his parent node. Consequently, the addition of all the local weights that belong to *S* has to sum 1 given that each criterion contributes in a diverse proportion such objective. This result will be justified in the following stages of the explanation.

Once it has been defined what needs to be found, let's link what we know about the matrix $\{A\}$ and how does it relate to the intuitive ideas that have been presented previously in relationship with graph theory. Previously it has been highlighted that (i) pairwise comparison matrices can be studied by graph theory and (ii) that a given node i has a prevalence over a node j which is linked not only because of the direct pathway between them but also because of other pathways that 'visit' other nodes along the way.

Therefore, **priorities** (or dominion relationships) or relevance of certain elements can be quantified taking into consideration the arch intensities of a certain pathway that links two nodes of a certain network. The priority of a given pathway is defined as the product of the intensities of the arches that configure such pathway.

[14] A graph is said to be **strongly connected** if every vertex is reachable from every other vertex.

A criterion assumes more or less relevance in relationship to the degree in which it is able to exercise a certain impact (either direct or indirect) to a greater number of criteria that belong to the same network of influence. Thus, to establish the **local weight** of a criterion i it is necessary to determinate the impact that it has over all the other n criteria (itself included) taking into account all the different pathways of whichever length (greater than 1) that take from i to any other node.

Note how in order to take into account how a node i influences all the other nodes of the network it is necessary to study not only the matrix of pairwise comparisons $\{A\}$ but also the matrices $\{A\}^k$ obtained by raising $\{A\}$ to a k-potence.

The matrix $\{A\}$ contemplates only the influence that nodes create to other nodes in the network through pathways of length one. However, to express the influence that nodes in the graph associated to $\{A\}$ create to each other $\{A\}^k$ needs to be considered.

Each element ${a_{ij}}^k$ of an $\{A\}^k$ matrix represents and considers the aggregated influence that node i has over node j throughout all the different pathways of length minor or equal to k that lead from i to j.

Mathematically this can be expressed as a summation (equation (9)):

$$(9) \qquad w_i^{(k)} = \sum_{j=1}^{n} {a_{ij}}^{(k)}$$

Equation 9: k-dependent local weight of sub-criterion i

Intuitively it can be implied that if $w_i^{(k)}$ is the sum of all the elements of the i-th row of the $\{A\}^k$ matrix it also represents a good estimate of the combined or aggregated influence that criterion i effectuates over all the other criteria (itself included) through pathways of length non-superior to k.

The same argumentation may be used for all the other nodes of the network and from doing that a vector of local weights the so-called **local weights vector** (**LWV**) can be obtained.

$$w^{(k)} = \begin{pmatrix} w_1^{(k)} \\ \vdots \\ w_n^{(k)} \end{pmatrix}$$

Note how this vector is a vector of absolute value. To obtain a normalization of the vector and therefore be able to define the percentage influence of a certain node i we are able no quantify the influence the node effectuates over his group through pathways of length not superior to k. *Note: normalized values will be noted as $wn_i^{(k)}$.*

The normalization is defined as follows (equation (10)):

$$(10) \qquad wn_i^{(k)} = \frac{w^{(k)}}{\sum_{i=1}^{n} w_i^{(k)}}$$

Equation 10: k-dependent normalization of local weight vector

The **normalized local weights vector** (**NLWV**) results into:

$$wn^{(k)} = \lim_{k\to+\infty}\begin{pmatrix} wn_1^{(k)} \\ \vdots \\ wn_n^{(k)} \end{pmatrix}$$

The issue of determining the vector of local weights is reduced then to the calculation of the k-potency of the matrix $\{A\}$ when $k \to \infty$. From a computational point of view this problem can easily be solved with an iterative approach where successive potencies of matrix $\{A\}$ are calculated and their respective wn are found. The iteration process stops when the results yielded for two consecutive approximations of wn are 'similar enough' that is when the variation between them is inferior to a previously fixed determined value $\varepsilon > 0$.

4.5.2.3. An algorithmic method to find local weights of an MPC

A possible implementation of an algorithm to find the normalized local weights vector could be the following:

00 The matrix $\{A\}$ of pairwise comparisons is defined. Each element in the matrix a_{ij}represents the preference of the decision-maker considering exclusively the direct impact (pathways of length one) of criterion i over criterion j;

01 The value of the w vector is obtained. Each element w_i considered under w is obtained sum of the i-th row of $\{A\}$ and represents the influence of i over other nodes through length one pathways;

02 The vector w is then normalized and wn is obtained which represents the first approximation of the **NLWV** of the considered subset;

03 The value of ε is fixed.

For a k iteration (where $k \geq 1$):

1. A potency of $\{A\}$ (most likely $\{A\}^{2k}$) is calculated. Such matrix will take into account for every criterion i, the impact that this one creates over criterion j considering all the pathways of length no greater than 2^k;
2. The vector $w^{(k)}$ is defined and so is $wn^{(k)}$. The later represents the k-th approximation to the exact **NLWV;**
3. $wn^{(k)}$ is compared with $wn^{(k-1)}$ according to the formula (equation (11)):

$$\Delta = \frac{|wn^{(k)} - wn^{(k-1)}|}{wn^{(k)}} \qquad (11)$$

Equation 11: Percentage variation of wn^k in comparison to wn^{k-1}

The iteration process finishes when $\Delta < \varepsilon$ for a particular iteration.

4.5.2.4. Consistency evaluation of an MPC

Once the **NLWV** has been defined the AHP allows to evaluate the feasibility of the obtained values through the expression of a judgement regarding the consistency of the judgements expressed by the decision-maker. The matrix $\{A\}$ defines the diverse judgements the decision-maker has effectuated as far as the relative importance of the elements confronted is concerned. Although the AHP allows that judgements reported by the decision-maker are inconsistent, inconsistency needs to be controlled given that alternatively such inconsistency might imply that the decision-maker has a low comprehension of the problem or has analyzed the information available inadequately.

The evaluation of the overall consistence of the priorities expressed by the decision-maker is measured through the **Consistency Index (CI)** which indicates the variance between values obtained of a parameter in the case of a given matrix (where the decision-maker has committed inconsistencies) and an ideal matrix (in which the decision-maker has been perfectly consistent).

Case 1: Decision-maker is perfectly consistent

To explain how **CI** is measured let's assume a case in which the decision-maker is perfectly consistent. In order to do so let's also assume that the decision-maker has a measuring instrument and is able to know the actual value of v_i and v_j and therefore of each a_{ij} of $\{A\}$. Note how in this case, rows and columns are proportional to each other (each column can be obtained as multiplying the first column by a scalar factor). From this it is implied that the range Rg ($\{A\}$) = 1. In such cases it is trivial to demonstrate how the **LWV** is defined by an **eigenvector**[15] of $\{A\}$.The following example comes to demonstrate this fact:

$$\begin{pmatrix} v_1/v_1 & v_1/v_2 & \dots & v_1/v_n \\ v_2/v_1 & v_2/v_2 & \dots & v_2/v_n \\ \dots & \dots & \dots & \dots \\ v_n/v_1 & v_n/v_2 & \dots & v_n/v_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix} = \begin{pmatrix} (v_1/v_1).v_1 + (v_1/v_2)v_2 + \dots + (v_1/v_n).v_n \\ (v_2/v_1).v_1 + (v_2/v_2)v_2 + \dots + (v_2/v_n).v_n \\ \dots \\ (v_n/v_1).v_1 + (v_n/v_2)v_2 + \dots + (v_n/v_n).v_n \end{pmatrix} = \begin{pmatrix} n.v_1 \\ n.v_2 \\ \dots \\ n.v_n \end{pmatrix}$$

The condensed expression of the previous matrices can be expressed as (equation 11):

$$(12) \qquad A.v = n.v$$

Equation 12: Matrix eigenvalue/vector problem in MPC consistent matrices

Note how in the previous case the **LWV** is easily obtained by multiplying any matrix column j by v_j. Remember how this is only valid when the pairwise comparison matrix is perfectly consistent.

The matrix $\{A\}$ as a square-matrix of order n has n **eigenvalues**[16] (among which some ones can be equal or alternatively having multiplicity greater than 1). As a **consistent** matrix $\{A\}$ has o unique eigenvalue different to 0. This eigenvalue will be referred to as λ_{max} given that is the eigenvalue of greater modulus. Mathematically this is expressed as:

$$\lambda_{max} \neq 0 \; and \; \lambda_i = 0, \forall i = 1 \dots n \; with \; i \neq max.$$

The **eigenvalues** of matrix $\{A\}$ are the same than those of its Diagonal or Jordan transformation. Moreover the **trace** of $\{A\}$ is invariant through a Diagonal or Jordan transformation process. Mathematically (equation (12)):

[15] An **eigenvector** is defined in linear algebra when as a result of multiplying a matrix $\{A\}$ for a vector (v), the same vector multiplied by a scalar value (called **eigenvalue**) is obtained.

[16] **Eigenvalue** (See previous note).

$$(13)\quad tr\ (A) = tr\ (D) = \lambda_{max} + \sum_{i \neq max}^{n} \lambda_i\ = \lambda_{max}$$

Equation 13: Matrix trace relationship with its eigenvalues

Each one of the elements over the diagonal of $\{A\}$ results to be 1 so from that $tr(A) = n$ and $\lambda_{max} = n$

To summarize, a **consistent MPC** (after applying principles of linear algebra) comes to have a series of **characteristics**:

1. *Proportional columns to one another. This means that its rang (rg) is the unit and as a consequence there is only one eigenvalue* $\lambda_{max} \neq 0$ *.*
2. *The eigenvector associated to* λ_{max} *is the* ***LWV*** *and consequently all of its components are positive.*
3. *The* ***LWV*** *can be obtained from any column of the matrix if it is multiplied by element* v_j*.*
4. *The normalization of* ***LWV*** *is the* ***NLWV*** *and can be obtained normalizing any of the columns of the matrix of pairwise comparisons.*
5. *The* ***LWV*** *is an eigenvector associated to the eigenvalue* $\lambda_{max} = n$

Therefore, in the case of consistency (equation (12)) is particularized as it follows (equation (14)):

$$(14)\qquad A.v = \lambda_{max}.v$$

Equation 14: Particularized matrix eigenvalue/vector problem in MPC consistent matrices

Case 2: Decision-maker commits certain inconsistencies

Having studied the case of consistency let's now take a look at the case of non-consistency in which relative priority between elements cannot be determined objectively and the decision-maker will be compelled to supply subjective judgements in regard to every a_{ij}.

In this case the matrix $\{A\}$ would not necessarily respect the relationship $A.v = n.v$; and consequently some **MPC characteristics** (i.e. *1,3…*) seen previously would not allow their application. However, this situation still makes it possible to apply the algorithm to find **NLWV** presented in section *4.5.2.3.*

In fact, it is possible to demonstrate (mainly throughout the application of linear algebra principles) the fact that even in the case of non-consistency equation (12) is satisfied. Nonetheless, in this scenario the **LWV** is not the eigenvector associated to the biggest module eigenvalue (λ_{max}) but to a different one (λ_i) which in general differs from (λ_{max}). The Consistency Index (CI), the Random Index (RI) and the Consistency Ratio (CR)

Let's approach the MPC $\{A\}$ as if an eigenvalue/eigenvector problem needed to be solved.

The trace of $\{A\}$ is equal to n ($tr\ \{A\}\ = n$) given that all the elements of the diagonal of an MPC are 1 independently of whether it is consistent or inconsistent.

In the case of consistency, the range rg$\{A\}$=1 and thus, there will only be an eigenvalue λ_{max} that solves equation (11). All the other eigenvalues $\lambda_{i \neq ma}\ = 0$.

In the case of non-consistency, the range rg$\{A\}$ is not necessarily 1.This means that there is at least another eigenvalue (λ_i) other than (λ_{max}) different to 0.

In the case of consistency let's decompose (equation (12)) in its components (equation (15)):

$$(15) \qquad \sum_{j=1}^{n} a_{ij} \,.\, v_j = \lambda_{max} . v_i$$

Equation 15: MPC decomposition as a summation

If from the previous equation λ_{max} is isolated (equation (16)):

$$(16) \quad \lambda_{max} = \sum_{j=1}^{n} a_{ij} \,.\, \frac{v_j}{v_i}$$

Equation 16: Development of equation 14

If equation (12) and (15) are combined we obtain (equation (17)):

$$(17) \;\; \lambda_{max} = n - \sum_{i=1, i \neq m}^{n} \lambda_i$$

Equation 17: Combination of equations 12 and 15

Now the objective is to find the average value of the diverse eigenvalues other than λ_{max}. In order to do so a new variable µ needs to be created. Mathematically this variable µ is defined as (equation (18)):

$$(18) \quad \mu = \frac{\sum_{i \neq max} \lambda_i}{n-1}$$

Equation 18: Definition of µ

Then, equation (16) can be rewritten as (equation (19)):

$$(19) \;\; \lambda_{max} = n - \mu(\mathrm{n} - 1)$$

Equation 19: Elaboration of equation 16

At this point the CI has been defined combining eq. (17) and eq. (18) as (equation (20)):

$$(20) \qquad |\mu| = \frac{\lambda_{max} - n}{n-1}$$

Equation 20: Definition of µ *in relationship to the maximum eigenvalue and the dimension of the matrix*

Note how given this formula in the case on non-consistency it is not necessary to find all the eigenvalues of the MPC given that the average value of eigenvalues other than λ_{max} has already been defined by $\mu_{inconsitent}$.

Note as well that in the case of consistency the value for μ_{cons}=0.

The **consistency index (CI)** has only sense in the case of inconsistency given that if the **MPC** is consistent its value will always be 0. The CI is equal to $\mu_{inconsitent}$ and is defined by equation (20). The consistency index gives a numeric value that gives an idea of degree of inconsistency of a given matrix. However, the consistency index alone is not enough to evaluate whether the judgements in the **MPC** are sufficiently consistent.

Intuitively it is easy to understand that the bigger the dimension of the MPC is, the greater are the chances that the judgements supplied by the decision-maker are inconsistent or contradictory to one another.

Therefore, it is not the CI of a given **MPC** what measures the consistency of the judgements expressed inside such matrix but the degree in which an **MPC** matrix (let's call it **A**) is more consistent than another **MPC** matrix (B) of the same dimension in which the judgements have been inserted randomly (that is without information nor knowledge of the phenomenon that is being analyzed).

Thus, in order to evaluate if a given **MPC** $\{A\}$ is consistent enough, we need to compare its **CI** with a **CI** called **random index (RI)**. A **random index (RI)** of matrices of dimension **n** is obtained by calculating the **CI** of a sufficiently big number of matrices (i.e. 50000) equal in dimension that have been filled at random. Once those (i.e. 50000) **CI**'s have been obtained they are averaged and as a result the **RI** is defined.

The RI's have been calculated for matrices of very different dimension. In the following table the RI's for matrices of dimension$\leq n = 10$ are attached:

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0,52	0,89	1,11	1,25	1,35	1,4	1,45	1,49

Table 5: Values of RI for matrices of n ≤ 10

The **consistency ratio (CR)** is what really defines whether an MPC is indeed consistent. This ratio is defined by the following equation (equation (21)):

$$(21) \quad CR = \frac{CI}{RI}$$

Equation 21: Consistency Ratio

To assure that the judgements are reliable enough it is necessary that the judgement inconsistency is kept in an inferior order-of-magnitude that the judgement themselves. This is equivalent to saying that the value that **CI** takes is maintained under the 10%, in other words that **CR**$\leq 10\%$.

If the **CR** remains under this 10% the judgements expressed by a given matrix are acceptable whereas in other cases the judgements delivered will have to be revised by the decision-maker.

The process of matrix construction and evaluation needs to be done repeatedly throughout all the hierarchy till all the **NLWV** have been obtained and the consistency of all the **MPC** matrices has been verified.

What to do if a matrix of judgements is not consistent enough?

If the inconsistency is greater than acceptable it is necessary to review the judgement's expressed by the decision-maker. Saaty (2006) proposes a technique to improve the consistency of the MPC through the identification of the most inconsistent judgement. Mathematically improving inconsistency means to induce λ_{max} tend to the dimension n of the matrix. The matrix element that influences the difference $\lambda_{max} - n$ the most is the element for which the magnitude $a_{ij}.\frac{wn_j}{wn_i}$ is maximum or minimum. The proposition made by Saaty is to substitute that element for another one defined as $a_{ij} = \frac{wn_i}{wn_j}$. An application example of this technique is seen in section 4.6.6.2.

4.5.3. The hierarchic synthesis

Once the two previous stages have been carried out (hierarchy construction and judgement comparison) the decision-maker has at disposal the hierarchic structure of the problem and the local weight of each criterion that describes the priority of each of them in respect to the respective 'father' control node. The local weights however still represent the priority of elements inside their group but not in respect to the father of the overall hierarchy. To evaluate the best strategy it is necessary to disassociate the priority of local weights in respect of their control criterion and traduce them into weights in the overall hierarchy. To do so a top-down logic procedure is followed. Starting from the father node (objective) we descend to the first level. The global weights of the first level are already defined (those are equal with their local weights). Then, we descend to the second level. At this stage, to define a global weight of a given element we multiply its local weight by the local weight of his control node and operating in this way its absolute weight is defined. This has to be repeated till all the absolute weights of all the elements of the hierarchy have been defined.

Example 5

The figure below shows an example of a problem that has been approached with the AHP methodology. In this case the objective is the acquisition of a car for the Jones family. To illustrate how to define absolute weights we will consider the sub criterion 'Purchase Price'. The local weight of that criterion is 0.25 and the local weight of its father criteria 'Cost' is also 0.25. The absolute weight of criterion 'Purchase Price' will be obtained as the product of those two local weights resulting into an absolute weight of 0.125.

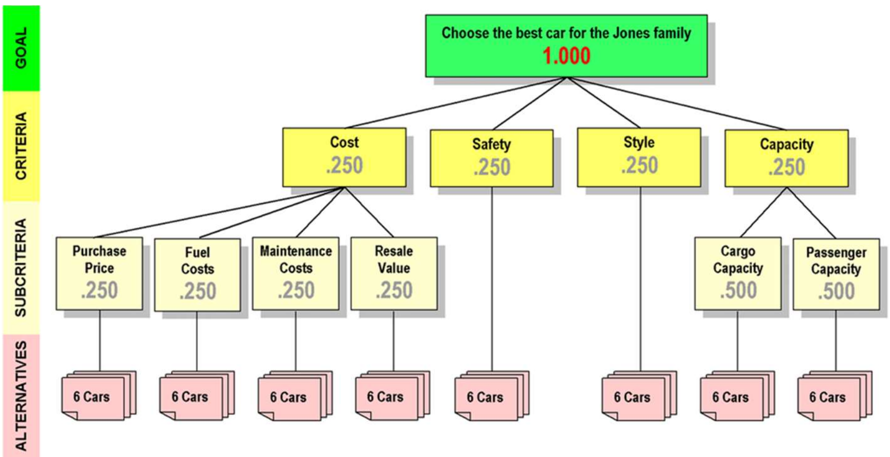


Figure 6: Example of an AHP hierarchic synthesis process

4.5.4. Making a decision

Once absolute weights have been defined for all the elements it is necessary to evaluate the different alternatives available in order to choose among them. Every alternative needs to be evaluated in respect to every criterion in the lowest level of the hierarchy. As it can be seen in figure 6 the alternatives are linked directly with the subsets at the very bottom of the criteria hierarchy. The evaluation of an i alternative in respect to a control criterion j allows to assign to the alternative i a 'mark' of how well does it performs as measured by the individual criterion (j) considered. Such judgement will then be weighted according to the importance that criterion j has into achieving the overall hierarchy's objective.

Once again, judgements need to be delivered by the decision-maker in order to define the capacity of a given alternative to respond to the requirements defined by a certain criterion of the hierarchy. Whereas the AHP provides a methodology analogous to the one that has been explained previously that operates using MPCs the evaluation in the present application of the AHP will not be made using MPCs. Instead, surveys will be used at this point to evaluate the performance of each alternative into the overall goal achievement.

The way alternatives will be evaluated in our methodology is explained in section 4.6.7.

4.5.5. Steps to be followed in when applying the AHP to a real case

In order to make the best decision the AHP method requires a series of steps that need to be defined:

1. **A definition of the problem:**

   At this stage in needs to be clarified the general objective of the decision process together with the actors involved in it. A description of the atmosphere in which the study will be developed is also required at this stage.

2. **A definition of the criteria involved:**
   This stage is where the array of decision criteria is selected.
3. **A definition of the actors involved:**
   The participants involved in the process of decision need to be carefully selected because of them depends the validity of the obtained results.
4. **Selection of the feasible alternatives:**
   This stage is where the different alternatives are defined and those that are feasible are chosen.
5. **Construction of the hierarchic model:**
   The problem is approached as a criteria and alternatives hierarchy. The criteria selected in step 2.
6. **Judgement setting:**
   Based on info obtained or the perception of the actors of the process the judgements are inserted inside all the MPC of the hierarchy. The local weights of each criterion of the hierarchy are obtained.
7. **Judgement synthesis:**
   The local weights obtained in 6 are used to define the absolute weight of each criterion. Once this is done the alternatives are compared in relationship with their performance according to the weights of the decision criteria involved. A ranking of alternatives is set.
8. **Decision making:**
   To validate the ranking of alternatives a sensitivity analysis is carried out to set the interval in which a change of the weights assigned to the multiple criteria does not imply a change in the ranking of the alternatives.
9. **Solution evaluation:**
   Once the decision has been taken it is possible to analyze whether the solution that has been obtained is suitable for the case of analysis.

4.6. Application of the AHP to the problem of evaluating Academic Performance at Engineering Schools

4.6.1. A definition of the problem

An analysis of major University Rankings has brought to light that in their pursue of objectivity they mostly forget to take into account the opinion of students which are one of their greatest (if not the greatest) priorities. University rankers focus their Rankings mostly in scientific performance and academic reputation without considering what is happening inside the university in relationship to teaching. Students have an imperfect knowledge about how well does a certain University teach and it is very difficult to get information about how Universities perform in teaching which is probably the most important asset of a university from a student's perspective. The fact that the field of study of the author of this thesis is Engineering, the study will try to evaluate the academic performance of a set of Engineering Schools (alternatives in the AHP parlance) to determine which one of them performs better. This process should not only serve to establish a rank of priorities between Engineering Schools but will also be useful for universities to notice what students value most and to try to focus their policies towards improving their performance specially in criteria that having been indicated as 'relevant' by students the University does not accomplish in a satisfactory way.

4.6.2. A definition of the criteria involved

In order to provide an accurate selection of criteria a public online brainstorming has been carried out. Colleagues in different engineering schools all around Europe have been asked to share their opinion on the main topic 'What constitutes a high quality Engineering School as far as teaching is concerned'.

Given that the analyst is an ETSEIB student and a former member of ESN (Erasmus Student Network) he has taken advantage of his personal network to acquire the opinions of engineering students of his network in different countries.

The criteria selected, thus, has been obtained through the analysis of spontaneous opinions and ideas of 24 engineering [17]students all around Europe (8 UPC, 3 TUM, 3 Université de Lorraine, 1 DTU, 2 HEC, 2 Roma Tor Vergata, 2 Roma La Sapienza, 2 Università di Bologna and 1 TU Delft). The requisites to be included inside the group of ideas creators were the following:

1. Being an engineering student.

2. Being a student having participated in Erasmus or double degrees programs or having been a student in at least two different engineering schools.

[17] The number of students involved is arbitrary. The analyst has contacted around 60 Engineering students though the number of final collaborators has resulted into 24. The greater the sample the higher the degree of 'collective intelligence'. However, it seems that taking into account the opinion of 24 people many ideas are repeated so probably that might be an indicator that the sample is 'big enough'.

In order to obtain the students' ideas a variation of the **nominal group technique** [18] has been used. In this particular study, students have been asked to share their ideas throughout a questionnaire of 4 main themes which had been previously sent to them.

Such 4 topics are presented below:

- Factors that contribute to high quality education in engineering fields.
- Practices you have seen in your own University that in your opinion contribute to achieve higher degrees of teaching quality.
- Practices you have seen in other Universities you have studied that in your opinion contribute to achieve higher degrees of teaching quality.
- Factors that assure Well-Being at University

After that, the analyst has collected all the ideas and together with other actors (presented in the following section) has decided which ones of them were relevant. The determination of their relevancy has been based in the number of apparitions of each idea. Every idea that has been expressed at least 2 times has automatically been considered a relevant criterion for further analysis. Discarding or accepting ideas when ideas have only been added for one person has been rather complicated. The difficulty has not been deciding which criteria were indeed relevant (because nearly all of them were) but synthesizing intuitive general ideas expressed by students into a concise criterion which would accurately describe the situation students were referring to. The work thus has been to redefine vague ideas into concrete concepts that would allow measuring and evaluation. At the end of the process, however, almost all the ideas have been included in some way. Considerations relative to how engineering should be taught presented at subsection 3.4 have been very useful at this stage.

After redefining, developing and concreting raw ideas a definitive criteria list has been obtained. Below the 33 criteria obtained are presented together with a brief description of each of them:

1. **Accessibility to facilities (FA):** This criterion relates to the amount of available computers, studying spots (i.e. in the library) group-working spaces and resting areas.
2. **Accessibility to labs (LA):** This item relates to the existence of laboratories prepared to carry out research experiments and the degree in which it is possible to access to such spaces anytime. The existence of practical courses in the curriculum might enhance higher scores in this criterion.
3. **Accessibility to older exams (AOE):** This criterion defines whether students have (in general) access to previous exams in an open way.
4. **Appropriate Bibliography (AB):** This dimension relates to the degree in which a complete, appropriate-for-the-course (that means it is not too broad nor too scarce) bibliography is defined by faculty staff to follow a certain course.
5. **Balance between Done-in-Class and appears in the exam (BCE):** Degree in which the exam evaluates the content of a particular course (and whether the distribution of the exam questions corresponds to the amount of time dedicated to each particular topic/theme).

[18] The reason for using this technique is that it is a variant of brainstorming that has already been used in in curriculum design and evaluation of educational institutions (Lomax & McLeman (1984), Chapple, M., & Murphy, R. (1996)). It is a group process that involves problem identification, solution generation and decision-making. In the present case, multiple changes have been made to the original version.

6. **Balance do-by-hand/ Software Approach (BHS):** Most engineering fields have developed faster ways to solve problems using computing. Engineering schools often try to teach their students the manual way to solve a problem and often this might derive into a rather out-of-date approach. However, only learning how a particular software is used might also be a wrong strategy because the student fails to understand what is 'behind'. The degree in which faculties achieve to prepare students not only to be able to solve problems manual but also to do it in an efficient, fast, real-world way (through software) is what this item is trying to measure.
7. **Balance Oral/Written Exams (BOWE):** Though evaluation is normally carried out through a written exam, oral exams provide also a high degree of info when it comes to evaluating students. A good engineering faculty should combine this two evaluating methodologies in order to properly prepare their students. The encouragement of oral presentations and oral exams are included into this item.
8. **Creative factor (CF):** The degree in which creativity is included in the competences acquired by the student. Creativity can either be fostered in the development of innovation-driven projects but also in exams where the student might have to imply facts that he intrinsically does not know and link them to other facts he has learnt in a particular course. Other creativity indicators might be the design or synthesis in contrast to the analysis (i.e. in mechanism theory the synthesis of mechanisms in contrast to the analysis of existing ones or in electrical engineering the design of circuits instead of the analysis of existing ones…). Practical design-of-experiments in statistics related courses might also be indicators of creativity.
9. **Case method (CM):** The case-method is a teaching approach that uses decision-forcing cases to put students in the role of people who were faced with difficult decisions anytime in the past. The degree in which engineering school asks students to devise and defend solutions to the problems at the heart of each case is considered under this parameter.
10. **Continuous Evaluation (CE):** CE refers to the degree in which a university examines its students continuously over most of the duration of their education. It is used as an alternative to a final examination system. The degree in which the university accomplishes to evaluate its students in such mode is included inside this dimension.
11. **Courses timetable/Overlapping problem (CTO):** This domain relates to the degree in which courses organization allows students to take the courses they want without experimenting overlapping problems.
12. **Deep analytical approach to problems (DAAP):** This item relates to whether students are compelled to solve problems analytically as a whole and 'till the end'. This item might be rather ambiguous or confusing but it tries to determine whether students are generally given the tools to solve a multi-disciplinary engineering problem without the need to resort to help of engineers of other disciplines, that is if their preparation enables them to divide problems in different kinds and solve each of those problems autonomously till the end (till the actual result). This relates with the capacity of integrating knowledge of other fields (i.e. fluid mechanics problems require a certain knowledge of differential equations to be developed appropriately) and whether the students are challenged to apply previous knowledge to actual problem solving. Students will often complain that a certain exam is evaluating aspects of other courses which have not been taught specifically in a certain course but it is important to prepare students to understand both the small sub-problems (and their specific solutions) and the 'big-picture' in which all the small analytical problems fit together.

13. **Degree of Students Participation (DSP):** This item is aimed at measuring whether students are active during lessons and encouraged to participate (giving their opinion, answering questions or even presenting how they solved a particular problem to the class).
14. **Direct access to publication and specialized magazine (PA):** The degree in which professors encourage students to revise academic publications, articles, bibliography and whether the university has subscriptions (Open-to-students) to follow journals related to fields of knowledge covered in the courses.
15. **Theory linked to practice (TLP):** The degree in which theoretical knowledge links to practical knowledge. In general gives an indication of whether both theory and practice are considered and how do they collaborate to help the engineer student develop holistically.
16. **Flexible Syllabus (FS):** The fact that students are able to select such courses they are most interested in. The fact that syllabus are flexible and students' preferences are taken into account.
17. **Hands-on Practice (HOP):** The degree in which students are compelled/offered the possibility to apply his engineering knowledge to develop practical works. (i.e. applying Heat-transfer theory to evaluate the efficiency of a building or applying control theory to control an induction motor…). The existence of the so-called practicum/practical works in most courses is a good indicator of this item.
18. **Interactive Resources (IR):** The existence of interactive resources such as moodle or interactive exercises with feedback at students' disposal and the fact that faculty staff make use of those. Online questionnaires or applets are included under this item as is the availability of recorded lectures online.
19. **Justice Perception (JP):** The degree in which the work developed by the student is recognized by the university standards. Certain universities might be too demanding with their students whereas others could be less demanding. The influence of this differences in the students' motivation is evaluated under this item. For example a university in which most of its students often fail might have a negative impact in the motivation of the student body. The same might occur when grading is strict and a tremendous effort of a student is 'answered' with a mediocre mark. Note that the opposite case can also have negative outcomes if, for example, students effort is always overvalued.
20. **Open Discussion (OD):** The degree in which class discussion is encouraged (obviously when a certain topic allows to be discussed).
21. **Peer Interaction: competitive vs collaborative (PI):** The degree in which students are encouraged to work in teams and whether the students' body is rather collaborative or individualistic.
22. **Projects and associations (P):** The degree in which the university supports and creates the framework so that students' can develop projects and join associations at university. Are also included under this dimension the presence of students' associations (such as ESN, AIESEC, BEST or similar) or competition teams (such as Formula Student or Smart Moto challenge etc).
23. **Possibility to retake exams (PRE):** The degree in which exams can be retaken (i.e. if a student fails an exam how often can he/she retake it and what is the 'cost of opportunity' for him/her).
24. **Problem Solving Hours (PSH):** The existence of hours aimed at problem-solving in which students can solve academic problems with a professor that advises/guides students. Problem solving is one of the most notable attributes of an engineer and thus it is indispensable to foster such competence.

25. **Quality Synthesis Material (QSM):** The quality of the synthesized bibliography given/available to students (slides or similar).
26. **Soft Skill Enhancement (SSE):** The degree in which soft-skills are fostered in the courses. Soft-skills examples are public-speaking ease or communication abilities.
27. **Startup Incubators and Entrepreneurship (SIE):** The degree in which university gives support to startups or entrepreneurship projects of its students. The presence of abundant projects subjects throughout the degree fosters higher evaluations of this item. The fact there is an infrastructure aimed to advice students when developing and materializing their projects is also included inside this item.
28. **Student and Professors Engagement (SPE):** The degree in which students and professors are engaged with the university. Note how a motivated student in an institution in which student engagement is not fostered might end up performing worse than in a university with a higher degree of student engagement. Analogously motivated students with professors that are not fully involved might also diminish the motivation of students.
29. **Student/Professor Collaboration (SPC):** The degree in which students are given the possibility to collaborate with faculty staff (Professors or Ph. D students) in their research, their publications or their projects.
30. **Student/Professor Info Exchange Efficiency (SPIEE):** The degree in which students are welcome to interact with professors. Abundance of professor consultancy hours, flexible meeting time setting or efficient e-mail exchange are the main factors to be considered under this category.
31. **Technical-Professional Competence Acquired (TPCA):** The degree in which a student acquires the technical competence in a determined field. Often knowledge on a particular field can be scarce for an eminently technical application. Being able to design, identify malfunctioning of a system are factors associated to Technical-Professional Competence whereas theoretical academic case study maybe not (i.e. an electronics course that allows the student to design electronic circuits is technical-professional competent whereas one aimed at explaining digital transistors history and principle of functioning is not.)
32. **Workshop Access (WA):** The degree in which a given faculty organizes/ gives access to workshop-kind events inside or outside the courses related to the field of study .Specialized conferences or fairs access are also included under this category.

33. **Winter Semester and Summer Semester (WSSS):** The fact that courses can be taken both in the summer as in the winter semester.

Once defined the criteria they have been ordered in a hierarchy. This hierarchy will be presented in *section 4.6.5*.

4.6.3. A definition of the actors involved

The actors involved in the process are the decision-makers. At this stage a variant of the **Delphi method**[19] has been used in order to give consistency to the judgements delivered by a unique decision-

[19] The **Delphi Method** is an ordered communication technique, originally developed as a systematic, interactive forecasting method which relies on a panel of experts. However in the present case the 'forecasting' will be substituted by

maker. The analyst has felt that being the only decision-maker could possibly lead to cognitive or decisional biases (see Kahneman (2011)). For this reason peer students have been chosen to act as consultants and their judgements have been taken into account and been given a relative weight equal to those of the analyst. The panel of experts has been configured by 4 other engineering students that have helped to provide judgements based on their beliefs. The students have been selected taking into account their field of study, their faculty, their University and their expectations[20]. The panel of experts has been configured by Martí Valero (Industrial Engineering, ETSEIB, UPC), Elisabeth Assens (Civil Engineering, ETSECCPB, UPC), Fabrizio De Lucia (Management Engineering, Engineering Faculty, Università degli Studi di Roma Tor Vergata) and Edu Mañas (Mechanical Engineering, Fakultat für Machinenwesen, TUM) and the developer of the present study Javier Morales (Industrial Engineering, ETSEIB, UPC). The process how this process has been carried out will be detailed in *section 4.6.6.*

4.6.4. Selection of the alternatives

Given the limitations in terms of time and budget the present study focuses exclusively in the study of a set of 4 alternatives (4 different Engineering Schools). It is clear that this supposes a scarce scope of the study given that this might not deliver as much information as the analyst would have liked. However, the list of alternatives could be increased in the future and the work carried out inside the present thesis could be re-used with a bigger group of alternatives.
In the present study, the Engineering Schools considered are the following[21]:

- UPC (in particular its faculty Escola Tècnica Superior d'Enginyeria Industrial de Barcelona)
- Facoltà d'Ingegneria dell'Università degli studi di Roma:Tor Vergata
- Facoltà d'Ingegneria dell'Università degli studi di Roma: La Sapienza
- Techincal University Delft

4.6.5. Construction of the hierarchic model

The objective that needs to be achieved is **academic performance** as defined in *section 3.1.*
To do so the 33 criteria presented in section 4.6.2 need to be arranged under father categories that directly link with the overall **academic performance** objective. Therefore, each son criterion will be included inside a more general category (father) with which it is related.
The set of in the fathers' categories and their definition is presented below:

---

the relevance assignment to the different criteria used conform the hierarchy to use the AHP methodology. For further information about the Delphi method see Linstone & Turoff (1975).

[20] The panel of experts has been chosen so that all collectives in the University are taken into account. Whereas some might want their education to be enterprise-focused others might want the University to be an ideas-exchange place. Some want to create innovative products whereas others might be more interested in research. The panel of experts counts with students that have contraposed views on many topics so that their judgements can represent the overall opinions of engineering students.

[21] The election of the universities under study is rather arbitrary. The selection is strongly conditioned by the incapacity of the analyst to effectuate on-site data collection without having to assume notorious costs. However, the subset of Universities takes into account three important factors. In the first place, it includes institutions from different countries. In the second place, it includes the home institution of the analyst (ETSEIB) which has provided the analyst with the knowledge and competence to develop this analysis. In the third place it has been deemed crucial to include a highly reputed technology university (TU Delft) in the subset.

- **Social Interaction and Motivation:** This category includes the social and motivational aspects at University. 'Professional' relationships with peers and faculty staff are taken into account together with aspects related with fairness perception or engagement which can be indicators of the degree of motivation of the student. Are included inside this category Justice Perception (JP), Peer-interaction (PI), Student Engagement (SE), Student/Professor Info Exchange Efficiency (SPIEE) and Student-Professor Collaboration (SPC).

- **Students' Resources:** This category comes to evaluate the resources available at University for students to develop as engineers out of-the-classroom. Under this category are included the following criteria: Labs accessibility (LA), Facilities accessibility (FA), Projects (P), Interactive Resources (IR), Startups Incubators and Entrepreneurship (SIE), Publications Accessibility (PA) and Workshops Accessibility (WA).

- **Class Dynamics:** This category includes the aspects that have to do with how classes are imparted. The degree in which classes impact in the students learning processes is taken into account in this category. Here belong Open Discussion (OD), Hands-on Practice (HOP), Case Method (CM), Degree of Students Participation (DSP), Theory linked to Practice (TLP) and Problem Solving Hours (PSH).

- **Competences Development:** This category refers to whether the university provides the students with the competences, skills and knowledge necessary to become prepared engineers. Under this category can be found criteria such as Deep Analytical Approach to Problems (DAAP), Balance do-by-Hand and Software (BHS), Soft Skills Enhancement (SSE), Technical- Professional Competence Acquisition (TPCA) or Creativity Factor (CF).

- **Evaluation Methodology:** This category refers to the ways students are evaluated. Fall into this category Balance Oral-Written Exams (BOWE), Possibility to retake Exams (PRE), Continuous Evaluation (CE), Balance Done-in-Class appears in the Exam (BCE) and Accessibility to Older Exams (AOE).

- **Organization**: This category contemplates whether the organization of the courses is correctly done. Fall into this category aspects related with timetables, syllabus flexibility or quality support material. Under this category are considered: Course Timetables-Overlapping (CTO), Winter-Semester and Summer-Semester (WSSS), Flexible Syllabus (FS), Bibliography appropriateness (AB) and Quality of Synthesized Material (QSM).

The aggregation of these 6 different categories provides a composite definition of what academic performance is. A formal definition of academic performance could be given at this point. The **Academic performance of an Engineering School** is the degree in which it achieves to provide the engineer with tools and skills to carry out his duties in the professional world taking into consideration that he/she has availability of resources at disposal, is evaluated according to logical methodologies, is enhanced to work collaboratively, has a certain grade of motivation and is able to feel his institution is well organized.

The AHP hierarchy obtained to achieve the objective of academic performance is shown in figure 7. Level 0 (highest) is the objective (orange), level 1 the criteria (grey) and level 2 (lowest) the sub-criteria (yellow).

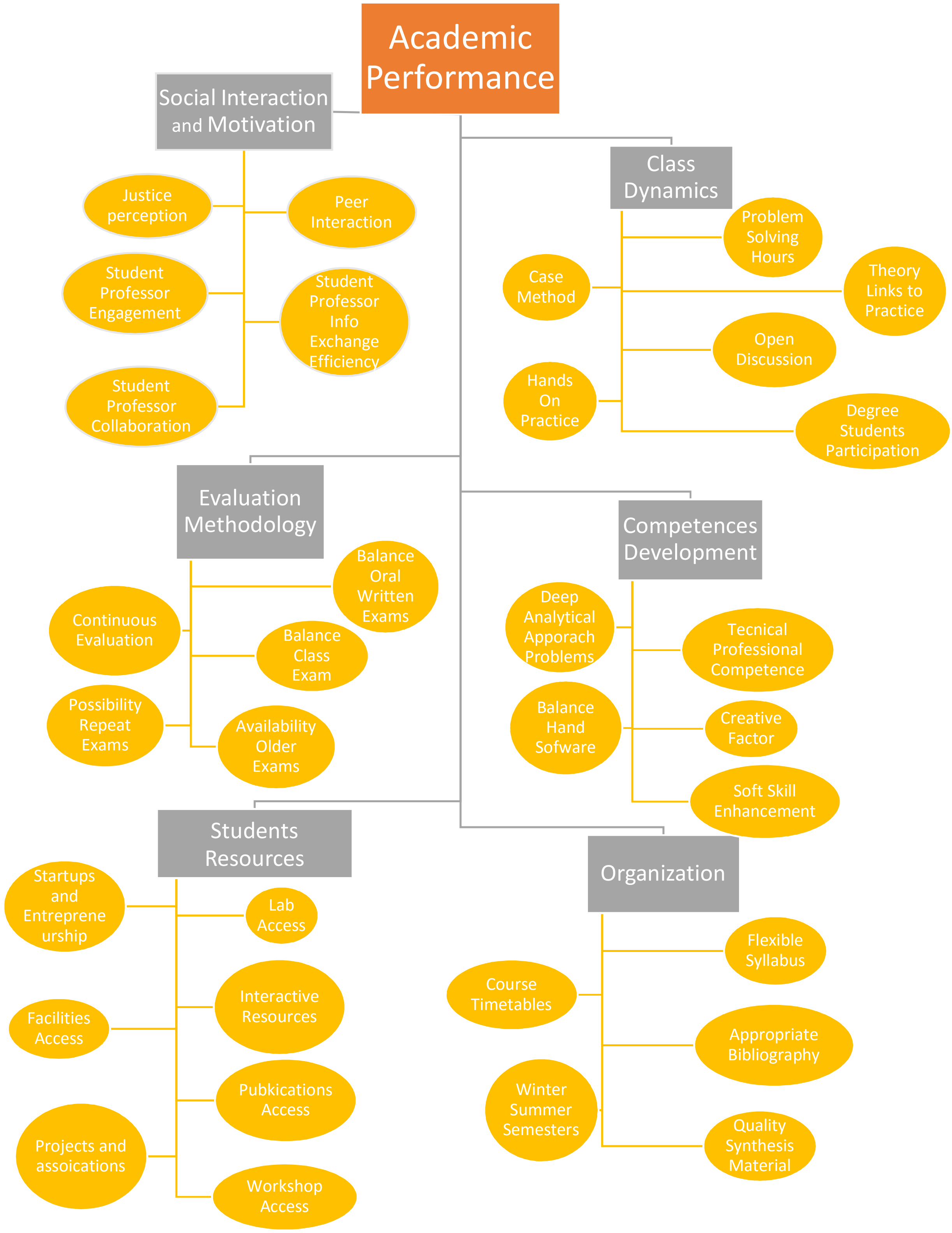


Figure 7: Organization chart of the Hierarchy. Objective in orange; Criteria in grey and Sub-criteria in yellow.

4.6.6. Judgement setting

How have judgements been inserted? Each expert has defined its preferences in relationship to the 6 nodes in the first level of the hierarchy. That means that a total of 30 matrices of pairwise comparisons have been filled (5 decision-makers have delivered judgements regarding the 6 sub-criteria in the second level of the hierarchy). The definition of pairwise comparison matrices and how they work has been explained in *section 4.5.2.1.*

An example will help to explain how the process has been carried out.

4.6.6.1. General process of judgement setting in relationship to a given node

Let's take the first father node of the hierarchy (**Social Interaction and Motivation)** and let's share the judgements expressed by the first decision-maker in the experts' panel in which he delivers his preferences between the sub-criteria associated to the father node. The set of sub-criteria to be compared are the following: Justice Perception (JP), Peer Interaction (PI), Students' Engagement (SE), Student Professor Info Exchange Efficiency (SPIEE) and Student Professor Collaboration (SPC). Given this number of sub-criteria the dimension of the MPC will be 5.

	JP	PI	SE	SPC	SPIEE
JP	1,0000	5,0000	3,0000	7,0000	3,0000
PI	0,2000	1,0000	1,0000	3,0000	1,0000
SE	0,3333	1,0000	1,0000	5,0000	1,0000
SPC	0,1429	0,3333	0,2000	1,0000	0,3333
SPIEE	0,3333	1,0000	1,0000	3,0000	1,0000

Table 6: MPC matrix of first decision-maker for father node Social Interaction and Motivation

The decision-maker considers JP to be considerably more relevant than PI and more relevant than SE to translate but a few of the numbers inside the matrix in *table 6* according to the fundamental scale of Saaty presented in section 4.5.2.1.

Now we will proceed to examine how this MPC is used to calculate the **normalized local weights** and the **consistency** of the judgements expressed inside the matrix.

Acquiring local weights:

In the first place a sum of the elements that conform each of the 5 rows is carried out. The first row is obtained by adding 1+5+3+7+3=19 and the second one by adding 0.2+1+1+3+1=6.2. We operate in this way for each row and a preliminary vector of local weights is obtained and presented in *table 7*:

JP	19,0000
PI	6,2000
SE	8,3333
SPC	2,0095
SPIEE	6,3333

Table 7: Preliminary local weights vector associated to node Social Interaction and Motivation (1st decision-maker)

Then it is necessary to normalize the vector. A variable that we will call vector sum is obtained by adding all the elements of the vector 19+6.2+8,333+2.0095+6.3333=41.8762. Each i-th element of the preliminary local weights vector is divided by variable vector sum: 19/41.8762=0.4537. The same is done with the second element 6.2/41.8762=0.1481. Operating in this way the preliminary **normalized local weight vector (NLWV)** wn^0 is obtained:

JP	0,4537
PI	0,1481
SE	0,1990
SPC	0,0480
SPIEE	0,1512

Table 8: Preliminary wn^0 normalized local weight vector for node Social Interaction and Motivation

Note how the sum of the elements in *table 8* equals to 1. See how the values obtained express the contribution of the sub-criterion i into achieving the objective defined by its father node **Social Interaction and Motivation.** The interpretation of the vector is quite intuitive: **JP** (Justice Perception) accounts for the 45.37% to the achievement of the objective Social Interaction and Motivation, **PI** (Peer Interaction) for the 14.81 % and so on.

However this is not the final normalized local weight vector (**NLWV**) yet because the study of the $\{A\}$ MPC of table 6 has only taken into account the length-1 interactions between sub-nodes in the **Social Interaction and Motivation** category. We need to take into account the interactions between nodes considering pathways of all lengths (why this is so has been explained in *section 4.5.2.1.*).

To do this we will make use of the algorithm presented in *section 4.5.2.3.*

We fix the value of $\varepsilon = 0.01$.

For a k iteration (where $k \geq 1$):

A potency of $\{A\}$ (most likely $\{A\}^{2k}$) is calculated. In this case we calculate A^2:

5,0000	18,3333	15,4000	53,0000	16,3333
1,4952	5,0000	4,2000	15,4000	4,6000
1,9143	6,3333	5,0000	18,3333	5,6667
0,5302	1,9143	1,4952	5,0000	1,6286
1,6286	5,6667	4,6000	16,3333	5,0000

Table 9:(MPC)^2 matrix of first decision-maker for father node Social Interaction and Motivation

We proceed to find the values for **NLWV** as explained previously (adding elements in each row of the column and normalizing the values obtained).
The results obtained for wn^1 are attached in the following table 8:

JP	0,49164488
PI	0,13964673
SE	0,16945652
SPC	0,04807984
SPIEE	0,15117203

Table 10: Preliminary wn^1 normalized local weight vector for node Social Interaction and Motivation

Finally we compare the results obtained in wn^0 with the ones obtained in wn^1

$$\Delta = \frac{|wn^1 - wn^0|}{wn^1}$$

And we obtain the following vector:

0,07714194
-0,0602145
-0,1743385
0,00192545
-0,0004462

Table 11: Comparison of vectors wn^0 and wn^1

Given that $\Delta > \varepsilon$ (clearly 0.077 and 0.174 are bigger than 0.01) we need to keep iterating. We repeat the process considering $\{A\}^4$ (k=2) and still $\Delta > \varepsilon$. The iterative process converges for (k=3) where $\Delta < \varepsilon$.

The values obtained for wn^3 are the definitive values for the normalized local weights vector associated to the node **Social Interaction and Motivation** according to the first expert 1. The results for wn^3obtained are the following:

0,48982466
0,13871511
0,17082447
0,04909452
0,15154124

Table 12: Definitive wn^3 normalized local weight vector for node Social Interaction and Motivation

However having obtained this vector does not warranty that this vector can be deemed as relevant. It is necessary to verify the consistency of the judgements to give credit to the obtained vector.

Evaluating the of the consistency judgements:

To evaluate the consistency of the judgments in the MPC we need to obtain the product presented in *equation (12)*:

$$A.v = \lambda_{max}.v$$

Where A is the MPC considered in table 6 and v is the normalized local weights vector obtained and presented in table 10. By doing this λ_{max} can be obtained easily as a sum of the elements of the $A.v$ product resulting vector.

$$A.v =$$

2,4942
0,7063
0,8698
0,2500
0,7716

λ_{max} is obtained as 2.4942+0.70630+0.8698+0.2500+0.7716=5.0919. Note how the matrix has a certain degree of inconsistency. In section 4.5.2.4 (case 1) we defined how the value of λ_{max} for consistent matrices was equal to the MPC dimension. Here, instead the λ_{max} = 5.0919 greater than 5. Now we can use the expression defined in *equation (20)* to find the value of the Consistency Index (CI) of the matrix.

$$\mathrm{CI} = \frac{\lambda_{max} - n}{n-1} = \frac{5.0919 - 5}{5-1} = 0.023$$

Now, using the Random Index (RI) value assigned to matrices of dimension 5 (with an RI value=1,11 as presented in *section 4.5.2.5*) it is possible to determine the value of the consistency ratio (CR):

$$CR = \frac{CI}{RI} = \frac{0.023}{1.11} = 2{,}07\ \% \ \leq 10\%$$

In this case, therefore the judgements expressed by the first decision-maker regarding the **Social Interaction and Motivation** node are consistent (or better said not-too-inconsistent).

In the example case the judgements expressed by the first expert have resulted to be non-inconsistent. What to do if MPCs result to be inconsistent?

4.6.6.2. Judgement revision in relationship to a given node

As explained in Section 4.5.2.5. there is a mechanism which can be used if the CR of an MPC of the hierarchy is above the 10%. We will take the example of a case found in the current study. When defining the MPC of father node **Organization** it has been found that decision-maker 3 has been inconsistent when giving his judgements.

	CTO	WSSS	FS	AB	QSM	LWV	Nlwv	λmax	
CTO	1,0000	3,0000	1,0000	0,2000	0,3300	5,5300	0,1210	0,63776	CI
WSSS	0,3333	1,0000	0,1429	0,1429	0,2000	1,8190	0,0398	0,2163	0,1643
FS	1,0000	7,0000	1,0000	1,0000	1,0000	11,0000	0,2406	1,23876	CR=CI/RI
AB	5,0000	7,0000	1,0000	1,0000	0,3333	14,3333	0,3136	1,53262	0,1480
QSM	3,0303	5,0000	1,0000	3,0000	1,0000	13,0303	0,2850	2,03189	
						45,7127	1,0000	5,6573	

Table 13: MPC of expert 3 of node father Organization. LWV, NLWV, CI and CR are calculated on the right-hand side.

The matrix presented in table 11 let's call it A appears to be too inconsistent. Note how the value of CR is 14,8%>10%. Thus, it is not possible to conclude that the judgements expressed by the decision-maker are coherent enough to be integrated into the overall hierarchy. Arrived at this point it is necessary to go back to the expert and ask him to re-evaluate the preference he has given to a particular confrontation. To find which pairwise comparison is the most responsible of A's inconsistency we will create a matrix B. B is a transformation of A under according to which each b_{ij} is defined as $b_{ij} = a_{ij} . \frac{wn_j}{wn_i}$. We will ask the expert to re-evaluate his judgements on the box b_{ij} that takes the maximum or minimum value. Matrix B is defined as:

1	0,98682511	1,98915009	0,51838457	0,77757685
1,01335079	1	0,86387435	1,12565445	1,43265112
0,50272727	1,15757576	1	1,3030303	1,184573
1,92906977	0,88837209	0,76744186	1	0,3030303
1,28604651	0,69800664	0,84418605	3,3	1

Table 14: Matrix B to re-evaluate the judgements expressed in MPC inconsistent matrix (Node Organization, decision maker 3)

It is observable that in B the element (i=4,j=5) takes the minimum value. From B we have noticed thus that a_{45} is the position that can be made responsible for A's high degree of inconsistence. We need to ask the expert if a_{45} could take the value $a_{45} = \frac{wn_4}{wn_5} = \frac{0,3136}{0,2850}$=1,11 where wn_4 and wn_5 are the **normalized Local Weights** of criteria **Appropriate Bibliography (AB)** (4-th row of A) and **Quality Synthesis Material (QSM)** (5-th row of A) respectively (table 11).

Note how asking the expert to change the value of a_{45} is equivalent to asking him whether QSM and AB could be deemed as equally relevant instead of considering QSM moderately more relevant than AB as the expert has considered in the first place. In this case the decision-maker agrees and a'_{45}=1,11 and A is transformed into A':

	CTO	WSSS	FS	AB	QSM	LWV	Nlwv	λmax	
CTO	1,0000	3,0000	1,0000	0,2000	0,3300	5,5300	0,1246	0,6447	CI
WSSS	0,3333	1,0000	0,1429	0,1429	0,2000	1,8190	0,0410	0,2158	0,0653
FS	1,0000	7,0000	1,0000	1,0000	1,0000	11,0000	0,2478	1,24588	CR=CI/RI
AB	5,0000	7,0000	1,0000	1,0000	1,1000	15,1000	0,3402	1,76885	0,0588
QSM	3,0303	5,0000	1,0000	0,9091	1,0000	10,9394	0,2464	1,38593	
						44,3884	1,0000	5,2612	

Table 15: Transformation of MPC of expert 3 of node father Organization to obtain a higher degree of consistency. LWV, NLWV, CI and CR are also calculated on the right-hand side.

In this new matrix the judgements delivered by the decision-maker are already consistent enough. Note how the value obtained for the CR is $5{,}88\ \% \leq 10\%$.

For now it has been explained how a decision-maker puts his preferences inside the **MPC**, how the **MPC** is used to define the **NLWV** and how the consistency of the **MPC** is evaluated. This process needs to be done for each one of the 30 matrices (6 father nodes in the first level per 5 decision-makers) that are built to articulate the hierarchy.

It is now necessary to consider the aggregation problem. How can judgements delivered by the different experts be aggregated into a synthesis MPC?

4.6.6.3. Synthesis matrices

Following with the example of the **Social Interaction and Motivation** node five matrices have been obtained, one for each decision-maker/expert. In order to put them together it is necessary to create a synthesis matrix. The methodology followed to build it is the **WPM** (Weighted product model) presented in section 4.4.2. based in assigning weights to each decision-maker's judgements and add them 'geometrically' (see eq. (21)). In the present study there will not be experts whose opinions are more valuable than the others so weights will be set equal for each one of the decision-makers. Each element of the synthesis matrix $a_{ij\,synthesi}$ will be obtained according to the following expression (equation (21)):

$$a_{ijsynthesis} = (a_{ijex} \;. a_{ijex} \;. a_{ijex} \;. a_{ijex} \;. a_{ijexp5})^{1/5}$$

Equation 22: Aggregation methodology to obtain the MPC synthesis matrix throughout the aggregation of each decision-makers' judgements

This way the 5 matrices will be aggregated to conform a unique synthesis matrix which will take into account the preferences expressed by the panel of experts.

The processes explained in sections 4.6.6.1 and 4.6.6.2 are also valid for synthesis matrices so they will also be applied to each synthesis matrix. The **Social Interaction and Motivation** node's synthesis matrix obtained through the application of equation (21) is:

					LWV	nLWV	λmax	
1	0,58181	1	1,01718	1,1847	4,7837	0,18037	0,905909	CI
1,71877	1	1,5518	2,37144	1,4509	8,0929	0,30515	1,526746	0,0058
1	0,64439	1	1,24573	1,1247	5,0149	0,18909	0,951274	CR=CI/RI
0,98311	0,42168	0,8027	1	1	4,2075	0,15865	0,783184	0,0052
0,84412	0,68925	0,8891	1	1	4,4225	0,16675	0,856088	
					26,521		5,023202	

Table 16: Synthesis MPC matrix for father node Social Interaction and Motivation. Local Weight Vector, Normalized Local Weight Vector, Coefficient Index and CR are also calculated on the right-hand side.

In table 16 the value of **nLWV** corresponds to its preliminary value. After the iteration process described in section 4.6.6.1 the definitive **nLWV** values for the **Social Interaction and Motivation** node are obtained (Table 17):

Social Interaction and Motivation	Local Weights Vector
Justice Perception	0,18056
Peer Interaction	0,30353
Student Engagement	0,18940
Student Professor Collaboration	0,15604
Student Professor Info Exchange Efficiency	0,17047

Table 17: Social Interaction and motivation node's local weights

Similarly all the synthesis matrices associated to the multiple father nodes have been determined as have their respective local weights vectors. The results are attached in the following tables:

Student Resources	Local Weights Vector
Accessibility to labs (LA)	0,16385
Accessibility to facilities (FA)	0,17859
Projects and associations (P)	0,18834
Interactive Resources (IR)	0,16418
Startup Incubators and Entrepreneurship (SI)	0,07566

Direct access to publications and specialized magazines (PA)	0,09019
Workshop Access (WA)	0,13919
Class Dynamics	**Local Weights Vector**
Open Discussion (OD)	0,13272
Hands-on Practice (HOP)	0,16635
Case Method (CM)	0,09280
Degree of students participation (DSP)	0,15377
Theory links to practice (TLP)	0,25190
Problem solving seminars (PSS)	0,20246
Competence Development	**Local Weights Vector**
Deep analytical approach to problems (DAAP)	0,16391
Balance do-by-hand/ Software Approach (BHS)	0,20448
Soft Skill Enhancement (SSE)	0,13865
Technical Professional Competence Acquired (TPCA)	0,31295
Creative Factor (CF)	0,18001
Evaluation Methodology	**Local Weights Vector**
Balance Oral/Written Exams (BOWE)	0,07916
Possibility to retake exams (ERP)	0,19969
Continuous Evaluation (CE)	0,25270
Balance between Done-in-Class and appears in the exam (BCE)	0,20741
Accessibility to older exams (AOE)	0,26104
Organization	**Local Weights Vector**
Courses timetable/Overlapping (CTO)	0,18671
WS vs SS (WSSS)	0,05670
Flexible syllabus (FB)	0,19441
Appropriate Bibliography (AB)	0,26925
Quality synthesis material (QSM)	0,29293

Table 18: local weights of all nodes in the second level of the hierarchy

The precedent tables present the local weight that each sub-criterion in the second level of the hierarchy has into achieving the objective expressed by its father node in the first level of the hierarchy. The following table (table 19) expresses, on the other hand, the local weights that each criterion in the first level of the hierarchy has into achieving the overall objective 'academic performance' in the level 0 of the hierarchy.

Academic Performance	**Local Weights Vector**
Motivation and Social Interaction (MSI)	0,13255
Students' Resources (SR)	0,11861
Class Dynamics (CD)	0,21798
Competences Development (DC)	0,23877
Evaluation Methodology (EM)	0,13587
Organization (O)	0,15622

Table 19: local weights of all nodes in the first level of the hierarchy

Note: All the matrices that have been used to obtain these results can be found in the Annexes.

4.6.7. Judgement synthesis:

4.6.7.1. Hierarchy synthesis

The normalized local weights obtained in *4.6.6* will be used to define the absolute weight of each criterion at the lowest level of the hierarchy. The absolute weight will be obtained as a product between local weights in the first level and local weights in the second level. For example, sub-criterion **Flexible Syllabus (FS)** which has a local weight of 0.19441 and its parent node **Organization** has a local weight of 0.15622. The absolute weight of **FS** will be 0.19441x0.15622= 0.030371.
According to this, all absolute weights for all the 33 nodes in the lowest level of the hierarchy can be obtained.

Social Interaction and Motivation	Justice Perception (JP)	3,009%
	Peer interaction: competitive vs collaborative (PI)	5,059%
	Student engagement (SE)	3,157%
	Student/Professor Collaboration (SPC)	2,601%
	Student/Professor Info Exchange Efficiency (SPIEE)	2,841%
Student Resources	Accessibility to labs (LA)	2,731%
	Accessibility to facilities (FA)	2,977%
	Projects and associations (P)	3,139%
	Interactive Resources (IR)	2,736%
	Startup Incubators and Entrepreneurship (SIE)	1,261%
	Access to publications and specialized magazines (PA)	1,503%
	Workshop Access (WA)	2,320%
Class Dynamics	Open Discussion (OD)	2,212%
	Hands-on Practice (HOP)	2,773%
	Case Method (CM)	1,547%
	Degree of students participation (DSP)	2,563%
	Theory linked to practice (TLP)	4,198%

	Problem solving seminars (PSS)	3,374%
Competences Development	Deep analytical approach to problems (DAAP)	2,732%
	Balance do-by-hand/ Software Approach (BHS)	3,408%
	Soft Skill Enhancement (SSE)	2,311%
	Technical Professional Competence Acquired (TPCA)	5,216%
	Creative Factor (CF)	3,000%
Evaluation Methodology	Balance Oral/Written Exams (BOWE)	1,319%
	Possibility to retake exams (ERP)	3,328%
	Continuous Evaluation (CE)	4,212%
	Balance between Done-in-Class and appears in the exam (BCE)	3,457%
	Accessibility to older exams (AOE)	4,351%
Organization	Courses timetable/Overlapping (CTO)	3,113%
	WS vs SS (WSSS)	0,943%
	Flexible syllabus (FB)	3,234%
	Appropriate Bibliography (AB)	4,270%
	Quality synthesis material (QSM)	5,107%

Table 20: Absolute weights (in percentage) of all decision criteria at the lowest level of the AHP hierarchy

The list of absolute weights of each node defines the relative weight that each criterion has into achieving the AHP hierarchy's objective (academic performance).
Once this has been done the alternatives need to be compared in relationship with their performance according to the weights of the decision criteria involved that have been presented in *table 20*. The following section 4.6.7.2. is intended to explain how performance evaluation of each university has been carried out.

4.6.8. Alternatives performance evaluation

Normally the AHP makes use of MPCs to evaluate the performance of a determined alternative in relationship to the different criteria involved in the decision-making process. That is, an expert or a set of experts are asked to deliver judgements regarding which alternative performs better in regard to a specific criterion. In the present case of study, however, this would mean asking to each of the 5 experts presented in *section 4.6.3* to deliver their judgements on how well each of the four Universities under study performs according to each of the decision criteria set in *section 4.6.2.* As the reader might imagine the experts have at this point abandoned their expert status. Experts on a subject or system are familiarized with how it operates, how it behaves and how it works etc. and it seems the panel of experts defined in *section 4.6.3* are not experts in evaluating an institution they have never been to nor know much about. Arrived at this point there is a need to make a controversial decision (which can certainly be questioned) that has, nonetheless, seemed the most appropriate in opinion of the analyst. The decision taken at this stage is particularly sensitive of creating controversy because the AHP does not normally need a statistical analysis to work given that alternatives (Higher Education Institutions in our case) are evaluated either by a single decision-maker or by a panel of experts. In this case it has been deemed appropriate to evaluate alternatives through a survey where former panel of 5 experts has been substituted by a new panel of experts for each University.

Performance evaluation has been carried out through a survey developed at each one of the four universities under study (ETSEIB, Tor Vergata, Sapienza and TU Delft). The respondents to such survey have been the 'new' experts in performance evaluation.

Note: At this point the analyst would have liked to take the time to approach the subject of the index creation using a statistical approach instead of a MCDM one. Some hindrances have appeared along the way and have made it impossible. Data collection presented in the following subsection should have been used as data to be analyzed with statistical techniques such as Principal Components Analysis (PCA), Factor Analysis (FA) or Cronbach Alpha (CA). These tests (PCA and FA) are aimed to reveal how different variables change in relation to each other and whether they are associated. The objective of such analyses is to transform a set of correlated variables into a set which is uncorrelated using a covariance matrix (or if standardized called correlation matrix). Similarly the CA estimates the internal consistency of items in a survey. These techniques would have been useful to unveil information hidden in the data. The analyst has noticed, however, that the data collected (presented in the following section) was too small to conduct such analyses and that not enough time was available to acquire the size samples that would allow further statistical based analyses. Statistical properties can be applied if the number of indicators (in the present case equal to 33) is significantly lower to the sample of the population (120 interviews as presented in the following section). Though no absolute rule is considered most of the recommendations (rules of thumb) found in literature require interviews-to-variables ratios no lower than 5 (Bryant & Yarnold, 1995; Nunnaly, 1978, Gorsuch, 1983) (in this dataset case is 4.84) or the rule of 200 (Gorsuch, 1983) which assures 200 cases are necessary regardless the cases-to-variables ratio (in this case the dataset is 160 observations big). However most authors recommend to significantly exceed such minimum conditions to assure reliable. Though regardless of the recommendations these tests have been conducted there is no proof that the statistical conclusions obtained are significant. This is explained in section 4.6.8.2)

4.6.8.1. Data collection

Stage 1: Definition of the population.

The total population regards the total amount of students of the Universities in the engineering faculties of the diverse disciplines. HEIs not strictly specialized in engineering have been restricted to their engineering faculties. In particular, the analyst has tried to analyze faculties which offer studies similar to those that are offered at ETSEIB (industrial engineering faculty that belongs to the UPC). At this point a relevant problem has appeared given the different approaches of the other three universities. Tor Vergata, La Sapienza and TU Delft adopt a policy of 'early specialization' and their undergraduates are already highly specialized. At ETSEIB specialization occurs at Master Level and this creates problems in defining which faculties in other Universities need to be taken into account to allow comparison. As a general rule, students in faculties of mechanical, materials, electrical, electronic, civil, biomedical, aerospace, maritime, nuclear, energetic, management engineering fields have been deemed appropriate for comparison. The dimension in number of students that constitutes the population is also difficult to determine because though it is easy to find the distribution of students in each faculty it is not so easy to find the distribution of students of a given faculty by their field of study. However, some estimates in the number of students of engineering faculties may be of help to have an idea of the population under study: ETSEIB (3170[22]); Tor Vergata (5366[23]); La Sapienza (10.286[24]) and TU Delft (19613[25]).

Stage 2: Definition of the sampling frame

The sampling frame does not contemplate all the students of a certain University but only a subset. The limitations in the population are the following:

- Only undergraduate (Bachelor) and graduate (Master) are taken into account. (Ph. D. students are not attending to lectures nor are subject to exams so the scope of the present study does not concern to them).
- Only students from year 2 onwards have been considered (first year students are not yet ready to evaluate their HEIs given the scarcity of time they have spent at university).
- Only students that were at university by the time the survey are potential answerers. This is called convenience sampling in data analysis parlance and refers to the *convenience* of generating a sample out of people who are in the right place at the right time. (This practice can obviously lead to significant biases but the analyst has not found a more rigorous methodology with the time and resources in reach).

Stage 3: Sampling Technique

The population sample has been carried out making use of the technique of **random simple sampling** [26] which has, however, been biased by two phenomena (i) the reluctance of around one third (see stage 4 for justification) of the potential selected to respond the survey and (ii) the impossibility to judge whether people responding were being sincere in their answers.

The sample has been stratified by University and by year of study of the participants [27]

[22]Obtained from ETSEIB http://bit.ly/1RoqshC Guia Estudiante Master/Doctrorado 2011/2012
[23]Obtained from http://ing.uniroma2.it/ in all kinds of engineering studies
[24]Obtained from http://www2.uniroma1.it/infostat/facolta.php in industrial engineering, civil engineering fields
[25] Obtained from http://www.tudelft.nl/en/about-tu-delft/facts-and-figures/faculty/io/page/18/

[26] **Random simple sampling** is a statistical method to obtain a random sample. Each individual is chosen randomly and entirely by chance so that each individual has the same probability of being chosen at any stage during the sampling process. In this case each person has been found at several areas of the university (cafeteria, library, laboratory, computer rooms, study rooms and corridors).

[27] Initially also the field of study had been considered as a stratification variable but given the preponderance of industrial engineering students at ETSEIB which are mostly non-specialized at undergraduate level has made it impossible.

Stage 4: Determination of sample size

The size of the sample has been defined arbitrarily (40 people for each university making a total sample of 120 interviews). As stated in the note at the precedent section statistical theory recommendations might consider the dimension of the sample too scarce, the reason for this decision though questionable has an explanation based in three factors (i) quality over quantity; (ii) time constrictions; (iii) verification that once some particular observations have been made, additional observations do not lead to significantly different results.

(i) The analyst has deemed it very necessary to stablish 'relationships' with each of the respondents. This means it has been deemed relevant to carry out a face-to-face data collection. This way he has been able to receive feedback and clarify on-site any possible doubt a respondent might have in regard to any of the questions stated in the questionnaire attached at the stage 5 of the present subsection. Probably online questionnaires would have allowed greater size samples but also would have increased misunderstandings and provoked lower degrees of reliability.

(ii) The analyst has tried to assess the amount of time taken by the respondents which has resulted in around 12 minutes each. This amount of time considers the introduction to the study, the delivery of a questionnaire and its compilation, the question-answering and a little feedback recollection. Times comprised between two samples has been fixed on average at around 3 minutes and failed attempts have accounted for the 36.51% of the attempts (252 attempts have been necessary to obtain a 160 people sample). Some basic arithmetic operations reveal that the analyst has used around 49 hours to obtain the samples (in practice three whole mornings at each university). If time was not a constriction the sample should have had to be according to Israel (1992) around 1000 people big. With a $\pm 3\%$ and a 95% confidence interval the present study would require a sample of 1,087 interviews.

	Avg Marginal time(min)	Total (min)
Useful measures	12	1920
In-between time	3	477
Unsuccesful attempts	3	276
Unsuccesful In-btw time	3	273
	Total time (min)	2946
	Total time (hour)	49,1

Table 21: Amount of time required to ensure current data collection

Considering a constant percentage of unsuccessful attempts and that no experiences curves experimented by the analysis the amount of time to collect the required amount of data would be no less than 334 hours (definitely out of reach in this project).

(iii) An accurate analysis of the samples (carried out in the following 4.6.7.3 subsection) reveals that no significant change is perceived in the results (inferior to 5%) when considering only the first 20 samples instead of the final 40 observations sample. That might be a proof that most likely greater samples will not deliver significant changes in the obtained results.

Stage 5: Questionnaire design

A questionnaire has been designed to make it easier for students to 'rate' their institutions. A copy of the questionnaire is attached below.

Questionnaire: Academic Performance in Engineering Studies

University: ____
Year of study and typology: (i.e. 1rst year master):__________________
Student's field of study: ____

Please answer the following questions in regard to your university giving each aspect a **grade from 1 to 10** (1 is very poor and 10 is very high or very good). **Answer all the questions**

Social Interaction and Motivation	
Fairness of the grading of the exams and assignments	
Encouragement to work in teams	
General engagement to the university both by staff, faculty members and students	
Possibility of collaborating with faculty staff (research, publications, projects...)	
Interaction with professors (consultancy hours, meetings and mail communication)	
Student Resources	
Accessibility to Labs (experiments, courses or research)	
Accessibility to facilities (availability of computers, working places, resting areas)	
Personal projects (Student associations, project teams (build a car or a boat with a team...)	
Existence and quality of interactive resources (e.g. Moodle, online questionnaires with feedback, recorded lectures...)	
Support to startups and entrepreneurship of the students (abundance of projects in courses and motivation to join entrepreneurial activities)	
Direct access and encouragement to read publications and specialized magazines	
Accessibility to workshops, conferences and fairs inside or outside the field of study	
Class Dynamics	
Encouragement of class discussion (when the subject allows it)	
Application of theoretical concepts to practice (e.g. applying Heat-transfer theory to evaluate the efficiency of a building or applying control theory to control an induction motor)	
Approach of the concepts in class by cases (review of past complex engineering or management problems and how where they solved)	
Degree of students participation during lessons	
Link between theoretical knowledge and practical one	
Existence of problem-solving hours where students can solve academic problems with the advice of the professor	
Competences Development	

Deep analytical approach to problems (solve a problem 'till the result' integrating knowledge of different disciplines)	
Balance of do-by-hand and Software Approach (know both what a software does and know the 'code' or principles behind)	
Soft Skills enhancement (e.g. Public speaking or communication techniques)	
Technical-professional competence acquisition in the field (acquiring 'knowledge' as well as 'methodology')	
Presence of creativity in the courses	
Evaluation Methodology	
Balance between oral and written exams (the oral component is taken into account (i.e. oral presentations, oral exams…)	
Possibility to retake exams	
Continuous evaluation (the mark is not defined exclusively defined in a 'final exam')	
Balance between Done-in-Class and appears in the exam	
Accessibility to older exams	
Organization	
Organization of the courses (a good organization does not suppose overlapping)	
Availability of the courses in both Summer and Winter semesters	
Flexibility of the syllabus that takes into account the student's preferences on courses election	
Appropriateness of the bibliography suggested in the courses	
Quality of the synthesized bibliography given/available to students (i.e. slides or similar)	

Table 22: Satisfaction questionnaire delivered to interviewed students

Note how *table 22* closely relates with *table 20* in structure and content. They both will be used in subsection *4.6.10* to obtain the results.

Note: The results of the survey carried out at the different universities is attached in the annexes.

4.6.8.2. Survey data statistical evaluation

In order to check the possibility to obtain statistically relevant information out of the data some statistical tests have been carried out:

(i) **Normality of the data:** Each variable has been exposed to a normality test to verify whether data behaved following a normal distribution. Normality is the number one condition to apply multivariate analysis. It has been found that the samples acquired for most criteria (or variables) did not accomplish the normality assumption. Thus, the application of PCA or FA multivariate analyses with not-normal data will most likely derive into flawed results. An application of the CLT (Central Limit Theorem) could allow the samples to be analyzed anyways and indeed this is what the analyst has tried to do knowing that one should be skeptical about the reliability of the results obtained. Resampling techniques such as bootstrapping[28] could be used but the best alternative

[28] The basic idea of bootstrapping is that inference about a population from sample data (sample → population) can be modeled by *resampling* the sample data and performing inference on (resample → sample).

would definitely be increasing the sample dimension which has been impossible with the amount of time given.

(ii) **Non-correlation verification:** To ensure that every criterion 'adds value' a non-correlation verification needs to be done. If two items are too correlated it might be possible that two items are referring to the same phenomenon and that the model could be less 'complicated'. This fact has been verified by two different methodologies. A Principal Components Analysis has been carried out using the Minitab Software. (Though the dimension of the samples is too small to give credit to the results obtained using such methodology.) The results obtained show that the model could not be significantly reduced if the model is intended to be accurate. Assuming that we want to explain the 90% of variability of the model, the variable (criteria) reduction using multivariate analysis will not be greater than 12 (that is the model could be reduced to 21 core indicators instead of the 33 included in the present model).

Though no statistical analyses can be properly carried out because the samples do not completely fulfill statistics' assumptions it is interesting to see the characteristics of the scores given by the survey respondents to their respective Universities. The 160 interview survey has delivered 40 judgements per university regarding the 33 criteria for evaluation. Each one of the 33 criteria has obtained a unique overall evaluation at each University obtained as the average of the 40 collected singular evaluations on that specific item. Once this has been done all the items belonging to a same category (first level of the AHP hierarchy as defined in section 4.6.5.) have been averaged together. The following figure shows the distribution into the 6 main categories (first level of the hierarchy) of the performances of each university without taking into account the weights defined in section 4.6.7. The diagram can help to understand differences in Universities performances in relationship to the different categories and might be relevant to conduct ranking forecasts prior to the weighted aggregation.

Note: The statistical data used and the results obtained can be found in the annexes.

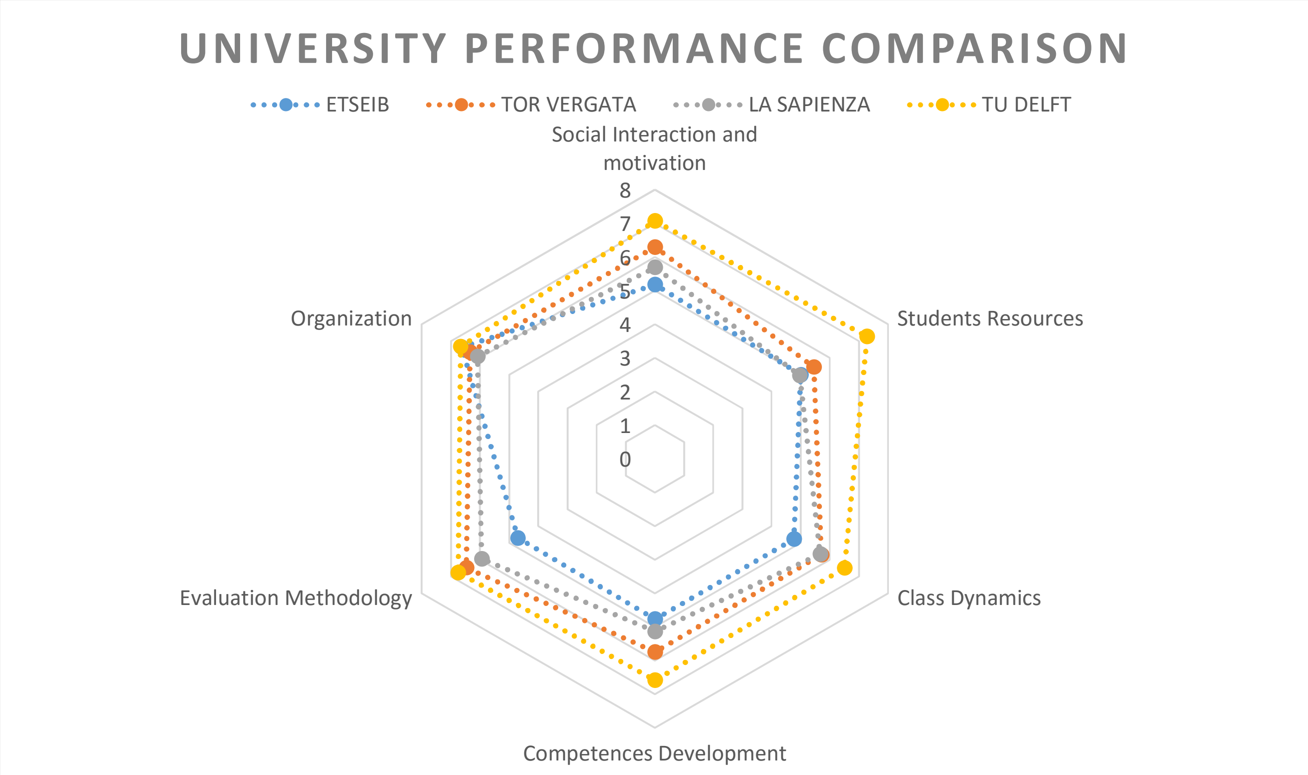


Figure 8: University Comparison (without taking weights into account)

The chart in *figure 8* shows how TU Delft performance reports are better in every category making the university become the first ranked regardless of the weights assignation (the weighted aggregation will be presented in section 4.6.9. and conducted in section 4.6.10. L'Università degli Studi di Roma Tor Vergata is ranked second in most categories so most likely will rank after TU Delft. Between L'Università degli Studi di Roma La Sapienza and ETSEIB the later has a disadvantage but for very high weights of the Organization category might end up ranking prior to the former. The *section 4.6.10* will help to clarify the a priori forecasts. Other aspects that deserve to be highlighted are the low reports ETSEIB, Tor Vergata and La Sapienza students give to the category of 'Students' Resources'. An explanation to this phenomenon can be found if we consider that Students' Resources is probably the one dimension that requires higher university's expenses. Most likely a high underlying correlation exists between University available budget and 'Students' Resources' expenses.

Another particular observation refers to the similar reported performance of all four universities under study as far as Organization is concerned. A possible explanation might be found if we take into account that organization is not as subject to variability as other categories. Organization is not as sensitive to extreme evaluation and probably does not suffer from the non-*conformist-effect*. This concept refers to students who polarize their assessments evaluating variables with very low scores when their expectations are not met. Organization is not as sensitive to nonconformist evaluation and this might be the explanation for the observed phenomenon. This observation might also imply diminishments into the results reliability because assessments are not being objective and are being conditioned by **cognitive biases** (see Kahneman 2011).

4.6.9. Survey data and criteria weights aggregation

Data presented in subsection 4.6.8.1. has been used to obtain the overall results for each university. The method of aggregation used has followed the WSM presented in *subsection 4.4.1*. The formula has been applied as follows:

$$University\ Score_i = \sum_{j=1}^{N} a_{ij}.W_j \quad for\ i = 1,2,3,4$$

Equation 23: University score definition

Where a_{ij} is the performance of university i as measured by the criterion j based on the 40-interview assessment results and W_j is the absolute weight of each j criterion.

Each a_{ij} is obtained as the average (mean) of the 40 evaluations for each university in regard to each decision criterion j. Thus, a_{ij} belongs to the following set $\{a_{ij} \middle| 1 \le a_{ij} \le 10\ and\ a_{ij} \in \mathbf{Z}\}$. $W_j's$ have been obtained and presented in section 4.6.7. Taking into account the nature of each of the inputs, each University score will be comprised between 1 and 10 where 1 would represent the worst performance and 10 the best one.

Once the four University Scores have been obtained the alternatives (Universities) can be compared in relationship with their performance (University Score) and a ranking can be set up.

4.6.10. Getting the results

The recurrent utilization of *equation 22* delivers the 4 Universities Scores and allows to define a ranking between them. The Ranking has been named the '**Engineering Teaching Ranking'**.

The results obtained are included in the following figure:

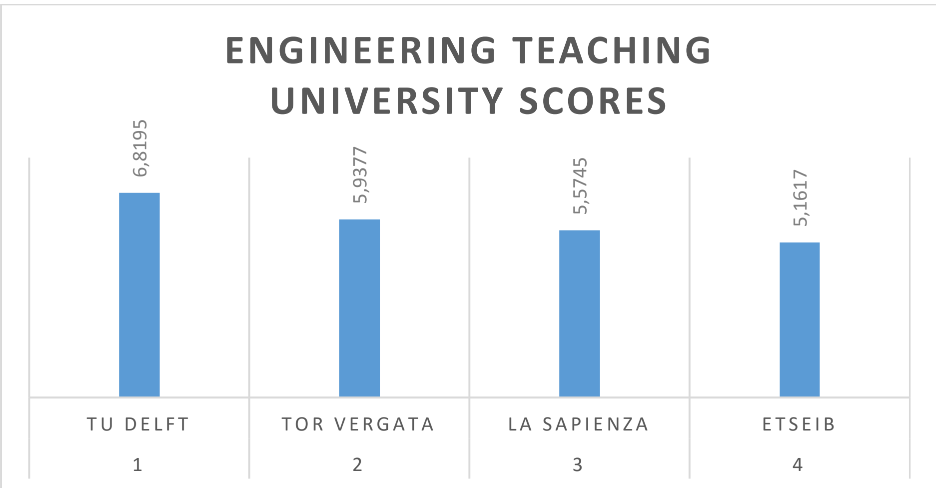
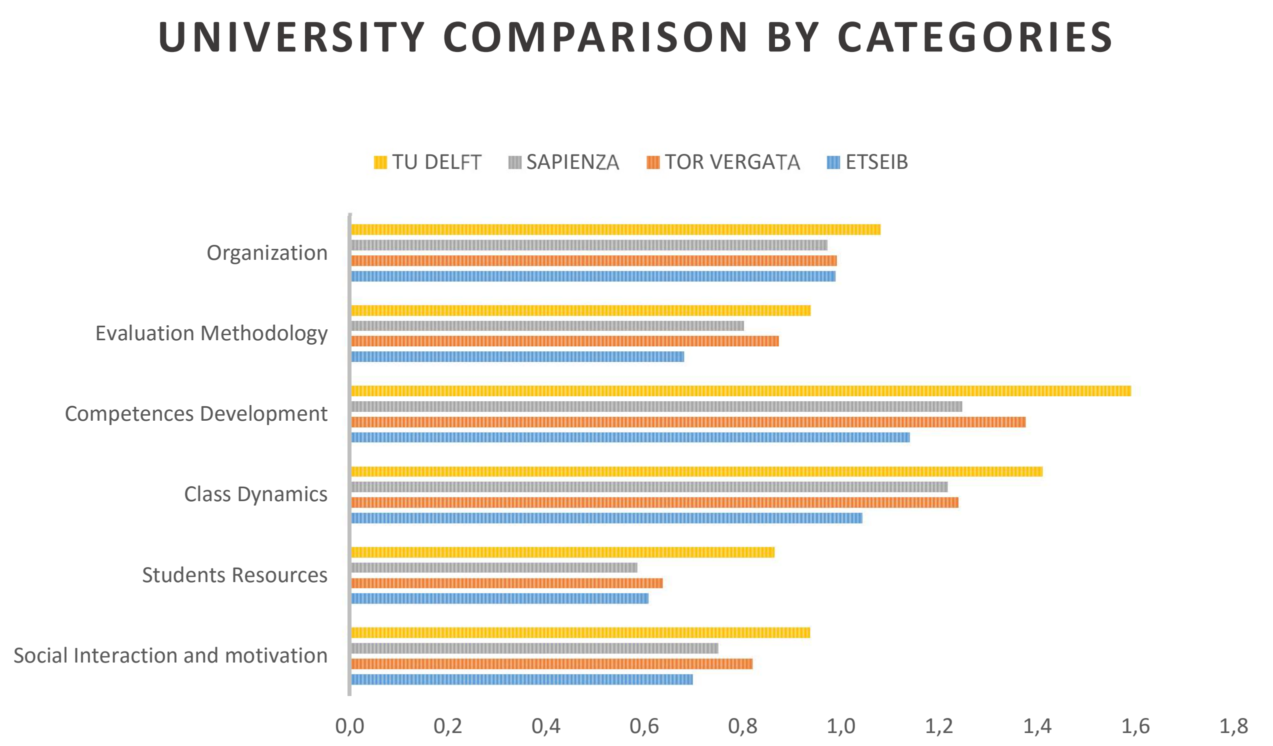


Figure 9: Ranking of Universities' scores according to the 'Engineering Teaching Ranking'

As the reader may appreciate TU Delft appears ranked in first position with a score of 6.82/10. Similarly Tor Vergata ends up second with a score of 5.94. La Sapienza appears in third position with a 5.57 score and ETSEIB fourth with a 5.16 score. An additional chart decomposes the overall scores into the 6 main categories. The following figure shows such decomposition. The categories considered have already been presented many times throughout the document



	Social Interaction and motivation	Students Resources	Class Dynamics	Competences Development	Evaluation Methodology	Organization
TU DELFT	0,9364	0,8643	1,4103	1,5900	0,9382	1,0802
SAPIENZA	0,7506	0,5851	1,2167	1,2470	0,8021	0,9729
TOR VERGATA	0,8206	0,6379	1,2389	1,3757	0,8733	0,9913
ETSEIB	0,6982	0,6084	1,0441	1,1405	0,6809	0,9895

Figure 10: Overall Universities' scores decomposed by category.

Note how the aggregation of bars of the same color (corresponding to the same university) results into the overall university scores presented in *figure 9.*

It is interesting to see the influence of weighting and aggregating in the outcomes in relationship to what has been obtained in *figure 8* prior to the weighted aggregation. TU Delft as appears as the highest scoring University in all categories as forecasted previously and no significant changes occur in other universities scores as a consequence of the weights assignation. ETSEIB comes out last in 5/6 categories and Tor Vergata maintains the second position as it could have been predicted.

5. Conclusions

5.1. Comparison to other Major Rankings

In the first place it seems interesting to compare the position each university receives in the rankings in comparison to that in which they appear in major Rankings. The following table offers a comparison between the ranking positions for the 4 Universities under study.

Note: ETSEIB has been taken into account considering it as a part of UPC. Given that there is no ranking that differentiates performance by faculty, scores and rank for the overall University (UPC) have been used to refer to the faculty of Industrial Engineering in Barcelona (ETSEIB).

.

	ARWU RANKING (2015) ENGINEERING FIELD	**LEIDEN RANKING (2015) ENGINEERING**	**QS RANKING (2015/2016) ENGINEERING &TECHNOLOGY**	**THE RANKING (2015/2016) ENGINEERING & TECHNOLOGY**	**ENGINEERING TEACHING RANKING**
TU DELFT	101-150 (1)	79 (1)	19 (1)	19 (1)	1
TOR VERGATA	Out of TOP 200 (3)	579 (4)	302 (4)	Out of TOP 100 (2)	2
LA SAPIENZA	151-200 (2)	344 (3)	149 (3)	Out of TOP 100 (2)	3
ETSEIB (UPC)	101-150 (1)	268 (2)	82 (2)	Out of TOP 100 (2)	4

Table 23: Comparison of Ranking Position of the institutions under study in major Rankings and in the Engineering Teaching Ranking. In parenthesis the relative Ranking considering only the Universities under study

Absolute positions cannot be discussed because the Engineering Teaching Ranking only observes the behavior of 4 Universities. However, it is remarkable to see how all Universities in the study rank lower in rankings which are based on objective research metrics (such as the Shanghai ARWU and the Leiden Ranking) than they do in multi-indicator rankings systems (QS and THE rankings).

Table 23 reveals some information about how the Universities under study perform according to rankings aimed with different approaches and objectives. The data in the table reveals how TU Delft apart from being a quite well-positioned research University (ARWU) performs equally well when multiple indicators are considered (THE and QS) or in teaching-focused ranking methodologies (**Engineering Teaching Ranking**). That might make the reader infer that there is a correlation between Scientific Performance and Teaching Performance and that higher ranks in research focused rankings provoke higher ranks in Teaching focused ranking. When looking at the rankings of the other Universities, however, we notice that inferring that better scientific performance leads to better teaching quality might be inappropriate. The cases under study reveal how a low-scientific performance University (Tor Vergata) receives, in contrast, a high teaching quality rank according to the Engineering Teaching Ranking. This phenomenon has two different explanations that are opposed to one another.

The first one outlines that it is possible that students in lower scientific performance institutions are less critical with the way 'things' are handled at their University and thus they report the University to teach better than it actually does. The second explanation, on the other hand, sustains that when focusing in high-quality research Universities might regard 'teaching' as a secondary objective and thus provoke their students perceive their education is not up to their expectations.

In any case, looking at the results it can be stated that both explanations bear a certain degree of truth. The case of ETSEIB is quite particular because unlike Tor Vergata it obtains relatively high ranks in Scientific Performance and multi-indicator Rankings but does not do very well as far as teaching is concerned. The reason why this is so should be analyzed in further studies which are not included in the present one.

As a conclusion of this comparative study between results obtained in Major Rankings and the Engineering Teaching Ranking some mysteries remain undisclosed. The following subsection might help understanding why some results that are indeed surprising can be explained.

5.2. Limitations to the use of purely subjective data

5.2.1. Inadequacy to compare institutions across countries/cultures

As it has been explained in *section 3.3* there are certain limitations to the Engineering Teaching Ranking which might make it inappropriate to make direct comparisons between Universities in different countries. The problem of taking into account exclusively subjective (though quantitative) parameters when evaluating a University is that the lack of objectivity might cause a considerable diminishment of the reliability that may be given to the results obtained. In *section 3.3* it has been highlighted how cultural differences might have a tremendous undesired incidence in the results and after having obtained such results it seems that such predictions have come true. Though it is possible that the results obtained are actually accurate it seems that the difference between the scores different universities receive is too broad and this might imply that there are underlying variables that might have affected to the scores.

Why is it then that the results have come out like this?

The explanation to this might be found in the performance evaluation survey presented in *section 4.6.9.*

In such survey two kind of respondents (acquiescent and critical) have been identified by the analyst. The *acquiescent respondent* seems to evaluate all the items with magnitudes that for him/her represent a 'fair' assessment for its institution but does not usually dare to give extreme scores in any criterion. Some proof of that is that when asking Tor Vergata students for feedback about the questionnaire presented in *section 4.6.8.1* many of them would argue that their University was not focusing or giving attention to many aspect included on the questionnaire but still gave relatively high scores to its University's performance according to such items. The most cited items which in their opinion were not being considered but that got relatively high scores were: Labs Access (there is only an electronics laboratory in the faculty) but still that item scored 4,4 /10 and Publications' Access that scored on average 5,475/10 even though some students (more than 10/40 interviews) did not know what publications were. These examples might help to understand the high results of Tor Vergata in the Teaching Engineering Ranking but also seem to undermine the reliability of the Ranking.

The identification of the *acquiescent respondent* unveils why the use of subjectivity can create significant biases in the results. Measuring exclusively subjective criteria has had a much bigger negative impact than the analyst could have forecasted at the beginning. As explained in section 3.3 cultural influence has played a central role in such subjective reports.

The other kind of respondent which has been named *critical respondent,* contrarily, tends to give the score the University deserves despite the repercussion this might have for the overall score of the University. The main characteristic of the *critical respondent* is that his answers are more polarized and thus he/she will deliver very good assessments when, according to its opinions, a certain criterion is highly achieved. Analogously *critical respondents* are more disposed to appraise with very low scores.

It could be inferred that the likelihood of coming across each kind of respondent is random and that in any case all universities should have the same *acquiescent-critical respondent* proportion.

However, a basic descriptive statistical analysis can provide the proof of why that is not like this. In this analysis all the scores delivered by all the students of a certain university have been put together as a variable. Each variable (population) has been represented with a sample (obtained out of 40 individuals attending to such university each one of which has responded to 33 question). We note that the resulting variance for each group/variable (University) differs. In order to know whether this is a reasonable observation or not a hypothesis test (ANOVA) would be the best methodology to prove that.

Nonetheless, the assumption of normality of observations (as explained in section 4.6.8) is not accomplished and thus such methodology cannot be used. However, the following figure and table show substantial differences that might be enough to conclude that apart from a performance difference there might be also a difference on respondents based on the university they study at.

Variable	N	N*	Mean	Mean Standard error	Std.Dev	Minimum	Q1	Median	Q3	Maximum
ETSEIB	1320	0	5,1530	0,0633	2,3009	1,0000	3,0000	5,0000	7,0000	10,0000
SAPIENZA	1320	0	5,5455	0,0522	1,8979	1,0000	5,0000	6,0000	7,0000	10,0000
TOR VERGATA	1320	0	5,9606	0,0539	1,9578	1,0000	5,0000	6,0000	7,0000	10,0000
TU DELFT	1320	0	6,5947	0,0490	1,7815	1,0000	6,0000	7,0000	8,0000	10,0000

Table 24: Descriptive statistics of the total amount of answers (40 respondents for each university x33 questions for each respondent=1320 observations) obtained in the survey

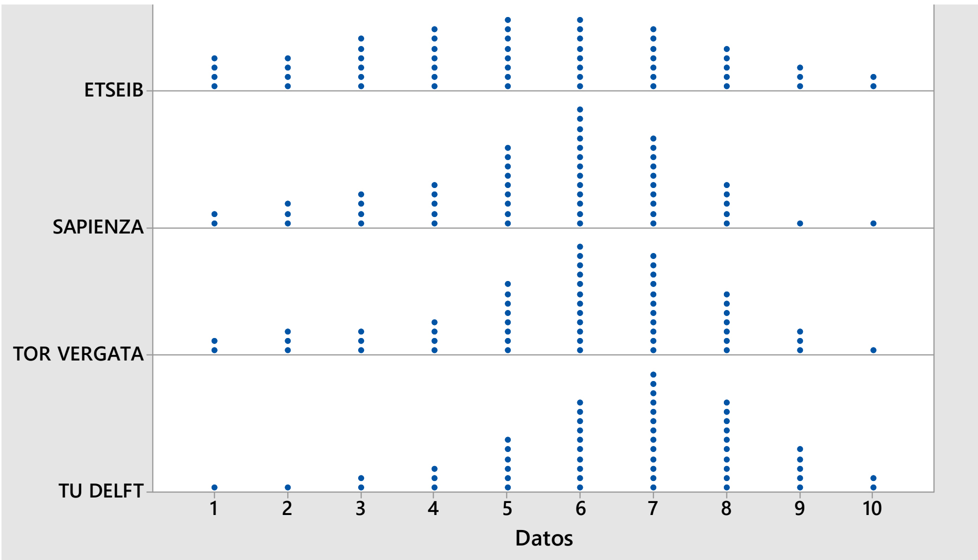


Figure 11: Distribution of scores given by students to each University. Each dot represents up to 27 observations.

From *table 24* there are two main things that deserve attention (i) variance difference between universities and (ii) quartiles analysis. (i) First of all the variance of the observations differs and it seems that ETSEIB has certainly a greater variance than the other institutions (if data was coherent confidence intervals as well as hypothesis testing would help us verify the the degree of confidence of such statement) and (ii) note how the first quartile from the bottom is set at 3 for ETSEIB at 5 for Tor Vergata and Sapienza and at 6 for TU Delft. Though this difference could be objectively true it seems to be too big considering the dimension of the samples (1320 evaluations (40 people x 33 questions) per university). Figure 11 illustrates this situation in a graphic way which reflects the conclusions obtained through *table 24.*

With this in mind it seems that it is likely that ETSEIB students' proportion is more biased towards the critical ones that in other institutions and that not only students are dissatisfied with their alma mater (which is undeniable) but also that they have higher levels of exigence when appraising it.

As it can be observed from the results the preponderance of *critical respondents* can (and it indeed does) provoke significant differences in this fact diminishes enormously the reliability that can be given to the results because such differences make the comparison almost impossible.

5.2.2. Inadequacy to build indexes based on purely subjective data

It seems that the fact that all data is subjective induces to results that do not seem to be correct. The fact that TU Delft (with a score of 6.8195) distances ETSEIB (with a score of 5.1617) by 1.6578 points on a scale of 10 seems a too big difference. One would expect quite tight results or at least that different Universities change their positions in the ranking if different weights were assigned to each one of the 6 main categories (Social Interaction and Motivation, Students' Resources, Class Dynamics, Competences Development, Evaluation Methodology and Organization) but surprisingly this has not occurred. As it may be observed in *figure 8* TU Delft results the first ranked according to its students' performance evaluation. That is, it appears ranked first in all categories regardless of the weights assigned. The analysis of the results shows that when a university is ranked in a certain overall position, it is because it ranks in such position in all categories and not because it is the best in the most important (the one which accounts for greater weights as presented at *table 19*) categories. The analysis of Tor Vergata's results also shows a similar trend. It is ranked second in all categories prior to weighting and aggregation (check *figure 8)* and thus it stays in second position once weighting and aggregation has been carried out (check *figure 10*). Some oscillations happen between La Sapienza and ETSEIB but the general trend also applies for them.

This fact has made it unnecessary and banal (unfortunately) to carry out a sensitivity analysis which would have been interesting to see how small modifications in the weights of each category (or even to each decision criterion) would affect to the overall score of a certain University and subsequently to the rankings of the Universities under study according to the Engineering Teaching Ranking methodology.

The fact that when a university is ranked in a certain overall position, it is in such position in all categories diminishes a lot the possibility of further development of the indicator and makes the analysis of the results somehow dull. In trying to understand the reasons which have led to such results the survey study has appeared to be the critical element which has conditioned the final and unexpected results as it has been pointed out previously.

5.3. Policy making and decisions

After having tried to condense the information of University Performance in a unique number (that is indeed the objective of a composite indicator) it now seems that probably a single composite indicator is not enough to evaluate academic performance and that relying into a single numeric value to make important decisions may often be too reductionist and risky. Policy decisions need to be made taking into account many and diverse factors and probably a dashboard would be more useful to that objective. If a composite indicator (as we have seen) makes use of single indicators and synthesize them into a numeric value, dashboards take into account the many single indicators separately and follow their dynamic evolution in time with the objective of identifying those indicators that cause low-performance. Probably a dashboard which would include different categories such as the 6 categories presented in the composite indicator could serve to help decision making.

Policy then should try to take into account the information delivered by the dashboard and allocate resources to boost the performance in such categories that seem to be responsible (according to data represented by the indicators that conform the dashboard) for low performance. However it is important to create metrics and consider indicators which take into account both subjective and objective parameters because as we have seen relying exclusively in subjective metrics might lead to undesired results.

5.4. Guidelines for the future

Having seen that relying in subjective assessment has had such a negative impact it is clear that any further study carried out should try to balance both objective and subjective modalities of assessment.

It seems also clear that the broader the scope of a certain study is the less accurate it becomes. If the objective of league tables are intended to classify Universities according to their performance it seems logical that each field of knowledge or each discipline has its specific demands in relationship to what is important. It is intuitive to sustain that responding the question of which is the best in the field of electronic engineering can be responded more easily than the question of which is the best overall University.

Following this logic it seems that probably the scope of the thesis has been too broad and too general. Engineering is certainly a field of knowledge but in a highly specialized globalized world it appears to be still too wide. The trend in University quality evaluation to allow comparison across Universities is increasingly switching its objective and the question is no longer which University is the best in general terms or which one is the best in Engineering in general but rather which one is the best in a certain field of knowledge such as electrical engineering to name but one. This way even raw variables and indicators may be adapted in relationship to which field of knowledge is being assessed or ranked and a higher reliability of the results can be obtained.

Also it is important to keep in mind that data sets need to be sufficiently big so that statistical methods can be used to evaluate the suitability of the data acquired and the relevance of all the indicators to be included in the composite index.

The main problem, however, will always remain how to choose adequate variables and how to measure them appropriately.

6. Bibliography

1. Bhushan, N. and Ria,K. (2004) '*Strategic Decision Making: Applying the Analytic Hierarchy Process'*, London: Springer-Verlag London Limited.

2. Billaut J.C., D. Bouyssou, P. Vincke. (2010) '*Should you believe in the Shanghai ranking?*' Scientometrics, Springer Verlag, 2010, 84 (1), pp.237-263

3. Blaug, M. (Sept. 1997) *"Ugly Currents in Modern Economics",* Policy Options.

4. Bouyssou, D.Th. Marchant, M. Pirlot, P. Perny, A. Tsoukias, and Ph. Vincke (2000) '*Evaluation and decision models: A critical perspective.*' Kluwer, Dordrecht.

5. Bryant F.B., and Yarnold P.R. (1995), '*Principal components analysis and exploratory and confirmatory factor analysis. In Grimm and Yarnold 'Reading and understanding multivariate analysis'*. American Psychological Association Books

6. Chapple, M., & Murphy, R. (1996). *'The Nominal Group Technique: extending the evaluation of students teaching and learning experiences*'. Assessment & Evaluation in Higher Education, 21(2), 147-160.

7. Chen, S.J. and C.L. Hwang, (1992) '*Fuzzy Multiple Attribute Decision Making: Methods and Applications'* Springer-Verlag New York, Inc. Secaucus, NJ, USA.

8. Diener, E., & Suh, E. M. (Eds.). (2000). '*Culture and subjective well-being'* Cambridge,MA: MIT Press

9. Dill, D. and Soo, M. (2005) '*Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems'* Higher Education, 49:495–533.

10. Dixon, Huw D. (1997), ‘*Controversy: Economics and Happiness’*, Editorial note, *Economic Journal 107 (November)*: 1814-14

11. Donovan, S., Springer, L., Stanne, M. (1999) ‘*Effects of small-group learning on undergraduates in science, mathematics, engineering and technology: a meta-analysis’. Rev Educ Res* 1999 Vol 69 (21-52)

12. Enserink M. (2007). ‘*Who ranks the university rankers?*’ Science, 317(5841):1026–1028, 2007. 24th August 2007. Retrieved 1st December 2015

13. Felder R.M., and G.S. Huvard, (1993) ‘*Make Your Technical Training Effective.’ Chem. Engr., 100*(6), 133-136.

14. Felder R.M.,(1987) "*On Creating Creative Engineers," Engr. Education, 77*, 222.

15. Felder, R.M., & Brent, R. (1994). ‘*Cooperative learning in technical courses: Procedures, pitfalls, and payoffs.’* ERIC Document Reproduction Service ED-377038). View at < www2.ncsu.edu/effective_teaching/ >.

16. Felder, R.M., and Silverman, L.K., ‘*Learning and Teaching Styles in Engineering Education’* Engineering Education, Vol. 78, No. 7, 1988, pp. 674–681. Online at http://www.ncsu.edu/felder-public/Papers/LS-1988.pdf

17. Felder, R.M., Brent, R., and Prince, M.J. (2011). ‘*Engineering instructional development: Programs, best practices, and recommendations,*’ J. Engr. Education, 100(1), 89–122.

18. Fishburn, P.C. (1967), *‘Additive Utilities with Incomplete Product Set: Applications to Priorities and Assignments’* Operations Research Society of America (ORSA) Publication, Baltimore, MD.

19. Galbács, Peter (2015). *‘The Rational Expectations Hypothesis as a Key Element of New Classical Macroeconomics. The Theory of New Classical Macroeconomics. A Positive Critique’.* Heidelberg/New York/Dordrecht/London: Springer. pp. 53–90. *ISBN 978-3-319-17578-2*.

20. Gertner, Jon (May 2010) ‘*The rise and fall of the GDP’*, The New York Times Magazine, page 60: http://www.nytimes.com/2010/05/16/magazine/16GDP-t.html?_r=0

21. Gorsuch R. L. (1983) ‘*Factor Analysis’* Hillsdale, NJ: Lawrence Erlbaum. Orig. ed. 1974.

22. Hare, J. (2013) ‘*Ranking of unis is 'bad science*', The Australian, 16th October 2013. Retrieved 30th Nov. 2015.

23. Hazelkorn (2014) “*Rankings and the Reconstruction of Knowledge in the Age of Austerity”*, In: Z. Fadeeva, L. Galkute, C. Madder and G. Scott (Eds), Assessment for Sustainable Transformation: Redefining Quality of Higher Education, Basingstoke, Palgrave MacMillan.

24. Hazelkorn, E. (2009) Impact of Global Rankings on Higher Education Research and the Production of Knowledge. *Unesco Forum on Higher Education, Research and Knowledge,*Occasional Paper No. 18. Available from http://unesdoc.unesco.org/images/0018/001816/181653e.pdf

25. Hazelkorn, E. (2013) *‘Higher Education’s Future: A new global order?’* In: R. Pritchard and J. E.

26. Hazelkorn, E. (2014) *“Rankings and the Global Reputation Race*”, In: L. Portnoi and S. Bagley (2014) *New Directions for Higher Education*. Jossey-Bass/Wiley. In Press.

27. Hazelkorn, E. (2014) “*Reflections on a Decade of Global Rankings: What we’ve learned and outstanding issues”*, *European Journal of Higher Education,*49(1)12-28.

28. Hazelkorn, E. (2014) ***"****Using U-Multirank to Enhance the European Higher Education Area (EHEA)"*, Science|Business, 5 June. http://bulletin.sciencebusiness.net/news/76582/Use-U-Multirank-to-enhance-the-European-Higher-Education-Area

29. Hazelkorn,E. (2011) '*Rankings and the Reshaping of Higher Education. The Battle for World-Class Excellence'* Basingstoke, Palgrave MacMillan http://www.palgrave.com/products/title.aspx?pid=391266

30. Hazelkorn,E. (2014) *"'Devil is in the detail of global university rankings?"* The Irish Times, 7 October.http://www.irishtimes.com/news/education/devil-is-in-the-detail-of-global-university-rankings-1.1949588

31. Hazelkorn,E. (2015) "*Rankings and Quality Assurance: Do Rankings Measure Quality?"*, Policy Brief #4, CHEA International Quality Group, Washington DC.

32. Israel, Glen D. (1992) *"Determining Sample Size."* Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD-6.

33. Kahneman, D (2011). '*Thinking, Fast and Slow'*. New York: Farrar, Straus and Giroux.

34. Keeney, R.L., (1992) *'Value-focused thinking. A path to creative decision making',* Harvard University Press, Cambridge.

35. Linstone,H., Turoff M., *(1975), 'The Delphi Method: Techniques and Applications' Reading, Mass.: Addison-Wesley, ISBN 978-0-201-04294-8.*

36. Liu, N.C. and Cheng, Y.(2005) *'The Academic Ranking of World Universities'*, Higher Education in Europe, 30: 2, 127 — 136.

37. Lomax, P., & McLeman, P. (1984). *'The uses and abuses of nominal group technique in polytechnic course evaluation'.* Studies in Higher Education, 9(2), 183-190. http://dx.doi.org/10.1080/03075078412331378834

38. Marshall, Alfred (1890), *'Principles of Economics'*, London: Macmillan.

39. Marzsal, A. (2012) *"University rankings: which world university rankings should we trust?" The Telegraph. 4th October 2012. Retrieved 14th Nov. 2015*

40. Miller, D.W., and M.K. Starr (1969) '*Executive Decisions and Operations Research*' Prentice-Hall, Inc., Englewood Cliffs, NJ.

41. Nunnaly J. (1978) '*Psychometric theory*' New York: McGraw-Hill.

42. OECD (2008), '*Handbook on constructing composite indicators'.*

43. Oswald, Andrew (1997), *'Happiness and Economic Performance'*, Economic Journal 107 (November): 1815-31

44. Redden, E. (2013) '*Scrutiny of QS Rankings',* Inside Higher Ed*, May 31, 2013. Retrieved 30th Nov. 2015:* http://www.theaustralian.com.au/higher-education/university-rankings/scrutiny-of-qs-rankings/story-fna15id1-1226653550158

45. Saaty T.L. (2000). '*Fundamentals of the Analytic Hierarchy Process'*. RWS Publications, 4922 Ellsworth Avenue, Pittsburgh, PA 15413, 2000.

46. Saaty T.L. (2006) '*Rank from comparisons and from ratings in the analytic hierarchy/network processes'* European Journal of Operational Research, 168(2):557-570.

47. Saaty, T.L. (1994) '*How to make a decision: the analytic hierarchy process'* Interfaces, Vol. 24, No. 6, pp.19–43

48. Saaty, T.L. and Vargas, L.G. (2000) '*Models, Methods, Concepts and Applications of the Analytic Hierarchy Process*' Boston: Kluwer Academic Publishers.

49. Saaty, Thomas L. (1999). *'Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World*.' Pittsburgh, Pennsylvania: RWS Publications*. ISBN 0-9620317-8-X.*

50. Saisana, M. and D'Hombres B. *(2008) 'Higher education rankings: Robustness issues and critical assessment. How much confidence can we have in higher education rankings?'* Technical Report EUR 23487 EN 2008, IPSC, CRELL, Joint Research Centre, European Commission.

51. Sen, Amartya (1985) *'Commodities and Capabilities'*, Amsterdam: North-Holland.

52. Simon, Herbert (1957). "*A Behavioral Model of Rational Choice*", in Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. New York: Wiley.

53. Stiglitz, Joseph. (2001) *"Information and the change in the paradigm of economics",* Nobel Prize Lecture.

54. Thompson, Nick (25 October 1996). *'Down with Rankings' Summit: Stanford's Newsmagazine of Progressive Politics.*

55. Times Higher Education *'THE World University Rankings (2015-2016)'. Retrieved 30 October 2015.*

56. Triantaphyllou, E., Shu B., Nieto Sanchez, S., and Ray, T. (1999) '*Multi-Criteria Decision Making: An Operations Research Approach'* J. Wiley, New York.

57. Usher, A., and Savino, M. (2006). '*A World of Difference: A Global Survey of University League Tables'* Toronto, ON: Educational Policy Institute

58. Van Raan, A.F.J. (2005) '*Fatal attraction: Ranking of universities by bibliometric methods*.' Scientometrics, 62:133–145.

59. Van Raan, A.F.J. (2006) '*Challenges in the ranking of universities*.' In J. Sadlakand N.C.Liu, editors, World-Class University and Ranking: Aiming Beyond Status, pages 81–123, Bucharest, 2006. UNESCO-CEPES. ISBN 92-9069-184-0.

60. Wildavsky, Ben *(2010). 'The Great Brain Race: How Global Universities are Reshaping the World' Princeton University Press.*

61. Zimmermann, H.-J., (1996) '*Fuzzy Set Theory and Its Applications'* Kluwer Academic Publishers, Second Edition, USA 1996.

62. Zitt, M., Ramanana-Rahary, S., and Bassecoulard, E. (2005) *'Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation*.' Scientometrics, 63(2):373–401.