POLITECNICO DI MILANO

Facoltà Di Ingegneria Dei Sistemi

Corso Di Laurea In
Ingegneria Gestionale

SIMILARITIES ANALISYS FOR CLUSTERING GREAT ENGINEERING PROJECTS WITHIN THE OIL AND GAS SECTOR

RELATORE: PROFE. FRANCO CARON

TESI DI LAUREA DI:
PEDRO HENRIQUE ENDO NICOLINI
MATRICOLA 795477

2013 - 2014

Astratto

Quando si parla di soldi e l'opportunità di guadagnare in più, l'oggetto è visto con grande attenzione. Molti investimenti che generanno ritorni finanziari e anche vantaggi competitivi a un'azienda sono basati in progetti. Indipendenti della dimensione del progetto, quello che è più importante è avere successo.

Il successo di un progetto è un incentivo per fare in più. Ma come si può garantire che i progetti successivi avranno successo, solo perché l'azienda ha avuto successo col primo? Innanzitutto è necessario imparare con i progetti passati, avendo successo prima oppure no. Ma chi ha detto che imparare è un compito facile? Un processo di ottimizzazione di apprendimento appare come importante strumento per ottenere successo, tanto nell'imparare con gli altri progetti, come nell'esecuzione del progetto per se.

Comunque, per stabilire un processo di apprendimento, è necessario innanzitutto sapere che progetti siano paragonabili. Non fa senso imparare con i progetti che non hanno nulla di simile. Qui, emerge la parola "chiave": simile. La similarità tra progetti è il punto di convergenza per questo studio. Partendo delle similarità tra progetti che aggruppamenti diversi, ma con caratteristiche uguali all`interno, sorgono e permettono lo studio approfondito delle caratteristiche macro di progetti dentro un dato campione.

Così, si propone una metodologia capace di studiare diverse tipologie di progetti, guidati da una serie di variabili, e aggruppare quelli progetti che sembrano simili secondo le sue variabili. Quello è il grande interesse primario di questo studio. L'interesse secondario è quello di permettere che un progetto ottenga successo per una corretta pianificazione, prima dell'esecuzione.

Alla fine, è importante rilevare che i progetti che saranno utilizzati per lo studio sono quelli del settore di Oil & Gas. Quello è una sfida in più, una volta che sono progetti che coinvolgono grandi investimenti, molte incertezze e considerevoli ritorni finanziari.

Parole-chiavi: successi di progetti, apprendimento, clustering, Oil & Gas.

Abstract

When we talk about money, and opportunities to raise more, a great attention is given. Huge investments that generates financial return (or capital gain) or even competitive advantage for a company are always based on internal or external projects. It doesn't matter the size, duration, cost, or requirements – the key issue is to achieve success.

The project success serves as in incentive or the company to do more. But, how do we guarantee success for subsequent projects? First, it is necessary to obtain knowledge from past projects, being those successful or not. Learning is not an easy task, it is also necessary to improve and optimize this process and have consequently significant gains from similar projects.

Measuring similarities between projects is not a trivial argument, though. It is necessary first to establish a rigorous criteria and common points that should group the projects. The clustering methodology seems to give a proper way to do so, and this is the focal point of this paper. A robust clustering process gives us a better understanding about a range of projects, and a better understanding leverages the future projects assessment.

Given that, we can say that a cluster methodology is for great interest from those who are concerned about project success. By success we can read return and opportunities.

Finally, we will approach this methodology from the view of Oil & Gas industry. Normally they tend to be huge projects in terms of size, cost, return and opportunities. Moreover, they tend to carry a great portion of uncertainty. So it will be our challenge the attempt of give this industry a different approach for their projects.

Key words: project success, knowledge, project, clustering, Oil & Gas.

SOMMARIO

Il progetto è uno sforzo temporaneo intrapreso in modo da creare un prodotto, servizio o risultato unico. Questa definizione mette in foco due caratteristiche distinte. La prima riguarda il carattere temporaneo dei progetti, o sia, succedono per un determinato spazio di tempo, e non assume caratteristica di continuità. La seconda cosa da distinguere sui progetti riguarda l'unicità del risultato.

Fare una considerazione di queste due caratteristiche, mettendo insieme i costi dei progetti, è possibile dire che un progetto è irreversibile in quanto l'utilizzo di risorsa, e qualche sbaglio può costare un grande ammontare di soldi, tempo, mano d'opera, e altre risorse. Tutto posto insieme costituisce un importante aspetto del cosiddetto "successo del progetto".

Dato questa considerazione iniziale, sarebbe ragionevole sistemare uno studio capace di mitigare i rischi dei progetti di non avare successo. Però come si deve confrontare questo problema? Siccome ci sono tante variabili che compongono il successo dei progetti, la scelta è stata fatta su una di queste dimensioni. Per arrivare a questo punto, innanzitutto si deve capire bene quelle che sono le fasi dei progetti e come sono connesse con l'idea di successo di progetto.

Le fasi dei Progetti e il Successo del Progetto

Il progetto segue una logica di quattro fasi distinte: l'iniziazione, la pianificazione, l'esecuzione e la chiusura. Ogni fase ha il suo costo di cambiamento e la sua influenza sul risultato. Nella fine, cambiare qualcosa diventa più costoso in modo che il progetto avanza. La Fig. 1 traduce quella relazione.

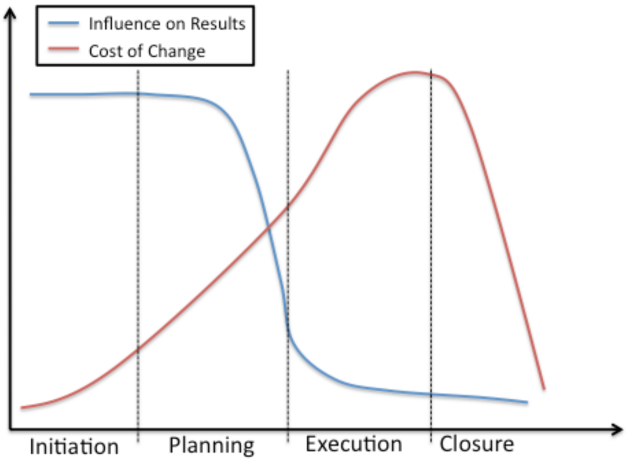


Fig. 1 – Il costo di cambiamento e l'influenza sul risultato.

Si vedi che anche la probabilità di successo del progetto diventa minore in caso di uno errore nelle fasi più avanzate del progetto. È quel lì il ragionamento più forte dello studio, cambiare qualcosa nell'inizio del progetto porta tre benefici: costo di cambiamento più ridotto, maggiore impatto sul risultato del progetto, e l'aumento della probabilità di successo del progetto.

La motivazione dello studio sarà pertanto i miglioramenti nel processo di pianificazione del progetto, più precisamente l'impatto dell'apprendimento nel processo de elaborazione della *baseline* di costi basata in progetti passati.

Sorge qui una questione, che è l'obietto di studio: come si può definire quali progetti passati che hanno comportamento simile dal progetto in pianificazione. Il punto chiavi qui sarà distinguere come si può definire cos'è similarità di progetti, e come aggruppare progetti che sono simile in modo da facilitare il processo di apprendimento con i progetti passati, e conseguentemente migliorare il processo di pianificazione. Alla fine, l'impatto atteso è quello di aumentare la probabilità di successo del progetto.

L'analisi di Similarità

Allora, la similarità tra progetti è il punto principale dello studio. Come fu mostrato prima, l'apprendimento dipende della scelta giusta dei progetti che possono essere comparabili. Così, si può pensare di una forma quantitativa per misurare il grado di somiglianza tra due osservazioni.

Sorgono così i metodi di calcolo delle distanze tra due punti, che possono essere due progetti, ad esempio. Le distanze che sono grandi rappresentano progetti non simili, e distanze piccole l'opposto. In altre parole, la somiglianza tra due progetti è rappresentata da un numero che misura la prossimità di questi progetti.

Le principali misure di prossimità sono: Distanza Euclidea, Distanza Statistica e Distanza di Minkowski.

Distanza Euclidea

Questa è la più conosciuta misura di distanza tra due osservazioni. Semplicemente data come:

$$\boldsymbol{d_E}\,(\mathbf{x},\mathbf{y}) = \sqrt{(\boldsymbol{x_1}-\boldsymbol{y_1})^2 + (\boldsymbol{x_2}-\boldsymbol{y_2})^2 + \cdots + (\boldsymbol{x_p}-\boldsymbol{y_p})^2} = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

Distanza Statistica

La distanza statistica è una derivazione della Distanza Euclidea, che mette nei calcoli la correlazione tra le diverse variabili del modello. Il punto d'attenzione qui dev'essere la matrice di covarianza S.

$$d_S\ (\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})}$$

Distanza di Minkowski

La Distanza di Minkowski è la generalizzazione della Distanza Euclidea. Il suo uso, però è nel caso in cui si vuole minimizzare una differenza troppo elevata sul qualche elemento dello standard.

$$d_M\ (\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{\frac{1}{m}}$$

Relazione di Similarità

Alla fine, c'è la relazione tra le distanze tradotta da un indice di similarità, che può essere calcolato come:

$$s_{ik} = \frac{1}{1 + d_{ik}}$$

In realtà quell'indice non serve nulla per il modello di Clustering, ma fornisce una buona visualizzazione di come i dati si comportano. Il valore della similarità tra due osservazioni è compreso tra zero e uno, dove zero significa progetti non simili, e uno progetti con similarità massima.

Il Modello di Clustering

Una volta calcolata la distanza tra i diversi progetti, è necessario aggruppare quelli che sono simili tra di loro. Un compito che sembra semplice, ma che domanda l'utilizzo di robusti algoritmi che sono chiamati Algoritmi di Clustering.

Comunque, innanzitutto si deve avere in testa la definizione e le caratteristiche del Clustering.

> *"Il Clustering è una tecnica di analisi multivariata attraverso la quale è possibile raggruppare le unità statistiche, in modo da minimizzare la 'lontananza logica' interna a ciascun gruppo e di massimizzare quella tra i gruppi."*[1]

Questa definizione ci permette ricavare alcune caratteristiche della teoria di Clustering:

1. Il Clustering è un tipo di aggruppamento di oggetti che hanno simili caratteristiche;
2. Gli oggetti dei Cluster seguono determinati standard (spesso sono vettori di misure);
3. L'obiettivo del Clustering è minimizzare la distanza tra oggetti.

[1] Fonte: http://host.uniroma3.it/facolta/economia/db/materiali/insegnamenti/185_903.pdf

Allora, in tutto questo posto, è possibile utilizzare questa metodologia per arrivare all’obiettivo proposto di aggruppare i differenti progetti, ma che sono simili tra di loro, e poi ricavare informazioni importanti per il processo di pianificazione del progetto nuovo.

Il metodo che se propone in questo lavoro sono due: il single linkage e il k-means. La Fig. 2 dimostra il risultato del single linkage, il quale sarà input per il k-means.

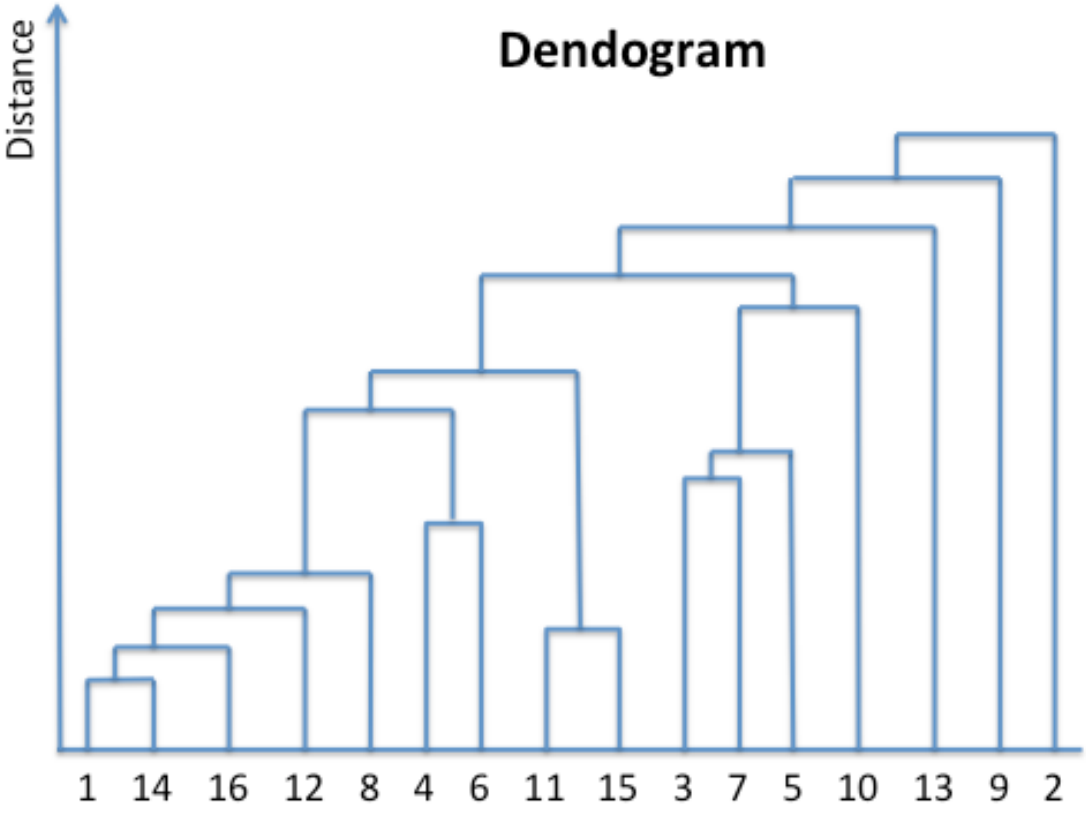


Fig. 2 – Dendograma risultante dell’applicazione del single linkage.

Poi di applicare il single linkage, è stato proposto il k-means, che segue le seguenti fasi:

1. Assegnare ogni osservazione a uno cluster;
2. Computare le medie di ogni cluster B(L, j) e anche l’errore iniziale;
3. Per il primo caso, computare ad ogni cluster il valore

$$ie = \frac{N(L)D(1,L)^2}{N(L)+1} - \frac{N[L(1)]D[1,L(1)]^2}{N[L(1)]-1}$$

   Che significa l’incremento d’errore dato il trasferimento del primo caso del cluster L(1) per il cluster L.

4. Aggiustare le medie dei cluster e l’errore;
5. Riperete le fasi 3 e 4 per ogni osservazione;
6. Fermare il metodo se nessuna movimentazione è vista.

Così finisce tutta la modellazione dello studio. Quello che segue è l'applicazione del modello e i risultati otenuti.

L'applicazione del Modello

Il modello ha seguito una strutturata metodologia di base, come si può vedere nella Fig. 3.

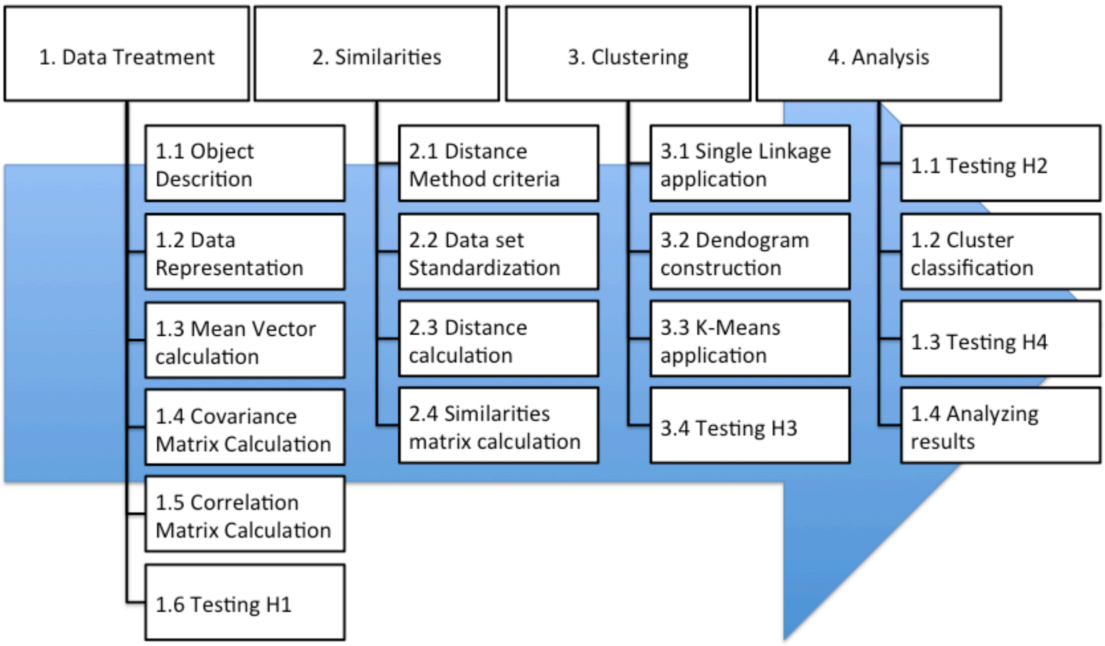

1. Data Treatment
2. Similarities
3. Clustering
4. Analysis
1.1 Object Descrition
1.2 Data Representation
1.3 Mean Vector calculation
1.4 Covariance Matrix Calculation
1.5 Correlation Matrix Calculation
1.6 Testing H1
2.1 Distance Method criteria
2.2 Data set Standardization
2.3 Distance calculation
2.4 Similarities matrix calculation
3.1 Single Linkage application
3.2 Dendogram construction
3.3 K-Means application
3.4 Testing H3
1.1 Testing H2
1.2 Cluster classification
1.3 Testing H4
1.4 Analyzing results

Fig. 3 – La metodologia generale del studio sviluppato.

E i risultati sono stati come atteso. Prima, se deve mostrare la rappresentazione di ogni cluster che è stato formato, secondo il grafico di dispersione della Fig. 4.

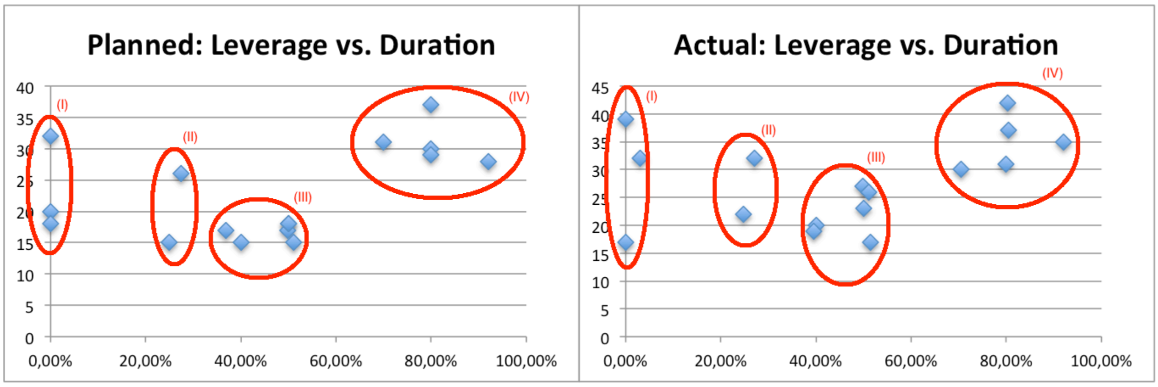


Fig. 4 – L'indicazione del grafico di dispersione che esistono aggruppamenti naturali dei progetti.

Alla fine, sono stati ricavati più precisamente tutti i aggruppamenti (oppure i clusters) che appariscono nel grafico di dispersione. Sono state ricavate quattro tipologie diverse di clusters:

I. Basso grado di leverage e alta duration.
II. Basso grado di leverage e bassa duration.
III. Alto grado di leverage e bassa duration.

IV. Alto grado di leverage e alta duration.

Alla fine, per validare lo studio, è stato mostrato secondo la rappresentazione di stelle come i progetti sono in, infatti, simili.

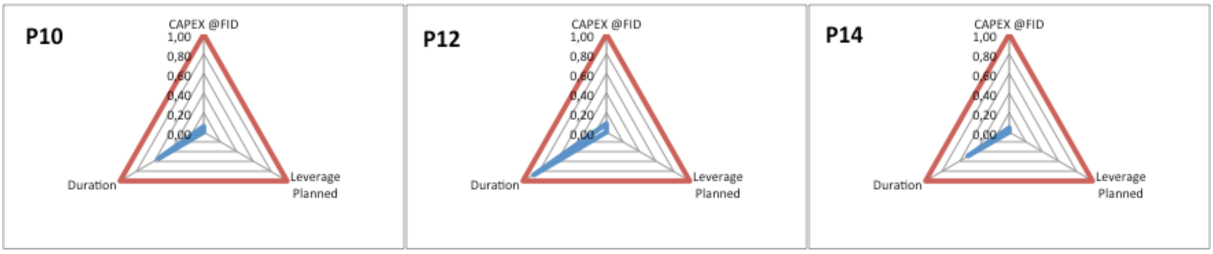


Fig. 5 – Il primo cluster, corrispondente a un basso grado di leverage e alta duration.

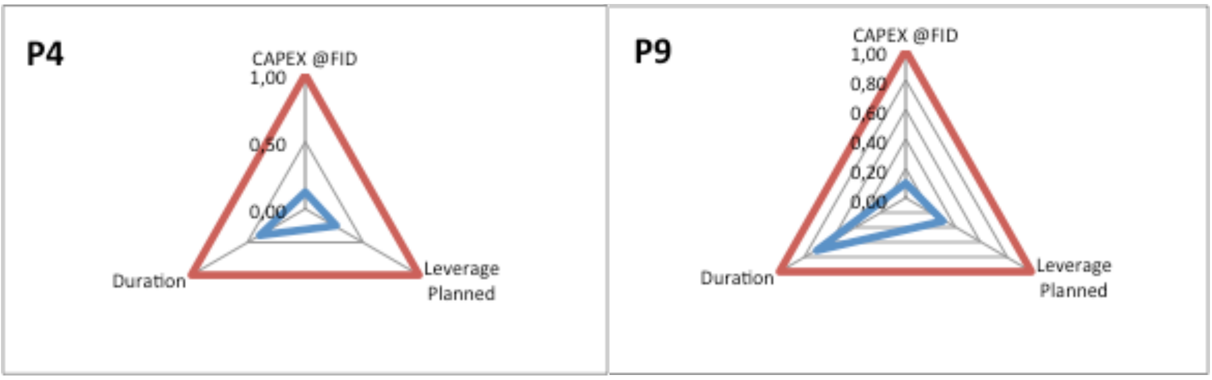


Fig. 6 – Il secondo cluster, corrispondente a un basso grado di leverage e bassa duration.

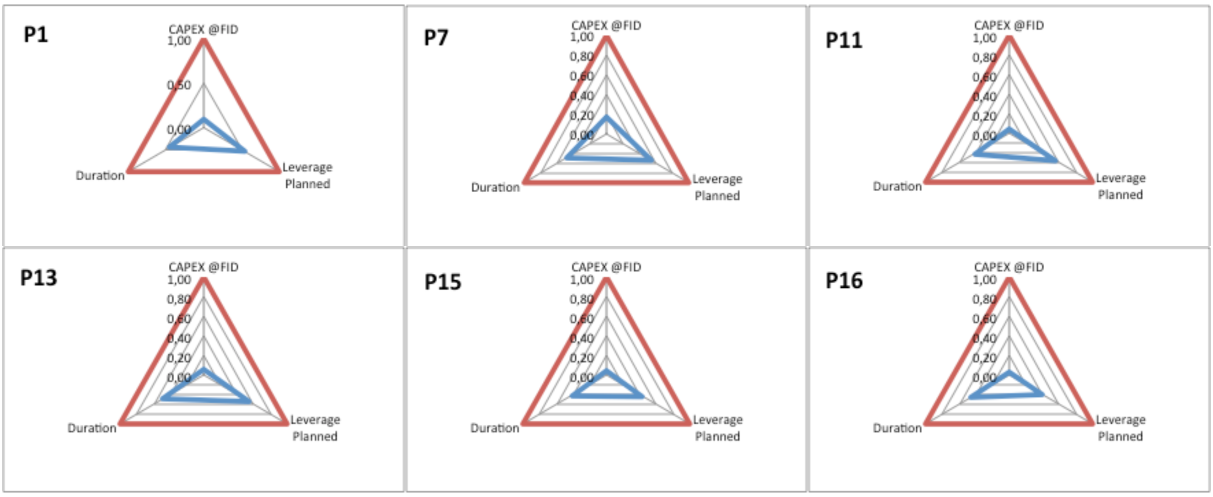


Fig. 7 – Il primo cluster, corrispondente a un alto grado di leverage e bassa duration.

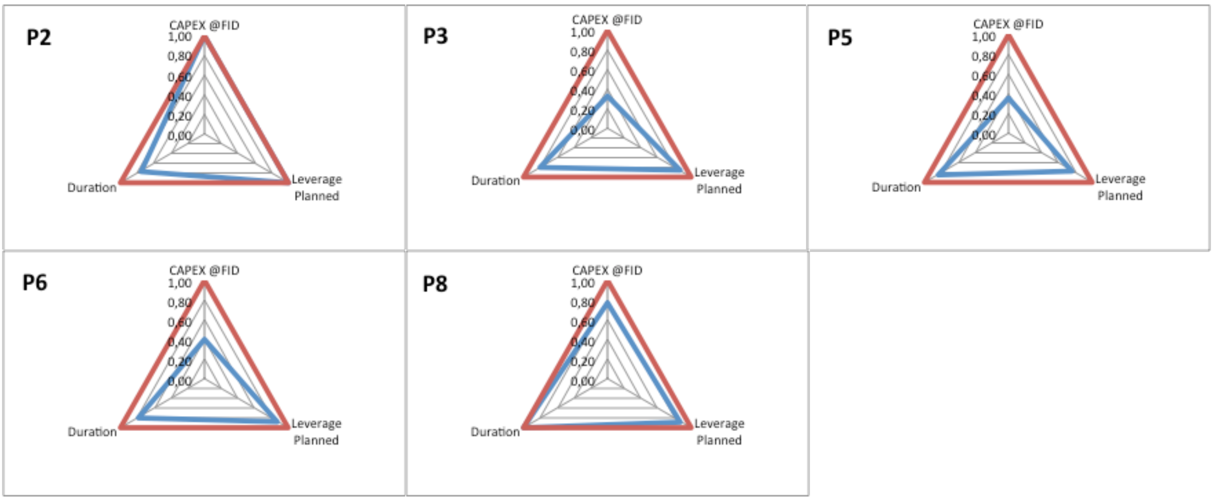


Fig. 8 – Il primo cluster, corrispondente a un alto grado di leverage e alta duration.

Così si può confermare che il modello funziona bene per aggruppare progetti che sono simili. Per avere una formalizzazione matematica di questi risultati, tutte le ipotesi intermedie sono state verificate.

Conclusioni

Le conclusioni fatte dallo studio sono state tre, e riguardano le seguenti parole chiavi:

I. Flessibilità

II. Conformità

III. Continuità

La prima conclusione ci dice che il modello sviluppato è abbastanza flessibile, essendo passibile d'applicazione in diversi settori economici. In altre parole, è possibile replicare il modello con diversi tipi di progetti, non solo quelli del settore di Oil & Gas.

La seconda conclusione ci dice che il modello è stato sviluppato per uno proposito specifico: essere un aiuto per i modelli più complessi. In questo senso si può dire che il modello non è gestionale, ma si uno strumento di supporto a quelli modelli gestionali (come il EVMS – Earned Value Management System).

L'ultima conclusione ci dice che per validare il modello è necessario fare un'ulteriore studio che possa essere capace di verificare se ha l'incremento di successo nei progetti che utilizzano il modello di clustering per aiutare nella pianificazione dei nuovi progetti. Il prossimo obiettivo sarebbe misurare l'impatto del cluster nell'apprendimento e gestione.

Table of Contents

List of Figures 19

List of Tables 21

1 Introduction 23
1.1 Motivation 24
1.2 The work Structure 24
1.2.1 Literature Review 25
1.2.2 Model development, application and conclusion 26

2 Literature Review about Project Management 27
2.1 Understanding a Project 27
2.1.1 How to define a Project 28
2.1.2 Groups of Variables to consider 28
2.1.3 The phases of a Project 30
2.1.4 Defining Success for a Project 32
2.1.5 Ten Drivers for Success 36
2.1.6 Five Immutable Principles 37
2.1.7 Five Immutable Practices 40
2.1.8 Summary 42
2.2 From the Object to the Motivation 42
2.2.1 The view of a single project 44
2.2.2 Project Portfolio Management 51
2.2.3 Summary 55
2.3 Chapter Summary 56

3 Literature Review: Multivariate Statistics and Clustering Theory 58
3.1 Statistical Tools and Data Representation 58
3.1.1 Basic Descriptive Statistics 58
3.1.2 Graphical Techniques 60
3.1.3 Distance Methods 63
3.1.4 Final Comments 64
3.2 Clustering Theory 65
3.2.1 The Cluster and the Clustering Analysis 65
3.2.2 Data Requirements 67
3.2.3 Clustering Algorithms 69
3.2.1 Algorithm Robustness 72
3.2.1 Discriminant Analysis 73
3.3 Chapter Summary 74

4 Objective, Hypothesis and Model Development 75
4.1 The Objective 75
4.2 The Hypothesis 76
4.2.1 Hypothesis One 76
4.2.2 Hypothesis Two 78
4.2.3 Hypothesis Three 78
4.2.4 Hypothesis Four 79
4.3 Model Development 80
4.3.1 Data Treatment 81
4.3.2 Similarities 82
4.3.3 Clustering 83
4.3.4 Analysis 84

5 Application and Results 86
5.1 Data Gathering and Context 86
5.1.1 Data Gathering 86
5.1.2 Context 86
5.2 Running the model 90
5.2.1 Phase One: Data Treatment 91
5.2.2 Phase Two: Similarities 95
5.2.3 Phase Three: Clustering 96
5.2.4 Phase Four: Analysis 99
5.3 Analysis and Discussion 101
6 Conclusion 103
Appendix A 105
Appendix B 109
Appendix C 110
Appendix D 112
Appendix E 114
Appendix F 115
Appendix G 117
7 Bibliography 119

List of Figures

Figure 1.1 – The work structure. 25
Figure 2.1 – Flowchart to guide the literature review of projects. 27
Figure 2.2 – Fundamental variables of projects. 29
Figure 2.3 – Project Variables expressed by diferente paradigms. 30
Figure 2.4 – Influence on results and cost of change during the life cycle of a Project. 31
Figure 2.5 – Hierarchy of Critical Factors for Success. 33
Figure 2.6 – Different measures for success in three different perspectives. 34
Figure 2.7 – Methodology for Project Success. 35
Figure 2.8 – The 10 Drivers for Project success. 36
Figure 2.9 – Five gaps at the Project life cycle. 39
Figure 2.10 – The Five Immutable Practices for Project Success. 41
Figure 2.11 – Macro-structure for the study motivation. 43
Figure 2.12 – EVM progress timeline 45
Figure 2.13 – EVMS curves. 47
Figure 2.14 – CPI and SPI along a Project execution. 49
Figure 2.15 – CPI and SPI in a stable form. 49
Figure 2.16 – Single project vs. multiple projects. 52
Figure 2.17 – Impacts of PPM at projects. 52
Figure 2.18 – Reasons to practice the PPM. 53
Figure 2.19 – First three steps of PPM life spam. 54
Figure 2.20 – Single and Multiple Projects views. 56
Figure 3.1 – Scatter Plot for two and three dimensions. 61
Figure 3.2 – Example of use for Star Representation. 62
Figure 3.3 – Nine standards for Chernoff Faces. 63
Figure 3.4 – Data Representation in scatter plot for pre-analyze of data. 67
Figure 3.5 – Example of a Dendogram Representation 70
Figure 3.6 – Graphical interpretation of Single and Complete Linkage. 71
Figure 3.7 – Graphical interpretation of membership analysis. 73
Figure 4.1 – Model structure. 81
Figure 5.1 – Scatter plot for all pair of variables strictly quantitative. 92
Figure 5.2 – Scatter plot for confront between typologies of projects. 93
Figure 5.3 – Dendogram for planned values. 97
Figure 5.4 – Clusters found in scatter plots. 98
Figure 7.1 – Linear Transformation for the data set. 111
Figure 7.2 – Star representation for all projects, by their planned values 114

List of Tables

Table 2.1 – Terminology used at EVMS. 47
Table 2.2 – Operations and Projects Management issues. 52
Table 3.1 – Representation of the data. 59
Table 5.1 – Data set of all projects in planning phase. 88
Table 5.2 – Data set of all projects after execution. 89
Table 5.3 – T-values for all correlations between variables. 94
Table 5.4 – MANOVA calculated for the clusters C1, C2, C3, and C4. 99
Table 5.5 – Distances from projects to clusters calculated for discriminant analysis. 101
Table 7.1 – MANOVA table for comparing population mean vectors. 112
Table 7.2 – Similarities computed for projects' planned values. 115
Table 7.3 – Similarities computed for projects' actual values. 116
Table 7.4 – Evolution of clusters in k-means algorithm for planed values. 117
Table 7.5 – Evolution of clusters in k-means algorithm for actual values. 118

1 Introduction

Great engineering projects do not take place without a good thinking about the subject. They normally involve a large amount of resources and have no tolerance for mistakes. This concept is not from our time, it is well known by many civilizations in many periods of time.

Let's recall the PDCA Cycle for quality management: plan, do, check, act (Shewhart, 1931, 1939). Huge projects take those steps into consideration, but first we need to weight them. The plan phase takes great importance here, because it is the time that you have the gap for doing mistakes – everything is on the paper, so the cost of error is minor. The do phase as well takes great relevance because what you've planned should be well executed. Here the cost of error is considerable. The check phase is the worse. If you have trouble planning and doing, you will check that there is nothing much to do now – your project is a failure. Finally, the act phase will determine whether you are satisfied with the outcomes or not. There is nothing to do but accept you final product that derives from your project.

Imagine then that you are an architect and have designed an amazing building. You take hours drawing your sketches, which evolve to a CAD[2] drawing, which evolve to a plan of the building. As an architect you are not hundred percent concerned about the feasibility of your project. You don't know, and don't want to know if it could be done, you just want it done.

But as an engineer, you have to think about all details, from the structure to the finishing, all project is on your hand. The drawings from the architect most of the times cannot be built. Obviously there is a conflict of interest. Both parties want the project to be done, but sometimes it cannot be done as both parties want. That struggle can be fixed at planning phase. This phase is the round for negotiations and the time to change, and change, and change again the project structure.

Simplifying, the project can be understood as something planed, executed, checked, and certainly made by some specific purpose. Besides that, is an endeavor taken by consideration from many different stakeholders. Finally, the project uses many quantities of resources, so it is something that every actor should be concerned.

Doing wrong is not just a matter of misuse of resources, but a waste in time and credibility of the maker. The project, though, is something really complex, and really volatile to errors. The question that arises from this brief discussion is: how can we turn a project into

[2] CAD: Computer Aided Design, a class of software that helps developing technical drawings of products and huge projects.

something successful, or in other words, something with minimal errors, from the point of view of every stakeholder?

1.1 Motivation

This study`s motivation emerges exclusively from one word: error. Projects, as we said are endeavors with no tolerance to mistakes, after all, they are customized outcomes – therefore irreversible process – that take a lot of resources to be done.

A great way to eradicate the errors within a project is by learning all that is possible about them. A good start is to look back into past projects and learn from them. If they were not successful, we should study the reasons why it not as we expected. If they were successful, we could replicate the actions.

By those lines that this work should walk. Knowledge and learning are the very strong words among the next pages. But it will not be easy, though. Trying to explain future projects with past ones is a formula that had already been tried. There are some details that we will modify in this point of view in order to change the usage of knowledge and learning at the planning phase of projects.

The great discussion – and we can say motivation – will be the right choice of projects that should be compared in order to learn and obtain knowledge as a consequence. A key word here is “similarity”. The comparison will be based on the similarities between projects, that make them comparable. The idea behind is that we cannot join two projects just because we think that they are equal, and should behave the same. We need to investigate carefully if they are actually comparable before using one or another as a model.

1.2 The work Structure

This work will be divided in five parts: the project management review, the mathematical review, the model development, the model application, and the conclusion. The Figure 1.1 shows the path in a pictured way, so we can see clearly what will happen in the next pages. The crucial point here is that this work will not be simply based on quantitative analysis, but a huge effort will be made to balance quantitative and qualitative analysis. In this sense, a boarder line is proposed to make this distinction clear.

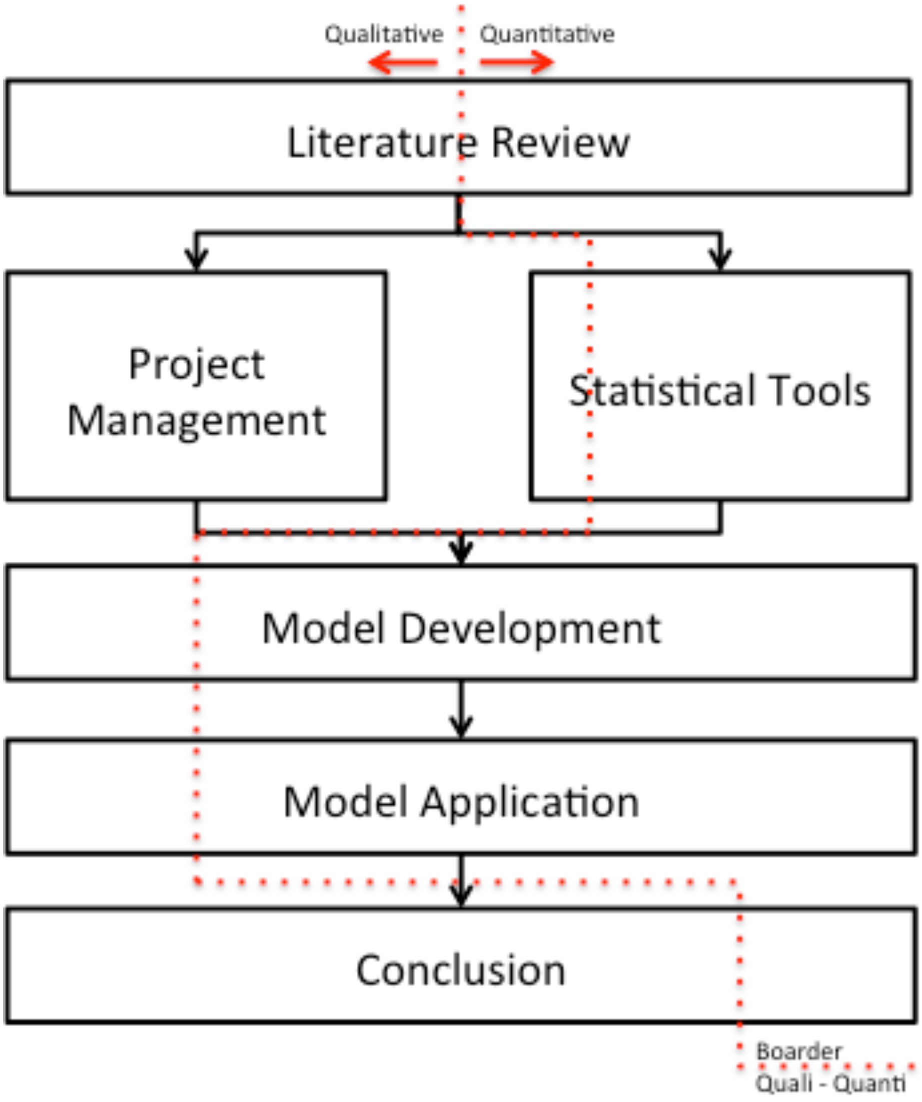


Figure 1.1 – The work structure.

So let`s give some details about the work structure, and see how we will evolve along the pages.

1.2.1 Literature Review

The project management review consists in a literature review in order to establish some concepts that should be known within the Project Management environment. We will start with project conceptualization, and then we will discuss the project phases, and finally what defines project success. The second part of this chapter will focus on tools of project management.

In this very first part the most important topics to be discussed are the project variables and the project success. The variables are for sure the main guidance for everything with the model, since the model development until the application and discussion – in other words, the variables are the foundation of this work. The project success is something that we are trying to increase, or increase the probability of success. We will not measure success in this work, but it is for sure a goal.

The mathematical review will focus on statistics and clustering theory. This chapter is crucial for our study because it will give us the robustness for our model. We need to tight up the concepts with numbers, so the tools necessary for doing that will be shown in this chapter.

1.2.2 Model development, application and conclusion

The concepts and statistical tools seen on the literature review will draw the model development. A strong point here is the boarder line between qualitative and quantitative analysis that will make the model more robust. Moreover, we will try to follow a linear approach that test the hypothesis – that work as milestones –, and if it is true, we can move forward.

The model application will take into consideration the model we've just developed and the actual data. This model will be tested with Oil & Gas data, given by the professor Franco Caron. The data was gathered from past studies of several Oil & Gas companies that developed great projects within this sector.

Finally, we will conclude our work by analyzing the outcomes of our model, and how we can improve this study with further papers. We think, so far, that our study will need some following studies in order to develop a robust theory. We think that we have this potential, but our goal here is just to give the kickoff.

2 Literature Review about Project Management

In this Chapter we are going to start presenting the object and motivations for our study in order to have a robust qualitative basis to develop properly our model and reach our objective. It should be highlighted that the purpose of this literature review is not to get into details of models and methods for managing a project. There will be a brief discussion about certain topics, which are important for our model development.

2.1 Understanding a Project

The main objective of this paragraph is to understand the major concepts about projects and also to establish the boundaries of our study. It is important to give a little recall about the main objective of this thesis:

The objective of this study is to perceive similarities throughout a range of projects, put these projects into groups (clusters) accordingly those similarities, study the characteristics of each cluster and classify as a typology of project.

Moreover, we will start this thesis with the OBJECT of our study, which is Project Management. In a more specific definition, we can say that our OBJECT is the project success based in a robust statistical model for a better project management.

For this matter, we will approach of the literature in order to cover all the points shown at the Figure 2.1 above. This approach will provide a linear understanding about the main topics that guide the model, always having in mind the objective of the study.

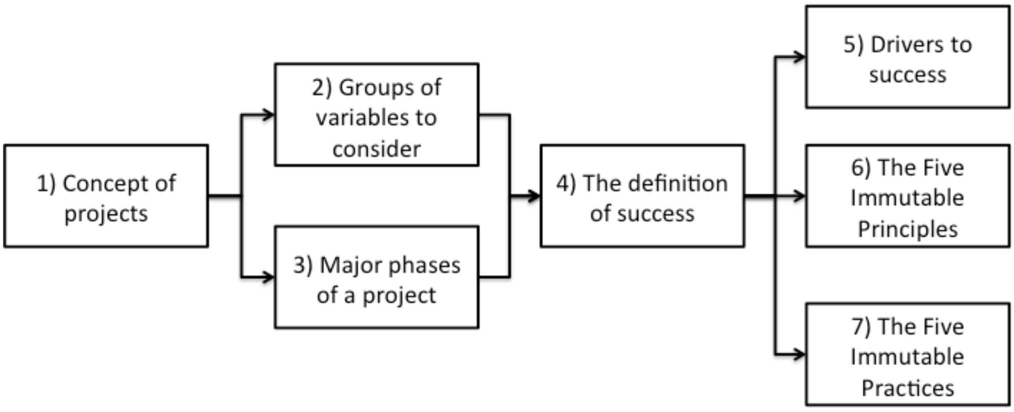

1) Concept of projects
2) Groups of variables to consider
3) Major phases of a project
4) The definition of success
5) Drivers to success
6) The Five Immutable Principles
7) The Five Immutable Practices

Figure 2.1 – Flowchart to guide the literature review of projects.
(Developed by the author)

It's possible to see that there are seven steps of this literature review. To make the comprehension easier, we will divide this paragraph in sub-paragraphs that will cover each point given. Then, a little conclusion will give us the summary of all information.

2.1.1 How to define a Project

A project is non-repetitive process[3], conducted in order to provide a product or a service. For example, a project could be a simple summer trip with your friends, or even a complex project of a power plant.

Moreover, the Project Management Institute[4] defines a project as:

> *"A project is a temporary endeavor undertaken do create a unique product, service or result."*
>
> *(Project Management Institute – Book of Knowledge, 2008)*

It becomes clear by this definition that a project has two singular characteristics. The first is related to its timeframe: a project develops a certain outcome within a certain amount of time. The second characteristic is related to the outcome itself: a project is a one-time occurrence that generates a customized product or service (Caron, 2009).

Those characteristics are founded in two attributes that every project has. The first is that every project is a "one chance to get it right". The second attribute is that a sort of variables governs a project – some are dependent, some are independent, some we can control, some we can't control (Alleman, 2014).

2.1.2 Groups of Variables to consider

The variables that govern a project could be classified in three different fundamental categories: Cost, Schedule and Technical Performance Measures (TPM), as shown on Figure 2.2 above. They are related to the following questions:

- How much will the project cost?
- How long it will take to be done?
- Will the deliverables behave accordingly to requirements?

[3] It could be distinguished from repetitive processes, like the Ford production line in the beginning of 20th Century.

[4] Project Management Institute: world's leading not-for-profit professional membership association for the project, program and portfolio management profession. It has more than 2.9 million professionals working in nearly every country in the world through global advocacy, collaboration, education and research. Source: http://www.pmi.org/About-Us.aspx

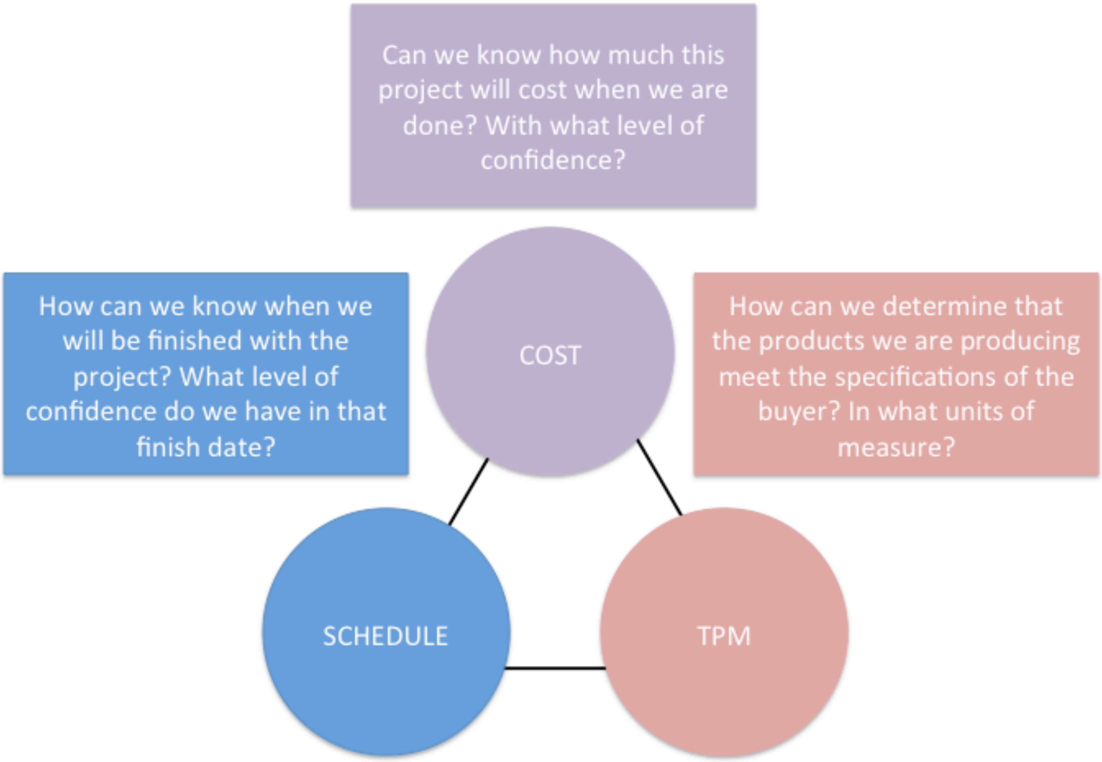

Can we know how much this project will cost when we are done? With what level of confidence?
How can we know when we will be finished with the project? What level of confidence do we have in that finish date?
COST
How can we determine that the products we are producing meet the specifications of the buyer? In what units of measure?
SCHEDULE
TPM

Figure 2.2 – Fundamental variables of projects. (Adapted from ALLEMAN, 2014.)

Our model and analysis will be based on some variables within these three groups. For example, we will take into consideration the total cost, total duration and some qualitative variables such as typology of product delivered and where it took place.

Although, it is worth saying that a good manage of a project do not take into consideration this paradigm of Cost, Schedule and Technical Performance Measure (traduced by quality of the project deliverables). There had been two evolutions of this paradigm until the right explanation about what is really important when managing a project. The former – Cost, Schedule and TPM – gives us a relation between three categories that are interdependent, or "non-MECE[5]".

The first shift introduces a certain degree of dynamicity to the project control. It states that the groups of variables that should be controlled belong to the following categories: Progress, Change and Risk. Note that the Cost, Schedule and TPM are part of this new point of view, although they are present in different groups of variables.

A third paradigm collects the variables and put them into the following groups: Risk, Value and Baseline. This new paradigm considers the precedents paradigms, but proposes

[5] MECE: it is a concept used mostly by consulting firms to describe a way of organizing information that is "Mutually Exclusive, Collectively Exhaustive". In other words, no element should represent other element (or overlap), and all elements, taken together, should fully categorize the problem as a total. Source: www.caseinterview.com/mece

some differentiations. The Project Value describes the how the customer's requirements are being satisfied, the Project Baseline gives the information about the Cost and Schedule to accomplish the project's objective and the Project Risk gives the information of how thing can get wrong, and what will be the impact (Caron, 2009). The evolution of the paradigms can be seen at Figure 2.3 above.

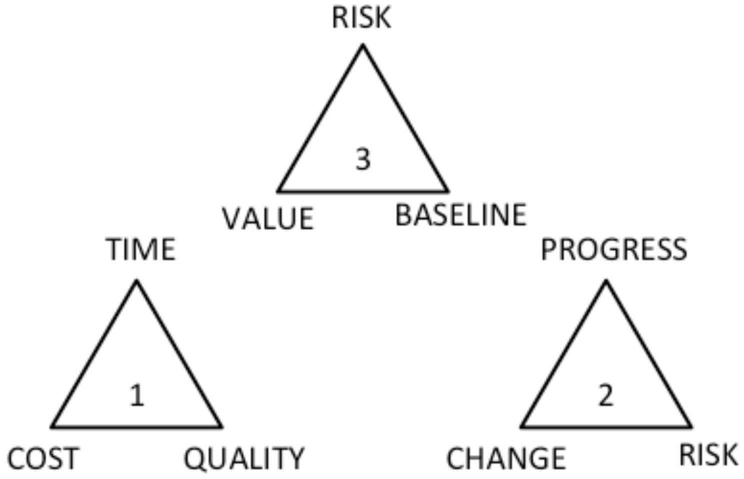


Figure 2.3 – Project Variables expressed by diferente paradigms. (Adapted from CARON, 2009)

For our purposes, we are going to express the model according the first paradigm. The reason why we will take this path is because we are not interested at the dynamic process within a project deployment. Our objective here is to deliver a result based on comparative statics[6], or in other words, we want to compare planned variables such as cost, duration and typology of product and sort this results into clusters.

Although, it is worth saying that in some moments, especially at the conclusion, we will have to recall the paradigm number three because it has some nuances that could improve this study in terms of its range of application in a future work.

2.1.3 The phases of a Project

Every project has a standard set of phases, defining its life cycle. The Project Management Institute proposes the following set of phases:

I. Initiation

II. Planning and Organizing

III. Execution

[6] Comparative Statics: an analysis tool used to compare two situations in equilibrium (without further changes). It is commonly used in the field of economics (Boitani, 1988). For our purposes, we will compare different projects with different sets of variables before something changes (planning phase, or initial equilibrium) and after everything has changed (final result, or final equilibrium).

IV. Closure

The first phase – Initiation – is related to the definition of the project. In this phase we develop the main objective and the strategy behind the project. The product or service is discussed in order to know the requirements of the client[7]. The second phase – Planning and Organizing – defines with a great degree of details the outcomes of the project and how they will reach those outcomes. The third phase – Execution – is where the resources will be consumed, in terms of work force and budget to construct the actual product or service. The final phase – Closure – is the moments that the outcome of the project is transferred to the client, and the knowledge received by the completion of the project is documented (Caron, 2009).

This work is mostly related to the phases I, II and IV. The Execution is not part of our scope. Just to recall, our objective is to compare projects before the execution, or at FID[8], and group them into clusters. Moreover, do the same for projects after execution and group them into clusters. By this approach we can see the similarities and improve our learning process.

The reason why we focus on this objective, and consequently in phases I, II and IV, is because the learning process, and the achievement of knowledge for future projects is critical for their success. We can see by the Figure 2.4 how phases II and I can influence the results of a project, without being high costly.

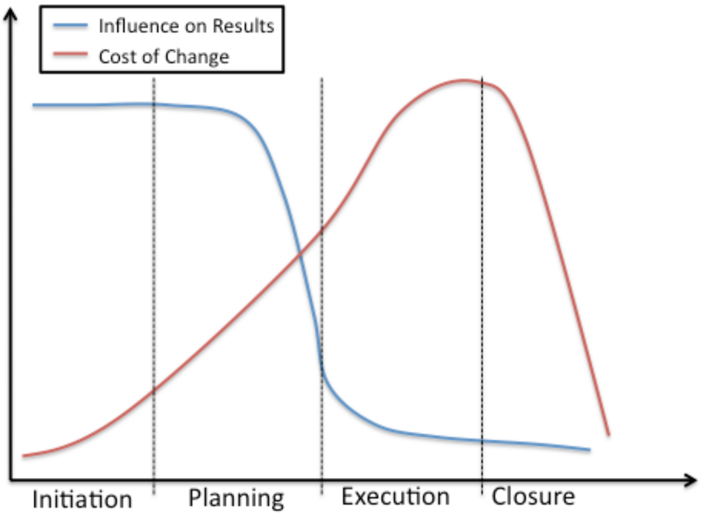


Figure 2.4 – Influence on results and cost of change during the life cycle of a Project. (Adapted from CARON, 2009)

[7] We will use the term “client” to define the entity that is requesting a project. It could be an external client (e.g. a family that want to build a house) or internal client (the board of directors that strategically introduces a project for increase manufacturing capacity).

[8] FID: Final Investment Decision is the trigger to project execution, after the feasibility study. This term is commonly used at the Oil, Gas & Petrochemical Industry.

Notice how great is the importance on results of a well-structured and developed Initiation and Planning phases, and by synergic effect of low cost of changes, how great is the overall importance of those phases. And to reach this structure and develop it is important also the Closure phase, where the knowledge will be documented to a proper use in next projects.

2.1.4 Defining Success for a Project

Many authors indicate that success of a project is usually seen by the analysis of the groups of variables we have mentioned before: cost, schedule and quality. When the actual value of those three groups of variables meets the planned ones, the project is considered successful.

On the other hand, we've seen projects that did not met their objectives in terms of cost, schedule and quality, and even though were successful. They are cases like Windows from Microsoft, Macintosh from Apple, and Taurus from Ford. All those three examples reached success not by the project variables, but from other sources (Shenhar & Dvir, 2007).

So, those three groups of variables must not fulfill our analysis of a project success. We have to take into consideration also the effectiveness of a project developed, by taking the correspondence between the outcome, and how this outcome supports the corporate strategy and stakeholders' interests (Cserháti & Szabó, 2013).

Actually the answer if a project was successful or not is a very complex analysis. Takes into consideration the success criteria and success factors[9] that the project team should address to the project itself at the beginning. Every type of project has its own criteria and factors for success. For example, a bank industry project considers countries' political situation a very strong factor for success (Ika, Diallo, & Thuillier, 2012). In the case of Events Projects, the stakeholder's satisfaction would be one factor to consider (Cserháti & Szabó, 2013). For urban projects one important factor for success is considering the "minimization of conflict between stakeholders" (Yu & Kwon, 2011).

As we can see, it is not trivial to address a sort of criteria and critical factors for success at a project. But, as we could see from the research is that both this aspects are generally the same, what matters instead is the importance each one is given for each type of industry.

[9] The difference between a criterion and a factor is that a criterion is a rule or principle for evaluating or testing something. A factor instead is one of the elements contributing to a particular result or situation.

To make our comprehension more generalized and applicable for a broad range of industries we will divide our critical factors for success into some categories (Shenhar & Dvir, 2007). The Figure 2.5 above shows how to engage the success of a project by looking at the critical factors for success.

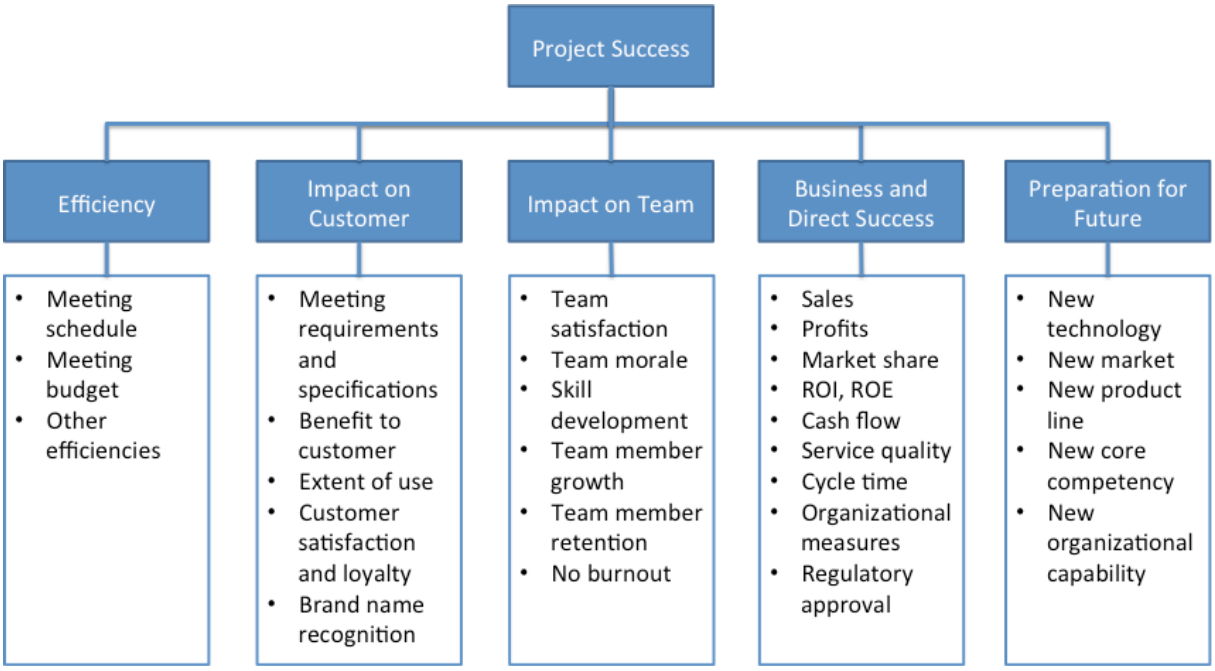


Figure 2.5 – Hierarchy of Critical Factors for Success.
(Shenhar & Dvir, 2007)

The Figure 2.5 indicates in one-word basis what are the most common measures for a project success. The attempt here is to be MECE. The figure also shows a several critical factors of success (CFS). It is worth saying that every project, belonging to a certain industry reveals its own set of factors. For example, if we take a look at the Los Angeles subway system we can infer that was not a successful project (Shenhar & Dvir, 2007). Let us analyze all the CFS`s:

- Efficiency: the project met all requirements in terms of budget and schedule, and was voted as “Project of the Year” by the PMI in 1993;
- Impact on Customer: that was the problem. Given that L.A. is a city of automobile, the customer did not see the metro as an alternative way of transport;
- Impact on Team: as a project, the subway system was successful for the team, as it generated knowledge for new projects;
- Business and Direct Success: because no customer tended to use the metro in L.A., the project did not achieve a great return. The misuse generated great losses;

- Preparation for Future: in this area the project was successful because led to new technologies of developing subway systems. Moreover, it led to the idea that cost, budget and quality are not the only CFS`s.

Notice that even though it met some CFS`s, the project was not successful. So, we can derive a question from that: there is something missing on our analysis? Or any project should present all CFS`s in order to achieve success?

As our later examples showed, it is not necessary to present all CFS`s in order to achieve success, but there is for sure something missing in our analysis.

More than having a static perspective of CFS's, we should also address three more variables on analyzing them. They are not really variables, but perspectives that engage different CFS's for different projects (Samset, 1998). The Figure 2.6 shows the dynamics of the process.

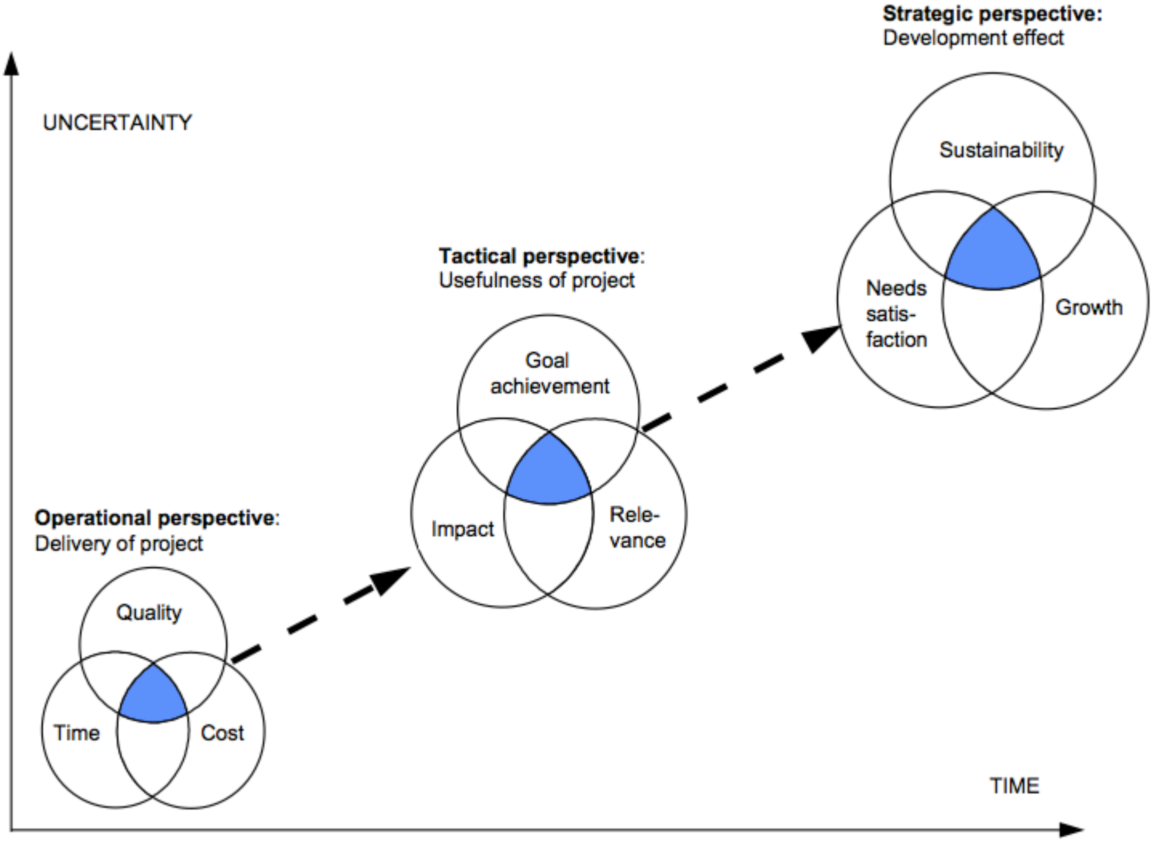


Figure 2.6 – Different measures for success in three different perspectives.
(Samset, 1998)

Notice that for different time frames and different levels of uncertainty, the measures that most matter changes. So, the Figure 2.5 and Figure 2.6 should be interpreted together in order to give a better support for our analysis.

Given that, we can say infer two main comments:

I. Meeting resources constraints indicates well-manage and efficient projects, but not necessarily successful ones;

II. When using different time frames and different level of uncertainty, project success becomes a dynamic concept with both short and long term implications.

There are several frameworks utilized in order to achieve project success. The frameworks should be in line with the typology of project that you have, and the purpose, or objective of your project. By the fact that we are dealing with Oil & Gas industry, and by the fact that we have a range of sixteen projects to analyze, we should consider some parameters for our project success framework.

We will follow the traditional mind-set that says that project success depends on the old paradigm Cost, Schedule and Quality. This school says that all other CFS`s should be addressed to high level of the company's hierarchy. The project manager should be just concerned about doing the project within the budge, time and requirements constraints (Shenhar & Dvir, 2007).

We chose to walk through this path because our model will be developed not from the point of view of the firm, but the point of view of the single project. That is why we can, and we should, aggregate our efforts the maximum on the project and project variables, giving by side the strategic sense of all.

So, finally, once that we reached a great view of the ways to measure success and the dynamics that involve those measures in different perspectives we can show a structured way to achieve project success by those means. Remember that only analyzing the CFS's is difficult and not a structured way to guarantee the success of a project. For that purpose we will analyze a methodology created by Alleman (2014) that gives a step-by-step process to achieve the success. The Figure 2.7 shows how it will work.

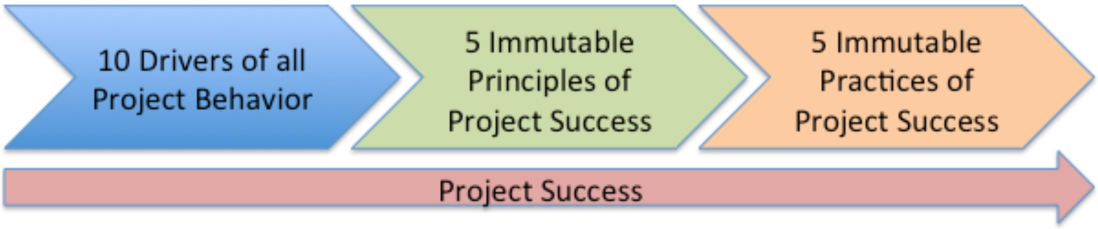

10 Drivers of all Project Behavior
5 Immutable Principles of Project Success
5 Immutable Practices of Project Success
Project Success

Figure 2.7 – Methodology for Project Success.
(Adapted from ALLEMAN, 2014)

The reason why we are going to spend some time to grasp this process is because we need to understand what phases of the project are important to achieve high levels of success, and focus our model to pursue better results in those phases.

2.1.5 Ten Drivers for Success

The first aspect that we should be aware to achieve the project success is to consider the Ten Drivers that govern the project behavior. And those ten drivers are organized in three different groups, related to the project's life cycle phases: Planning, Executing (related to deliverables) and Performance Management (related to Initiation, Planning, Execution and Closure phases).

Instead of listing all Drivers, for better comprehension we will show them within their categories in a pictured way. The Figure 2.8 above is a good form to visualize the system that we are analyzing.

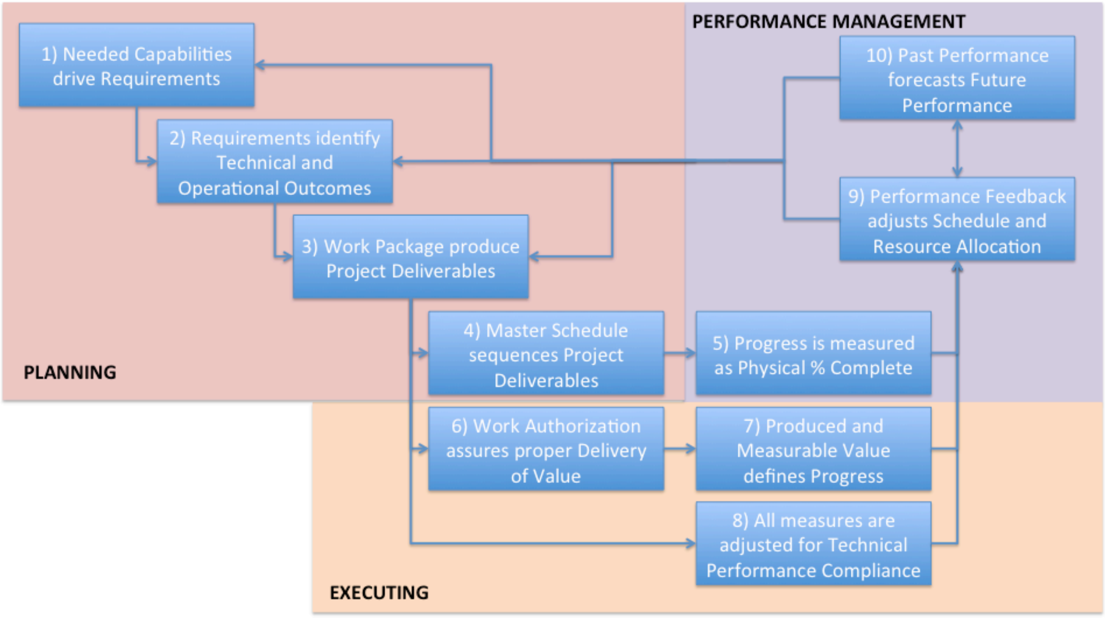

PERFORMANCE MANAGEMENT
1) Needed Capabilities drive Requirements
2) Requirements identify Technical and Operational Outcomes
3) Work Package produce Project Deliverables
4) Master Schedule sequences Project Deliverables
5) Progress is measured as Physical % Complete
6) Work Authorization assures proper Delivery of Value
7) Produced and Measurable Value defines Progress
8) All measures are adjusted for Technical Performance Compliance
9) Performance Feedback adjusts Schedule and Resource Allocation
10) Past Performance forecasts Future Performance
PLANNING
EXECUTING

Figure 2.8 – The 10 Drivers for Project success.
(Adapted from ALLEMAN, 2014)

At a first glance we can notice that the drivers (9) and (10) feed the drivers (1), (2) and (3). Moreover it seems that this approach is cycled between PLANNING, EXECUTING and PERFORMANCE MANAGEMENT.

In reality it is not properly a cycle, because we are talking about a project (as we said before, it is a one-time occurrence). But it is important to highlight the notion of gathering data of other projects or within the same project to forecast and/or predict the behavior of our current job. The Figure 2.8 expresses this relation, not a cycled relation.

This is extremely correlated to the main objective of this thesis: use other projects' data, group them by similarities and utilize this knowledge to help forecasting and predicting the behavior of our project, comparing it to similar groups of projects.

Though, the drivers (9) and (10) will be extremely useful to determine the boundaries of our study. They concentrate the core of our model, and because of that they should be remember throughout the chapters (and we will recall them every time that is necessary).

It is also important to mention the role of understanding what the client want: what are his objectives, and how it will transform into requirements for the project development. Understanding the capabilities that the client is looking for is another variable to categorize and group similar projects. So, the driver (1) will be useful during this work.

Finally, we can assume that the drivers (2) through (8) are not important to our study. They have actually great role determining the success of a project, and must not be misjudge. But, having in mind our objective, it does not aggregate much talking about those drivers.

It is time to continue our path for project success and deepen on the principles and practices for reach this success.

2.1.6 Five Immutable Principles

The Five Immutable Principles are governed by the Ten Drivers. Besides that, they are highly correlated to the project phase. Actually, added to the Ten Drivers they can be understood as a transcription of the phases in a form that fit better to the purposes of our study. Their approach is clearer to the proposed clustering process.

Alleman (2014) proposes Five Immutable Principles (5IP) that are exposed in a form of questions. The 5IP are critical aspects that must have great attention during the planning process of every project. The 5IP are:

I. What the client is asking for?

II. How do we get what the client is asking on Time and on Budget with acceptable outcomes?

III. How to assess resources?

IV. How to handle impediments and assess risk?

V. How to measure progress to plan?

Again, we will focus our attention just for the principles (I), (II) and (III). The principle (IV) is related to eliminate or control the risks. That principle is very important, indeed, but for our purposes we are just take a glance on it. There are several types of risks, which derivate from different forms of uncertainty: natural variation, foreseen uncertainty, unforeseen uncertainty and chaos. Moreover, there are different sources of risk: external world, our knowledge of this world and our perception of this world. Our model do not utilizes as input the information of this principle, but can be expanded to supply information for a better risk analysis. Unfortunately this is not our goal, so we will let this principle aside.

Taking a glance at the fifth principle, the author exposes the necessity of having a measure for physical progress, in terms of deliverables that will correspond to the actual progress of the project. He also alerts the fact that is important to measure performance related to this physical progress. Again, this is not our goal because this principle is highly correlated to the Execution phase of a project.

Let's focus, than, at the first three principles. Starting with (I) what the client is asking for? This principle reveals the importance of a good start for the project. It is necessary to understand perfectly what your client is requesting, in terms of final product and its capabilities.

The reason why this is important is because knowing the requirements of the client we are able to put boundaries and isolate our system. We know what to produce, nothing more and nothing less. We are able to develop a well-structured Planning Phase.

In terms of our study, the typology of product and its qualitative characteristics are some possible variables for group the project into cluster. Similarity between projects initiate with a proper comparison in terms of characteristics and capabilities.

Many authors utilize for this principle the term: project definition.

Project Definition is a process used to identify the needs of the stakeholders and the specifications of the product or service are defined. It usually belongs to the first phase of a project, and has the purpose to establish its size, scope and complexity (Cano & Lidón, 2011).

There are also some models that have a precise approach to limit the errors that can be done on project definition. One good example is a well-known model used to help improve quality in services called SERVQUAL. The model tells us about focusing in five gaps between client's expectations, producer perceptions and service delivered (Westbrook & Peterson, 1998). The simplicity of this model made it capable of approaching also the problem of project definition. The Figure 2.9 above shows the five possible gaps within a project lifecycle.

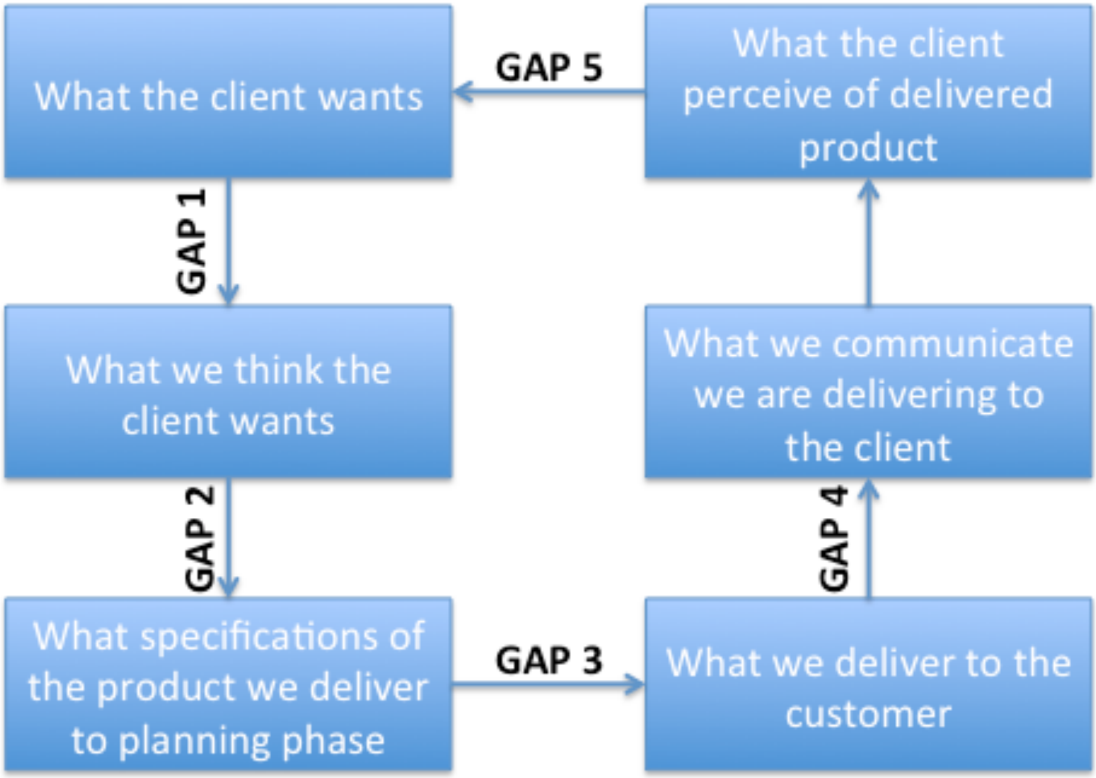

What the client wants
GAP 5
What the client perceive of delivered product
GAP 1
What we think the client wants
What we communicate we are delivering to the client
GAP 2
GAP 4
What specifications of the product we deliver to planning phase
GAP 3
What we deliver to the customer

Figure 2.9 – Five gaps at the Project life cycle.
(Developed by the author)

It is possible to note that the Gap 1 is the problem of project definition. If we have a miscomprehension about the Gap 1, the role project has the risk to be doomed, and consequently the project success tend to go zero. That is why we should give a good attention on project definition or the IP1.

Talking about the IP2, instead, we can immediately notice that it is related to the variables Cost, Schedule and Technical Performance Measures we've mentioned before. This principle assesses the ability of developing a proper estimation of these variables, during the Planning phase.

For our purpose, this principle has a central role to develop the model. Together with the first principle we are able to touch all variables to analyze similarities between project ex-ante[10]. In other words, we will be able to compare projects by comparing the variables that arise from these two principles (e.g. typology of product, typology of client, cost planned, duration planned, etc.).

The principle number 3 (IP3) is related to the resources used by the project. Seems not to be our objective here, but if we want to know better the similarities between projects, one possible variable that could arise is the effort in terms of resources to deliver the result.

[10] Ex-ante here means before the Project is executed, or simply during the Planning phase, or at FID.

Later we will discuss that in terms of our data, this is not really important, but the model is constructed in a flexible way, so for other projects, in other circumstances this principle can be really useful and give us some very strong insights.

From now, we are going to finish this discussion about the 5IP enunciating that the model will only utilize the first three principles, and in a flexible way, depending on industry, project and client we are dealing with.

2.1.7 Five Immutable Practices

The Ten Drivers and the Five Immutable Principles gave us the information of what kinds of variables are important for our clustering model. But for the project work properly and behave as expected we should be aware of the Five Immutable Practices. But why this is important? It is important because if we discuss the similarities between projects ex-ante, but each project is governed by different practices, they will not deliver the expected results, and the model will collapse.

Moreover, the model will collapse because the groups' patterns ex-ante will certainly be different from the groups' patterns ex-post[11].

The Five Immutable Practices (5IPr) are also strictly correlated to every phase in the project life cycle. But not as a form of transcriptions of each phase, but rather the actions that should be taken in every step of project life cycle.

It is worth comment about the world "immutable". Alleman (2014) consider both the principles as the practices as immutable in the sense they are permanent and equal for every type of project. It is not an affirmative implying that this is the right way to analyze and approach a project. Instead, this is part of a methodology of increasing the project success that rewrites the old paradigms (e.g. how the PMI approach the project management).

So, the practices that Alleman (2014) proposes can be synthesize by the Figure 2.10 above, as well the steps to apply each practice and their benefits.

[11] Ex-post here means the results that the projects generate, in terms of values of the variables we are analyzing.

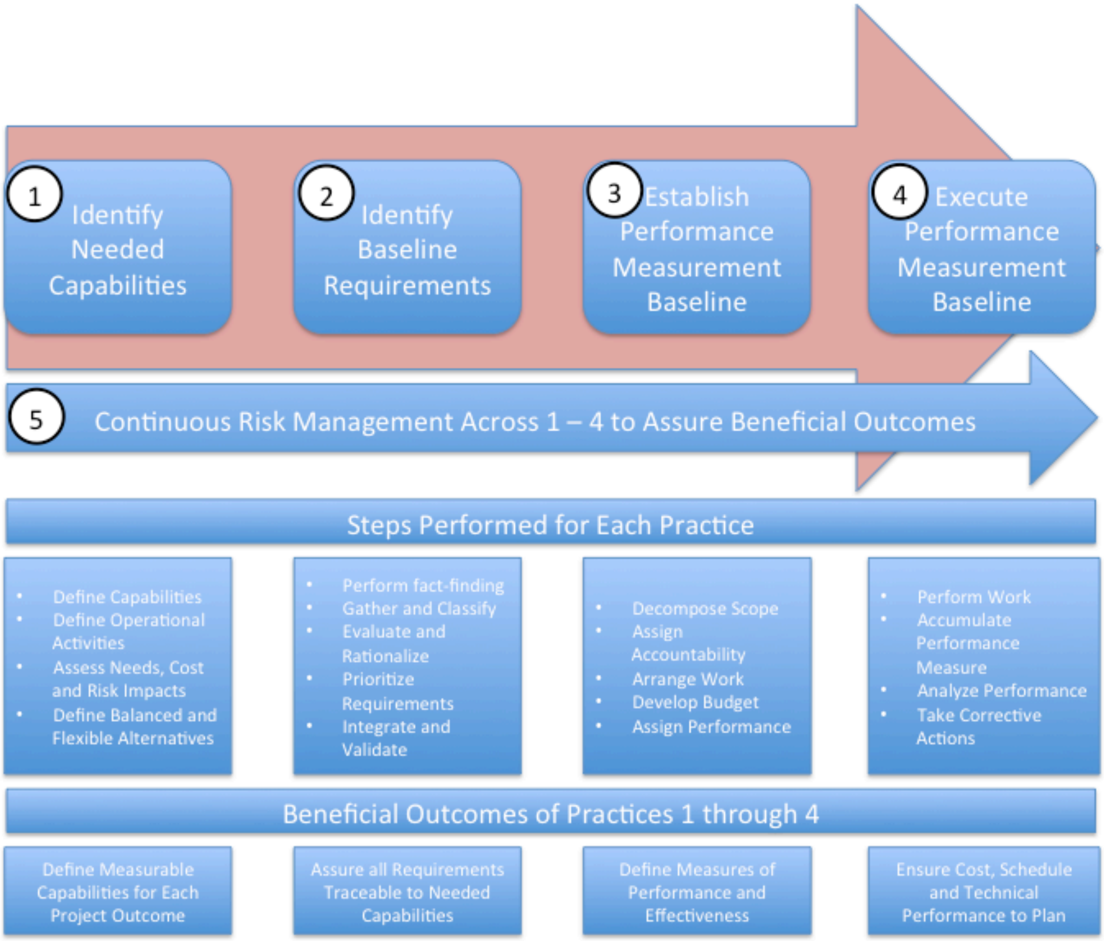

1 Identify Needed Capabilities
2 Identify Baseline Requirements
3 Establish Performance Measurement Baseline
4 Execute Performance Measurement Baseline
5 Continuous Risk Management Across 1 – 4 to Assure Beneficial Outcomes
Steps Performed for Each Practice
Define Capabilities
Define Operational Activities
Assess Needs, Cost and Risk Impacts
Define Balanced and Flexible Alternatives
Perform fact-finding
Gather and Classify
Evaluate and Rationalize
Prioritize Requirements
Integrate and Validate
Decompose Scope
Assign Accountability
Arrange Work
Develop Budget
Assign Performance
Perform Work
Accumulate Performance Measure
Analyze Performance
Take Corrective Actions
Beneficial Outcomes of Practices 1 through 4
Define Measurable Capabilities for Each Project Outcome
Assure all Requirements Traceable to Needed Capabilities
Define Measures of Performance and Effectiveness
Ensure Cost, Schedule and Technical Performance to Plan

Figure 2.10 – The Five Immutable Practices for Project Success. (Developed by ALLEMAN, 2014)

It is possible to notice that the practice (1) is related with the phase of Initiation at the project life cycle. The practices (2) and (3) are related to the phase of Planning. And the phases (4) and (5) are related to the phase of Execution.

Again, our focus is just on the phases of Initiation, Planning and Closure, so the practices (4) and (5) is not something that we will care along this study. The practices (1), (2) and (3) instead will be mention during the development of the model.

Also, it is worth saying that the author forgot to mention one practice that have always been extremely important, and should be considered as part of the “Immutable” Practices. We are referring to the practice of Document and Learning, belonging to the last phase (Closure) of project life cycle.

This practice is important because will lead future projects for the correct, or more precise process of estimation and planning of the cost and schedule baselines.

Moreover, it is good to remember that this methodology is focus on project success. So it will only guide us on the development of our model. Any change on this methodology will be considered as an adaptation necessary to best absorb the literature about project success.

2.1.8 Summary

In this paragraph we've started restating our objective, and going to the literature to have a robust background to support our model development and reach our goal.

The first part of this paragraph was based on project concepts: what it is, what are the variables that govern it, and what are the phases that compose a project from the beginning until the end.

The next step was based on the word "success", and our attempt to define what does it means into the project environment. Our model should be used to increase probability of a project success, by helping in some areas that are not well supported by a robust mathematical analysis.

The third part of this literature review was based on a methodology to increase project success' probability proposed by Alleman (2014). The methodology that consist in a linear thought of understanding what are the drivers that govern a project, what principles should those drivers be following and what practices will make a good development of a project to reach success.

However this methodology should be adapted to our purposes because it not takes into consideration the fact that Documentation and Learning are vital, not for a perfect cycle within a project in development. But to boost the success for the next project that will be done. In this aspect the methodology seemed a little flawed.

For the next paragraph we will examine a portfolio management. The purpose of doing so is to get the knowledge acquired about a singular project and see the synergy with other projects that are being developed ate the same time.

In this sense, it becomes clear how important is the practice of Document and Learning. It certainly help to improve the success of other projects that are being developed, and the gain for the portfolio would be much higher.

2.2 From the Object to the Motivation

In the previous paragraph we've discussed the project concept and development and correlate our objective with the project management. We've argued that a good process of forecasting and learning could be vital to project success.

We've also discussed the relation between the phases of a project and the Five Immutable Principles and Practices. In summary, we have exposed our OBJECT of study.

Now, it is time to continue our analysis by the side of MOTIVATION of the study. In other words, once we know that "project success based on a robust process of forecasting and estimating project variables" is our object of study, we want to know "where" can we apply our study, "how" we can do it, and "why" we want to do it. Moreover, we want to express the importance of the model.

To accomplish this mission, we will start our discussion in this paragraph with "where" and "how" we apply our model. For doing so we will recall the Drivers, Principles and Practices shown in the previous paragraph. But the essence now is to show how the Project Management Office (PMO)[12] uses the learning acquired in other projects to develop a proper planning process on new projects. In other words, the first part of this paragraph will focus on the Project itself as a motivation of the study.

The second part will take into consideration a sort of projects that, together, form a Portfolio. We will show how important is the success of a project for the success of the overall, or the portfolio. The success of the portfolio could mean the success of the entire company. This section could be seen as the "why" for our study.

It is worth reinforcing here our objective for this study. Our model is not focus on demonstrating the importance of a good process of planning for the success of the project, or the portfolio or the company. This relation will be shown by the literature that already exists. We are actually trying to develop a model that helps the process of planning, based on statistical tools for helping forecasting and estimating some variables of the project. The Figure 2.11 above exposes more clearly the macro-structure of this paragraph.

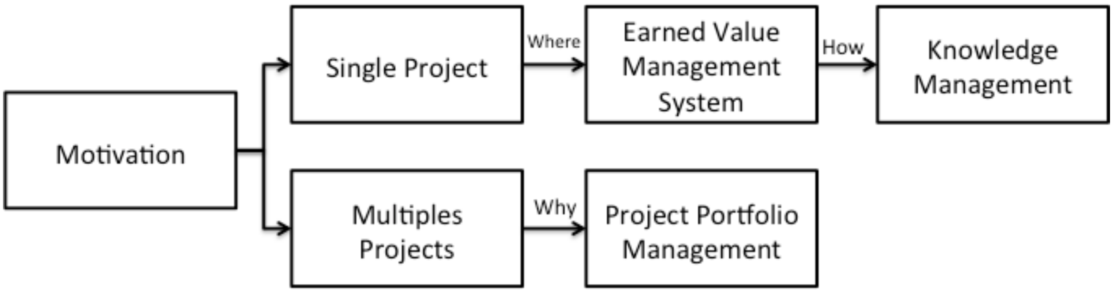


Figure 2.11 – Macro-structure for the study motivation.
(Developed by the author)

[12] Project Management Office: department of a company that provides project-related services as a temporary entity established to support a specific project or program. May include supporting data management, coordination of governance and reporting, and administrative activities to support the project or program team (Project Management Institute, 2013).

As we can see, there are two main topics to cover, which are defined by Single and Multiples Projects. Then they open the discussion to a more detailed level, which represents the Earned Value Management System (EVMS) as best practice for executing well a project. Later we will discuss about the Knowledge Management, as being a support for data used at EVMS. Finally the Project Portfolio Management (PPM) as a company tool that brings all together for better results to the entire company.

2.2.1 The view of a single project

In the previous paragraph we have said that we will consider as variable of a project the Cost, Schedule and Quality (or Technical Performance Measures), as we can see on Figure 2.2.

The success of a project is based on how we identify the TPM's, how we plan and estimate the Cost and Schedule, how we control these variables during the execution of a project and how we document the project life cycle for a learning process for future projects (as we saw during the discussion of the Five Immutable Principles and Practices).

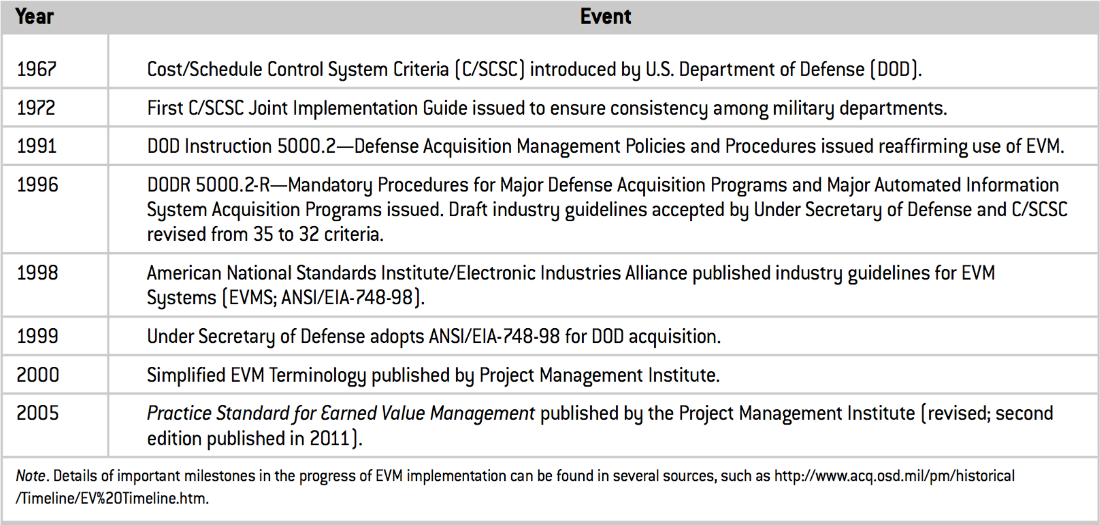
The most common tools for control and monitoring of a project during the execution are the Critical Path Method (CPM) and the Earned Value Management System (EVMS). The first is related to the schedule of a project, in terms or work packages[13], but do not take into consideration (in a robust way) the costs of a project either the TPM issues. For that matter the Earned Value Management (EVM) was developed: a model that integrates scope, cost and schedule under the same framework. Moreover it allows the project managers to verify the progress of a project and confront it with what was projected on the planning phase (Acebes, Pajares, Galán, & López-Paredes, 2013). To understand well this model as an integrative part of our study, we need to understand first its history and criteria. Later we will discuss how it works.

2.2.1.1 Earned Value Management System

The first occurrences of the EVM can be traced back in the late 1800s, as a basic methodology for controlling the financial aspects – such as cost and cost-in-time – of a project. In 1967, the U.S. federal government introduced the EVM to understand the financial aspects of programs and to have a consistent methodology for controlling such programs (Kwak & Anbari, 2011).

[13] Work Packages: dissociation of the total work into small elements of work with the purpose of a better management (Caron, 2009).

The EVM had progressed along some lines until its standardized practice – and consequently wider use – from the Project Management Institute in 2005, as we can see from Figure 2.12 the EVM progress timeline.



Year	Event
1967	Cost/Schedule Control System Criteria (C/SCSC) introduced by U.S. Department of Defense (DOD).
1972	First C/SCSC Joint Implementation Guide issued to ensure consistency among military departments.
1991	DOD Instruction 5000.2—Defense Acquisition Management Policies and Procedures issued reaffirming use of EVM.
1996	DODR 5000.2-R—Mandatory Procedures for Major Defense Acquisition Programs and Major Automated Information System Acquisition Programs issued. Draft industry guidelines accepted by Under Secretary of Defense and C/SCSC revised from 35 to 32 criteria.
1998	American National Standards Institute/Electronic Industries Alliance published industry guidelines for EVM Systems (EVMS; ANSI/EIA-748-98).
1999	Under Secretary of Defense adopts ANSI/EIA-748-98 for DOD acquisition.
2000	Simplified EVM Terminology published by Project Management Institute.
2005	*Practice Standard for Earned Value Management* published by the Project Management Institute (revised; second edition published in 2011).

Note. Details of important milestones in the progress of EVM implementation can be found in several sources, such as http://www.acq.osd.mil/pm/historical/Timeline/EV%20Timeline.htm.

Figure 2.12 – EVM progress timeline
(Kwak & Anbari, 2011)

Within the time period of late 1800's, until 2005, the EVM suffered several modifications and standardizations. We can attribute a massive importance for the occurrence of 1998. The American National Standards Institute/Electronic Industries Alliance (ANSI/EIA) published that year the guidelines for the EVM System, as an effort to give the EVM a wider use. They've identified 32 criteria (see Appendix A) that an EVMS should have in order to perform well. Those criteria were divided into five categories [took from (Kwak & Anbari, 2011)]:

- *Organization*: Activities that define the scope of the effort and assign responsibilities for the work;
- *Planning and Budgeting:* Activities for planning, scheduling, budgeting, and authorizing the work;
- *Accounting:* Activities to accumulate the costs of work and material needed to complete the work;
- *Analysis:* Activities to compare budgeted, performed, and actual costs; analyze variances; and develop estimates of final costs;
- *Revisions and Data Maintenance:* Activities to incorporate internal and external changes to the scheduled, budgeted, and authorized work.

Note that those categories encompass all the three first phases of the project life cycle (Initiation, Planning and Execution). Hence, it is notorious how the EVMS have earned the importance and usage that it has nowadays. Besides that, it is possible to notice why the EVMS is a motivation for our study (where we can, or should, use our model as support for a better control and monitoring).

For our purposes, we should focus on the *Organization* and *Planning and Budgeting* categories, which imply that for a good better use of the EVMS framework we must be concerned about the variables paradigm Cost, Schedule and Quality. Furthermore, we should take into consideration the scope of the project, and forecast the cost and schedule, guided by the TQM's.

Moreover, to successfully implement the EVMS we should be aware some factor, that are divided in three categories: Factors for better acceptance of EVMS, Factors for better use of EVMS, and Factors for better performance of EVMS (Kim, Wells Jr., & Duffey, 2003).

In terms of better acceptance, we can list the believe of top management on the EVMS, based on proven track records exposing the utility of EVMS in achieving project success. Besides that, the acceptance of project managers and their team, also based on records of EMVS functionality. Notices that both of these factors are related to a good documentation process of past projects for a better understand of future projects. Finally, the third factor could be the flexibility to project managers on selecting their own framework of EVMS, accordingly to their projects (Kim, Wells Jr., & Duffey, 2003).

For the factors for better use of EVMS, Kim et al. states:

- Long experience using the EVMS: as well, this factor is contributed by the documentation process;
- Use of integrated teams;
- Use the CPM as a complementary tool;
- Cross-organizational communication within the company.

Finally, the factor for better performance of EVMS could be seen as the project managers' expertise, use of good computer software for supporting the EVMS and high level of communication within the organization.

The more important here is to notice that the EVMS is also related to the Closure phase of project life cycle. We could see by the study of Kim et al. that for the EMVS work properly, it should be founded on the documentation of past projects. This inference closes the gap and shows how the EVMS encompass all four phases of project life cycle. Beyond that, it suggests the importance of EVMS, and why it is a motivation for our study.

So, to give a more strict view about the motivation, taking into consideration the vision of a single project, we will explore the relation between the project life cycle, and EVMS framework and documentation process.

To understand the relation between EVMS and project life cycle, and consequently, the relation with our study, we need to understand this framework. In this sense, we will start this part of discussion by explaining briefly how the EVMS operates.

Basically what the EVMS does is control and monitoring the cost and schedule of a project by comparing three major values: the planned cost, the actual cost and the earned value, which is the planned cost of the work that was actually done. In other words, at the Time Now[14] (TN), we have the data of work that have been already done, which is a percentage of total work, and a planned cost for this percentage. This is the earned value.

Graphically, the Figure 2.13 above shows the relation between this three curves of planned and actual cost, and the earned value.

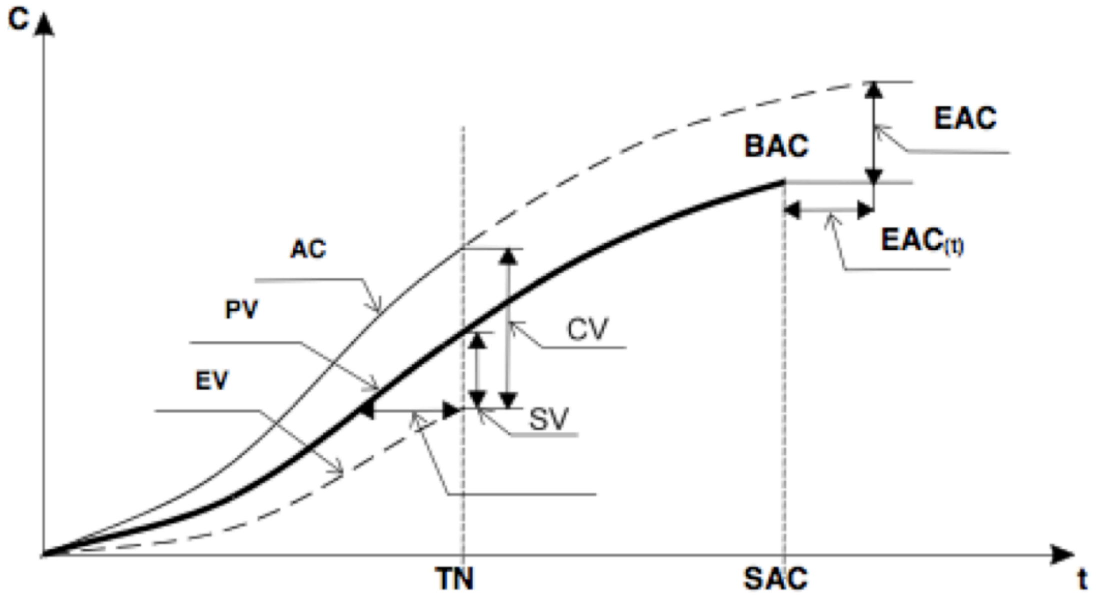


Figure 2.13 – EVMS curves.
(Merli, 2010)

To better understand the Figure 2.13 it is necessary expose the terminology used into EVMS framework. The Table 2.1 was crafted to show such terminology.

Table 2.1 – Terminology used at EVMS.

PV	Planned Value
EV	Earned Value
AC	Actual Cost
CV	Cost Variance (CV = EV – AC)

[14] Time Now: generic instant of time during the execution of a project, used to control and monitor the variables that compound the project.

SV	Schedule Variance (SV = EV – PV)
CPI	Cost Performance Index (CPI = EV/AC)
SPI	Schedule Performance Index (SPI = EV/PV)
BAC	Budget At Completion
SAC	Schedule At Completion
EAC	Estimate At Completion

Source: developed by author

The Planned Value (PV) is also known as Budget Cost Work Schedule, and it finishes at the BAC. This curve is what we are aiming. It corresponds to everything that was done during the Planning Phase of the project (in terms of cost and schedule), based on the requirements specified on the Initiation Phase. As said before, the EVMS encompass every phase of a project development. Moreover, recalling our object of study, our attempt here is to offer a proper deployment of the PV, based in the similarities of a project with past projects, to finally encounter project success.

Beyond that, it is worth saying that every slip of the project from what were planned shifts the AC curve and the EV curve. Such shifts aggregate the value of CV and SV, as well decrease the CPI and SPI values. The CPI and SPI are together valuable fonts of data for control and monitoring the project progress, in terms of cost and schedule. Furthermore, they are important also for calculating the EAC, which is now different from BAC (when the slip occurs).

A lot of studies were made in order to calculate the EAC based on CPI and SPI. Caron et al. (2012) and Merli (2010) are good examples of a statistical approach for the calculation of EAC. Both studies focus on the dynamic progress of a project, in the sense that at the Time Now it is possible to calculate more accurately the EAC.

To estimate the duration of a project, Lipke (2003) proposed the Earned Schedule, a new approach that does not take into consideration the use of SPI as calculated as usual for estimation. What Lipke found is that the value of SPI, because it is calculated from the relation between EV and PV, it converges to the value of 1, even for projects that not following correctly the planned schedule. For this purpose, the author proposes a methodology that uses the projection of the curve PV in terms of time. The new value for SPI would be:

$$SPI = \frac{Te}{T}$$

Where *Te* is the Time Earned, gathered from the value of EV at the Time Now, projected on the PV curve.

Beyond that, Lipke (2006) refers the Earned Schedule as a connection of EVMS and the schedule estimation. Together they are the foundation of control and monitoring a project during execution.

Our motivation in this sense is to give a statistical background to plan the cost and schedule baseline (or the PV) properly. If we accomplish that task, we could infer that the CPI and SPI would reduce in terms of variance and the EAC and SAC would be easily calculated. The Figure 2.14 below shows an example of a project such as the CPI and SPI were calculated along the project progress and computed to originate a graphic. The Figure 2.15 shows the CPI and SPI stabilized around the value 1 for a different project.

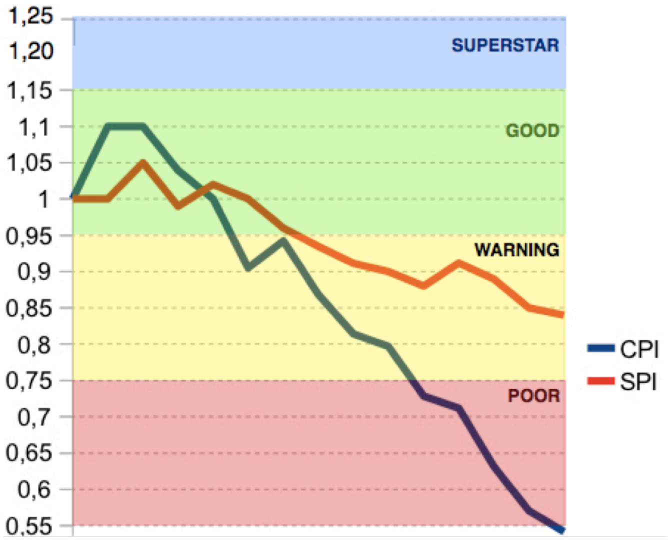


Figure 2.14 – CPI and SPI along a Project execution.
(Merli, 2010)

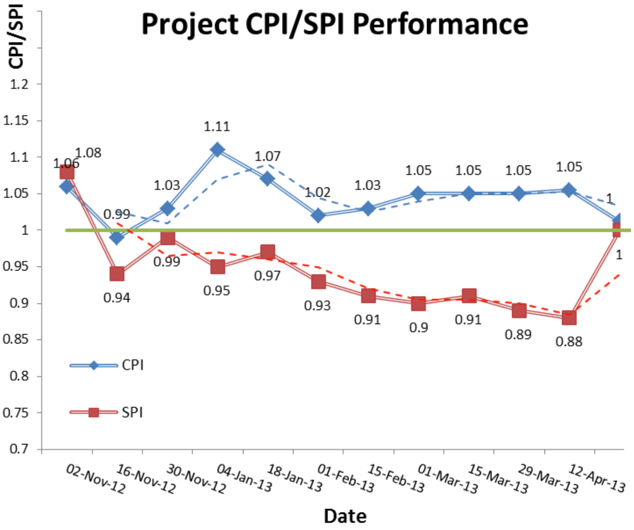


Figure 2.15 – CPI and SPI in a stable form.
Source: www.projectcontrolacademy.com

Our model, if well developed, would be able to stabilize the value of CPI and SPI during the project execution because the PV would contain the information of previous projects about their variations from the AC seen. Because of that, the AC and PV would have a similar form for this new project, and consequently the EV will move along the same lines.

2.2.1.2 Knowledge Management

To EVSM work properly, though, it is necessary a good planning process. To have a good planning process, it is necessary the knowledge acquired from other projects. So, the concept of Knowledge Management gains strength in this context.

The broad concept of Knowledge Management refers to the identification of the collective knowledge in an organization (for example an Engineering & Contracting organization) in order to leverage the competitiveness. Moreover, the Knowledge

Management System (KMS) refers to the information system applied to manage organizational knowledge (Alavi & Leidner, 2001).

The massive data within the KMS comes from the process known as Knowledge Creation: "*Organizational knowledge creation involves developing new content or replacing existing content within the organization's tacit and explicit knowledge. Through social and collaborative processes as well as an individual's cognitive process (e.g. reflection), knowledge is created, share, amplified, enlarged and justified in organizational settings*" (Alavi & Leidner, 2001).

Applying to our study, we can say that the Knowledge Creation consists in learning from the past projects. The documentation generated from the Closure phase of project life cycle must be put into the KMS. Hence, a good management of KMS consists in take this data, process it and put in use for the next project that the organization is developing.

The importance of all this is not just by the perspective of data, but also the knowledge for using the tools and models for project management. A good use of this knowledge, translated through the *PMBOK Guide*®, and the constant renew of such knowledge are vital for project success and not many studies were made in the context of Knowledge Management (Chou & Yang, 2012). In other words, the Knowledge Management is related to:

I. Gathering data to Initiation and Planning phases;
II. Assessing efficacy of project management techniques;
III. Developing skills for improving the correct and efficient use of tools, models and techniques;
IV. Improving the criteria of knowledge retained into the KMS.

As a conclusion, we can say that all phases of project life cycle are subjected to the KMS. The Initiation will take into consideration the requirements and gather data to estipulate if the organization is able to do the project.

The Planning phase will extract the data from past projects and turn it to baselines of cost and schedule. Moreover, will plan all the execution of the project. The Execution by itself will apply techniques such as EVMS that are known by the organization (acquired by the Knowledge Creation process), based on data and baselines made during the Planning Phase.

At the Closure, the KMS will be fed with new data as an attempt of creating, enlarging and amplifying the knowledge that the organization already has. Beyond that, the necessity of usage of KMS comes from the simple projects to complex projects (Ahern, Leavy, & Byrne, 2014).

Finally it is worth to say that the KMS involves the entire organization. Another concept that rises from this is Knowledge Alignment. It states that for the KMS has the role of spreading the knowledge throughout the organization. The alignment of knowledge of the company is necessary because the entire company must know its limits and boundaries, in terms of skills, technology and resources (Reich, Gemino, & Sauer, 2013).

This organization-based approach of KMS and the Knowledge Alignment concept are also important when we discuss our next topic, the Project Portfolio Management. By now we can say that we had the view of a single project, from the initiation to closure, and why it is important to use store the data from past projects and use it properly on new projects. Moreover, how we can use this data.

2.2.2 Project Portfolio Management

So far we have discussed the motivation for our study in terms of a single project. Starting with the EVMS, which is a model for execution of a project, but also depends on a robust Initiation, Planning and Closure Phase. To gather data and knowledge about how to use the EVMS we discover the importance of Knowledge Management (KM), which involves the Knowledge Creation, Knowledge Alignment and Knowledge Management System.

Thus, we can finally sew up the following argument: from the data of past projects containing at the KMS we can group then into clusters. The next project to be made will be planned and the EVMS will be executed from the similarities with the cluster that it belongs, and at its closure, it will feed again the KMS.

But what about the other projects that the company is developing? How they are connected together? For answering those questions we should introduce the concept of Project Portfolio Management and finally understand the final argument of this paragraph that is related to the passage of single project view to multiple projects.

> *"Project Portfolio Management is a set of business practices that brings the world of projects into tight integration with other business operations. It brings projects into harmony with the strategies, resources, and executive oversight of the enterprise and provides structure and processes for project portfolio governance."*

(Levine, 2005)

From this definition, we can say that by doing multiple projects, the management process is not anymore just in charge of the PMO, but there are lots of new roles being played in order to achieve success. The Figure 2.16 above shows the main difference between doing a project alone, and doing it in terms of portfolio.

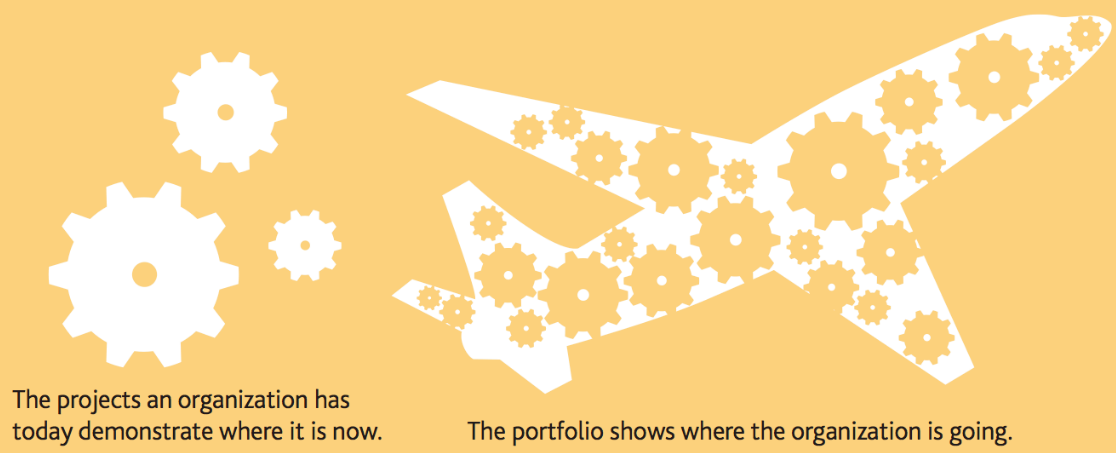
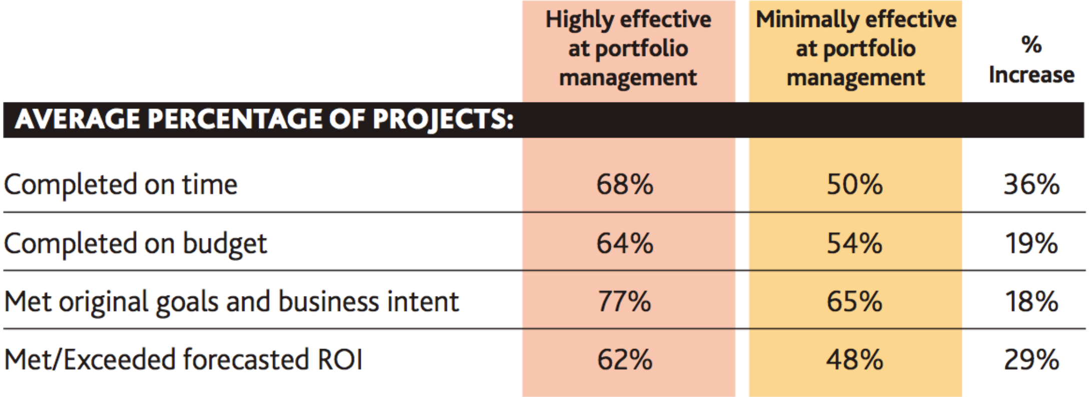

The projects an organization has today demonstrate where it is now.
The portfolio shows where the organization is going.

Figure 2.16 – Single project vs. multiple projects.
(Project Management Institute, 2014)

A perspective of a single project shows the "time now" of the company, but does not express the strategy behind, or the synergies with other projects in terms of value added to the organization. These synergies bring results to the entire mix of projects, as effective is the PPM, as shown on Figure 2.17.



AVERAGE PERCENTAGE OF PROJECTS:	Highly effective at portfolio management	Minimally effective at portfolio management	% Increase
Completed on time	68%	50%	36%
Completed on budget	64%	54%	19%
Met original goals and business intent	77%	65%	18%
Met/Exceeded forecasted ROI	62%	48%	29%

Figure 2.17 – Impacts of PPM at projects.
(Project Management Institute, 2014)

But who is in charge of PPM? As defined by Levine (2005) the PPM brings integration between the Operations Management and the Project Management. The Table 2.2 above shows the issues related to Operation and Project Management.

Table 2.2 – Operations and Projects Management issues.

Operations Management	Project Management
Strategies	Schedule/time

Objectives, goals	Project Cost
Business Performance	Project Performance
Stockholder Satisfaction	Stakeholder Satisfaction
Project Selection and Mix	Scope/change Control
Resource Availability	Resource Utilization
Cash flow, Income	Cash Usage

Source: (Levine, 2005)

The main function of PPM is to bridge, or bring together those issues: plan where the organization is going, taking the view from the PMO and OMO[15]. Moreover, the intensity of engagement and role clarity from different stakeholders of the organization are vital to PPM success, and consequently for the success of the entire organization (Beringer, Jonas, & Gemünden, 2012).

The point is, together, all the stakeholders must put some effort to bring unified the projects into a singular goal or objective. The reasons why an organization practices the PPM can be seen on Figure 2.18.

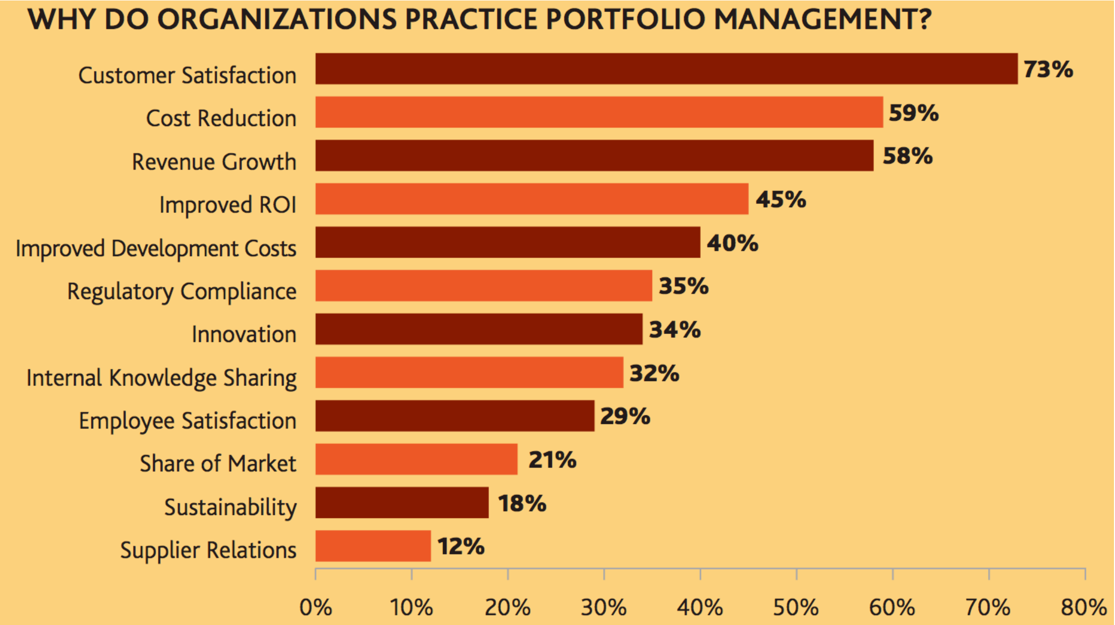


Figure 2.18 – Reasons to practice the PPM.
(Project Management Institute, 2014)

The figure above shows the objectives that an organization must aim in order to practice the PPM. For our purposes we will focus on the Cost Reduction. It represents a great driver in terms of percentage. It is worth saying that the data from Figure 2.17 and Figure 2.18

[15] PMO and OMO: related to Project Management Office and Operations Management Office. The OMO does not necessarily exist within an organization; it is usually referred as the board of directors.

were gathered from the *Pulse of Profession*, an annual report brought by Project Management Institute® with more than a thousand surveys with project managers.

Our study is focused on improving the planning phase, and consequently execution and closure of a single project, and all together means the improvement of project success. But when we talk about multiple projects, our purpose is in fact to bring cost reduction to the organization. This is possible if we can help to improve the PPM.

So, if we talk about PPM, and moreover if we talk about cost reduction brought by PPM we need to understand how it works, and how our study can support this model. The first thing to know about the PPM, once we already know its definition, is the life spam and the phases of PPM.

The Project Portfolio life spam is composed by five actions:

I. Identification of needs and opportunities

II. Selection of best combinations of projects

III. Planning and execution of the projects

IV. Bring out the outcomes

V. Realization of benefits

Notice that the actions (I) and (II) are engaged substantially with the Operations Management. The actions (III) and (IV) are engaged with the Project Management. The action (V) is engaged with both, and it is highly important because it is the main argument of coherence within the organization in terms of PPM (Levine, 2005). The Figure 2.19 bellow shows how starts the PPM and is actually what we are interested on.

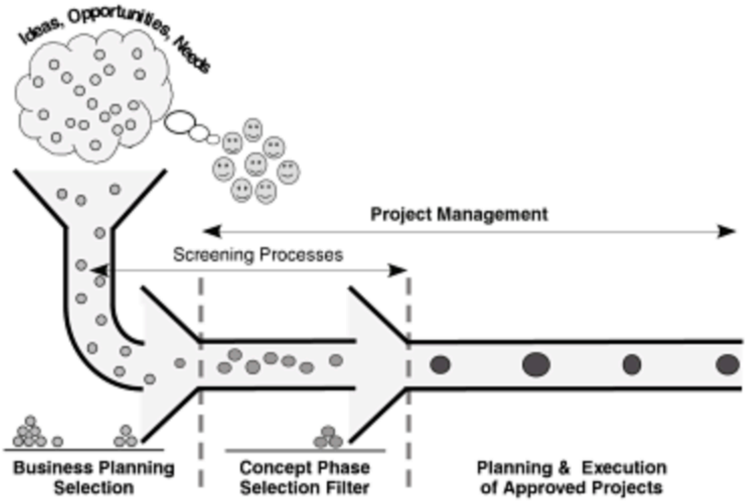

Ideas, Opportunities, Needs
Project Management
Screening Processes
Business Planning Selection
Concept Phase Selection Filter
Planning & Execution of Approved Projects

Figure 2.19 – First three steps of PPM life spam.
(Wideman, 2004)

It is possible to see that there are several projects that could be developed, but only a few will be taken to the pipeline. The process of selection is supported by the data of each project in terms of cost, schedule, resources needed and value for the organization. So it is important in this phase to have a good estimative about these variables, in order to provide a robust selection.

Moreover, we can breakdown this phase in the following structure:

- Preparation of project proposals
- Evaluation of project value and benefits
- Evaluation of risks that might modify the benefits
- Alignment of candidate projects with organization strategy
- Determine the use of resources
- Ranking projects according a set of selection criteria
    - Execution of strategic and tactical guidelines
    - Maintaining of an inventory of available resources
    - Establishment of budget buckets
    - Decision of optimum size of pipeline
    - Setting of boundaries of acceptance risk
- Selection of projects

For a proper selection of projects it is necessary a robust deployment of data about the project. So, once more, we can relate to our study, because if we can associate a new project with past project, we can have a more precise estimative of cost, schedule, risks and value add to the organization. That is directly correlated with the portfolio success, and consequently with organization success.

2.2.3 Summary

Through this paragraph we have seen the motivation for our study. We've discussed the view of a single project, and the importance along its life cycle of having a robust instrument of prediction (of cost and schedule). We talk about the EVMS, and how important it is as a model of execution of a project. We've also introduced the concept of Knowledge Management, and how the KMS feed the whole life cycle of a project and how is fed along the same lines, a continuous process of learning.

Finally we introduced the concept of Project Portfolio Management and showed the importance of managing multiple projects and the gain from the synergy between them, all of this in order to achieve success for the entire organization. To expose the summary in a pictured way, we developed the following diagram that expresses the argument present in this paragraph.

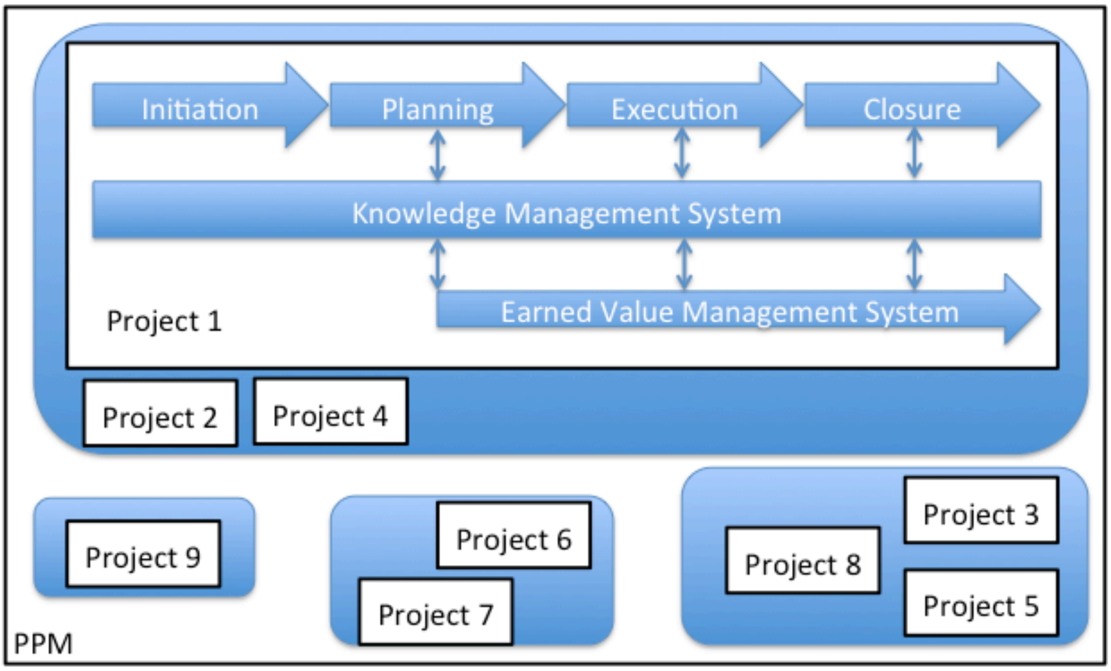


Figure 2.20 – Single and Multiple Projects views.
(Developed by author)

Note that similar project are grouped, but this is only in reference of typology of project (e.g. internal IT projects, new product development projects). This is how they compose the PPM, and this is a visual way to understand the system.

2.3 Chapter Summary

This Chapter was focused in viewing the Object of our study and then the Motivation for it.

In terms of Object, we've discussed the project itself. The paragraph was divided in seven pieces: conceptualization, variables and phases, definition of success and drivers, principles and practices to achieve success. We say that this is the object of study because we will focus on the single project, and its variables and phases. Our mission here is to develop a model able to improve the success of a project, considering the behavior of its variables.

In terms of Motivation, we've discussed two views of motivation. The first one is the view of a single project, and how important is to plan well, in order to have a good execution, and consequently achieve success. This part was express with the EVMS model, which takes into consideration what was done during the planning phase and will be used to design and manage the execution phase. All of this supported by Knowledge Management, which has the purpose of feeding the EVMS and be fed during the execution of it. Furthermore, the KMS is where we put the documentation resulted from past projects.

Still during the Motivation, we got a step forward and discussed the view of multiple projects. One project that achieves success could mean nothing for the company if it utilizes the entire organization's resources. For that purpose, we talked about the Project Portfolio Management. This model includes all projects that a company is developing and takes advantage from the synergy between them.

But why we did such literature review? The Purpose of that was to give enough qualitative background to deploy our model. Our mission is to compare projects, and group them into clusters, to help developing future projects from the database established with past performances.

Note that our mission evolves the notion of phases, variables and success (object of study). Moreover it has a huge impact on execution of EVMS, the usage of KMS and the development and management of PPM. This notions, though, sew up our argument exposed this chapter.

3 Literature Review: Multivariate Statistics and Clustering Theory

In this Chapter we are going to start presenting the mathematical tools used to build the model. The main purpose here is to develop a proper knowledge that will be the foundation of our quantitative analysis. It should be highlighted that the purpose of this chapter is not to cover all the subjects related to each point shown above. There are some main topics that will be covered, which are important for our discussion and model development.

3.1 Statistical Tools and Data Representation

In this paragraph we will expose the statistical background necessary to understand the model. The paragraph will flow along those lines: Basic Descriptive Statistics, Graphical Techniques for exposing the data ex-ante, Distance Methods and Final Comments.

The exposure of descriptive statistics is placed here in order to provide a substantial capability for us to manipulate the data of past projects, in terms of their variables. It is worth saying that every project is a multivariable event, with specifics outcomes.

In terms of graphical techniques, we will show how is possible to perceive similarities between projects before treating the data, and applying a robust statistical model (in our case the clustering theory). The graphical scheme will provide a visual support for clusters established from projects' data.

Following graphical techniques, we have distance methods. They are vital for the cluster theory. Actually, distance methods are the core argument within our model. We will show possibilities, and argue about our choice, giving pros and cons of typology of distance that we might use.

Finally, we will give some final comments about the data and statistics, and the importance of treating the relation. For example, we will discuss about independence of variables and observations, and the influence on results that a flawed data can have.

3.1.1 Basic Descriptive Statistics

In this part we will discuss three points: how to present the data, how to calculate and present the sample mean, and how to calculate and present the sample variance and covariance.

But first of all, we need to discuss briefly about the data. Once that we said we will present the sample mean, variance and covariance, it becomes inductive that we are not talking about a population. It is easy to understand this argument because we will present a group of data about projects that belong to a single company, and won't present all projects done in certain sector. That is why we refer to a sample of projects. But in fact, there are no meaningful differences, because the data calculus of variance will be corrected in order to provide a proper estimation of the population variance.

The easiest way to present an amount of data, for multivariate statistical analysis is by the use of a matrix, as shown in Table 3.1, where:

$$x_{jk} = measurement\ of\ the\ k^{th}\ variable\ on\ the\ j^{th}\ project$$

Table 3.1 – Representation of the data.

$$\mathbf{X} = \begin{array}{c|ccccccc|} & \text{Variable 1} & \text{Variable 2} & \dots & \text{Variable k} & \dots & \text{Variable p} \\ \text{Project 1} & x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ \text{Project 2} & x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{Project j} & x_{j1} & x_{j2} & \dots & x_{jk} & \dots & x_{jp} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{Project n} & x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{array}$$

Source: developed by author

Considering this representation, we will show now how to calculate the sample means. The mean is usually calculated in terms of a single variable, given a sort of observations. These values will be extremely necessaries for our model development. The mathematical way of sample mean calculus is given by:

$$\boxed{\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad\qquad k = 1, 2, \dots, p}$$

Every k-value of mean represents a subset of the full set of measurements that might have been observed. By representation of every p-variables, we have p-values of mean.

Moreover, the sample means will be represent in terms of an array that expose all values. This representation eases the mathematical procedure for the whole model development.

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Finally we will show how to calculate the sample variance and covariance. It is important to remember the concept behind. The variance within a variable is a value that shows the dispersion of data, within a range of values. Moreover, it shows the average shift of an observation from the sample mean (Costa Neto, 2002).

The covariance, in other hand express the linear association between the measurements of variables *j^th^* and *j+1^th^*. It is a very important concept for our purposes because it will show how the variables interact with each other, and will guide the calculus of a proper statistical distance between projects (Johnson & Wichern, 2007). The mathematical way of expressing the variance and covariance is:

$$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2 \qquad k = 1, 2, \dots, p$$

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \qquad i \neq k, \qquad i, k = 1, 2, \dots, p$$

In terms of visual representation, the variance and covariance will be grouped at the same array, called from now as just covariance array. Once more, this representation was chosen because it eases the mathematical handling of data for our model construction.

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

In terms of matrix manipulation, we can make the calculus of mean array and covariance array even easier. The resulted equations that we will utilize are the following:

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1} \qquad (1)$$

$$\mathbf{S}_n = \frac{1}{n}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\,\mathbf{1}'\right)\mathbf{X} \qquad (2)$$

The calculus behind the results shown can be seen on Appendix B.

3.1.2 Graphical Techniques

Plots are important, but frequently neglected, aids in data analysis (Johnson & Wichern, 2007). Because of that, we will discuss now some techniques to have a visual approach of the data. Basically there are three different techniques that we can talk about, in terms of our model: the scatter plot, stars representation and Chernoff faces.

3.1.2.1 Scatter Plot

The first graphical technique is related by scatter plot (SP). Basically, a SP is a visual representation that takes into consideration the relation between two or three variables. The observations are put into the graphic form, in a dispersive way, and the picture resulted is examined. Figure 3.1 can express an example of SP in two or three dimensions.

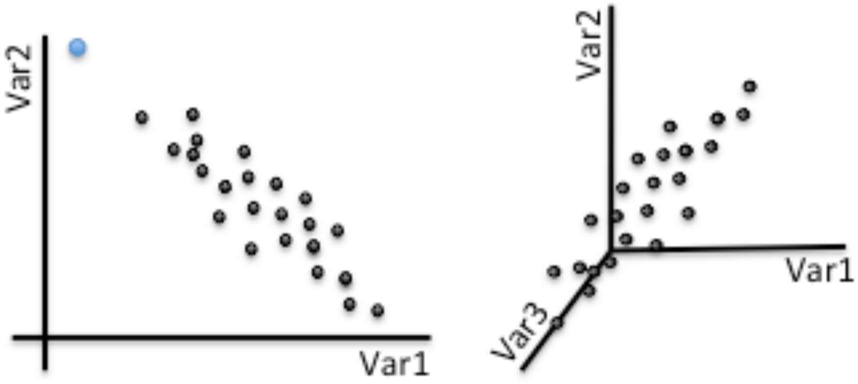


Figure 3.1 – Scatter Plot for two and three dimensions. (Developed by the author)

For our purposes, the scatter plot would be a good tool for perceiving visually if the variables of our model have some correlation. For example, it is possible to see that the bi-dimensional SP in Figure 3.1 expresses the negative correlation between two variables.

Moreover, the scatter plot is a powerful tool for perceiving outliers within our data pool. Again, our example shows a possible outlier within our bi-dimensional scatter plot, highlighted by blue color.

3.1.2.2 Star Representation

The second graphical technique that we would like to present is the star representation. It belongs to a classification of graphical technique called pictorial representation. These pictures are valuable aids in understanding data and often prevent many false starts and subsequent inferential problems (Johnson & Wichern, 2007).

The star representation itself can only be used when the data consists in nonnegative observations and the number of variables is such that $p \geq 3$.

The construction is very easy and the results are very visual. Every axes represent a variable, and the center represent the value of zero. If the variable takes a great value, the star will point for it.

It is also worth saying that each star represents an observation. So, this pictorial representation is better used for a small sample, in our case it fits well because we have a small range of projects to analyze and group. Moreover, for the construction of a star, a good practice would be standardizing the values in every observation. By doing this, the axes would be easily read.

Figure 3.2 exposes a good example for star representation.

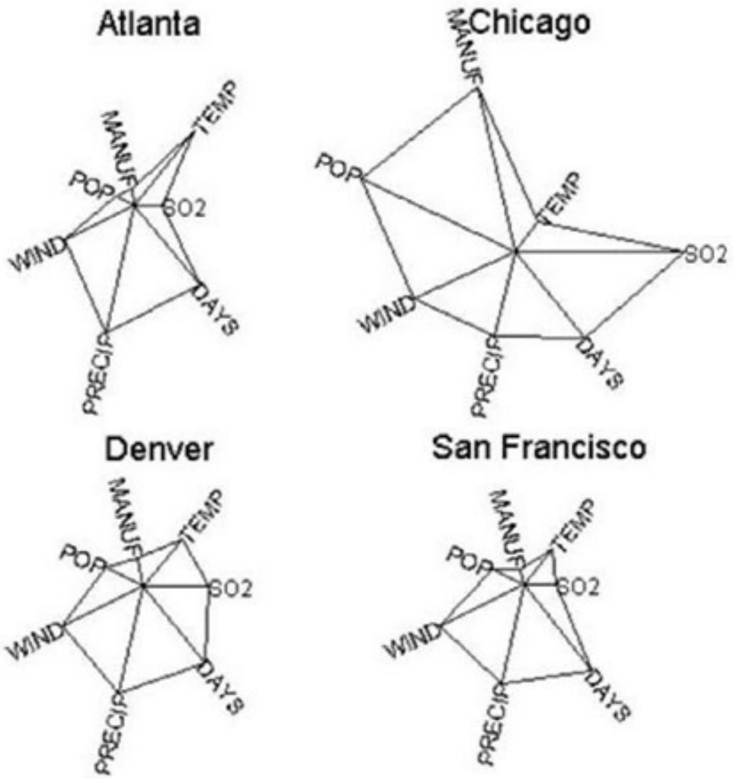


Figure 3.2 – Example of use for Star Representation.
(Source: http://what-when-how.com/statistics/skewness-to-systematic-review-statistics)

3.1.2.3 Chernoff Faces

The third graphical technique is also a pictorial representation, but with some differences in terms of usage that the star representation. It was originally design in 1973, by Herman Chernoff, a mathematician and physicist graduated at MIT and currently professor at Harvard University.

The purpose of Chernoff Faces was initially to aid the visual aspects at multivariate analysis and were designed to analyze at maximum eight variables at the same time (Johnson & Wichern, 2007).. It was a common agreement within the academic field that people are able to process information translated by faces with more speed and accuracy (Sivagnanasundaram, Chaparro, & Palmer, 2013).

The agreement that people recognize more easily differences into faces is not actually true. In fact, we are not able to group different observations by perceiving similarities at this kind of representation (Sivagnanasundaram, Chaparro, & Palmer, 2013). But still, the Chernoff Faces (CF) has a practical use: it helps to understand changes in time of a set of variables. In other words, if the observations have a correlation because of time, it becomes visual by using CF.

Figure 3.3 exposes nine standard for Chernoff Faces.

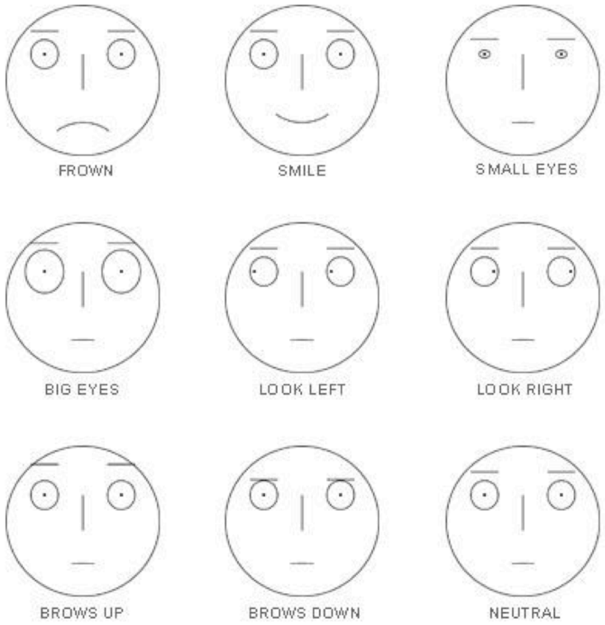


Figure 3.3 – Nine standards for Chernoff Faces.
(Sivagnanasundaram, Chaparro, & Palmer, 2013)

3.1.3 Distance Methods

The Distance Methods are the core of cluster theory. To produce a group structure from any sort of data we need first have a measure of "closeness", or "similarity" between the observations. The main issue here is to calculate the distance between observations, in our case, between projects.

One factor that may difficult the problem of measuring distance is the subjectivity that the calculus brings. It is important to include some thoughts such as "nature of variables", "scale of measurement" and "subject matter knowledge" (Johnson & Wichern, 2007). For our purpose, the more important here is to scale the measurements in order to compare different variables within the same distance calculus. We can see a way to scale the measurements in the Appendix C.

We will present here three ways of measuring distances: Euclidean Distance, Statistical Distance and Minkowski Distance. They have the same general idea, but with some subtle differences.

The Euclidean Distance is the most famous, and most common for clustering practices. It is also known as the straight-line distance, because measure takes into consideration the minimum between two p-dimensional observations (a straight-line) **x’** = [x$_1$, x$_2$, …, x$_p$] and **y’** = [y$_1$, y$_2$, …, y$_p$]. The Mathematical formulation is:

$$\boldsymbol{d_E}\ (\mathbf{x}, \mathbf{y}) = \sqrt{(\boldsymbol{x_1} - \boldsymbol{y_1})^2 + (\boldsymbol{x_2} - \boldsymbol{y_2})^2 + \cdots + (\boldsymbol{x_p} - \boldsymbol{y_p})^2} = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} \quad (3)$$

The Statistical Distance takes into consideration the Euclidean Distance, adjusted by covariance matrix (relations between variables). The mathematical formulation is:

$$\boldsymbol{d_S}\ (\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{y})} \quad (4)$$

Finally, the Minkowski Distance that incorporates to the Euclidean Distance the weight given to larger and smaller differences between variables at two p-dimensional observations. The Mathematical formulation is:

$$\boldsymbol{d_M}\ (\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} |\boldsymbol{x_i} - \boldsymbol{y_i}|^{\boldsymbol{m}}\right]^{\frac{1}{\boldsymbol{m}}} \quad (5)$$

The distance method that we will utilize in our model will depend on the data we have. If we notice that we have no relation between variables, the Statistical Distance becomes useless. If we think that we want to give great importance on huge differences between observations, the Minkowski Distance takes *m = 1* or *m = 2* (Euclidean Distance).

Finally, we can say that for purposes of calculation, we will use the pure distance methods. But to present the results and conclusion we will use the concept of similarity. The similarity between two p-dimensional observations is the reverse relation of distance. Mathematically speaking the similarity is:

$$\boldsymbol{s_{ik}} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{d_{ik}}} \quad (6)$$

Where, $0 < s_{ik} \leq 1$. For this reason we can have a standardized view about the closeness between two p-dimensional observations.

3.1.4 Final Comments

In this paragraph we have seen some basic descriptive statistics, graphic representation and distance methods. However, we need to expose some comments about the applicability of this literature, in order to deploy a proper model around the cluster theory.

The first comment we should expose is related to the independency of the model variables. If the variables were, indeed, independent the covariance matrix would be a diagonal matrix[16]. For that reason, the Euclidean or Minkowski Distance would be a better choice for calculating the similarities.

The second comment is about the independency of the observations (in our case, the projects). The measurements from different observations must be independents in order to guarantee the existence of inverse for the covariance matrix. This is very important if we chose to utilize the Statistical Distance.

Furthermore, if the sample size is minor than the number of variables, the covariance matrix does not accept inverse. Once more, it could compromise the functionality of our model.

3.2 Clustering Theory

In the previous paragraph we discussed about the statistical treatment for the data, before its utilization on Clustering Analysis, as well we discussed the different distance methods. The pros and cons of each distance method, and how we will determine which one is proper for our model.

This paragraph will be devoted to the Clustering Analysis. We will understand its conceptualization, the principles of utilization, the requirements for data treatment ex-ante and the major class of algorithms that are in vogue for Clustering Theory.

In sequence we will discuss the robustness of the algorithms that we've seen. Once we've applied one of the algorithms, we must check if the data is well partitioned and the results are consistent. This part is crucial for our study to move on for the outcome analysis.

Finally, we will discuss the Discriminant Analysis, which is a model for classification of observations, once we have the possible groups established. This analysis will be vital for our model development, as being part of the final arrangement for a project to be properly estimated.

3.2.1 The Cluster and the Clustering Analysis

"Even though there is an increasing interest in the use of clustering methods in pattern recognition, image processing and information retrieval, clustering has a rich history in other disciplines such as biology, psychiatry, psychology, archaeology, geology, geography, and marketing" (Jain, Murty, & Flynn, Data Clustering: a Review, 1999).

This is a good way for introducing the Clustering Analysis. The range of application gives us the opportunity to apply it in our model. Note that, in terms of the citation, we will

[16] Diagonal matrix refers to a matrix that has values in the main diagonal and all the rest is equal zero.

explore the side of “pattern recognition”. Our first goal is to identify patterns along the projects data, and then group those patterns into clusters. We will also utilize the clustering theory capability of “data retrieval”. Once we have determined the patterns, and consequently determined the clusters, we will be able to recall those characteristics to compare new projects with past projects.

We have, then, exposed the usage for the Clustering Analysis. But, what really stands for “cluster”? In other words, what exactly does “cluster analysis” mean? Several authors determine the concept of cluster and clustering analysis, for example:

> *“[Cluster]… is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity.”*
> (Jain, Murty, & Flynn, 1999)

> *“Cluster analysis seeks to reduce the dimensionality of data by grouping objects into a small number of groups, or clusters, whose members are more similar to each other than they are to objects in other clusters.”*
> (Manton, Lowrimore, Yashin, & Kovtun, 2005)

> *“La cluster analysis è una tecnica di analisi multivariata attraverso la quale è possibile raggruppare le unità statistiche, in modo da minimizzare la ‘lontananza logica’ interna a ciascun gruppo e di massimizzare quella tra i gruppi.”*[17]

Those definitions make possible to understand some vital issues about the Clustering Theory. These issues will guide the usage of this type of analysis into our model. In summary, all that we need to say about Clustering Theory is the following:

a. Cluster is the group of objects with similar set of values, among a set of variables;
b. Clustering Analysis is the procedure of grouping a set of observations into clusters with the purpose of describe and study different groups of patterns among the range of data.

[17] Fonte: http://host.uniroma3.it/facolta/economia/db/materiali/insegnamenti/185_903.pdf

Moreover, a typical procedure of clustering analysis is based on the following standard steps (Jain & Dubes, Algorithms for Clustering Data, 1988):

I. How to represent the data;

II. How to define the similarity between observations;

III. How to apply the correct algorithm of data partition;

IV. How to analyze the results of data partition;

V. How to study the clusters obtained.

In the previous paragraph we have devoted some space for discussing Data Representation and Distance Methods. Our task now is to put it all together, and show how to work with the steps (III), (IV) and (V).

3.2.2 Data Requirements

Before discussing the Algorithms and Analysis of clusters, we should recall some concepts that we've seen on the last paragraph regarding Data Representation and Distance Methods. The importance of doing this is because the right choice of clustering algorithm, and consequently a robust cluster structure depends on some data treatment. Moreover, the analysis of data partition that result from the algorithm and the study of the clusters could be compromised if we don't have a decent pre-analysis of the data set.

The first point, though, is about the importance of graphical techniques in pre-understanding the sort of data we are dealing with. For example, if we have an amount of data, represented by n-observations in two-dimensional variables we can do a scatter plot to have a first idea about data behavior. The Figure 3.4 shows an example of data behavior that a scatter plot can capture.

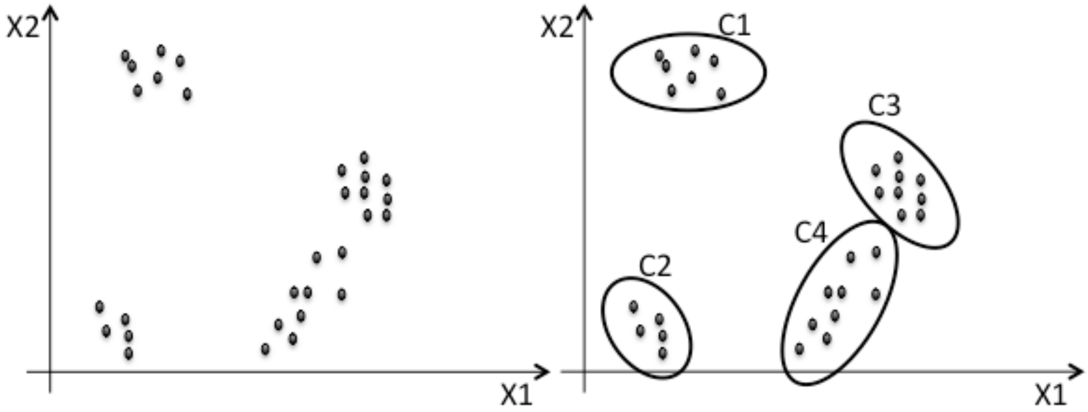


Figure 3.4 – Data Representation in scatter plot for pre-analyze of data.
(Developed by the author)

Notice that in this pre-analyze we captured the data behavior, and created an idea that maybe there are four patterns in this population, or in other words, there are four clusters that represent this data. This kind of analysis is useful especially when we do not have a clear idea about how many clusters we should compute in order to divide the data.

The second point that deserves our attention is related to the Distance Methods. We have exposed in the previous paragraph three different methods for perceiving similarities within a data set: Euclidean Distance, Statistical Distance and Minkowski Distance. Each of these methods has the pros and cons and it depends on us leveraging our model with the proper choice of distance method.

There are some clues for choosing well. The first one is related with the covariance matrix. We can derive from that the Correlation Matrix, which gives us the information about how correlated are the variables in confront with each other. The mathematical way of calculating the correlation between variables is given by:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

In terms of matrix operation, we can derive the Correlation Matrix as being:

$$\mathbf{R} = \mathbf{D}^{-1/2}\,\mathbf{S}\,\mathbf{D}^{1/2} \qquad (7)$$

Where D is *sample standard deviation matrix*, defined by all variables standard deviations in the main diagonal, and all other values equal zero.

But what is the importance of the correlation matrix? The main information that it could give us is about the independence of variables. If the correlation between is equal zero, it means that one variable does not provide any information about the other one. In this case, the Statistical Distance loses it strength, and we should utilize the Euclidean Distance or the Minkwoski Distance.

For choosing between these two methods, we should take a look on context. We should analyze what typology of data it is, and how the data was gathered. The main reason of doing so is because the Minkowski Distance has the capability of weakening high distances between observations. If our data is difficult to gather, and the variances are high, the Minkowski Distance should be applied because it can diminish the bias.

It is worth saying that the standard is to utilize the Euclidean Distance because it is the middle of the three methods, and bring together the pros and cons of each. But if we have a special sort of data, we must pre-analyze it in order to choose correctly the distance method and consequently have a robust model.

3.2.3 Clustering Algorithms

The main objective of any clustering algorithm is to minimize the distances (or the sum of all distances between observations). We've already discussed the techniques for measuring distances, and already know the concept of similarity. So basically, the algorithms that we will see will follow the objective function:

$$\min f = \sum Distances$$

Will discuss two different classes of methods, each with the pros and cons: the hierarchical and nonhierarchical clustering methods. Each class is used with a certain objective, and with a certain data set. The correct choice of algorithm for our study will depend on our pre analysis of input and desirable outputs.

3.2.3.1 Hierarchical Clustering Methods (HCM)

"Hierarchical Clustering Techniques proceed by either a series of successive mergers or a series of successive divisions." (Johnson & Wichern, 2007). This is the core of hierarchical clustering methods. They are not tied to a pre-defined number of clusters.

There are two main algorithms within the HCM that translate well the core issue: the single linkage algorithm and the complete linkage algorithm. They technically arrive at the same result, but take different premises. They also belong to a subcategory of HCM: Agglomerative Hierarchical Methods. The main steps to follow in order to apply these algorithms are (Johnson & Wichern, 2007):

1. Starts with N clusters, each containing a single entity and an *N x N* symmetric matrix of distances D = $\{d_{ik}\}$.
2. Search the distance matrix for the nearest pair of clusters. Compute that distance between clusters U and V as being d_{uv}.
3. Merge the clusters U and V and label it (UV).
4. Update the distance matrix by deleting the rows and columns corresponding U and V and adding the row and column corresponding the cluster (UV)
5. Repeat steps 2, 3 and 4 a total of *N – 1* times. Record for each merge the entities merged and the level (distance) at which the merge took place.

To represent an Agglomerative Methods we utilize a dendogram. The Figure 3.5 represents an example of dendogram.

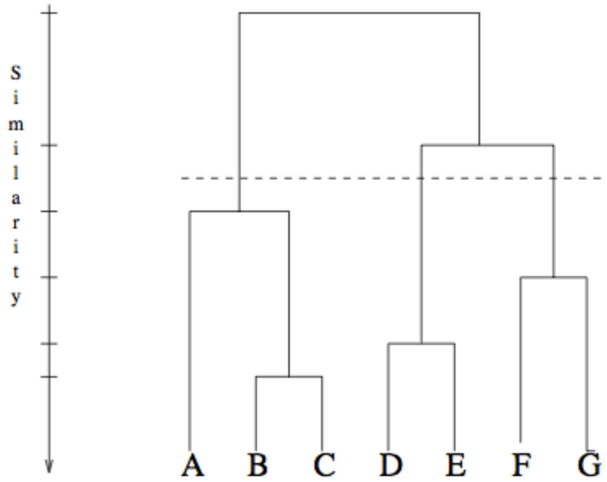


Figure 3.5 – Example of a Dendogram Representation
(Jain, Murty, & Flynn, 1999)

Just notice that the number of clusters that result from an Agglomerative Method depends on the level of similarity that we are looking for. Hence, for the purpose of our study, these methods can be useful, if we do not know for sure in how many clusters we want to divide our dataset.

3.2.3.1.1 Single Linkage Algorithm

The mechanism of the Single Linkage Algorithm follows basically the steps of any Agglomerative Method. The difference here is that for step 4 we will recalculate the distance between the cluster (UV) and a generic observation W by:

$$d_{(UV)W} = min\,\{d_{UW}, d_{VW}\}$$

3.2.3.1.2 Complete Linkage Algorithm

The mechanism of the Complete Linkage Algorithm also follows the steps of any Agglomerative Method. The difference here is that for step 4 we will recalculate the distance between the cluster (UV) and a generic observation W by:

$$d_{(UV)W} = max\,\{d_{UW}, d_{VW}\}$$

3.2.3.1.3 Difference between Single and Complete Linkage

The Figure 3.6 can graphically express the difference between the two Agglomerative methods that we just have presented.

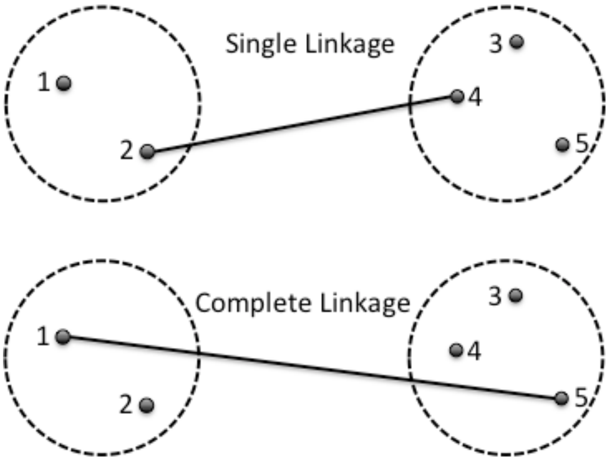


Figure 3.6 – Graphical interpretation of Single and Complete Linkage.
(Johnson & Wichern, 2007)

Notice that in the Single Linkage the distance between clusters in translated by the distance d_{24} (minimal distance). The case of Complete Linkage, the distance between clusters is expressed by the distance d_{15} (maximal distance).

Moreover, it is worth saying that the intermediate results of these methods, expressed by the time when observations are grouped in a moderate number of clusters, are the real point of interest. The initial output (each observation is considered one cluster) and the final output (all observations compound on cluster) do not carry much information (Johnson & Wichern, 2007).

3.2.3.1 Nonhierarchical Clustering Methods

Nonhierarchical Clustering Methods (NHCM), also known as Partitional Clustering Algorithms, is designed to group items into a collection of K clusters. Differently of HCM, these methods have a pre-established number of clusters, or this number is part of the procedure. In other words, we can't escape from the output of K clusters. To mitigate this problem, the NHCM are usually run several times in order to produce a robust outcome.

3.2.3.1.1 K-means Algorithm

The most popular NHCM is the k-means algorithm, and will be the one that we will study in order to develop our model. There are some preliminary concepts to understand the algorithm (Hartigan, 1975):

- There are a determined number of clusters K;
- Let L(i) be the Lth cluster that contain the ith observation;
- Let N(L) be the number of observations that the Lth contain;
- Let B(L, j) be the mean of the jth variable over the cases in the Lth cluster;
- Let D(i, L) be the distance between the ith case and the Lth cluster.

The objective of this algorithm is to minimize the error of the partition. Mathematically speaking:

$$\min error = \sum_{i=1}^{n} D[i, L(i)]^2$$

The steps of this algorithm is given by:

7. Assign every observation for one of the clusters.
8. Compute the clusters means B(L, j) and the initial *error*.
9. For the first case, compute for every cluster L

$$ie = \frac{N(L)D(1, L)^2}{N(L) + 1} - \frac{N[L(1)]D[1, L(1)]^2}{N[L(1)] - 1}$$

Which is the increase in *error* in transferring the first case from cluster L(1) to cluster L. If *ie* is negative, transfer the first case to the most negative *ie*.

10. Adjust cluster means and the *error*.
11. Repeat steps 3 and 4 for every case.
12. If there is no movement to an observation to a cluster to another, stop. Otherwise start at step 3 again.

The main issue of this algorithm is the first partition of the data. The question that it raises is: how to allocate an observation to a determined cluster? To answer this question, we can just simply utilize a graphical representation of the data and allocate the observations accordingly to the patterns we see by eyes.

3.2.1 Algorithm Robustness

In order to achieve a great level of robustness, we will implement an algorithm that mixes the Hierarchical and Non Hierarchical Clustering Methods. At the first level of analysis we can implement the single linkage algorithm in order to map the possible clusters to be formed.

The second level will be applying the k-means algorithm based on the clusters that arose from the first level.

Moreover, we need to test if the result from this double-leveled algorithm is robust. For doing this, we can run the algorithms several times, changing the initial partition. The results could be analyzed in terms of basic statistics. For testing the robustness we can apply a ratio based on:

$$Robustness = \frac{Max\ cluster\ appearance}{Total\ of\ algarithm\ runs}$$

We can say that if $Robustness \geq 90\%$ our algorithm is robust, and we can accept the outcomes.

3.2.1 Discriminant Analysis

After we have established the cluster from the algorithms we just have seen, we need to find a way to classify a new object to a given cluster. This analysis requires the study of the clusters we have obtained and an analysis of membership for new observations.

In theory, the procedures for classification of more than 2 groups (clusters) have not been fully investigated, and a there is a lot more to be studied. If we talk about non-normal distributions the proprieties developed for a two groups classification cannot be generalized for more groups (Johnson & Wichern, 2007).

However, we can determine if an observation belong to a certain group, with a certain probability. The analysis that we will do here will take into consideration the distance between the new observation and the cluster mean array.

If we consider that the cluster population is normally distributed, we can say that with a certain degree of certainty that the distance between the cluster means and the observation is such that:

$$(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi_p^2(\boldsymbol{\alpha}) \qquad (8)$$

Where Σ^{-1} is the inverse covariance matrix of the population, and $\chi_p^2(\alpha)$ is the chi-square distribution with p degrees of freedom (correspond to p-variables) (Johnson & Wichern, 2007).

By doing this, we can say that with a certain probability the new observation belongs to the cluster analyzed. Graphically, we have something like the Figure 3.7.

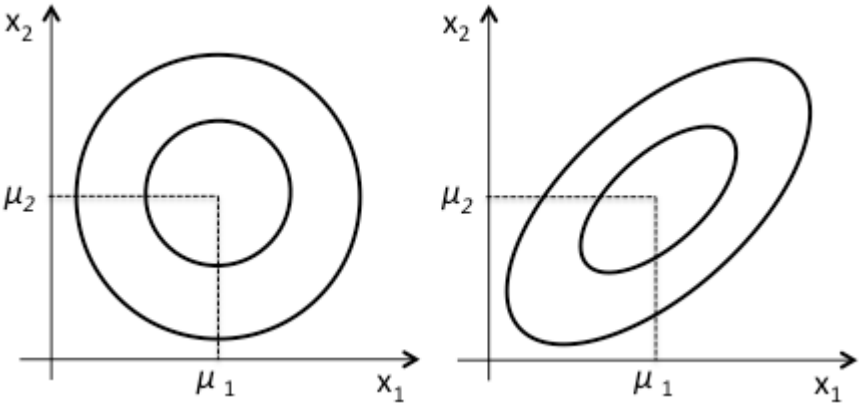


Figure 3.7 – Graphical interpretation of membership analysis.
(Johnson & Wichern, 2007)

Just note that for the first case we have that Σ^{-1} is a diagonal matrix, or in other words there is no correlation between the variables. Furthermore, both cases represent the low probability (bigger radius) and high probability (smaller radius) of membership for an observation.

3.3 Chapter Summary

In this chapter we have discussed all the mathematical support for our study. The main importance for this chapter is to have a robust background to develop a well-structured model.

The five standard steps for clustering procedures can be seen as a good summary for this entire chapter. The rest of the chapter is here to support the 5-steps procedure. All the data analysis, graphical techniques and algorithms that we have seen will be valuable tools for developing our own model.

Hence, the next chapter will try to take all the mathematical background, and mix with the project management theory in order to provide a robust model and consistent results.

4 Objective, Hypothesis and Model Development

The core of this chapter is our model deployment. But first, we need to recall what is going to be our objective, and what hypothesis we will try to test in order to pursue our goal. Then we will be able deploy a model for testing the hypothesis and finally reach our objectives.

4.1 The Objective

Now that we have shown all the theory that is the foundation for our model, it is time to devote a whole chapter developing it. But first we need to understand well what is our objective, and how our model will achieve this objective.

We commented in the first chapter that our main objective is: perceive similarities throughout a range of projects and put these projects into groups (clusters) accordingly those similarities in order to facilitate the learning process and give a better information about forecasting future projects.

Hence, our model will focus on the application of clustering theory, as we saw in the last chapter. The clustering theory will group the projects and consequentially will estipulate typologies of projects. To be clearer, once we have our data about several projects, we can aggregate them into several groups with similar characteristics. Those groups are different from each other but contain projects that are similar. So, we can imply that each group is a typology of project, with some characteristics in common.

Through this new point of interest, we can change a little our objective: "perceive similarities throughout a range of projects and put these projects into groups (clusters) accordingly those similarities, study the characteristics of each cluster and classify as a typology of project. All of this in order to facilitate the learning process and give a better information about forecasting future projects".

The last comment is related to the second part of our objective that is "facilitate learning process" and "help forecasting variables for future projects". Indeed, that was suppose to be a continuation for our model, but as we told on the second chapter, about the motivation, we already know how important it is to utilize the data from pas projects into future projects. So, it is not our scope studying this relation between outcomes of past and future projects.

Our scope is really to understand the variables of a project, analyze their values, combine projects based on these values and verify typologies of projects. And the motivation for that is the gain in the learning process, and planning and forecasting the values of these variables for future projects. Hence, notice that we managed to disaggregate the scope of this study from the motivation of it. By doing this we can focus on the substance of the model, which is the cluster theory and left aside the discussion of impact on the learning process, which is a totally different analysis.

In summary, we can rewrite our objective, and propose it as a final argument for our model development:

The objective of this study is: (I) perceive similarities throughout a range of projects, (II) put these projects into groups (clusters) accordingly those similarities, (III) study the characteristics of each cluster and (IV) classify as a typology of project.

4.2 The Hypothesis

Once we have sharpened the discussion of our objective for this study, we have to establish some hypotheses to verify if we can or not achieve our objectives. The main purpose here is to put some boundaries around our model and see if we are going through the right path.

Moreover, the hypotheses will guide the discussion and model development, in the sense that we can made several tests along the model's steps. The hypotheses will serve, though, as milestones of model application.

There will be three hypotheses in total that we will verify:

I. The variables of projects have some correlation.

II. The clusters formed with planned values of projects variables are the same for the projects after conclusion, given a certain dataset.

III. The clusters obtained are statistically different from each other.

IV. Every future project must belong to one, and just one cluster.

4.2.1 Hypothesis One

Let first state the Hypothesis One, in order to have a clear notion of what does it means, how we will test it, and most important, what is its relevance for the whole model. The Hypothesis One says:

H1: There is some degree of correlation between the variables of a project, for both planned and actual values.

This first hypothesis estate that a given data set, ruled by variables that differ in value from observation to observation have a specific statistic propriety: the variables are correlated if analyzed in pairs. Moreover, given that our study is focused in projects variables, and a whole data set of observations, we can estate that the correlation remains from planned values and actual values.

The importance of testing this hypothesis results in understanding our data set mathematically, and understanding the implications for our model development of the first critical decision in our model: the choice of distance method to apply. If we have that this hypothesis is true, for example, we have a concise argument for using the Statistical Distance, instead of Euclidean Distance.

The implications of our distance method's choice affect the model as a total because the way we calculate the distances, which measure the degree of similarity between observations, drives the correct formation of cluster. In other words, the correct choice of distance method is a critical factor for success within our model.

For testing this hypothesis is very simple. We first need to calculate the covariance matrix S, as the equation $\mathbf{S}_{\boldsymbol{n}} = \frac{\mathbf{1}}{\boldsymbol{n}} \mathbf{X}' \left(\mathbf{I} - \frac{\mathbf{1}}{\boldsymbol{n}} \mathbf{1}\,\mathbf{1}' \right) \mathbf{X}$ (2) implies. Then, we use this matrix for calculating the correlation matrix, as given by the equation $\mathbf{R} = \mathbf{D}^{-1/2}\,\mathbf{S}\,\mathbf{D}^{1/2}$ (7).

Once we have the correlation matrix, we need to analyze each of the correlation values. They should be comprised between -1 and 1. If the value is zero, it means that there is no correlation between variables. But, as we know, statistically speaking, we have to consider a certain variation upon the data. So, our analysis will be based on probabilistic distribution, in other words we will consider values that within a threshold to be equal zero. The threshold will be based on Pearson Product-moment Correlation[18], and we will test a hypothesis of correlation of the population equal zero. For example, if our sample correlation is 0,1 and the number of observations is 20, then with 5% of significance level we can say that our correlation is equal zero.

The values of threshold vary accordingly the chi-square distribution, but we will utilize a proxy because we do not know for sure if our population (from which derives our

[18] The Pearson assume normality within the population (which can be consider by the central limit theorem) and apply the t-student distribution by comparing the value $T(r) = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with t-distribution with $n - 2$ degrees of freedom.

data set) is normally distributed, and our focus is not analyzing the population and its characteristics, but analyzing their internal groups.

4.2.2 Hypothesis Two

The Hypothesis Two also should be study, in order to have a clear notion of what does it means, how we will test it, and most important, what is its relevance for the whole model. The Hypothesis Two says:

H2: The clusters that arise from the planned values of project variables are the same that arise from the actual values of project variables.

The second hypothesis implies that the groups that are formed from the data of projects in planning stage are the same that are formed from the data of the same projects, but in closure stage (once that all variables have their actual values computed, accordingly how the project was developed).

The impact of this hypothesis is tremendous on our final discussion. Once that we have analyzed the data, and applied the correct clustering algorithm, we must be sure that the groups that derive from the planned values of each project variables are the same that derive from the actual values. If they are not the same, we are not able to make the correlation between what is planned and what really happens within a certain project development. That implication results in an uncertainty about the process of learning, because similar project in planning stages don’t follow similar paths, and it is impossible to predict the behavior of a future projects based on past projects.

To test this hypothesis is very simple. We just need to form the clusters of planned observations and actual observations and compare both. If the clusters are the same, then we can say that H2 is true. If the clusters are not the same, then H2 is false.

When we are verifying this hypothesis, we just need to be careful if there is any outlier. If, for example, the clusters are the same, in planned and actual values, by the difference of just one project, before stating that H2 is false, we need to analyze if this single project is not an outlier. We need to check its behavior, and see if any event occurred during its development.

4.2.3 Hypothesis Three

Now, let's state the third hypothesis. We need also to understand well the phrase that compounds this hypothesis and discuss its importance within our model development. Finally, we need to know how we will verify if this hypothesis is true or false. The Hypothesis Three says:

H3: The groups obtained from the clustering algorithm are statistically different from each other.

The third hypothesis implies that every group formed from any clustering algorithm should have different characteristics. The other way to say that is analyzing the behavior of the data set variables for each group, and sees if they can be considered different from a statistical perspective.

The impact of this hypothesis is also big. If we can't affirm that the clusters are different, it does not make sense to group the project in the first place. Moreover, we can't analyze the behavior of different groups of projects because they would be technically the same.

In summary, this hypothesis serves as a stage gate between two events: the final application of clusters algorithms and the initiation for analyzing the results. In other words, the immediate point that the part (II) of our objective becomes the part (III).

For testing the third hypothesis is a little bit more complicated than the hypothesis one and two. For this case we will need to utilize a method for comparing several multivariate population means, called One-way MANOVA. The assumptions for this analysis and the structure of the algorithm could be seen at Appendix D. The MANOVA give us a robust statistical analysis of whether we are dealing with a whole population or separate sub-populations within our data set. In our case, H3 tests if every cluster is a different sub-population, and if H3 is true we can move forward with our study.

4.2.4 Hypothesis Four

Finally we can discuss the last hypothesis. Again, let's state this hypothesis and verify its impact on our study. We also should mention how we will test it, and continue our path for achieving the objective. So, the Hypothesis Four says:

H4: Every future project should belong to just one cluster, accordingly its planned value for each project variable.

The fourth hypothesis follows up the third hypothesis and means that, once we have established the clusters, and verified that they are different, we should now address new observations to the right cluster. And this new observation should belong to just one cluster, because we want to analyze the characteristics of it and predict its behavior.

The impact of H4 into our study is strictly related to Discriminant Analysis. We want to address a new project to a singular cluster and facilitate the planning phase. But to do so, we must guarantee the unity of this observation in front of all population and possible clusters.

In summary, the H4 is the final part of our objective (IV), and if we can verify that H4 is true, we can estate that our model works because it would be able to label a new project as being part of a singular cluster with singular characteristics.

For testing this hypothesis we will use the discriminant analysis algorithm described on the last chapter for every cluster, and see where the new observation fits. Because we are talking about a probabilistic environment, we know that we have some uncertainty. The discriminant analysis will give us the most probable fit for an observation, but this fit must be unique for every new project.

If H4 is false, though, we will need to reconfigure the clusters in order to have groups with distinguished characteristics and behaviors. In summary, we can notice that the H4 is the closure on our mathematical model, and a stage gate for starting a subject analysis.

4.3 Model Development

Now it is time to discuss our model. We will focus on the steps that will guide our study towards the objective. We should also remember that we have stated four hypotheses that will serve as stage gates during the model development and evolution.

The most important about this topic is to bring all we have seen so far together. The first part, when we discussed about the project conceptualization, the phases and variables will serve as a theoretic background. The part that we have discussed about EVMS, Knowledge Management and Portfolio Management will serve as our motivation, and the focus for our objective. The part we have discussed about multivariate statistics and clustering theory will serve as tools for modeling mathematically our model. Finally, as we said, the path for our model will be guided by our objective and the boundaries will be represented by the hypotheses we have stated.

We will divide our model into four main phases, which are: Data Treatment, Similarities, Clustering and Analysis. Each phase will be divided in some activities that will represent the model as a total. To be clearer, we can show the Figure 4.1 that expresses the structure of our model.

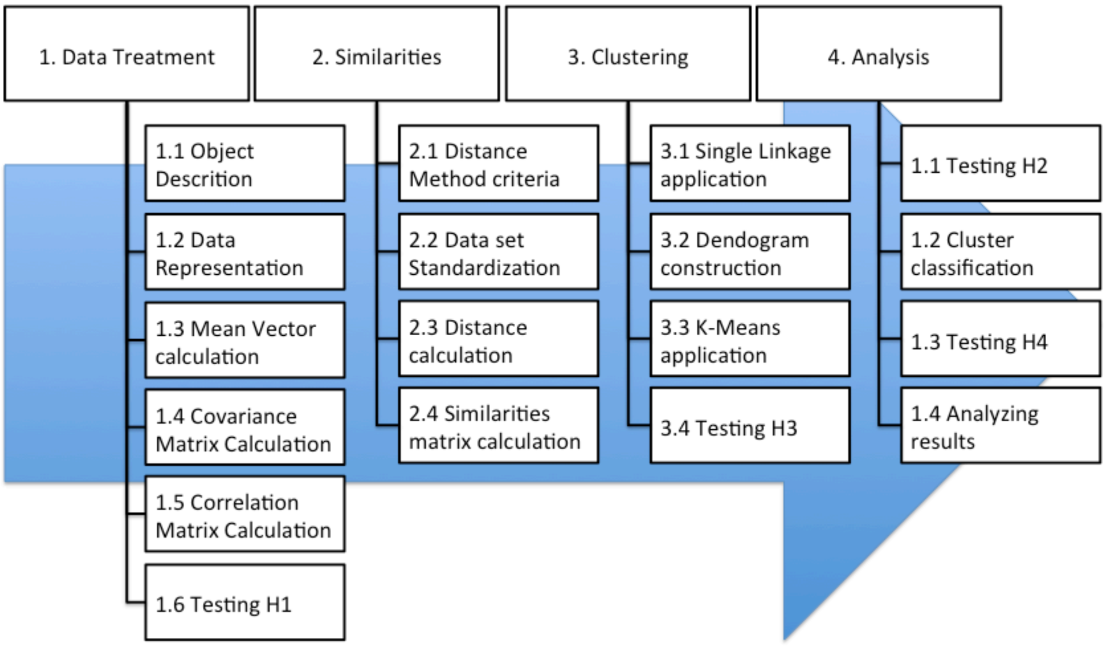

1. Data Treatment
2. Similarities
3. Clustering
4. Analysis
1.1 Object Descrition
1.2 Data Representation
1.3 Mean Vector calculation
1.4 Covariance Matrix Calculation
1.5 Correlation Matrix Calculation
1.6 Testing H1
2.1 Distance Method criteria
2.2 Data set Standardization
2.3 Distance calculation
2.4 Similarities matrix calculation
3.1 Single Linkage application
3.2 Dendogram construction
3.3 K-Means application
3.4 Testing H3
1.1 Testing H2
1.2 Cluster classification
1.3 Testing H4
1.4 Analyzing results

Figure 4.1 – Model structure.
(Developed by the author)

We should now clarify what does every phase means and why they are important for the model development. Moreover we should explain each activity that compound each phase in order to make the model comprehensible and more susceptible to argumentation for improvements of this study.

4.3.1 Data Treatment

The first step for our model is related to data treatment. Once we have a specific amount of data, we need to know what we are dealing with. We need to understand our data set qualitatively and quantitatively.

4.3.1.1 Object Description

To make this model more general, we can apply a sort of checklist to understand qualitatively our environment and our data set. The questioning should be:

- What is our object?
- What are the variables that represent this object?
- How many observations do we have?

4.3.1.2 Data Representation

Considering now that we know what is this all amount of data about, we can start doing some representation of the data, which represents the activity 1.2 from our guidelines.

The data representation is also a tool for understanding qualitatively the data set, but it is focused in showing the behavior prior to a mathematical treatment. This activity helps to bring some thoughts to the model that might be helpful for other activities. For example, a scatter plot might let us think that the variables are correlated, or even some clusters that might be formed prior to the structured algorithm. Also, a star representation might indicate that some projects are similar, prior to our clustering analysis.

4.3.1.3 Mean Vector, Covariance Matrix and Correlation Matrix

Once we have studied the data qualitatively, we can proceed with the data treatment studying it quantitatively. The main aspects that we should look in order to have a good perspective of our data set is: mean vector, covariance matrix and correlation matrix.

Both, the mean vector and the covariance matrix have the importance along the model development, not just here, for a quantitatively analysis. We will use their values in all phases of our model. To obtain those values we must recall the Equations $\bar{\mathbf{x}} = \frac{1}{n}\,\mathbf{X}'\mathbf{1}$

(1) and $\mathbf{S}_n = \frac{1}{n}\,\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\,\mathbf{1}\,\mathbf{1}'\right)\mathbf{X}$ (2).

The correlation matrix, which derives from the covariance matrix is important in this phase, because is the input necessary for testing the Hypothesis One. The correlation matrix can be obtained with the Equation $\mathbf{R} = \mathbf{D}^{-1/2}\,\mathbf{S}\,\mathbf{D}^{1/2}$ (7).

4.3.1.4 Testing H1

To close the first phase we need to check if hypothesis one is true. As we've been mentioning, the hypotheses are the stage gates to move forward along the model. So, here we want to verify if we may proceed or not.

The H1 tests if the correlation matrix is a diagonal matrix or not. Hence, once we have already calculated the correlation matrix, we just need to analyze it. The way we will do it is a proxy of Pearson product-moment correlation discussed earlier: if the correlation is a small value, we can say by a certain probability that is equal zero. Remembering that we can be flexible with this analysis because of the central limit theorem.

4.3.2 Similarities

The second phase of our model is devoted to perceiving similarities between observations. Here we will state our criteria for selecting the distance method that best fit with our data, and then we will standardize our data set for application of distance method. Finally, we will calculate the distances between observations and consequently calculate the similarities matrix.

4.3.2.1 Choosing the right distance method

The first activity is regarding the criteria for selecting the distance method. It should involve the following aspects:

- Are the data very sparse (we can see by the scatter plot)?
  - If the data is very sparse we can utilize the Minkowski method for bringing the data closer.
- Are the variables correlated (tested by H1)?
  - If the variables are not correlated, we have no means for utilizing the Statistical Distance method.
- What level of accuracy are we trying to reach?
  - If we want to reach a high level of clustering accuracy, it is preferred to not utilize the Minkowski method because it will bring the data closer and will be more difficult to group it.
- How great is the number of clusters we are pursuing to achieve?
  - Again, the number of clusters may be smaller if we utilize the Minkowski method, by the same reason mentioned early.

4.3.2.2 Distance and Similarities

After establishing the criteria, we can start the distance calculations. But, before doing that, we need to standardize the data in order to have it all in the same scale. We saw early the importance of standardization, and in the Appendix C we saw some methods for doing so. Hence, we just need to apply them here.

Finally, we can choose our distance method for calculating similarities between observations, giving the criteria and our data set. The three options that we have are: Euclidean Method, Statistical Method and Minkowski Method. All of them have their pros and cons, but we will choose what fit better for our model.

Once we have our distance matrix, we can build the Similarity matrix, which comprehends the same information as the distance matrix, but in a format that is easier to analyze qualitatively. For this calculation we jus need to use the Equation $s_{ik} = \frac{1}{1+ d_{ik}}$

(6).

4.3.3 Clustering

Once we have all data treated and all distances computed it is time to apply the clustering algorithms.

In order to have a robust model, and consequently a robust result we will apply two clustering algorithms: the single linkage and the k-means algorithm. Since they are different approaches and are founded in two different concepts of clustering, we will try to complement our study by bringing together these two different analyses with their pros and cons.

Hence, the first step in this phase will be applying the single linkage algorithm. We have already discussed the steps in the previous chapter, so it's just applying it given our data set.

The second activity is the dendogram construction, as we've seen on Figure 3.5. The dendogram is one more visual representation that we will utilize in our model. The main contribution here is to visualize how the distances are structured. Moreover, we can have a signal about "natural clusters"[19] that may be formed, and that will guide the first choice of clusters in the k-means algorithm.

Now we can apply the k-means algorithm that will provide the final arrangement of clusters, given our data set. The algorithm used will be the one we've discussed earlier. But there are some criteria that we must establish before running the algorithm.

- How many clusters do we have?

➢ To answer this, we need to look at the arguments described on first phase: data treatment.

- What is the first arrangement, or the initial groups?

➢ To answer this, we look at the dendogram.

Finally, we should test the Hypothesis 3, in order to have a robust clustered environment. It is worth saying that for testing the H3 we will utilize the MANOVA described on the Appendix D. If H3 is true, we can move to the final phase, if H3 is false, we run the phase three again.

4.3.4 Analysis

The last phase of our model starts with the Hypothesis's Two test. It is very important to understand now, that as we are talking about projects, we have the planned values for the variables, and the actual values. So, in order to have a proper conclusion about the strength of

[19] In this sense, natural cluster means observations that are very similar and have a natural direction to be united.

our model we need to certify that the clusters obtained from the planned values are the same that obtained from the actual values. And this is the grand opening for our analysis.

4.3.4.1 Consequences of testing H2

Let’s discuss briefly about the consequences of testing H2.

If H2 is false, them we can reach two conclusions:

1) The variables that we have chosen do not follow a standard path. They behave randomly and do not offer a possibility for planning according the pas projects behave.
2) There must be any outlier that disorients the correct cluster formation.

For both cases we need to examine well our model. The first case gives us important information about projects, but also brings the necessity of running again the model with different variables. The second case could be a bad data collection, or maybe the occurrence of a special cause that affected the normality on the project life cycle.

If H2 is true, then, we can move with our model and finish with the activities that are missing.

4.3.4.2 Labeling Clusters and future projects

Given that H2 is true, we can label every cluster by its characteristics. This activity structures the population of projects into possible types of project with a distinguished behavior. Once we know by hypothesis two that each cluster is different for each other, we now observe that there are some expected paths that a new project may follow, accordingly what was planned. So, we must label every clusters type by the characteristics of projects within.

Once we have labeled that clusters, we can now associate new projects with the determined group that it belongs. However, this activity requires that we test the hypothesis four, which say that a project should just belong to one cluster.

Testing H4 is very important because our conclusion about the behavior of new project take into consideration the expected behavior of the cluster that it is included. I conclusion, the new project is also labeled.

Finally, we can analyze the results of our model. We have tested all hypotheses and reach our objective. We need also to understand the typologies of project and what is suppose to happen to each category of cluster. The final analysis should also encounter a criticism about the model.

5 Application and Results

This chapter consists in the model's application. But first we need to understand from where our data was gathered, and what is the context that we are dealing with. Later, we will exploit the model we've just developed, trying to be clear and concise.

5.1 Data Gathering and Context

The first thing to do, before applying the model and showing the conclusions, is the explanation about the precedence of our data. Moreover, it is necessary a contextualization about the project environment. Both analyses give a concise foundation for some subjective decisions that we will have to take during the model application.

5.1.1 Data Gathering

The professor Franco Caron, from Politecnico di Milano, provided the data regarding our study. The data was the foundation of another student's thesis and given to our research in order to explore other ways of comparing the similarity between projects.

All the data given can be seen bellow on Table 5.1 and Table 5.2, and will be used in our model. Furthermore, we would like to clarify that we will just consider the data given, not adding or changing anything. Any deviation from some standard behavior or any value that seems odd we will analyze subjectively or will arbitrarily consider an outlier.

Because we don't know the procedures of data gathering, we will also consider that it followed a structured process and was made accurately and efficiently.

5.1.2 Context

The context of our data is the Oil & Gas industry. The projects are directly related to this sector of the economy. We know that the Oil & Gas Industry is very complex and carry with it an atmosphere of uncertainty. For this reason we know that the model should be robust enough to retain the variances within the data and smart enough to deal with possible outliers.

We will just show a brief history about the Oil & Gas Industry and then some comments about the data, explaining the variables and the observations that we have.

5.1.2.1 Oil & Gas Industry

The Oil & Gas Industry is an extremely large industry, which comprehends a widely source of energy for our globe. For example, more than 85% of USA's energy consumed in 2008 was strictly related to fossil fuels (e.g. coal, oil and natural gas).

There are two major sectors within this industry: upstream and downstream. The upstream refers to the exploration and production of oil and gas. The downstream refers to the operations after the production phase and through to the point of sale[20].

Our focus here is the projects for the upstream sector of Oil & Gas Industry. They are based in two main types of activities: Oil Drilling & Services, and Oil Refining. Both of them require an enormous amount of capital investment, and because of that they are very error sensitive.

These types of projects have some special characteristics that make them complex and uncertain (Costa Lima & Suslick, 2006):

- They are in part irreversible: in case of unsuccessful investment, the corporation cannot recover the capital invested.
- The future uncertainty affects the elements of cash flows (tool used to study the project viability): the average oil prices, the economic growth (controls the demand) and the interest rates can affect both revenues as well the costs in project investments (Jafarizadeh, 2010).

Given that, we can say the margin for project deviation is very narrow. We should have a robust model that takes into consideration the complexity and uncertainties of this industry. The new project should incorporate in the planning phase the events that happened with past projects, in order to have a good estimative about costs, risks and revenues. We can consider as ”new information” that could aggregate value for the new project (Amstrong, Galli, Bailey, & Couët, 2004).

That is our objective, then. We will turn our model more specific to the Oil & Gas Industry, in order to mitigate the risks within a new project development, and aggregate value for it.

5.1.2.2 Variables to describe the Model

Given the context, we need to choose carefully the variables for our study because the impacts could affect even our model than our analyses.

As we said, the context is very complex and uncertain. Moreover, the investment decisions should take that uncertainty into consideration. For that reason, we will try to include into our model a variable that contains the uncertainty and complexity by the point of view of the company that is developing the project. We should also include variables that

[20] Source: www.investopedia.com/features/industryhandbook/oil_services.asp

explain a project, as the Figure 2.2 shows. Hence, we will apply the paradigm Cost, Schedule, Performance.

So, to contain all this information and remain simple, our model will include the following variables: total cost, total duration and percentage of debit capital.

Just a brief explanation, we think that the percentage of leverage in the project is a good indication of how risky the project is, by the view of the company. As we said, an Oil & Gas project is relatively irreversible, cost an enormous amount of capital, and have the uncertainties of interest rates (that affect the cost of investment). For this reason, we think that a more leveraged project could be more risky, and consequently should be analyzed differently.

The cost and total duration are both project variables that should be analyzed because they translate certain patterns of projects. Moreover, they carry the information of how big is the project. In summary, the variables chosen have some synergies, and together could give information about different types of projects.

5.1.2.3 The data set and statistics strength

In terms of observations, we have a data set of sixteen projects and we will use them all. They are two types of projects: offshore projects and subsea projects. We will apply the model for the whole data set, not distinguishing the difference between onshore and offshore.

Moreover, it is worth saying that this binary variable does not affect the results, as we will see later. So, to summarize all the data set, we can present it as the following TABLE 1 and TABLE 2.

Table 5.1 – Data set of all projects in planning phase.

Projeto	Type	Country	CAPEX @FID	Leverage	Duration	CAPEX Equity
1	Subsea	Nigeria	$434.433.000,00	49,81%	17	$218.042.000,00
2	Subsea	Norway	$4.812.281.000,00	92,10%	28	$380.223.000,00
3	Subsea	Angola	$1.550.955.000,00	80,00%	30	$310.191.000,00
4	Subsea	USA	$587.600.000,00	25,00%	15	$440.700.000,00
5	Subsea	Norway	$1.724.064.000,00	70,00%	31	$517.219.000,00
6	Subsea	Angola	$1.941.010.000,00	80,00%	29	$388.202.000,00
7	Subsea	Egypt	$785.370.000,00	50,00%	18	$392.685.000,00
8	Subsea	Angola	$3.749.855.000,00	80,00%	37	$749.971.000,00
9	Offshore	Italy	$518.508.000,00	27,36%	26	$376.621.000,00
10	Offshore	Congo	$289.896.000,00	0,00%	20	$289.896.000,00
11	Offshore	Tunisia	$253.060.000,00	51,00%	15	$124.000.000,00
12	Offshore	Australia	$445.870.000,00	0,00%	32	$445.870.000,00
13	Offshore	Egypt	$326.340.000,00	50,00%	18	$163.170.000,00
14	Offshore	Congo	$227.507.000,00	0,00%	18	$227.507.000,00

15	Offshore	Egypt	$202.633.000,00	40,00%	15	$121.580.000,00
16	Offshore	Egypt	$190.253.000,00	36,82%	17	$120.200.000,00
T1	Subsea	Nigeria	$536.553.000,00	30,00%	17	$375.587.000,00
T2	Offshore	Tunisia	$214.980.000,00	51,00%	19	$105.341.000,00

Table 5.2 – Data set of all projects after execution.

Projeto	Type	Country	CAPEX Actual	Leverage	Duration	CAPEX Equity
1	Subsea	Nigeria	$486.351.000,00	49,81%	27	$244.100.000,00
2	Subsea	Norway	$7.732.216.000,00	92,10%	35	$610.845.000,00
3	Subsea	Angola	$1.810.417.000,00	80,25%	42	$357.525.000,00
4	Subsea	USA	$525.400.000,00	24,78%	22	$395.200.000,00
5	Subsea	Norway	$1.623.758.000,00	70,53%	30	$478.466.000,00
6	Subsea	Angola	$2.120.870.000,00	80,00%	31	$424.174.000,00
7	Subsea	Egypt	$1.009.860.000,00	50,00%	23	$504.930.000,00
8	Subsea	Angola	$4.030.929.000,00	80,53%	37	$785.000.000,00
9	Offshore	Italy	$598.549.000,00	27,00%	32	$436.933.000,00
10	Offshore	Congo	$549.103.000,00	2,94%	32	$532.950.000,00
11	Offshore	Tunisia	$345.667.000,00	51,00%	26	$169.377.000,00
12	Offshore	Australia	$1.063.100.000,00	0,00%	39	$1.063.100.000,00
13	Offshore	Egypt	$550.639.000,00	51,44%	17	$267.395.000,00
14	Offshore	Congo	$310.137.000,00	0,00%	17	$310.137.000,00
15	Offshore	Egypt	$264.900.000,00	40,02%	20	$158.900.000,00
16	Offshore	Egypt	$199.670.000,00	39,45%	19	$120.910.000,00
T1	Subsea	Nigeria	$689.500.000,00	28,56%	21	$492.600.000,00
T2	Offshore	Tunisia	$476.802.000,00	49,12%	30	$242.610.000,00

Given that we have few projects for very complex and sophisticated statistical tools, we have to give a qualitative analysis to validate our model, before applying the methodology. So, we will explain every statistical passage and show why we can apply such a sophisticated tool without losing strength.

The first thing we should analyze is the basics: mean vector and covariance matrix. As we said, the projects are related to the Oil & Gas Industry, a very complex sector that is influenced by many factor, for example: government, energy sector, politics, economy and so on. But as we simplified, the level of leverage will include all uncertainty.

As we minimized the number of variables, the need for a great number of observations diminishes. The following relation gives the robustness of covariance matrix:

$$If\ n\ \leq p, that\ is\ (sample\ size) \leq (number\ of\ variables), then$$

$$|S| = 0, for\ all\ samples$$

This relation means that if the number of variables is larger, we need more observations in order to have an existing inverse of covariance matrix. More than that, assuming that we have three variables, we can certainly say that each subsample of more than three observations will have an existing inverse of covariance matrix. So, as we minimize the number of variables, we can express the results with fewer observations (Johnson & Wichern, 2007).

The second point is related to the distances methods. As we gave a relation between the number of variables, observations and correlation matrix, we can say that the distance method that includes it is also robust. The other methods are the pairing of two observations, so this number does not influence.

As we run through our methodology, we can say that the number of observations will not affect the data representation. As well the clustering methodology because it is an algorithm that takes as input the correlation matrix, by the distance methods.

The challenge will surge next, when we talk about the MANOVA and Discriminant Analysis. Both utilize as a proxy that the population is normal-distributed. The problem is that we cannot guarantee the normality for the population. But we can relax this hypothesis for two reasons:

I. The Oil & Gas projects tend to be great projects, in terms of cost, schedule and return. The error would minimize the potential gains, and because of that a huge effort is made in order to standardize the projects with the minimal variance between similar projects. If the company does not do the same as the competitor, it would be difficult to follow up in the industry;

II. There are several clusters within the same typology of project (Oil & Gas projects). So, as we can relax the Central Limit Theorem by saying that a sum of different populations will give a normal distributed population;

Given all this explanation, we can move forward and continue our model without compromising the methodology and consequently, the results.

5.2 Running the model

Now it is time to run the model, which we've developed in the previous chapter. The guidelines for our model are shown in the Figure 4.1, and will be exactly our procedure. We will follow the four phases of model development and during the execution of the phases we will propose some "mini conclusions" in order to have a better understanding during the process.

Later, we will end up our model with a broad analysis and a more detailed discussion that will take into consideration all mini conclusions that we find during our model development.

5.2.1 Phase One: Data Treatment

The first phase our model is a qualitative and quantitative analysis of our data set. The application of the activities expressed on the Figure 4.1 is a good start for our model because gives us a possibility to contextualize and have a wide vision of our problem.

5.2.1.1 Object Description

The object description is based in three questions:

- What is our object?

  Our Object is a project of Oil & Gas Company. More than that, is a range of several projects that we will try to group in order to perceive similar behaviors and help with the complexity and uncertainty of the system.

- What are the variables that represent this object?

  To represent our object, we utilize four variables: cost, duration, leverage and type of project. The cost is represented in U$ dollars, the duration in months and the leverage in percentage of debit capital investment. More particular is the type of project, which is represented by a binary number. If the project is an offshore project the value is zero, if it is a subsea project, the value is one.

- How many observations do we have?

  We currently have eighteen observations, equally distributed: nine for offshore projects and nine for subsea projects. Two of those observations we will randomly extract from our data set in order to test the Hypothesis Four.

5.2.1.2 Data Representation

The data representation gives us the opportunity to have some ideas about the data. It is a systemic view for the problem. Let’s check the scatter plots that confront pair of variables. There are two main aspects that we are trying to investigate now: some visual correlation between variables and the presence of clusters (prior to algorithm application). We will do this analysis in two different moments, the first just the confront between variables that are strictly quantitative. Then, we will analyze the confrontation between typologies of projects. The Figure 5.1 bellow shows the first analysis.

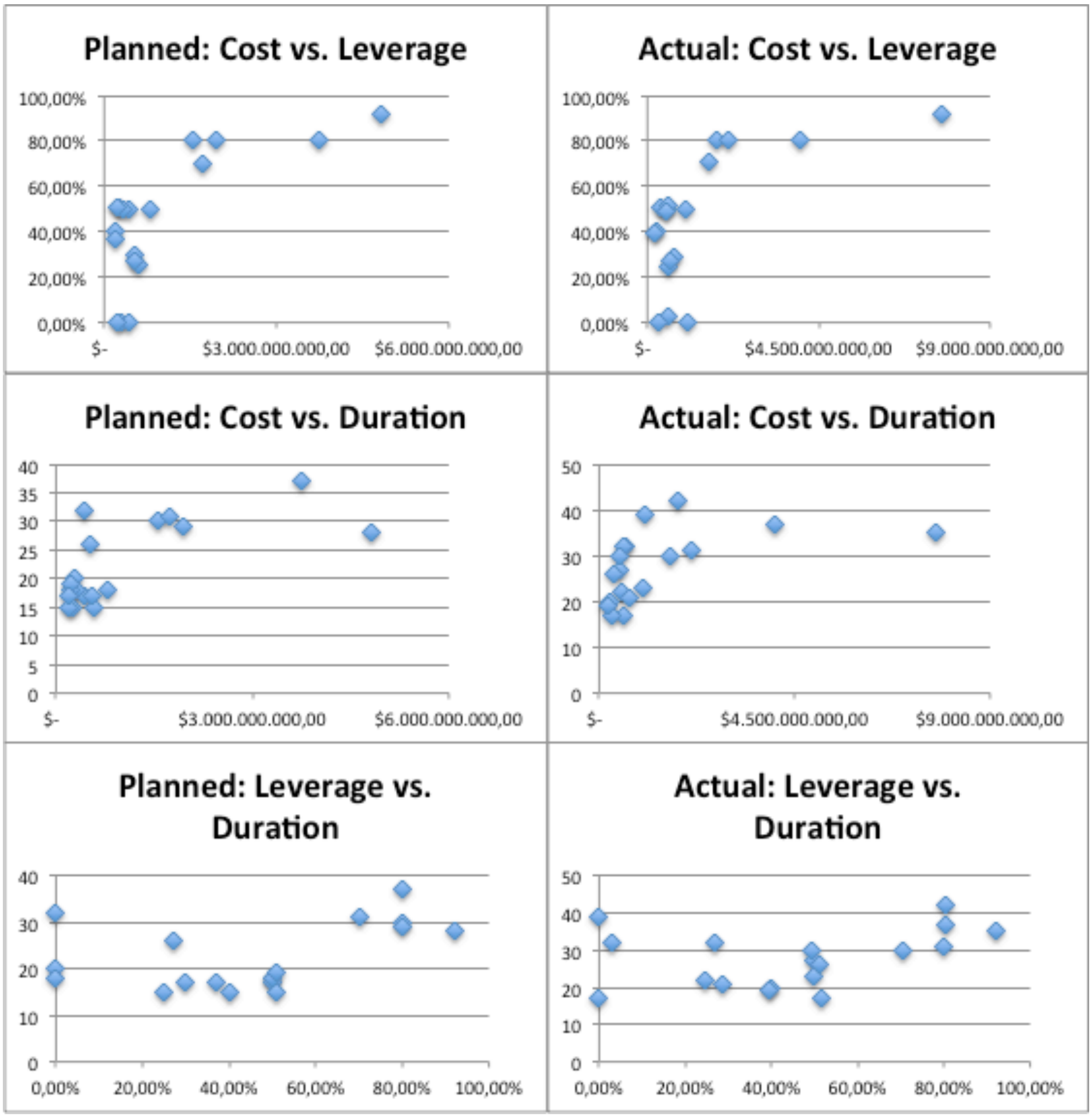


Figure 5.1 – Scatter plot for all pair of variables strictly quantitative.

The first thing that comes to mind is the similarity between the results of planned and actual values. They seem to have the same configuration, and this might give a good perspective for verifying the Hypothesis Two.

At a second glance, we can sense that there is a little correlation between Cost and Leverage, Cost and Duration, and Leverage and Duration. This analysis will be more robust right after the representation. But, we have a signal here that the Hypothesis One might be also true.

The third comment is about some clustering pattern. We can see from every confrontation that at some areas we have a hard density of points. That's a good indication that we might have statistically different clusters, and the Hypothesis Three might be true.

Now, let's check the Figure 5.2 for the confrontation of typologies of project with quantitative variables.

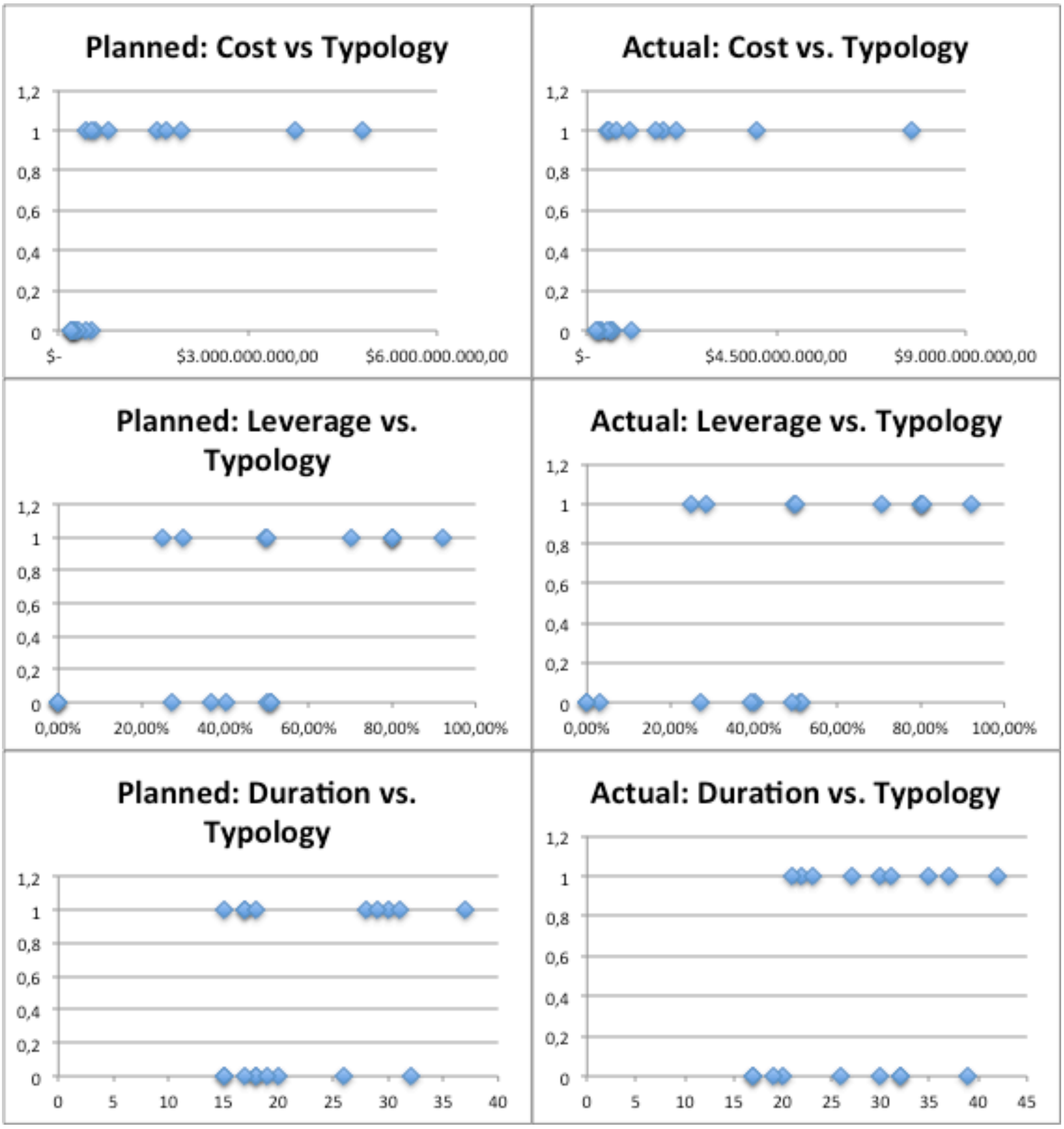


Figure 5.2 – Scatter plot for confront between typologies of projects.

In this case, we can see that there is not so much difference between typology of projects. The one thing that comes to mind is that offshore projects tend to be less costly, and a little bit less leveraged. In terms of duration, we can't really say by just examining visually. In other words, we can't say that typology of projects are a natural cluster.

For this reason, we will abandon the "typology of project" variable because it does not carry much information, but it could be a heavy weight on our distance calculation. So, we have our first mini conclusion:

MC1: The typology of project does not affect the clustering procedure because the offshore projects and subsea projects have similar distributions of cost, leverage and duration.

5.2.1.3 Basic Statistics

The basic statistics comprehends: mean vector, covariance matrix and correlation matrix. The calculi for those values are very simple, as we saw in chapter three. Hence, we had the following planned values:

$$\bar{\mathbf{x}} = \begin{bmatrix} \$1.127.500{,}00 \\ 46\% \\ 23 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 1{,}8E+18 & 2{,}8E+08 & 6{,}3E+09 \\ 2{,}8E+08 & 8{,}4E-02 & 8{,}8E-01 \\ 6{,}3E+09 & 8{,}8E-01 & 5{,}1E+01 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1{,}00 & 0{,}73 & 0{,}67 \\ 0{,}73 & 1{,}00 & 0{,}43 \\ 0{,}67 & 0{,}43 & 1{,}00 \end{bmatrix}$$

And for the actual values:

$$\bar{\mathbf{x}} = \begin{bmatrix} \$1.450.000{,}00 \\ 46\% \\ 28 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 3{,}5E+18 & 3{,}5E+08 & 7{,}3E+09 \\ 3{,}5E+08 & 8{,}2E-02 & 7{,}2E-01 \\ 7{,}3E+09 & 7{,}2E-01 & 5{,}9E+01 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1{,}00 & 0{,}65 & 0{,}51 \\ 0{,}65 & 1{,}00 & 0{,}33 \\ 0{,}51 & 0{,}33 & 1{,}00 \end{bmatrix}$$

It is interesting to note here that the values for cost and duration had gotten higher in terms of mean and variance. It is another evidence of how complex and uncertain is the Oil & Gas Industry, and how much this affects the project development.

5.2.1.4 Testing H1

The next step is testing H1 to see if there is any correlation between variables. The test we mentioned earlier is the Pearson Product-moment correlation, and the value T(r) calculated should be compare to a t-student value, with significant level α and degrees of freedom of $n-2$. If any value is greater than the t-student value, we can say that the correlation is different from zero. Otherwise, we will consider as being equal zero.

With significance level of $\alpha = 5\%$ and $n-2 = 14$ degrees of freedom, we find the value for t-student distribution equal 1,76. In this case, let's compare the values that we'd found for T(r).

Table 5.3 – T-values for all correlations between variables.

Correlation Test – T(r)			
	Planned Values for the variable		
	V1	**V2**	**V3**
V1	0,00	3,18	2,20
V2	3,18	0,00	1,29
V3	2,20	1,29	0,00
	Actual Values for the variables		
	V1	**V2**	**V3**
V1	0,00	4,05	3,39

V2	4,05	0,00	1,76
V3	3,39	1,76	0,00

Source: developed by author

By testing H1, we can say with 95% sure that there is a correlation between the V1 & V2, and V1 & V3. This result in the second mini conclusion that we can infer:

MC2: With 95% sure we can verify the correlation between the cost variable with both leverage and duration. But we can't estate the same for the correlation between leverage and duration variables.

5.2.2 Phase Two: Similarities

Now that we already have a good understanding about our data, we can start the mathematical aspects of our model. The second phase is related to the similarities analyses, and has as inputs the data set and the qualitative analysis that we just have done. So, let's follow our guidelines and discover more about our data.

5.2.2.1 Choosing the right distance method

The first activity is related to some criteria that we will use to define the best distance method. We need to answer the following questions:

- Our data is very sparse?

  From the scatter plots we can say that we don't have a very sparse data set. Some projects that tend to be a little distant, but it seems to be normal.

- There is any correlation between variables?

  The costs are correlated with leverage and duration variables. It's is a good indicative that we should use the statistical distance.

- What level of accuracy are we trying to reach?

  Given that our data set is not really sparse, we should give more weight to the distances measures in order to have a more accurate analysis.

- How great is the number of clusters we are trying to achieve?

  We didn't go through this analysis so far, but considering our variables and our scatter plot diagram, we could say that our focus is four clusters. The reason we will explain later.

5.2.2.2 Distance and Similarities

Once we answered the questions, we automatically established the criteria for our distance method. The first characteristic is that our distance method should take into consideration the correlation between variables, eliminating the Euclidean Distance. The second aspect is that we should weight our distance properly, eliminating the Minkowski Distance. So, our best choose for distance method is the Statistical Method.

The distances were calculated and computed and will be utilized in the clustering algorithm. But we also computed the similarities (with values comprised between 0 and 1) that are displayed in Appendix F. These values are more visual than the distances and we can take some analysis from them after the clustering method.

If we take a glance on those values we can predict that the projects P1, P14 and P16 may be at the same cluster, as well the projects P11 and P15. It is important to say that not necessarily this visual analysis will conclude at the exactly clustering. For example, the projects P1 and P14 may be similar, but P14 and P16 very discrepant. In that situation, we cannot form a cluster {P1, P14, P16}. So, let's process the clustering algorithms in order to have a robust result.

5.2.3 Phase Three: Clustering

We already have all distances computed, and a good understanding about our context. So, we can know apply the third phase of our model, which is the clustering process.

As we told earlier, we will try to apply two algorithms: the single linkage and k-means. The first is a hierarchical procedure, and will give us a good visualization about the clusters that might be formed. Moreover, will give us a good start for the k-means algorithm. Hence, let's see how can we group our projects.

5.2.3.1 Single Linkage and Dendogram Representation

Before we move forward with the clustering algorithms we need to comment some important thoughts that might be helpful for our data analysis. As we said, the single linkage will be an aiding tool for correctly choosing the inputs for the k-means algorithm. So, it is not necessary to develop the single linkage for both planned and actual values. Once we have an initialization for the k-means, we will standardize it for the planned and actual values of the variables, in order to prevent variances within the model.

The usage of single linkage algorithm followed the steps that we've told before in the Chapter Three. The Statistical Method was used for calculating the distances. Furthermore, for doing the calculi and computation of results, we've utilized the MS Excel®. Finally, the result can be represented by a dendogram that we can see on the Figure 5.3.

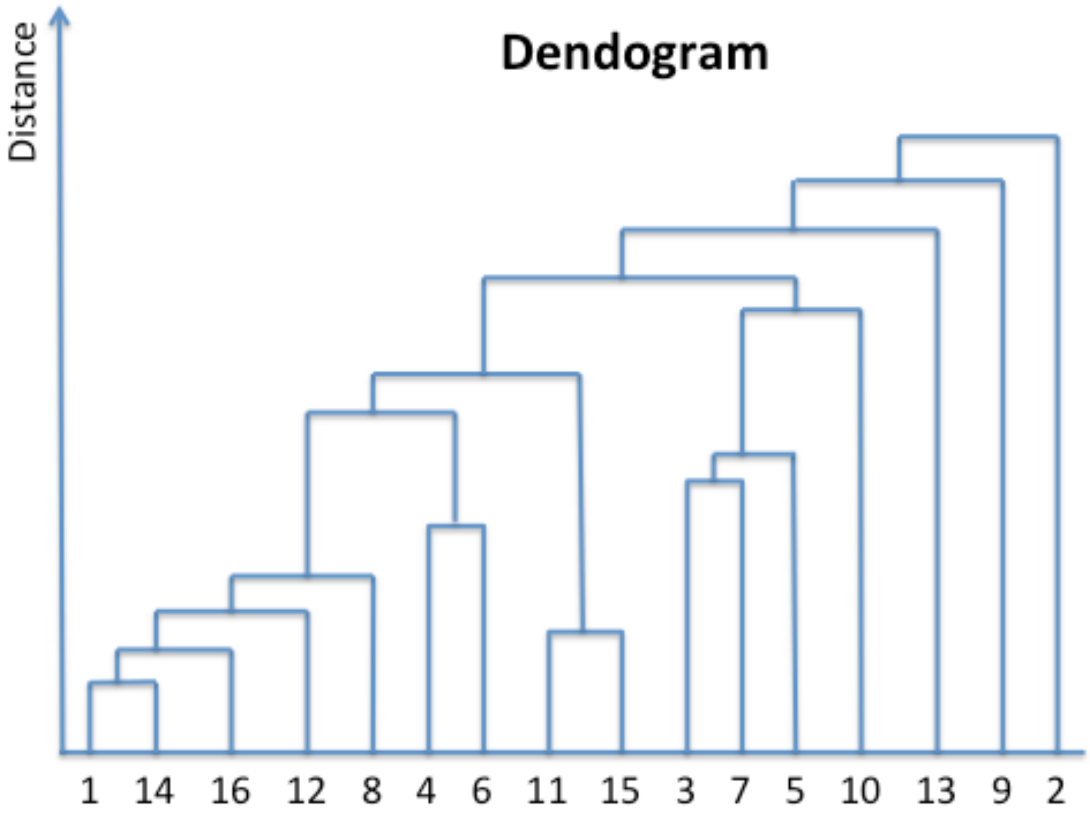


Figure 5.3 – Dendogram for planned values.

We can notice that the projects 1 and 14 are very similar, and one of them could be on pivot for the k-means. The same analysis is possible for the projects 11 and 15, 4 and 6. The other extreme is also true; the projects 9 and 2 could also be another pivot, once they are most distant of the other projects.

5.2.3.2 K-means Algorithm

There are some issues to deal with, before entering the k-means algorithm for our final clustering configuration. As we've been saying, it is a non-hierarchical method, and because of data there are some inputs for the model that we have to decide arbitrarily.

The first decision we should make is the number of clusters we want to have. By the analysis of our scatter plot, and the variables themselves, we can have some thoughts about the clusters configuration. We've that there are a correlation between Cost & Duration and Cost & Leverage. So, looking for scatter plots that confront these two pairs do not show us much information about clusters. But, if we look at the graph that represents Leverage vs. Duration we can have some ideas of groups that would form. The Figure 5.4 above shows some interesting arguments.

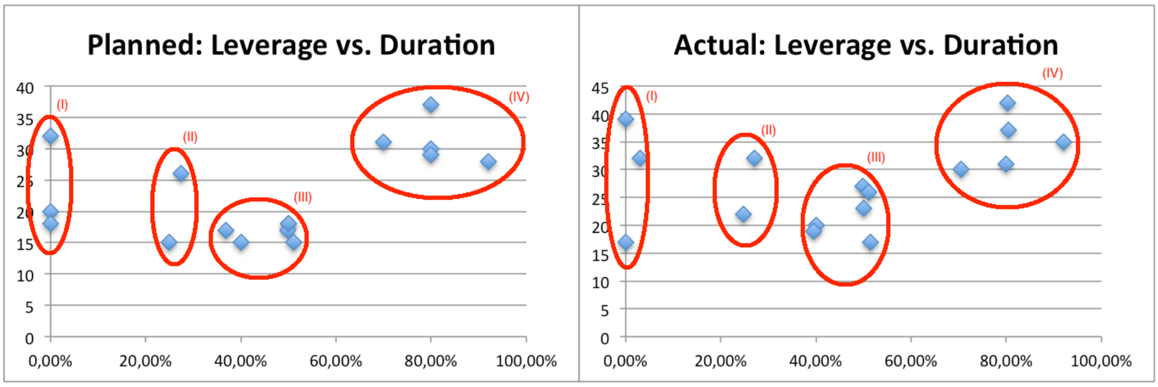


Figure 5.4 – Clusters found in scatter plots.

The picture shows some natural groups that may represent the following results:

V. Low level of leverage and high level of duration.

VI. Low level of leverage and low level of duration.

VII. High level of leverage and low level of duration.

VIII. High level of leverage and high level of duration.

Notice that the cost does not take particular relevance now because its value is correlated with the other two variables, as we stated on MC2. Those arguments are enough for developing the Mini Conclusion Three.

MC3: There are four natural clusters in our study, depending on duration and level of leverage. The cost here does not play a main role because it is correlated to both duration and leverage.

The second issue that we have to deal is the initial cluster arrangement. But this task became easy to resolve because we already treated our data and utilized the single linkage algorithm to establish a first combination of clusters. So, we can apply the k-means algorithm with the following inputs:

- Number of clusters: four.
- Initial cluster arrangement: (P1, P14, P16, P12, P8, P4, P6), (P11, P15), (P3, P7, P5) and (P10, P13, P9, P2).

The algorithm was developed by a VBA code for MS Excel® and the results can be seen with more detail on Appendix G. The clusters that resulted from the k-means are: (P1, P7, P11, P13, P15, P16), (P2, P3, P5, P6, P8), (P4, P9) and (P10, P12, P14). They are a little different from the initial inputs, but the reflect a robust conclusion that we will discuss later.

5.2.3.3 Testing H3

But before we do the final discussion, we need to test the Hypothesis Three to see if the clusters obtained are statistically different from each other. We must now apply a two-steps methodology in order to validate H3. First, we calculated the mean, or the center of each cluster, and then, we applied the MANOVA, as we referred before. The procedure can be seen on the Appendix D and the results are shown in Table 5.4 above.

Table 5.4 – MANOVA calculated for the clusters C1, C2, C3, and C4.

Source of Variation	Matrix of sum of squares and cross products (SSP)	Degrees of freedom (d.f.)
Treatment	$\mathbf{B} = \begin{bmatrix} 0{,}84 & 0{,}93 & 0{,}54 \\ 0{,}93 & 1{,}54 & 0{,}42 \\ 0{,}54 & 0{,}42 & 0{,}42 \end{bmatrix}$	$4 - 1 = 3$
Residual (error)	$\mathbf{W} = \begin{bmatrix} 0{,}37 & 0{,}09 & 0{,}03 \\ 0{,}09 & 0{,}05 & -0{,}01 \\ 0{,}03 & -0{,}01 & 0{,}17 \end{bmatrix}$	$16 - 4 = 12$
Total (corrected for the mean)	$\mathbf{B} + \mathbf{W} = \begin{bmatrix} 1{,}21 & 1{,}02 & 0{,}57 \\ 1{,}02 & 1{,}59 & 0{,}41 \\ 0{,}57 & 0{,}41 & 0{,}59 \end{bmatrix}$	$16 - 1 = 15$

Source: developed by author

The Wilk's Lambda for the given table is equal 0,0069, and consequently, the Barlett Index is 57,24. Moreover, the upper 5th percentile of a chi-square distribution with 3(4 – 1) = 9 degrees of freedom is $\chi_9^2(5\%) = 23{,}59$. So, we can reject the hypothesis that the populations have the same means.

Moreover, we reached the conclusion that, with a 95% of certainty, the clusters obtained are different from each other. By this reason, we say that H3 is true and we are now able to move along with our model, directly to the analysis.

MC4: The four natural clusters are different from each other, and are expected to behave differently.

5.2.4 Phase Four: Analysis

The last phase of our model is the Analysis. Here we will finalize the last issues until we achieve our objective. To do so, we need to test the Hypothesis H2 and H4, and classify our cluster accordingly their characteristics and probable behavior.

5.2.4.1 Testing H2

To test H2 is very simple, we just need to verify the outputs from the clustering algorithm. In other word, we just need to observe what are the final clusters that arrived from the k-means algorithm, and see if they are the same, planned and actual values.

By analyzing the k-means outputs we can say that the clusters are exactly the same: (P1, P7, P11, P13, P15, P16), (P2, P3, P5, P6, P8), (P4, P9) and (P10, P12, P14). So, we can affirm that the Hypothesis Two is true, so we can continue our model until reaching the objective.

5.2.4.2 Clusters Classification and Testing H4

Now we face our last challenge before completing the model and achieving our objective. We need to classify the clusters we've obtained and test the hypothesis four.

For the first task we will try to utilize the classification we said during the k-means algorithm:

I. Low level of leverage and high level of duration.
II. Low level of leverage and low level of duration.
III. High level of leverage and low level of duration.
IV. High level of leverage and high level of duration.

At the Appendix E we can see the star representations for each project in the planned values. Once that we reached the conclusion that H2 is true, and the clusters with planned value are the same that with actual values, the analysis could be made by just one side. Let's call the clusters:

- C1 = {P10, P12, P14};
- C2 = {P4, P9};
- C3 = {P1, P7, P11, P13, P15, P16};
- C4 = {P2, P3, P5, P6, P8};

By the star representations we can see clearly that C1 is Low Leverage and High Duration, C2 is a cluster of Low Leverage and Low Duration, C3 is a cluster of High Leverage and Low Duration, and C4 is High Leverage and High Duration. So, we conclude that we found some patterns that could help to contain the complexity and uncertainty of projects in Oil & Gas Industry.

The last task is to test the hypothesis four. We left two projects of our data set for doing the test of H4, we will call them Test Projects PT1 and PT2. We can see by the star representation that PT1 probably belongs to the cluster C2 and PT2 probably belongs to C3.

But affirming this would be just guessing. So, we need to apply the Discriminant Analysis that we've discussed earlier to be sure if the Test Projects belong to one, and just one cluster. The distances from each Test Project to each cluster center were calculated, and the results can be seen on Table 5.5.

Table 5.5 – Distances from projects to clusters calculated for discriminant analysis.

Clusters	**Distance to Cluters' center**	
	PT1	**PT2**
C1	4,93	2,45
C2	0,44	4,25
C3	0,98	0,50
C4	10,22	7,32

Source: developed by author

Comparing those values to chi-square distribution with 3 degrees of freedom $\chi_3^2(90\%) = 0{,}58$. We choose the significant value of 90% because we are dealing with a small sample of projects, so we are not able to guarantee a great level of precision. As we discussed earlier, greater is the significant level, closer to the cluster center is the Test Project located.

The Table 5.5 shows that our visual analysis was right, and PT1 belongs to C2 and PT2 belongs to C3. Moreover, they belong to just one cluster. For that reason, we can state that the Hypothesis H4 is True, and conclude our analysis with the mini-conclusion MC5.

MC5: The Discriminant Analysis showed that future projects belong to one, and just one cluster with a significant level of 90%. Which means that our cluster structure is robust enough.

5.3 Analysis and Discussion

To fulfill our study, and close model application we should recall the five mini-conclusions that we've stated during this chapter. By doing this, we will be able to analyze the results in a robust way and develop some conclusions about our model. The five MC were:

1) The typology of project does not affect the clustering procedure because the offshore projects and subsea projects have similar distributions of cost, leverage and duration.
2) With 95% sure we can verify the correlation between the cost variable with both leverage and duration. But we can't estate the same for the correlation between leverage and duration variables.
3) There are four natural clusters in our study, depending on duration and level of leverage. The cost here does not play a main role because it is correlated to both duration and leverage.
4) The four natural clusters are different from each other, and are expected to behave differently.

5) The Discriminant Analysis showed that future projects belong to one, and just one cluster with a significant level of 90%. Which means that our cluster structure is robust enough.

If we take into consideration those mini conclusions and the hypothesis that we stated earlier, we can easily see that our model worked in a systemic way, designed with some boundaries (the hypothesis) that led the conclusions to a certain path within a closed system.

Bringing all together, we arrived to a more complex conclusion that in a context of Oil & Gas Projects, we have an array of variables (cost, schedule and leverage) that follows some patterns accordingly the values that these variables assume. There are four typologies of Oil & Gas projects, and they are different from each other. Moreover, every future project will behave similarly to one typology, and consequently will have similar characteristics.

So, we can say that our conclusion is aligned with our objective, and reach the goal that we first proposed. Restating our object, we said that:

The objective of this study is: (I) perceive similarities throughout a range of projects, (II) put these projects into groups (clusters) accordingly those similarities, (III) study the characteristics of each cluster and (IV) classify as a typology of project.

The task (I) was completed by the qualitative analysis and the similarities and distance methods that we applied, comprised by the first and second phases of our model. The clustering algorithms in the third phase, like the single linkage and the k-means, guaranteed the Task (II). The task (III) was aided by the graphical representation, and a bibliographic study about the context. The task (IV) was also based in the cluster analysis and discriminant analysis.

In conclusion, we can say that our model fulfills the requirements that we imposed in the first place, and is robust enough for two reasons: the first reason is that we guided our model with some strict hypothesis, and the second reason is that we could verify our hypothesis with some strong statistical tools. So we can say that we can assume the conclusion as true, and that our objective was achieved.

6 Conclusion

As we concluded in the last chapter, our model achieved our objective, and more than that, gave us certain thoughts about Oil & Gas projects. But we were not aiming on the first place to develop a model just to verify some peculiarities about the Oil & Gas projects.

The objective that we stated was: (I) perceive similarities throughout a range of projects, (II) put these projects into groups (clusters) accordingly those similarities, (III) study the characteristics of each cluster and (IV) classify as a typology of project.

So, we can see the openness of our study, because we are dealing with a broad model that can be adjusted to a great range of projects' classes. The variables that we assumed in the Oil & Gas projects can be different from, for example, hospital projects, or software projects.

Moreover, the algorithms are flexible enough to analyze different number of projects and variables. All that we've developed is flexible enough to be adjusted to other systems and other environments. If we take into consideration everything different from Oil & Gas projects, we can still have our objective reached.

That is our first comment for this entire study:

The model is flexible enough to be applied to different classes of projects, and will arrive to the objective no matter the environment and system we are analyzing.

But once we've reached our objective, it does not mean that our job is finished. Our model is not a way to manage projects; it is more a tool to help a project to achieve success, as we said in the first chapter. The outcomes of our model are the following: clusters with similar characteristics and consequently a range of projects with similar behavior grouped and easy to manipulate.

So, our model does not say how to manage a project, neither how the EVMS or PPM will work with a project in the pipeline. Our project says that a range of project with same inputs will end up having the same outcomes. We can infer, though, that a future project with some characteristics, or some variable values will behave approximately equal as a project with similar attributes.

That is the beauty of our model, because we achieved a conclusion that the inputs and outcomes are related. Consequently, as Machiavelli words were interpreted: the end justifies the means. Which means that our model utilizes the result of a project to infer and justify the usage of some tools and managing model. Therefore we have the second comment of our entire study:

The model is not a managing tool. But should be used as an aid tool for helping to achieve project success. It cannot walkthrough the project predicting the events, but can give a good view about its behavior and its outcomes.

Finally, as we said, we have a flexible tool that helps the project managers to predict some behaviors of their projects; we need now to propose a following study about those appliances of our model. We summarized the potentiality of our study, but those were not verified, as they were not the scope of our study.

Our main objective was reached, but we need to do a step-up in order to promote something that could help in a long extension the project management. Imagine that we have a mechanism that predict not the events that might happen to a project, but something that can infer statistically how those events will influence a project on average.

More and less, we developed this tool, but we did not mix it from the beginning to the end of a project, coming along the EVMS. More broadly, we did not see our study working within a PPM and seeing the effectiveness for the Portfolio success as a whole.

So, we can say as a third comment for our study that:

A following study should be conducted in order to see how our clustering model would connect with the EVMS to achieve project success. Moreover, we should conduct a study with several projects that form a portfolio in order to analyze the impact of the clustering method for the PPM success.

So, we end this study calling for a follow up, and saying what were the key points that can make it a strong and robust model. The first thing that comes to mind is that our model is tight up with a well-managed statistics. A well-known statistical tool supports every passage, and we tried to be flawless in terms of mathematical approach. Moreover, we've recalled several times our objective, so we wont lose our path along the model development. The idea of having hypothesis and hypothesis tests also helped in this sense.

Finally, we start our study by showing the characteristics of project success and the tools for project management. By doing this we could introduce in this conclusion (as we did) the next steps for our model: the junction of our structure with EVMS and moreover, the PPM.

Appendix A

Organization

1. Define the authorized work elements for the program. A work breakdown structure (WBS), tailored for effective internal management control, is commonly used in this process.
2. Identify the program organizational structure including the major subcontractors responsible for accomplishing the authorized work, and define the organizational elements in which work will be planned and controlled.
3. Provide for the integration of the company's planning, scheduling, budgeting, work authorization and cost accumulation processes with each other, and as appropriate, the program work breakdown structure and the program organizational structure.
4. Identify the company organization or function responsible for controlling overhead (indirect costs).
5. Provide for integration of the program work breakdown structure and the program organizational structure in a manner that permits cost and schedule performance measurement by elements of either or both structures as needed.

Planning, Scheduling, and Budgeting

6. Schedule the authorized work in a manner, which describes the sequence of work and identifies significant task interdependencies required to meet the requirements of the program.
7. Identify physical products, milestones, technical performance goals, or other indicators that will be used to measure progress.
8. Establish and maintain a time-phased budget baseline, at the control account level, against which program performance can be measured. Budget for far- term efforts may be held in higher-level accounts until an appropriate time for allocation at the control account level. Initial budgets established for performance measurement will be based on either internal management goals or the external customer negotiated target cost including estimates for authorized but undefinitized work. On government contracts, if an over target baseline is used for performance measurement reporting purposes, prior notification must be provided to the customer.
9. Establish budgets for authorized work with identification of significant cost elements (labor, material, etc.) as needed for internal management and for control of

subcontractors.

10. To the extent it is practical to identify the authorized work in discrete work packages, establish budgets or his work in terms of dollars, hours, or other measurable units. Where the entire control account is not subdivided into work packages, identify the far term effort in larger planning packages for budget and scheduling purposes.
11. Provide that the sum of all work package budgets plus planning package budgets within a control account equals the control account budget.
12. Identify and control level of effort activity by time-phased budgets established for this purpose. Only that effort which is unmeasurable or for which measurement is impractical may be classified as level of effort.
13. Establish overhead budgets for each significant organizational component of the company for expenses, which will become indirect costs. Reflect in the program budgets, at the appropriate level, the amounts in overhead pools that are planned to be allocated to the program as indirect costs.
14. Identify management reserves and undistributed budget.
15. Provide that the program target cost goal is reconciled with the sum of internal program budgets and management reserves.

Accounting Considerations

16. Record direct costs in a manner consistent with the budgets in a formal system controlled by the general books of account.
17. When a work breakdown structure is used, summarize direct costs from control accounts into the work breakdown structure without allocation of a single control account to two or more work breakdown structure elements.
18. Summarize direct costs from the control accounts into the contractor's organizational elements without allocation of a single control account to two or more organizational elements.
19. Record all indirect costs, which will be allocated to the contract.
20. Identify unit costs, equivalent units costs, or lot-costs when needed.
21. For EVMS, the material accounting system will provide for:

    I. Accurate cost accumulation and assignment of costs to control accounts in a manner consistent with the budgets using recognized, acceptable, costing techniques.

    II. Cost performance measurement at the point in time most suitable for the category of material involved, but no earlier than the time of progress payments or actual receipt of material.

III.Full accountability of all material purchased for the program including the residual inventory.

Analysis and Management Reports

22. At least on a monthly basis, generate the following information at the control account and other levels as necessary for management control using actual cost data from, or reconcilable with, the accounting system:
    I.Comparison of the amount of planned budget and the amount of budget earned for work accomplished. This comparison provides the schedule variance.
    II.Comparison of the amount of the budget earned the actual (applied where appropriate) direct costs for the same work. This comparison provides the cost variance.
23. Identify, at least monthly, the significant differences between both planned and actual schedule performance and planned and actual cost performance, and provide the reasons for the variances in the detail needed by program management.
24. Identify budgeted and applied (or actual) indirect costs at the level and frequency needed by management for effective control, along with the reasons for any significant variances.
25. Summarize the data elements and associated variances through the program organization and/or work breakdown structure to support management needs and any customer reporting specified in the project.
26. Implement managerial actions taken as the result of earned value information.
27. Develop revised estimates of cost at completion based on performance to date, commitment values for material, and estimates of future conditions. Compare this information with the performance measurement baseline to identify variances at completion important to company management and any applicable customer reporting requirements including statements of funding requirements.

Revisions and Data Maintenance

28. Incorporate authorized changes in a timely manner, recording the effects of such changes in budgets and schedules. In the directed effort prior to negotiation of a change, base such revisions on the amount estimated and budgeted to the program organizations.
29. Reconcile current budgets to prior budgets in terms of changes to the authorized work and internal re-planning in the detail needed by management for effective control.
30. Control retroactive changes to records pertaining to work performed that would

change previously reported amounts for actual costs, earned value, or budgets. Adjustments should be made only for correction of errors, routine accounting adjustments, effects of customer or management directed changes, or to improve the baseline integrity and accuracy of performance measurement data.

31. Prevent revisions to the program budget except for authorized changes.
32. Document changes to the performance measurement baseline.

Appendix B

Consider the sample mean array and the covariance array as given by

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad\qquad \mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

As well, consider an array **1** (*n x 1*) that every data within this array is given by the value one and **I** as the identity matrix (*n x n*). For the Proof #1:

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} x_{j1} \\ \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} x_{j2} \\ \vdots \\ \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} x_{jp} \end{bmatrix} = \frac{1}{n}\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n}\,\mathbf{X}'\mathbf{1}$$

For the Proof #2, which is the equation to obtain the covariance array we have the following:

$$\mathbf{1}\,\bar{\mathbf{x}}' = \frac{1}{n}\,\mathbf{1}\,\mathbf{1}'\mathbf{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix}$$

$$\mathbf{X} - \frac{1}{n}\,\mathbf{1}\,\mathbf{1}'\mathbf{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

And since,

$$\mathbf{S}_n = \frac{1}{n}\begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_1 & \cdots & x_{n1} - \bar{x}_1 \\ x_{12} - \bar{x}_2 & x_{22} - \bar{x}_2 & \cdots & x_{n2} - \bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} - \bar{x}_p & x_{2p} - \bar{x}_p & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$\times \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$= \frac{1}{n}\left(\mathbf{X} - \frac{1}{n}\,\mathbf{1}\,\mathbf{1}'\mathbf{X}\right)'\left(\mathbf{X} - \frac{1}{n}\,\mathbf{1}\,\mathbf{1}'\mathbf{X}\right) = \frac{1}{n}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\,\mathbf{1}\,\mathbf{1}'\right)\mathbf{X}$$

Appendix C

Suppose that we have a data set of n observations at p-dimensional variables. This data set could be put into a matrix with n rows and p columns.

$$X = \begin{array}{|lllllll|} & \text{Variable 1} & \text{Variable 2} & \dots & \text{Variable k} & \dots & \text{Variable p} \\ \text{Observation 1} & x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ \text{Observation 2} & x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{Observation j} & x_{j1} & x_{j2} & \dots & x_{jk} & \dots & x_{jp} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{Observation n} & x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{array}$$

Let's say that each array represents one variable, in other words, an array represents a column and has n rows. To standardize the columns so that they are comparable and able to interact with each other to create proper distances we can utilize two different algorithms, both with a linear approach.

Algorithm 1 – Minimal Zero

1. Search for the maximum value of each variable
2. Implies the minimum value equal zero to each variable
3. For every variable, divides each observation by the maximum value

   The result will be the Euclidean Distance looking like:

$$d_{ij} = \sqrt{\frac{(x_{i1} - x_{j1})^2}{x_{1max}^2} + \frac{(x_{i2} - x_{j2})^2}{x_{2max}^2} + \dots + \frac{(x_{ip} - x_{jp})^2}{x_{pmax}^2}}$$

Algorithm 2 – Max Min Values

1. Search for the maximum and minimum values of each variable
2. For each variable, subtract for each observation the minimum value
3. For every variable, divides each observation by the difference between maximum value and minimum value

   The result will be the Euclidean Distance looking like:

$$d_{ij} = \sqrt{\frac{(x_{i1} - x_{j1})^2}{(x_{1max} - x_{1\square in})^2} + \frac{(x_{i2} - x_{j2})^2}{(x_{2max} - x_{2min})^2} + \dots + \frac{(x_{ip} - x_{jp})^2}{(x_{pmax} - x_{pmin})^2}}$$

Differences in Algorithm 1 and Algorithm 2

These two algorithms for standardize the data set follow two different approaches of linear transformation of variables. The Figure 0.1 above shows how each algorithm transforms the observations of one variable in order to standardize the data set.

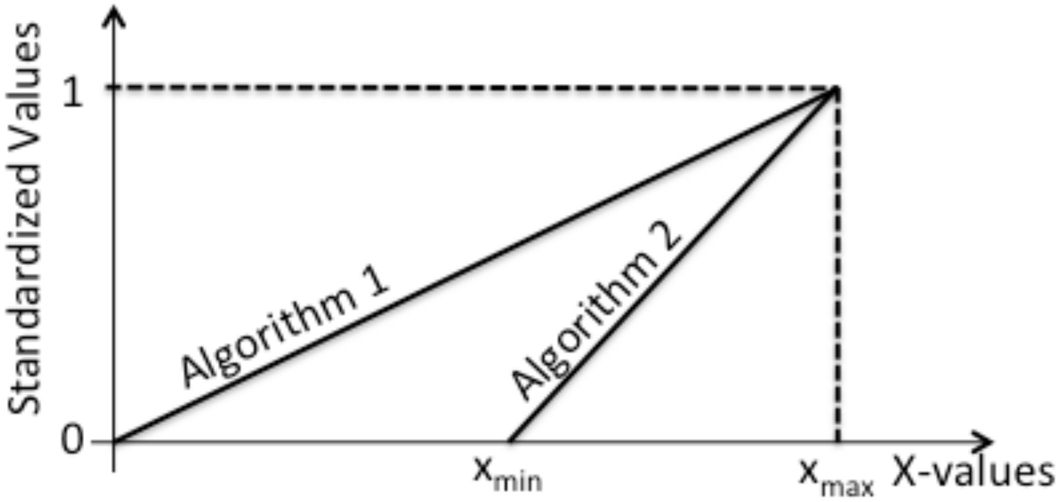


Figure 0.1 – Linear Transformation for the data set.
(Developed by the author)

Moreover, we can say that the Algorithm 2 gives a greater weight for greater distances between variables. By this reason, the distances for the case of Algorithm 2 can be greater than one. This result does not happen if we utilize the Algorithm 1.

Finally, we can say that if we utilize the Algorithm 2, the Minkowski Distance may result redundant, because this method of distance is utilized when we want to give a greater weight for greater distances.

Appendix D

The Multivariate Analysis of Variance (MANOVA) is a method for comparing several multivariate population means. The main objective of MANOVA is to see whether the population mean vectors are different or not, and if they are, which component differ significantly.

There are some assumptions about the data for applying the MANOVA, which are:

I. The random samples from different populations are independent.

II. All populations have a common covariance matrix Σ.

III. Each population is multivariate normal.

The third assumption can be relaxed by appealing the Central Limit Theorem.

The algorithm is very simple, and will be computed at MS Excel®, following Johnson & Wichern (2007). The Table 0.1 above pictures the summary of MANOVA algorithm.

Table 0.1 – MANOVA table for comparing population mean vectors.

Source of Variation	Matrix of sum of squares and cross products (SSP)	Degrees of freedom (d.f.)
Treatment	$\mathbf{B} = \sum_{\ell=1}^{g} n_\ell (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})'$	$g - 1$
Residual (error)	$\mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\bar{x}_{\ell j} - \bar{x}_\ell)(\bar{x}_{\ell j} - \bar{x}_\ell)'$	$\sum_{\ell=1}^{g} n_\ell - g$
Total (corrected for the mean)	$\mathbf{B} + \mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\bar{x}_{\ell j} - \bar{x})(\bar{x}_{\ell j} - \bar{x})'$	$\sum_{\ell=1}^{g} n_\ell - 1$

Source: (Johnson & Wichern, 2007)

The main objective for this algorithm is testing the hypothesis that all mean vectors are the same:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_g$$

For testing this hypothesis we will utilize the Wilk's Lambda parameter:

$$\Lambda^* = \frac{|W|}{|B + W|}$$

With Wilks Lambda we can calculate the Barlett index, which is:

$$-\left(n - 1 - \frac{(p + g)}{2}\right) ln\Lambda^*$$

Finally, we can reject H_0 if we have the following occurrence:

$$-\left(n-1-\frac{(p+g)}{2}\right) \; ln\Lambda^{*} > \; \chi^{2}_{p(g-1)}(\alpha)$$

Appendix E

From P1 to P8 we can see all subsea projects, and from the P9 to P16 all offshore projects. Moreover, PT1 is a subsea project and PT2 is an offshore project.

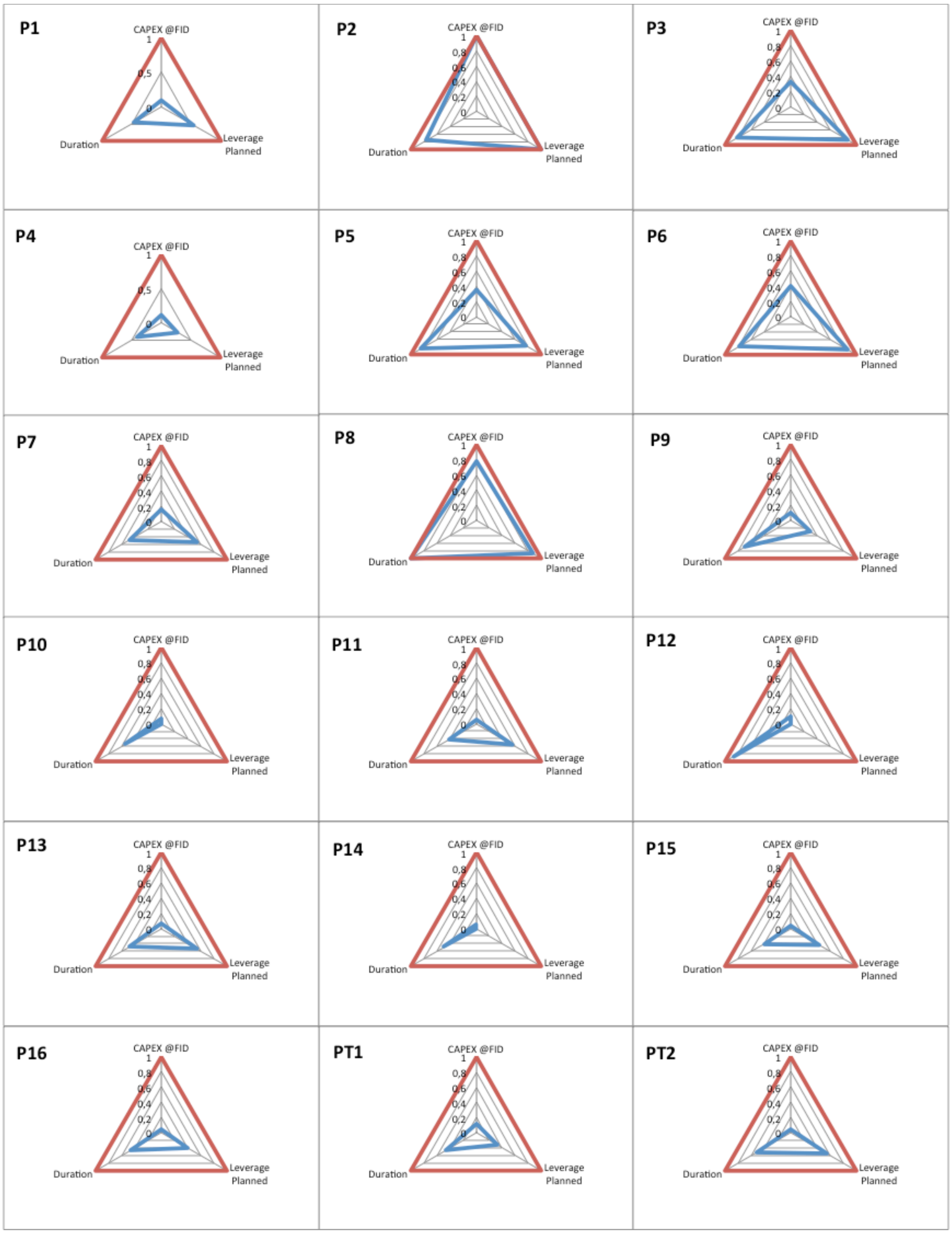


Figure 0.2 – Star representation for all projects, by their planned values

Appendix F

Table 0.2 – Similarities computed for projects' planned values.

	Planned															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1,000	0,375	0,523	0,774	0,536	0,509	0,907	0,391	0,983	0,671	0,927	0,883	0,993	0,642	0,827	0,839
2	0,375	1,000	0,570	0,338	0,555	0,588	0,390	0,902	0,377	0,317	0,364	0,357	0,376	0,310	0,348	0,350
3	0,523	0,570	1,000	0,454	0,956	0,949	0,553	0,607	0,528	0,417	0,503	0,489	0,525	0,405	0,472	0,476
4	0,774	0,338	0,454	1,000	0,464	0,443	0,718	0,351	0,764	0,835	0,825	0,863	0,770	0,790	0,923	0,909
5	0,536	0,555	0,956	0,464	1,000	0,909	0,567	0,591	0,541	0,425	0,514	0,500	0,538	0,413	0,482	0,486
6	0,509	0,588	0,949	0,443	0,909	1,000	0,537	0,628	0,513	0,407	0,489	0,477	0,511	0,397	0,460	0,464
7	0,907	0,390	0,553	0,718	0,567	0,537	1,000	0,407	0,922	0,628	0,847	0,810	0,913	0,603	0,763	0,773
8	0,391	0,902	0,607	0,351	0,591	0,628	0,407	1,000	0,394	0,328	0,379	0,372	0,392	0,321	0,362	0,364
9	0,983	0,377	0,528	0,764	0,541	0,513	0,922	0,394	1,000	0,664	0,913	0,870	0,990	0,635	0,816	0,827
10	0,671	0,317	0,417	0,835	0,425	0,407	0,628	0,328	0,664	1,000	0,709	0,737	0,668	0,937	0,781	0,770
11	0,927	0,364	0,503	0,825	0,514	0,489	0,847	0,379	0,913	0,709	1,000	0,949	0,921	0,677	0,885	0,899
12	0,883	0,357	0,489	0,863	0,500	0,477	0,810	0,372	0,870	0,737	0,949	1,000	0,877	0,702	0,930	0,945
13	0,993	0,376	0,525	0,770	0,538	0,511	0,913	0,392	0,990	0,668	0,921	0,877	1,000	0,640	0,823	0,834
14	0,642	0,310	0,405	0,790	0,413	0,397	0,603	0,321	0,635	0,937	0,677	0,702	0,640	1,000	0,742	0,733
15	0,827	0,348	0,472	0,923	0,482	0,460	0,763	0,362	0,816	0,781	0,885	0,930	0,823	0,742	1,000	0,983
16	0,839	0,350	0,476	0,909	0,486	0,464	0,773	0,364	0,827	0,770	0,899	0,945	0,834	0,733	0,983	1,000

Table 0.3 – Similarities computed for projects' actual values.

	Actual															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1,000	0,387	0,538	0,722	0,693	0,612	0,975	0,493	0,898	0,724	0,972	0,847	0,825	0,555	0,768	0,746
2	0,387	1,000	0,579	0,336	0,467	0,512	0,383	0,642	0,370	0,337	0,382	0,361	0,357	0,295	0,346	0,342
3	0,538	0,579	1,000	0,446	0,707	0,817	0,531	0,854	0,507	0,446	0,530	0,490	0,483	0,376	0,463	0,455
4	0,722	0,336	0,446	1,000	0,547	0,495	0,735	0,414	0,786	0,996	0,737	0,830	0,852	0,706	0,922	0,956
5	0,693	0,467	0,707	0,547	1,000	0,840	0,681	0,630	0,642	0,548	0,679	0,616	0,604	0,445	0,573	0,561
6	0,612	0,512	0,817	0,495	0,840	1,000	0,602	0,717	0,572	0,496	0,601	0,551	0,542	0,410	0,517	0,506
7	0,975	0,383	0,531	0,735	0,681	0,602	1,000	0,487	0,919	0,737	0,996	0,866	0,843	0,563	0,784	0,761
8	0,493	0,642	0,854	0,414	0,630	0,717	0,487	1,000	0,466	0,415	0,486	0,452	0,446	0,353	0,429	0,422
9	0,898	0,370	0,507	0,786	0,642	0,572	0,919	0,466	1,000	0,789	0,922	0,938	0,911	0,593	0,842	0,816
10	0,724	0,337	0,446	0,996	0,548	0,496	0,737	0,415	0,789	1,000	0,739	0,833	0,855	0,704	0,926	0,960
11	0,972	0,382	0,530	0,737	0,679	0,601	0,996	0,486	0,922	0,739	1,000	0,868	0,845	0,564	0,786	0,763
12	0,847	0,361	0,490	0,830	0,616	0,551	0,866	0,452	0,938	0,833	0,868	1,000	0,969	0,617	0,892	0,862
13	0,825	0,357	0,483	0,852	0,604	0,542	0,843	0,446	0,911	0,855	0,845	0,969	1,000	0,629	0,918	0,887
14	0,555	0,295	0,376	0,706	0,445	0,410	0,563	0,353	0,593	0,704	0,564	0,617	0,629	1,000	0,667	0,684
15	0,768	0,346	0,463	0,922	0,573	0,517	0,784	0,429	0,842	0,926	0,786	0,892	0,918	0,667	1,000	0,963
16	0,746	0,342	0,455	0,956	0,561	0,506	0,761	0,422	0,816	0,960	0,763	0,862	0,887	0,684	0,963	1,000

Appendix G

We will present here the details about the results of k-means algorithm.

- There were in total seventeen interactions until arriving to the final result.
- The final error for the planned values was 0,60 and for the actual values 0,77.
- The initial error for planned values was 2,97 and for the actual values was 2,70.
- The clusters configurations for the 10 last interactions were:

Table 0.4 – Evolution of clusters in k-means algorithm for planed values.

Interaction	7	Error	1,94	Interaction	8	Error	1,69	Interaction	9	Error	0,86
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
14	11	3	10	14	11	3	10	14	11	3	10
16	15	7	13	16	15	5	13	16	15	5	13
12	1	5	9	12	1	2	9	12	1	2	9
8		2	4	8	7	6	4		7	6	4
		6								8	
Interaction	10	Error	0,86	Interaction	11	Error	0,77	Interaction	12	Error	0,77
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
14	11	3	10	14	11	3	13	14	11	3	13
16	15	5	13	16	15	5	9	16	15	5	9
12	1	2	9	12	1	2	4	12	1	2	4
	7	6	4	10	7	6		10	7	6	
		8				8				8	
Interaction	13	Error	0,77	Interaction	14	Error	0,73	Interaction	15	Error	0,73
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
14	11	3	13	14	11	3	9	14	11	3	9
16	15	5	9	16	15	5	4	16	15	5	4
12	1	2	4	12	1	2		12	1	2	
10	7	6		10	7	6		10	7	6	
		8			13	8			13	8	
Interaction	16	Error	0,73	Interaction	17	Error	0,60				
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4				
14	11	3	9	14	11	3	9				
16	15	5	4	12	15	5	4				
12	1	2		10	1	2					
10	7	6			7	6					
	13	8			13	8					
					16						

Table 0.5 – Evolution of clusters in k-means algorithm for actual values.

Iteraction	7	Error	1,91	Iteraction	8	Error	1,70	Iteraction	9	Error	1,09
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
8	3	9	11	8	3	9	11	12	3	9	11
12	5	10	15	12	5	10	15	14	5	10	15
14	7	13	1	14	2	13	1	16	2	13	1
16	2	4		16	6	4	7		6	4	7
	6								8		
Iteraction	**10**	**Error**	**1,09**	**Iteraction**	**11**	**Error**	**1,00**	**Iteraction**	**12**	**Error**	**1,00**
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
12	3	9	11	12	3	9	11	12	3	9	11
14	5	10	15	14	5	13	15	14	5	13	15
16	2	13	1	16	2	4	1	16	2	4	1
	6	4	7	10	6		7	10	6		7
	8				8				8		
Iteraction	**13**	**Error**	**1,00**	**Iteraction**	**14**	**Error**	**0,94**	**Iteraction**	**15**	**Error**	**0,94**
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
12	3	9	11	12	3	9	11	12	3	9	11
14	5	13	15	14	5	4	15	14	5	4	15
16	2	4	1	16	2		1	16	2		1
10	6		7	10	6		7	10	6		7
	8				8		13		8		13
Iteraction	**16**	**Error**	**0,94**	**Iteraction**	**17**	**Error**	**0,77**				
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**				
12	3	9	11	12	3	9	11				
14	5	4	15	14	5	4	15				
16	2		1	10	2		1				
10	6		7		6		7				
	8		13		8		13				
							16				

7 Bibliography

Acebes, F., Pajares, J., Galán, J. M., & López-Paredes, A. (2013). Beyond Earned Value Management: a Graphical Framework for Integrated Cost, Schedule and Risk Monitoring. *Procedia - Social and Behavior Sciences , v. 74*, 181-189.

Ahern, T., Leavy, B., & Byrne, P. J. (2014). Knowledge Formation and Learning in the Management of Projects: a Problem Solving Perspective. *International Journal of Project Management* , 1-9.

Alavi, M., & Leidner, D. E. (2001). Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *Mis Quarterly , v. 25*, 107-136.

Alleman, G. B. (2014). *Perfomance-based Project Management: increasing the probability of project success.* USA: Amacom.

Amstrong, M., Galli, A., Bailey, W., & Couët, B. (2004). Incorporating technical uncertainty in real option valuation of oil projects. *Jounal of Petroleum Science & Engineering , v. 44*, 67-82.

Beringer, C., Jonas, D., & Gemünden, H. G. (2012). Establishing Project Portfolio Management: an Exploratory Analysis of the Influence of Internal Stakeholders' Interactions. *Project Management Journal , v. 43*, 16-32.

Cano, J. L., & Lidón, I. (2011). Guided Reflection on Project Definition. *International Journal of Project Management , v. 29*, 525-536.

Caron, F. (2009). *Gestione dei Grandi Progetti di Ingegneria: il project management in azione.* Novara, Italy: isedi.

Chou, J.-S., & Yang, J.-G. (2012). Project Management Knowledge and Effects on Construction Project Outcomes: an Empirical Study. *Project Management Journal , v. 43*, 47-67.

Costa Lima, G. A., & Suslick, S. B. (2006). Estimation of volatility of selected oil production projects. *Journal of Petroleum Science & Engineering , v. 54*, 129-139.

Costa Neto, P. L. (2002). *Estatística* (2nd Edition ed.). São Paulo: Edgard Blucher.

Cserháti, G., & Szabó, L. (2013). The Relationship between Success Criteria and Success Factors in Organisational Event Projects. *Internationa Journal of Project Management* , 1-12.

Hartigan, J. A. (1975). *Clustering Algorithms.* USA: Wiley-Interscience.

Ika, L. A., Diallo, A., & Thuillier, D. (2012). Critical Success Factors for World Bank Projects: an empirical investigation. *International Journal of Project Management , v. 30*, 105-116.

Jafarizadeh, B. (2010). Financial factor models for correlated inputs in the simulation of project cash flows. *Journal of Petroleum Science & Engineering , v. 75*, 54-57.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data.* New Jersey: Prentice-Hall Inc.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: a Review. *ACM Computing Surveys , Vol. 3*, 264-323.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (Sixth Edition ed.). New Jearsey, USA: Pearson.

Kim, E. H., Wells Jr., W. G., & Duffey, M. R. (2003). A Model for Effective Implementation of Earned Value Management Methodology. *International Journal of Project Management , v. 21*, 375-382.

Kwak, Y. H., & Anbari, F. T. (2011). History, Practices and Future of Earned Value Management in Government: Perspectives from NASA. *Project Management Journal , Vol. 43*, 77-90.

Levine, H. A. (2005). *Project Portfolio Management: a practical guide to selecting projects, managing portfolios, and maximizing benefits.* San Francisco, USA: Jossey-Bass.

Manton, K. G., Lowrimore, G., Yashin, A., & Kovtun, M. (2005). Cluster Analysis: Overview. In B. Everitt, & D. Howell, *Encyclopedia of Statistics in Behavioral Science* (Vol. Vol. 1, pp. 305-315). Wiley.

Merli, A. (2010). *I Giudizi Soggetivi nel Calcolo delle Stime a Finire di un Progetto: l'applicazione della statistica bayesiana all'Earned Value Management.* Politecnico di Milano, Ingegneria Gestionale, Milan.

Project Management Institute. (2014). *Pulse of the Profession: in-depth report.* Pennsylvania: Project Management Institute.

Reich, B. H., Gemino, A., & Sauer, C. (2013). How Knowledge Management Impacts Performance in Projects: an Empirical Study. *International Journal of Project Management* , 1-13.

Samset, K. (1998). *Project Management in a high-uncertainty situation: Uncertainty, Risk and Project Management in International Developments Projects.* Ph. D. Dissertation, Norwegian University of Science and Technology, Faculty of Civil and Environmental Engineering.

Shenhar, A. J., & Dvir, D. (2007). What makes a Project Successful. In *Reinventing Project Management: The Diamond Approach to Successful Growth and Innovation* (pp. 1 - 12). Boston, MA: Havard Business Press.

Sivagnanasundaram, N., Chaparro, A., & Palmer, E. (2013). Evaluation of the Presence of a Face Search Advantage in Chernoff Faces. *Human Factors and Ergonomics Society Annual Meetings. v. 57*, pp. 1609-1614. San Diego: Sage.

Westbrook, K. W., & Peterson, R. M. (1998). Business-to-Business Selling Determinants of Quality. *Industrial Marketing Management , v. 27*, 51-62.

Wideman, R. M. (2004). *A Management Framework: for project, program and portfolio integration.* Trafford.

Yu, J.-H., & Kwon, H.-R. (2011). Critical Success Factors for Urban Regeneration Projects in Korea. *International Journal of Project Management , v. 29*, 889-899.