University of Alberta

Analysis of Industrial Construction activities using Knowledge Discovery Techniques

by

Carlos Vicente Gonzalez-Villalobos

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Science

In

Construction Engineering and Management

Department of Civil and Environmental Engineering

Fall, 2011

Edmonton, Alberta

To Veronica,

The love of my life

My wife and best friend

For your support, advice and care

I will always be grateful

We did this together!

Abstract

An industrial construction enterprise operating in the City of Edmonton wants to improve its bidding strategies that are currently plagued with uncertainty, lack of information and historical price variability. The present research studies a compilation of documents obtained from company archives detailing previous pipe fabrication performances in order to improve and support decisions during tendering. Two stages constitute the present study: 1) the creation of industrial indicators accumulating company data into a single source facilitating past event consultations; and 2) analysis of data comprising these indicators using a combination of *Clustering, Association Rules* and *Distribution Fitting* techniques designed to detect embedded trends and arrangements, enhancing previous performance comprehension. Results obtained in this research ranged from industrial indicator creation to constitution of multiple project profiles, project characterization and statistical distribution fitting reflecting different fabrication aspects and historical knowledge present in previous pipe module fabrication projects.

Acknowledgements

First of all, I would like to thank God for guiding and giving me the opportunity of discovering new facets of my professional life during these years. Definitively this was a life-changing experience.

Second, my respect and admiration goes to Dr. Yasser Mohamed, who in March 2009 offered me the opportunity of continuing my education and working with him. I consider Dr. Mohamed my friend and guide. I thank him for all the opportunities given to me during my graduate years in the University of Alberta's prestigious Hole School of Construction.

I would like to thank the company that supported this study: KBR Canada Ltd. Thanks to Mr. Rishad Hakkim and Mr. Hal Williams for their patience and opening their minds to alternative analysis approaches designed to improve Industrial Operations. Their contribution represents a foundation of all studies present in this Thesis.

Moreover, I want to thank my office mates. They shared their thoughts, experiences and time with me during the last couple of years. Thanks to Mr. Di Hu, Ms. Farzaneh Saba and Mr. Amr El Nimr. We came from different parts of the world, but our time together has proven that we have many things in common. Thank you all for receiving me as one of your own. I will always be grateful for that.

I would like to recognize and dedicate this work to my beloved wife, Veronica Bohorquez. She has given me unconditional love, support and strength at all times. Thank you very much for being there always.

Finally, I would like to thank my parents Mrs. Gladys Villalobos and Mr. Romer Gonzalez. Thank you for all your advice and being there without caring about distances and time zones. Your effort and dedication in this experience are something I will remember.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Amt/Hr	*Amount of executed quantities per hour*
ARFF	*Attribute Relationship File Format*
CERP	*Construction Enterprise Resource Planning System*
CS %	*Carbon Steel percentage*
DI	*Diameter Inches*
DIW	*Diameter Inches Weld*
DSS	*Decision Support System*
ESP	*Estimating Summarizing Program*
ExtraHeavyPct	*Extra-Heavy Steel percentage*
HeavyPct	*Heavy Steel percentage*
Hr/Ft	*Hours per foot*
Hr/Qty	*Amount of hours per executed Quantity*
JTD	*Job to Date quantities*
LengthPipeFt	*Total length of pipe in feet*
LengthPipeM	*Total length of pipe in meters*
LightPct	*Light Steel percentage*
MediumPct	*Medium Steel percentage*
MH	*Manhours involved in an activity*
Mh/DIW	*Manhours per Diameter Inch Weld*
MhrsperDI	*Manhours per Diameter Inch*
MiscPct	*Miscellaneous Steel percentage*

Mod	*Module*
OLAP	*Online Analytical Process*
PCMS	*Project Cost Management System*
PNN	*Probabilistic Neural Networks*
SizeDesc	*Size of an element in inches*
WCB	*Work content budget quantities*
WEKA	*Waikato Environment for Knowledge Analysis*

Chapter 1: Introduction

1.1. Problem Statement

An industrial construction enterprise with base operations in the City of Edmonton wants to improve its bidding strategies. Essentially, current company decisions during development of bidding proposals are based on a combination of in-house standard factors and records from past experiences. During this process, special percentages known as discount/adders factors are adjusted. To do this, the company relies on information that may not totally reflect actual operations. Influential aspects such as disorganization during project documentation and lack of records related to fabrication operations can negatively affect decisions, representing a risk during determination of an activity's final price.

The company is looking to explore and analyze its historical data to augment its inner-operations knowledge. Increasing it not only represents a direct benefit but also improves the company's own tendering processes and enhances its competitiveness.

The scope of this research involves a collection of records from different departmental areas of the company and its past projects, containing data from previous pipe module fabrication assignments. Capturing, cleaning and organizing this vital information represent a challenge. Finding coherency between different departmental records adds difficulty to the decision making process, due to a combination of instability and lack of criteria present in historical data.

The desired outcome is to observe previous project trends and compare them with new ones generated by potential projects. By doing this, decisions will be supported by past documented experiences, determining new factors affecting estimation.

Positive results can be achieved through documenting and organizing knowledge, thereby producing a solid foundation to eventually guide a company through many decisions in the immediate future.

1.2. Research Objectives

In order to review historical data to determine a solid knowledge base of previous performances to be used by the company, three objectives need to be defined:

1. Create a set of Industrial Indicators reflecting past operations.
2. Consolidate data from different departmental areas of the company related to these indicators into a single source to facilitate understanding of previous project performances.
3. Analyze data constituting these Industrial Indicators, to detect embedded trends and arrangements in previous performance records, enhancing project comprehension.

1.3. Thesis Organization

Chapter No. 2 includes a detailed literature review of concepts relevant to the present research. *Industrial Construction, Construction Project Bidding, Data Warehousing and Data Mining* concepts are reviewed in this chapter. It serves as knowledge base for analysis techniques applied during the course of the present Thesis.

Chapter No. 3 introduces an overview of different company departments and some of the documents produced by each operational area. It also presents a description of all projects constituting the scope of this research and their distinct characteristics. Furthermore, a research process performed during visits to this company is depicted in conjunction with alternative Industrial Indicators.

Chapter No. 4 explores five alternatives specifically designed to analyze data from departmental records, using a combination of *Clustering, Association Rules* techniques and *Distribution Fitting*. In addition, alternative sources of information are analyzed, enhancing research robustness and the understanding of fabrication operations.

Chapter No. 5 contains Conclusions, Recommendations and Research Contributions to the study of industrial construction activities, specifically pipe module fabrication processes.

Appendices can be found at the end of this Thesis. All remaining charts, tables and other types of graphs that were removed from previous chapters due to space restrictions are presented in this section.

Chapter 2: Literature Review

2.1. Introduction

The subsequent chapter is structured in three main sections: *Industrial Construction*, *Construction Project Bidding* and *Data Mining*. Research work performed by different authors in each of these areas has been reviewed.

Firstly, *Section 2.2* presents *Industrial Construction* concepts. Through this section, a definition of this construction discipline is presented. Secondly, *Section 2.3* introduces different research associated with *Construction Project Bidding* and certain elements potentially influencing bidding decisions.

Thirdly, *Section 2.4* describes *Data Warehousing* by elucidating support and contribution of information sources within an industrial construction company. This section presents an outline on *data handling and pre-processing*. Fourthly, *Section 2.5* discusses certain applications of *Data Mining* in construction and their different outcomes. This section includes a literature review on *Data Mining* techniques applied in this study: *Clustering* and *Association Rules*. Main concepts related to these techniques are explained. Lastly, *Conclusions* are presented at the end of this chapter.

2.2. Industrial Construction

Industrial construction represents a branch of the Construction Industry that involves a mixture of specialized activities, characterized by its unique techniques and processes. This area of construction is often used in major-sized projects, such as today's industrial facilities.

Furthermore, in this engineering area, modular construction is engaged. This construction technique generates large industrial facilities such as oil processing plants and refineries built with single entities comprised by pipe spool modules (Mohamed et al., 2007). In addition, one of the highest degrees of complexity within construction disciplines is found in industrial construction (Wang et al., 2009).

2.2.1. Pipe Spool and Steel Fabrication

Pipe spool modules are structures usually formed by different arrangements of equipment, pipes, steel structures and other assorted elements (Mohamed et al., 2007). Their fabrication is characterized by presenting multiple stages (Sadeghi & Fayek, 2008), (Song et al., 2009):

- The module construction process begins with the production of shop drawings, which will include spool and steel production data. This process is known as drafting.
- Pipe Spools are generated in a fabrication shop, using industrial techniques of cutting, welding and fitting of pipes. The processed pipes will constitute spools made of different materials, weights and lengths. Furthermore, steel members are also produced to be used as structural elements according to specifications.

- Fabricated spools are transported to the module yard and grouped with steel structures that have been previously erected. These items will be joined sequentially, initially assembling steel structures to be superseded by the installation of pipe spools and equipments, forming decks. Combination of assembled decks constitutes a complete pipe module.
- Once construction of pipe modules is completed, these are transported and later installed on-site.

Figure 2-1 depicts the fabrication process of a pipe module (Sadeghi & Fayek, 2008), (Song et al., 2009):

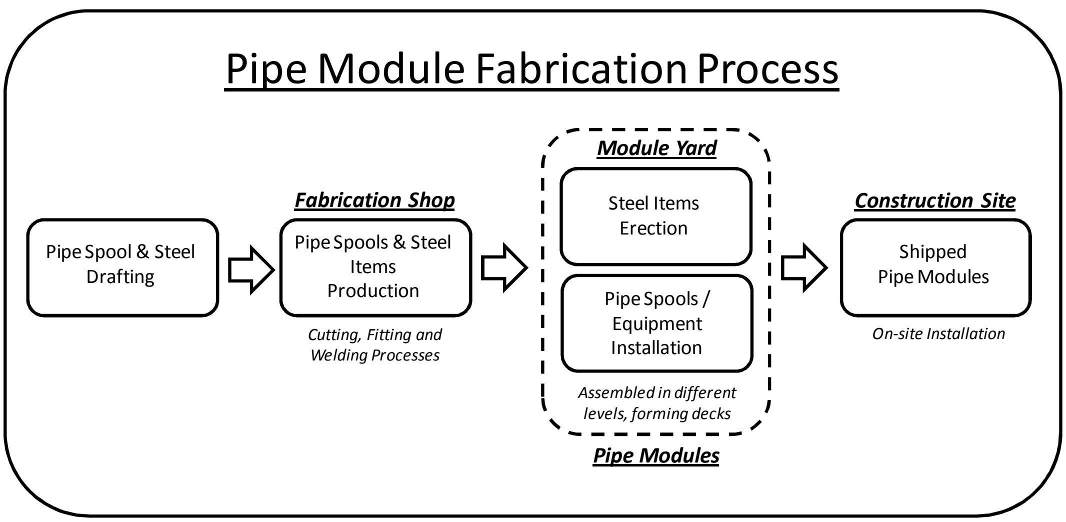

Figure 2-1: Pipe Module Fabrication process (Sadeghi & Fayek, 2008), (Song et al., 2009)

Due to diverse features of each produced item in a fabrication shop, precise construction methods are demanded (Sadeghi & Fayek, 2008). As a consequence of this, distinctive modules are created in terms of composition and enclosed arrangements (Mohamed et al., 2007).

Modular construction has proven to be a technique that efficiently reduces costs and is frequently used in industrial construction sites located outside urban areas (Taghaddos et al., 2010).

2.3. Construction Project Bidding

With the use of tendering, innumerable tasks have been and are being acquired worldwide (Seydel, 2003).

When determining bidding prices, a correct estimation of an industrial construction project represents a highly intensive and time-demanding phase. Diverse factors could potentially impact an estimator's perspective about a determined production rate, for example weather impact and workforce expertise availability (AbouRizk et al., 2001).

Halpin & Woodhead (1998) mentioned unit pricing and resource enumeration as two common approaches used to evaluate construction costs. More to the point, there are other existing factors that can affect a bid's final price (Flanagan & Norman, 1985):

- The quality of tendering data. With a greater quality of information, an optimal final bid price tends to be lower.
- Competence of other companies participating in the bid. Experienced competition likely diminishes bid prices.

Flanagan & Norman (1985) mentioned in their research that contractors' knowledge of their own processes (effectiveness and familiarity with construction techniques) has greater influence in bidding prices than tactically made decisions against competitors during tendering.

Experience represents an important influence while developing bidding proposals. According to Seydel (2003), in single auctions the main difficulty in bidding lies in establishing the most favorable gain for a company while presenting a lower tender to achieve it. To overcome this, experience is one important feature to acquire for any construction company. Fayek (1998) stated experience and instinct support numerous conclusions concerning calculation of a final bid's price.

Fan et al. (2007) expressed in their research that a combination of long practical expertise and skilled instruction are unique factors needed to build experience in decision makers, either from company managerial or operative areas. However, this research presents an additional alternative to build experience in a construction company. Experience can be stored in documented events translating situations lived during a particular endeavor especially when reviewing previous performances of past projects.

Enhancing a company's knowledge about its inner operations represents an opportunity for increasing its competitiveness. Through continuous learning, auditing and improvement of construction operations a company will reduce uncertainty in its tenders by having at hand previous patterns and action paths to consider during current bidding processes.

Different approaches have been performed by researchers in the field of construction bidding and mark-up estimation:

- Dozzi et al. (1996) developed a multi-criteria utility theory model, created in Visual Basic® programming language. It included 21 different bidding arguments, classified in a hierarchal arrangement. Using utility theory functions, the model was capable of calculating bidding mark-ups incorporating different types of individual and assessable data through the entire decision-making process.
- Fayek (1998) created a tendering approach model to facilitate margin size calculation, to fulfill construction companies' requisites during tendering. This model used Fuzzy Set Theory concepts to determine suitable mark-up boundaries.
- Chua & Li (2000) designed a multi-attribute bid reasoning model that comprised four different contractor judgments during bidding: rivalry, company's situation during bidding, risk scope and urge for work. The authors also created hierarchic configurations for each judgment to identify the factors with greater control in bid decisions.
- Christodoulou (2000) created a probabilistic neural network (PNN) model that merges qualitative/quantitative figures with factual information. This

model adjusts to a specific margin, consequently determining likelihood of success of a particular bid.

These applications considered internal and external factors influencing a construction company's bidding decisions. Criteria, need of work, resource availability and risk represented some factors explored by researchers. Nonetheless, the main focus of this research involves historical records obtained from an industrial construction company. Characteristics of past projects and their resemblance to conditions found in potential projects will dictate bidding decisions during estimation processes.

Construction companies need to have as a main priority reorganizing and documenting previous project data before participating in new tendering opportunities. Proposal Managers require the best data available to execute proper decisions adjusted to company's requirements by creating and consolidating a good knowledge base to analyze incoming projects, constituting a guideline which will guarantee success in the development of solid bidding proposals.

2.4. Data Warehousing

Construction projects normally comprehend several parties coordinated by trading copious quantities of single documents (Zhiliang et al., 2005). Although most systems do not reach absolute synergy, information technologies commonly sustain the greater part of construction operations (Bouchlaghem et al., 2004).

One of the techniques used to handle information derived from multiple areas is Data Warehousing. Ahmad et al. (2004) defined its concept as the result of incorporating several information sources into a single database, capable of handling and manipulating diverse quantities of data.

Azhar et al. (2010) mentioned in their research additional applications of a Data Warehouse. According to these authors, a Data Warehouse can be applied to store and analyze historical information, arrange data in a constant form and consequently simplify its usage and manipulation by a user.

Data Warehousing is an appropriate technique for those activities requiring information synchronization between managerial areas. Some of its applications are reviewed by the following researchers:

- Zhang (2010) recommended the creation of a platform to administer logistics information. This system is essentially assembled by a combination of two elements. First, metadata is used to implement data organization in documentation processes. Second, an Oracle database is employed to store information from diverse areas. Within this database, a Data Warehouse is found. The proposed combination enhances data review by homologizing its input into a Data Warehouse and enabling its analysis through *Data Mining* and other applications, properly sustaining decisions in transport operations.

- Ahmad et al. (2004) applied Data Warehousing techniques in the creation of a Decision Support System (DSS) in residential construction. This system collected information from different sources and formats, stored its data in a Data Warehouse and manipulated its output using online analytical process (OLAP). The purpose of this system was to provide additional support to construction companies interested in choosing potential sites for future developments.
- Shi and Halpin (2003) stated in their research the significance of creating a computer-integrated system to manage construction company resources. The authors recommended a system that combines company data stored in a Data Warehouse, process models containing managerial duties and input provided by classified user interfaces (according to their functional areas), creating a system known as Construction Enterprise Resource Planning System (CERP).

Due to the nature of the construction business, individual decision making approaches generated in environments with numerous and diverse types of information represents an unsuitable solution (Shi and Halpin, 2003). A unique informational display containing trustworthy information can be generated by Data Warehousing, in which construction company data is integrated from distinct functional areas (Ahmad, 2000).

2.5. Data Mining

Data Mining is a practice dedicated to the withdrawal of knowledge from records. The knowledge obtained is generally substantially beneficial but unfamiliar (Witten & Frank, 2005). As Soibelman & Kim (2002) mention, the purpose of their application of *Data Mining* in construction activities is to produce a tool that extracts arrangements, elucidating and forecasting trends within these ventures. The product of evidence extraction is converted, creating facts. Afterwards, those facts are transformed into knowledge (Witten & Frank, 2005).

Data Mining involves heavy computing using specific algorithms which can vary depending on the problem to be analyzed. In some scenarios, data can be scattered and incomplete. To overcome this, *Data Mining* can establish correlations between different items to properly classify them and create knowledge.

In construction, knowledge derived from *Data Mining* can become a vital decision-support tool to construction managers. Its application can improve project understanding. For example, with a broader project comprehension, managers will be able to evade potential issues during project building by making correct decisions (Soibelman & Kim, 2000). In addition, *Data Mining* applications represent an exceptional answer to the problem of studying complex information present in construction records (Buchheit et al., 2000).

Some of the algorithms built in *Data Mining* are classified by their outcomes (Lee et al., 2008). These are the following:

- Classification
- Clustering
- Association Rules

As Zhang et al. (2004) discussed in their study, *Data Mining* presents two main duties: first, it can describe generalities and gather hidden rules from figures; secondly, depending on data and required specifications, *Data Mining* can anticipate significant outcomes.

Fan (2007) declared *Data Mining* models hold two distinctive characteristics are build and processed mainly from acquired data through specialized algorithms and are capable of obtaining outcomes in those situations in which the complexity of information is problematic for alternative analysis approaches. According to the author, support for these models is skewed towards assumed details in data instead of opinions or professional expertise.

In today's world, data mining has been used in different areas as a support tool, motivated by the need of discovering significant arrangements in factual sets of data (Soibelman & Kim, 2002)

Several authors have researched different applications of data mining in the construction industry:

- Fan et al. (2008) modeled an auto-regressive tree to forecast residual value of construction equipment. Their research focused on wheel loader's residual values. Data was collected from different sources including an online database (LastBid® software) and other historical information sources such as U.S. Bureau of Economics and Statistics Canada. Their model detected relationships between records efficiently calculating as output the residual value of a wheel loader with unique characteristics. This model offered a visual interface to the user, in which parameters and analysis results of historical data could be observed.
- Song and AbouRizk (2008) suggested a new modeling method to determine productivity, gathering information from past projects and creating new productivity models based on previous experiences. According to the authors, a consistent estimate can be derived from a company's historical records. In addition, they stated the significance of historical records in a company's potential ventures, because of their relevant forecasting content. In this research, an artificial neural network (ANN) was the tool selected to deal with large amounts of data coming from steel drafting and fabrication tasks.
- Lee et al. (2008) applied data mining techniques to generate knowledge from service records of a construction enterprise. The author's obtained 7790 different cases collected from the maintenance department of the company. Decision-tree analysis was performed to identify elements that

could be generating concrete cracks in structural elements of high-rise buildings. This analysis obtained positive outcomes identifying potential causes for this issue.

- Hammad (2009) proposed a framework to improve labor asset management. In addition, a feature produced by this study was a Data Warehouse conception to properly manage company information. Furthermore, *Data Mining* techniques were applied to historical data derived from this archetype warehouse. Unit costs, resource requirements and durations were determined for three diverse case studies, representing knowledge to be used in future scenarios.

Science, farming and health have been different areas in which machine learning through *Data Mining* has been applied in the past. Its effect in increasing business knowledge has been positive, resulting in many people considering its use (Witten & Frank, 2005). Furthermore, as Hammad (2009) expressed in his study, managers will become more interested in *Data Mining* if the cost of its application in industrial construction is minimal as it generates improved productivities, accentuating its expenditure compensation.

2.5.1. Clustering

Clustering is a *Data Mining* method that has the capability of grouping records by their resemblance: It magnifies both agglomeration of comparable records into groups and diversification of such groups by their distinctness (Foss & Zaïane, 2002). Ankerst et al. (1999) portrayed clustering as one of the main techniques to analyze databases. Likewise, these authors mentioned two different applications of clustering in *Data Mining*: Firstly, it is a technique suitable to initially prepare a set of data into different groups to apply further analysis using additional *Data Mining* algorithms. Secondly, as another application it can be implemented independently to obtain an overview of a data set's composition.

To analyze pipe module fabrication data obtained in this study and implement *Data Mining*, two different clustering algorithms have been selected due to their efficiency and reliability*: K-Means* and *DBScan algorithms*. A literature review of both techniques of analysis is described in the following sections. Their application in four different *Case Studies* is in *Chapter No. 4*.

2.5.1.1. Clustering Algorithm: K-Means

Foss & Zaïane (2002) mentioned a well-known approach used to divide data: *K-Means algorithm*, in which clusters of records are symbolized by their respective centroids. In addition, in the survey paper by Wu et al. (2007), *K-Means algorithm* is chosen as one of the 10 most important algorithms in *Data Mining*. MacQueen (1967), detailed in his work the procedure followed by *K-Means algorithm*:

- First, *k* groups are initially established. These groups are formed by unique arbitrary points.

- Secondly, new points are attached to those groups with means closer to the value of the new point.
- Thirdly, to consider a new point included, the group's mean is adapted.

However, *K-Means algorithm* presents two disadvantages (Pavan et al., 2010):

a. Initial quantities of *k* groups specified by a user prior cluster determination.
b. Primary seeds choice is arbitrary.

Extended iteration cycles and deficient outcomes can be generated by unsuitable selection of cluster amounts (Pham et al., 2005) and primary seeds.

2.5.1.2. Clustering Algorithm: DBScan

Density Based Spatial Clustering of Applications with Noise or *DBScan* is an algorithm with the capability of detecting randomly-shaped clusters (Ester et al., 1996). This method is capable of dealing with noise in data (Rehman & Mehdi, 2006). Moreover, in this algorithm the determination of a single point's density is estimated through detection of items in its vicinities, delimited by a radius (Ertöz et al., 2003). The same authors defined three point types estimated during the implementation of the *DBScan algorithm*:

- *Core Points* are those characterized by having a density higher than initially established. These points remain inside the area correlated to a point.

- *Non-Core Points* are those points lacking related *Core Points.* These are also known as *Noise Points*.
- *Non-Noise, Non-Core Points* are points implanted to those clusters present in their vicinities. These are also identified as *Border Points*.

DBScan is a density-based algorithm that presents the benefits of not requiring an initial input for number of clusters and also detects randomly-shaped arrays (Zaïane et al., 2002). Nevertheless according to researchers, one of the downsides of the *DBScan algorithm* is observed when cluster densities are drastically fluctuating: it does not accurately perform cluster detection (Zaïane et al. 2002, Ertöz et al. 2003).

2.5.2. Association Rules

Witten & Frank (2005) defined *Association Rules* as a *Data Mining* technique similar to *Classification* algorithms. In addition, these authors highlighted particular differences between these methods: one of them lies in the capacity of *Association Rules* to predict attributes and their multiple permutations within a dataset without regarding sample sizes. Furthermore, the same researchers stated that detection of congruity values governing a data set is another function of *Association Rules*. The distinct results can be obtained through their application on a data set and only those rules covering a larger amount of records with greater precision need to be considered for analysis.

In order to define *Association Rules*, two factors must be taken into consideration (Witten & Frank, 2005):

- *Support:* Indicates the number of occurrences that are accurately covered by the requirements of a particular rule.
- *Confidence:* States the amount of occurrences appropriately foreseen under a rule from all attributes detected. It can be represented by a percentage, indicating rule precision.

To evaluate relationships between fabrication attributes and review the effect of their diverse arrangements in a set of data, two different *Association Rules* have been selected: *Apriori* and *Predictive Apriori* algorithms. These techniques are applied during *Case Study No. 3* and *Case Study No. 4* and are explained thoroughly in *Chapter No. 4* of the present thesis.

These algorithms are used as a secondary method for analysis of pipe fabrication data. Once a large group of records has been extracted using *Clustering* algorithms, *Association Rules* take effect. This improves analysis of data obtained from pipe spool fabrication operations and serves as base for chart development, creating a data display, improving results visualization and benefitting decision makers in the Company.

2.5.2.1. Apriori Algorithm

According to Agrawal & Srikant (1994), this association algorithm performs the following tasks:

- Firstly, minimum support and confidence measures must be defined prior algorithm application. This establishes a limit that will be imposed on the algorithm in order to purge all not complying rules.
- Secondly, the algorithm performs an initial pass into the database detecting attribute frequencies, generating a large list containing all items and their respective occurrences.
- Thirdly, this algorithm performs additional passes searching for all existing attribute combinations, according to initial conditions.
- Fourthly, once these combinations are found, a list of attributes with their different *support* and *confidence* values will be calculated and displayed.

2.5.2.2. Predictive Apriori Algorithm

This algorithm is similar to *Apriori* in terms of rule generation. Nevertheless, Scheffer (2004) highlights in his research some qualities of *Predictive Apriori algorithm*:

- *Predictive Apriori* is a method that attempts to discover those association rules considered best among all present in a set of records.
- It does not involve a user establishing either minimum S*upport* or *Confidence* values. Instead, these measures are automatically determined.

- *Support* and *Confidence* measures are entered in an equation within this algorithm calculating *predictive accuracy* values.
- One single requirement that has to be inputted by the user: total number of rules to be generated.
- As a final product of analysis, each of this algorithm's rules will present maximum *predictive accuracy* values.

2.6. Conclusions

In *Chapter No. 2,* concepts and perceptions from different authors in *Industrial Construction* have been presented. In addition, some interesting applications of *Data Mining* techniques in the construction area have been reviewed.

Professional experience is the main topic mentioned by referred authors when developing applications to improve the quality of bid decisions. Some researchers integrated experience with techniques such as Utility Theory Models, Fuzzy Set Theory concepts, Contractor Judgments and Probalistic Neural Networks (PNN). On the other hand, when observing applications of *Data Mining* in consulted works, a common element was noticed: historical information. This was present during the entire development of decision-support tools.

Even though these studies and their applications proved to generate knowledge in their respective areas, their utilization was focused to forecast residual equipment values, determining productivity, event investigation and resource allocation respectively. None of the consulted literature was dedicated to project profile determination to support bidding decisions. Furthermore, no studies dealt with project characterization using fabrication quantities of industrial fabrication activities, representing an area of opportunity for analysis using *Clustering* and *Association Rules* algorithms.

Using these *Data Mining* techniques it is possible to identify and classify new information derived from historical records. This represents an interesting area to explore: comparing data from a potential projects versus historical records can enhance a construction company's competitiveness during bidding, by having knowledge of its previous performances at hand. It can also enable a construction company to establish pricing of potential projects with a starting baseline supported by profile comparison and identification of average characteristics within a project.

Chapter 3: Creating Industrial Indicators

3.1. Introduction

An industrial construction enterprise has its base of operations in the City of Edmonton. This company has more than thirty years of experience in the construction industry with an important level of participation in oil-related projects, specializing in pipe module fabrication.

Fabrication shop and module yard are facilities specifically designed for the execution and progress of this type of industrial construction. Within the fabrication shop activities such as cutting, fitting and welding of different pipe diameters and materials are performed continuously as demanded. In the module yard, located next to the fabrication shop, assembly and erection of steel structures is executed progressively. These structures will hold in place diverse arrangements of pipes and equipments. This chain of events occurs on an uninterrupted basis during the duration of a project. It represents multi-million dollar activities that will permit construction of industrial facilities using modular construction techniques.

The Company wants to improve its current bidding strategies through analysis of previous operations. It has chosen a sample of the five best construction projects performed within the last 10 years, according to certain parameters:

- Quality of the information (complete and reliable data).
- Performance during construction.

- Obtained profits.

These projects will represent a benchmark used by the Company to compare performances and make decisions regarding final bidding prices in future tendering processes. A description of analysis performed during the last eight months for this corporation is presented in corresponding sections:

Section 3.2 describes *Actual Scenario* and *Current Practices* of different departments in charge of elaborating bidding proposals. Duties, documentation used and responsibilities of each area are explained.

Section 3.3 presents an overview of the projects integrating the *Research Scope*. Five different industrial construction projects are described throughout this section. Due to confidentiality reasons, all graphs, figures and tables present in this chapter have been scaled.

Section 3.4 details diverse *Information Sources* and documentation provided supporting development of Industrial Indicators detailed in this chapter.

Section 3.5 explains *Data Gathering* and *Cleaning* approaches used in acquired data to properly apply *Data Mining* techniques for further analysis.

Section 3.6 introduces *Industrial Indicators* elaborated during this research to review past performances of the company. These become visualization tools to improve analysis of specific characteristics between industrial construction projects.

Section 3.7 displays the *Research Process* used to review results derived from application of Industrial Indicators.

Section 3.8 presents *Conclusions* about explored topics in this chapter.

3.2. Actual Scenario

Personnel from different departments of the company schedule meetings to discuss bidding strategies in their proposals. These teams include multiple individuals from the *Project Controls*, *Project Estimation* and *Project Execution* departments. Previous and current project performances, market demand, competition and current workload (both for fabrication shop and module yard) are discussed and reviewed. During these sessions each department provides different insights concerning fabrication activities.

3.2.1. Project Controls Department

The Project Controls department collects actual quantities performed in each project. It is the department responsible to account for all projects being carried out. One of its duties is to gather factual data in a report system known as Project Cost Management System (PCMS).

PCMS collects data from projects and it is in complete synergy with the company accounting. Furthermore, reports produced by this system include information from different project areas such as direct and indirect labor, construction management personnel and other overhead costs related to fabrication activities.

In addition, Work Content Budget (WCB) and Job to Date (JTD) quantities are presented for these operational areas in .pdf reports, developed on a monthly basis accumulating data from both fabrication shop and module yard. A high-level view of project information is introduced in PCMS reports. Fabrication activities are depicted generally. Single tasks are grouped into larger categories. Measures of quantities performed and man-hours invested in different activities are presented using ranges (e.g. pipe welding 2.5” to 10”).

This level of perspective does not support deep analysis of fabrication activities. These reports are not suitable for those cases in which productivities of single activities and their behaviors in different projects are reviewed and compared in a detailed way.

Finding out which unit type was measured in an activity represents a challenge when reviewing PCMS reports. In these documents, quantities are presented as general figures (unit values) without distinction (e.g. quantities for pipe cutting activities are presented as unit values, rather than linear feet or linear meters). Only personnel related to these reports know which unit type was measured. This increases the risk of comparing two distinct registered measures when analyzing project behaviors, causing irregularities and questionable results.

In some occasions, uncertainty was present in company personnel at the moment of reviewing factual data. Sometimes people from this department were not sure about which unit type was measured for certain activities at particular times.

Alternative reports generated by *Project Controls* are spreadsheets derived from a FoxPro® database managing system, known as Quantity Takeoffs. These are spreadsheets that specify all items involved in fabrication of pipe modules. In these documents, items are classified by activity (piping, welding, handling, hydrotesting, etc), item type (pipes, valves, bolt-ups and supports) and can contain additional measures such as unit price, man-hours invested in a particular activity and diameter inches weld (DIW).

Information related to man-hours was frequently not present in these reports. Estimated man-hours were registered in some fabrication activities. No record of actual man-hours is present in these files. However, Quantity Takeoffs reports provide more robust data related to fabrication activities than PCMS reports. These reports contain a higher level of detail for analysis of specific tasks.

3.2.2. Estimation Department

The *Estimation* department gathers all information related to project costs in another database system that has a spreadsheet designed to facilitate input and output of information and prepare estimates for bids. This system is called ESP (Estimating Summarizing Program) and is an add-on application to MS Excel®.

In a highly detailed manner, all requirements constituting a bid's final price are listed: execution areas, items and quantities, specifications, factors and currency exchange rates (for those projects located outside Canada). Furthermore, ESP holds macros designed to link and calculate figures between different tabs containing estimation data. In addition, these macros generate complete estimate reports and different data outputs included in tenders, according to the client's specifications.

3.2.3. Project Execution Department

Project Execution is the department responsible for tender elaboration and revision. This department gathers information from both *Estimation* and *Project Controls* areas. In addition, legal company documentation such as company qualifications, subcontract information and other requirements are included in tender packages by this department. Furthermore, one of the functions of the *Project Execution* department is to monitor previous performances and compare them with potential opportunities.

Recently, this department has been using alternative analysis methods to review past information. Via industrial indicators, multiple comparisons between projects characteristics are being made. As an example of this, Figure 3-1 associates data from past averages (tons of steel used in modules in Projects A to E) against potential averages forecasted in new projects (Project F packages 1 to 6).

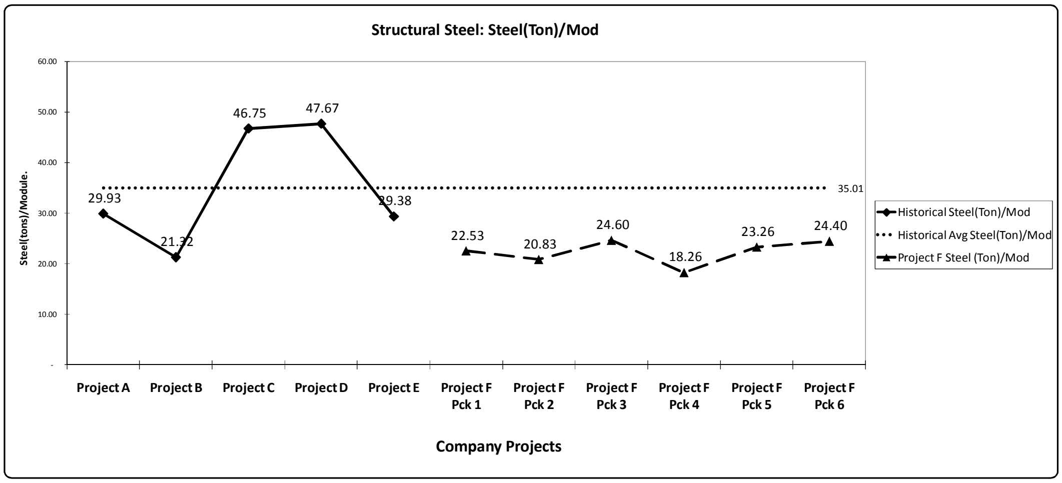

Figure 3-1: Historical and Potential project data comparison

According to this data, the closest match between Project F packages and historical data is Project B. This basic analysis supports personnel in the decision to use, as a starting base of estimation, prices derived from Project B tenders. Nevertheless, these types of analysis are characterized by being one-dimensional approaches, in which single measures are merely compared. This currently represents a potential risk when making bidding decisions, because this method does not necessarily reflect the entirety of the variables contained in a particular scenario. Multiple variables affecting fabrication activities could if omitted affect negatively a project’s bid price. Because of this simplistic approach, the quality of final decisions can be jeopardized.

For visualization purposes, graphs and charts provide an acceptable representation of how a project performs through its lifecycle. But some of the current views do not necessarily represent whole project profiles, specifically when single characteristics are compared without creating relationships with other factors. The goal of this study is to enhance a company's knowledge and awareness of its operations to become more competitive during bidding.

The necessity of creating new ways of observing and reviewing project characteristics is imperative. Upgrading comparison methods for different characteristics using data discovery techniques can diminish decision risks by having good quality data on time during development of bidding proposals.

Adding new data tools to this process will increase the company's operational knowledge. This research presents a decision support tool based on industrial indicators and knowledge discovered through implementation of *Data Mining* and *Distribution Fitting* techniques.

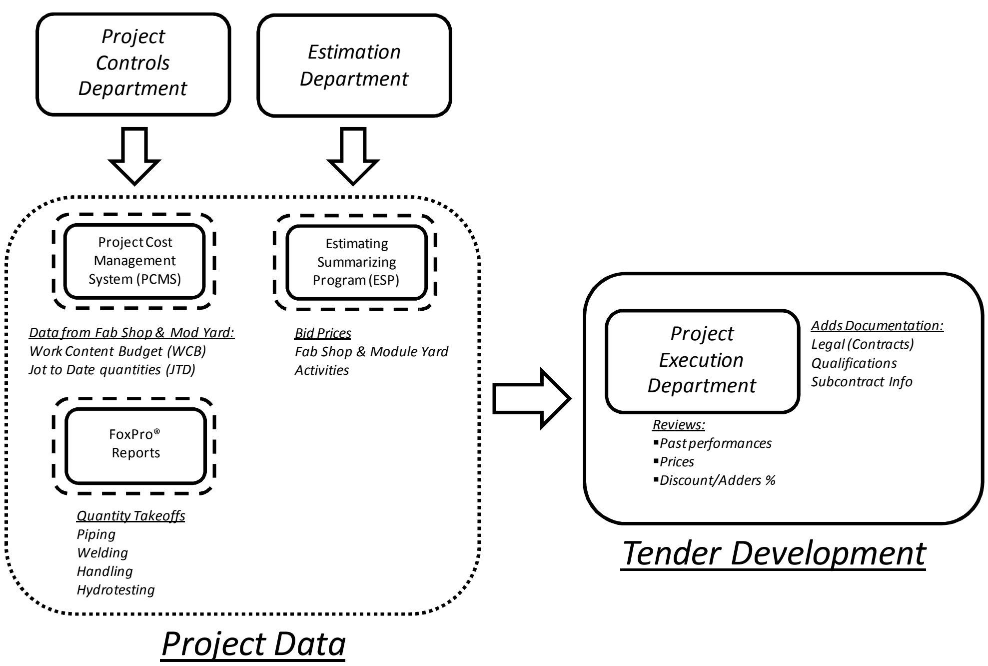

Figure 3-2: Interaction between Company Departments

3.3. Research Scope

Five different industrial projects were selected by managerial personnel to be used as benchmarks during proposal elaboration. These projects and their characteristics are presented below:

3.3.1. Project A

- Location: 75 Km at the northern area of Fort McMurray, Alberta.
- Project created to support an additional development for an Oil Sands treatment plant.
- 40 modules were fabricated. Most material used in fabrication was carbon steel (90% CS).
- Pipe size distribution: 10% small bore – 90% large bore.
- 27,310.80 man-hours invested between steel assembly and pipe installation activities.

3.3.2. Project B

- The project was a secondary upgrader unit for a larger project.
- 95 modules built.
- It presented several delays caused typically by a lack of materials.
- Originally scheduled for 4 months. Total project completion time: 12 months.
- Pipe size distribution: 20% small bore – 80% large bore pipe.

- 53,399.50 man-hours spend in pipe installation and steel assembly activities.

3.3.3. Project C

- Project designed to improve and enlarge a refinery close to the city of Edmonton.
- 40 modules were built for this project.
- 80% of the pipe material used in this project was carbon steel. The remaining 20% were alloys.
- Pipe size distribution: 75% small bore – 25% large bore pipe.
- 61,706.70 man-hours spent in pipe installation and steel assembly.

3.3.4. Project D

- Designed to be a part of an oil upgrader.
- Took 12 months to be completely built.
- 123 modules were fabricated.
- Material mix: 80% carbon steel – 20% alloys applied in fabrication.
- Pipe size distribution: 40% small bore – 60% large bore pipe.
- 134,839.38 man-hours entrusted in fabrication activities (steel assembly and pipe installation).

3.3.5. Project E

- 45 modules built.
- Project scheduled to be completed in 8 months.
- Pipe material mix: 85% carbon steel – 15% alloys.
- Pipe size distribution: 30% small bore – 70% large bore pipe.
- 31,956.38 man-hours invested during its fabrication between steel assembly and pipe installation activities.

3.4. Sources of Information

Two different information sources were selected for the development of this study:

3.4.1. Company records

Company records were obtained from the following sources:

- Project Controls
- Project Execution
- Field Operations
- Estimation

Project Controls provide quantity takeoff data from fabrication activities of previous projects. This information is specific and can be classified by activity (handling, welding, etc). Quantity takeoffs are presented in MS Excel® spreadsheets. In addition, PCMS reports are generated using an in-house database management system. These reports present in a high-level perspective the amount of man-hours and cost involved in previous projects.

Project Execution gathers information coming from *Project Controls* and *Estimation* departments. Meeting regularly, their responsibility is to organize data to revise tendering decisions. Furthermore, this department collects different spreadsheet reports and flat text files producing tendering packages and proposals.

Field Operations are handed by project managers, generating different types of spreadsheets containing information about fabrication of pipe spools, steel assembly and module loading.

The Estimation department calculates tender financial outcomes. By using a specifically designed spreadsheet called ESP this department calculates final prices on items belonging to fabrication activities of a particular project.

The first three sources provided data to analyze case scenarios. *Estimation* department data was used to understand certain unit measures and their associated costs. It was only used for reference.

Each document presented its own structure and method of reporting. Data organization from all sources was deficient because there was no consistency during record review. Absence of responsible personnel who prepared previous reports represented an additional obstacle during research. People occupying key roles during data documentation were no longer working with the company. This increased the difficulty level of data processing because of a lack of understanding on how certain figures and factors were calculated.

3.4.2. Meetings with company personnel

Weekly meetings were held to discuss aspects related to the creation of alternative indicators to portray past project experiences of the company. In these series of meetings, the following participants were present:

Representing the company:

- Vice-President of Operations
- Director of Operations
- Project Execution Manager
- Project Controls Manager
- Field Project Manager

3.5. Data Warehousing

In the present study, diverse types of information associated with pipe module fabrication are collected individually from various departments of the company. Data from both the fabrication shop and module yard is uniquely treated according to its usage. For data analysis purposes, information was collected and stored following a star schema in a data base. According to Zhang et al. (2004), realistic and descriptive tables are present in a star schema, representing two different types of storing arrangements. These can contain accurate and explanatory data, respectively.

This is the case for the present study, in which spreadsheets containing valuable data from different departments were transferred to a main database created in MS Access 2007®. This database is static and will generate reports in spreadsheets according to *Data Mining* needs and approaches.

In this research, information sources inside the company were identified as follows:

- *Project Controls Department:* Produces Project Cost Management System reports (PCMS) on a monthly basis including Work Content Budget (WCB) and Job to Date (JTD) measures for each activity belonging to a project (quantities and hours performed, respectively). These reports are stored in flat text files (.pdf reports).

- *Project Execution Department:* Elaborates proposals for different potential projects. During this process, company experience documents, legal documentation and bid prices are assembled together forming tender packages. These are generally the product of a combination of flat text files, presentations and spreadsheets.

- *Estimation Department:* It produces calculations of every item present in tenders. The combination of these elements generates tender prices. Most of the documents created within this department are spreadsheets with multiple uses such as price bidding and benchmarking.

All data coming from these sources is gathered and stored in network drives within the company's local area network (LAN). There are several access levels to this network and its information is considered to be sensitive and strictly confidential.

Once data was available from sources, a database created in MS Access® using queries was used to collect and pre-process information for *Data Mining* purposes. Furthermore, formatting of these files was modified, creating alternate versions analyzed by *Data Mining* software selected: WEKA® software. An overview of a case study involving *Clustering* and pipe module fabrication database queries is presented in *Chapter No. 4*.

3.6. Data Gathering and Cleaning

Reports were obtained from *Project Execution*, *Project Controls* and *Field Operations*. Depending on use, records were organized and cleaned for either indicator creation or *Clustering* and *Association rules* analysis. It is important to mention that data tables, graphs and others presented in this thesis have been scaled due to confidentiality reasons. Some of the cleaning and organizing techniques used in the acquisition of data were the following:

- PCMS reports generated MS Excel® documents. Fabrication measures for both the Fabrication Shop and Module Yard were selected, obtaining as final result a condensed spreadsheet. Due to space limitations, this file was divided into two sections, observed in Table 3-1 and Table 3-2.

		Work Content Budget				
Project	Item	WCB Qty	WCB Hr	WCB Amount	WCB Hr/Qty	WCB Amt/Hr
PROJECT A MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	121074.8	12034	656689	0.11	60.027
PROJECT A MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	223744.4	19817.6	1081450.7	0.099	60.027
PROJECT A MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	0	4541.9	247892.7	0	60.038
PROJECT A MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	134191.2	14588.2	796088.7	0.121	60.027
PROJECT A MODULE YARD	A/G PIPE WELDING 12" AND LARGER	2983.2	2139.5	116778.2	0.792	60.038
PROJECT B MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	14060.2	5140.3	140197.2	0.407	29.997
PROJECT B MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	104089.7	9983.6	272280.8	0.11	29.997
PROJECT B MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	3227.4	7275.4	198413.6	2.475	29.997
PROJECT B MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	57968.9	5166.7	140908.9	0.099	29.997
PROJECT B MODULE YARD	A/G PIPE WELDING 12" AND LARGER	1922.8	3911.6	106683.5	2.233	29.997
PROJECT C MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	36735.6	10304.8	392304	0.308	41.877
PROJECT C MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	204267.8	10491.8	399424.3	0.055	41.877
PROJECT C MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	2956.8	2985.4	113654.2	1.111	41.877
PROJECT C MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	75324.7	5998.3	228356.7	0.088	41.877
PROJECT C MODULE YARD	A/G PIPE WELDING 12" AND LARGER	610.5	1304.6	49666.1	2.354	41.877
PROJECT D MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	23466.3	3340.7	122733.6	0.154	40.414
PROJECT D MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	102047	5981.8	219754.7	0.066	40.414
PROJECT D MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	123.2	145.2	5332.8	1.298	40.403
PROJECT D MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	232247.4	7858.4	288686.2	0.033	40.414
PROJECT D MODULE YARD	A/G PIPE WELDING 12" AND LARGER	79.2	64.9	2383.7	0.902	40.403
PROJECT E MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	53710.8	6958.6	260112.6	0.143	41.118
PROJECT E MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	237794.7	13670.8	511013.8	0.066	41.118
PROJECT E MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	484	1360.7	50987.2	3.091	41.228
PROJECT E MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	153450	6741.9	252012.2	0.044	41.118
PROJECT E MODULE YARD	A/G PIPE WELDING 12" AND LARGER	308	466.4	17475.7	1.661	41.228

Table 3-1: PCMS spreadsheet registering Work Content Budget figures (WCB)

		Job To Date					
Project	Item	JTD Qty	JTD Hr	JTD Amount	JTD Hr/Qty	JTD Amt/Hr	Utilization
PROJECT A MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	40723.1	29903.5	1786524.3	0.803	65.714	0.12023
PROJECT A MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	140642.7	35426.6	2108493.2	0.275	65.472	0.53669
PROJECT A MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	0	5492.3	342956.9	0	68.695	0.63987
PROJECT A MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	109083.7	11817.3	703973.6	0.121	65.527	0.44297
PROJECT A MODULE YARD	A/G PIPE WELDING 12" AND LARGER	783.2	2478.3	156487.1	3.476	69.454	0.46013
PROJECT B MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	7026.8	3612.4	201809.3	0.561	61.446	0.12023
PROJECT B MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	52021.2	9298.3	542047	0.198	64.119	0.53669
PROJECT B MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	1613.7	702.9	48068.9	0.484	75.229	0.63987
PROJECT B MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	28971.8	2249.5	128020.2	0.088	62.612	0.44297
PROJECT B MODULE YARD	A/G PIPE WELDING 12" AND LARGER	961.4	265.1	17758.4	0.308	73.689	0.46013
PROJECT C MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	36671.8	18012.5	670695.3	0.539	40.964	0.12023
PROJECT C MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	188843.6	23331	850705.9	0.132	40.106	0.53669
PROJECT C MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	1441	4439.6	171141.3	3.388	42.405	0.63987
PROJECT C MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	71612.2	4842.2	177493.8	0.077	40.315	0.44297
PROJECT C MODULE YARD	A/G PIPE WELDING 12" AND LARGER	501.6	932.8	37247.1	2.046	43.923	0.46013
PROJECT D MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	23466.3	3111.9	104920.2	0.143	37.092	0.12023
PROJECT D MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	102047	7175.3	238472.3	0.077	36.564	0.53669
PROJECT D MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	123.2	553.3	18458	4.939	36.729	0.63987
PROJECT D MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	232247.4	5449.4	182397.6	0.022	36.817	0.44297
PROJECT D MODULE YARD	A/G PIPE WELDING 12" AND LARGER	79.2	134.2	4103	1.859	33.638	0.46013
PROJECT E MODULE YARD	A/G PIP LAB SPL ERT 2" & LESS	53710.8	11161.7	408824.9	0.231	40.293	0.12023
PROJECT E MODULE YARD	A/G PIP LAB SPL ERT 2 1/2" - 10"	237794.7	17600	684358.4	0.077	42.768	0.53669
PROJECT E MODULE YARD	A/G PIPE WELDING 2 1/2" - 10"	484	1426.7	56570.8	3.245	43.637	0.63987
PROJECT E MODULE YARD	A/G PIP LAB SPL ERT 12" - 34"	153450	8560.2	311832.4	0.066	40.073	0.44297
PROJECT E MODULE YARD	A/G PIPE WELDING 12" AND LARGER	308	565.4	23657.7	2.024	46.024	0.46013

Table 3-2: PCMS spreadsheet registering Job to Date figures (JTD)

- Quantity takeoff spreadsheets were organized and categorized according to processing requirements, removing special characters (Table 3-3, raw data). Any item affecting functionalities of .ARFF files used by WEKA® was eliminated. A final extract is portrayed in Table 3-4 (cleaned data).

PROJECT	DESC	SIZE	QTY	SIZEDESC	UNITPRICE
PROJECT A	PIPE, STD, ASTM A53 Gr B ERW	16.000 0.000	54	16"	172.03
PROJECT A	ELBOW 90 DEG LR BW, STD, ASTM A234 Gr WPB	16.000 0.000	2	16"	331.22
PROJECT A	ELBOW 45 DEG LR BW, STD, ASTM A234 Gr WPB	16.000 0.000	2	16"	242.10
PROJECT A	NIPPLE, XS, PE, ASTM A106 Gr B Smls	0.500 6.000	2	1/2" x 6"	3.70
PROJECT A	NIPPLE, XS, PE, ASTM A106 Gr B Smls	0.500 3.000	2	1/2" x 3"	2.75
PROJECT A	ORIFICE FLANGE SET WELDNECK, STD, 300 lb, RF, 125-250 AARH, ASTM A105N	16.000 0.000	2	16"	458.91
PROJECT A	GASKET, 300 lb, 316SS FLEXGRAPH-CS O/R-1/8"	16.000 0.000	3	16"	24.84
PROJECT A	PIPE SUPPORT LS13	16.000 0.000	3	16"	919.91
PROJECT A	HANDLE PIPE, STD, CS	16.000 0.000	57	16"	266.01
PROJECT A	HANDLE PIPE, XS, CS	0.500 0.000	3	1/2"	29.15
PROJECT A	BOLT-UPS, 300 lb, CS	16.000 0.000	2	16"	687.71
PROJECT A	BOLT SET(SPEC BLIND, 1" THICK), 300 lb, Bolt-A193 GrB7/Nut-A194 Gr2H (Liqd Q&T)	16.000 0.000	2	16"	149.60
PROJECT A	HANDLE... FO300	16.000 0.000	2	16"	1489.85
PROJECT A	PIPE, STD, ASTM A106 Gr B Smls	4.000 0.000	58	4"	37.66
PROJECT A	HANDLE PIPE, STD, CS	4.000 0.000	183	4"	69.23
PROJECT A	PIPE SUPPORT GU11	4.000 0.000	14	4"	178.87
PROJECT A	PIPE, STD, ASTM A333 Gr 6 SMLS	4.000 0.000	144	4"	44.90
PROJECT A	PIPE, STD, ASTM A333 Gr 6 SMLS	2.000 0.000	158	2"	15.49
PROJECT A	TEE (REDUCING) BW, STD, ASTM A420 Gr WPL6	4.000 2.000	7	4" x 2"	38.75
PROJECT A	FLANGE WELDNECK, STD, 150 lb, RF, 125-250 AARH, ASTM A350 Gr LF2,Class 1	2.000 0.000	24	2"	16.86
PROJECT A	PIPE SUPPORT GU04-A1	4.000 0.000	14	4"	225.53
PROJECT A	HANDLE PIPE, STD, LOW TEMP CS	4.000 0.000	145	4"	69.23
PROJECT A	HANDLE PIPE, STD, LOW TEMP CS	2.000 0.000	81	2"	40.08
PROJECT A	ELBOW 90 DEG LR BW, STD, ASTM A420 Gr WPL6	2.000 0.000	35	2"	4.33

Table 3-3: Quantity takeoffs for Project A (raw data)

PROJECT	DESC	QTY	SIZEDESC	UNITPRICE
PROJECTA	PIPESTDASTMA53GrBERW	54	16	172.03
PROJECTA	ELBOW90DEGLRBWSTDASTMA234GrWPB	2	16	331.22
PROJECTA	ELBOW45DEGLRBWSTDASTMA234GrWPB	2	16	242.10
PROJECTA	NIPPLEXSPEASTMA106GrBSmls	2	0.5	3.70
PROJECTA	NIPPLEXSPEASTMA106GrBSmls	2	0.5	2.75
PROJECTA	ORIFICEFLANGESETWELDNECKSTD300lbRF125250AARHASTMA105N	2	16	458.91
PROJECTA	GASKET300lb316SSFLEXGRAPHCSOR18	3	16	24.84
PROJECTA	PIPESUPPORTLS13	3	16	919.91
PROJECTA	HANDLEPIPESTDCS	57	16	266.01
PROJECTA	HANDLEPIPEXSCS	3	0.5	29.15
PROJECTA	BOLTUPS300lbCS	2	16	687.71
PROJECTA	BOLTSETSPECBLIND1THICK300lbBoltA193GrB7NutA194Gr2HLiqdQ&T	2	16	149.60
PROJECTA	HANDLE...FO300	2	16	1489.85
PROJECTA	PIPESTDASTMA106GrBSmls	58	4	37.66
PROJECTA	HANDLEPIPESTDCS	183	4	69.23
PROJECTA	PIPESUPPORTGU11	14	4	178.87
PROJECTA	PIPESTDASTMA333Gr6SMLS	144	4	44.90
PROJECTA	PIPESTDASTMA333Gr6SMLS	158	2	15.49
PROJECTA	TEEREDUCINGBWSTDASTMA420GrWPL6	7	4	38.75
PROJECTA	FLANGEWELDNECKSTD150lbRF125250AARHASTMA350GrLF2Class1	24	2	16.86
PROJECTA	PIPESUPPORTGU04A1	14	4	225.53
PROJECTA	HANDLEPIPESTDLOWTEMPCS	145	4	69.23
PROJECTA	HANDLEPIPESTDLOWTEMPCS	81	2	40.08
PROJECTA	ELBOW90DEGLRBWSTDASTMA420GrWPL6	35	2	4.33

Table 3-4: Quantity takeoffs for Project A (cleaned data)

3.7. Industrial Indicators

Over a number of months, Industrial Indicators were assembled, each customized to reflect and register individual characteristics of a particular project. In addition, through the use of graphs, they become visualization tools to review past company performances.

3.7.1. Production Indicator

The purpose of this indicator is to calculate a historical average number of modules fabricated (average production) in all previous projects and review which projects surpassed an estimated average due to their unique characteristics. Modules fabricated in each project are represented with columns, while the historical average is portrayed as a benchmark line in Figure 3-3.

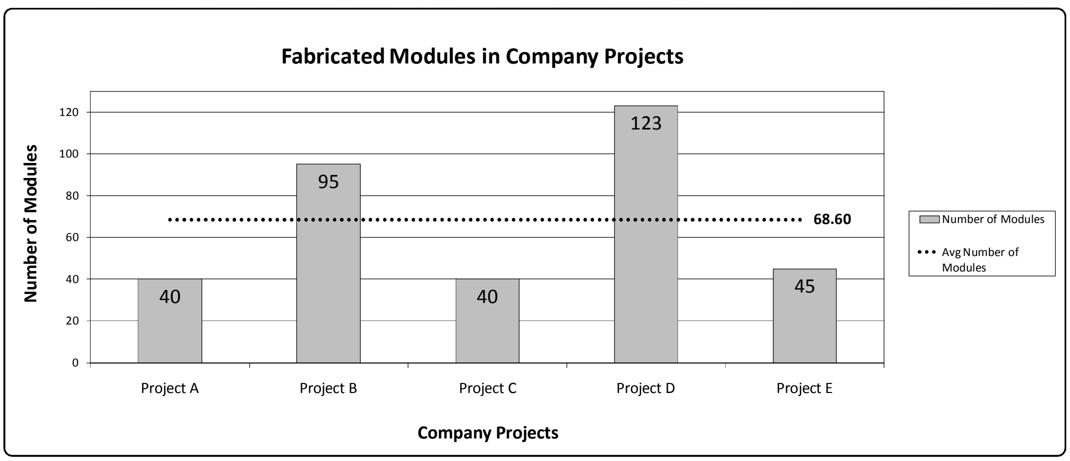

Figure 3-3: Fabricated Modules per Project

3.7.2. Steel Assembly Indicators

3.7.2.1. Steel (Ton)/Mod. Tons of steel per module fabricated. An average number of Tons of Steel per Module fabricated is represented in Figure 3-4. This was calculated using the following equation:

$$\frac{Steel(Ton)}{Mod} = \frac{Total\ Number\ \ of\ Tons\ of\ Steel\ in\ a\ project}{Number\ of\ Modules\ fabricated}$$

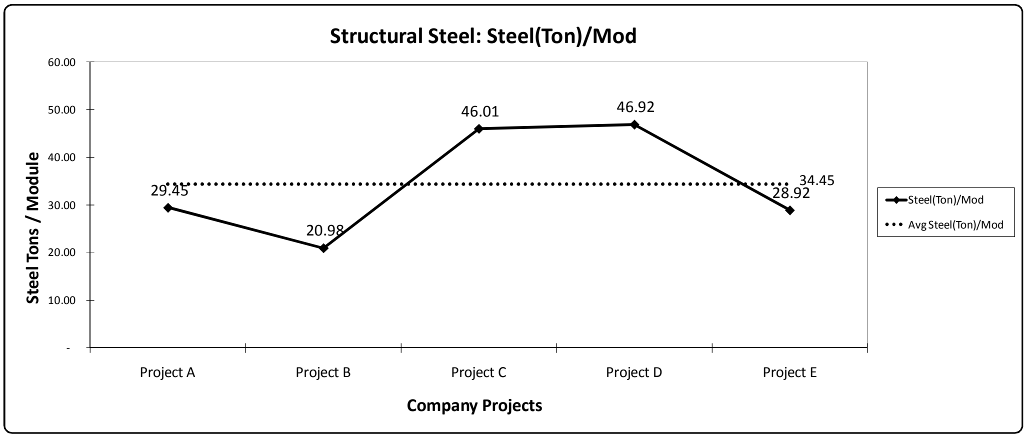

Figure 3-4: Steel Assembly Steel(Ton)/Mod Indicator

3.7.2.2. Direct Hrs/Steel Ton. Direct Fabrication Man-Hours per Ton of Steel. Its purpose is to reflect how many direct man-hours were invested in the steel assembly of a single module, portrayed in Figure 3-5. This indicator was built using the next equation:

$$\frac{Direct\ Hrs}{Steel\ (Ton)} = \frac{Total\ Number\ of\ Direct\ Man-hours\ invoved\ in\ Steel\ Assembly}{Total\ \ Number\ of\ Tons\ of\ Steel\ in\ a\ Project}$$

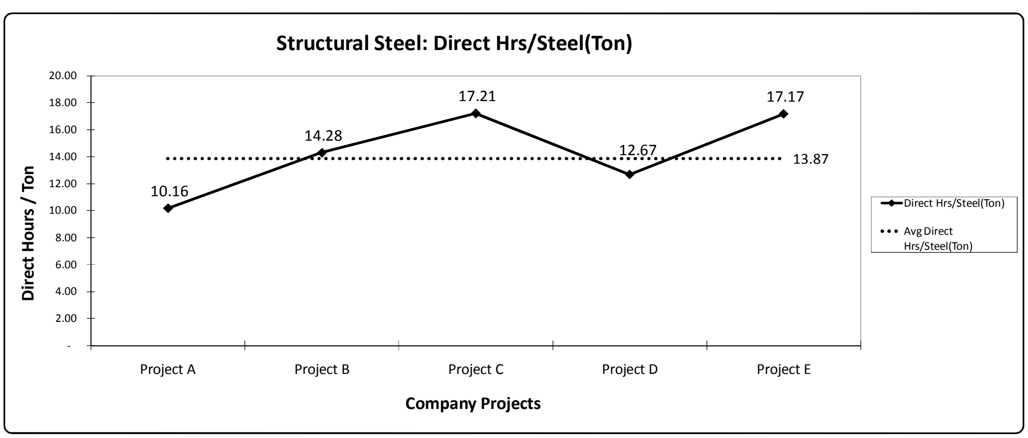

Figure 3-5: Steel Assembly Direct Hrs/Steel(Ton) Indicator

3.7.2.3. Steel member usage per project. Indicates the utilization percentage of structural steel (according to specifications) used in a pipe module construction project. For this indicator, records from steel assembly activities were gathered and ordered and are depicted in Figures 3-6 to 3-10.

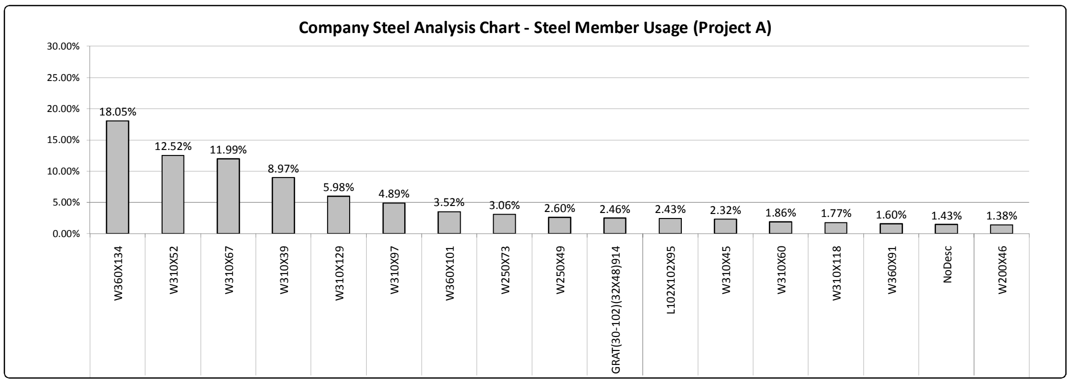

Figure 3-6: Steel analysis chart – Project A

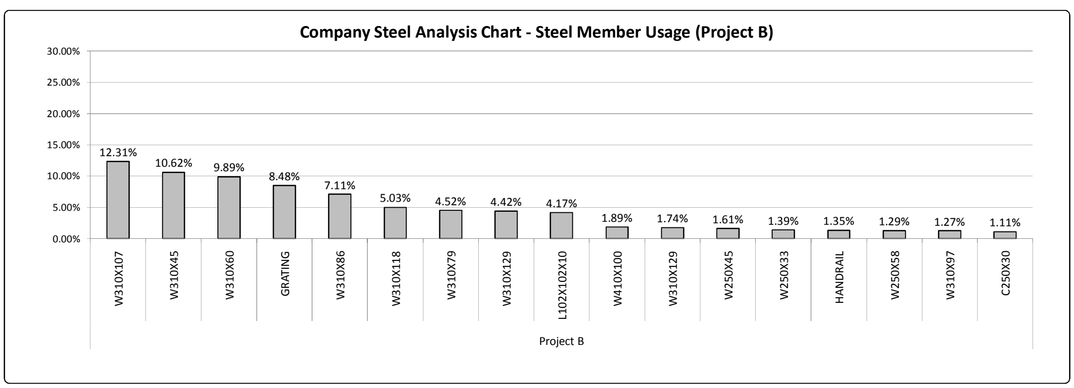

Figure 3-7: Steel analysis chart – Project B

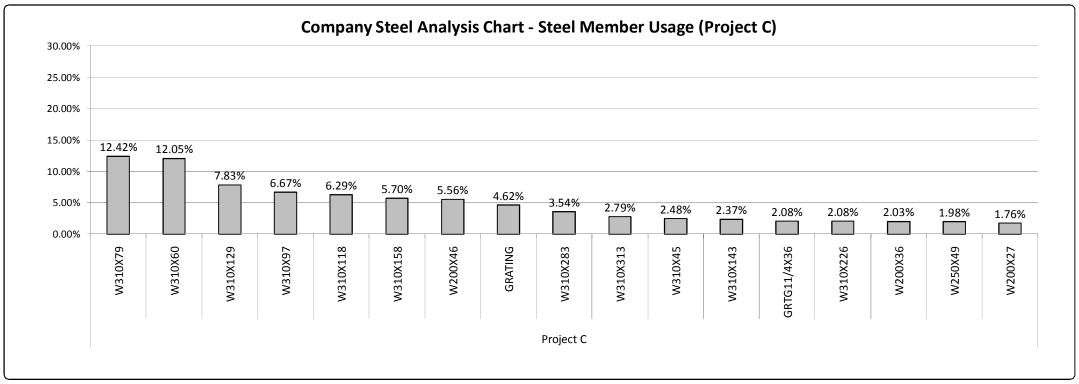

Figure 3-8: Steel analysis chart – Project C

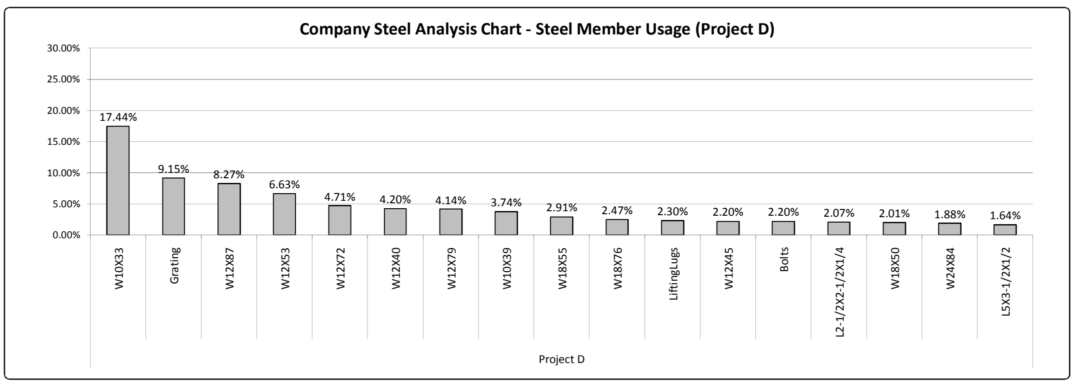

Figure 3-9: Steel analysis chart – Project D

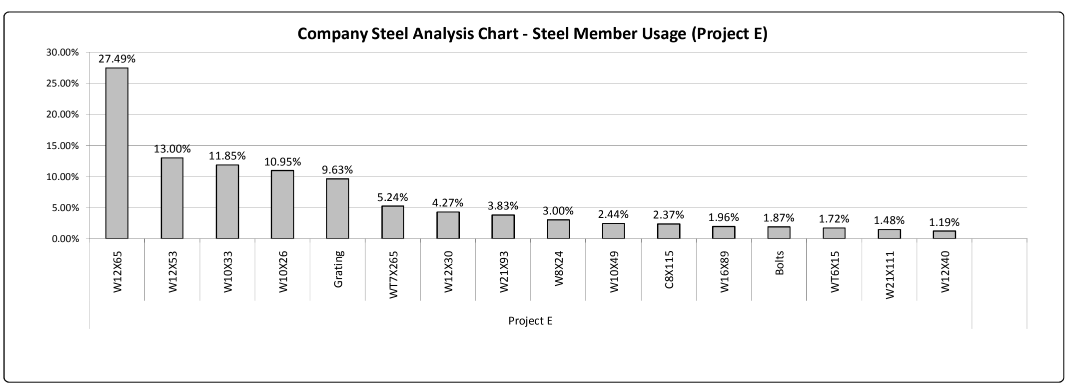

Figure 3-10: Steel analysis chart – Project E

3.7.2.4. Items used per Steel Category. This measure expresses how many structural steel items were used during steel assembly phase. In Table 3-5, steel items present in construction were quantified according to their weight following company standards.

Structural steel is categorized in the following denominations:

- *Extra Heavy Steel:* More than 90 kg per linear meter of steel.
- *Heavy Steel:* 60 to 90 kg per linear meter of steel.
- *Medium Steel:* 30 to 60 Kg per linear meter of steel.
- *Light Steel:* Less than 30 Kg per linear meter of steel.
- *Miscellaneous Steel:* Present in different fabricated items such as ladders, handrails, grating and plates.

Items used per Steel Category - Company Projects

Project	Description	Number of Items Used	Total Items
PROJECT A	Extra Heavy Steel	9	191
	Heavy Steel	12	
	Light Steel	81	
	Medium Steel	26	
	Miscellaneous Steel	63	
PROJECT B	Extra Heavy Steel	33	472
	Heavy Steel	23	
	Light Steel	224	
	Medium Steel	188	
	Miscellaneous Steel	4	
PROJECT C	Extra Heavy Steel	9	365
	Heavy Steel	51	
	Light Steel	87	
	Medium Steel	106	
	Miscellaneous Steel	112	
PROJECT D	Extra Heavy Steel	4	217
	Heavy Steel	54	
	Light Steel	84	
	Medium Steel	29	
	Miscellaneous Steel	46	
PROJECT E	Extra Heavy Steel	0	51
	Heavy Steel	12	
	Light Steel	17	
	Medium Steel	15	
	Miscellaneous Steel	7	

Table 3-5: Steel Assembly Items Used per Steel Category indicator

3.7.2.5. Top 5 Steel Items utilization. In Table 3-6, a ranking of the most used structural steel items during assembly is depicted with a description and a utilization percentage for each item.

Top 5 Items Utilization - Company Projects

Project	Description	Item Description	Weight (tons)	Utilization %
PROJECT A	Extra Heavy Steel	W360X134	204.92	18.05%
	Medium Steel	W310X52	142.20	12.52%
	Heavy Steel	W310X67	136.14	11.99%
	Medium Steel	W310X39	101.85	8.97%
	Extra Heavy Steel	W310X129	67.92	5.98%
PROJECT B	Extra Heavy Steel	W310X107	212.45	12.31%
	Medium Steel	W310X45	183.24	10.62%
	Medium Steel	W310X60	170.63	9.89%
	Miscellaneous Steel	Grating	146.40	8.48%
	Heavy Steel	W310X86	122.62	7.11%
PROJECT C	Heavy Steel	W310X79	205.88	12.42%
	Medium Steel	W310X60	199.78	12.05%
	Heavy Steel	W310X129	129.88	7.83%
	Heavy Steel	W310X97	110.54	6.67%
	Heavy Steel	W310X118	104.24	6.29%
PROJECT D	Medium Steel	W10X33	719.87	17.44%
	Miscellaneous Steel	Grating	377.40	9.15%
	Heavy Steel	W12X87	341.10	8.27%
	Heavy Steel	W12X53	273.53	6.63%
	Heavy Steel	W12X72	194.18	4.71%
PROJECT E	Heavy Steel	W12X65	297.89	27.49%
	Heavy Steel	W12X53	140.88	13.00%
	Medium Steel	W10X33	128.40	11.85%
	Medium Steel	W10X26	118.72	10.95%
	Miscellaneous Steel	Grating	104.38	9.63%

Table 3-6: Steel Assembly Top 5 Items Utilization chart

3.7.2.6. Steel involved in Module Fabrication. Figure 3-11 represents a steel mix distribution present during steel assembly. All observed percentages were categorized by their weight.

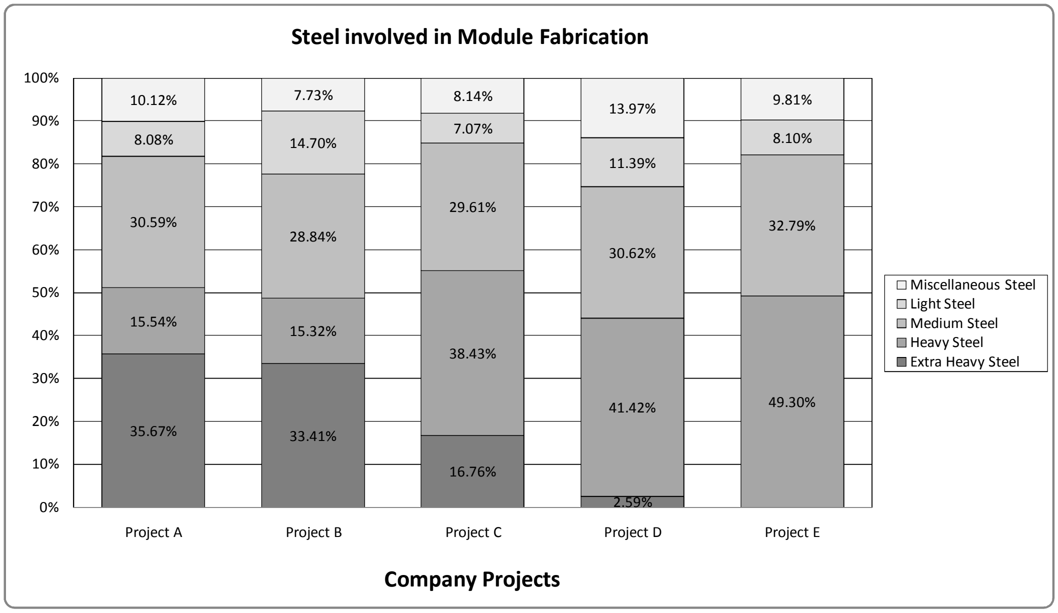

Figure 3-11: Steel percentages involved in Module Fabrication

3.7.2.7. Major frame structures. Figure 3-12 indicates a measure of man-hours per ton of steel invested during steel structure assembly and erection phases.

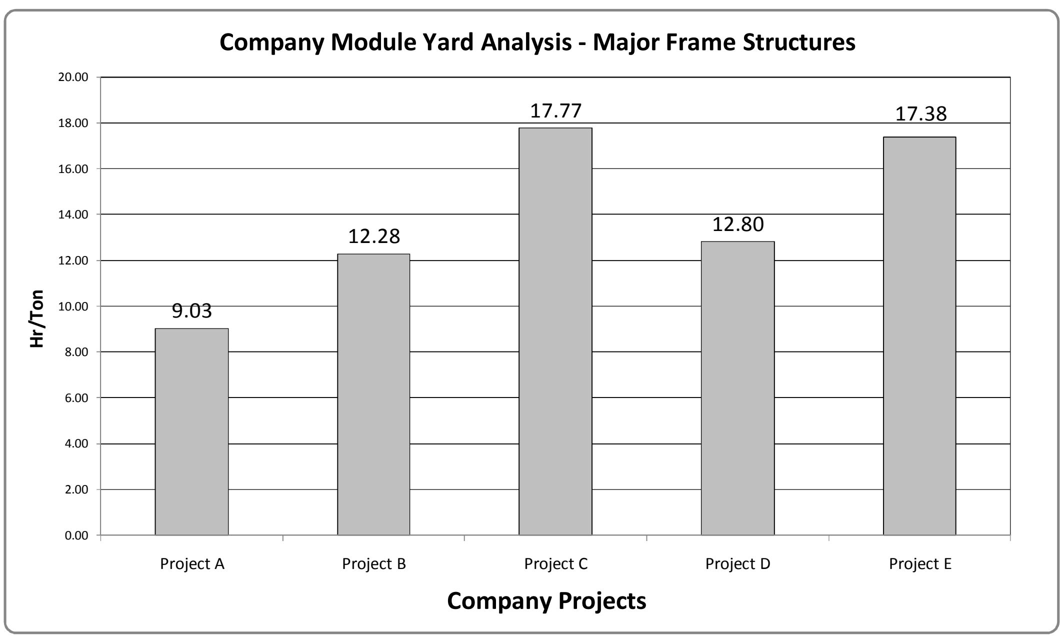

Figure 3-12: Number of man-hours per steel ton invested in assembly of Major Frame Structures

3.7.3. Pipe Spool Installation Indicators:

3.7.3.1. Pipe(m)/Mod. In Figure 3-13, this measure represents average linear meters of pipe used in the fabrication of a single module. It was calculated using the following equation:

$$\frac{Pipe(m)}{Mod} = \frac{Total\ Linear\ Meters\ of\ Pipe}{Total\ number\ of\ Modules\ fabricated}$$

This graph has two different lines: a continuous line represents average linear meters of pipe used per module for different projects. The dotted line depicts the total average linear meters of pipe used in a single module.

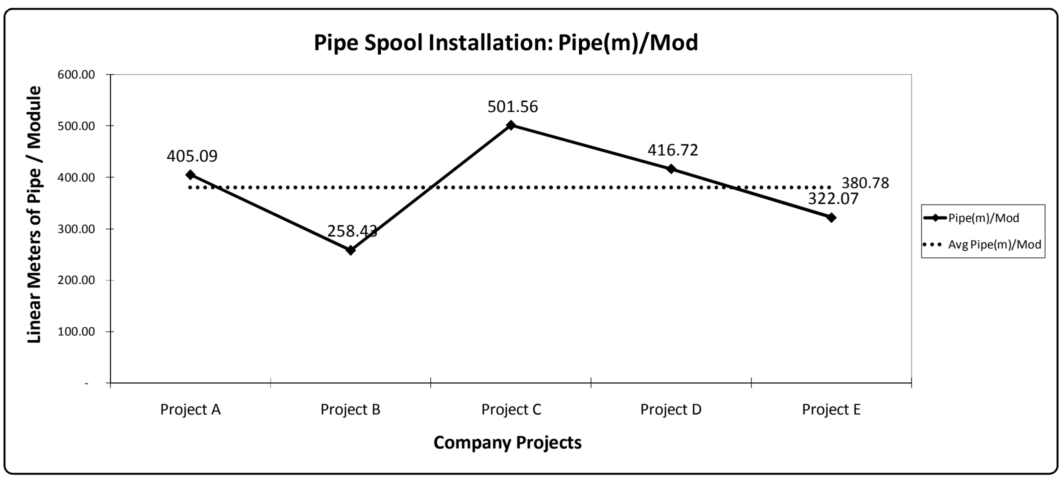

Figure 3-13: Pipe Installation Pipe(m)/Mod Indicator

3.7.3.2. Pipe(ft)/Mod. Figure 3-14 reflects the total number of linear feet of pipe used in fabrication of a module. It was elaborated using a similar equation:

$$\frac{Pipe(ft)}{Mod} = \frac{Total\ Linear\ Feet\ of\ Pipe}{Total\ number\ of\ Modules\ fabricated}$$

The only difference remains in its units. Instead of calculating linear meters of pipe per module, this graph presents the average linear feet of pipe per module fabricated.

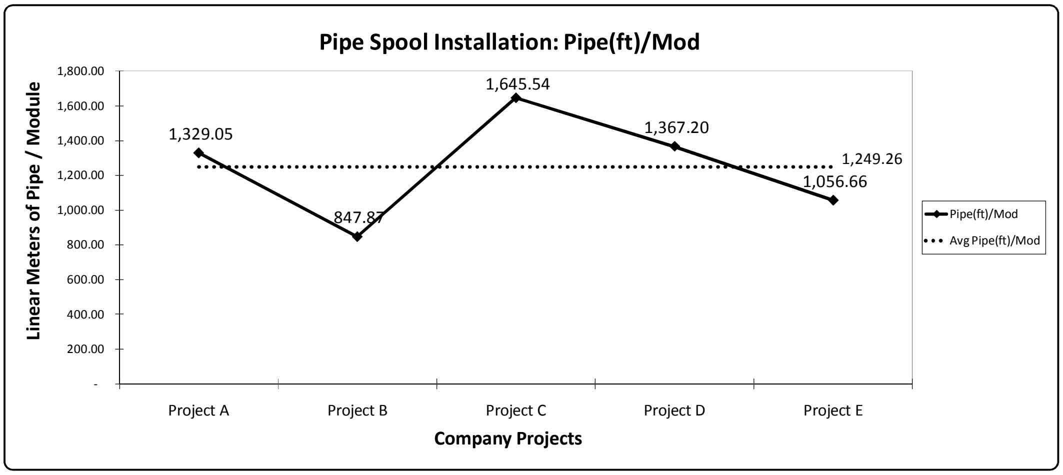

Figure 3-14: Pipe Installation Pipe(ft)/Mod Indicator

3.7.3.3. Direct Hrs/Pipe(m). This indicator was made to express a total amount of direct man-hours per linear meter of pipe installed (Figure 3-15).

$$\frac{Direct\ Hrs}{Pipe(m)} = \frac{Total\ direct\ man-\ hours\ involved\ in\ Pipe\ Installation}{Total\ Linear\ Meters\ of\ Pipe}$$

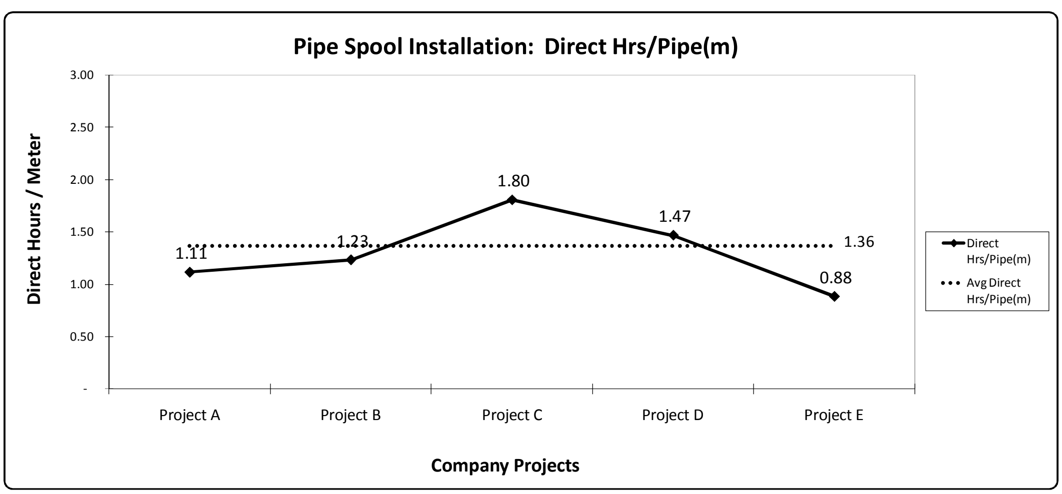

Figure 3-15: Pipe Installation Direct Hrs/Pipe(m) Indicator

3.7.3.4. Direct Hrs/Pipe(ft). It is the equivalent of the above mentioned indicator, with the only difference of measuring a total number of direct man-hours per linear foot of pipe installed (Figure 3-16).

$$\frac{Direct\ Hrs}{Pipe(ft)} = \frac{Total\ direct\ man - hours\ involved\ in\ Pipe\ Installation}{Total\ Linear\ Feet\ of\ Pipe}$$

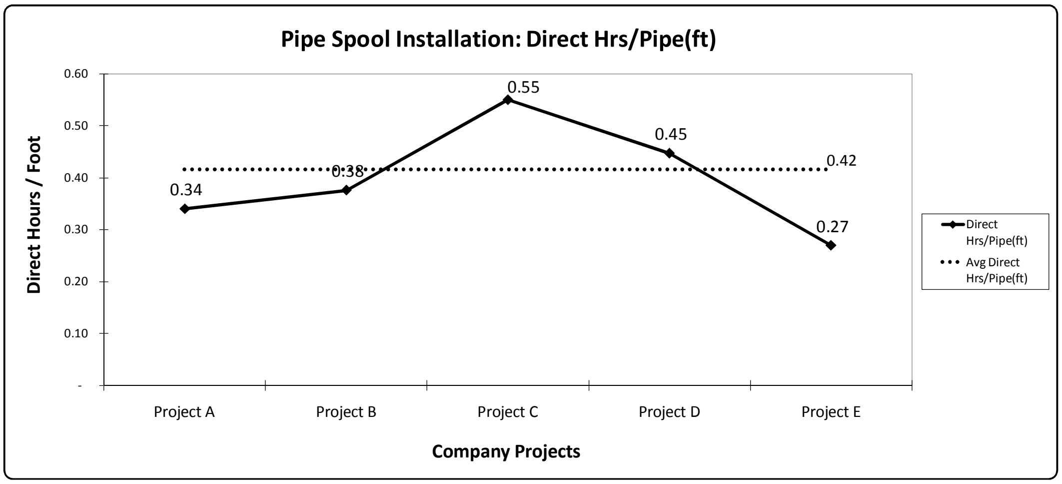

Figure 3-16: Pipe Installation Direct Hrs/Pipe(ft) Indicator

3.7.3.5. Performance Hr/Ft. It is a measure classified in three different categories to determine performances during pipe spool installation.

These categories were established according to PCMS reports representing performances for three different pipe diameter ranges (Figure 3-17):

- Spools erected in the Module Yard, averaging a diameter between 12”-34”.
- Spools erected in the Module Yard, averaging a diameter between 2 1/2”-10”.

- Spools erected in the Module Yard, averaging a diameter between 2" & less. This presented higher performance measures, possibly due increased complexity in spool arrangements for this category.

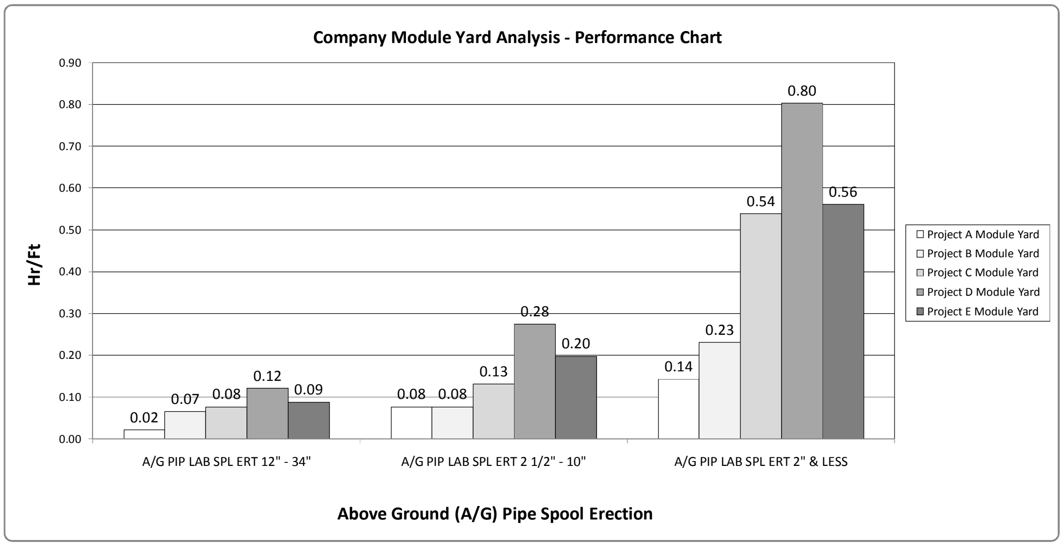

Figure 3-17: Pipe Installation Performance (Hr/Ft) Indicator

3.7.4. Fabricated Spools Indicators

3.7.4.1. Number of Spools per Module. It reflects an average number of spools fabricated for a single module (continuous line in Figure 3-18).

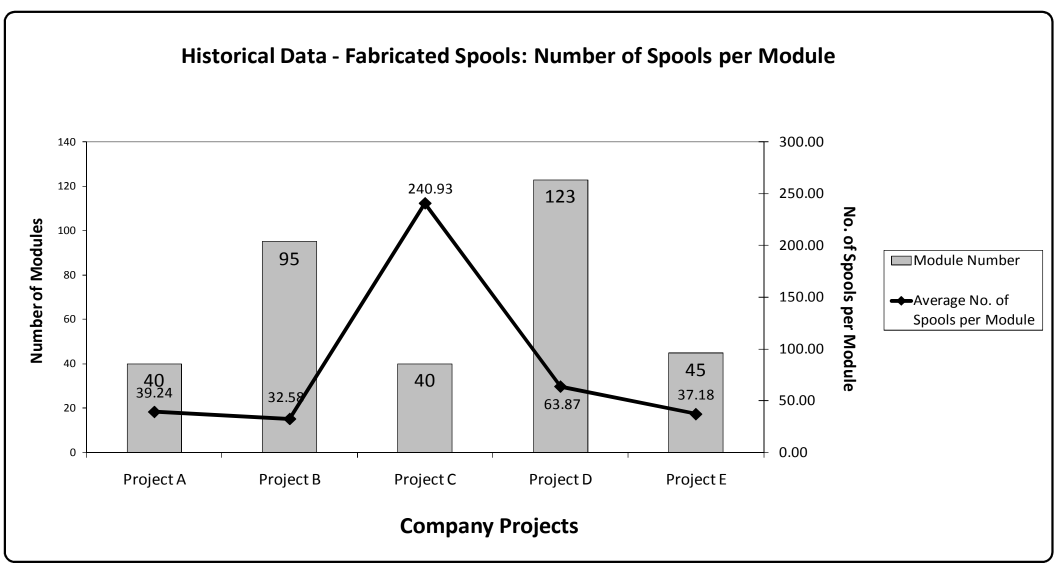

Figure 3-18: Average number of Spools fabricated per Module Indicator

3.7.4.2. Diameter Inches per module. Indicates an average welding diameter inches measure used in the fabrication of a module. It is symbolized with a dark line in Figure 3-19.

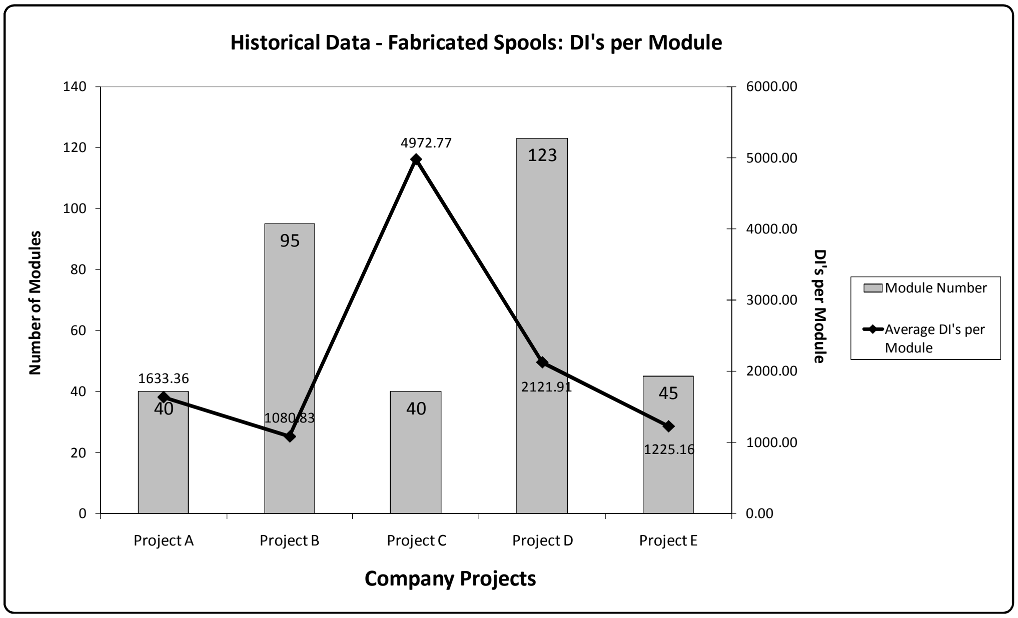

Figure 3-19: Average diameter inches (DI) fabricated per Module Indicator

3.7.4.3. Multi-level modules number. Designates how many multiple deck modules existed in a single project. This number is presented in Table 3-7.

3.7.4.4. Direct Hrs/Spool. Denotes average man-hours invested in a single spool of a module. As its predecessor, it can be observed in Table 3-7.

Fabricated Spools

Project	Module Number	Spool Number	Total DI's	Average No. of Spools per Module	Average DI's per Module	No. of Multi-level Modules	Average Direct Hrs/Spool
Project A	40	1,570	65,335	39.24	1633.36	0.00	11.51
Project B	95	3,095	102,678	32.58	1080.83	16.00	13.97
Project C	40	9,637	198,911	240.93	4972.77	0.00	5.78
Project D	123	7,856	260,995	63.87	2121.91	25.00	12.76
Project E	45	1,673	55,132	37.18	1225.16	20.00	10.21
Sub-Totals	**343**	**23,832**	**683,051**	**28.66**	**1991.40**	**20.33**	**10.85**

Table 3-7: Fabricated Spools Data

3.7.5. Historical Piping Average Project Profile Curve

Quantity takeoffs for each project were extracted from company records. In these files, pipes used in fabrication were classified by diameter size and their quantities were summed in linear meters. In addition, percentages of each diameter present in fabrication were calculated.

Figure 3-20 presents a diameter distribution line that reflects how much each pipe diameter was used during pipe module fabrication. Furthermore, for the development of this profile curve, all quantities derived from those projects integrating the research's scope were taken into consideration. It represents the historical average distribution of different pipe diameters implicated in this variant of industrial construction.

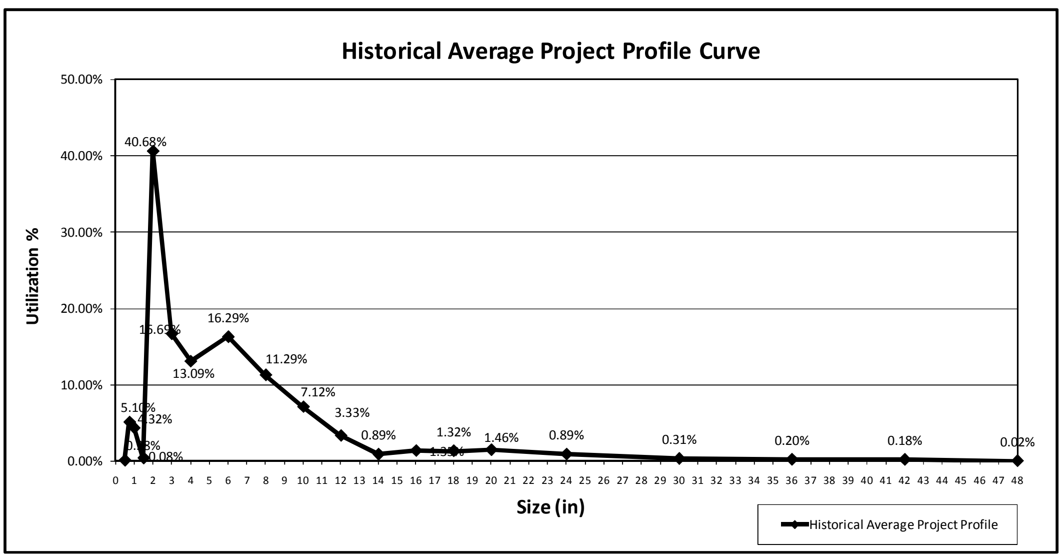

Figure 3-20: Historical Average Project Profile Indicator

An additional graph (Figure 3-21) was used for profile comparison by the managerial team of the company. It reflects the average piping project profile curve with the addition of new curves representing different pipe utilizations in each project. By comparing profiles, the managerial team can decide which project characteristics could be closer to those reflected in historical company data.

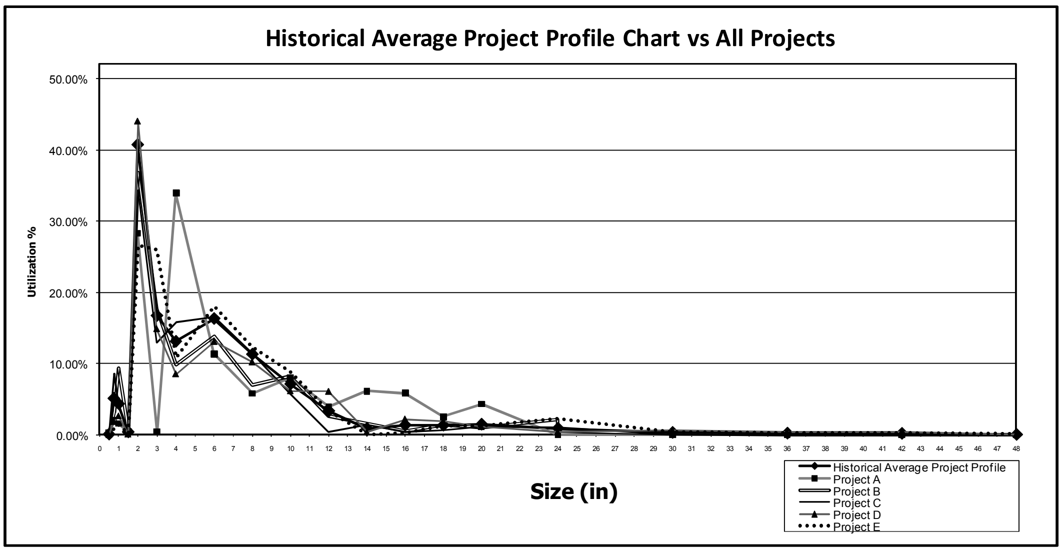

Figure 3-21: Historical Average Project Profile vs. Project Profiles Indicator

3.8. Research Process

Weekly meetings reviewed and monitored the research process for industrial indicators creation. Once a week, select personnel from different departments of the company observed work performed and provided feedback.

In addition, clarifications regarding how certain indicators were built were given. Indicator results were discussed, according to personnel expectations and operational knowledge.

Most of their recommendations included remarks about their experiences in the field. Those indicators requiring adjustments were extensively reviewed, consulting all available data sources. In other occasions, observations were merely related to aesthetical changes in some indicators, to establish consistency in the way these were presented.

Irregularities were eliminated in each case as required. New ways of depicting historical information were considered reliable only when experts in the field felt satisfied with results. Figure No. 3-22 explains this cycle, representing this process completely.

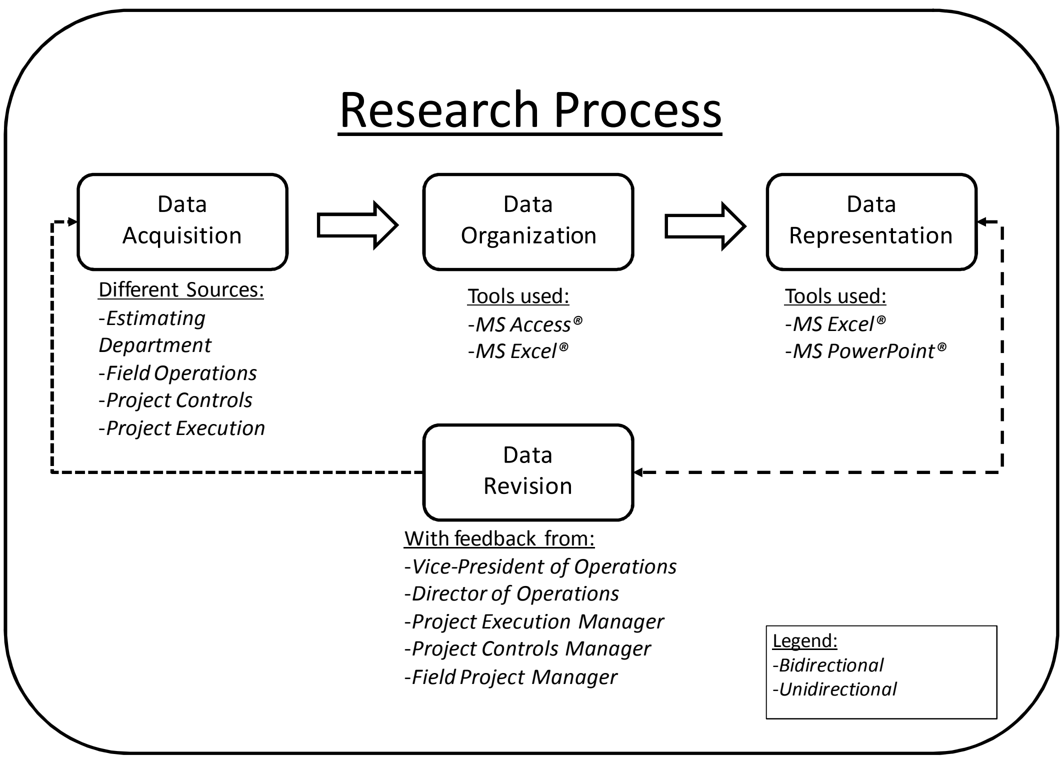

Figure 3-22: Research Process

3.9. Conclusions

Discussed in this the present chapter, data from different departments of an industrial construction company was collected in an eight month study. This information was characterized by having deficient organization and lack of consistency in their representation. No homologated structures between documents were observed. Moreover, the absence of those responsible for the development of previous documents caused delays when understanding reports, representing a limitation for data analysis.

The necessity of creating alternative ways to depict and organize historical data is important. Industrial Indicators were generated and reviewed in weekly meetings with company personnel, receiving positive reviews. In this way, additional value and knowledge is created by having tools designed to visualize past performances improving company decisions, specifically during proposal elaboration.

Upgrading a company’s operational knowledge with Industrial Indicators raises its business vision, integrating new information facets to support assessment during revision of historical performances. These alternative Industrial Indicators were adopted by the company and are currently used by their personnel as a reference of previous performances. Company knowledge has been documented in a compilation that can be consulted whenever required.

Chapter 4: Knowledge Discovery Techniques applied in case studies

4.1. Introduction

The present chapter constitutes the core of this thesis, introducing two main sections delimiting its content. Firstly, *Section 4.2* introduces a *Data Overview* of all obtained records for case study development: how these records are composed and which are the variables considered for analysis. Five industrial construction projects related to pipe module fabrication activities are presented. Alternatively, additional data samples were extracted from previous fabrication studies. As a product of this development, different spreadsheets, graphs and databases were assembled. Due confidentiality reasons, all graphs and tables present in the development of this chapter have been scaled.

Secondly, *Section 4.3* presents four case studies with application of *Clustering* and *Association rules* techniques using factual data from Company records. A description of applied methods and their outcomes is thoroughly presented. Its intention is to discover important characteristics of past projects and distinguish possible profiles and patterns. In addition, *Distribution Fitting* is applied to the last case, reviewing how distributions can resemble records present in raw data virtually creating input models for further construction simulation research.

Lastly, *conclusions* for all experiments are presented at the end of this chapter.

4.2. Data Overview

Analysis of company records was performed during the course of one year, once data from different departments of the company was acquired. Documentation from *Project Controls* and *Project Execution* departments was collected acting as raw data for three case studies. Two additional analysis scenarios were included in this chapter, using alternative data samples collected by previous researchers.

Data Ordering and *Cleaning* techniques applied to different spreadsheets and flat text files originated new records in the form of additional spreadsheets, graphs and databases. Furthermore, usage of these tools generated alternative ways of reviewing past performances through distinct Industrial Indicators creation.

However, certain characteristics of specific fabrication processes such as welding, handling Pipe, supports and handling valves were difficult to analyze by using simple approaches and visual displays. Scattered records and multiple behaviors and trends in each category were observed, challenging proper analysis. These records quantified several items and materials used during fabrication. Different item measures captured during fabrication of *piperack, pipe* and *stair modules* were contained in these sets of data.

For confidentiality reasons, all project names, quantities and other measures concerning historical project data have been scaled during discussion of case studies presented in this chapter.

4.2.1. Data Ordering and Cleaning

Project Execution and *Project Controls* departments are the prime sources of information studied for *Data Mining* and *Distribution Fitting* applications used in this chapter.

Project Execution supplied multiple files containing different fabrication characteristics for each scope project, including the following:

- Number of fabricated modules
- Length of pipe installed in modules
- Direct manhours for fabrication activities (both pipe installation and steel assembly)
- Carbon steel mix present in pipe fabrication (CS %)
- Man-hours per diameter inch weld fabricated (Mh/DIW)

These characteristics were extracted forming new arrangements called *Project Descriptions*.

Project Controls possessed files which included details related to *steel assembly* and *pipe fabrication* activities. For *steel assembly*, *Project Controls* supplied data which included different structural steel records, captured and quantified by field personnel.

Lists of several steel items used during assembly of steel structures supporting pipe modules were observed, detailed on large spreadsheets. Within these files, steel items used were classified according to their weight and by module number as their final destination. In addition, costs associated with each item were noted within these records. Extracting, ordering and cleaning information from only steel quantities created supporting versions representing percentages of utilization for each steel category:

- Extra Heavy %
- Heavy Steel %
- Medium Steel %
- Light Steel %
- Miscellaneous Steel %

These records are called *Structural Steel Utilization* and each steel category in steel assembly activities is represented by a percentage. Another important collection of data supported by *Project Controls* are *Quantity Takeoffs* files. These records included different configurations of pipes, valves, supports and others used during pipe spool fabrication.

Quantity Takeoffs represented outputs of a database management system run in FoxPro®, used by company personnel. These files contained unpolished and unordered data: disorganized raw outputs of pipe spool fabrication activities. *Welding*, *handling pipe*, *pipe supports* and *handling valves* were activities portrayed in these records.

Using spreadsheets, information derived from these files was ordered and cleaned of particular characters affecting negatively *Data Mining* software performance. Once this process was finished two different data sets were obtained.

The first data set collected and quantified information about all pipe diameters used during fabrication, for each project. Within this data set a utilization measure was calculated. In this way, each pipe diameter was represented by a percentage of utilization. This measure was called *Pipe Diameter Utilization Percentage*. The other data set contained ordered information related to four fabrication areas. It is represented by the following groups, classified by project:

Project A

Welding	33
Handling Pipe	39
Supports	51
Handing Valves	11
Sub-total Records:	*134*

Project B

Welding	99
Handling Pipe	282
Supports	1138
Handing Valves	169
Sub-total Records:	*1688*

Project C

Welding	132
Handling Pipe	140
Supports	589
Handing Valves	129
Sub-total Records:	*990*

Project D

Welding	131
Handling Pipe	166
Supports	843
Handing Valves	97
Sub-total Records:	*1237*

Project E

Welding	80
Handling Pipe	127
Supports	67
Handing Valves	305
Sub-total Records:	*579*

Table 4-1: Summary of records in each fabrication area

These data groups were called *Fabrication Main Characteristics*.

A third information source was defined by a previous study, *Fabrication Time Studies*, performed by PhD students of the Hole School of Construction of University of Alberta. Data was compiled in a MS Access® database, containing detailed fabrication information related to an external project outside the present research scope. This database included detailed fabrication data, comprising multiple attribute information specifying man-hours performed in each fabrication activity (e.g. cutting, welding and fitting). Required data for further analysis was obtained from this database through query development and data tables. Additional modifications to the data were performed using spreadsheets. As a result of *Data Ordering and Cleaning* techniques, five *Case Studies* were arranged.

Case Study No. 1 was formed by grouping Project Descriptions, Structural Steel Utilization and Pipe Diameter Utilization percentages. This case study attempts detection of different project profiles by creating comparisons between historical and potential projects using factual data. Through project profiling, the company will have an alternative view to review the closest match between current opportunities and past performances and potentially decide its bid starting prices.

Case Study No. 2 collects information from activities performed during pipe spool fabrication in past projects: *welding, handling pipe, supports and handling valves*. The purpose of this case study is to detect main trends embedded in these activities for each historical project. Its detection will increase analysis and comprehension of past performances. By comparing characteristics of activities included in potential projects with previous executions, the company reinforces its bidding strategies by strengthening knowledge of previous operations.

Case Study No. 3 is similar in structure to *Case Study No. 2*. Both cases also acquired data from the same data source. The difference between these cases lies in the approach used while ordering data. In this case, all project data is merged as one knowledge source. Once this step is performed, data is then divided into four fabrication areas. *Clustering* techniques, *Association Rules* and graphical representations are applied to enhance understanding of these fabrication activities.

Case Study No. 4 introduces an extract of data related to pipe spool fabrication activities collected from the *Fabrication Time Studies* database. This data has 295 unique records which combine six different attributes. In this case study, *Statistical Analysis* is performed to detect particular distributions of man-hours performed in each weld. In addition, *Clustering* is applied in conjunction with *Association Rules* to detect knowledge enclosed in data.

Case Study No. 5 presents the largest data sample from all experiments. 19,960 records are gathered in a table containing five attributes. This is the most detailed data comprising welding information classified by fabricated spools. The purpose of this experiment is to determine the best distributions derived from raw data resembling project performances for certain data attributes.

The process of *Data Ordering and Cleaning* for all *Case Studies* involving historical data is presented in Figure 4-1 and Figure 4-2. These figures display all knowledge contributions from different departments in each of the cases that are explained in the present chapter.

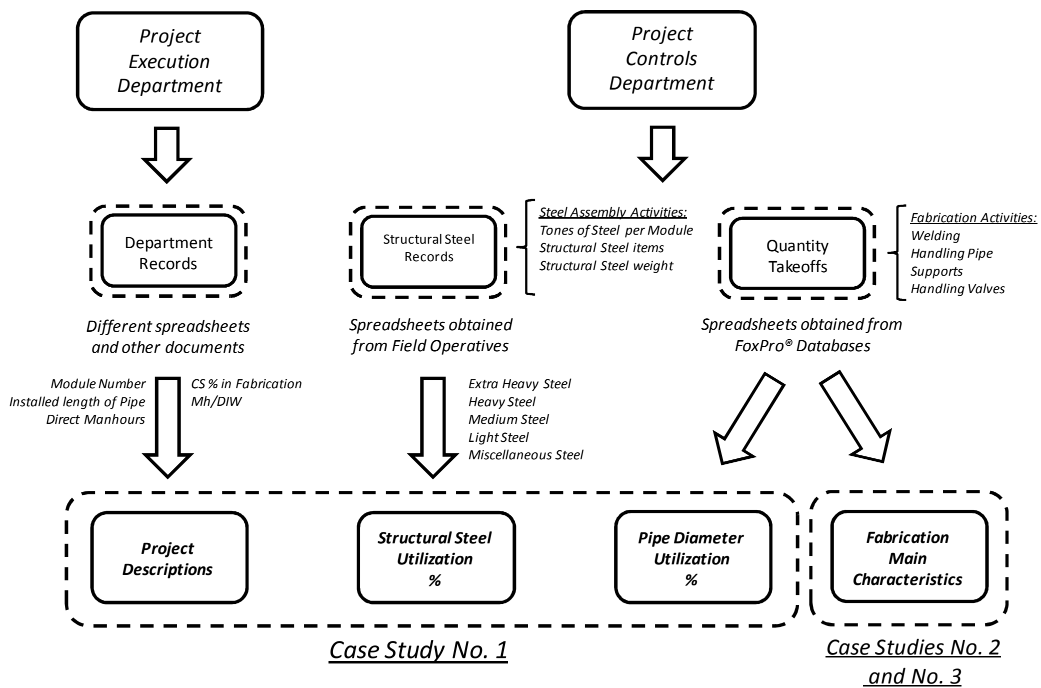

Figure 4-1: *Data Acquisition, Ordering and Cleaning* processes for Cases No.1 to No.3

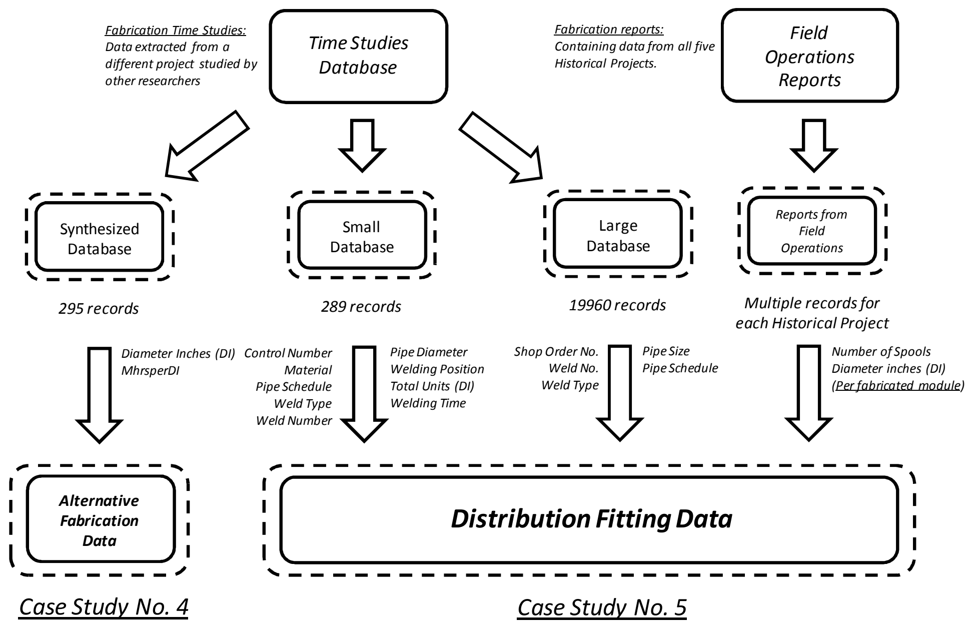

Figure 4-2: *Data Acquisition, Ordering and Cleaning processes* for Cases No.4 and No.5

4.2.2. Limitations of acquired data in Historical Projects

Data acquisition was a process filled with incertitude: multiple formats and different ways of presenting data measures embodied some of the difficulties found during this process. Information concerning man-hours invested in each activity was not entirely present in historical records. Only theoretical information was portrayed in these files. This data is derived from performance tables developed in-house by the company. These are still used today as starting point when developing bidding proposals.

Missing information quantities varied between historical projects. Certain files and spreadsheets contained more complete records than others, depending on activities performed and sources of information. Actual measures of man-hours spent in fabrications activities were not collected by the company. This represented a constraint for analysis of past performances. The company lacked the technology necessary to capture real-time performances, specifically during particular fabrication tasks.

Managerial reports contained information regarding fabrication but were collected in an aggregated manner, establishing an amount of man-hours per category at a very high level (e.g. pipe spool erection 2”-10”). This does not reflect all behaviors observed in each fabrication operation.

4.3. Analysis of Industrial Construction activities

To review the information and discover new knowledge from company records, WEKA® software was chosen as *Data Mining* tool. This program has the capability of running different types of *Data Mining* algorithms and it possesses a friendly user interface for those that are not experts in the Knowledge Discovery field. As Witten & Frank (2005) mentioned in their work, WEKA® has an integrated collection of algorithms and pre-processing tools designed to recognize best analysis alternatives for particular scenarios, through approach association between different algorithms. In addition, the authors mentioned some of the algorithms embedded in WEKA®:

- Classification
- Association
- Clustering

- Data Visualization

Examples of some of the algorithms that can be run in WEKA® are presented in Figure 4-3.

Examples of algorithms used by WEKA® Software:

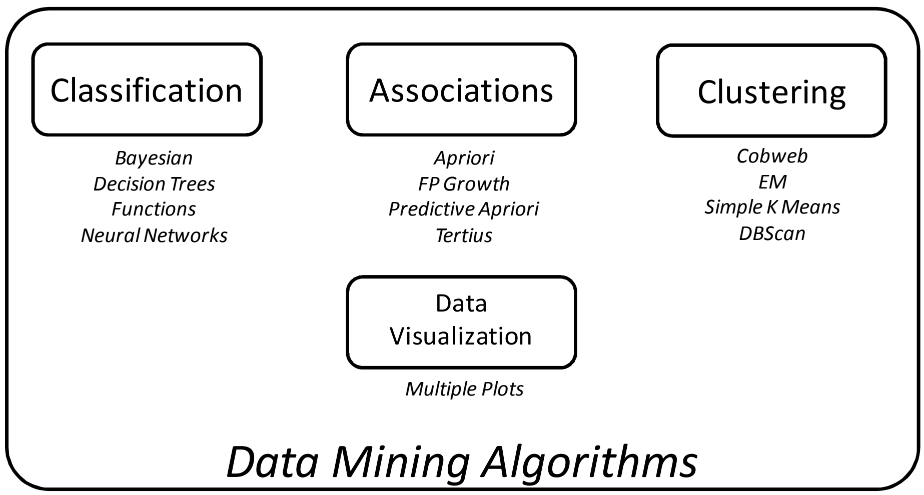

Figure 4-3: Different Tasks performed by WEKA® (Witten and Frank, 2005)

Depending on analysis scenario, quality and number of data items certain algorithms will be more suitable than others while mining data.

Analysis of four different case scenarios through application of *Clustering* algorithms is presented. These methods are implemented in those occasions in which attributes are associated into spontaneous groups according to their similarities, for cases in which a class prediction is not required (Witten & Frank, 2005).

4.3.1. Case Study 1: Project profiling

Data was extracted from company records, after which ordering and cleaning techniques were applied creating multiple spreadsheets. These spreadsheets were later uploaded into a MS Access® database, constituting tables used for further data manipulation. In this experiment, a single query was created using MS Access®, reflecting the following fields:

- Project name
- Module number
- Pipe Diameter Utilization: pipe diameters between < 2” and 48”
- Structural Steel Utilization: percentages of Extra-Heavy, Heavy, Medium, Light and Miscellaneous steel used during Steel assembly
- Installed length of pipe (both in meters and feet)
- Direct man-hours involved in steel assembly
- Direct man-hours involved in pipe installation
- Carbon steel percentage present in fabrication (CS %)
- Man-hours per diameter inch of weld produced (Mh/DIW)

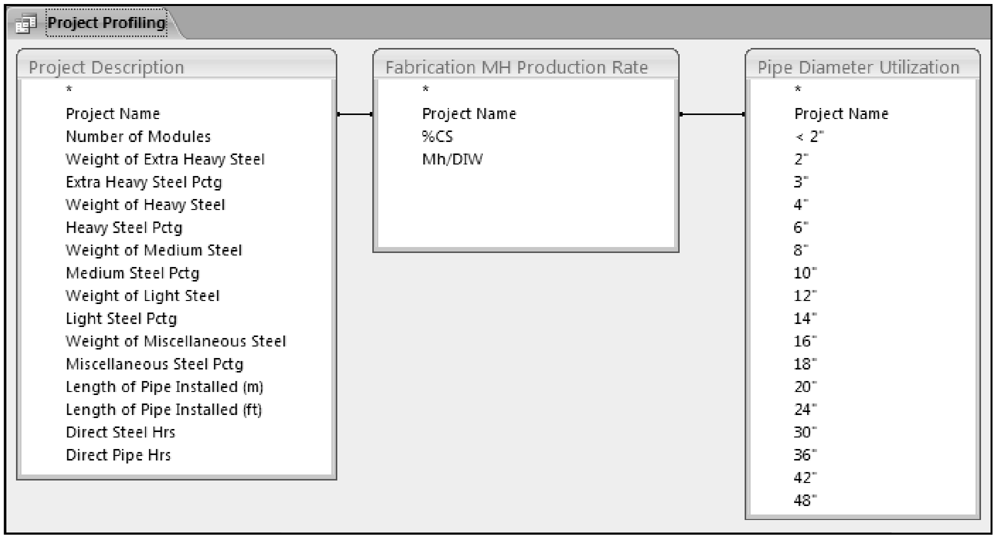

Figure 4-4: Database Query for Case Study 1

As an output of this query, a spreadsheet was obtained and special characters and spaces were removed to prepare it for *Data Mining* analysis using WEKA®. The *Simple K Means* algorithm was the tool chosen in this first *Case Study*. It is used specifically to create five different clusters, resembling each project constituting research scope. Each cluster represents unique historical characteristics, forming diverse project profiles.

According to Witten & Frank (2005), this clustering technique allocates different items in randomly-chosen centroids by calculating the *euclidean distance* between items and centroids. As initial step during this analysis, the number of centroids in this algorithm has been set by the user, defining an initial *k* factor. As mentioned by these authors, once cluster assignments are finalized, means for each cluster are calculated, reassigning new cluster centroids. Once this feature has been performed, the entire process will be repeated until the value of the cluster centers becomes constant.

By training this algorithm to specify *five* cluster centers, a base structure with *five* different project profiles is trained. This enables *Simple K Means* algorithm to make profile comparisons by testing it against new data derived from potential projects.

It is important to mention that new data must have a similar structure to those observed during training. In other words, to make proper comparisons the same distribution of features as described at the beginning of this *Case Study* must be present. The experiment performed in the analysis of *Case Study No. 1* exhibits a comparison of singular project profiles and also a classification of potential project profiles according to similarities of previous company experiences.

Moreover, as a result of this experiment, the company will know the closest match between characteristics of a potential project and those previously performed.

With this, the company can begin defining its base starting prices having at hand useful information about possible trends present in current characteristics.

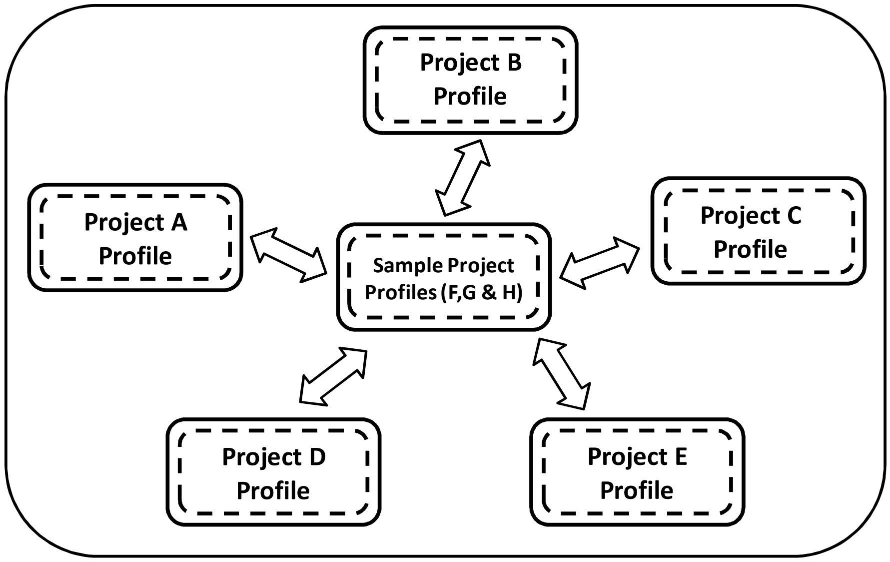

Figure 4-5: Profile comparison between Historical and Potential projects

This technique improves the way the company analyzes and compares potential projects with its previous performances. It is based on a multi-dimensional analysis in which different characteristics are compared simultaneously rather than uni-dimensionally as currently performed by personnel through use of simpler charts derived from previous executions. Using past data, the algorithm was trained to create five different projects as clusters. These were randomly-created, assigning different cluster numbers to previous projects:

Project Profile A: Cluster 2

Project Profile D: Cluster 0

Project Profile B: Cluster 1

Project Profile E: Cluster 4

Project Profile C: Cluster 3

Each of the five clusters created presents a structure depicted in Figure 4-5. In addition, clustering assignments can be observed in Figure 4-6, in which five different centroids have been defined, corresponding to each historical project.

Moreover, six different Industrial construction projects have been chosen as samples to test this clustering algorithm:

- Project Profile F
- Project Profile G
- Project Profile H
- Project Profile I
- Project Profile J
- Project Profile K

Clusterer output

Cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(5)	(1)	(1)	(1)	(1)	(1)
ProjectName	ProjectA	ProjectD	ProjectB	ProjectA	ProjectC	ProjectE
NumberofModules	68.6	123	95	40	40	45
Less2	0.0724	0.044	0.1265	0.0417	0.1252	0.0248
2	0.3252	0.4205	0.3521	0.2704	0.3285	0.2544
3	0.1376	0.1423	0.1688	0.0039	0.1241	0.2488
4	0.151	0.0817	0.094	0.3245	0.1507	0.1041
6	0.1394	0.1249	0.1328	0.1088	0.1579	0.1725
8	0.0893	0.0977	0.0664	0.0555	0.1093	0.1177
10	0.0709	0.059	0.0789	0.0766	0.0551	0.0846
12	0.0315	0.0581	0.0248	0.0377	0.0044	0.0322
14	0.0162	0.0036	0.0052	0.0588	0.0131	0.0001
16	0.0195	0.0212	0.0135	0.0561	0.0042	0.0023
18	0.0141	0.0181	0.0096	0.0245	0.0063	0.012
20	0.0194	0.0109	0.0207	0.0415	0.0114	0.0127
24	0.0081	0.0036	0.0052	0	0.0097	0.022
30	0.0023	0.006	0.0014	0	0	0.0043
36	0.0015	0.0042	0	0	0	0.0032
42	0.0013	0.0038	0	0	0	0.0028
48	0.0004	0.0002	0	0	0	0.0015
ExtraHeavySteelPct	0.183	0.0262	0.3466	0.3706	0.1714	0
HeavySteelPct	0.3507	0.4566	0.1734	0.1762	0.4287	0.5184
MediumSteelPct	0.3434	0.3376	0.3263	0.3468	0.3303	0.3761
LightSteelPct	0.1111	0.1256	0.1663	0.0916	0.0789	0.0928
MiscSteelPct	0.1119	0.154	0.0875	0.1147	0.0908	0.1126
PipeLengtM	25313.288	51256.81	24550.9	16203.748	20062.482	14492.5
PipeLengtFt	83049.0584	168165.382	80547.577	53161.9	65820.7	47549.733
DirectSteelHrs	30468.68	66474.1	25879.7	10886.7	28788.1	20314.8
DirectPipeHrs	42611.58	91153.7	39314	16424.1	50644	15522.1
CSPercentage	0.994	0.99	0.99	1	0.99	1
MhperDIU	0.781	0.77	1.045	0.715	0.715	0.66

Figure 4-6: Cluster creation. Initial Centroids

4.3.1.1. Results obtained during Project Profiling

As a main result of the application of *Simple K Means* algorithm, different *Clustering* assignments were observed for each test project. Each was run independently, obtaining unique allocations.

All measures involved in project profiling for Projects E, F and H are depicted in Table 4-2. Missing records are depicted with a question mark (*?)* character.

	Project Name	***Project F***	***Project G***	***Project H***
	Module Number	29	68	17
Pipe Diameter Utilization	*<2"*	17.95%	6.27%	2.13%
	2"	23.47%	5.73%	23.62%
	3"	23.41%	2.00%	17.43%
	4"	15.95%	0.99%	11.07%
	6"	8.56%	0.00%	0.00%
	8"	9.09%	0.16%	0.00%
	10"	5.66%	0.37%	0.00%
	12"	0.00%	0.46%	0.00%
	14"	0.00%	0.20%	0.00%
	16"	0.02%	0.20%	0.00%
	18"	4.50%	0.21%	0.00%
	20"	1.40%	0.00%	0.00%
	24"	0.00%	0.14%	0.00%
	30"	0.00%	0.00%	0.00%
	36"	0.00%	0.00%	0.00%
	42"	0.00%	0.00%	0.00%
	48"	0.00%	0.00%	0.00%
Steel Utilization	ExtraHeavyPct	?	0.00%	?
	HeavyPct	?	0.00%	?
	MediumPct	?	97.60%	?
	LightPct	?	0.00%	?
	MiscPct	?	12.40%	?
Characteristics	LengthPipeM	3,786.65	22,061.09	6,532.47
	LengthPipeFt	12,423.40	72,378.90	21,432.00
	DirectSteelHrs	7,308.40	21,227.80	935.00
	DirectPipeHrs	32,639.20	34,927.20	10,644.00
	%CS	99%	81%	100%
	Mh/DIW	0.94	1.27	0.78

Cluster Classification Method			
Simple K Means	**3**	**4**	**2**

Table 4-2: Potential Projects F, G and H. Measures involved during *Clustering*

The closest profile matches between historical and Projects F, G and H were the following:

- For *Project Profile F*, its closest match found by *K Means Algorithm* is Cluster 3, represented by *Project Profile C*.
- In the case of *Project Profile G*, its closest match is Cluster 4, portrayed by *Project Profile D*.
- Finally for *Project Profile H*, its equivalent is determined by Cluster 2, containing information of *Project Profile B*.

Cluster 3		**Cluster 4**		**Cluster 2**	
Project F	Project C	Project G	Project D	Project H	Project B
29	40	68	123	18	95
17.95%	12.52%	6.27%	4.40%	2.13%	12.65%
23.47%	32.85%	5.73%	42.05%	23.62%	35.21%
23.41%	12.41%	2.00%	14.23%	17.43%	16.88%
15.95%	15.07%	0.99%	8.17%	11.07%	9.40%
8.56%	15.79%	0.00%	12.49%	0.00%	13.28%
9.09%	10.93%	0.16%	9.77%	0.00%	6.64%
5.66%	5.51%	0.37%	5.90%	0.00%	7.89%
0.00%	0.44%	0.46%	5.81%	0.00%	2.48%
0.00%	1.31%	0.20%	0.36%	0.00%	0.52%
0.02%	0.42%	0.20%	2.12%	0.00%	1.35%
4.50%	0.63%	0.21%	1.81%	0.00%	0.96%
1.40%	1.14%	0.00%	1.09%	0.00%	2.07%
0.00%	0.97%	0.14%	0.36%	0.00%	0.52%
0.00%	0.00%	0.00%	0.60%	0.00%	0.14%
0.00%	0.00%	0.00%	0.42%	0.00%	0.00%
0.00%	0.00%	0.00%	0.38%	0.00%	0.00%
0.00%	0.00%	0.00%	0.02%	0.00%	0.00%
?	17.14%	0.00%	2.62%	?	34.66%
?	42.87%	0.00%	45.66%	?	17.34%
?	33.03%	97.60%	33.76%	?	32.63%
?	7.89%	0.00%	12.56%	?	16.63%
?	9.08%	12.40%	15.40%	?	8.75%
3,786.65	20,062.48	22,061.09	51,256.81	7,185.72	24,550.90
12,423.40	65,820.70	72,378.90	168,165.38	23,575.20	80,547.58
7,308.40	28,788.10	21,227.80	66,474.10	1,028.50	25,879.70
32,639.20	32,918.60	34,927.20	68,365.28	11,708.40	27,519.80
99.00%	80%	88%	80%	110%	80%
0.9427	0.715	1.2661	0.77	0.7766	1.045

Table 4-3: Potential Projects F, G and H cluster assignments. *Simple K Means* algorithm implementation (results in right columns)

Cluster 3		Cluster 4		Cluster 4	
Project I	Project C	Project J	Project D	Project K	Project B
108	123	73	45	78	45
0.19%	4.40%	1.59%	2.48%	0.17%	2.48%
5.64%	42.05%	29.41%	25.44%	22.96%	25.44%
23.04%	14.23%	6.00%	24.88%	14.16%	24.88%
5.26%	8.17%	8.02%	10.41%	10.49%	10.41%
10.93%	12.49%	9.80%	17.25%	13.17%	17.25%
6.12%	9.77%	9.53%	11.77%	16.68%	11.77%
8.42%	5.90%	7.37%	8.46%	6.08%	8.46%
7.28%	5.81%	13.90%	3.22%	12.77%	3.22%
7.63%	0.36%	0.92%	0.01%	6.17%	0.01%
11.12%	2.12%	9.55%	0.23%	4.13%	0.23%
1.85%	1.81%	4.29%	1.20%	0.65%	1.20%
2.98%	1.09%	5.08%	1.27%	0.33%	1.27%
3.76%	0.36%	3.78%	2.20%	2.23%	2.20%
7.64%	0.60%	0.77%	0.43%	0.00%	0.43%
3.76%	0.42%	0.00%	0.32%	0.00%	0.32%
4.39%	0.38%	0.00%	0.28%	0.00%	0.28%
0.00%	0.02%	0.00%	0.15%	0.00%	0.15%
0.00%	2.62%	0.00%	0.00%	0.00%	0.00%
33.70%	45.66%	21.22%	51.84%	11.11%	51.84%
50.87%	33.76%	55.84%	37.61%	63.99%	37.61%
17.31%	12.56%	19.10%	9.28%	22.21%	9.28%
8.12%	15.40%	13.85%	11.26%	12.69%	11.26%
36569.53	51256.81	19195.60	14492.50	36769.57	14492.50
119977.30	168165.38	62976.92	47549.73	120633.60	47549.73
34325.28	66474.10	19497.46	20314.80	23255.23	20314.80
64812.31	68365.28	43873.28	11641.58	49131.94	11641.58
?	80%	?	83%	?	83%
?	0.77	?	0.66	?	0.66

Table 4-4: Potential Projects I, J and K cluster assignments. Simple K Means algorithm implementation (results in right columns)

4.3.1.2. Further analysis using alternative *Clustering* techniques

To review the results of this experiment, different runs with four alternative clustering algorithms were performed:

- *Simple EM*
- *Farthest First*
- *Filtered Clusterer*
- *Make Density Based Clusterer*

A similar procedure of training WEKA® to create historical clusters stated in the first part of this experiment was engaged, selecting this time different algorithms. For each case the amount of clusters to be created remained equal, testing a new sample of data at a time to establish its classification.

Once testing ended, results of this classification were obtained. In some occasions, clusters assignments were identical to those accomplished during *Simple K Means* testing. In others, these allocations were different but consistent to those project profiles initially acquired. For example, during *Simple K Means* testing, Cluster 3 was assigned to *Project C*, defining the closest historical project profile to *Project F*. Furthermore, while verifying this hypothesis using *Farthest First* algorithm, a different cluster appeared. This cluster was Cluster 4.

However, all results for this example remained consistent, determining *Project C* as the closest historical project profile to *Project F* no matter cluster allocation. The same behavior was observed during testing of the remaining five sample projects. Depending on the *Clustering* algorithm, centroid calculation and clustering assignments occurred differently. The results of these algorithms are represented in Table No. 4-5:

Project Name	Clustering Algorithms				
Project F	**K Means**	EM	Farthest First	Filtered	Make Density
Cluster Assignment	**3**	**3**	**4**	**3**	**3**
Project Name	***Project C***	***Project C***	***Project C***	***Project C***	***Project C***
Project Name	***Clustering Algorithms***				
Project G	K Means	EM	Farthest First	Filtered	Make Density
Cluster Assignment	**4**	**2**	**2**	**4**	**4**
Project Name	***Project E***	***Project E***	***Project E***	***Project E***	***Project E***
Project Name	***Clustering Algorithms***				
Project H	K Means	EM	Farthest First	Filtered	Make Density
Cluster Assignment	**2**	**1**	**0**	**2**	**2**
Project Name	***Project A***	***Project A***	***Project A***	***Project A***	***Project A***
Project Name	***Clustering Algorithms***				
Project I	K Means	EM	Farthest First	Filtered	Make Density
Cluster Assignment	**0**	**0**	**1**	**0**	**0**
Project Name	***Project D***	***Project D***	***Project D***	***Project D***	***Project D***
Project Name	***Clustering Algorithms***				
Project J	K Means	EM	Farthest First	Filtered	Make Density
Cluster Assignment	**4**	**2**	**2**	**4**	**4**
Project Name	***Project E***	***Project E***	***Project E***	***Project E***	***Project E***
Project Name	***Clustering Algorithms***				
Project K	K Means	EM	Farthest First	Filtered	Make Density
Cluster Assignment	**4**	**2**	**2**	**4**	**4**
Project Name	***Project E***	***Project E***	***Project E***	***Project E***	***Project E***

Table 4-5: Results comparisons of different *Clustering* algorithms

To summarize all results obtained, Figure 4-7 presents results of application of *Clustering* algorithms to determine closest matches between potential and historical project profiles:

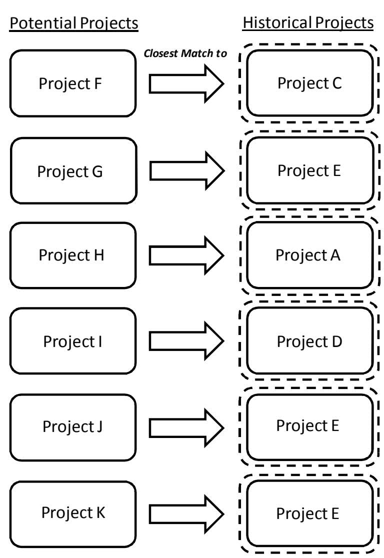

Figure 4-7: Suggested profiles by *Clustering* Algorithms

4.3.1.3. Case Study 1 Limitations

Project profiling through *Clustering* application presents two major limitations:

- In this multi-dimensional analysis, all attributes are considered equal. Their unique importance is not defined within this analysis. This does not represent factual conditions present in pipe module fabrication. Establishing weight percentages for each fabrication attribute can improve profile determination by creating a more realistic model with an improved fabrication structure, in which attributes are evaluated differently and aligned to company requirements.
- Fixing cluster amounts to determined quantities are not recommended. The purpose of this unsupervised data mining technique is to create groups from multiple records, randomly. In this case study, clustering has been determined through unique and finite historical project characteristics. Comparisons between historical and potential project data are based on an installed structure created by the user. Even though project profiling represents an interesting topic, other analysis methods must be explored to reach improved outcomes in this area.

4.3.2. Case Study 2: Project Characterization using fabrication quantities

Only Quantity Takeoffs files were considered during development of this particular experiment. These represented one of the most detailed samples of data found in the company archives. It depicts every item used in pipe module fabrication activities of historical projects.

These files were ordered, cleaned and divided in four main fabrication areas:

- *Welding.*
- *Pipe supports.*
- *Handling pipe.*
- *Handling valves.*

Welding includes all different weld types used to join different metallic pieces, forming a larger fabrication component (e.g. butt welds, o-let welds quantified by unit). *Handling pipe* depicts different pipe types and materials used in fabrication of pipe modules (e.g. handle pipe XS CS: pipe with presenting a XS schedule type and made of carbon steel, in meters). *Pipe supports* shows different parts used to sustain pipe spool configurations (e.g. Pipe support AC70, according to company's specifications). *Handling valves* presents all valve types used during fabrication of pipe spools (e.g. handle manual valve, 150 lb made of carbon steel, determined by piece). Sizes for all measures are represented in inches (").

Once this categorization process was completed, *Simple K Means and DBScan* algorithms were used to compile records and detect unique patterns that may highlight main characteristics of each fabrication area in an industrial construction project. The *Clustering* structure is explained in Figure 4-8.

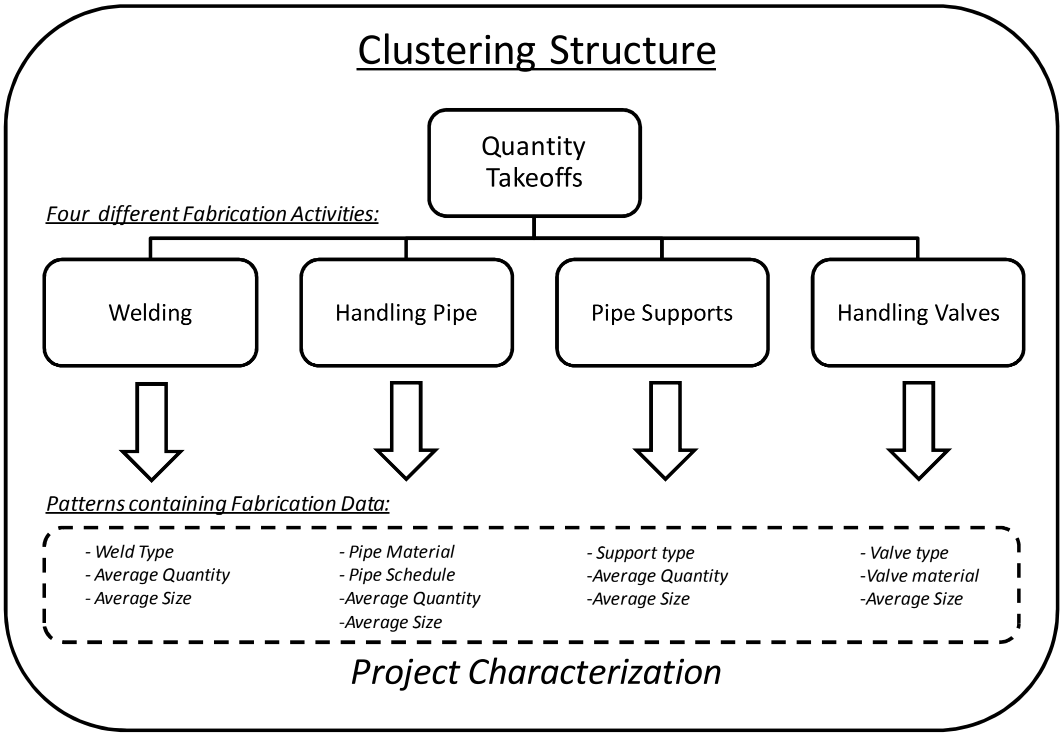

Figure 4-8: Clustering Structure for Project Characterization

Clustering of historical projects was performed individually analyzing single fabrication areas one at a time. Once this process is finished, weighted averages of predominant groups within all fabrication areas were determined and presented as key factors. Clusters with great presence within these areas represent average project characteristics. These can highlight unique project features to decision makers, supporting bidding decisions when tendering.

By consulting these main groups, the company will know which characteristics were often present establishing patterns and identifying key factors in past event fabrication activities.

4.3.2.1. Parameters specified in clusters derived from this study

Results for each historical project are presented using tables containing the following parameters:

- Total Clusters: Represents the total amount of clusters created by the selected *Clustering* algorithm. Depending on the algorithm type, cluster creation can be supervised specified in *Simple K-Means algorithm* (MacQueen, 1967) or unsupervised, created automatically by *DBScan algorithm* (Zaïane et al., 2002).
- R squared error: Specifies the R Squared Error percentage obtained during application of clustering algorithms. It measures in what manner forecasted values in a trendline are aligned to current figures (Winston, 2011).
- Cluster No.: Denoted the cluster number with highest percentage of presence in the data sample.
- Presence: A percentage representing the amount of items clustered from total data.
- Description: Illustrates which are the characteristics of the item present in the cluster with greatest presence.
- Avg. Quantity: Calculates a measure representing the average quantity of an item representing a cluster's center.

- Average Size: similar to the previous measure, calculates an average size for a particular item (in inches).

4.3.2.2. Project A Characterization

Project A Main Characteristics. Fabrication Activities.

Welding	*Total Clusters*	12
	R Square Error	27.43%
	Cluster No.	5
	Prescence	19%
	Description	FLANGE WELDNECK, STD, 150 lb
	Avg Quantity	16.86663
	Avg Size	2.2
Handling Pipe	*Total Clusters*	4
	R Square Error	28.85%
	Cluster No.	14
	Prescence	37.00%
	Description	HANDLE PIPE, XS, CS
	Avg Quantity	15.50219
	Avg Size	1.73646
Supports	*Total Clusters*	22
	R Square Error	27%
	Cluster No.	20
	Prescence	10%
	Description	PIPE SUPPORT, LS12
	Avg Quantity	2.64
	Avg Size	6.6
Handling Valves	*Total Clusters*	7
	R Square Error	30.7%
	Cluster No.	3
	Prescence	20%
	Description	HANDLE MANUAL VALVE, 150lb, CS
	Avg Quantity	172.205
	Avg Size	1.65

Table 4-6: Project A Characterization, using a *Simple K Means algorithm*

Table 4-6 presents compiled results derived from the utilization of *Simple K Means algorithm.* Clusters having the largest presence within total data in each fabrication areas were selected.

From all produced clusters, only those fulfilling the above mentioned requirement were considered as representative samples for each fabrication activity. This restriction is also applied in Projects B to E, in which *DBScan algorithm* was used to analyze scattered data present in each fabrication sample.

4.3.2.3. Project B Characterization

Project B Main Characteristics. Fabrication Activities.

Welding	*Total Clusters*	3
	Cluster No.	0
	Prescence	43%
	Description	BUTTWELD, STD, LOW TEMP CS
	Avg Quantity	1.32
	Avg Size	6.424
Handling Pipe	*Total Clusters*	15
	Cluster No.	13
	Prescence	10%
	Description	HANDLE PIPE, STD, LOW TEMP CS
	Avg Quantity	8.536
	Avg Size	16.522
Supports	*Total Clusters*	11
	Cluster No.	7
	Prescence	11%
	Description	HANDLE PIPE, Sch80, CHROME MOLY
	Avg Quantity	1.892
	Avg Size	2.629
Handling Valves	*Total Clusters*	9
	Cluster No.	2
	Prescence	13%
	Description	HANDLE MANUAL VALVE, 150lb, CS
	Avg Quantity	3.586
	Avg Size	1.584

Table 4-7: Project B Characterization, using a *DBScan algorithm*

4.3.2.4. Project C Characterization

Project C Main Characteristics. Fabrication Activities.

Welding	Total Clusters	2
	Cluster No.	0
	Prescence	67%
	Description	HYDROTEST END WELDCAP, CS
	Avg Quantity	3.333
	Avg Size	6.281
Handling Pipe	Total Clusters	6
	Cluster No.	5
	Prescence	18%
	Description	HANDLE PIPE, Sch80, CS
	Avg Quantity	155.639
	Avg Size	2.871
Supports	Total Clusters	8
	Cluster No.	0
	Prescence	14%
	Description	PIPE SUPPORT INSTALL, SA6CS
	Avg Quantity	23.287
	Avg Size	5.346
Handling Valves	Total Clusters	8
	Cluster No.	0
	Prescence	19%
	Description	HANDLE MANUAL VALVE, 300lb, CS
	Avg Quantity	7.766
	Avg Size	3.663

Table 4-8: Project C Characterization, using a *DBScan algorithm*

4.3.2.5. Project D Characterization

<u>Project D Main Characteristics. Fabrication Activities.</u>

Welding	*Total Clusters*	6
	Cluster No.	5
	Prescence	20%
	Description	BUTTWELD, STD, CS
	Avg Quantity	5.753
	Avg Size	8.041
Handle Pipe	*Total Clusters*	9
	Cluster No.	0
	Prescence	16%
	Description	HANDLE PIPE, STD, CS
	Avg Quantity	450.571
	Avg Size	8.734
Supports	*Total Clusters*	11
	Cluster No.	1
	Prescence	14%
	Description	PIPE SUPPORT INSTALL, SHIMCS
	Avg Quantity	12.122
	Avg Size	21.043
Handle Valves	*Total Clusters*	7
	Cluster No.	3
	Prescence	19%
	Description	HANDLE MANUAL VALVE, 300 lb, CS
	Avg Quantity	27.126
	Avg Size	5.555

Table 4-9: Project D Characterization, using a *DBScan algorithm*

4.3.2.6. Project E Characterization

		Project E Main Characteristics. Fabrication Activities.
Welding	*Total Clusters*	1
	Cluster No.	0
	Prescence	100%
	Description	BUTTWELD, STD, CS
	Avg Quantity	12.76
	Avg Size	6.831
Handling Pipe	*Total Clusters*	5
	Cluster No.	0
	Prescence	25%
	Description	HANDLE PIPE, STD, CS
	Avg Quantity	182.677
	Avg Size	8.437
Supports	*Total Clusters*	3
	Cluster No.	1
	Prescence	44%
	Description	PIPE SUPPORT INSTALL, SHIMCS
	Avg Quantity	5.841
	Avg Size	16.973
Handling Valves	*Total Clusters*	3
	Cluster No.	0
	Prescence	35%
	Description	HANDLE CONTROL VALVE, 300lb, CS
	Avg Quantity	3.784
	Avg Size	2.959

Table 4-10: Project E Characterization, using a *DBScan algorithm*

4.3.2.7. Case Study 2 Limitations

This experiment involved application of two different *Clustering* algorithms in historical data, divided in four fabrication areas:

- *Simple K Means algorithm* (used during analysis of *Project A* characteristic determination)

- *DBScan algorithm* (used for the analysis of *Projects B to E*)

Clustering techniques were chosen to simplify the analysis of scattered data and to observe key points that could help define trends within fabrication areas. Normally for validation purposes, 80% of the total data is used for training and 20% for testing. However, due to the limited amount of records present in historical projects (specified in Table 4-1), the totality of data was taken into account for analysis. The entirety of records was used as training data during clustering, impeding proper validation.

A low number of records were observed in each project. In addition, most of these were incomplete and disperse. Limitations were found due to these circumstances. Furthermore, another limitation encountered was the effectiveness of the *Simple K Means* clustering algorithm in limited and scattered amount of data. Large cluster numbers (surpassing 20 clusters) and unexpected *R squared error* measures (above 100%, indicating possible association problems between expected and actual values) were obtained when analyzing project data. This was a sign indicating *Simple K Means* was not an algorithm suitable for analysis, due to highly sparse data. Therefore, *Simple K Means Algorithm* was only used during analysis of data related to *Project A*, in which plausible results were collected.

The decision on how many initial clusters were going to be specified was based on two particular conditions. After several iterations experimenting with different *k* clusters values, the following conditions were established by the user:

- Obtained *R Squared Error* must be below 30%.
- The *cluster number* created has to be inferior to 25 clusters.

Due to inadequate results retrieved during application of *Simple K Means* in analysis of remaining projects (*Projects B to E*), an alternative analysis instrument was chosen. *DBScan algorithm* was selected as an additional method of reviewing and detecting trends in past performances. As Ester et al.(1996) mentioned in their research, this algorithm has the capability of detecting clusters characterized by having random outlines. In addition, the same authors mentioned this algorithm has the capability of detecting noise in a database. Results achieved in analysis of project data using *DBScan algorithm* can be observed in Tables 4-7 to 4-10. This method identified main clusters present in scattered data under the same conditions specified in the application of *Simple K Means* algorithm, in terms of *R squared errors* result and *cluster number*.

4.3.3. Case Study 3: Analysis of aggregated quantities of fabrication areas

The techniques used during buildup of this case study are similar to those observed during the development of *Case Study No. 2*. Certain steps were required to achieve relevant results.

Firstly, data belonging to historical projects contained in Quantity Takeoffs files was compiled into a single database. In this way, all project data was unified. Secondly, the database was divided into four main fabrication areas. Each section contained condensed data from multiple historical projects. These fabrication areas were:

- Welding
- Supports
- Handling pipe
- Handling valves

Thirdly, a *DBScan algorithm* was chosen as tool of analysis, similar to previous experiments. Using this algorithm an unsupervised quantity of clusters was created in each area. The largest cluster from all groups was selected. Table 4-11 presents a sample of the clustered data for *handling pipe* activities. All examples presented in this case study belong to this fabrication area. Table 4-12 depicts all clusters created with their respective assigned instances in the *Handling pipe* area.

DESC	QTY	SIZEDESC	UNITPRICE	MH		Cluster
HANDLEPIPESTDCS	59.2	16	278.1	0.81	-->	0
HANDLEPIPEXSCS	2.7	0.5	30.48	0.09	-->	1
HANDLEFO300	1.2	16	1557.57	2.63	-->	0
HANDLEPIPESTDCS	190.9	4	72.38	0.23	-->	6
HANDLEPIPESTDLOWTEMPCS	151.8	4	72.38	0.23	-->	6
HANDLEPIPESTDLOWTEMPCS	84.8	2	41.91	0.13	-->	2
HANDLESHIPLOOSE150VLV	4.6	2	84.79	2.63	-->	2
HANDLEPIPELESS20FTSTDLOWTEMPCS	29.9	2	167.62	0.13	-->	2
HANDLEPIPESTDCS	98.6	2	41.91	0.13	-->	2
HANDLEPIPESTDCS	62.1	14	251.44	0.75	-->	13
HANDLEPIPELESS20FTSTDCS	21.5	2	167.62	0.13	-->	2
HANDLEMANUALVALVE150lbCS	20.7	2	84.79	0.81	-->	2
HANDLEPIPESTDCS	31.7	8	156.19	0.46	-->	4
HANDLEPIPEXSCS	7.3	1	41.91	0.1	-->	9
HANDLEPIPELESS20FTSTDLOWTEMPCS	4.5	4	289.54	0.23	-->	6
HANDLEMANUALVALVE150lbLOWTEMPCS	3.5	4	155.64	1.5	-->	6
HANDLEPIPESTDCS	39.8	12	224.78	0.69	-->	5
HANDLEPIPESTDCS	1.3	3	57.14	0.18	-->	8
HANDLEPIPELESS20FTSTDCS	20.1	12	899.09	0.69	-->	5
HANDLEMANUALVALVE150lbCS	4.6	3	120.8	1.15	-->	8
HANDLEMANUALVALVE300lbCS	1.2	3	144.03	1.38	-->	8
HANDLEBLINDSPACER150lbCS	3.5	3	120.8	0.58	-->	8
HANDLEPIPELESS20FTSTDCS	0.7	8	624.8	0.46	-->	4
HANDLEPIPELESS20FTSTDCS	11.2	3	228.59	0.18	-->	8
HANDLEMANUALVALVE150lbCS	2.3	12	540.1	5.18	-->	5

Table 4-11: *Clustering* results from implementation of a *DBScan algorithm* in aggregated fabrication data (Handling pipe area)

Cluster	Instances	Percentage
2	295	15%
3	220	11%
4	176	9%
6	174	9%
8	176	9%
10	172	9%
9	156	8%
12	162	8%
5	91	5%
11	79	4%
0	65	3%
7	51	3%
1	39	2%
13	32	2%
14	45	2%
15	48	2%
16	14	1%
17	8	0%

Table 4-12: Clusters created by the *DBScan algorithm* (Handling pipe area)

Fourthly, cluster data was transformed into graphs (Figure 4-9) to enhance data visualization. Table 4-13 presents a percentage distribution that became source data for development of these graphs.

PROJECT	DESC	Schedule	Material	QTY	%
PROJECTD	HANDLE PIPE	XS	CS	19056.42	52.06%
PROJECTD	HANDLE PIPE	XS	LOWTEMPCS	9359.28	25.57%
PROJECTC	HANDLE PIPE	Sch80	CS	1864.27	5.09%
PROJECTD	HANDLE PIPE	Sch40S	SS304304L	1556.07	4.25%
PROJECTC	HANDLE PIPE	Sch80	LOWTEMPCS	1217.62	3.33%
PROJECTB	HANDLE PIPE	STD	CS	1167.6	3.19%
PROJECTC	HANDLE PIPE	Sch160	CS	591.91	1.62%
PROJECTB	HANDLE PIPE	Sch40S	SS	358.34	0.98%
PROJECTC	HANDLE PIPE	Sch80S	SS	292.33	0.80%
PROJECTD	HANDLE PIPE	Sch80S	SS316316L	262.66	0.72%
PROJECTA	HANDLE PIPE	STD	LOWTEMPCS	199.07	0.54%
PROJECTC	HANDLE PIPE	Sch160	CHROMEMOLY	168.82	0.46%
PROJECTB	HANDLE PIPE	Sch160	LOWTEMPCS	131.45	0.36%
PROJECTD	HANDLE PIPE	Sch40S	SS316316L	120.98	0.33%
PROJECTB	HANDLE PIPE	XXS	CS	106.61	0.29%
PROJECTD	HANDLE PIPE	Sch80S	SS304304L	64.98	0.18%
PROJECTC	HANDLE PIPE	Sch160	CrGROUP3	31.74	0.09%
PROJECTD	HANDLE PIPE	Sch40S	SS	24.15	0.07%
PROJECTB	HANDLE PIPE	Sch80S	SS347347H	8.51	0.02%
PROJECTB	HANDLE PIPE	Sch40S	SS347347H	4.49	0.01%
PROJECTB	HANDLE PIPE	Sch160	SS347347H	4.03	0.01%
PROJECTD	HANDLE PIPE	Sch160	SS	2.07	0.01%

Table 4-13: Handling pipe percentage distribution results

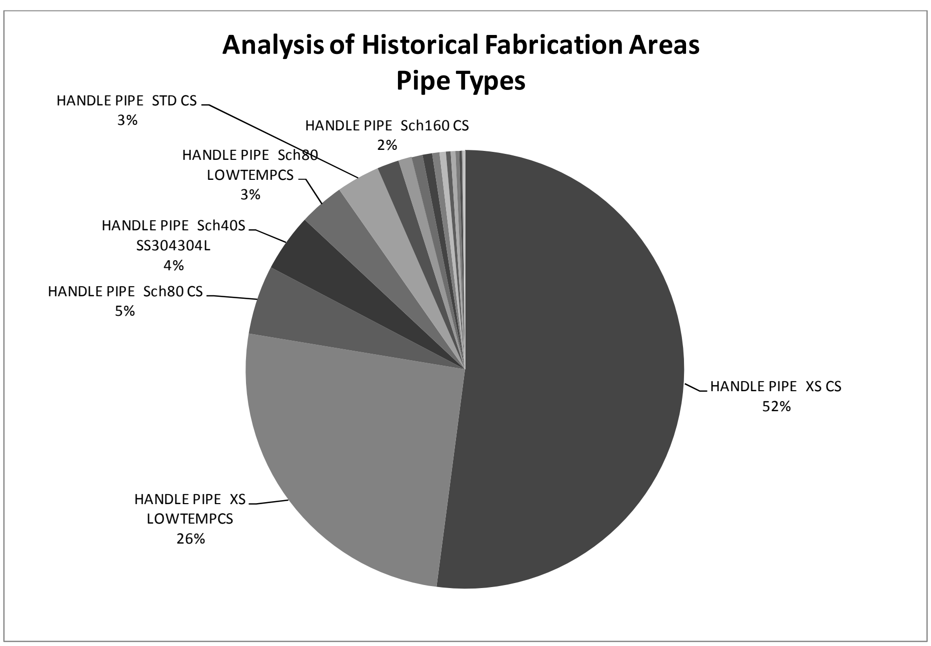

Figure 4-9: Pipe types Composition Chart

Finally, a *Predictive Apriori algorithm* was implemented in the largest cluster to detect association rules that could represent important knowledge from fabrication processes. Table 4-14 presents twenty rules created in Cluster No. 2, obtained from *Handling Pipe* activity analysis.

Rule No.	*Attribute No. 1*	*Attribute No. 2*	*Support*		*Attribute No. 3*	*Attribute No. 4*	*Support*	*Confidence*
1	Material=SS347347H	3		==>	Prescence=Small		3	89.07%
2	PipeSchedule=Sch40	2		==>	Prescence=Small		2	83.27%
3	Material=SINCONEL625	2		==>	Prescence=Small		2	83.27%
4	PipeSchedule=XS	Material=LOWTEMPCS	9	==>	Prescence=Large		8	77.97%
5	Material=CHROMEMOLY	7		==>	Prescence=Small		6	71.57%
6	PipeSchedule=XS	25		==>	Prescence=Large		18	68.53%
7	Material=LOWTEMPCS	17		==>	Prescence=Large		12	64.77%
8	PipeSchedule=XXS	5		==>	Prescence=Small		4	64.06%
9	PipeSchedule=Sch160	18		==>	Prescence=Small		12	61.35%
10	Material=LOWTEMPCS	Prescence=Large	12	==>	PipeSchedule=XS		8	59.86%
11	Material=SS316316L	4		==>	Prescence=Large		3	59.71%
12	Material=SS304304L	Prescence=Large	4	==>	PipeSchedule=Sch40S		3	59.71%
13	PipeSchedule=Sch160	Prescence=Large	6	==>	Material=CS		4	57.09%
14	Material=CS	33		==>	Prescence=Large		20	56.45%
15	Material=SS	8		==>	Prescence=Small		5	55.96%
16	Material=SS	Prescence=Large	3	==>	PipeSchedule=Sch80S		2	54.43%
17	PipeSchedule=XXS	5		==>	Material=CS		3	53.27%
18	Material=LOWTEMPCS	Prescence=Small	5	==>	PipeSchedule=STD		3	53.27%
19	PipeSchedule=XS	25		==>	Material=CS		14	53.07%
20	Material=SS304304L	7		==>	PipeSchedule=Sch40S		4	52.81%

Table 4-14: Rules extracted from Cluster No. 2 (Handling pipe area)

Figure 4-10 presents an overview of the analysis process for *Case Study 3*:

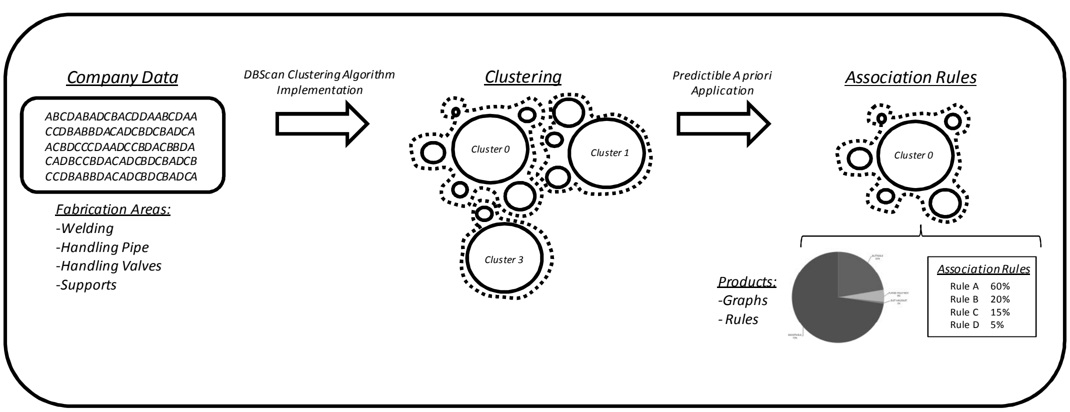

Figure 4-10: Case Study 3 Workflow

4.3.3.1. Results obtained from application of *Clustering* and *Association rules*

Clustering techniques applied in this experiment obtained some notable results in each fabrication area:

Welding area:

- 69 instances were classified in Cluster No. 1, out of 16 created clusters. This represented 15% of the total sample for this area. All welds grouped within this cluster presented a diameter of 2 inches (2”). 72.65% of the observed welds were from the socketweld type. Within this type, 56.34% of these welds were performed on carbon steel (CS) pipes. Another relevant percentage observed in data classified under Cluster No. 1 was buttweld weld type, representing 22.31% of the total data present in Cluster No. 1. 98.76% of welds used in carbon steel pipes presented large usages. Socketwelds had a 97.34 % probability of being applied in pipes presenting 3000 lb schedule (Sch 3000lb).

Handling pipe area:

- The cluster with greater size for this fabrication area was Cluster No. 2 (out of 18 created clusters). This cluster displayed 295 instances, representing 15% of the total clustered data for this activity. All items within this cluster had a diameter of two inches (2”). The material with most presence in pipe spool composition was carbon steel (CS), with 92%. 52.06% of produced pipes had carbon steel and extra-strong (XS) pipe schedule. According to rules obtained using *Predictive Apriori algorithm*, there is a 54.45% chance of observing large pipe quantities in those records that had carbon steel as prime material. In addition, in records with an extra-strong pipe schedule there was a 68.53% chance of presenting large quantities.

Handling valves area:

- Cluster No. 0 grouped the largest amount of records, having 77 instances representing 17% of the total data. 14 clusters were created for this fabrication area by the implementation of a *DBScan algorithm*. In addition, all items included in this cluster presented a valve diameter of 2 inches (2”). Furthermore, 84.15% of the total cluster data were manual valves made of carbon steel (CS). Only 7.62% of records present in this cluster were control valves fabricated with the same material. In terms of important rules observed within this cluster, 150lb manual valves made of carbon steel (low temp CS) had a 94.56% probability of being largely used in pipe spool fabrication processes.

Supports area:

- All supports classified in Cluster No. 8 presented a diameter of 2 inches (2”). The total amounts of records grouped in this cluster was 587, representing 20% of the total fabrication data. Furthermore, 18 clusters were obtained. 92% of the supports present in Cluster No. 8 presented carbon steel (CS) as material. Only 8% of the supports present in this cluster were made of stainless steel (SS). In addition, there is a 62.85% chance of using large quantities of supports made of carbon steel (low temp CS) during fabrication.

4.3.4. Case Study 4: Analysis of fabrication man-hours

A different sample of pipe spool data was obtained from the *Fabrication Time Studies* database (previously developed by PhD students of the Hole School of Construction of the University of Alberta). This data presented 295 different records with the following attributes:

- Material
- Size
- Pipe Schedule
- Weld Type
- Diameter Inches (DI)
- Manhours per Diameter Inch (MhrsperDI)

Material	Size	PipeSchedule	WeldType	DI	MhrperDI
Cr	2	XS	BW	2.00	0.36
Cr	2	XS	BW	2.00	0.39
Cr	2	XS	BW	2.00	0.42
Cr	2	XS	BW	2.00	0.48
Cr	3	XS	BW	3.00	0.24
Cr	4	XXS	BW	4.00	0.63
Cr	4	XXS	BW	4.00	0.65
Cr	6	1.25	BW	6.00	0.05
Cr	6	XXS	BW	6.00	0.65
Cr	6	Sch160	BW	6.00	1.14
Cr	11	0.25	FW	3.50	0.15
CS	1	3000	SW	1.00	0.08
CS	1	3000	SW	1.00	0.09
CS	1	3000	SW	1.00	0.09
CS	1	3000	SW	1.00	0.14
CS	1	3000	SW	1.00	0.15
CS	1	3000	SW	1.00	0.17
CS	1	3000	SW	1.00	0.18
CS	1	3000	SW	1.00	0.18
CS	1	3000	SW	1.00	0.20
CS	1	3000	SW	1.00	0.20
CS	1	3000	SW	1.00	0.21
CS	1	3000	SW	1.00	0.23
CS	1	3000	BW	1.00	0.24
CS	1	3000	SW	1.00	0.27
CS	1	3000	SW	1.00	0.29
CS	1	3000	SW	1.00	0.29
CS	1	3000	SW	1.00	0.29
CS	1	3000	SW	1.00	0.30
CS	1	3000	SW	1.00	0.30

Table 4-15: Data sample obtained from *Fabrication Time Studies*

To analyze its data, this case study is constituted by three unique stages.

In the first stage, data will be analyzed using a *DBScan algorithm*. Data clusters will be formed and subsequently studied using *Statistical Analysis*. In the second stage of the present case study, *Association Rule* known as *Apriori algorithm* is enforced to detect rules potentially characterizing clustered data. The third part of this study is to translate detected rules into graphical representations of data (pie charts). This is done to facilitate visualization of data distributions in pipe spool fabrication within a unique cluster. A brief description of this process is presented In Figure 4-11.

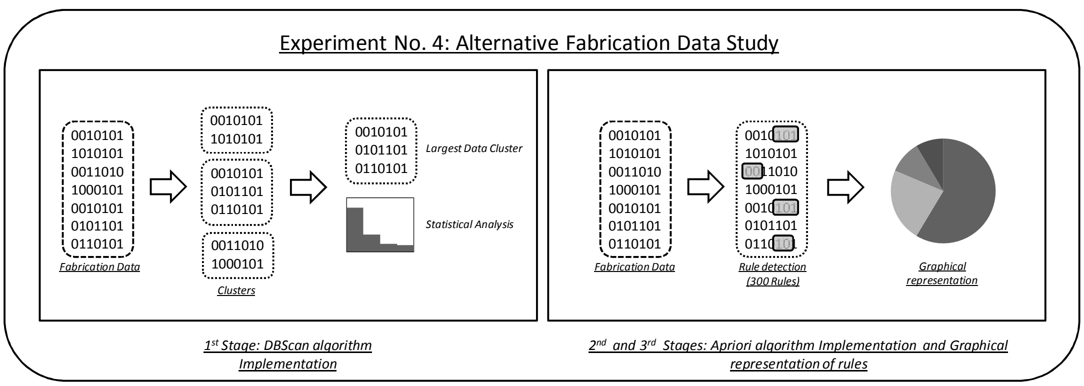

Figure 4-11: Case Study 4 Workflow

This data set was extracted from an alternative information source. Data present during development of this experiment does not belong to any historical project previously discussed. This data is part of a time studies project performed by other students to analyze pipe spool fabrication productivities.

4.3.4.1. Case Study 4: First Stage results. Statistical Analysis

For the first stage of this experiment *DBScan* was the algorithm chosen to create clusters in this data set. In addition, it established assignment distributions for each of those clusters conforming to 69% of total pipe spool fabrication data. Table 4-16 depicts these distributions in those clusters considered important, due to their size.

Cluster	Assignments	Percentage	
2	54	25%	69%
0	35	17%	
8	29	14%	
6	26	13%	

Table 4-16: Cluster assignments as result of implementation of a *DBScan algorithm*

Cluster No. 2 presents a greater number of records from all created clusters. In this cluster, a total assignment of 54 records was found representing 25% of the total Pipe Spool Fabrication data. *Statistical Analysis* was performed to further study data present in this particular cluster. Man-hours per diameter inch attribute (MhrsperDI) were selected to study its behavior within this cluster. By building a histogram, frequency observations related to man-hours per diameter inches (MhrsperDI) can be observed. Their results are portrayed in Table 4-17.

Ranges	No. Obs	Rel. Freq.	Cum. Rel. Freq.
0.05-0.09	20.5	37.96%	0.38
0.09-0.13	14	25.93%	0.64
0.13-0.17	5.5	10.19%	0.74
0.17-0.21	7	12.96%	0.87
0.21-0.25	4.5	8.33%	0.95
0.25-0.29	1.5	2.78%	0.98
0.29-0.33	0	0.00%	0.98
0.33-0.37	1	1.85%	1.00

Table 4-17: Cluster 2 frequency results

Using data extracted from the above table a histogram was build and is presented in Figure 4-12. In this graph, it can be noted that 37.96% of the total cases included in Cluster No. 2 belonged to those pipe fabrication records that were in the ranges of 0.05-0.09 man-hours per diameter inch.

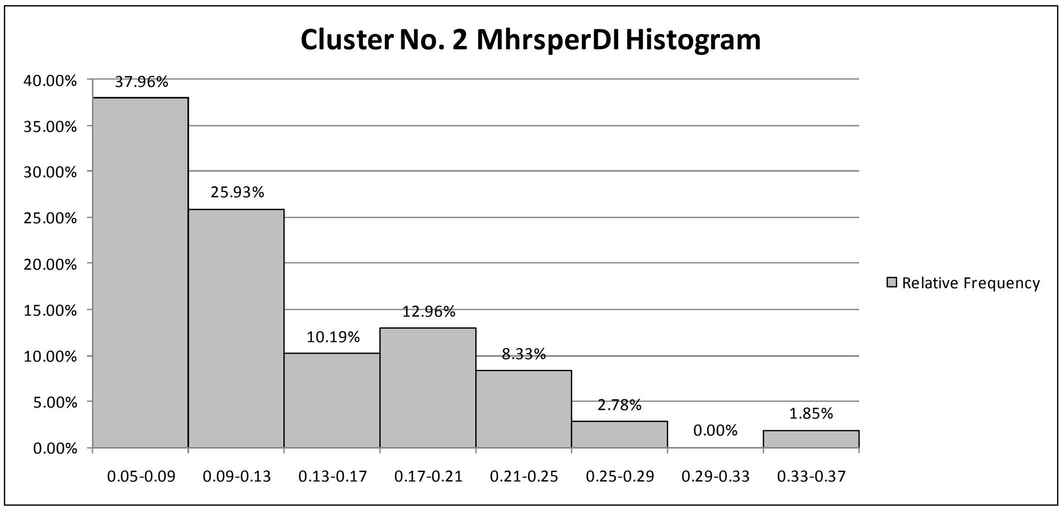

Figure 4-18: Cluster No. 2 Histogram

Furthermore, 25.93% of the cases observed in Cluster No. 2 expressed values within a range of 0.09-0.13 man-hours per diameter inch. These two categories represented 63.89% of the total cases included in Cluster No. 2. In addition, all records classified in Cluster No. 2 presented consistency in the values of three different attributes (material, pipe schedule and weld type). In these records, values such as carbon steel (CS), pipe schedule 3000 (Sch 3000) and saw weld type (SW) continuously appeared. Such attributes resemble unique characteristics present during pipe spool fabrication activities in this particular project.

4.3.4.2. Case Study 4: Second Stage results. *Apriori* algorithm

The second stage of this experiment starts with the application of an *Apriori algorithm* to the entire pipe spool database. In this case, 300 rules were set to be determined by the algorithm using Weka® software. A minimum confidence factor was defined at 60% in each discovered rule. In addition, a minimum support value was established at 0.03, generating rules involving at least 9 records. The attributes selected for rule generation were Material, Size, Pipe Schedule, Weld Type and MhrsperDI. These can be observed in Figure 4-13.

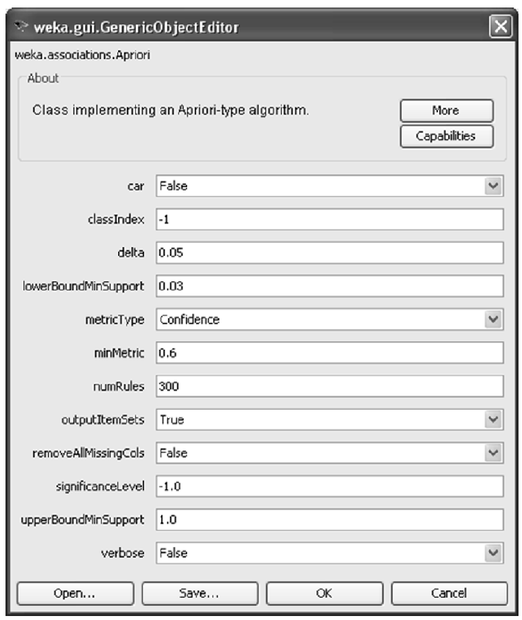

Figure 4-19: Parameters set in *Apriori algorithm* (Weka® software)

Furthermore, an extract of the 300 rules generated by *Apriori algorithm* is observed in Table 4-18.

Rule No.	Attribute No. 1	Attribute No. 2	Attribute No. 3	Attribute No. 4	Support		Attribute No. 5	Attribute No. 6	Attribute No. 7	Support	%
120	PipeSchedule=3000				103	==>	WeldType=SW			91	88%
96	WeldType=SW				99	==>	PipeSchedule=3000			91	92%
209	WeldType=SW				99	==>	Material=CS			62	63%
215	WeldType=SW				99	==>	Size=2			61	62%
188	Size=2				91	==>	WeldType=SW			61	67%
202	Size=2				91	==>	PipeSchedule=3000			58	64%
122	Material=CS	WeldType=SW			62	==>	PipeSchedule=3000			54	87%
216	Material=CS	WeldType=SW			62	==>	Size=2			38	61%
123	Size=2	WeldType=SW			61	==>	PipeSchedule=3000			53	87%
211	Size=2	WeldType=SW			61	==>	Material=CS			38	62%
115	Material=CS	PipeSchedule=3000			60	==>	WeldType=SW			54	90%
105	Size=2	PipeSchedule=3000			58	==>	WeldType=SW			53	91%

Table 4-20: Sample rules generated by *Apriori algorithm*

4.3.4.3. Case Study 4: Third Stage results. Graphical representations

In this final stage, information derived from detected rules was transformed into pie charts. These charts facilitate data review by providing an accessible and simple way of portraying data. Figures 4-12 and 4-13 present some of the graphs created during this experiment stage.

Material	Size	Observed	Schedule	Weld Type	Observed	%
CS	1"	31	Sch 3000	SW	24	77%
			Sch 3000	OL	4	13%
			Sch 3000	BW	1	3%
			Sch 160	BW	1	3%
			Sch 160	OL	1	3%

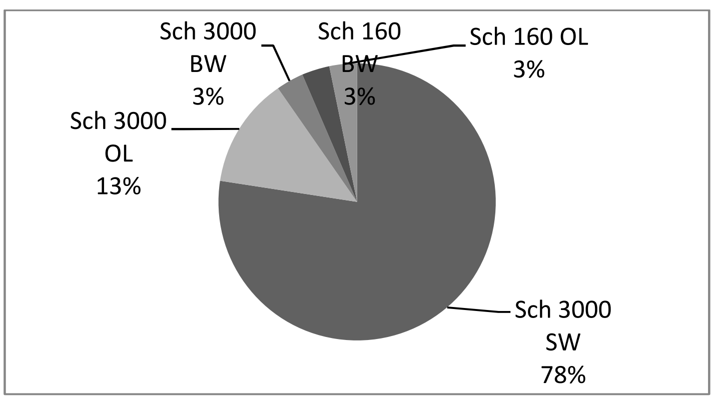

Figure 4-12: Schedule type distribution used in fabrication of a carbon steel pipe (1 inch)

Material	Size	Observed	Schedule	Weld Type	Observed	%
SS	2"	42	Sch 160	BW	10	23.81%
				OL	2	4.76%
			Sch 3000	SW	23	54.76%
				OL	3	7.14%
				BW	1	2.38%
			Sch XXS	FW	2	4.76%
				BW	1	2.38%

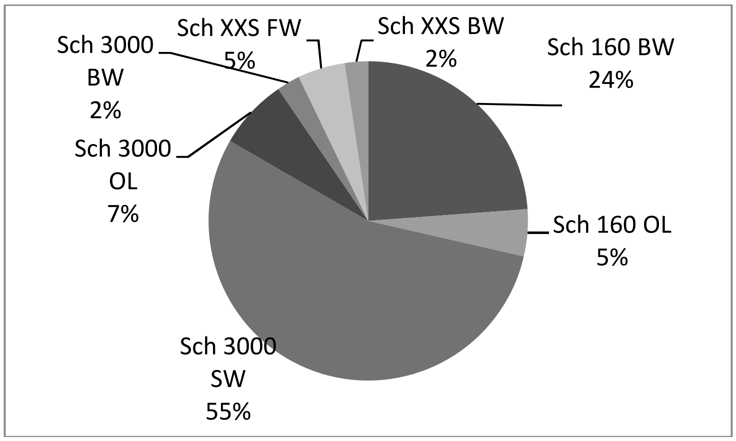

Figure 4-13: Schedule type distribution used in the fabrication of a stainless steel pipe (2 inches)

4.3.5. Case Study 5: Modeling data through probability distributions

This case study has the purpose of analyzing from a different perspective fabrication processes, specifically pipe spool fabrication. *Distribution Fitting* of data is performed using two sources:

1. Production reports obtained from Field Project Manager. These reports presented two sets of quantities, classified per each fabricated module: *number of spools* and *diameter inches* (DI) produced during module fabrication. Table 4-20 illustrates these measures using Project A as a sample. The same structure of data was adapted in all five historical projects.

2. Another source to be used for distribution fitting was derived from a *Fabrication Time Studies* database. Data capturing was executed by PhD Students of the Hole Construction School at University of Alberta. Multiple tables and queries are found in this database. Only two different outputs were taken into consideration for this section:

- A file containing 289 records with multiple attributes (material, pipe schedule, weld type, weld number, pipe diameter, welding position, diameter inches and welding time). This file was later named *small database.*
- A larger database with 19,960 records. This document presented six different attributes (shop order number, material, weld number, weld type, pipe size and pipe schedule). Because of the quantity of records embedded in this database, it was named *large database*.

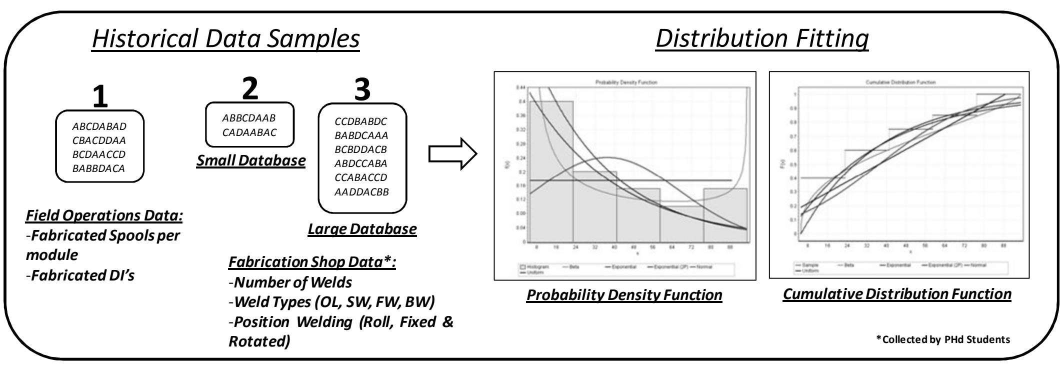

Figure 4-14: Case Study 5 Workflow

4.3.5.1. Production Reports analysis

Once historical project reports were obtained from the company's Field Project Manager, two measures were extracted: number of spools and diameter inches

(DI). Each of these quantities was classified by fabricated module. As example, Table 4-19 presents quantities performed for these two measures during fabrication of pipe spools present in Project A.

	Spools	
No.	Total	DI
1	40	876
2	40	1732
3	53	2,905
4	78	4,317
5	101	5634
6	78	3,879
7	36	1760
8	58	3335
9	87	2,964
10	43	1,871
11	20	1365
12	131	2,752
13	77	3499
14	48	3583
15	38	2,344
16	8	634

	Spools	
No.	Total	DI
17	22	1336
18	122	3,876
19	18	1,060
20	25	1,375
21	98	2,855
22	91	2,858
23	56	1,787
24	46	886
25	26	538
26	38	1,173
27	9	628
28	24	1097
29	12	263
30	13	336
31	53	2,800
32	2	329
33	49	1,657

Average Spools per Mod	49.70
Average DI's per Spool	41.65
Average DI's per Module	2,069.82
Average Welds per Spool	0.05
Average Welds per DI's	0.0012

Table 4-21: Project A fabricated spools and diameter inches per module

Using EasyFit 5.5 professional software (Mathwave Technologies, 2010) quantities derived from the number of spools and diameter inches performed were studied. Diverse distributions were fitted into these data sets in an attempt to find the best statistical option that could embody ad-hoc data from these attributes. Figures 4-17 and 4-18 present statistical distributions fitted to spool data for Project A.

- Normal Distribution
- Beta Distribution
- Exponential Distribution
- Triangular Distribution
- Uniform Distribution

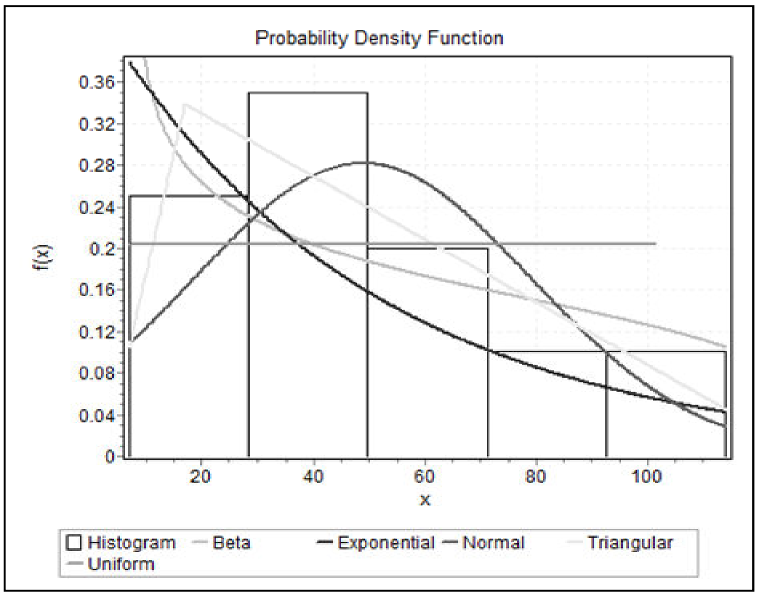

Figure 4-15: Project A spools Probability Density Function

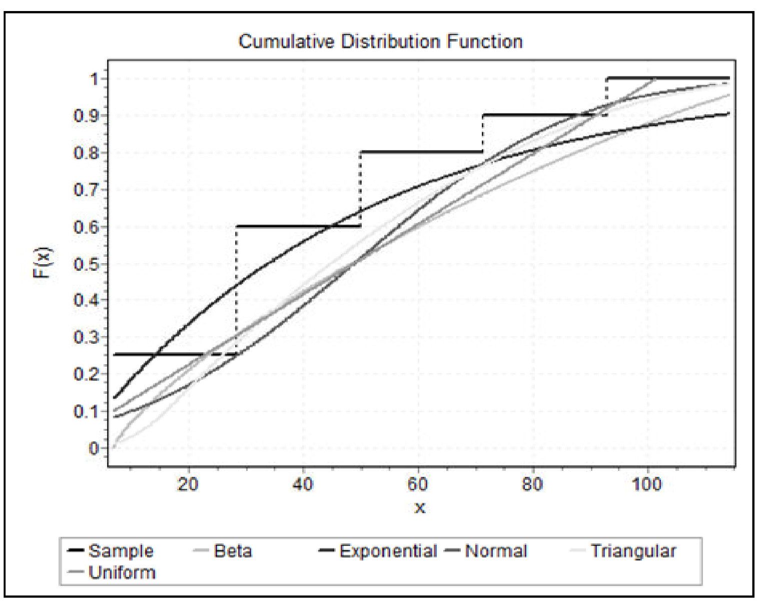

Figure 4-16: Project A spools Cumulative Distribution Function

4.3.5.2. Results obtained from distribution fitting

For each historical project a statistical distribution fit was obtained. For number of spools fabricated per module attribute, best fits are highlighted and presented in Tables 4-20 to 4-22:

Project A Spools		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.13835	2	2.1596	4	0.39541	2
2	Exponential	0.2293	6	1.3741	3	4.5333	5
3	Exponential (2P)	0.18644	5	2.2209	5	1.3606	3
4	Normal	0.15211	4	0.52327	2	2.0154	4
5	Triangular	0.09731	1	0.23146	1	0.34669	1
6	Uniform	0.13905	3	7.5322	6	N/A	

Project B Spools		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.11701	2	1.495	4	0.20837	2
2	Exponential	0.25569	6	1.9449	5	7.07	6
3	Exponential (2P)	0.1874	5	2.3319	6	0.85564	4
4	Normal	0.13066	3	0.40077	2	1.293	5
5	Triangular	0.14129	4	1.4108	3	0.62469	3
6	Uniform	0.11284	1	0.35649	1	0.19886	1

Table 4-22: Best fit results (spools fabricated per module) for Projects A & B

Project C Spools		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.15641	1	2.4146	4	0.375	2
2	Exponential	0.19069	3	0.73456	1	0.17335	1
3	Exponential (2P)	0.16448	2	2.1174	3	3.7596	4
4	Normal	0.19935	4	1.0332	2	1.1495	3
5	Triangular	0.30047	6	5.6957	6	5.25	5
6	Uniform	0.23085	5	4.7483	5	N/A	

Project D Spools		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.16074	3	1.3105	3	0.37501	1
2	Exponential	0.12667	2	0.4798	1	1.4374	4
3	Exponential (2P)	0.10092	1	1.8879	4	0.50559	2
4	Normal	0.21922	5	0.98896	2	0.81215	3
5	Uniform	0.18981	4	10.893	5	N/A	
6	Triangular			No Fit			

Table 4-23: Best fit results (spools fabricated per module) for Projects C & D

Project E Spools		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.5776	6	39.832	4	N/A	
2	Exponential	0.3	5	3.0378	2	2.1816	4
3	Exponential (2P)	0.3	3	60.403	5	2.1816	3
4	Normal	0.17281	1	0.86663	1	1.4027	1
5	Triangular	0.3	4	61.463	6	2.1154	2
6	Uniform	0.21435	2	4.434	3	N/A	

Table 4-24: Best fit results (spools fabricated per module) for Project E

In the scenario in which diameter inches attribute was analyzed, best fit results are depicted in Tables 4-23 to 4-25:

Project A DI's		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.09927	2	1.8567	2	0.375	2
2	Exponential	0.23254	6	2.037	3	2.6834	5
3	Exponential (2P)	0.19816	5	2.6771	4	0.32921	1
4	Normal	0.10983	3	0.26559	1	1.1406	3
5	Triangular	0.19158	4	3.4533	5	2.25	4
6	Uniform	0.09912	1	3.9135	6	N/A	

Project B DI's		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.12592	2	1.9061	2	0.375	1
2	Exponential	0.32749	6	3.0969	4	2.1824	4
3	Exponential (2P)	0.20421	5	2.6863	3	1.0932	3
4	Normal	0.10694	1	0.28155	1	0.47007	2
5	Triangular	0.19492	4	3.3154	5	3.8333	5
6	Uniform	0.14191	3	7.3719	6	N/A	

Table 4-25: Best fit results (diameter inches per module) for Projects A & B

Project C DI's		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.10959	1	1.2192	3	8.523E-06	1
2	Exponential	0.23592	5	0.90708	1	0.07097	2
3	Exponential (2P)	0.14762	2	2.4201	4	1.5525	3
4	Normal	0.19272	3	1.0271	2	2.3511	4
5	Triangular	0.25241	6	4.5671	5	5	5
6	Uniform	0.22853	4	7.8168	6	N/A	

Project D DI's		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.19982	4	3.0703	3	1.5	3
2	Exponential	0.27417	6	1.3445	2	1.7616	4
3	Exponential (2P)	0.2653	5	3.1453	4	1.4846	2
4	Normal	0.13904	1	0.3557	1	0.49692	1
5	Triangular	0.17562	3	3.9293	5	2.2499	5
6	Uniform	0.14198	2	4.0799	6	N/A	

Table 4-26: Best fit results (diameter inches per module) for Projects C & D

Project E DI's		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.55706	6	36.202	5	N/A	
2	Exponential	0.3	3	3.3602	2	3.9271	3
3	Exponential (2P)	0.3	4	3.3602	3	3.9271	4
4	Normal	0.152	1	0.68396	1	0.87756	1
5	Triangular	0.3	5	138.47	6	1.5	2
6	Uniform	0.19832	2	7.611	4	N/A	

Table 4-27: Best fit results (diameter inches per module) for Project E

Multiple statistical distributions represented the best fit for each attribute. Consistency of distributions in historical projects was not present, obtaining different results when fitting.

However, according to each project profile defined by the number of fabricated spools and diameter inches performed, input models for a simulation project can be established using these results. Five different models can potentially be used to simulate different alternatives derived from project profiles.

4.3.5.3. Time Studies: small database

289 records were obtained introducing nine attributes. Data depicting position weld types (rotated, fixed and rolled) and man-hours performed for each weld were present in this database.

Control Number	Material	PipeSchedule	WeldType	WeldNumber	PipeDiameter	WeldingPosition	TotalUnits(DI)	Welding Time
A	CS	STD	BW	1	18	Rotated	50	0.95
A	CS	STD	BW	2	18	Rotated	50	0.94
A	CS	0.25	BW	3	14	Fixd	50	0.59
B	CS	STD	BW	1	3	Roll	23	0.25
B	CS	STD	BW	2	3	Roll	23	0.45
B	CS	0.25	FW	3	24	Fixd	23	0.24
B	CS	STD	BW	4	3	Roll	23	0.23
B	CS	STD	BW	5	3	Roll	23	0.32
B	CS	STD	BW	6	3	Roll	23	0.28
C	CS	Sch120	BW	1	10	Rotated	42	1.85
C	CS	3000	OL	2	2	Fixd	42	0.78
C	CS	3000	SW	3	2	Fixd	42	0.45
C	CS	3000	SW	4	2	Rotated	42	0.18
C	CS	0.25	FW	5	27	Fixd	42	0.61
C	CS	0.25	FW	6	27	Fixd	42	0.53
C	CS	Sch120	BW	7	10	Rotated	42	1.62
D	CS	0.25	FW	1	24	Fixd	22	0.28

Table 4-28: Small database sample

As observed in the data sample shown in Table 4-26, for each produced spool there is a combination of attributes defining unique records. Each spool possesses a control number for identification, pipe material, pipe schedule, weld type, weld number, pipe diameter, weld position, total units and man-hours performed during welding.

4.3.5.4. Preparing data for distribution fitting

To study the distribution of welds per fabricated spool, this database was modified considerably. Only two attributes are used to reflect data and perform distribution fitting. These attributes are control number and number of welds. To determine quantities for the last mentioned attribute, this database was consolidated using an MS Excel® spreadsheet in which the number of welds per spool is counted. The final product of this process is a table outlining total weld numbers executed in fabricated pipe spools.

Table 4-27 presents a sample of the mentioned table produced by this data cleaning and ordering process.

Control Number	Welds
A	3
B	6
C	7
D	11
E	9
F	10
G	1
H	8
I	14
J	7
K	3
L	1
M	16
N	11
O	2
P	2
Q	1
R	2
S	4

Table 4-29: Number of welds per fabricated spool

Similar means were performed while preparing data related to position welds and weld types. Alternative spreadsheets were created in MS Excel® for each case. One table detects how many position welds were used in the fabrication of a particular spool. The other accumulates quantities of weld types per pipe spool.

In the case of position welds, some of the obtained results are presented in Table 4-28.

Control Number	Roll?	Fix?	Rotated?
A	3	0	0
B	3	3	0
C	6	1	0
D	9	2	0
E	0	5	4
F	0	1	9
G	0	0	1
H	0	0	8
I	0	1	13
J	0	2	5
K	0	1	2
L	0	0	1
M	0	2	14
N	0	3	8
O	0	2	0
P	0	0	2
Q	0	0	1

Table 4-30: Weld types per fabricated spool sample (rolled, fixed and rotated position welds)

All tables and their contents were analyzed using EasyFit 5.5 professional. Different statistical distributions were tested against raw data to determine a best fit. Figures 4-19 and 4-20 display results obtained in number of welds per fabricated pipe spool attribute. The rest of the graphs produced in this case study can be observed in the Appendices chapter of this thesis.

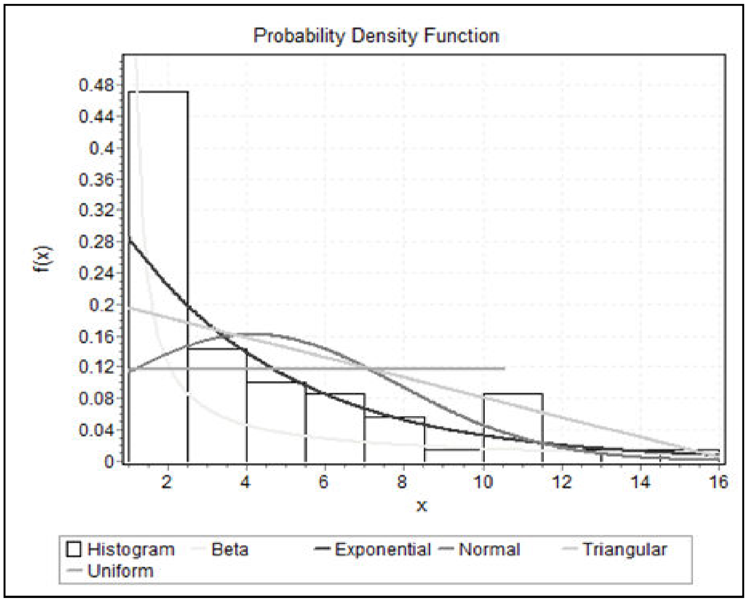

Figure 4-17: Number of welds per fabricated pipe spool Probability Density Function

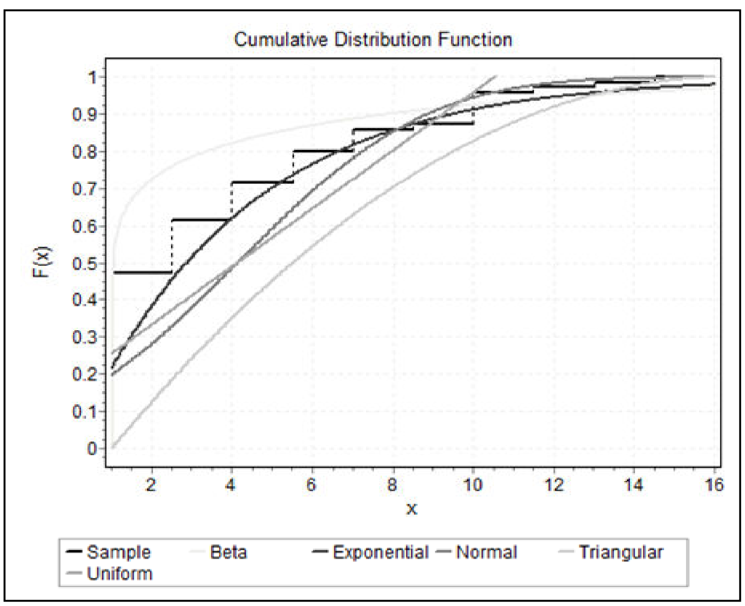

Figure 4-18: Number of welds per fabricated pipe spool Cumulative Distribution Function

4.3.5.5. Results obtained from distribution fitting of a small database of records

Results for each of the three categories mentioned during the development section of *Case Study No. 5* were obtained using *Distribution Fitting*. In each case a best fit between acquired data and a statistical distribution was found.

Welds per Spool		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.42561	5	22.678	3	31.356	3
2	Exponential	0.21511	1	2.3333	1	8.6487	1
3	Normal	0.23416	2	4.7444	2	11.933	2
4	Triangular	0.37107	4	83.508	5	65.604	4
5	Uniform	0.25581	3	31.441	4	N/A	

Table 4-31: Distribution results for number of welds used in pipe spool fabrication (small database)

Fixed Welds		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.54662	5	194.96	4	N/A	
2	Exponential	0.4	4	11.289	2	38.423	3
3	Normal	0.29118	1	6.0315	1	26.411	1
4	Uniform	0.30147	2	20.368	3	N/A	
5	Triangular			NO FIT			

Table 4-32: Distribution results for number of fixed welds used in pipe spool fabrication (small database)

Roll Welds		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.67702	3	35.261	4	100.39	2
2	Exponential	0.91429	6	-5.9156	1	340.2	4
3	Normal	0.51863	2	21.796	2	52.393	1
4	Triangular	0.91429	5	1884.1	6	340.29	5
5	Uniform	0.49067	1	30.753	3	N/A	

Rotated Welds		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.47116	6	84.302	4	N/A	
2	Exponential	0.25714	3	13.941	2	10.952	2
3	Normal	0.24539	1	5.4885	1	16.117	3
4	Triangular	0.39942	5	162.69	6	75.671	4
5	Uniform	0.26336	4	23.646	3	N/A	

Table 4-33: Distribution results for number of roll and rotated welds used in pipe spool fabrication (small database)

BW per Spool		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.50876	5	78.929	4	N/A	
2	Exponential	0.22857	2	15.246	2	6.7353	2
3	Normal	0.25953	4	5.2262	1	45.661	3
4	Uniform	0.25604	3	20.032	3	N/A	
5	Triangular			NO FIT			

FW per Spool		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	N/A		N/A		N/A	
2	Exponential	0.5143	5	-14.135	1	119.57	4
3	Normal	0.34178	2	7.7331	2	16.852	1
4	Triangular	0.57143	4	712.13	5	105.13	2
5	Uniform	0.28638	1	17.222	3	N/A	

Table 4-34: Distribution results for number of BW and FW weld types used in pipe spool fabrication (small database)

OL per Spool		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	N/A		N/A		N/A	
2	Exponential	0.82857	4	-11.289	1	268.82	3
3	Normal	0.49356	2	17.881	2	45.37	1
4	Uniform	0.45158	1	26.189	3	N/A	
5	Triangular			NO FIT			

SW per Spool		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.63876	3	47.561	4	N/A	
2	Exponential	0.71429	6	8.9669	1	187.75	3
3	Normal	0.38408	2	15.27	2	25.107	1
4	Triangular	0.71429	5	1179.4	6	247.11	4
5	Uniform	0.37317	1	37.473	3	N/A	

Table 4-35: Distribution results for number of OL and SW weld types used in pipe spool fabrication (small database)

The *Normal* distribution was predominantly observed during *Distribution Fitting* of fabrication attributes. In only a few occasions did *Exponential* distribution prove to obtain better results (in attributes such as number of welds and buttwelds executed per fabricated pipe spool).

However, when reviewing Cumulative Distribution Function graphs, the Normal distribution can be used as the best fit. Because of its shape, it can be associated to project data. It has an approximated pattern closer to attribute distributions.

4.3.5.6. Time Studies: large database

A larger quantity of records containing fabrication data was obtained from the time studies database. 19,960 records were present in this document detailing five attributes: shop order number, weld number, weld type, pipe size and pipe schedule.

The goal of this particular section of *Case Study 5* is to fit statistical distributions for *number of welds* and *weld types* used in the fabrication of a pipe spool. Because of the larger quantity of records present, a more representative fit of a statistical distribution is expected.

Shop Order No	Weld	Weld Type	Size (Pipe)	Schedule (Pipe)
A	1	SW	2	3000
A	2	SW	2	3000
A	3	SW	2	3000
B	1	FW	24	0.25
B	2	SW	2	3000
B	3	SW	2	3000
B	4	SW	2	3000
B	5	SW	2	3000
B	6	SW	2	3000
B	7	SW	2	3000
B	8	SW	0.75	3000
B	9	SW	0.75	3000
B	10	SW	0.75	3000
B	11	SW	0.75	3000
B	12	SW	0.75	3000
B	13	SW	0.75	3000
B	14	SW	0.75	3000
C	1	SW	2	3000
C	2	SW	2	3000
C	3	SW	2	3000
C	4	SW	2	3000
C	5	SW	2	3000
C	6	SW	2	3000
C	7	SW	2	3000
C	8	SW	2	3000
C	9	SW	0.75	3000
C	10	SW	0.75	3000
C	11	SW	0.75	3000
C	12	SW	0.75	3000

Table 4-36: Large database sample

The procedure used to extract data from both number of welds and weld types executed per fabricated pipe spool attributes is similar to the one described in the previous section of this case study.

Using a MS Excel® spreadsheet, information pertinent to these attributes was extracted and ordered as samples shown in Tables 4-35 and 4-36.

Shop Order No	Welds
A	3
B	14
C	15
D	21
E	4
F	11
G	4
H	21
I	0

Table 4-37: Number of welds per fabricated pipe spool (large database)

Shop Order No	BW	SW	OL	FW
A	0	3	0	0
B	0	13	0	1
C	0	15	0	0
D	0	20	0	1
E	0	4	0	0
F	0	11	0	0
G	0	4	0	0
H	0	6	0	0
I	0	0	0	0

Table 4-38: Weld types used per fabricated pipe spool (large database)

Furthermore, another similarity from previous section was found in its analysis tool. For all cases, distributions were fitted using EasyFit 5.5 Professional.

This database contained the most detailed sample of data in *Time Studies* section. Once all records were ordered, *Distribution Fitting* was performed. The results are arranged in Figures 4-21 and 4-22.

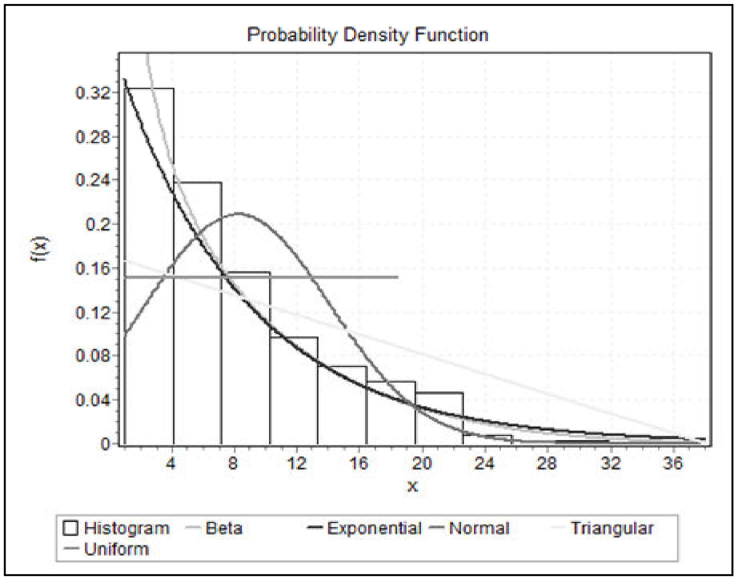

Figure 4-19: Number of welds per spool Probability Density Function

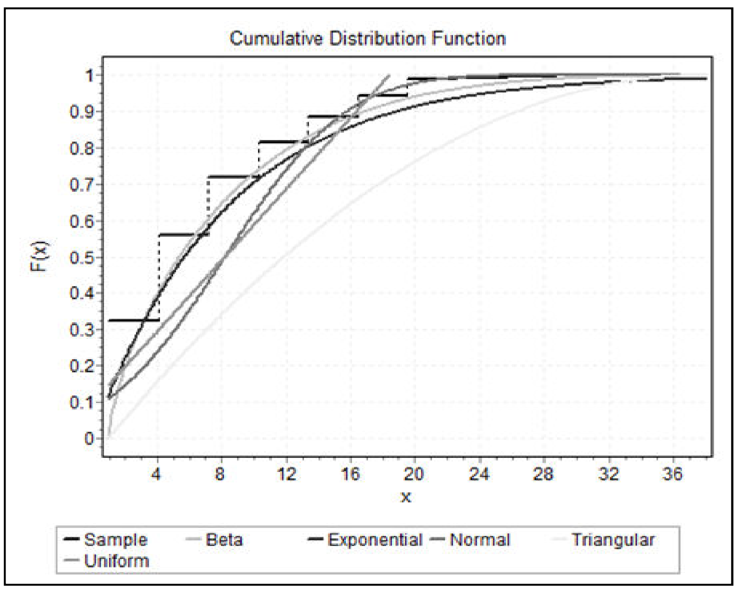

Figure 4-20: Number of welds per spool Cumulative Distribution Function

4.3.5.7. Results obtained from distribution fitting of a large database of records

As result of data analysis using EasyFit 5.5, new distributions were found resembling the characteristics of raw data. In almost all cases, Normal distribution presented best results in each attribute. These are displayed in Tables 4-37 to 4-39.

Welds per Order		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.1661	4	789.71	5	N/A	
2	Exponential	0.15562	3	78.625	2	429.02	1
3	Normal	0.14352	1	62.919	1	584.41	2
4	Triangular	0.29064	5	592.74	3	1043.7	3
5	Uniform	0.14612	2	767.33	4	N/A	

Table 4-39: Distribution results for number of welds used in pipe spool fabrication (large database)

BW per Order		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	N/A		N/A		N/A	
2	Exponential	0.59638	4	-144.56	1	8857.8	3
3	Normal	0.31941	1	304.23	2	793.95	1
4	Uniform	0.32914	2	894.9	3	N/A	
5	Triangular			NO FIT			

FW per Order		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	N/A		N/A		N/A	
2	Exponential	0.61985	5	-440.69	1	9464.1	4
3	Normal	0.3679	2	310.19	2	1090.7	1
4	Triangular	0.61985	4	30097	5	9436.2	2
5	Uniform	0.31279	1	883.59	3	N/A	

Table 4-40: Distribution results for number of BW and FW weld types used in pipe spool fabrication (large database)

OL per Order		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.43564	1	615.12	2	1167.1	1
2	Exponential	0.86656	6	-307	1	19498	4
3	Normal	0.50211	3	662.91	3	2065.6	2
4	Triangular	0.86656	5	57931	6	19521	5
5	Uniform	0.46661	2	974.83	4	N/A	

SW per Order		Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
#	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Beta	0.60344	6	1765.6	4	6832.4	5
2	Exponential	0.26936	4	454.76	2	1444.8	3
3	Normal	0.1782	1	114.59	1	289.5	1
4	Triangular	0.36716	5	6554.6	6	2860.2	4
5	Uniform	0.23377	2	853.26	3	N/A	

Table 4-41: Distribution results for number of OL and SW weld types used in pipe spool fabrication (large database)

4.4. Conclusions

Five unique case studies were reviewed during development of this chapter. Information collected in all experiments was acquired from different company reports. *Data Ordering and Cleansing* techniques were applied, forming multiple groups of records. *Clustering* algorithms were used to analyze project data, detecting key characteristics defining fabrication areas.

Case Study 1 created historical project profiles through *Clustering* using *Simple K Means algorithm*. In addition, it attempts to make comparisons between historical and potential project profiles to determine correspondent matches. Once similarities have been defined, a decision maker can potentially assign base prices in future tendering processes using historical data. However, establishing fixed structures based on historical data resembling project profiles increases supervision during clustering. This can generate misleading results while attempting to produce matches between historical and potential projects. Even though this approach is more practical than using a uni-dimensional analysis currently performed by the company, it represents a deterministic path not recommended.

Case Study 2 involved an application of two different clustering methods to define characteristics present in pipe module fabrication activities of historical projects. The analysis methods chosen were *Simple K Means* and *DBScan* algorithms, depending on the amount and quality of collected data. Characterizations were found in each of the four fabrication areas defined within this case study. However, for this *Data Mining* analysis records proved to be sparse. Most of the classification results obtained during characteristic determination presented a high average number of incorrectly classified instances (66.55% during clustering with *DBScan algorithm*). This proved that large amounts of very diverse records were present within data samples. Furthermore, insufficient data for testing algorithms was one limitation encountered during this case study progress.

Case Study 3 demonstrated additions representing alternative analysis approaches. Initial data observed in *Case Study No. 2* was merged into a single source condensing information in each fabrication area. Besides applying a *DBScan algorithm* to identify main characteristics of each fabrication area, other data analysis techniques such as *Association Rules* were applied to discover knowledge hidden in company records. Furthermore, graphs reinforced present analysis providing information displays facilitating comprehension of company data.

Case Study 4 studied man-hours involved in welding activities of fabricated pipe spools from a detailed insight. Frequencies between records were detected using simple *Statistical Analysis* means such as histograms. In addition, *Apriori algorithm* was applied detecting relationships between attributes present in fabrication. It is important to mention that initial parameters of *Association Rules* must be aligned to company requirements. Depending on the conditions specified by a user, results derived from *Association Rules* can vary.

Case Study 5 presents the largest collection of records acquired during this research. In addition, historical data was examined using *Distribution Fitting*. *Fabrication Time Studies* represented an alternative data source. Two results are derived from this case study. First, it is possible to obtain statistical distributions representing historical project characteristics. Second, when analyzing pipe spool fabrication databases, Normal distribution seems to be the best choice representing data for each observed attribute.

To conclude, the result of applying *Clustering, Association Rules* and *Distribution Fitting* techniques depends directly on data quality. Beneficial results can be obtained from those records providing truthful and reliable insights about fabrication operations. Additional data is needed to efficiently define project profiles and main characteristics for case studies handling historical knowledge. This was the case of *Case Studies 1 to 3*. Supplementary data is required to increase robustness of these experiments. A different situation is observed in *Case Studies 4 and 5*, in which data forming statistical distributions provided an important contribution to build input models for computer simulation. These present real fabrication data from field operations.

Chapter 5: Conclusions and Recommendations

5.1. Research Summary

Over the course of months, information from multiple data sources and departments of an industrial construction Company was analyzed to study at a more detailed level pipe module fabrication processes.

Research was performed in two different stages. The first stage was directed to enhance a company's perception about its previous performances. During this phase, a great degree of difficulty was found when comparing data obtained from multiple sources. Each document produced by operational areas presented its own impression about fabrication operations. One obstacle encountered during research was consolidating knowledge produced by all departments into a single data source, to simplify its review by decision makers. Alternative Industrial Indicators were designed to comply with this requirement. Their development was comprised of previous historical performances using a combination display of graphs and figures, accumulating past experiences.

The second stage of this research presented numerical analysis of data obtained from company records constituting Industrial Indicators. Three alternatives were explored to detect special relationships between numerous data quantities constituting previous performances. These were the following:

1. Clustering (using *Simple K Means* and *DBScan* algorithms)
2. Association Rules (implementing *Apriori* and *Predictive Apriori* algorithms)
3. Distribution Fitting

The application of these techniques generated various results:

- A first attempt to define project profiles using *Simple K-Means Clustering*, using a structure based on a combination of multiple attributes describing certain characteristics of pipe module fabrication projects. In this study further analysis is needed to determine how attribute weights are assigned, defining their importance. Alternative approaches exploring project profile determination are recommended due to excessive clustering supervision during implementation of this case study.
- Data from fabrication activities was highly dispersed. This was experienced during analysis of fabrication activities through application of *Clustering* techniques. Because of this, two different algorithms were used in fabrication data: *Simple K Means* and *DBScan*. During *Simple K Means* experimentation, high amounts of clusters and *R squared error* percentages were obtained. In addition, an average of 66.55% of incorrectly classified instances during implementation of a *DBScan algorithm* was acquired, proving existence of deficient and scattered amounts of records.
- The diameter observed in the entirety of clusters representing majorities within all fabrication areas (with average presence of 16.75%) was two inches

(2”). This indicated the most prevalent size used in fabrication of pipe spool modules.

- The material with the greatest occurrence during pipe module fabrication was carbon steel (CS). This item displayed presence values above 50% in all fabrication areas. *Association Rules* such as *Predictive Apriori* algorithm estimated a 56.45% probability of using large quantities of carbon steel pipes during module fabrication activities.

- When executing *Distribution Fitting* in fabrication data, the number of spools per fabricated module between historical projects does not pursue steady results according to a particular distribution. No consistency between project statistics was found.

- On the other hand, the number of diameter inches per fabricated module presented a consistent trend. Its values can be represented with a Normal distribution. In those few cases in which Beta distribution represented the first choice (Project A and C), Normal distribution can be used as secondary alternative.

- When analyzing an industrial construction small database containing 289 records, the number of welds per spool observed is fitted by an Exponential distribution. An alternative for this case exists using a Normal distribution because it is considered the second best option for fitting. This was experienced when expanding obtained knowledge by analyzing a large database of records.

5.2. Research Contributions

Analysis of pipe module fabrication activities from an alternative point of view has been one of the contributions of the present research. Operations were broken down by areas, achieving great level of details. Unique project characteristics were determined from multiple sources with variant structures, contents and numbers of records.

The analysis performed by this research generated alternative Industrial Indicators which are currently used by the company as a decision support tool to its managerial team in charge of developing tenders during bidding processes. From multiple factors developed, historical distributions of pipe diameters used in fabrication and steel mix percentages present in module steel structures represented items of relevant interest to the company.

Another contribution of this research is detection of statistical distributions present in fabrication activities of historical projects. Analysis of pipe spools, diameter inches and weld types in historical data represents a significant addition to spool fabrication research by supporting its study with a primary source of obtained knowledge that resembles real life conditions.

5.3. Future Research

Pertinent considerations are proposed as a continuation of this research:

- Include additional projects in its scope. This can improve results obtained during application of *Clustering* and *Association Rules* techniques.

Distribution Fitting using supplementary data can reflect more robust results related to fabrication operations, by using an amplified knowledge base. More data can support validation of *Data Mining* techniques used in this research.

- Develop input models using statistical distributions results derived from this study. This can reflect additional aspects of fabrication operations. Each historical project and their distinct distributions can emulate unique behaviors in pipe spool fabrication activities, representing versatile scenarios benefitting the study of this industrial discipline through application of computer simulation.
- Explore other *Data Mining* techniques to discover additional insights of pipe fabrication activities. Decision Trees, Linear Regression and Neural Networks are some of the algorithms representing alternatives to study fabrication data.

6. References

AbouRizk, S., Knowles, P., & Hermann, U. (2001). Estimating labor production rates for industrial construction activities. *Journal of Construction Engineering and Management* , 127(6),502-511. DOI:10.1061/(ASCE)0733-9364(2001)127:6(502).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB.* , 487-489. Retrieved from http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf.

Ahmad, I. (2000). Data warehousing in construction organizations. *Proceedings of Construction Congress VI: Building Together for a Better Tomorrow in an Increasingly Complex World*, 194-203. DOI:10.1061/40475(278)22.

Ahmad, I., Azhar, S., & Lukauskis, P. (2004). Development of a decision support system using data warehousing to assist builders/developers in site selection. *Automation in Construction* , 13(4). 525-542. DOI:10.1016/j.autcon.2004.03.001.

Ankerst, M., Breunig, M., Kriegel, H., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, 49-60. Retrieved from http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/OPTICS.pdf.

Azhar, S., Ahmad, I., & Sein, M. K. (2010). Action research as a proactive research method for construction engineering and management. *Journal of Construction Engineering and Management*, 87-98. DOI:10.1061/(ASCE)CO.1943-7862.0000081.

Bouchlaghem, D., Kimmance, A., & Anumba, C. (2004). Integrating product and process information in the construction sector. *Industrial Management and Data Systems*, 104(3), 218-233. DOI:10.1108/02635570410525771.

Buchheit, R., Garret, J., Lee, S., & Brahme, R. (2000). A knowledge discovery case study for the intelligent workplace. *Proceedings of the Eight International Conference on Computing in Civil and Building Engineering*, 279(40513),119. DOI:10.1061/40513(279)119 .

Chirstodoulou, S. (2000). The science in human intuition: optimum bid markup calculation in competitive bidding environments using probabilistic neural networks. *Computing in Civil and Building Engineering*, 574-579. DOI:10.1061/40513(279)75.

Chua, D. K., & Li, D. (2000). Key factors in bid reasoning model. *Journal of Construction Engineering and Management*, 126(5), 349-357. DOI:10.1061/(ASCE)0733-9364(2000)126:5(349).

Dozzi, S. P., AbouRizk, S. M., & Schroeder, S. L. (1996). Utility-theory model for bid markup decisions. *Journal of Construction Engineering and Management*, 119-124. DOI:10.1061/(ASCE)0733-9364(1996)122:2(119).

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. *Proceedings of Second SIAM International Conference on Data Mining*, 1-12. Retrieved from http://www-users.cs.umn.edu/~kumar/papers/kdd02_snn_28.pdf.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226-331. Retrieved from http://www2.cs.uh.edu/~ceick/7363/Papers/dbscan.pdf.

Fan, H. (2007). Leveraging operational data for intelligent decision support in construction equipment management. *PhD Thesis*. University of Alberta. Edmonton, Alberta. Canada.

Fan, H., AbouRizk, S., & Kim, H. (2007). Building Intelligent Applications for Construction Equipment Management. *Journal of Computing in Civil Engineering* , 192-199. DOI:10.1061/40937(261)24.

Fan, H., AbouRizk, S., Kim, H., & Zaiane, O. (2008). Assessing residual value of heavy construction equipment using predictive data mining model. *Journal of Computing in Civil Engineering*, 22(3), 181-191. DOI:10.1061/(ASCE)0887-3801(2008)22:3(181).

Fayek, A. (1998). Competitive bidding strategy model and software system for bid preparation. *Journal of Construction Engineering and Management*, 124, 1-10. DOI:10.1061/(ASCE)0733-9364(1998)124:1(1).

Flanagan, R., & Norman, G. (1985). Sealed bid auctions: an application to the building industry. *Construction Management and Economics*, 3(2), 145-161. DOI:10.1080/01446198500000011.

Foss, A., & Zaïane, O. (2002). A parameterless method for efficiently discovering clusters of arbitrary shape in large datasets. *In IEEE Intl. Conf. on Data Mining*, 179-186. Retrieved from http://webdocs.cs.ualberta.ca/~zaiane/postscript/icdm02-2.pdf.

Halpin, D. W., & Woodhead, R. W. (1998). *Construction management, 2nd Edition*. Wiley, New York.

Hammad, A. (2009). An integrated framework for managing labour resources data in industrial construction projects: a knowledge discovery in data (KDD) approach. *PhD Thesis*, University of Alberta. Edmonton, Alberta. Canada.

Lee, J. R., Hsueh, S. L., & Tseng, H. P. (2008). Utilizing data mining to discover knowledge in construction enterprise performance records. *Journal of Civil Engineering and Management*, 14(2), 79–84. DOI:10.3846/1392-3730.2008.14.2.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics.* , 281-297. Retrieved from http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&hanhan=euclid.bsmsp/1200512992.

Mohamed, Y., Borrego, D., Francisco, L., Al-Hussein, M., AbouRizk, S., & Herman, U. (2007). Simulation-based scheduling of module assembly yards: case study. *Engineering, Construction and Architectural Management*, 14(3), 293-331. DOI: 10.1108/09699980710744926.

Pavan, K., Rao, A., & Sridhar, G. (2010). Single pass seed selection algorithm for k-means. *Journal of Computing Science* , 60-66. DOI:10.3844/jcssp.2010.60.66.

Pham, D., Dimov, S., & Nguyen, C. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* , 219(1), 103-119. DOI:10.1243/095440605X8298.

Rehman, M., & Mehdi, S. (2006). Comparison of density-based clustering algorithms. *IADIS Virtual Multi Conference on Computer Science and Information Systems* , 218-222. Retrieved from http://www.iadis.org/Multi2006/Papers/16/S015_DS.pdf.

Sadeghi, N., & Fayek, A. R. (2008). A framework for simulating industrial construction processes. *Proceedings of the 2008 Winter Simulation Conference*, 2396-2401. Retrieved from http://www.informs-sim.org/wsc08papers/300.pdf.

Scheffer, T. (2004). Finding association rules that trade support optimally against confidence. 1-17. Retrieved from http://www.cs.uni-potsdam.de/ml/publications/ida-rules.pdf.

Seydel, J. (2003). Evaluating and comparing bidding optimization effectiveness*. Journal of Construction Engineering and Management*, 285-292. DOI:10.1061/(ASCE)0733-9364(2003)129:3(285).

Shi, J. J., & Halpin, D. W. (2003). Enterprise resource planning for construction business management. *Journal of Construction Engineering and Management*, 129(2), 214-221. DOI:10.1061/(ASCE)0733-9364(2003)129:2(214).

Soibelman, L., & Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering* , 39-48. DOI:10.1061/(ASCE)0887-3801(2002)16:1(39).

Soibelman, L., & Kim, H. (2000). Generating construction knowledge with knowledge discovery in databases. *Computing in Engineering*, 279(40513),118. DOI:10.1061/40513(279)118.

Song, L., & AbouRizk, S. (2008). Measuring and modeling labor productivity using historical data. *Journal of Construction Engineering and Management*, 134(10), 786-794. DOI:10.1061/(ASCE)0733-9364(2008)134:10(786).

Song, L., Mohamed, Y., & AbouRizk, S. (2009). Early contactor involvement in design and its impact on construction schedule performance. *Journal of Management in Engineering* , 12-20. DOI:10.1061/(ASCE)0742-597X(2009)25:1(12).

Taghaddos, H., AbouRizk, S., Mohamed, Y., & Hermann, U. (2010). Simulation-based multiple heavy lift planning in industrial construction. *Construction Research Congress* , 373(41109), 35. DOI:10.1061/41109(373)35.

Technologies, M. (2010). EasyFit 5.5 Professional Version.

Wang, P., Mohamed, Y., AbouRizk, S., & Rawa, A. R. (2009). Flow production of pipe spool fabrication: simulation to support implementation of lean technique. *Journal of Construction Engineering and Management*, 135(10), 1027-1038. DOI:10.1061/(ASCE)CO.1943-7862.0000068.

Winston, W. L. (2011). *Microsoft Excel 2010: data analysis and business modeling*. Microsoft Press.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques, second edition*. San Francisco, CA: Morgan Kauffman Publishers.

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., et al. (2007). Top 10 algorithms in Data Mining. *Knowledge and Information Systems* , 14(1),1-37. DOI:10.1007/s10115-007-0114-2.

Zaïane, O. R., Foss, A., Lee, C., & Wang, W. (2002). On data clustering analysis: Scalability, constraints and validation. *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 28-39. http://webdocs.cs.ualberta.ca/~zaiane/postscript/pakddZaiane.pdf.

Zhang, J., Ma, T., & Shen, Q. (2004). Application of data warehouse and data mining in construction management. *Proceedings of 10th International Conference on Computing in Civil and Building Engineering*, 1-12. Retrieved from http://e-pub.uni-weimar.de/volltexte/2004/127/pdf/icccbe-x_182.pdf.

Zhang, Y. (2010). The logistics information sharing platform of comprehensive transportation. *ICLEM 2010: Logistics for Sustained Economic Development-Infrastructure, Information, Integration*. *Proceedings of the 2010 International Conference of Logistics Engineering and Management.*, 2158-2164. DOI:10.1061/41139(387)300.

Zhiliang, M., Wong, K., Heng, L., & Jun, Y. (2005). Utilizing exchanged documents in construction projects for decision support based on data warehousing technique. *Automation in Construction*, 14(3), 405-412. DOI:10.1016/j.autcon.2004.08.016.

7. Appendices

7.1. Case Study No. 3: Analysis aggregated quantities of fabrication areas

7.1.1. Supports area

DESC	QTY	SIZEDESC	UNITPRICE	MH		Cluster
PIPESUPPORTLS13	2	16	961.72	6	-->	0
PIPESUPPORTGU11	15	4	187.00	6	-->	1
PIPESUPPORTGU04A1	15	4	235.78	6	-->	1
PIPESUPPORTGU03A1	12	2	185.84	6	-->	8
PIPESUPPORTGU10	6	2	141.70	6	-->	8
PIPESUPPORTLS11	2	2	188.16	6	-->	8
PIPESUPPORTGU12	6	14	235.78	6	-->	14
PIPESUPPORTLS13	2	14	961.72	6	-->	14
PIPESUPPORTFC04A2	1	2	322.90	6	-->	8
PIPESUPPORTGU12	2	12	235.78	6	-->	2
PIPESUPPORTSHS0099	1	12	0.00	6	-->	2
PIPESUPPORTSHS0131	1	12	0.00	6	-->	2
PIPESUPPORTSH25A3	9	12	1475.11	6	-->	2
PIPESUPPORTSH26A7	1	12	1779.42	6	-->	2
PIPESUPPORTBS07A8	2	8	636.50	6	-->	4
PIPESUPPORTBS23C4	1	8	192.81	6	-->	4
PIPESUPPORTSHS0130	1	12	0.00	6	-->	2
PIPESUPPORTSHS0012	1	12	0.00	6	-->	2
PIPESUPPORTFC211	1	4	734.07	6	-->	1
PIPESUPPORTGU12	2	18	235.78	6	-->	12
PIPESUPPORTLS13	1	18	961.72	6	-->	12
PIPESUPPORTGU01	2	2	183.52	6	-->	8
PIPESUPPORTGU11	3	6	187.00	6	-->	5
PIPESUPPORTLS12	2	6	326.38	6	-->	5
PIPESUPPORTLS02A	1	4	351.93	6	-->	1
PIPESUPPORTGU03A1	1	1.5	185.84	6	-->	3
PIPESUPPORTFC401	2	2	318.25	6	-->	8

Table 7-1: Sample *clustering* results from implementation of a *DBScan algorithm* in unified fabrication data (supports area)

Cluster	Instances	Percentage
8	587	20%
5	316	11%
6	284	10%
9	288	10%
1	256	9%
4	233	8%
10	208	7%
11	199	7%
2	110	4%
0	95	3%
3	86	3%
7	66	2%
13	71	2%
12	43	1%
14	36	1%
15	16	1%
16	11	0%
17	8	0%

Table 7-2: Clusters created by the *DBScan algorithm* (supports area)

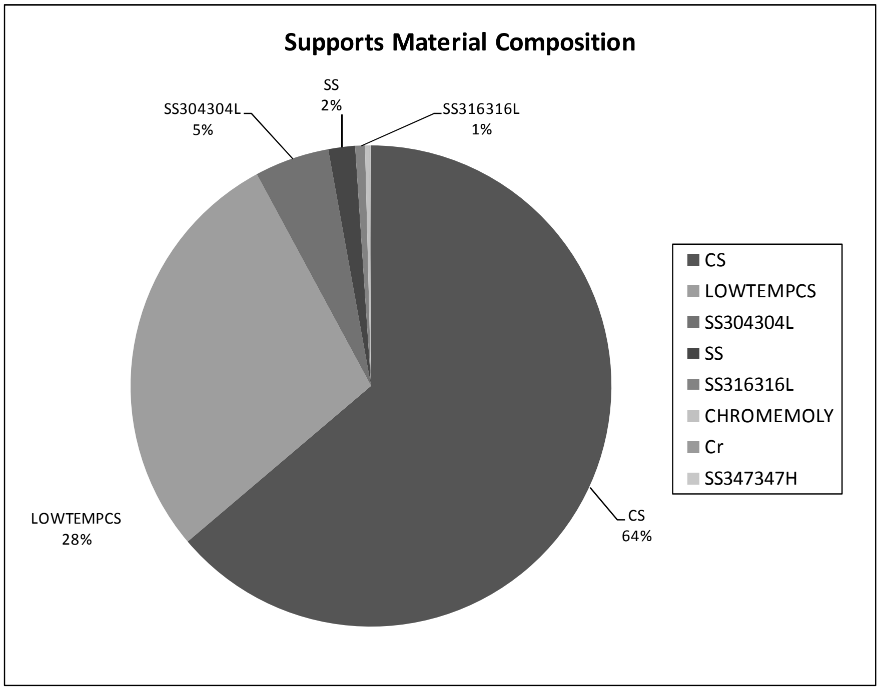

Figure 7-1: Supports material Composition Chart

Rule No.	*Attribute No. 1*	*Support*		*Attribute No. 2*	*Support*	*Confidence*
1	Material=Cr	2	==>	Presence=Large	2	70.12%
2	Material=LOWTEMPCS	170	==>	Presence=Large	108	62.85%
3	Material=CHROMEMOLY	10	==>	Presence=Small	6	55.34%
4	Material=SS316316L	17	==>	Presence=Small	9	53.41%
5	Material=SS	21	==>	Presence=Small	11	53.32%
6	Material=SS304304L	21	==>	Presence=Large	11	53.32%
7	Presence=Small	211	==>	Material=CS	113	53.25%
8	Material=SS	21	==>	Presence=Large	10	49.84%
9	Material=SS304304L	21	==>	Presence=Small	10	49.84%
10	Material=SS316316L	17	==>	Presence=Large	8	49.60%
11	Material=CHROMEMOLY	10	==>	Presence=Large	4	46.11%
12	Material=CS	331	==>	Presence=Large	218	45.01%
13	Presence=Large	362	==>	Material=CS	218	40.91%
14	Material=LOWTEMPCS	170	==>	Presence=Small	62	37.24%
15	Presence=Small	211	==>	Material=LOWTEMPCS	62	30.88%
16	Material=CS	331	==>	Presence=Small	113	21.01%
17	Presence=Large	362	==>	Material=LOWTEMPCS	108	15.96%
18	Presence=Small	211	==>	Material=SS	11	4.31%
19	Presence=Small	211	==>	Material=SS304304L	10	4.14%
20	Presence=Small	211	==>	Material=SS316316L	9	4.01%

Table 7-3: Association Rules extracted from Cluster No. 8 (supports area)

7.1.2. Welding area

DESC	QTY	SIZEDESC	UNITPRICE	MH	DIW		Cluster
ORIFICEFLANGESETWELDNECKSTD300lbRF125250AARHASTMA105N	1	16	479.77	6.08	7.94	-->	0
FLANGEWELDNECKSTD150lbRF125250AARHASTMA350GrLF2Class1	25	2	17.63	6.08	7.94	-->	1
OLETWELDOLETSTDASTMA105N	3	2	9.15	6.08	7.94	-->	1
FLANGEWELDNECKSTD150lbRF125250AARHASTMA105N	1	14	187.04	6.08	7.94	-->	9
FLANGEWELDNECKSTD150lbRF125250AARHASTMA105N	14	2	12.88	6.08	7.94	-->	1
OLETWELDOLETSTDASTMA105N	1	8	162.12	6.08	7.94	-->	12
OLETWELDOLETSTDASTMA105N	1	2	9.15	6.08	7.94	-->	1
FLANGEWELDNECKSTD150lbRF125250AARHASTMA350GrLF2Class1	7	4	31.23	6.08	7.94	-->	7
OLETWELDOLETSTDASTMA105N	5	3	24.27	6.08	7.94	-->	3
FLANGEWELDNECKSTD150lbRF125250AARHASTMA105N	2	3	18.76	6.08	7.94	-->	3
OLETWELDOLETSTDASTMA105N	1	12	419.80	6.08	7.94	-->	2
FLANGEWELDNECKXS150lbRF125250AARHASTMA105N	26	2	13.48	6.08	7.94	-->	1
OLETWELDOLETXSASTMA105N	1	2	15.94	6.08	7.94	-->	1
FLANGEWELDNECKSTD150lbRF125250AARHASTMA105N	2	6	35.96	6.08	7.94	-->	13
BUTTWELDSTDLOWTEMPCS	3	4	142.86	2.88	7.94	-->	7
ORIFICEFLANGESETWELDNECKSTD300lbRF125250AARHASTMA105N	1	18	921.97	6.08	7.94	-->	11
FLANGEWELDNECKSTD150lbRF125250AARHASTMA105N	1	18	301.23	6.08	7.94	-->	11
BUTTWELDSTDLOWTEMPCS	1	2	101.05	1.61	7.94	-->	1
ELBOW90DEGLRBWSTDASTMA420GrWPL6Welded	24	8	81.65	6.08	7.94	-->	12
FLANGEWELDNECKSTD150lbRF125250AARHASTMA350GrLF2Class1	21	8	110.33	6.08	7.94	-->	12
OLETWELDOLETSTDASTMA105N	1	12	397.69	6.08	7.94	-->	2
OLETWELDOLETSTDASTMA105N	1	3	25.19	6.08	7.94	-->	3
OLETWELDOLETXSASTMA105N	2	2	19.60	6.08	7.94	-->	1
FLANGEWELDNECKSTD150lbRF125250AARHASTMA105N	2	4	25.62	6.08	7.94	-->	7
OLETWELDOLETXSASTMA105N	2	2	15.94	6.08	7.94	-->	1
FLANGEWELDNECKSTD300lbRF125250AARHASTMA105N	2	2	16.99	6.08	7.94	-->	1
FLANGEWELDNECKSTD150lbRF125250AARHASTMA105N	1	20	371.48	6.08	7.94	-->	4

Table 7-4: Sample *clustering* results from implementation of a *DBScan algorithm* in unified fabrication data (welding area)

Cluster	Instances	Percentage
1	69	15%
12	48	10%
5	41	9%
6	41	9%
3	36	8%
7	36	8%
8	35	8%
13	33	7%
0	25	5%
2	21	5%
10	18	4%
4	15	3%
11	15	3%
14	12	3%
9	8	2%
15	6	1%

Table 7-5: Clusters created by the *DBScan algorithm* (welding area)

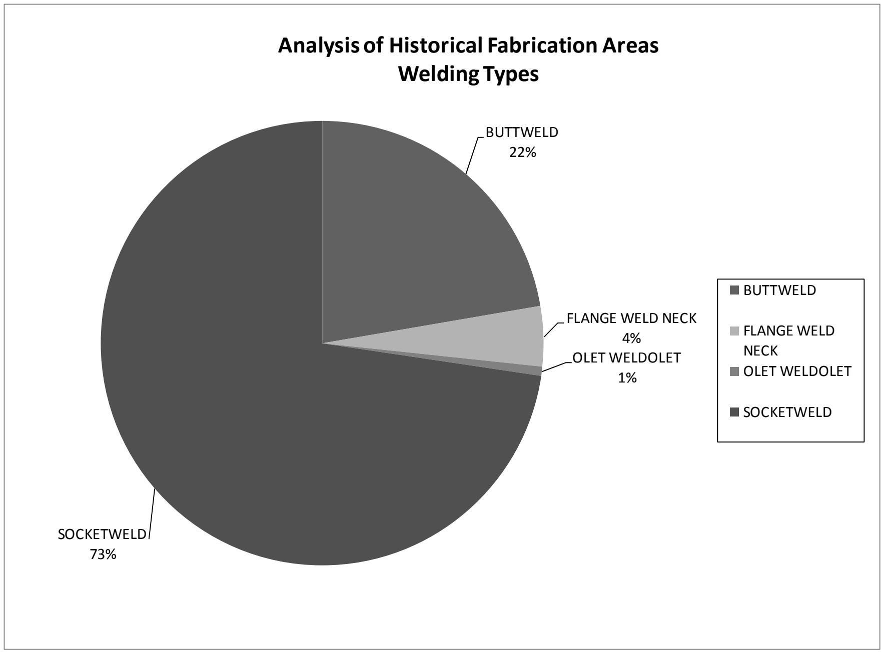

Figure 7-2: Welding types Composition Chart

Rule No.	*Attribute No. 1*	*Attribute No. 2*	*Support*		*Attribute No. 3*	*Attribute No. 4*	*Support*	*Confidence*
1	Material=CS		6	==>	Presence=Large		6	0.98761
2	Schedule=3000lb		4	==>	Weldtype=SOCKETWELD		4	0.97386
3	Schedule=Sch80		3	==>	Weldtype=BUTTWELD		3	0.95684
4	Material=150lb		3	==>	Weldtype=FLANGEWELDNECK		3	0.95684
5	Weldtype=OLETWELDOLET		2	==>	Presence=Small		2	0.92263
6	Schedule=Sch160		2	==>	Weldtype=BUTTWELD		2	0.92263
7	Material=SS304304L		2	==>	Presence=Small		2	0.92263
8	Weldtype=FLANGEWELDNECK	Presence=Small	2	==>	Schedule=STD		2	0.92263
9	Weldtype=FLANGEWELDNECK	Presence=Large	2	==>	Material=150lb		2	0.92263
10	Weldtype=SOCKETWELD		5	==>	Schedule=3000lb		4	0.64195
11	Weldtype=FLANGEWELDNECK		4	==>	Schedule=STD		3	0.57948
12	Weldtype=FLANGEWELDNECK		4	==>	Material=150lb		3	0.57948
13	Schedule=XS		4	==>	Presence=Large		3	0.57948
14	Material=LOWTEMPCS		4	==>	Weldtype=BUTTWELD		3	0.57948
15	Material=LOWTEMPCS		4	==>	Presence=Large		3	0.57948
16	Schedule=STD		6	==>	Presence=Small		4	0.5262
17	Material=CS		6	==>	Weldtype=BUTTWELD	Presence=Large	4	0.5262
18	Weldtype=BUTTWELD	Presence=Large	6	==>	Material=CS		4	0.5262
19	Schedule=Sch80		3	==>	Weldtype=BUTTWELD	Presence=Large	2	0.51472
20	Weldtype=FLANGEWELDNECK	Material=150lb	3	==>	Presence=Large		2	0.51472

Table 7-6: Association Rules extracted from Cluster No. 1 (welding area)

7.1.3. Handling valves area

DESC	QTY	SIZEDESC	UNITPRICE	MH		Cluster
HANDLEMANUALVALVE150lbCS	21	2	84.79	0.81	-->	0
HANDLEMANUALVALVE150lbLOWTEMPCS	3	4	155.64	1.50	-->	1
HANDLEMANUALVALVE150lbCS	5	3	120.80	1.15	-->	6
HANDLEMANUALVALVE300lbCS	1	3	144.03	1.38	-->	6
HANDLEMANUALVALVE150lbCS	2	12	540.10	5.18	-->	10
HANDLEMANUALVALVE150lbCS	2	6	204.42	1.96	-->	4
HANDLEMANUALVALVE150lbLOWTEMPCS	8	2	84.79	0.81	-->	0
HANDLEMANUALVALVE150lbCS	1	4	155.64	5.03	-->	1
HANDLEMANUALVALVE300lbCS	2	2	96.40	5.03	-->	0
HANDLEMANUALVALVE150lbLOWTEMPCS	1	3	120.80	5.03	-->	6
HANDLEMANUALVALVE150lbCS	30	2	84.54	0.81	-->	0
HANDLEMANUALVALVE150lbCS	3	3	120.44	1.15	-->	6
HANDLEMANUALVALVE150lbCS	1	8	298.79	2.88	-->	3
HANDLEMANUALVALVE150lbCS	6	12	538.52	5.18	-->	10
HANDLEMANUALVALVE150lbCS	5	16	836.15	8.05	-->	11
HANDLEMANUALVALVE300lbCS	12	2	96.13	0.92	-->	0
HANDLEMANUALVALVE300lbCS	3	4	179.50	1.73	-->	1
HANDLEMANUALVALVE300lbCS	3	6	311.54	2.99	-->	4
HANDLEMANUALVALVE300lbCS	5	10	621.91	5.98	-->	2
HANDLEMANUALVALVE600lbCS	1	8	778.55	7.48	-->	3
HANDLEMANUALVALVE900lbCS	1	3	227.70	2.19	-->	6
HANDLEMANUALVALVE900lbCS	2	10	1316.75	12.65	-->	2
HANDLEMANUALVALVE1500lbCS	1	2	119.60	1.15	-->	0
HANDLEMANUALVALVE150lbLOWTEMPCS	98	2	84.54	0.81	-->	0
HANDLEMANUALVALVE150lbLOWTEMPCS	2	3	120.44	1.15	-->	6
HANDLEMANUALVALVE150lbLOWTEMPCS	8	6	203.83	1.96	-->	4
HANDLEMANUALVALVE150lbLOWTEMPCS	6	8	298.79	2.88	-->	3

Table 7-7: Sample *clustering* results from implementation of a *DBScan algorithm* in unified fabrication data (handling valves area)

Cluster	Instances	Percentage
0	77	17%
4	70	15%
6	58	13%
1	50	11%
3	44	10%
2	36	8%
7	32	7%
9	23	5%
10	23	5%
5	16	3%
11	11	2%
12	8	2%
8	6	1%
13	6	1%

Table 7-8: Clusters created by the *DBScan algorithm* (handling valves area)

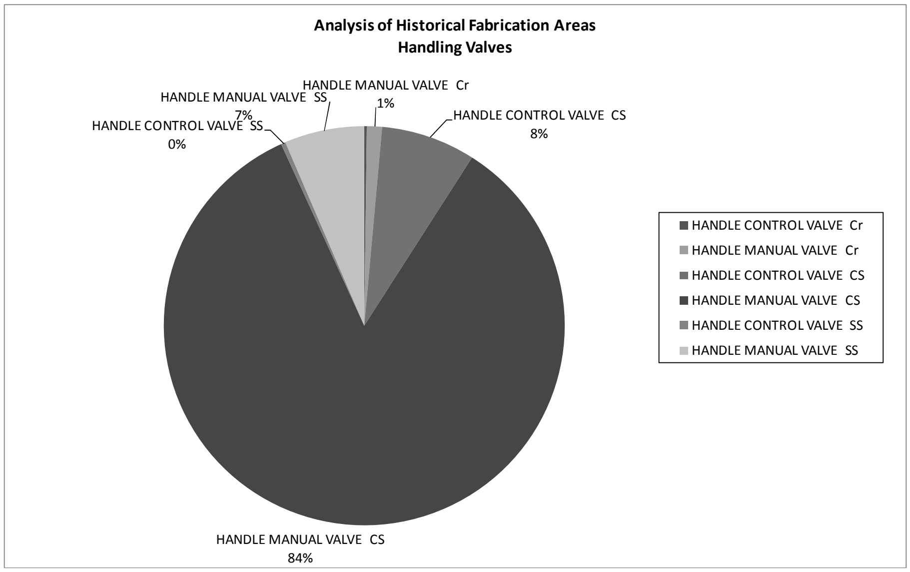

Figure 7-3: Handling valves Composition Chart

Rule No.	*Attribute No. 1*	*Attribute No. 2*	*Attribute No. 3*	*Support*		*Attribute No. 4*	*Support*	*Confidence*
1	Desc=HANDLEMANUALVALVE	Type=150lb	Material=LOWTEMPCS	5	==>	Presence=Large	5	0.94958
2	Desc=HANDLEMANUALVALVE	Type=150lb	Material=CS	5	==>	Presence=Large	5	0.94958
3	Type=150lb	Material=CS	Presence=Large	5	==>	Desc=HANDLEMANUALVALVE	5	0.94958
4	Material=SS304304L			4	==>	Desc=HANDLEMANUALVALVE	4	0.92867
5	Material=SS316316L			4	==>	Desc=HANDLEMANUALVALVE	4	0.92867
6	Type=600lb	Material=LOWTEMPCS		3	==>	Presence=Small	3	0.89497
7	Desc=HANDLEMANUALVALVE	Type=150lb		11	==>	Presence=Large	10	0.84575
8	Type=150lb	Presence=Large		11	==>	Desc=HANDLEMANUALVALVE	10	0.84575
9	Material=SS347347H			2	==>	Presence=Small	2	0.83877
10	Type=150lb	Material=CS	Presence=Small	2	==>	Desc=HANDLECONTROLVALVE	2	0.83877
11	Desc=HANDLEMANUALVALVE	Type=300lb	Material=CS	9	==>	Presence=Large	8	0.79541
12	Desc=HANDLEMANUALVALVE	Material=CS		21	==>	Presence=Large	18	0.78494
13	Type=600lb	Presence=Large		8	==>	Desc=HANDLEMANUALVALVE	7	0.76604
14	Presence=Large			41	==>	Desc=HANDLEMANUALVALVE	33	0.75702
15	Material=SS			7	==>	Desc=HANDLEMANUALVALVE	6	0.7336
16	Type=150lb	Material=LOWTEMPCS		7	==>	Presence=Large	6	0.7336
17	Type=600lb			21	==>	Desc=HANDLEMANUALVALVE	16	0.70305
18	Material=CHROMEMOLY			6	==>	Presence=Small	5	0.69749
19	Type=300lb	Material=CS		16	==>	Presence=Large	12	0.68258
20	Type=600lb	Material=CS	Presence=Small	3	==>	Desc=HANDLECONTROLVALVE	2	0.54636

Table 7-9: Association Rules extracted from Cluster No. 1 (handling valves area)

7.2. Case Study No. 4: Analysis of fabrication man-hours

Charts derived from Association Rules generated by an *Apriori algorithm.*

Material	Size	Observed	Schedule	Observed	%
CS	1"	31	Sch 3000	29	94%
			Sch 160	2	6%

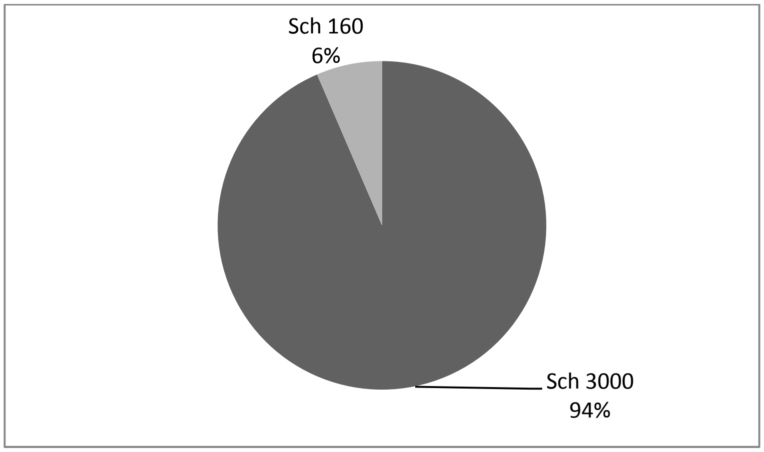

Figure 7-4: Material (CS), Size (1") and Schedule relationship

Material	Size	Observed	Schedule	Weld Type	Observed	%
CS	1"	31	Sch 3000	SW	24	77%
			Sch 3000	OL	4	13%
			Sch 3000	BW	1	3%
			Sch 160	BW	1	3%
			Sch 160	OL	1	3%

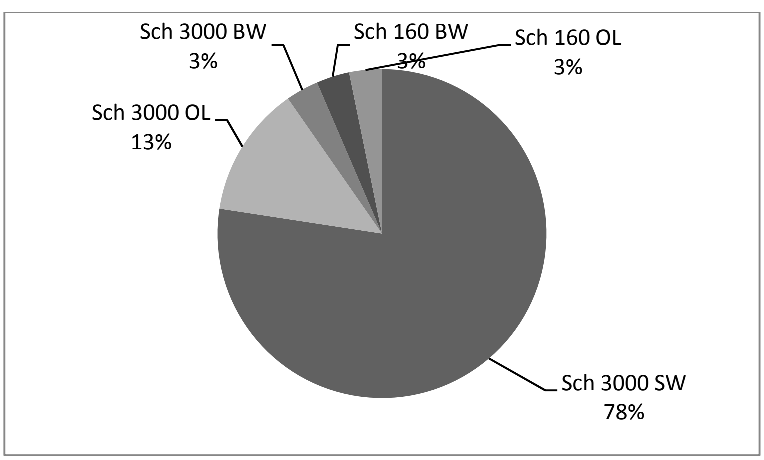

Figure 7-5: Material (CS), Size (1"), Schedule and Weld type relationship

Material	Size	Observed	Schedule	Observed	%
CS	2"	45	Sch 0.25	5	11%
			Sch 3000	31	69%
			Sch 6000	9	20%

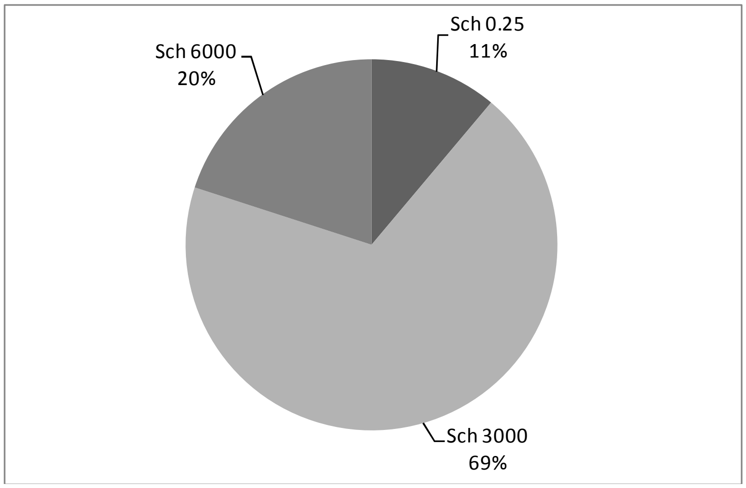

Figure 7-6: Material (CS), Size (2”) and Schedule relationship

Material	Size	Observed	Schedule	Weld Type	Observed	%
CS	2"	45	Sch 0.25	FW	5	11.11%
			Sch 3000	SW	30	66.67%
				OL	1	2.22%
			Sch 6000	BW	1	2.22%
				SW	8	17.78%

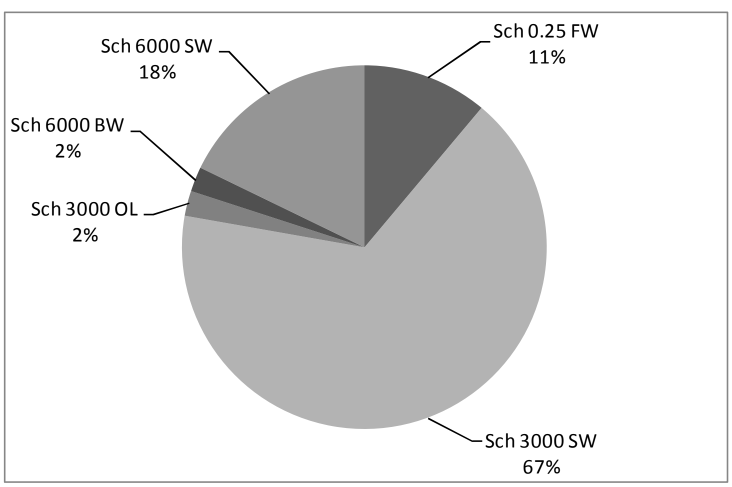

Figure 7-7: Material (CS), Size (2”), Schedule and Weld type relationship

Material	Size	Observed	Schedule	Observed	%
CS	6"	22	Sch STD	16	73.00%
			Sch 160	6	27.27%

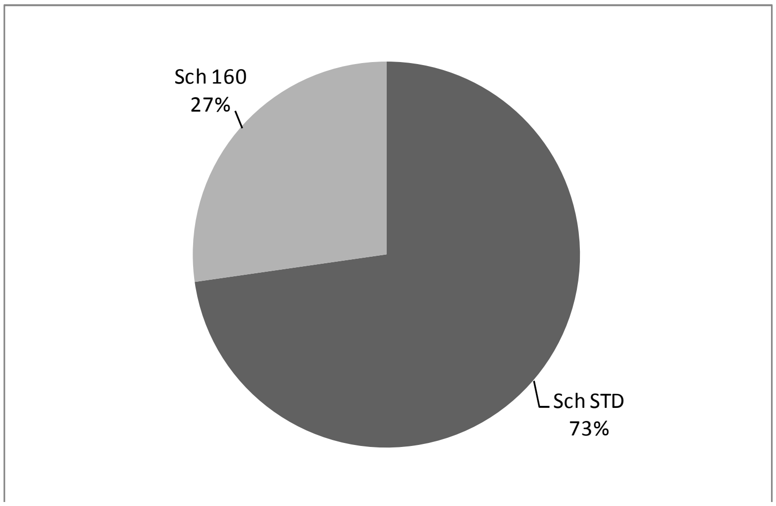

Figure 7-8: Material (CS), Size (6”) and Schedule relationship

Material	Size	Observed	Schedule	Weld Type	Observed	%
CS	6"	22	Sch STD	BW	16	72.73%
			Sch 160	BW	6	27.27%

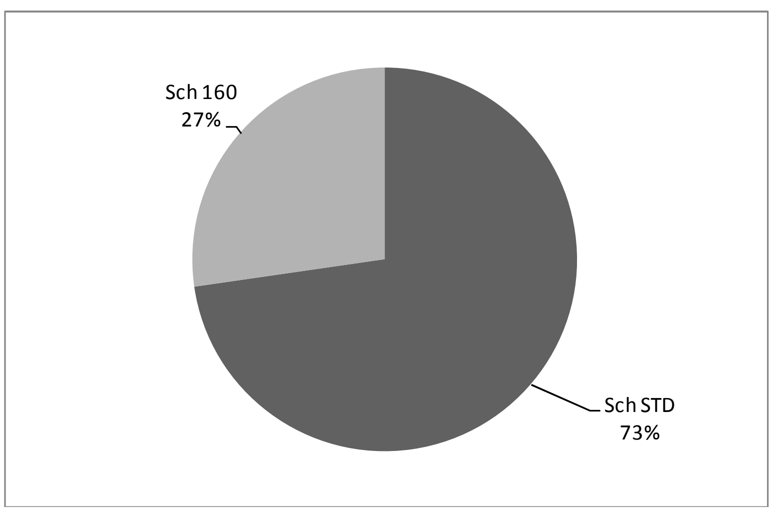

Figure 7-9: Material (CS), Size (6”), Schedule and Weld type relationship

Material	Size	Observed	Schedule	Observed	%
CS	8"	13	Sch 0.25	9	69.00%
			Sch 0.75	1	7.69%
			Sch 140	1	7.69%
			Sch STD	2	15.38%

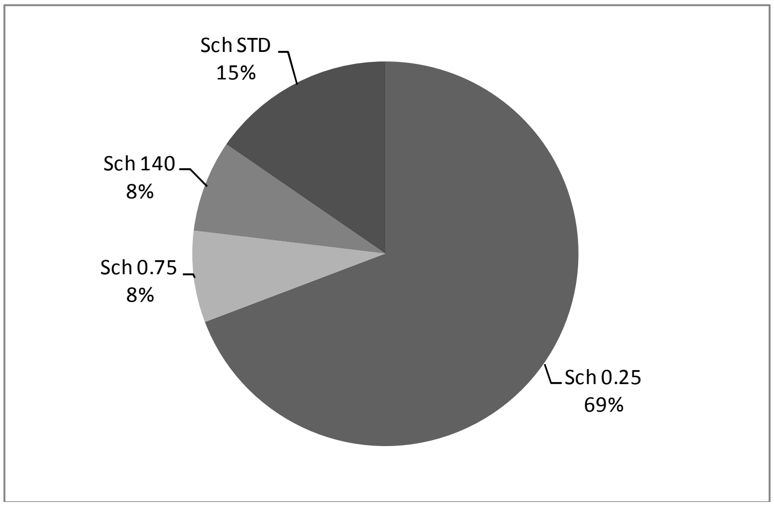

Figure 7-10: Material (CS), Size (8”) and Schedule relationship

Material	Size	Observed	Schedule	Weld Type	Observed	%
CS	8"	13	Sch 0.25	FW	9	69.23%
			Sch 0.75	FW	1	7.69%
			Sch 140	BW	1	7.69%
			Sch STD	BW	2	15.38%

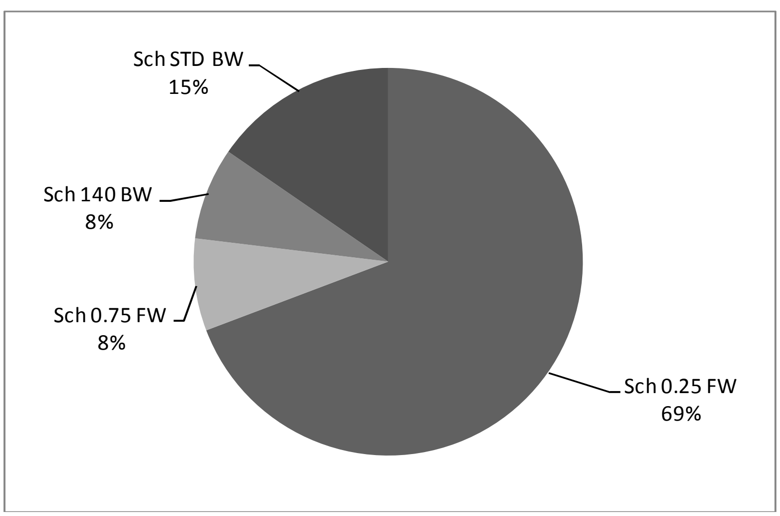

Figure 7-11: Material (CS), Size (8”), Schedule and Weld type relationship

Material	Size	Observed	Schedule	Observed	%
SS	2"	42	Sch 160	12	28.57%
			Sch 3000	27	64.29%
			Sch XXS	3	7.14%

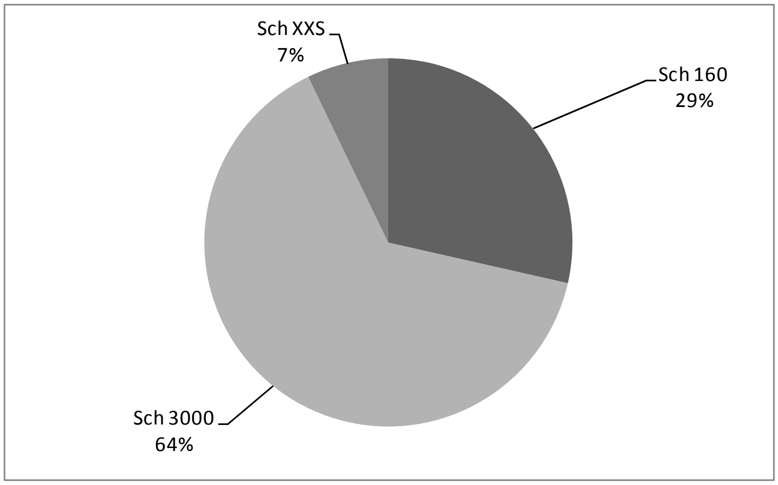

Figure 7-12: Material (SS), Size (2”) and Schedule relationship

Material	Size	Observed	Schedule	Weld Type	Observed	%
SS	2"	42	Sch 160	BW	10	23.81%
				OL	2	4.76%
			Sch 3000	SW	23	54.76%
				OL	3	7.14%
				BW	1	2.38%
			Sch XXS	FW	2	4.76%
				BW	1	2.38%

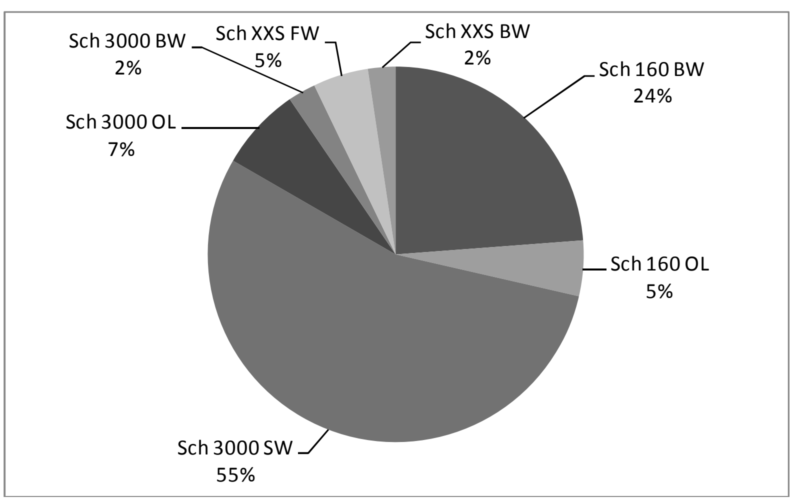

Figure 7-13: Material (SS), Size (2”), Schedule and Weld type relationship

Material	Size	Observed	Schedule	Weld Type	Observed	%
SS	21"	9	0.25	FW	9	100.00%

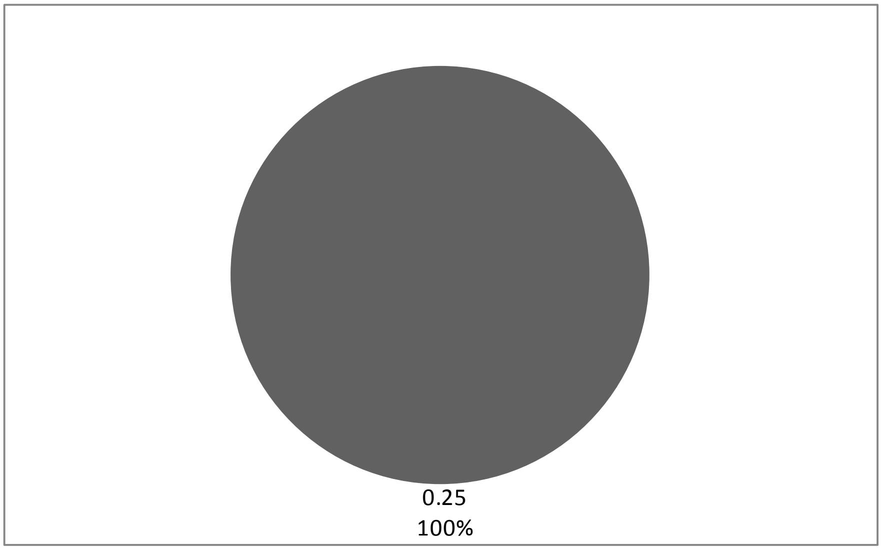

Figure 7-14: Material (SS), Size (21”), Schedule and Weld type relationship

Mhrs per DI	Observed	Material	Observed	%
0.04	12	SS	9	75.00%
		CS	3	25.00%

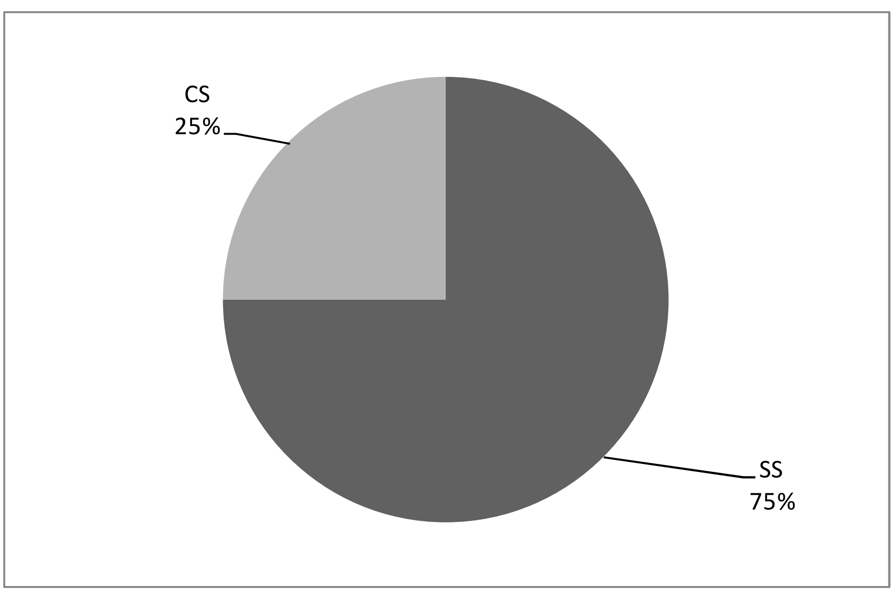

Figure 7-15: Man-hours (0.04) and Material relationship

Mhrs per DI	Observed	Material	Observed	%
0.06	22	CS	15	68.00%
		SS	6	27.27%
		LT	1	4.55%

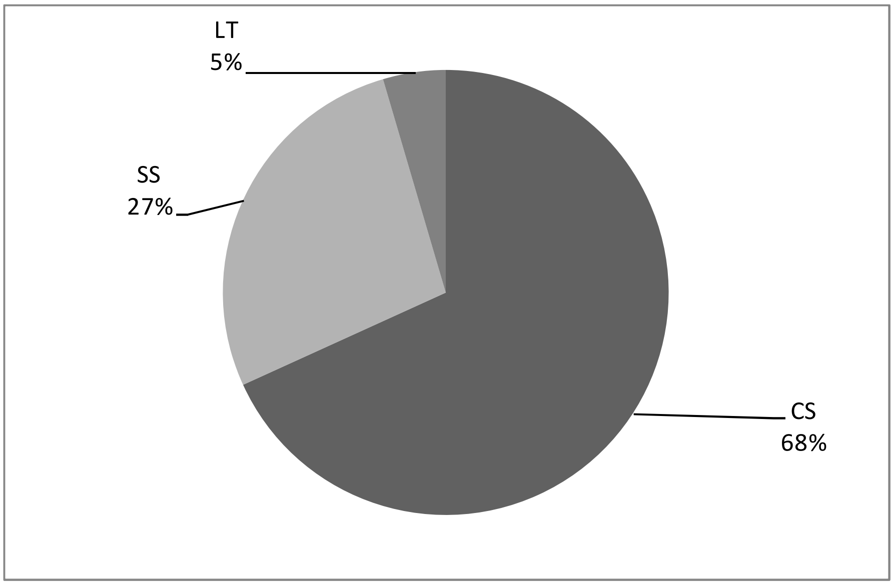

Figure 7-16: Manhours (0.06) and Material relationship

Mhrs per DI	Observed	Material	Observed	%
0.07	25	CS	22	88.00%
		SS	3	12.00%

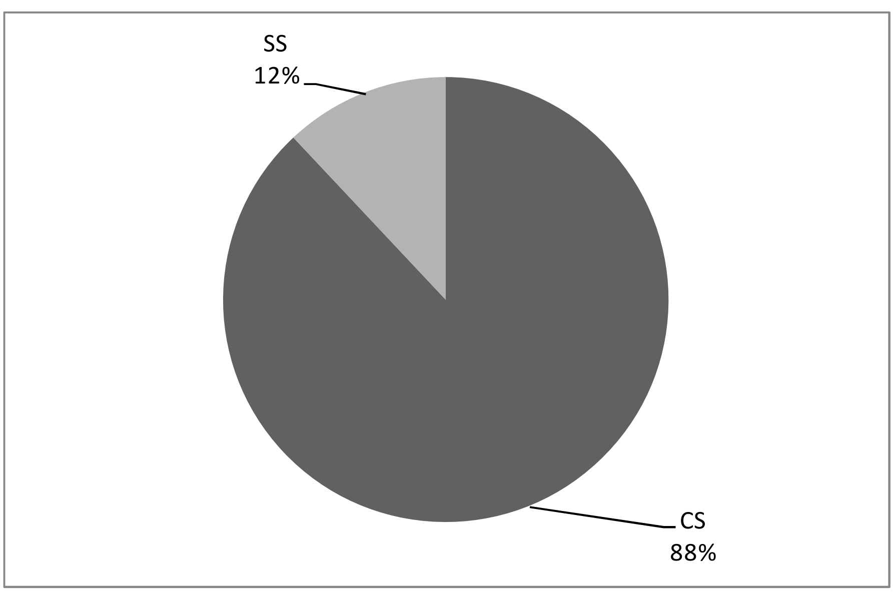

Figure 7-17: Man-hours (0.07) and Material relationship

Mhrs per DI	Observed	Material	Observed	%
0.08	13	CS	12	92.00%
		LT	1	7.69%

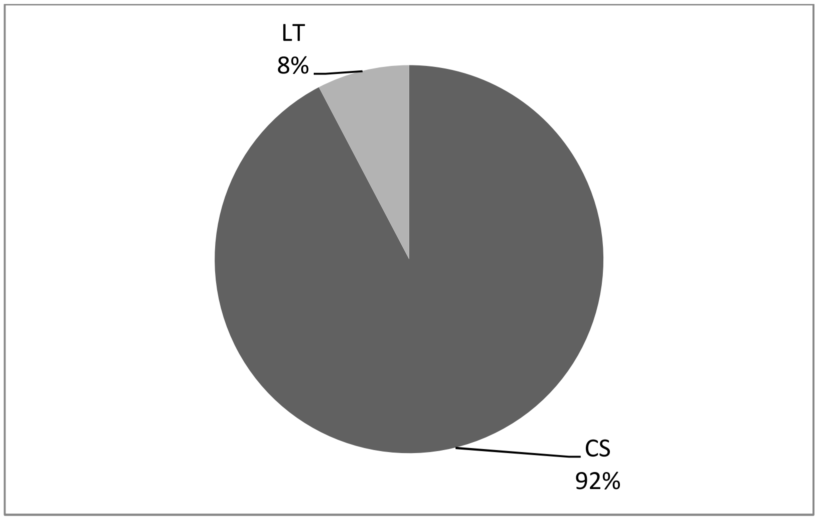

Figure 7-18: Man-hours (0.08) and Material relationship

Mhrs per DI	Observed	Material	Observed	%
0.16	19	SS	12	63.00%
		CS	5	26.32%
		Cr	1	5.26%
		LT	1	5.26%

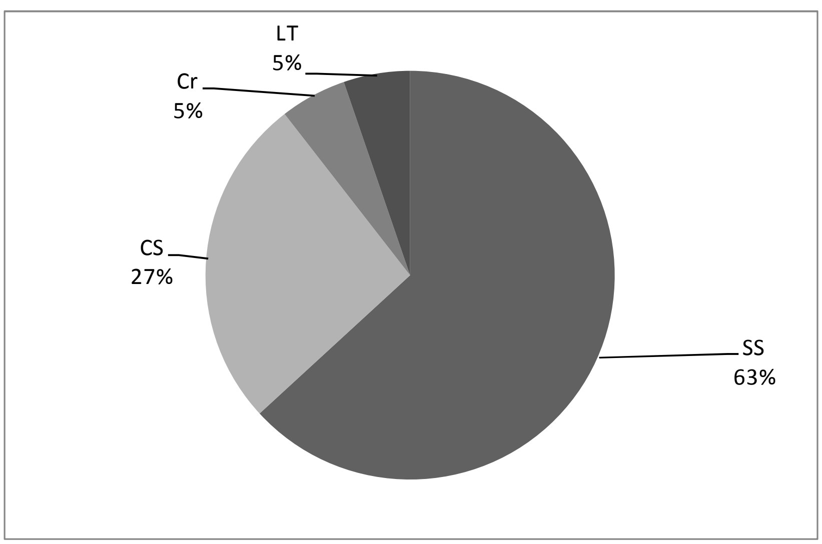

Figure 7-19: Man-hours (0.16) and Material relationship

7.3. Case Study No. 5: Modeling data through probability distributions

Production Reports analysis

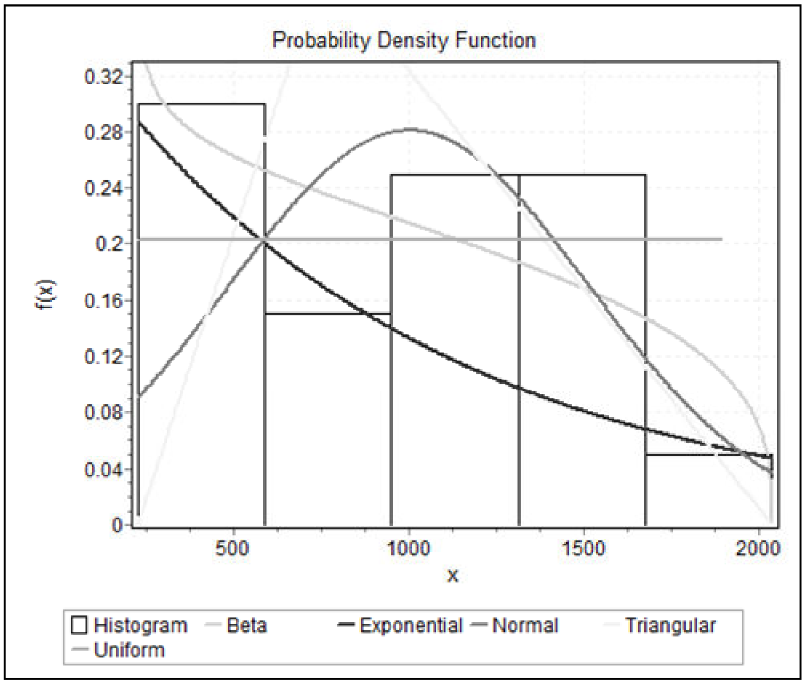

Figure 7-20: Project A DI's Probability Density Function

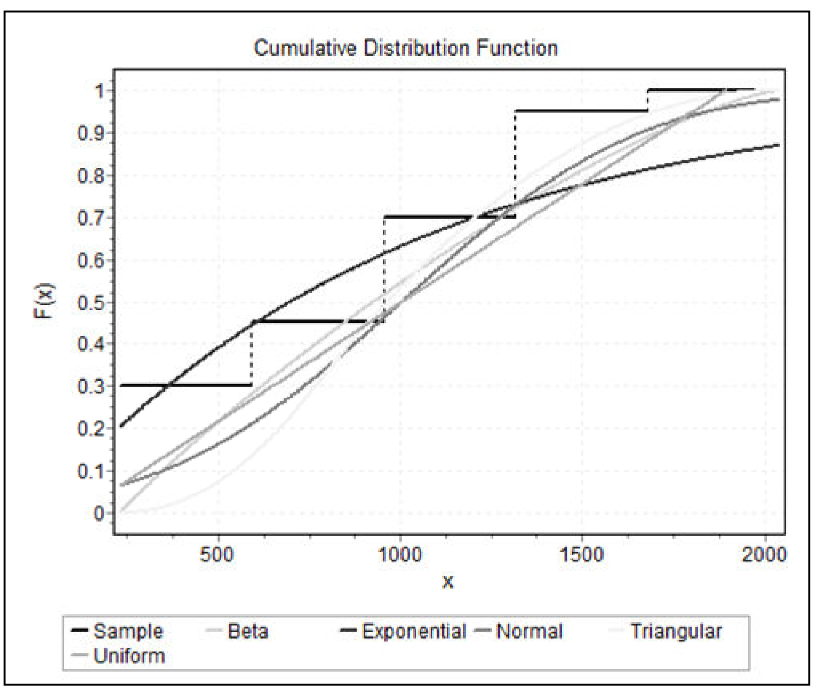

Figure 7-21: Project A DI's Cumulative Distribution Function

Project B fabrication data

	Spools	
No.	Total	DI
1	59	889
2	10	535
3	49	971
4	36	1089
5	36	1251
5	8	469
5	40	697
6	22	854
7	9	658
8	24	1129
9	61	1049
10	26	516
11	20	462
12	35	1785
13	24	1730
14	58	1413
15	53	1224
16	12	346
17	24	1273
18	38	1166
19	63	1226
20	20	369
21	67	1289
22	35	684
23	39	1090
24	22	815
25	15	476
26	8	348
27	31	1367
27	30	1258
27	21	721
28	28	807
28	5	323
28	21	703
29	20	731
29	26	461
29	21	1617
30	16	852
30	10	255
31	24	844
32	18	1166
32	10	606
32	23	889
33	16	447
34	20	1081
34	16	1011
34	17	638
35	29	634
36	24	966
36	13	354
36	13	529
37	26	651
38	38	2270
39	60	2892
40	21	1510
40	7	628
40	12	691
41	24	365
42	41	1206
43	17	130
44	30	1241
45	64	1159
46	129	3257
47	7	459
48	48	1668
49	32	1140
50	15	584
51	20	1213
52	20	1104
53	15	943
54	5	158
55	15	444
56	15	698
57	21	909
58	35	881
59	22	439
60	30	964
61	31	1045
62	15	562
63	68	1677
64	38	1257
65	46	1057
66	63	1006
67	14	521
68	43	1555
69	62	1124
70	25	568
71	25	1096
71	26	1178
71	24	940
72	15	500
73	17	1176
73	15	856
73	20	649
74	14	375
75	23	1290
75	29	688
75	23	1025
76	18	411
77	29	1226
78	23	1078
78	21	975
78	14	294
79	7	212
80	14	681
80	15	546
80	9	229
81	16	105
82	17	1045
82	22	863
82	16	831
83	21	161
84	16	531
84	14	547
84	12	366
85	29	297
86	29	1678
87	30	1726
88	24	785
89	30	300
90	47	646

Average Spools per Mod	35.96
Average DI's per Spool	33.17
Average DI's per Module	1,192.73
Average Welds per Spool	0.13
Average Welds per DI's	0.0039

Table 7-10: Project B fabricated spools and diameter inches per module

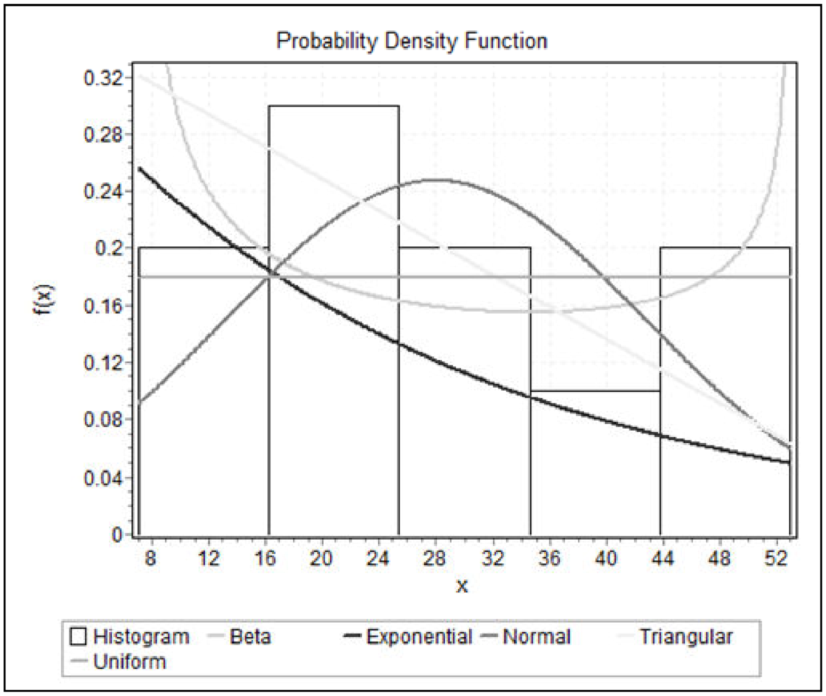

Figure 7-22: Project B spools Probability Density Function

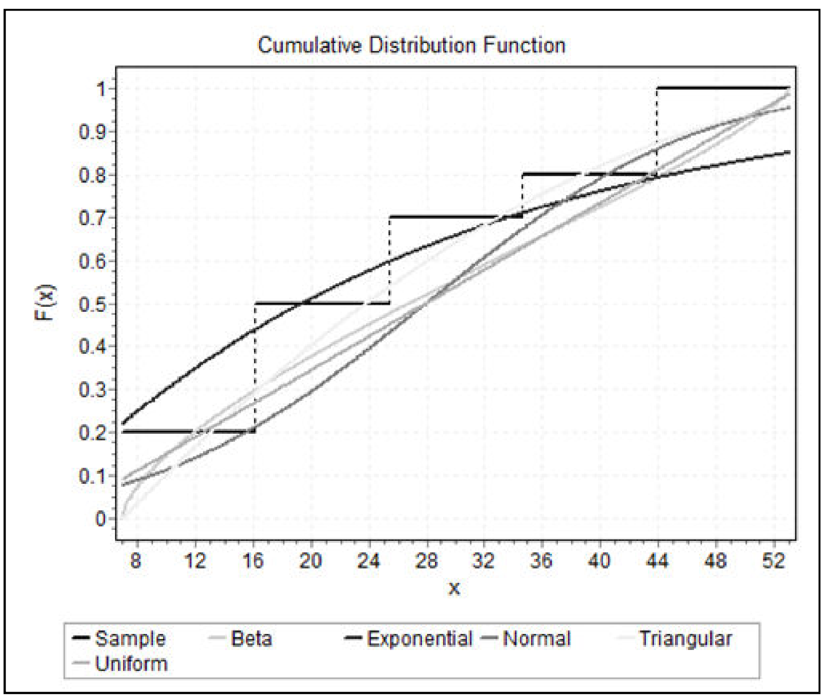

Figure 7-23: Project B spools Cumulative Distribution Function

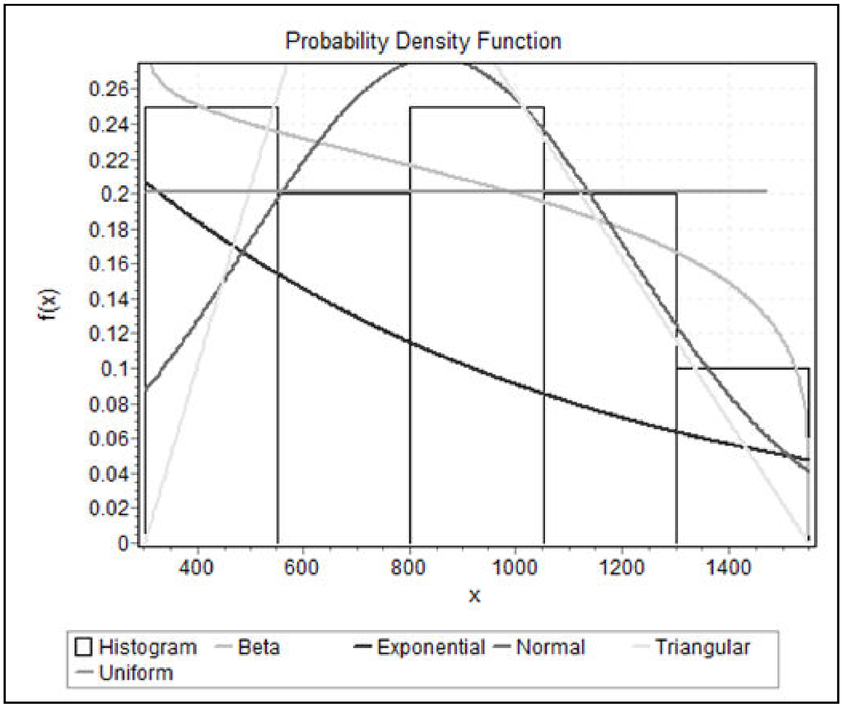

Figure 7-24: Project B DI’s Probability Density Function

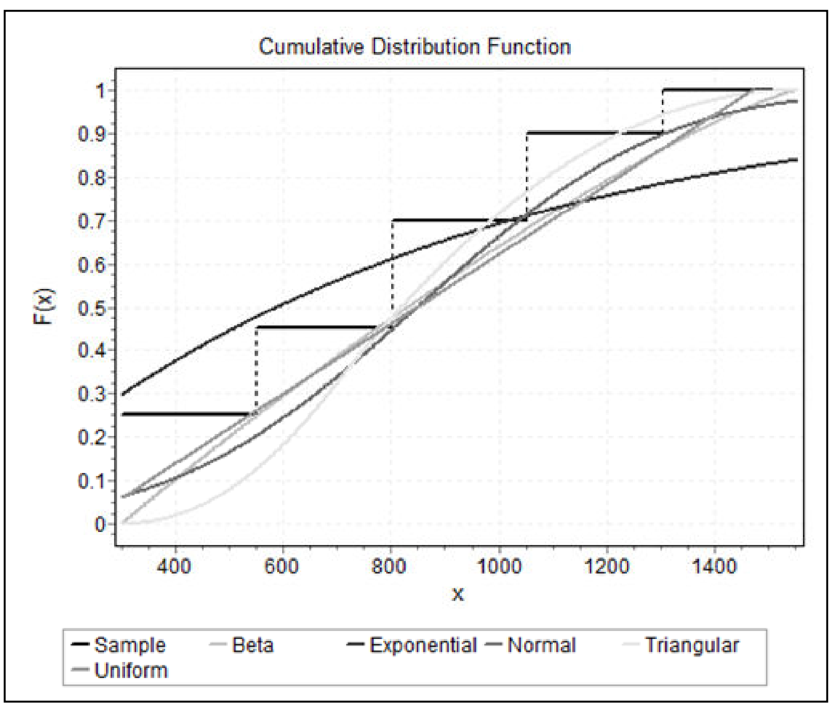

Figure 7-25: Project B DI's Cumulative Distribution Function

Project C fabrication data

	Spools	
No.	Total	DI
1	20	699
2	25	680
3	85	1467
4	98	1808
5	35	988
6	185	3819
7	76	2639
8	162	4402
9	56	1678
10	99	2977
11	32	1366
12	90	2171
13	28	1097
14	340	7604
15	49	833
16	102	3856
17	153	6088
18	21	914
19	35	836
20	158	4593
21	128	3840
22	61	940
23	61	2381
24	125	2122
25	125	4829
26	98	2407
27	131	2132
28	74	1325
29	24	687
30	92	2599
31	84	1750
32	49	2125
33	94	2437
34	90	1662
35	37	1442
36	150	2916
37	135	2506
38	53	1592
39	86	3107
40	70	3350
41	30	554
42	6	306

Average Spools per Mod	76.04
Average DI's per Spool	26.72
Average DI's per Module	2,031.74
Average Welds per Spool	0.01
Average Welds per DI's	0.0055

Table 7-11: Project C fabricated spools and diameter inches per module

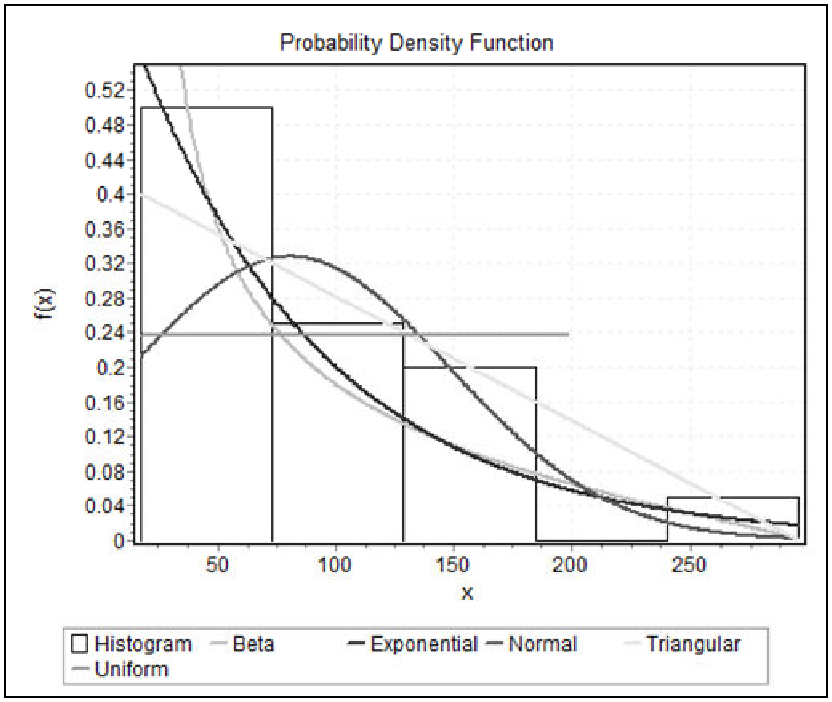

Figure 7-26: Project C spools Probability Density Function

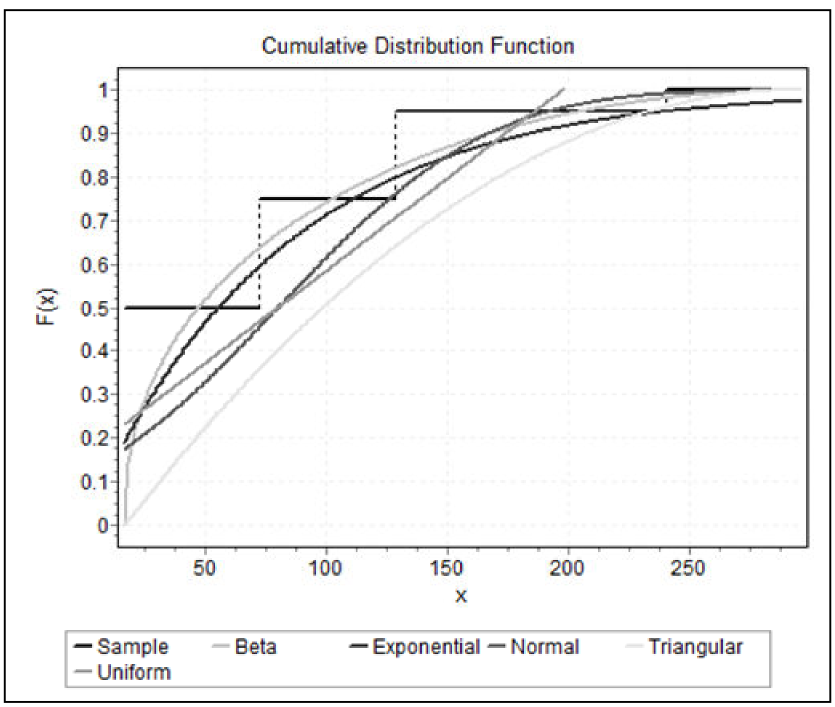

Figure 7-27: Project C spools Cumulative Distribution Function

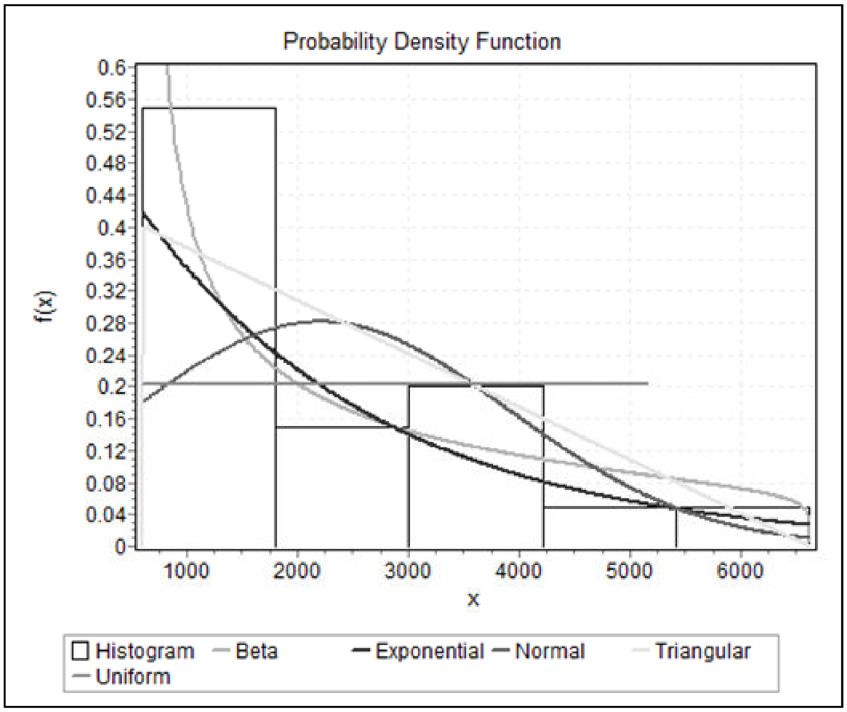

Figure 7-28: Project C DI’s Probability Density Function

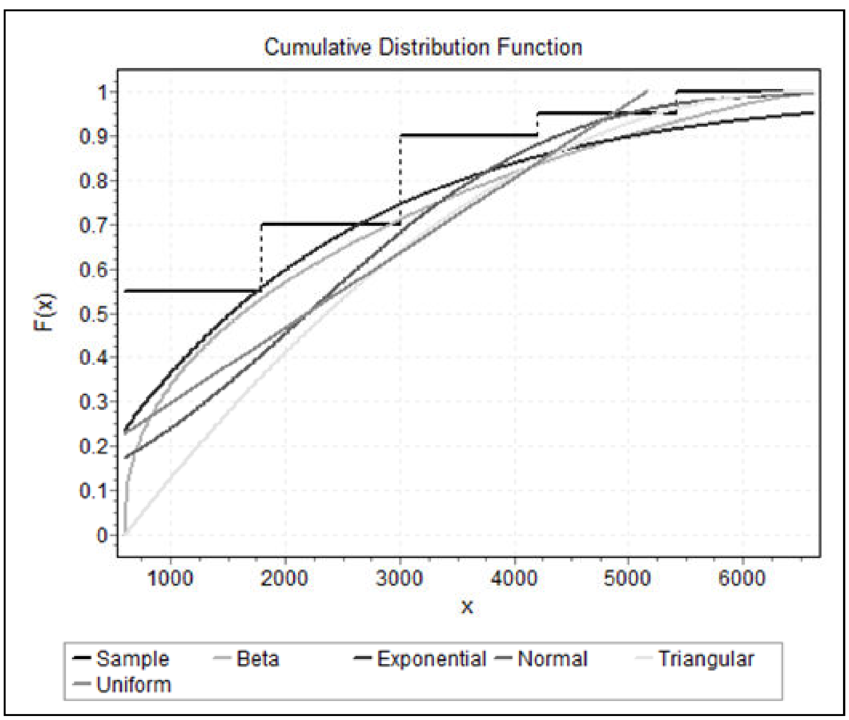

Figure 7-29: Project C DI’s Cumulative Distribution Function

Project D fabrication data

	Spools	
No.	Total	DI
1	109	3196
1	32	1302
1	15	1130
2	107	2929
2	55	2323
2	15	828
3	47	1409
3	17	933
3	7	105
4	102	2673
4	29	1318
4	6	113
5	77	2328
5	25	1012
5	8	49
6	76	2559
6	31	1239
6	14	966
7	48	1694
7	31	1900
7	24	2036
8	61	2256
8	23	1541
8	30	1636
9	46	1579
9	25	1278
9	17	1224
10	71	2416
10	21	822
10	21	998
11	58	2473
11	25	1225
11	24	1213

	Spools	
No.	Total	DI
12	76	2461
12	39	2060
12	17	1042
12	17	1042
13	16	903
14	21	1939
15	18	642
16	33	2406
17	6	405
18	29	1182
19	39	3014
20	25	1789
21	36	1878
22	39	1939
23	45	2508
24	20	1300
25	41	1426
25	22	1064
25	5	268
26	48	1431
26	17	1388
26	3	151
27	70	2701
27	22	1120
27	6	262
28	91	3334
28	17	1484
28	2	41
29	48	1824
29	30	1073
29	13	179
30	97	2482
30	45	1070
30	6	169
31	25	1001
32	20	918
33	30	1059
34	28	1283
35	20	1033
36	13	488
38	45	1125

	Spools	
No.	Total	DI
39	60	1810
40	71	1323
41	40	1561
43	44	2183
44	36	1167
44	25	1687
44	10	1157
45	31	777
45	30	1352
45	20	943
48	117	4009
49	135	4176
50	90	3773
52	63	2130
52	14	1048
52	3	353
53	58	1978
53	10	1352
53	1	58
54	128	2957
54	102	3198
55	163	4784
56	51	2101
57	91	2337
59	37	1079
60	47	1436
61	75	1196
62	40	775
63	49	1845
64	39	623
65	45	1021
66	32	3376
67	23	2179
68	35	3279
69	32	2842
70	110	2895
71	267	5582
72	6	141
73	6	141
74	71	2351
75	166	4327

	Spools	
No.	Total	DI
76	60	1163
77	21	1136
78	29	622
79	100	2113
80	52	1259
81	84	1174
82	74	2371
83	63	2372
84	22	403
85	26	1256
87	33	966
87	10	903
88	140	4353
89	147	3381
90	35	312
92	60	1111
93	35	1248
94	55	1382
95	46	1630
96	23	1190
97	24	758
98	13	654
99	53	1942
101	78	2095
102	141	3544
104	79	2321
105	62	1623
106	66	1674
107	79	2087
108	84	2857
109	63	1613
110	113	2398
111	60	1648
113	82	1819
113	147	3315
113	15	475
114	129	3306
115	139	2267
116	72	1365
117	109	3259
118	101	2076
119	56	3499
120	54	3647
121	290	3389
	8234	272858

Average Spools per Mod	68.05
Average DI's per Spool	33.14
Average DI's per Module	2,255.03
Average Welds per Spool	0.16
Average Welds per DI's	0.0048

Table 7-12: Project D fabricated spools and diameter inches per module

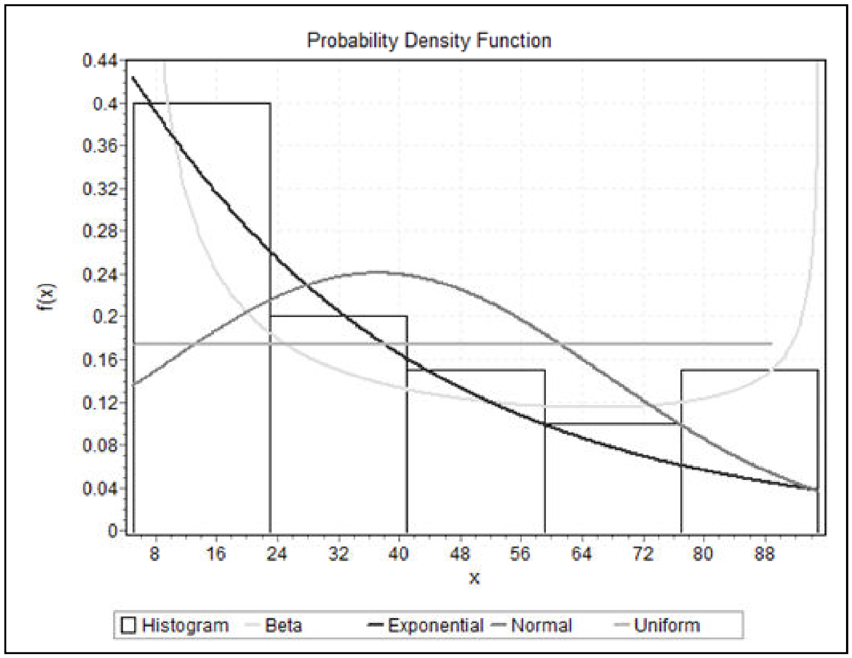

Figure 7-30: Project D spools Probability Density Function

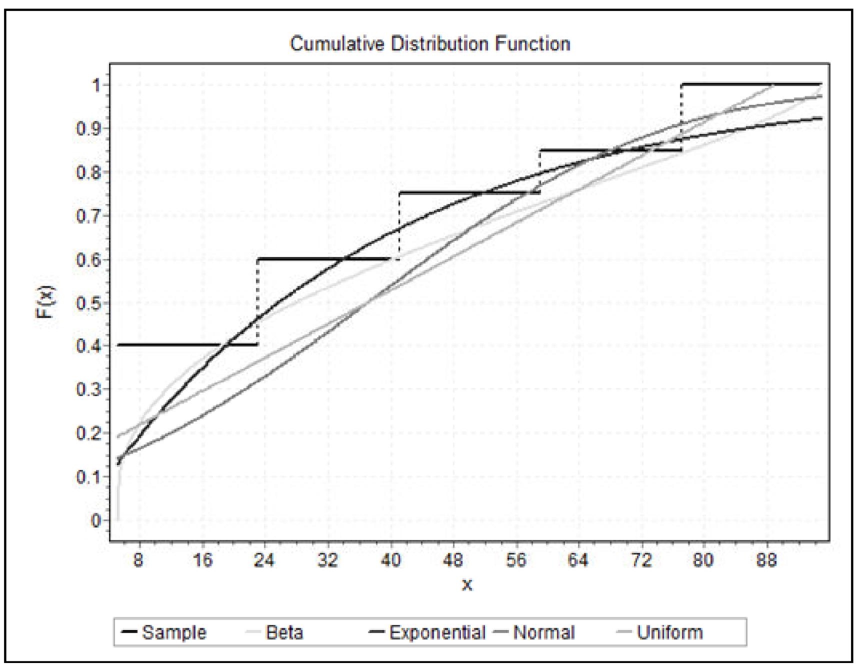

Figure 7-31: Project D spools Cumulative Distribution Function

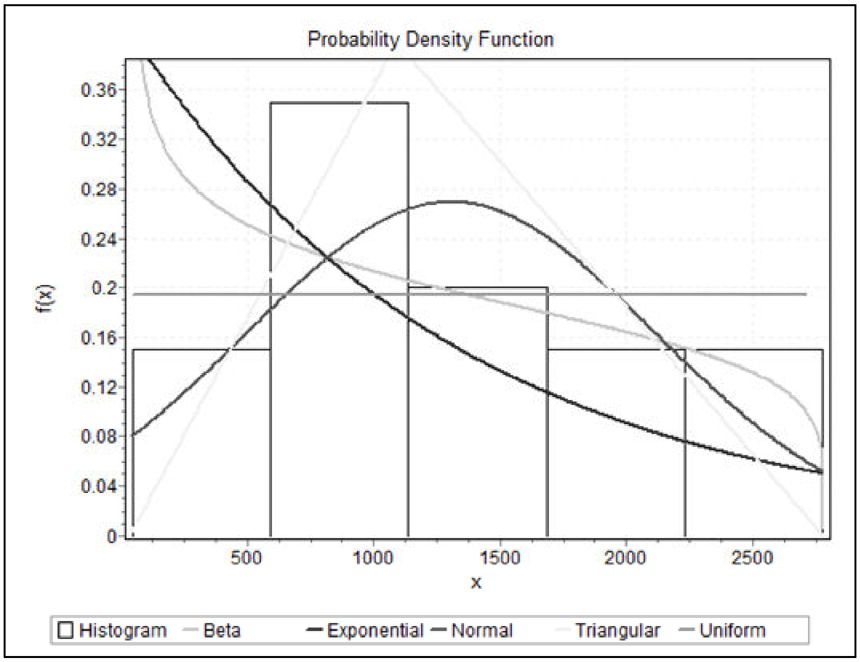

Figure 7-32: Project D DI's Probability Density Function

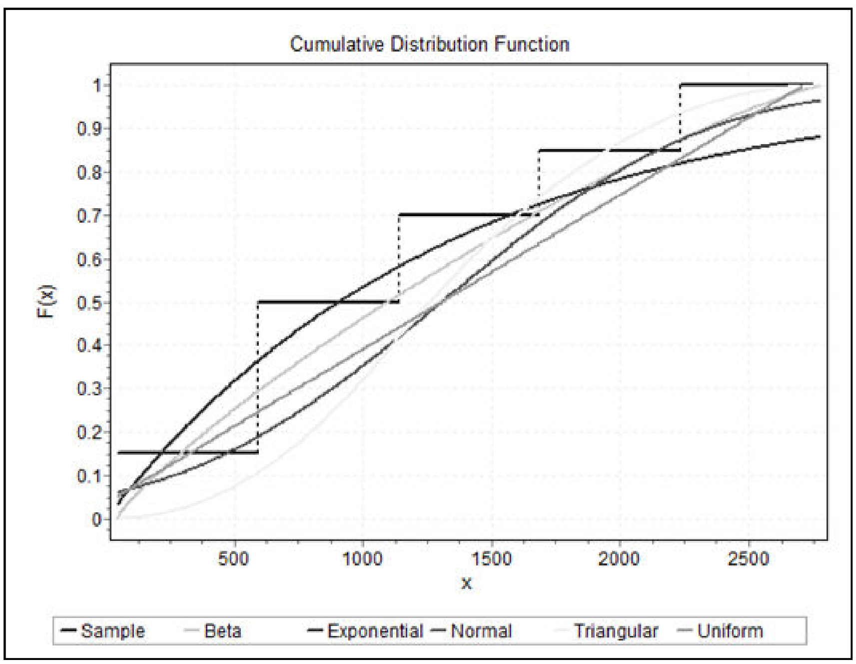

Figure 7-33: Project D DI's Cumulative Distribution Function

Project E fabrication data

	Spools	
No.	Total	DI
1	94	2601
1	20	850
2	38	1144
2	9	537
3	0	0
4	0	0
5	87	1685
5	8	340
6	37	795
6	1	81
7	0	0
8	0	0
9	0	0
9	0	0
10	49	1329
10	17	883
11	0	0
12	23	1298
13	55	1587
13	24	935
14	110	2486
14	30	1159
15	33	1201
16	21	585
17	68	2040
17	48	1971
18	39	1037
18	23	1048

	Spools	
No.	Total	DI
19	22	1025
20	24	653
21	77	1993
21	18	874
22	47	1302
22	29	1473
23	9	179
24	13	541
25	20	718
25	24	1040
26	20	710
26	31	1470
27	20	1010
27	56	2717
27	44	994
28	22	587
28	26	804
28	2	265
29	13	646
29	33	1278
30	28	988
30	29	845
30	14	844
31	22	1007
32	28	1254
33	40	574
34	62	960
35	56	1646
35	38	1347
35	7	384
36	6	213
36	9	414
36	5	321
37	10	378
37	6	270
37	5	324

Average Spools per Mod	47.27
Average DI's per Spool	32.95
Average DI's per Module	1,557.78
Average Welds per Spool	0.47
Average Welds per DI's	0.01

Table 7-13: Project E fabricated spools and diameter inches per module

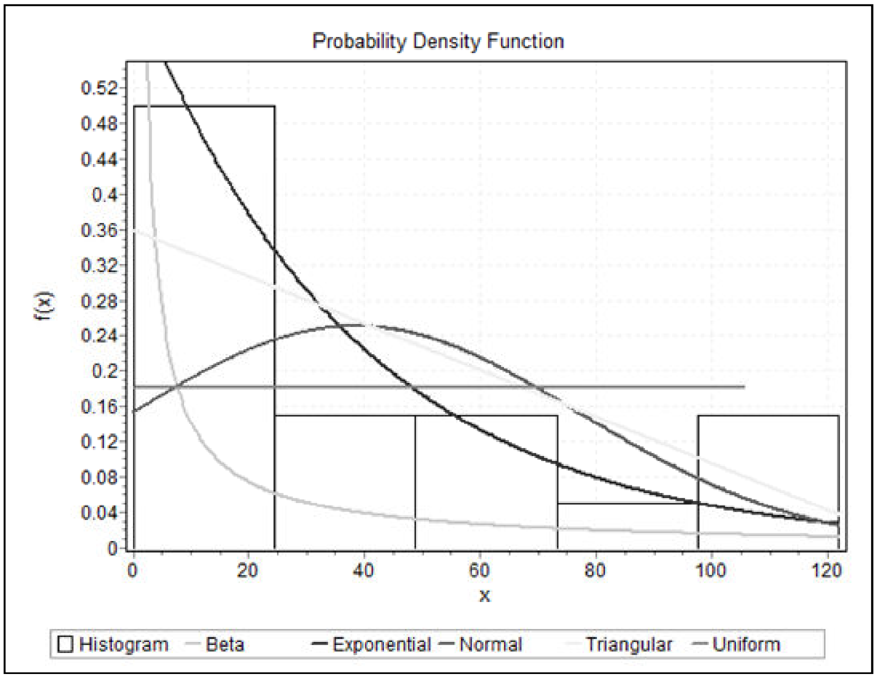

Figure 7-34: Project E spools Probability Density Function

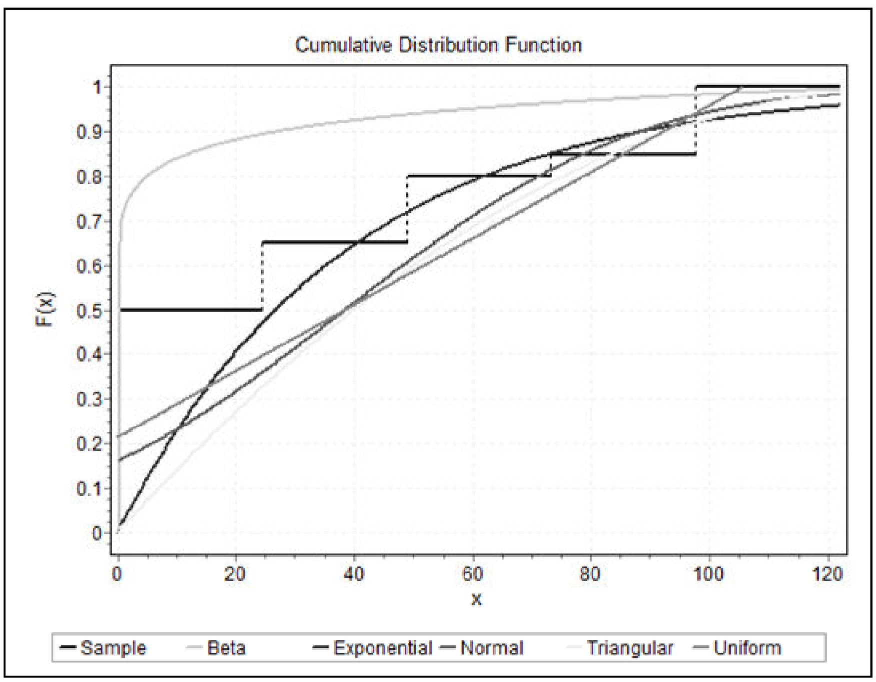

Figure 7-35: Project E spools Cumulative Distribution Function

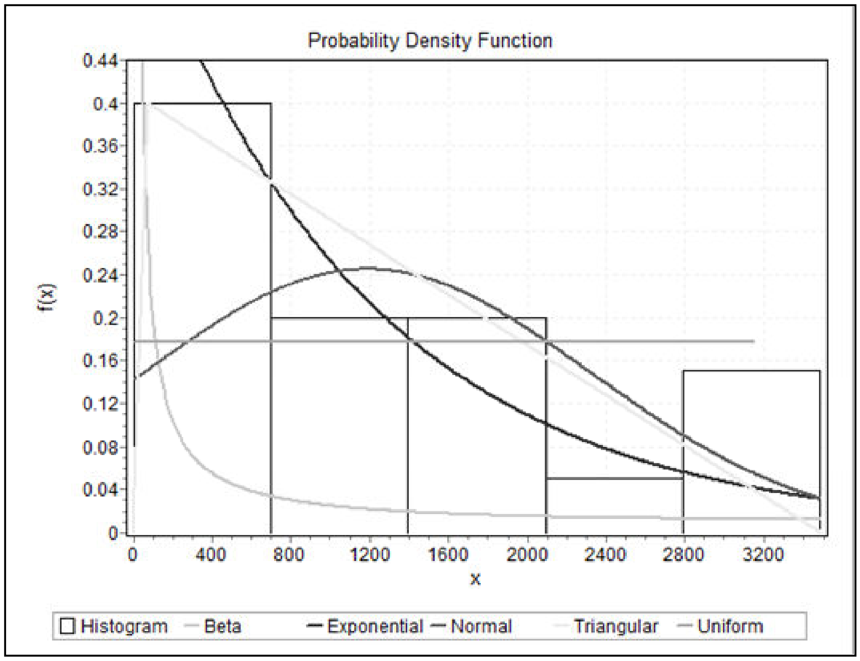

Figure 7-36: Project E DI's Probability Density Function

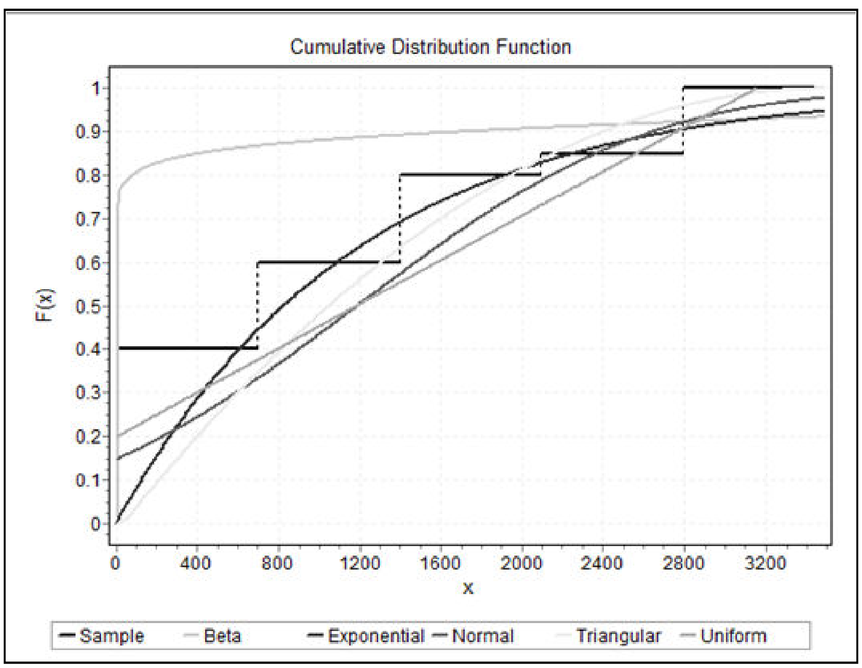

Figure 7-37: Project E DI's Cumulative Distribution Function