University of Alberta

Analyzing Scaffolding Needs for Industrial Construction Sites Using Historical Data

by

Lingzi Wu

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 Environment Engineering

Fall 2013

Edmonton, Alberta

To my beloved mother Zhenli Xu and father Jianlin Wu.

Abstract

Industrial construction covers a wide range of projects including petroleum refineries, chemical and power plants, which involve several disciplines such as civil, mechanical, and electrical. Different trades often depend on scaffolds to access their work areas. Quantification of scaffold requirements for large projects is difficult due to variability in work area heights and congestion, and availability of information. Metrics are generally based on a percentage of total direct trade man-hours.

This thesis presents research that aims at developing better understanding and estimates of scaffold needs for industrial construction projects, based on historical data from a mega industrial construction project over the course of two and a half years. This study seeks to discover hidden patterns and reliable correlations that may exist between required scaffold hours and other work attributes such as type of trade, height of scaffold, and other attributes that are relevant using data mining technique.

Acknowledgement

I would like to record my gratitude to a group of people who gave me encouragement, support and help during the time I am worked on my graduate degree at the University of Alberta in Canada.

First and foremost, I would like to give my most sincere thanks to Dr. Yasser Mohamed, my supervisor. Without his intellectual support, guidance, and encouragement, this work would not have been possible. I am also indebted to Dr. Aminah Robinson Fayek and Dr. Marwan El-Rich who reviewed my thesis and provided valuable comments in my thesis defense.

Also, particular thanks go to Mr. Rick Hermann, Mr. Hosein Taghaddos, Mr. Jamie Feuffel, Mr. Tyler Holdner, Ms. Brandy Holt and all the other employees at PCL Industrial Management Inc. who have participated in this research. I received sincere and warm help and support in all perspectives. Without their passion, no successful research could have been done within such a short period. Moreover, I wish to thank all the colleagues with whom I have had the pleasure to study with. Especially, I am truly indebted to Di Hu and Chandan Kumar who have given me valuable advice and kind encouragement.

I'd like also to thank all my friends who have been there for me when I needed them, especially Cristal Li and Berkley Laurin. My greatest thankfulness goes to Cristal, as my best friend, and my soul mate, for the endless understanding and support. I believe there is no words can express my love and gratitude to Berkley for loving me for who I am.

Finally, I want to take this opportunity to thank my family back in China; for my mother's infinite love and spiritual support for 25 years, for my father's respect of me as a unique individual, and constant tangible support without asking for anything in return.

Table of Contents

List of Tables

List of Figures

Chapter 1 – Introduction

1.1 Background

Industrial construction projects are usually large-scale, and involve gigantic structures. In addition, industrial construction projects often involve a wide range of different trades. Due to the resulting complicated nature, a single industrial construction project often heavily depends on several trades, like mechanical, electrical, pipe fitting, insulation, and/or iron work. However, there is one trade which receives little attention, though it plays an indispensable role in the whole project – that is the scaffold trade. On site, different trades focusing on their own work packages might have similar or completely different requirements for scaffolding to provide access to their working areas. From the project perspective, all these requirements from different trades contribute to a large scaffolding system.

According to OR-OSHA, scaffold refers to "any temporary elevated platform (supported or suspended) and its supporting structure (including points of anchorage), used for supporting employees or materials or both" (n.d.). Archaeological discoveries show even ancient Egypt, ancient China and Greece were using scaffolding-like structures to create access to high buildings. Illingworth (1987) summarized in his book that "scaffolding can truly be seen as the maid-of-all-work to the construction industry." Scaffolding is of the utmost importance; most permanent structures require its use, which means that the scaffolding crew interacts with most of the other trades on site. Nowadays, scaffolding is used in almost every construction project; no matter if it is residential, commercial, or industrial. Because of the complicated nature of industrial construction projects, the demand for scaffolding in these projects is more sophisticated.

Though indispensable, scaffolding systems are often taken for granted (Illingworth, 1987) and receive little attention; scaffolding systems are neither plotted on drawings, nor managed in a work package that could be included in the whole schedule. Current scaffolding practice within the industrial world is carried out in an ad-hoc way, which basically depends on the estimator's, planner's and scaffold foreman's experience, historical data, and the company's risk policy. Specifically, when it comes to estimation, scaffolding is treated as an indirect work, and is usually calculated as a percentage of the

total man-hours of direct work, which is hard to reflect accurately and specifically for each unique industrial construction project. Nevertheless, 30% to 40% of total direct man-hours is a large cost; if scaffolding is well estimated, scheduled, and managed, a significant improvement in productivity and reduction of project cost is feasible.

1.2 Research Motivation

Knowledge gathered based on previous research shows that there is need to devote more energy and effort into the temporary work of scaffolding. As more mathematical programming techniques and computer technologies have been applied in construction project management, cost has been reduced while efficiency, safety, and productivity have been increased markedly. However, most of the focus has been placed on permanent structures; very little effort has been given to temporary work.

Project scheduling, planning, and controlling have been treated as a major deficiency in construction management practices. One of the many key responsibilities of a project planner is to estimate and plan the possible temporary works needed on site. Temporary work plays a key role in construction projects regarding quality, safety, and productivity. Sometimes, temporary work can take up over 60% of the total project cost. Thus, decisions regarding temporary work are crucial to the success of a project (Proverbs, Holt, & Olomolaiye, 1998).

Hence, from a project planning and scheduling perspective, it is necessary and urgent for the project planner to properly estimate and plan scaffolding work.

1.3 Research Objectives and Methodology

This research is based on the analysis of a set of historical scaffolding data, provided by a typical industrial construction contractor (PCL Industrial Management Inc.), and aims at discovering the pattern of real scaffold requirements on mega industrial construction projects. The main objective of this research is to provide a scaffold estimate model, which is built on the previous scaffold data set. Using this estimate model, the manager and estimator can calculate the scaffold man-hours, not only for the whole project level, but also detailed down to each discipline on each construction area level. Figure 1 graphically shows the main objective of this research.

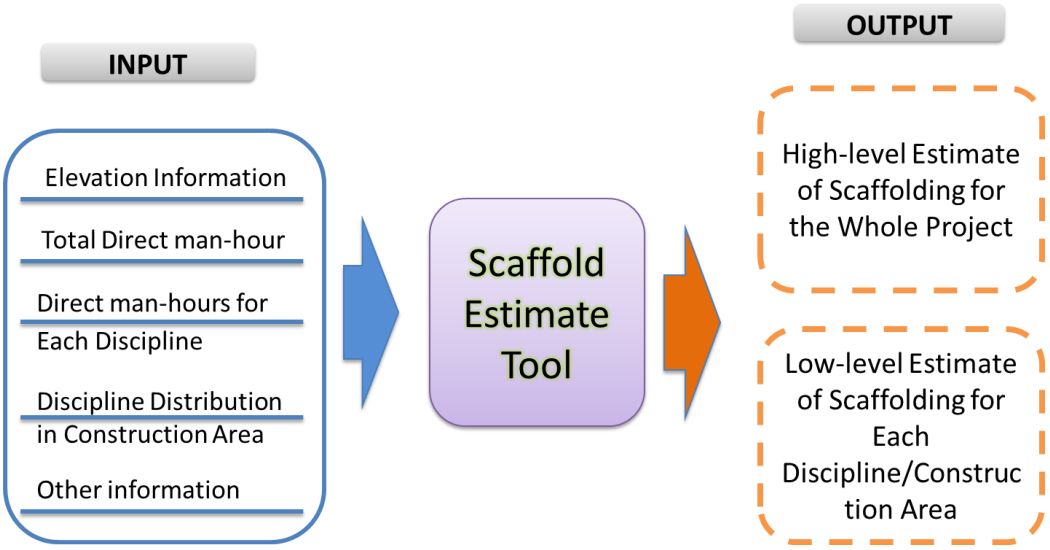


Figure 1 A graphical representation of scaffold estimate tool

The objective of this research is to build a scaffold estimate tool, which takes basic information, for example geometries from blue print and scaffold man-hours from schedule, to provide a more reliable and more detailed estimate for scaffold. The objective is achieved through several steps, listed below:

- Understand the basic knowledge of scaffolding and current practice of scaffold estimation and planning, then present the scaffold request process through a business model using IDEF0 diagram;
- Understand, examine, and prepare the scaffold data for data mining use;
- Design a set of data mining investigations, using trial and error method, based on the understanding of data, experts' advice, and existing data mining results;
- Evaluate the performance of each experiment from second phase data mining investigation based on three different methods of splitting data set, and comparing the result;
- Come to a conclusion and develop recommendations for future project use.

The research starts from gathering information from interviews of site experts and site visit. Based on the knowledge of the experts' experience, a hypothesis will be made.

Then data mining of the historical data will be involved as the major technique in this research to testify the hypothesis as well as providing a scaffold estimate tool. WEKA, an open source data mining tool is used to carry out the computer learning. An introduction of WEKA will be given in Chapter 2.

1.4 Thesis Organization

This thesis consists of six chapters. Chapter 1 provides the background, motivation, objectives, and methodology of this research.

Chapter 2 contains the literature review from three different angles. One describes the basic knowledge of scaffolding; the second part reviews the current research on estimation of temporary work; the third part explains the basic knowledge of computer learning and data mining tool – WEKA.

Chapter 3 explains the whole process of data preparation. Based on understanding of current scaffold estimation and planning practice at the industrial construction company, a business model reflects how the scaffold request process has been built. The most important and central component of this business model is the "scaffold request database," which is the main focus of this research. A preparation of the data from one historical scaffold request database is explained here to unveil the difficulties with data collection, cleaning, and input organization.

Chapter 4 reveals the computer learning process: the first phase is based on an initial input table of the data preparation, while the second phase is based on a modified input table which was adapted based on advice from the experts, and the third phase is based on the same input table as phase two, with a change in prediction class. Input table organization, and experimental results from all three phases of data mining investigation are presented in this chapter.

Chapter 5 provides evaluation of the model performance of the second phase of the data mining investigation, which is based on the expert modified input table.

Chapter 6 presents a conclusion of the research and proposes some recommendations for future research.

Chapter 2 – Literature Review

2.1 Introduction

This chapter gives a brief literature review to introduce basic concepts and knowledge related to this research. Section 2.2 contains the background of scaffolding; Section 2.3 briefly reviews current research regarding temporary work estimation, and Section 2.4 presents basic data mining knowledge.

2.2 Scaffolding in Industrial Projects

2.2.1 Basic Types of Scaffolds

According to Alberta Construction Safety Association, there are nine types of scaffolds which are commonly used in Canada (n.d.). They are standard tubular frame scaffold, standard walk-through or arch frame scaffolds, rolling scaffolds, fold-up scaffold frames, adjustable scaffolds, tube and clamp scaffolds, system scaffold components, mast-climbing work platforms and crank-up or tower scaffolds.

2.2.1.1 Standard Tubular Frame Scaffold

This type of scaffold is one of the most commonly used in construction projects. Standard tubular frame scaffold is usually fabricated in a variety of spans and configurations, which makes it easy to fit different site conditions. This kind of scaffold is easy for a scaffold crew to assemble, and simple for the trade workers to use. The most important strong point is its components can be lifted by the scaffold crew manually, which substantially increases productivity (Alberta Construction Safety Association, n.d.).

2.2.1.2 Standard Walk-through or Arch Frame Scaffolds

This kind scaffold is a variation of the first one. It is mainly used in masonry industry to meet their needs – providing larger height between each platform, as well as easier access of materials (Alberta Construction Safety Association (ASCA), n.d.).

2.2.1.3 Rolling Scaffolds

Rolling scaffolds are used when scaffolding needs to be moved around quite often. This type of scaffold is equipped with wheels. The advantage of this kind scaffold is it is flexible and cost efficient; once set up, it can be used in more than one location (ASCA, n.d.).

2.2.1.4 Folding Type Scaffold Frames

Trades like painters, electricians, and ceiling installers often need fold-up scaffolds. Similar to rolling scaffold, fold-up scaffold is easy to move and set up around the job site, or to travel from project to project. Nevertheless, its small dimension limits its usage (ASCA, n.d.).

2.2.1.5 Adjustable Scaffolds

Adjustable scaffolding is quite similar to fold-up scaffolding, though it takes some effort to erect. As the name implies, the height is adjustable; further, the whole system is relatively light and can be taken down into a limited number of components which are suitable for transporting (ASCA, n.d.).

2.2.1.6 Tube and Clamp Scaffolds

Tube and clamp scaffolds are often used for irregular forms. The advantage is their infinite adjustable ability in height and width. Generally, tube and clamp scaffolding has more flexibility, but is more complex and time-consuming to build than the other types (ASCA, n.d.).

2.2.1.7 System Scaffolds

Although system scaffolds are not as flexible as tube and clamp scaffolds, they are becoming increasingly popular on construction sites. They can be adjusted to a wide range of irregular shapes, for example, circular, dome, and non-rectangular structures (Infrastructure Health & Safety Association, n.d.).

2.2.1.8 Mast-climbing Work Platforms

Mast-climbing work platforms are most popular among the masonry industry. The advantage of this type of scaffold is the entire platform can be fixed to the exact height required, which satisfies the human physiological character and enhances the safety (Infrastructure Health & Safety Association (IHSA), n.d.).

2.2.1.9 Crank-up or Tower Scaffolds

Crank-up/tower scaffolds are used in some Canadian masonry projects, though they are more popular in America (IHSA, n.d.).

2.2.2 Scaffold Accessories

There are four major types of scaffold accessories, which cooperate with the major types of scaffolds to provide access for construction. They are sidewall brackets, platform components, ladders and stair section access, and guardrails (IHSA, n.d.).

2.2.2.1 Sidewall/Outrigger Brackets

Sidewall or outrigger brackets are most often used in masonry projects. They help to provide a fixable working platform at a convenient height (IHSA, n.d.).

2.2.2.2 Platform

Aluminum/plywood platform is the main part used to build a work deck. Usually, they vary in size, weight, strength and species. The load carrying capacities are one of the main indexes, which depend on the material and size of the platform, the span, and the location of the load, where regular inspection is required to ensure safety (IHSA, n.d.).

2.2.2.3 Letters and Stair Section Access

Letters and stairs are important methods to access platforms. Some of the letters and stairs are built into the platform system, while others are attached as a separate component (ASCA, n.d.).

2.2.2.4 Guardrails

Guardrails are a vital component to keep trade men safe. According to Alberta Construction Safety Association (n.d.), one of the major causes of trade men falling from work platforms is failure to install guardrails. Guardrail components are usually easy to attach to the scaffold platforms.

2.2.3 Scaffold Safety Issue and Design

Most of the existing research relative to scaffolding issues has focused on scaffold safety issues, scaffold structure analysis, and scaffold design. For example, Son and Park (2010) investigated steel pipe scaffolding in Korea. They designed specific tests based on different variables by checking the torque of the clamps being surveyed and the criteria specified from standard. Their results proposed discovery and recommendation for future use focused on marginal load of clamps. Peng et al. (1996) presented a simplified analysis system for high clearance scaffolds, which simplified the calculation of the critical loads in practical design. Due to limited access of relevant requirements for suspended scaffold structural design, as they are separately documented in more than one regulation, Hill et al. (2010) organized and provided information of key OSHA structural provisions considering suspended scaffold support elements design.

2.2.4 Scaffold Management

Ideally, good scaffold management could reduce scaffolding costs from around 25% of the total direct man-hours to about 15%. This big improvement could be realized through

data management. If all the direct work is effectively scheduled, planned, and collected in a central database, then it is possible to identify the scaffold needs and even erect scaffolding before a trade foreman requests it. A scaffold database is needed for this process to track information like who built the scaffolding, when it went up, components and materials used, and how long it lasted. Thus, the tracking of scaffold labor-hours, scaffold material rental cost, and management of scaffold yard materials would become an entire system (Ryan, 2009).

2.3 Temporary Work Estimate

To present temporary work estimation, basic knowledge of construction project estimates (Section 2.3.1), as well as indirect cost estimates (Section 2.3.2) must be introduced first.

2.3.1 Construction Project Estimates

Estimation is the procedure to provide a statement of the approximate cost, time or quantity of material needed to carry out a project. Estimation is usually related to decision making processes, for example bidding price, project cost, and project controlling and management policy (Carr, 1989). Estimation is a crucial process for the construction management team (Adeli, & Wu, 1998).

A large number of estimate tools and methods have been developed. These techniques range from simple floor area method to advanced intelligent systems. Recently, researchers have been focusing on the more sophisticated means, which tend to analyze a large amount of data using advanced computer technology and programming techniques (Kim, Seo, & Hyun, 2012). For example, Gunaydin & Dogan (2004) suggested a model trained by neural network methodology for cost estimating in early design stages.

At different project stages, information availability varies considerably, which directly affects the accuracy of a construction cost estimate. Usually, at the early stages of a project, the cost estimate is a ball-park figure. Before the beginning of construction, a fairly decent figure will be ready for budget control (Hendrickson & Au, 2008). Although the success of a project is largely affected by the accuracy of the construction cost estimate (Kim, An, & Kang, 2004), it is difficult and complicated to determine an accurate number, due to limitation of information in the early stages. Cost estimation can be classified into different categories based on the stages of project procession.

Generally, cost estimates could be grouped into preliminary cost estimates and the final cost estimate. The preliminary estimate is made during the project planning and design stage, when the project is not completely defined. Thus, preliminary cost estimates are approximate, and often called conceptual estimates. Basically, all the methods used for preliminary cost estimate are established on some measurement of gross unit costs generated from historical work data. The final cost estimate is proposed after the design is done, when drawings and specifications are prepared. The final cost estimate is more detailed than the preliminary cost estimate. The final cost estimate is carried out according to a complete and accurate work quantity takeoff. Nevertheless, the final construction estimate is not able to be exactly accurate due to many reasons, for example lack of standardization, uniqueness of each project, or different market conditions (Clough, Glenn, & Sears, 2000).

Beyond that, construction cost estimates can be classified into three major types – design estimate, bid estimate, and control estimate – according to the functions and objectives. Design estimate is targeted to owners or their assigned designing team. Usually, design estimate is carried on parallel with the planning and design process. During this period, based on different levels of information, design estimates can be sub-classified into screen estimate (order of magnitude estimate), preliminary estimate (or conceptual estimate), detailed estimate (or definitive estimate), and engineer's estimate. Bid estimates are for contractors which aim at bidding or negotiation. Control estimates are used for monitoring and controlling the project purpose, which applies to both owners and contractors. For contractors, usually after the award of contract, the bid estimate becomes their budge estimate, which will be treated as a controlling figure for the construction period. However, periodical update is necessary to reflect reality, show benefit or loss, and serve a better controlling role (Hendrickson & Au, 2008).

2.3.2 Indirect Cost Estimates

Besides the classification of estimates according to the chronological order of a project, generally estimates can be classified into two groups based on the nature of costs they deal with. These are direct cost and indirect cost. Simply, direct work can be traced down to a specific item which links to a cost code, while indirect cost is not traceable, and usually hard to assign to a cost system (Tah, Thorpe, & McCaffer, 1994). Usually, direct costs include labor, materials, supplies, equipment, and any expenses related to the final product. The indirect costs comprise two parts; one related to the costs that would happen

even if no specific activities had been carried out, for example overheads, profits, and contingence allowances; the other part is costs too trifling to allocate to a work break down system economically, for example temporary work costs (Carr, 1989). Indirect costs must not be neglected; surprisingly they might account for 35 to 55 percent of the total project cost (Clough et al., 2000).

Current practice of indirect work estimation for construction projects is highly subjective and heavily dependent on experience. Advanced quantitative methods have been developed, but are rarely used, for various reasons. Although efforts have been spent into the improvement of indirect cost estimates, its sensitive and confidential nature becomes a real obstacle to the research focused in this area. Tah et al.'s paper, suggests that future development of indirect cost estimation should adopt computer technologies to create a simple and straightforward method, and at the same time, to embrace the subjective nature of the activity (Tah et al., 1994).

2.3.3 Scaffold Estimates

Scaffolding is a typical type of temporary work. Estimates for temporary work are usually considered as part of indirect work estimates. Although usually treated as indirect work estimates, temporary work estimates have their own characteristics, which are different from other indirect costs, for instance, overheads, and profits.

Temporary work, for instance scaffolds, is only used during the construction period to support construction of the final product, so it is often torn down after it finishes its mission, or at the end of the project. In addition, very rarely is temporary work like scaffolding designed and planned into drawings. Because of these natures of temporary work, it is very hard to allocate temporary work into a schedule or cost system to track data during the project. Thus, most often, temporary work is treated as indirect work, though the actual costs of temporary work consist of labor, materials, equipment and all the other physical costs. It is not considered economical to trace temporary work costs (Carr, 1989). However, temporary work is a crucial factor to the success of a project (Proverbs, et al., 1998). Sometimes temporary work may take up to 60 percent of the total contract sum (Illingworth, 1987).

In the practical world, different companies have their own methods to calculate scaffold costs. However, it is recognized that scaffold estimates heavily depend on the estimator's judgments. A comparison between the coming project and the previous projects,

examination of historical data, and coming up with a bulky ratio of scaffold man-hours over direct man-hours is the most commonly used strategy.

2.4 Data Mining

2.4.1 Background

The prevalence of computer science and communication technology has produced a world dependent on information. Over recent years, the capability of creating and storing data rapidly increased (Chen, Han, & Yu, 1996). People living in modern society are overwhelmed with data, which seems to be ever-increasing. However, the majority of the information existing is in its raw form, which only qualifies as data. (Witten, Frank, & Hall, 2011)

Simply storing the data does not necessarily mean we understand it, or that we can take advantage of the useful information hidden inside; Witten et al. (2011) addressed this in their book:

> "We could all testify to the growing gap between the generation of data and our understanding of it. As the volume of data increases, inexorably, the proportion of it that people understand decreases alarmingly."

One of the reasons why we keep data is to solve problems by analyzing the data and discovering useful information or patterns within it. This piece of useful information or pattern can be used to make predictions in future cases. On one hand, data is under an exponential increase, but on the other hand, more advanced and sophisticated computer technologies are being introduced. Data mining is one of the methods to discover useful information within data. However, it can be dangerous when humans cannot understand data, but fully trust the results that come from a computer. However, people often blindly believe in the precise number and pattern, while ignore the true meaning of the data (Witten et al., 2011) Understanding the basic concepts of data mining is very important.

2.4.2 Data, Information, and Knowledge

Data can be defined as recorded facts, while information is what is gained from these recorded facts, or simply patterns found within the data, which can be learned and used in cases where unknowns are present. The definition of knowledge is the accumulation of information and the wisdom found in the process (Witten et al., 2011).

Technology allows people to collect and store data freely in large amounts. At the same time, transforming the vast amount of data into information, and finally converting it to knowledge is a universal challenge. Fortunately, advanced computer science makes possible for people to seek hidden patterns – locked up information – underneath the raw data (Witten et al., 2011).

2.4.3 Data Mining

> "Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic one" (Witten et al., 2011).

In other words, data mining is a technology aimed at bringing forth the hidden information, and articulating it in a structural way so people can explain the data, learn the useful information, and predict in the future cases (Witten et al., 2011).

2.4.4 Input

Input can be expressed in three elements – concepts, instances, and attributes. Concept is the hidden information to be learned; instance, sometimes called example, is an independent specific representative of the concept; attribute is a predefined feature, to characterize each instance. Usually the input ready for a data mining investigation is structured into a set of instances, with each instance containing a group of values with fixed, pre-set attributes. For example, if the instances are the rows of the input table, then the attributes are the columns (Witten et al., 2011).

Input is a key element of any data mining investigation experiment. How well an input is organized and prepared plays a crucial role to the success of the data mining results. Usually preparing input is time-consuming and labour-intensive work. Usually substantial effort has been devoted to the preparing input phase before a data mining investigation starts. Data preparing process starts by gathering data together into a set of instances, then building some simple histograms to show the distribution of values of nominal attributes, or charts and graphs for numeric values, which will be helpful to understand the data. One of the advantages of graphical visualization of data is it is an easy way to identify outliers. In most cases, outliers are errors or unusual situations. In any case, domain experts are needed to explain the data. Success of preparing input is the first step on the road to the success of the data mining (Pyle, 1999).

2.4.5 Algorithm

In this section, two basic algorithms are introduced: linear regression and M5 method. These are two basic computer learning methods, which were used in this research.

2.4.5.1 Linear Regression

In this research, one of the basic objectives is to train a model to be used in future projects to provide a better and more accurate estimate on scaffold activities (scaffold man-hours per trade man-hour). This problem is a typical numeric prediction problem in WEKA. When class is numerical, and most of the attributes are numerical as well, it is natural to try linear regression technique first (Witten et al., 2011). The main concept of linear model is to express the output of the model by combining all the attributes with certain weight, respectively. Giving y as the class to be predicted, $x_1, x_2, \ldots x_n$ are n attributes, the linear regression model can be written as:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \ ,$$

where $a_1, a_2, \ldots a_n$ are weights for each attribute.

In addition, linear regression can be used to deal with not only numeric attributes, but also nominal attributes. The basic concept is to treat every possible value of one nominal attribute as a binary case, if the attribute of this training instance belongs to this value, it will treat the value of this attribute as one, and otherwise its will be zero (Witten et al., 2011).

Linear regression is concise and frank, and according to Witten et al. (2011) "one of the most instructive lessons is that simple ideas often work very well. And we strongly recommend the adoption of a 'simplicity-first' methodology when analyzing practical datasets."

2.4.5.2 Attributes Selection

In WEKA, algorithm linear regression comes with two basic computer attribute selection methods. The performance of a selected attribute subset is measured by the classification performance of the model built upon this specific selection of attributes (Witten et al., 2011).

Most attribute selection methods use the concept of searching the space of attributes to find a best attribute subset, which builds the best performing model. Usually, there are

two directions which can be followed to search the spaces of attributes: one is from top to bottom – forward selection; the other is from the bottom up – backward elimination. If spaces are searched from top to bottom, it starts from an empty subset. As it goes towards the bottom, at each level, one attribute is added to the current attribute subset (like shown in Figure 2). Likewise, the backward elimination starts from the bottom with a full set of attributes; then as it moves upwards, at each stage, one attribute is moved out of the existing attributes subset (Witten et al., 2011). The figure below explains the space between attributes.

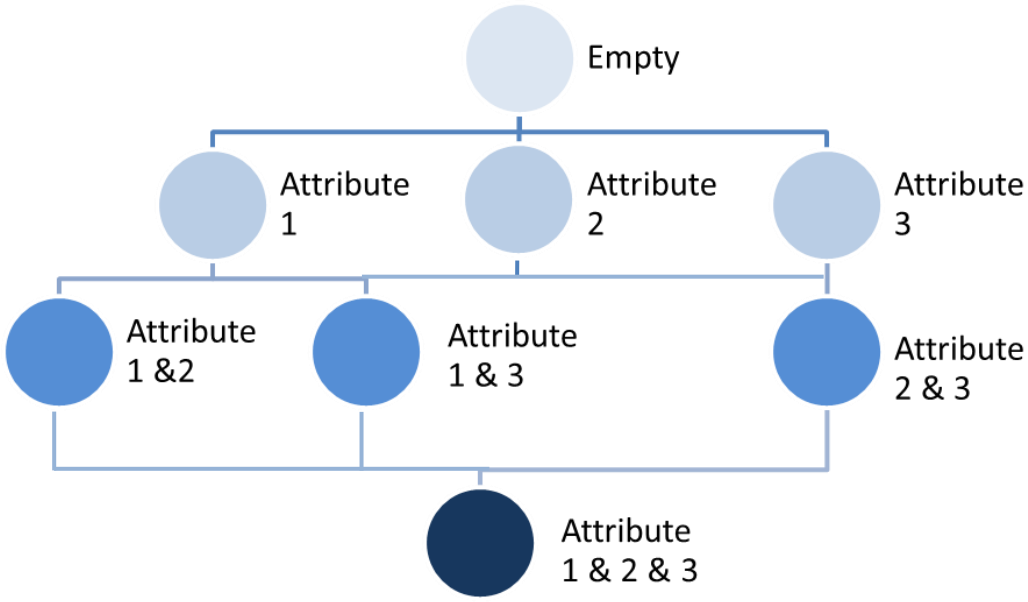


Figure 2Attribute space explanations

For each attribute subset, the quality is measured using information theory to judge the quality of a network. Two most common methods are Akaike information criterion (AIC) (Bozdogan, 1987) and minimum description length (MDL) (Stine, 2003).

Another method of selecting attributes is using the concept of pruning the decision tree. The theory of a decision tree is when a test instance is given, it comes down from the top of the tree, based on the value of the instance's attributes, and decisions are made at each node. For each interior node, in this tree, a linear model is built. This linear model could be expressed as below:

$$w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_n a_n$$

where $a_1, a_2, \ldots, a_n$ are attribute values, and $w_0, w_1, \ldots, w_n$ are weights for each attribute, respectively. On different notes, different subsets of attributes are used.

The pruning procedure is based on an error estimate. At each node, the absolute difference between actual value and the predicted value is calculated when each of the training instances reaches that node. For each node, the average of this absolute difference is calculated as an error estimate indicator. However, this error indicator is from the same dataset as the tree is trained, which makes it more optimistic for the unseen data. Thus, a factor $\frac{(n+v)}{(n-v)}$ is multiplied to correct the bias. Here, n is the number of the training instances that reach the node; v is the number of parameters in the linear model that gives the class value at the node. When considering the error estimate indicator, it consists of two parts, one is average of the absolute difference, the other is the compensation factor $\frac{(n+v)}{(n-v)}$. One way to minimize the estimate indicator, as well as simplify the tree, is to decrease the parameters or attributes used on one node. By dropping one attribute, the compensation factor is decreasing, while the average error is possibly increasing. Thus, attributes are eliminated one by one greedily, and the estimate indicator is calculated every time, until the increase of the average error out-weighs the decrease of the multiplication factor – the error estimate increases (Witten et al., 2011).

2.4.5.3 M5P trees

The classifier -- M5P tree in WEKA is based on M5 methods, which implement base routines for generating M5 Model trees and rules.

The classic environment where decision tree and decision rules are developed is when both the class value as well as the attribute values are discrete. However over past decades, developments have been made to extend decision trees and decision rules to handle continuous numeric attribute values and class value (Wang & Witten, 1996).

M5 method works in situations where both attribute values and class values are continuous numerical. This algorithm provides tree-based models. It has the tree structure of ordinary decision trees, however, unlike the traditional decision trees, at each leaf, a linear regression model is built to provide a prediction for the instances reaching the leaf. Thus, M5 can provide multivariate linear models, similar to piecewise linear functions (Quinlan, 1992). M5's flexibility and capacity to handle continuous numerical values make it a more powerful tool for real cases than regular linear regression.

The process of constructing a tree with M5 method can be expressed recursively. Instead of trying to maximize the information gain at each node like an ordinary decision tree, M5 trees use a splitting criterion that minimizes the intra-subset variation of the class value downward each branch. The splitting criterion can be simply called "standard deviation reduction (SDR)" (Wang & Witten, 1996). Specifically, assuming we have a set training data containing T instances, and each instance was defined by a set of unique and fixed attributes with a fixed class value; the target is to build a tree model to relate the class values of these instances to their attributes values. This T set of training data either belongs to one leaf or certain tests are conducted to split it into several subsets. Let T_i donate the ith subset out of one potential test. The reduction of the standard deviation (SDR) could be expressed as:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

Thus, after exhausting all the possible potential tests, one test with the maximization of SDR will be chosen to split the T set. This process repeats for each subset of T. The result of this process often leads to an over-elaborate tree, which needs to be pruned backward (Quinlan, 1992).

The pruning process depends on an error estimate of the accuracy of each node. This error estimate is expressed as a factor multiplying the average of the absolute difference between actual class values and predicted class values. The factor is used to underestimate the effect of unseen data. This factor can be expressed as below:

$$Factor = \frac{(n+v)}{(n-v)}$$

where, n is the number of training set, v is the number of parameters in this linear model at this node (Wang & Witten, 1996). The pruning process starts from the bottom of a tree, and each non-leaf node is examined. M5 method selects either the sub-tree structure or a simplified linear model with the lowest error estimate. For the latter, the original sub-tree will be replaced by this linear model, which means this node is pruned to a leaf.

2.4.6 Evaluating the Performance

According to Witten et al., "Evaluating is the key to making real progress in data mining" (2011), however, determining how to measure the performance, and how to compare one

algorithm with another on a specific case, is a challenge. Systematic methods are needed to help explain the results, and evaluate the performance of each experiment.

2.4.6.1 Repeated Cross-Validation

To evaluate a model's performance, error rate must be introduced. "The error rate is just the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier." (Witten et al., 2011). However, if the error rate is calculated based on the training data, which is also called re-substitution error, it is not necessarily a good indicator to the new data set, because the model was evaluated by the exact same data set that was used to train the mode. Usually, this rate is optimistic, sometimes even "hopelessly optimistic" (Witten et al., 2011).

To truly reflect the performance of a model, an independent data set which played no part of training the model, called the test set, is needed. Imaging both the train set and the test set is representative of the problem, then the error rate from test will be a good index to future data. A general method called holdout procedure is used to separate data into two sets, one for training, and the other for testing. As the term suggests, this entails holding out a certain amount of data as the test data set, and only using the remaining data set to train the model. Normally, the larger the training data set, the better the model. Likewise, this principle applies for test data set (Witten et al., 2011).

A problem occurs when considering dividing the whole dataset into two partitions. That is, it is hard to tell or justify whether the training or testing data is representative or not. On an extreme case, if all instances of a certain class were excluded by the training set, it is impracticable to expect the model generated from this training set to give a good performance on instances of that class. Thus, a random sampling, which makes sure every class is proportionally represented in both data sets, is needed. This procedure is called stratification. A universal way to reduce the bias caused by random sampling is repetition (Witten et al., 2011).

All the methods mentioned above are easy to practice, giving a huge amount of data available. However, when it comes to limited data, there is a dilemma of how to divide the data into two parts to obtain as much data as possible for training a good model, and at the same time reserving as much data as possible for obtaining a good error estimate. Especially in the real world, data is quite often limited (Witten et al., 2011).

According to Witten et al. (2011), "The question of predicting performance based on limited data is an interesting, and still controversial, one." Different techniques can be used to tackle this, among which, repeated cross-validation – "is probably the method of choice in most practical limited-data situations" (Witten et al., 2011).

The attempt of trying to maximize every single instance's inclusion in both training the model, as well as evaluating the performance of the model, leads to the feasible idea of switching the two data sets – that is, to train the model using testing data while reserving the training data for testing. After that, average the error rate to present the result, thus maximizing the usage of data (Witten et al., 2011).

Cross-validation method is a simple variation to apply both holdout method as well as switching roles of the testing and training data. For example, when using six-fold cross-validation, the entire data was split into six folds (partitions); each fold – one sixth of the data, in turn, is held as the test set and the remaining data is used to train the model. Thus, during the six experiments, every instance in the data set has been used to test just once. Usually, the standard way to cross-validate is to split the whole data set into ten folds, which is called 10-fold cross-validation. According to substantial experiments based on a variety of learning algorithms and different databases, ten is most likely the right number to get the best error index. Similarly, in order to obtain a reliable error estimate, the process is repeated. This process – 10-fold cross validation – is usually repeated ten times to get more accurate error estimation. Thus, this whole procedure carries a hundred times (Witten et al., 2011).

2.4.6.2 Measure the Performance of Numerical Prediction

To measure the performance among different data mining experiments on a given problem is difficult, since statistical testing is necessary to prove that apparent differences are not played by sudden factors. When it comes to numeric prediction situations, which is the case in this research, some evaluation measures have to be involved to measure the performance of each machine learning scheme (Witten et al., 2011).

Assume the predicted values on the test instances are $p_1, p_2, \dots, p_i, \dots, p_n$; while the actual values are $a_1, a_1, \dots, a_i, \dots, a_n$, then we have several ways to measure the performance, as listed below:

- Mean-squared error $\frac{(p_1-a_1)^2+\dots+(p_n-a_n)^2}{n}$

- Root mean-squared error $\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
- Mean-absolute error $\frac{\sum_1^n|p_i-a_i|}{n}$
- Relative-squared error $\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$
- Root relative-squared error $\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
- Relative-absolute error $\frac{\sum_1^n|p_i-a_i|}{\sum_1^n|a_i-\bar{a}|}$
- Correlation coefficient $\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i(p_i-\bar{p})(a_i-\bar{\bar{a}})}{n-1}$,

$$S_P=\frac{\sum_i(p_i-\bar{p})^2}{n-1}, S_A=\frac{\sum_i(a_i-\bar{\bar{a}})^2}{n-1}$$

Here, $\bar{a}$ is the mean value of the training data; $\bar{\bar{a}}$ is the mean value of the testing data (Witten et al., 2011).

Mean-squared error is the most common measurement for, many mathematical techniques, linear regression for instance. It behaves well, and it is easy to manipulate mathematically. Root mean-squared error is the square root of the mean-squared error, which provides the error measurement as the same dimension of the predicted case (Witten et al., 2011).

Mean-absolute error calculates the average of the absolute size of each individual error. This measurement shows the same dimension as the real value, but the weakness of mean-absolute error is it overstates the influence of outliers (Witten et al., 2011).

Other times, relative comparison matters over the absolute errors. For example, given a set of data ranging from 0.1 to 1000, where 10% of error is allowed, when the predicted value for 0.1 is 0.11 or the predicted value for 1000 is 900; whether this model is acceptable is not clear if looking at the mean-squared error or mean-absolute error, it is depends on the relative error. Thus, corresponding to mean-square error, root mean-square error and mean absolute error, relative-square error, root relative-square error and relative absolute error are introduced. Here, an assumption has been made for all relative error measurements, which is to treat average value ($\bar{\bar{a}}$) as the default prediction (Witten et al., 2011).

Relative square error provides a comparison between this model and the assumed predictor, which is the average value of the actual training data. It calculates the squared error and divides by the total squared error of the average value. Root relative square error is the result of square root of relative square error, which gives the same dimension as the data set. Likewise, relative absolute error is to compare the total absolute error from the model with the error of the assumed predictor. Unlike the first three evaluation equations, relative errors show the comparison between the model predictor and the average of the set of data. Specifically, if the relative error is bigger than 100%, it means the model predictor performs worse than average; if the relative error is smaller than 100%, it indicates the model performs better than average (Witten et al., 2011).

The last measure in the list is called correlation coefficient. This measure reveals the statistical correlation between the predicted values on the test instances $p_1, p_2, \dots, p_n$; and the actual values $a_1, a_1, \dots, a_n$. This measurement ranges from minus one to one. At both ends, one and minus one, it means two sets of values are perfectly correlated, except minus means they are negatively correlated, which should not happen for the reasonable numeric prediction scheme. The closer correlation coefficient is towards the middle point – zero, which means no correlation at all. The advantage of this measurement is it is scale independent. For example, if applying a constant factor to the values of prediction set, while leaving actual values unchanged, the correlation coefficient stays unchanged. However, this is not true for any other performance measures (Witten et al., 2011).

2.4.7 WEKA

WEKA is data mining software, developed at the University of Waikato in New Zealand. Its name stands for "Waikato Environment for knowledge Analysis." An affluent selection of state-of-art data mining algorithms and data processing tools is included in WEKA. It not only widely supports all kinds of machine learning processes, but also the preparation phase of input data, and measuring output statistically (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

2.4.8 Data Mining Practices in Construction Industry

Data mining techniques can be categorized differently according to the type of database to be learned, the type of knowledge to be discovered, and the type of techniques to be used. Most commonly, data mining techniques are classified by the type of knowledge to be learned, because this gives a clear description of data mining techniques and

requirements. Following this classification, data mining techniques can be categorized into: mining association rules, data generalization and summarization tools, perform classification, pattern-based similarity search, mining path traversal patterns and data clustering (Chen et al., 1996).

In the AEC (architecture, engineering and construction) industry, it has been acknowledged that effective exploration and utilization of historical project knowledge could improve business performance (Kamara et al., 2002). Data mining as a tool to seek information and knowledge from large database has been adopted by many researches as a major topic to help industrial companies to improve productivity or increase revenue. Researchers and developers in different fields have created various applications of data mining techniques to provide a better service to the customers (Chen et al., 1996).

Within the last several decades, construction companies have been under ever-increasing pressure to provide better-quality service, on time and within budget. Many of the responses to this pressure are supported by information technology. Usage of state-of-art information technology has contributed to improvement and development within the construction domain. In Rezgui's (2001) paper, a wide range of approaches and applications relative to the construction industry have been described. They include electronic document management system, product data technology, groupware technology, advanced decision support system, and data-warehousing techniques (Rezgui, 2001). Also, Piatetsky-Shapiro et al. (1996) reviewed the issues in developing data mining and knowledge discovery applications for industrial fields.

Data mining technique has been used to help industrial construction enterprises improve their bidding strategies. For example, in Gonzalez-Villalobos's (2011) research, a data warehouse was established to collect historical data from the client; then different methods of data mining, including clustering, distribution fitting, and association rules were used to help understand trends and arrangements of pipe module fabrication. Thus, a better understanding of pipe fabrication activities was achieved, and the indicator generated from the research was used as a decision support tool by the company in their bidding processes.

Construction equipment management is another area that has been improved by data mining technique. For instance, in Fan's (2007) research, advanced computer tools were used, based on nine-year-equipment-operation data, to improve the M-Track system,

which is the equipment information management system developed by NSERC/Alberta Construction Industry Research Chair. His research addressed the issue of design and implementation of information infrastructure among different companies for data sharing information retrieval, as well as knowledge discovery in construction equipment management. Utilization of data mining techniques helped discover information for the decision making process.

Data mining technique was also used to improve labour resources management for industrial construction projects. In Hammad's (2009) research, a framework was established to centrally store data, and provide dynamic reports. Additionally, data mining techniques were used to discover useful information from the real project data, which helped improve labour resources management practices.

Clearly, useful information from previous construction projects can offer a more objective, rather than subjective, decision (Moon, Kim, & Kwon 2007). Unfortunately, no specific examples of research were found which focused on generating information from historical scaffolding databases to better serve scaffolding needs, scaffolding controlling or scaffolding estimates.

2.5 Summary

This chapter provides a brief literature review to the basic knowledge of scaffold, temporary work estimation, as well as data mining information. However, no any previous research was found from any sources related to scaffold estimation.

Chapter 3 – Data Preparation

Data preparation contains a wide range of activities, including understanding the basic concept, and background of the research area, the practical operation of the client company and the database provided, which will be described in Section 3.1. Additionally, the introduction of the scaffold request database as well as data exploration using Excel will be covered in Section 3.2, followed by a brief summary in Section 3.3

3.1 Scaffold Practical Operations in Client Company

Based on the knowledge from Chapter 1 – Introduction and Chapter 2 – Literature Review, the scaffolding trade is unique and therefore faces challenges, for example, there are often insufficient resources to come up with an accurate estimation, or insufficient attention is given to scaffolding planning and controlling. Section 3.1 is divided into three parts, Section 3.1.1 introduces the current estimation method carried out in the client company; Section 3.1.2 briefs the current situation of scaffolding planning on site; Section 3.1.3 proposed a business model based on all the facts from the client company, to present the scaffolding request process in the mega industrial construction project being studied.

3.1.1 Estimate Phase

Scaffold estimates carried out in the client company depend on both historical data and experts' judgments. The process starts from the estimator comparing this coming project with all the projects that have been done, selecting the projects sharing similar features from database, and then looking up the ratio of total scaffold man-hour over total direct man-hours for these projects. Then, a factor will be decided to apply to the new project. Here, the major factors that estimators compare are type of the project, height of the project, level of modularization of the project, and congestion level of the project. Type of the project refers to the components of this project, which can be reflected directly from the proportion of each trade among the total direct man-hours. The height of the project can be reflected by average height, maximum height, or any other format. Level of modularization largely affects the direct man-hour needs on site. Specifically, the higher the modularization level for a project, the less direct man-hours needed for this module on site, which means less scaffold man-hours needed within this module as well. Congestion level is another important factor to the efficiency of trades on site, a site with

higher congestion level than its maximum capacity would severely influence the productivity of all trades working at the same time on site.

Scaffolding, treated as indirect work, is usually considered after the estimates of direct work. This also depends on time, type of contract, and scaffold estimate level divides. Sometimes, a bulk figure for the total scaffold man-hours is enough for the bidding process. Other times, estimates for scaffolding need to be detailed for each trade. Then, ratios of scaffold man-hours to each trade will be figured out by an experienced estimator from historical data for the scaffold estimate purpose.

From a risk management perspective, contract type is another factor which puts huge weight on to the estimates strategy. For example, if the contract is lump-sum, the contractor assumes more risk, which will lead them to a more conservative estimate; while for a reimburse contract, the situation is opposite and the estimator will try to squeeze the total direct man-hours. Nevertheless, this factor is subjective, and won't be considered in this research.

3.1.2 Planning Phase on Site

According to meetings and interviews with experts on field, it is clear that the current practice of scaffold planning is short-term. Usually, scaffolds requests are made two or three days in advance. Sometimes, in some special projects, like the project analyzed in this research, scaffold requests are demanded a week in advance. Scaffold requests are submitted by specific trade foremen, who are only concerned only with their own needs. As a consequence, the majority of scaffold requests are near-sighted, not planned into the big picture. It is fair to say that scaffold planning is done partially in an ad hoc manner, with lots of subjective factors affecting the decision, but not enough scientific and logical considerations.

3.1.3 Business Model

In order to achieve a better understanding of the scaffold request process on site, and how the scaffold request database was built, a series of interviews with an experienced project manager, scaffold foreman, and scaffold coordinator, who built and maintain the scaffold request database, were held. Also, a site visit lead by the scaffold foreman of a module yard was made. According to all the information gathered through those interviews, the site visit, and the meetings, a business model using IDEF0 diagram was created. The model is shown in section 3.1.3.2.

3.1.3.1 Processes

Within this business model, the most central part is the scaffold request database, and all the other processes directly or indirectly interact with this database. A list of processes involved in the business model is shown below:

- Receive scaffold request from foreman
- Conduct superintendent meeting
- Check and allocate scaffold request (which is further divided into three sub-activities: scaffold coordinator check scaffold requests; scaffold superintendent inspect existing scaffold, approve and prioritize scaffold requests; scaffold foreman site visit and allocation)
- Erect/modify approved scaffold request
- Use of scaffold by trade and weekly inspection
- Dismantle scaffold

An example of the scaffold request form is shown below. The figure is a typical scaffold request form used by the client for scaffold requests.

Requested By:			
Date Requested:			
Date Required:			
Purpose:		**ERECTION**	**TEAR DOWN**
	Mod Interconnects	☐	☐
	Mod Cross Connects & Loops (within Mod area)	☐	☐
	Insulation	☐	☐
	Equipment	☐	☐
	Piping	☐	☐

	Modification	☐	☐
	Electrical	☐	☐
	Others	☐	☐
Estimated Duration Req'd:			
Area & Location:			
Estimated Size:	Width (M) Length (M) Height (M)		
Review Alternate Access			
Approved / Tracking Number:			
Priority: **(by Scaffg Supt)**	**1**	**2**	**3**
Trades Using:	□ Labourers □ Carpenters □ Masons □ Millwrights □ Insulators	□ Ironworkers □ Pipefitters □ Electricians □ Painters □ Boilermakers	
Comments/Instructions:			

Figure 3 Example of a scaffold request form

3.1.3.2 Description of the Model

The scaffold circle starts with the foreman talking with his/her crew, and checking the construction schedule and drawings for scaffolding requirements (usually a week in

advance). Then he documents the information into a foreman's log. Based on the information in the foreman's log, as well as his experience, the foreman submits the scaffolding request. All the scaffolding requests will be stored into a database—scaffold request database.

At the same time, trade foremen from all trades, the scaffold superintendent, the scaffold foreman and the scaffold coordinator attend a superintendent meeting every morning. In terms of scaffold planning, this is a higher-level meeting, dealing with material availability, labor availability, and setting a general priority, which will be the higher-level control of the entire scaffold planning and scheduling.

A major process after the requests have been submitted is to check and allocate all the scaffold requests. This process is subdivided into the following three steps:

a. Scaffold coordinators input all the requests into the scaffold request database, and then sort them by required date and check for duplication.
b. Scaffold superintendent gets the requests list for a specific date (present or next day; scaffolding planning is short-term) and checks each request considering ground condition, site density, area platform, existing scaffolding, etc., making sure that the request is necessary, that there is enough material, and that the schedule requirement is met. Depending on the overall situation, some requests become modifications on existing scaffolds, some have to be erected from zero, others might be cancelled, and all the approved requests will be prioritized based on schedule. While checking scaffold requests, the scaffold superintendent checks existing scaffolding, and considers the overall situation as well, to decide whether dismantlement of some used scaffold is needed or not. All this information will update the scaffold request database immediately.
c. Next, the scaffold foreman goes through the approved scaffolding requests and assigns a scaffolding crew to each request. The approved scaffolding requests will update the scaffolding database.

After scaffold crews get the approved scaffolding request from the scaffolding database, they erect scaffolding according to the requirement. After completing the scaffold, a safety inspection will be done; if the scaffold is safe, it gets tagged. The erection information will update the scaffolding database. Next, the tagged scaffolds will be used

by originally requested trades and be inspected for safety weekly. Used scaffolds become a site condition and will be considered when the scaffold superintendent checks the site.

On the other hand, scaffold crews also get the dismantle requirements for certain existing scaffolds; they tear down this scaffolding, and release the material.

3.1.3.3 Model

The following two figures show the scaffold request business model in IDEF0 diagram.

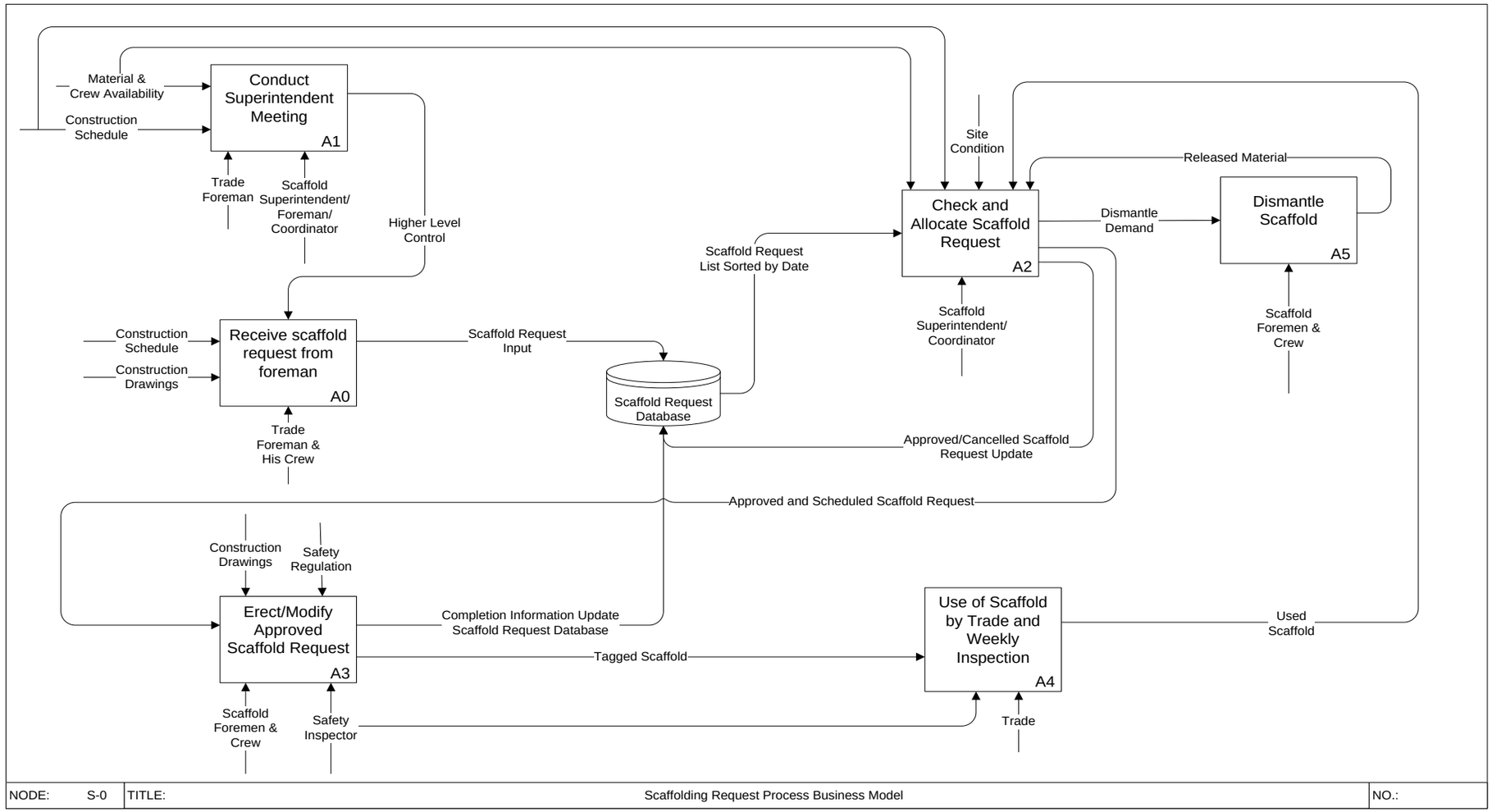


Figure 4 Scaffold request business model in IDEF0 diagram - 1

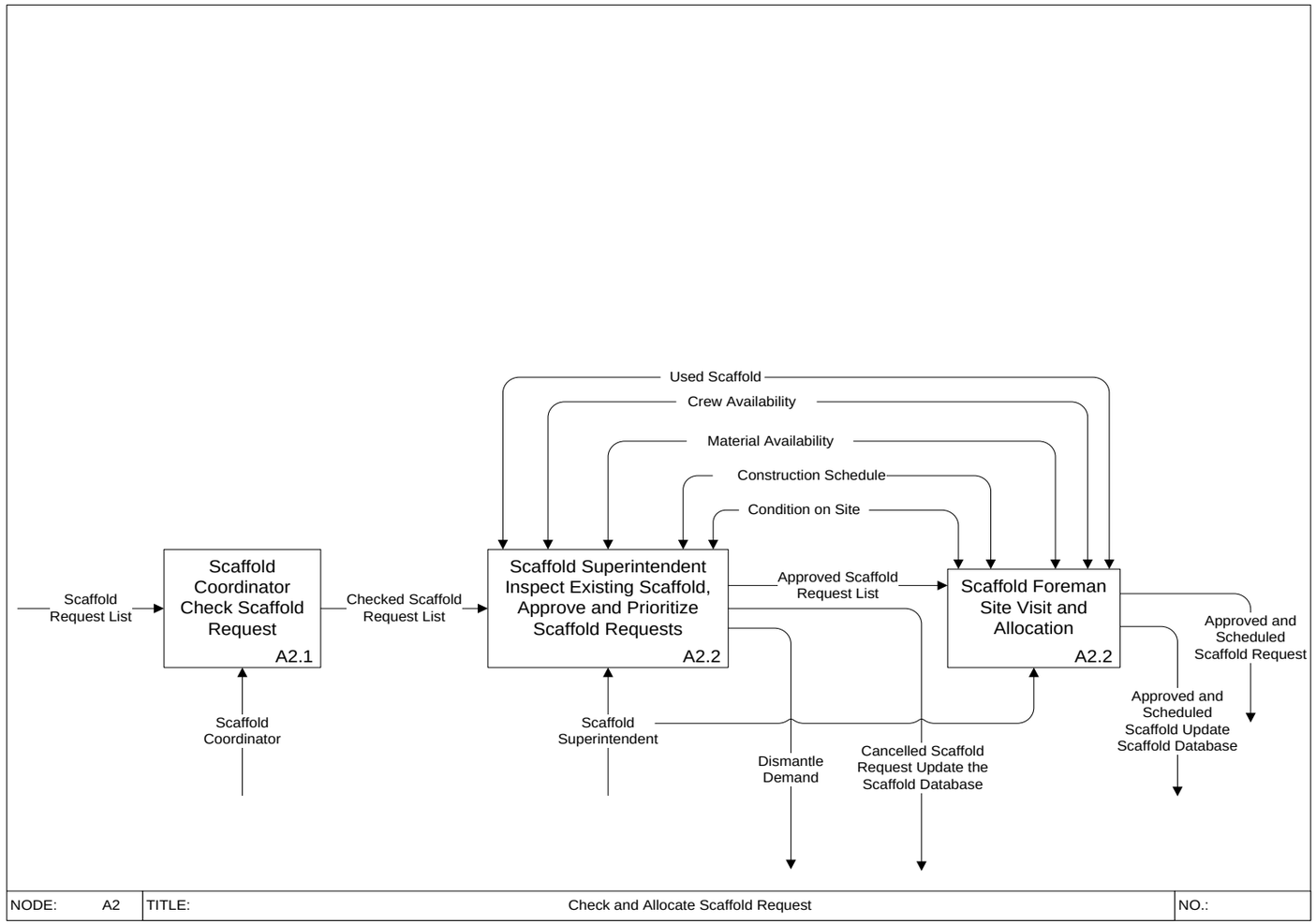


Figure 5 Scaffold request business model in IDEF0 diagram - 2

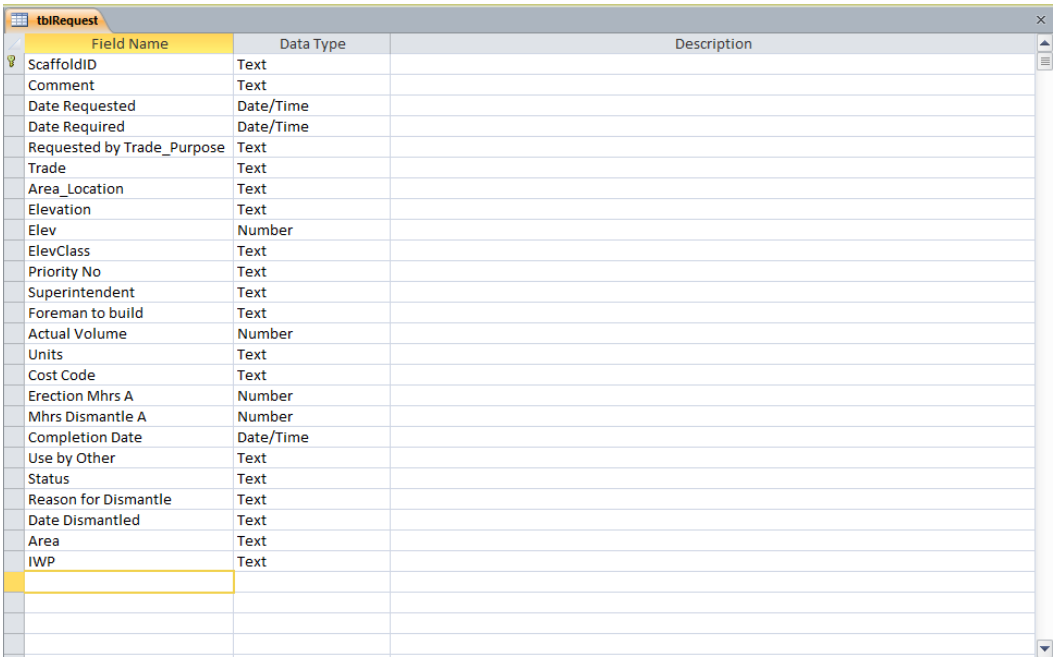
3.2 Data Pre-process

As explained and shown in the business model above, data has been collected and updated along with the project. However, collecting data is only the first part. With a huge amount of data, computer techniques must be involved in the data learning process to find the useful information lying behind the data, which can be used in prediction of future projects. This chapter provides a brief introduction of the scaffold database, and some analysis done using Excel and Access to reveal the features of this database, to prepare it for the following data mining experiments.

Section 3.2.1 is a detailed introduction of the scaffold database, and documents the definition of each column. Section 3.2.2 contains all the charts and figures generated from data analysis based on the Excel and Access platforms. This second part is sub-divided into three sections, which correspondingly present the analysis from the different perspectives.

3.2.1 Database Introduction

Before any data processing is done, the first and most important thing is to understand the data. The figure below is a screen shot of the "*tbRequest*" table, which the company was using to keep their scaffold activities information for a mega industrial project.


tblRequest

Field Name	Data Type	Description
ScaffoldID	Text	
Comment	Text	
Date Requested	Date/Time	
Date Required	Date/Time	
Requested by Trade_Purpose	Text	
Trade	Text	
Area_Location	Text	
Elevation	Text	
Elev	Number	
ElevClass	Text	
Priority No	Text	
Superintendent	Text	
Foreman to build	Text	
Actual Volume	Number	
Units	Text	
Cost Code	Text	
Erection Mhrs A	Number	
Mhrs Dismantle A	Number	
Completion Date	Date/Time	
Use by Other	Text	
Status	Text	
Reason for Dismantle	Text	
Date Dismantled	Text	
Area	Text	
IWP	Text	

Figure 6 Screenshot of the design view of "*tbRequest*" table

The primary key of this table is “ScaffolldID,” which represents the unique ID code for each scaffold request when it is submitted. The format of this “ScaffoldID” is “capital letter followed by dash, and then followed by a number,” for example “A-001.” The capital letter refers to the construction area this scaffold request is made within; the number automatically increases when new requests come in. However, this ID coding system only applies to the day shift scaffold activities. The scaffold requests that were assigned to scaffold night-shift crew, or done by night-shift crew, were recorded in the database following a different ID coding system. The ID for night shift starts with “N/S,” which stands for “night-shift,” and is then followed by a number, which ascends.

Comments document some special cases, for instance, the modification to the existing scaffold, or cancelled requests.

There are four dates tracked in this database; “Date Requested,” “Date Required,” “Completion Date,” and “Date Dismantled.” “Date Requested” is the date this scaffold request was submitted; “Date Required” is the date this specific scaffold was needed; “Completion Date” is the actual date when this scaffold request was completed; “Date Dismantled” is the date this scaffold was dismantled.

Information of elevation was kept in the database serving the purpose of reference for the scaffold crew. Here, elevation means the height level to which this scaffold needs to provide access, and most of them are kept in meters, which show the height above the sea level. Based on the knowledge that the project’s ground elevation is 622.5 meters, two fields were added to the original database, one is “Elev,” the other is “ElevClass.” The data type of “Elevation” is text and the records in this column are not consistent; some records contain more than one number, some are blank, and others are shown as “Various.” Thus column “Elev” was added to generally simplify records of column “Elevation.” Unlike “Elevation,” the data type of “Elev” is a number, and only contains one number for each record (the first number that appears in the Elevation column is used if more than one number is shows for one record). Then, according to the “Elev” records, a general classification has been done which is kept in “ElevClass.”

“Actual Volume” and “Units” were kept to serve the purpose of justifying the approximate quantity of materials and man-hours used in each scaffold request. These two columns are correspondent to each other: if the number in “Actual volume” is calculated by multiplying the area this piece of scaffold covered by the height of the

scaffold, then the record in "Unit" would be "m^3"; if the number in "Actual volume" is the area this piece of scaffold covered, then the "Unit" would be "m^2". These records kept the gross or overall volume of each scaffold, which doesn't accurately represent the actual volume of work, but it is a quick and easy method to handle on site.

"Erection Mhr A" is the column tracking the actual man-hours used in each scaffold request. When this database was designed and put into use, "Erection Mhr A" was expected to stay identical with the payroll sheet. However, due to multiple reasons, which will be explained in the following sections, the sum of Erection Mhr A was not consistent with the direct man-hours for scaffolds.

There are two columns – "Superintendent" and "Foreman to build" – that recorded all the supervisors' information for each scaffold request. The names of the scaffold superintendent and scaffold foreman who were directly responsible for the scaffold request were written down for each request.

After each scaffold had been used by all the possible trades, it was dismantled to release the material as well as reduce the site congestion. Dismantling information was documented – direct man-hours used for dismantling was kept in "Mhrs Dismantle A"; reason for dismantling was recorded in "Reason for Dismantle"; and when the scaffold was torn down was kept in "Date Dismantled."

Whether a scaffold request was approved or cancelled was kept in the "Status" column in the database. Generally there were three statuses – completed, cancelled, and in progress.

"Use by Other" column tracked information after each scaffold was used by its original requested trade, to see whether it was used by a different trade or a third party or not again. However, in this project, this information was not kept well, specifically, in the "Use by Other" column, only two types of records exist – "No" and blank.

Columns "Cost Code" and "IWP" were tracking cost and work package information, respectively. "Priority No" shows how important and urgent each scaffold request was – ranking from 1 to 3, the bigger the number the higher the priority. "Area_Location" documented where each scaffold request applied to, which is correspond to the serial number on drawings.

"Requested by Trade_Purpose" documented the trade (discipline) who submitted this scaffold request. However, because of the inconsistent records, a new column called "Trade" was added to clean, organize and group each existing record. In "Trade" column, there are only eight major disciplines confirmed, they are CIV – civil, EL – electrical, FP – fire proofer, INSTR – instrumentation, INSUL – insulation, IW – iron worker, ME – mechanical, PF – pipefitter. When doing grouping, some assumptions were made, which are listed below:

1. Trades BM (boilermaker), MW (millwrights) and mechanical equipment are grouped into ME (mechanical);
2. Trade ARCH (architecture), and all the blank records are grouped into PF (pipefitters);
3. EL (electrical) includes electrician, heat tracing and electrical heat tracing, because all the heat tracing in this project is electrical heat tracing.
4. PF (pipefitters) includes quality, hydrotesting, and all the sub-contractors related to pipe racks.

The table below shows the result of grouping column Requested by Trade_Purpose into eight major trades.

Table 1 Results of grouping trades

Requested by Trade_Purpose	Count	Sum Mhrs	Group	Trade
Labourers	1	15	CIV	CIV
Labourers, Carpenters, Pipefitters	1	48	CIV	
Labourers, Equipment	1	40	CIV	
Labourers/Carpenters/Masons	1	112	CIV	
Masons	1	22	CIV	
Carpenters	11	475	CIV	
Electrical Building	1	94.5	EL	EL
Electricians	2502	121779.75	EL	
Electricians/EHT	1	23	EL	
Electricians/HT	43	1133	EL	
Electricians/INSTRU	1	17	EL	

Heat Trace	7	569	EL	
Heat Tracing	1	69	EL	
Fireproofers	271	18859	FP	FP
Instrumentation	270	9115.75	INSTR	INSTR
Instrumentation(PF)	1	8.5	INSTR	
Instrumentation/Elec	5	109.5	INSTR	
Instrumentation/PF	2	31.5	INSTR	
Insulators	1370	38507.5	INSUL	INSUL
Insulators/Pipefitters	1	45	INSUL	
Insulators\Electricians	1	72	INSUL	
Ironworkers	209	7150.5	IW	IW
ARPIS	13	778	Mechanical Equipment	ME
Boilermakers	155	5641	BM	
Millwrights	263	11168	MW	
Millwrights/Pipefitters	1	50	MW	
P-426564-LAE	1	16	MW	
N.W.S.	1	86	ARCH	PF
NWS	6	114.75	ARCH	
(Blank)	10	725		
Other	1	15	PF	
Hydrotesting	48	974	PF	
Pipefitter/Insulators/Electricians	1	85	PF	
Pipefitters	8612	238192.75	PF	
Pipefitters/Electricians/Insulators	1	110	PF	
Pipefitters/Insulators/Electricians	60	4792	PF & EL	
Quality	88	2021	PF	
Continental Stress	45	1333.5	PF	
Ryan Ducholke	1	16	PF	
Simplex	2	41.5	PF	
Simplex Grinnell	1	5	PF	
TEAM	3	303	PF	
Scaffolder(s)	2	71.5		

3.2.2 Data Exploration of the "*tbRequest*" Database

A series of data exploration was done using Access and Excel manually, to reveal different aspects from this database. As a result, a better understanding and an overall idea of the data sets, input organization, initial model selection, and output class was gained.

3.2.2.1 Date Analysis

According to the scaffolding request business model, a basic analysis to indicate the running status of this scaffold requests system in this industrial construction mega project is conducted. Specifically, scaffold requests count and sum of scaffold man-hours are compared and made into charts, from two aspects.

a. Request in advance – lead time between required date and request date

This lead time is calculated by subtracting "request date" from "required date." The result – number of days, shows the time difference between each scaffold request's submission and required date; if it is negative, it means the actual request date was later than the date this scaffold was required. (The numbers of man-hours and counts on this figure, as well as all the numbers on the subsequent figures and tables, have been scaled for confidentiality reasons.)

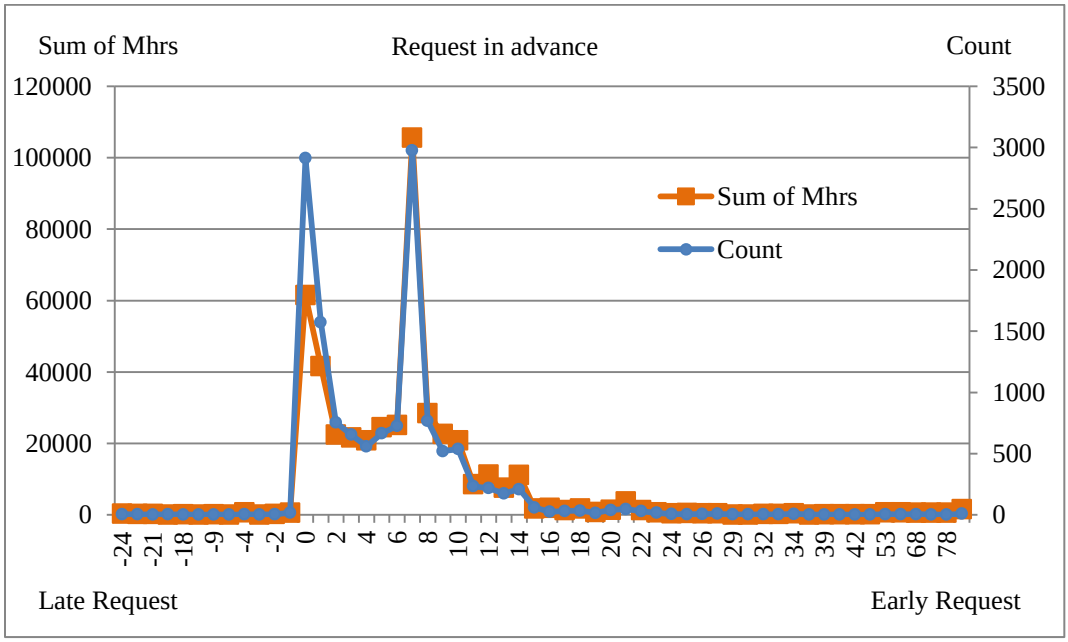


Figure 7 Man-hours and count of request in advance of both day-shift and night-shift

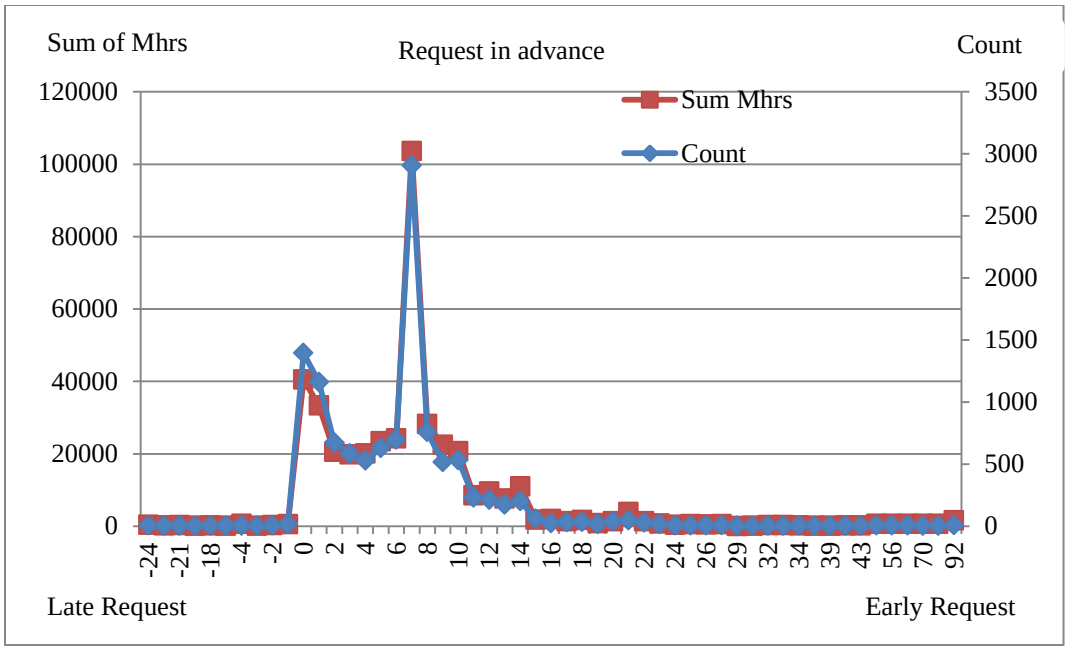


Figure 8 Man-hours and count of request in advance of day-shift

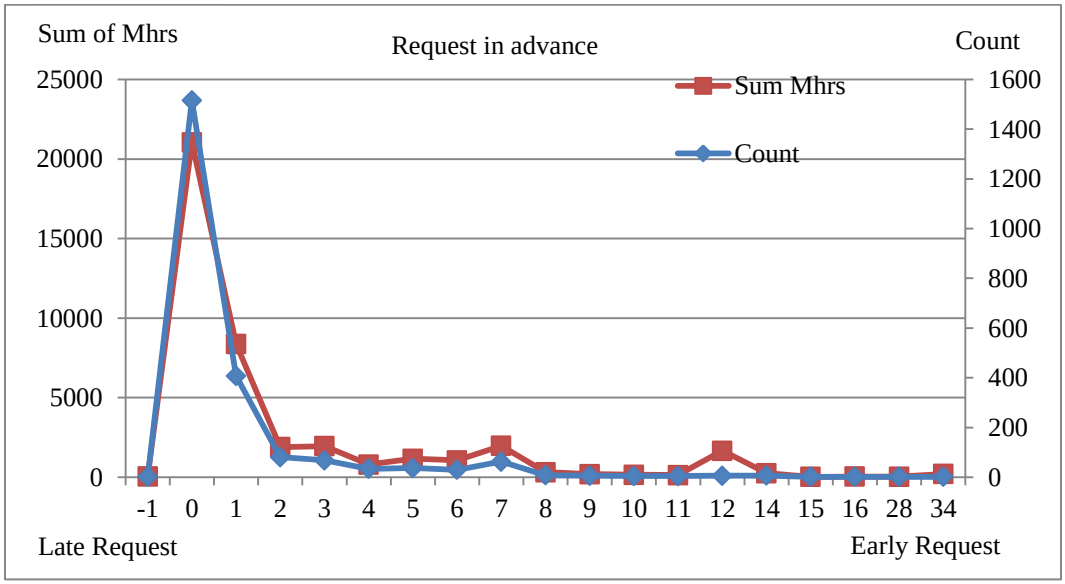


Figure 9 Man-hours and count of request in advance of night-shift

The three charts above are made based on day-shift, both day-shift and night-shift, and night-shift. From the first two charts, which show the lead time between request date and required date based on both day-shift and night-shift, and day-shift, respectively, it is clear that most data is concentrated among 0 days to 7 days, no matter the count or sum of man-hours. However, from the third chart, which shows only the night-shift cases, it is

manifest that most of the requests for night-shift were acted on in a more ad-hoc manner, and were submitted the exact day they were required.

b. Completion in advance – time difference between required date and completion date

The early completion is calculated by subtracting "completion date" from "required date." If the result is negative, it means the finishing date is behind schedule.

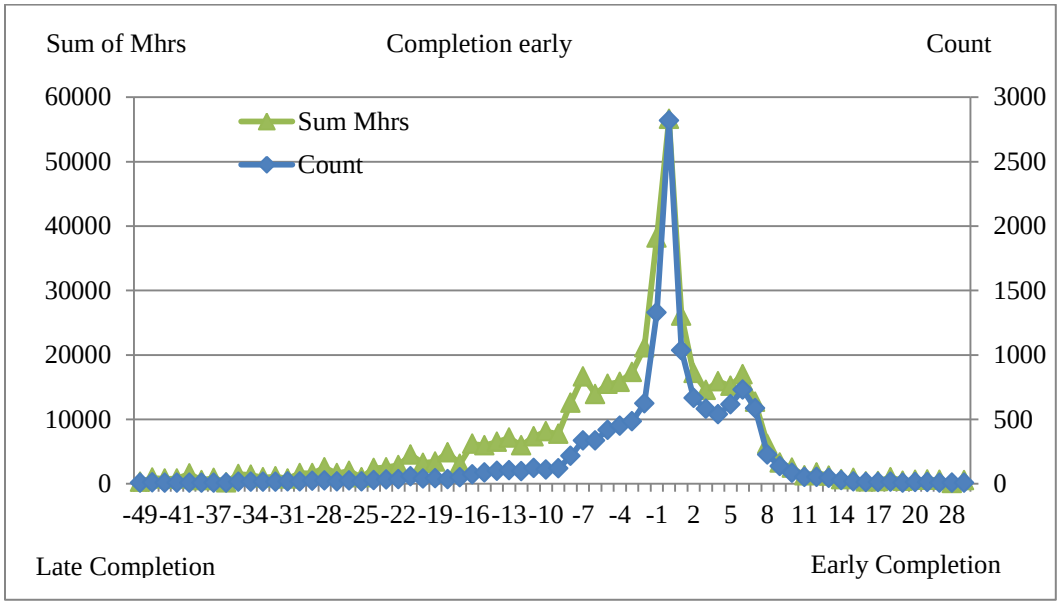


Figure 10 Man-hours and count of completion in advance of both day-shift and night-shift

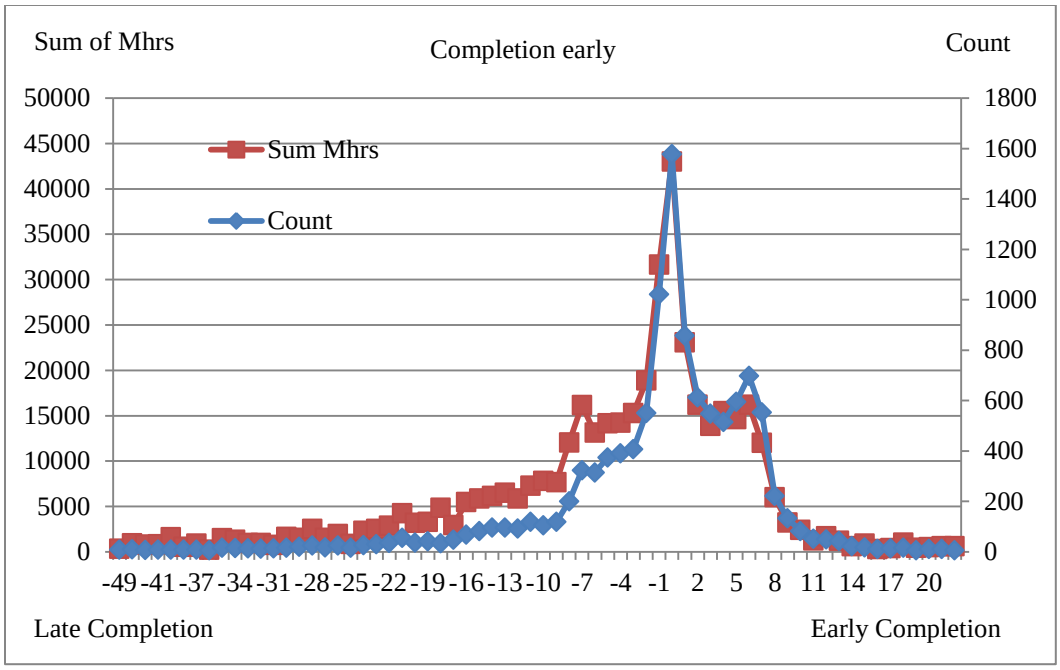


Figure 11 Man-hours and count of completion in advance of day-shift

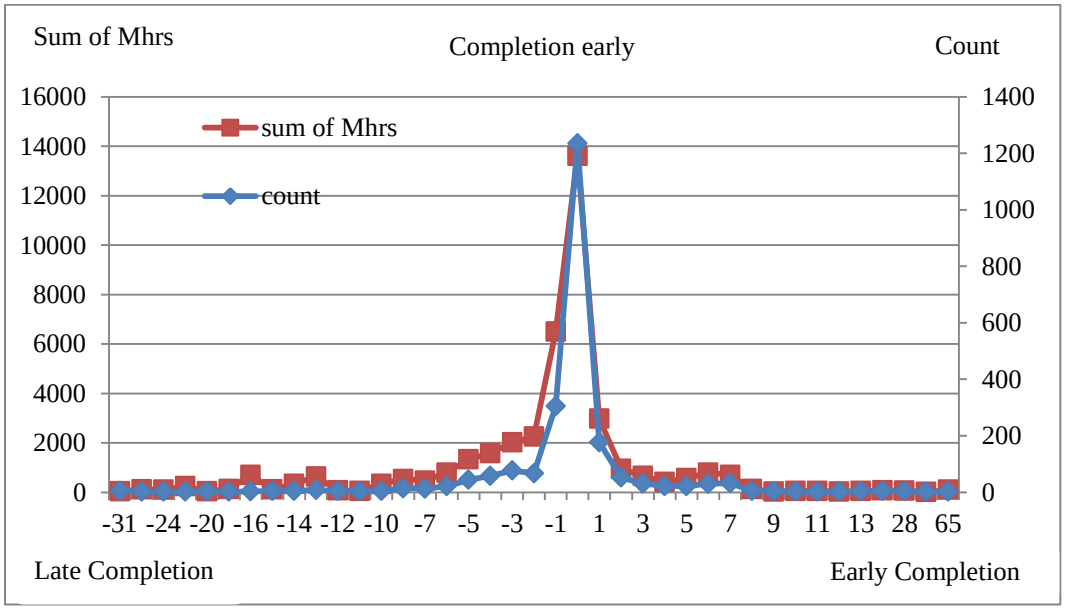


Figure 12 Man-hours and count of completion in advance of night-shift

The three charts above show how far ahead a scaffold request is usually completed before the required date. From the first two charts, it is clear that peaks of both are on 0 days, which means it was completed on the same day it was needed, although data quite largely rested between 7 days early in completion to 8 days behind schedule. However, the chart

of night-shift only shows a slight difference from the first two. Most of the night-shift scaffolds were completed among 1 day early to 1 day late, but the peak of the chart is still on 0 days in advance.

c. Conclusion

First, charts of day-shift, and both day-shift and night-shift show very similar trends in this analysis. This manifests that day-shift consumes most of the scaffold work, while night-shift only takes up a minor portion.

Second, generally, these six charts show how the scaffolding request system worked in this project, from one aspect. It also implicitly demonstrates how efficiently this system was functioning, as well as how accurately this database reflected the real world. Specifically, this analysis verified the basic scaffold management strategy, that weekly look ahead for scaffolding needs and any scaffold requests should be submitted at least 7 days in advance. Nevertheless, substantial requests were submitted the same date as they were required. This shows that scaffold planning is short-term in the real world, which changes from time to time.

According to the scaffold coordinator who built, updated, and maintained this scaffold database, one factor must be put into consideration when lead time between request date and required data is dealt with, that is – when filling in the scaffold request form, the foreman tends to put the required date a little bit earlier than the actual date that scaffold will be needed to create contingency time. After all, it is not hard to tell that scaffold planning is short-term. Some foremen didn't consider the whole picture when they are submitting scaffold requests; instead then only focused on the short-term benefit, and were constrained by their own trades.

3.2.2.2 Scaffolding Analysis on Different Construction Areas or Different Trades

1. Scaffold features on different construction areas

Construction areas were divided at the engineering design phase by the engineer, according to type of structures, their functions, and location. Thus, different construction areas have different characteristics, which also show different behaviours in terms of scaffolding requirements. The site was divided into 18 different construction areas in this construction project, which is a high level physical division. The following figures and tables show scaffold features from one construction area to another.

a. Man-hours analysis

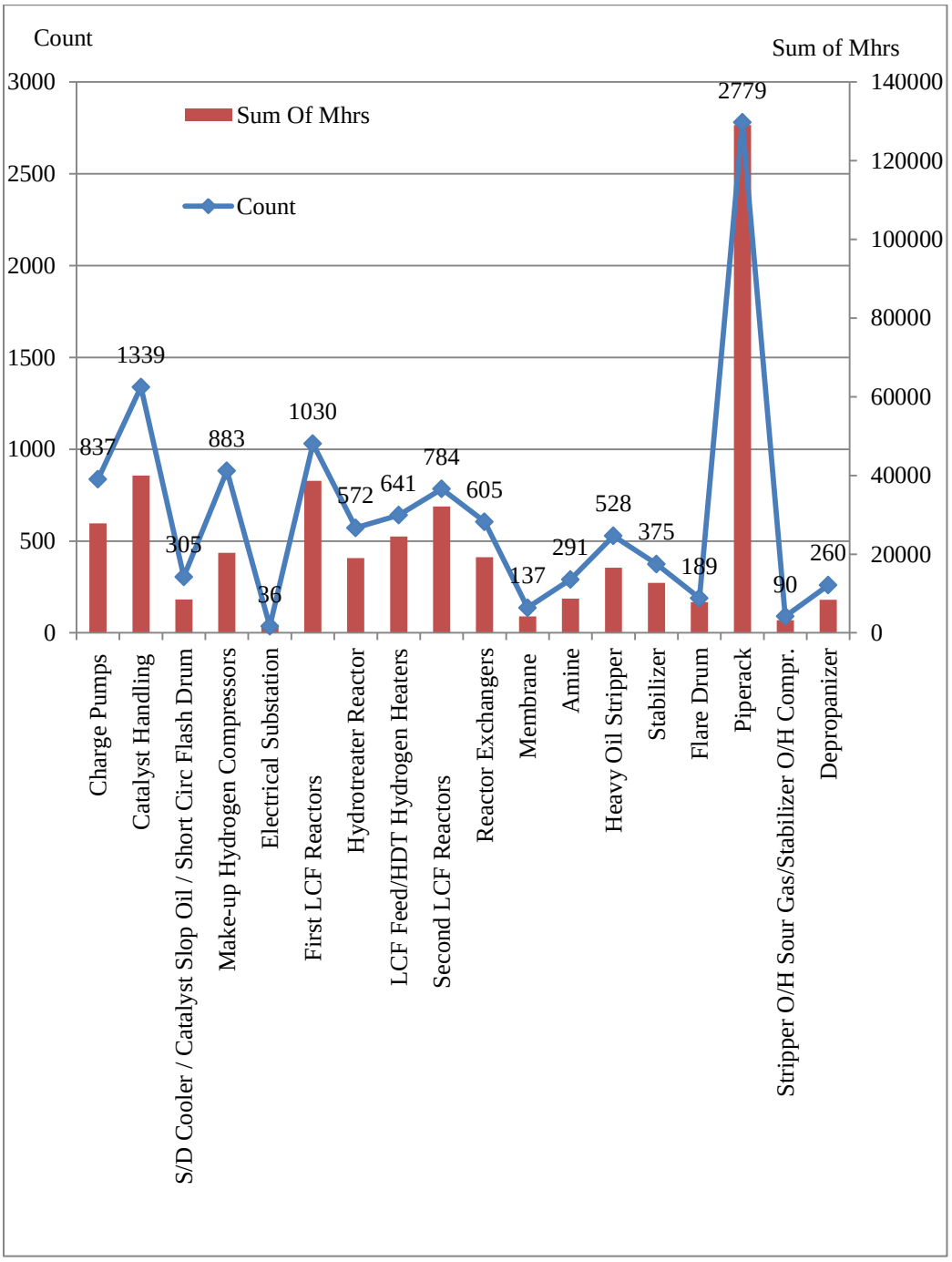


Figure 13 Scaffold man-hours and counts distribution on construction areas

b. Man-hour per volume analysis

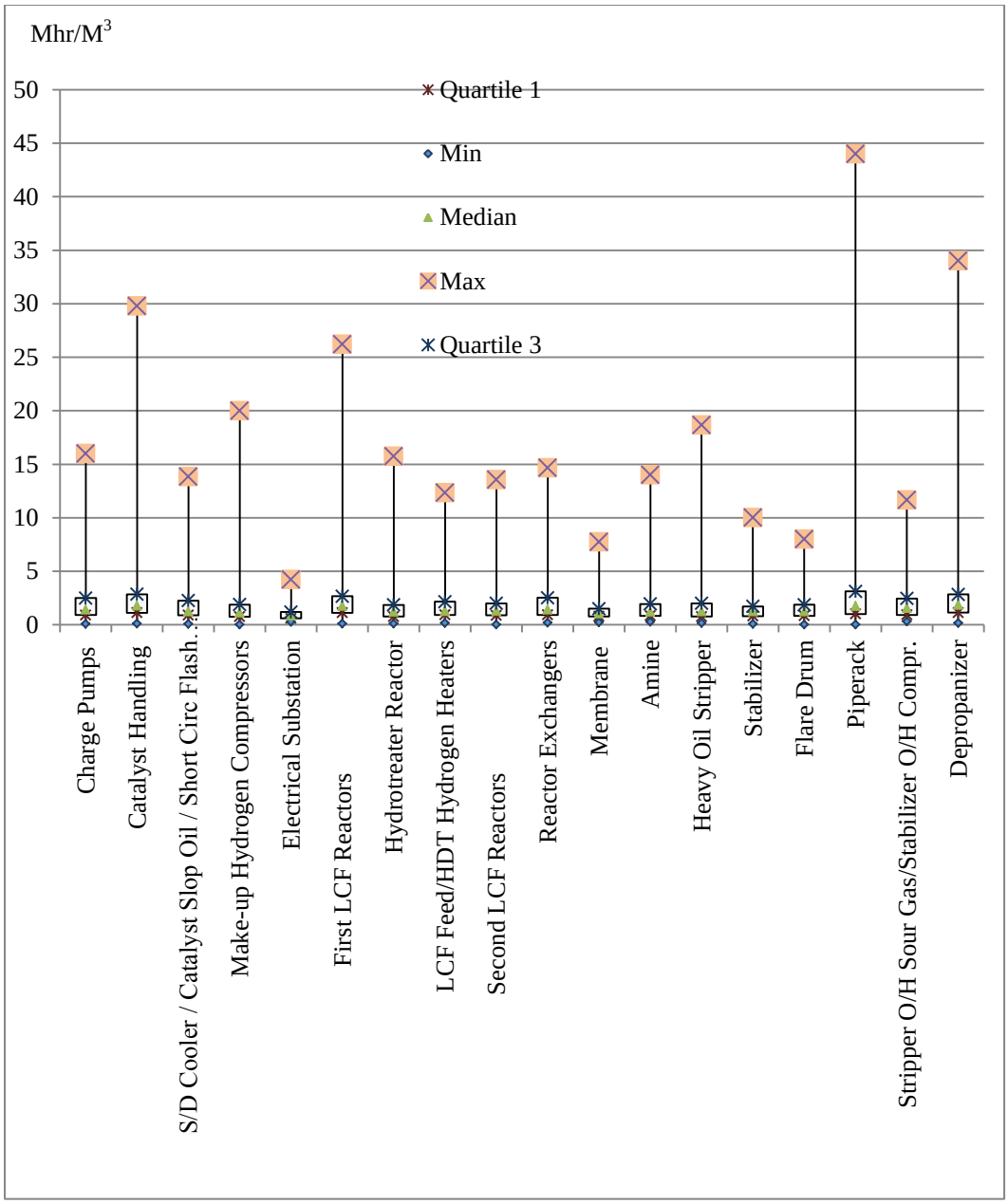


Figure 14 Box plot of man-hour per volume of each construction area

The figure above gives brief information about the average man-hour per volume distribution on each construction area. According to the chart, median, the first quartile, and the third quartile from each construction area are very close, though maximum numbers are all over the place. After further check and analysis of the database, several reasons, listed below, may explain the instability of the type of data:

- Some requests were met by changing or modifying the existing scaffold, which sometimes only took 2 or 3 man-hours, while the volume would show as "0."
- Some requests were located in a congestion area, although the volume of the scaffold was not too big, man-hours used were significantly more than usual.
- Other factors, for example elevation, and etc., influenced the man-hours per volume used.
- Volume recorded in "*tbRequest*" database was not accurate; lots of subjective measurements were involved, which created the noise in the data.

For these reasons, it is probably not wise to make a judgement based on the man-hour per volume from this database, or choose any class generated from volume data.

2. Scaffold features on different trades

Sorting scaffold activities according to different work areas is one way, organizing them based on different trades is another way to reveal scaffold features in mega industrial construction projects. The figure below shows the box plot of proportion of scaffold man-hours to direct trade man-hours for each trade.

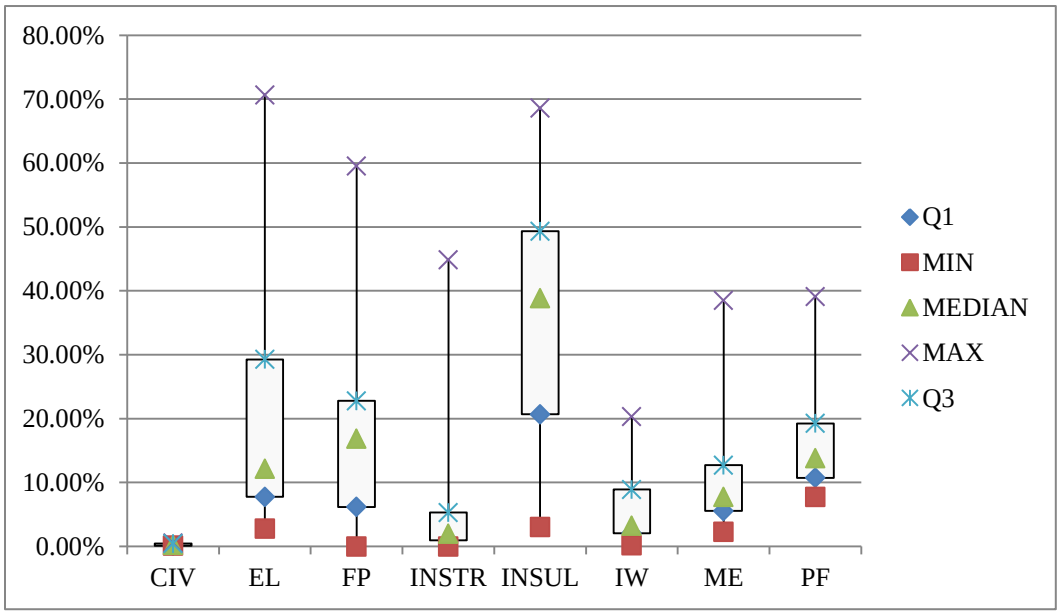


Figure 15 Box plot of proportion of scaffold man-hours to direct trade man-hours on each trade – excluding the three outliers

3. Conclusion

These three charts directly indicate: first, different trades have different requirements for scaffolding; second, different construction areas contain different features showing different scaffold performance. This conclusion agrees with the scaffold foreman, scaffold coordinator, and project manager's scaffolding experiences. For example, area "R" had more scaffolding activities than any of the other areas, and was where most of the pipe racks were located, which verifies the idea that around pipe racks there are heavy scaffolding requirements.

To conduct this research, more connection between the scaffold database and other types of data resources should be involved. Thus, the payroll sheet and as-built schedule were involved in this research as well. In the schedule, activities are organized by construction area, and within each area further down to each trade, and then each work package. The payroll sheet spreads out each individual who has worked on this project; it is a table recording actual man-hours based on individuals.

3.2.2.3 Timeline Analysis

This timeline analysis was conducted from two perspectives: one shows ups and downs of scaffold activity man-hours, by plotting both man-hours of scaffold erection and dismantle from "*tbRequest*" database; the other compares scaffold man-hours (including both erection and dismantlement) from different sources – "*tbRequest*" database, payroll sheet and as-built schedule.

a. Timeline of erection and dismantle man-hours from "*tbRequest*" database

In this industrial construction mega project, the scaffolding database started from September 2008 and went to December 2010. A timeline analysis was conducted to show the peak time of scaffolding activities during the whole project lifetime.

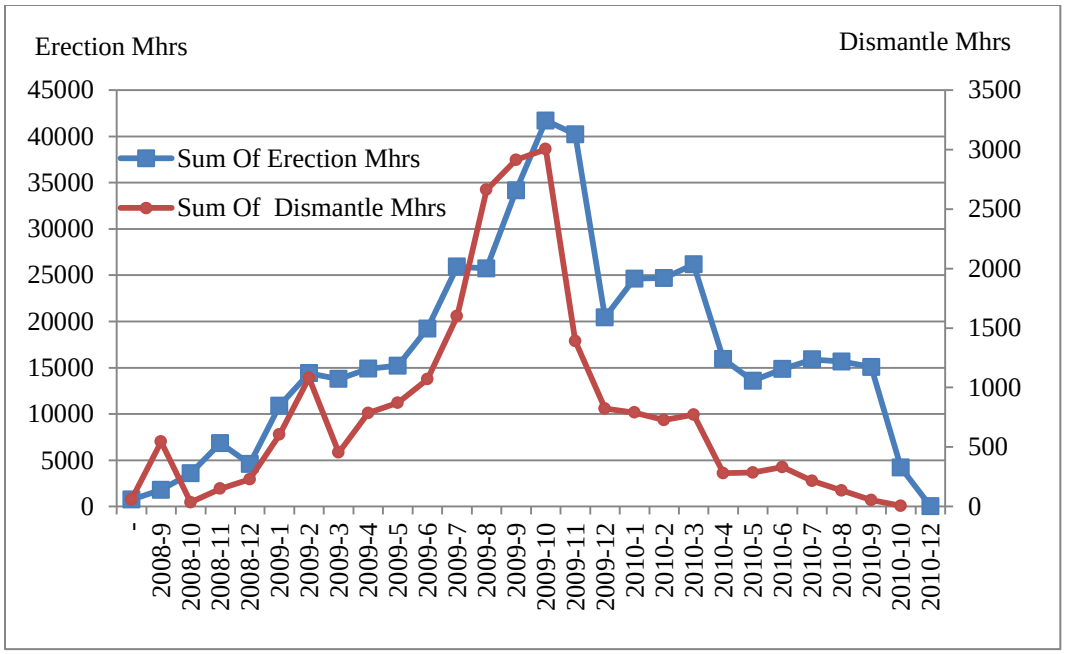


Figure 16 Erection and dismantle man-hours plotted in timeline from "*tbrequest*"

Note: There were 31 records without "completion date," though the status shows "completed" and certain "Erection Mhrs" were noted.

The chart above shows the scaffolding erection as well as dismantlement man-hours along the project life time, from the "*tbRequest*" database perspective. From the chart, it is noticeable that the busy time for scaffold activities of this project is during July, 2009 to March, 2010, for both erection and dismantlement. Further, the peak points of both two lines are on October 2009. The trends of both lines are close.

b. Comparison between "*tbRequest*", payroll, and as-built schedule

In the schedule, a bulk number shows the total scaffold man-hours, while in payroll, the sum of man-hours of all the workers who worked on scaffolding is calculated. These two numbers are very close (schedule is 2.5% higher than payroll). However, a certain gap between them and the sum of scaffold man-hours from scaffold request database is discovered. The table below shows the exact number from each source. The man-hours kept in "*tbRequest"* database are the man-hours used by laborers to erect or dismantle the scaffolding, which does NOT include the man-hours of the foreman who was in charge of the scaffolding activity.

Table 2 Scaffold man-hours comparison

Sources	Erection Man-hours	Dismantle Man-hours	Total Man-hours
"tbRequest"	574098	26998	601096
As-Built Schedule	N/A	N/A	838822
Payroll (without foreman)	N/A	N/A	767761

i. Man-hours plot on timeline

In order to show the difference of scaffold man-hours between the pay roll sheet and scaffold request database "*tbRequest,*" a timeline chart was made to show the monthly diversity. Assumption: the timeline of "*tbRequest*" database is based on completion date.

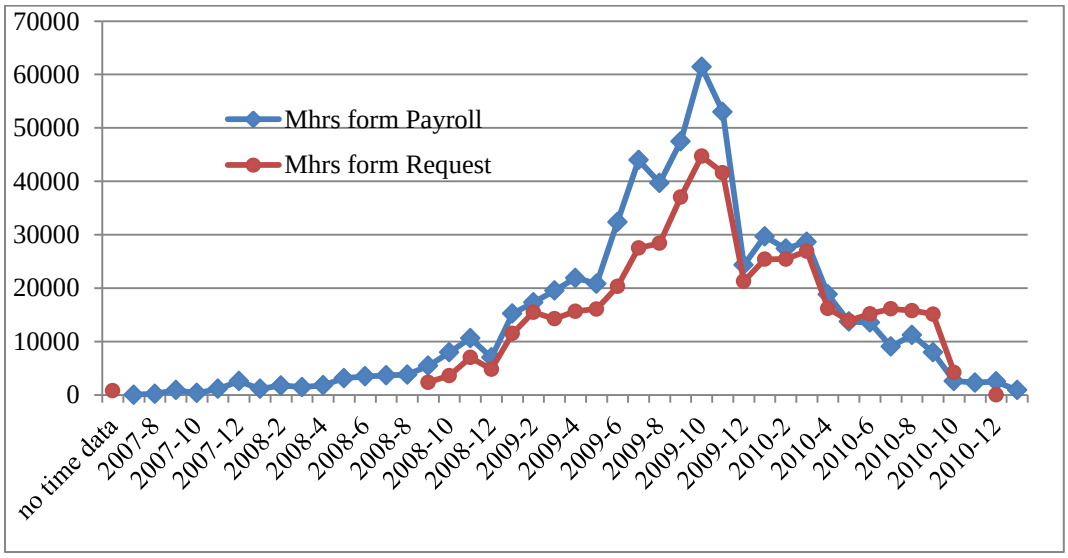


Figure 17 Man-hours plotted in timeline of both payroll sheet and "*tbRequest*"

The chart above shows a specific comparison made on equal ground: first, comparing over the same time period, the "*tbRequest*" scaffold database missed the starting point of this project, which is from August 2007 to August 2008; second, it excludes the scaffold foremen's and scaffold superintendents' man-hours, which are not recorded in the scaffold database. Generally, the man-hours from the payroll sheet are more than

"*tbRequest*," except from May, 2010 to October, 2010, which is the finishing phase of this project. Nevertheless, the trends from both documents are almost the same.

ii. Ratio = Mhrs from Payroll/Mhrs from tbRequest - 1

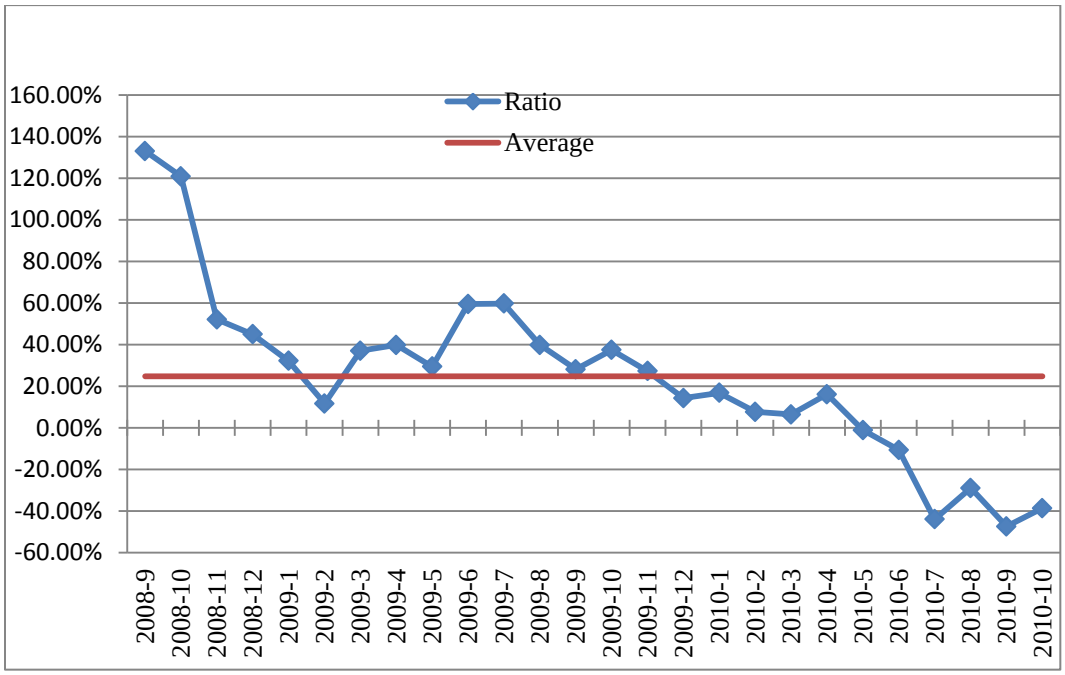


Figure 18 Ratio of man-hours difference between payroll and "*tbRequest*"

This chart shows the ratio (payroll/"*tbRequest*" minus 1) of two different sources regarding scaffold man-hours. It is clear that the ratio fluctuates by 24.76% -- the average of the differences, and during the first several months, the man-hours from payroll are significantly larger than *"tbRequest,"* however, towards the end, the gap becomes smaller, until it drops below 0, which indicates scaffold man-hours kept in the *"tbRequest"* is larger than that recorded in the payroll.

c. Conclusion

The total man-hours from payroll (foreman exclusively) is 1.28 times the total man-hours from "*tbRequest*" database. There are several reasons which contribute to this big difference:

- Some preparation time, which is sometimes a considerable amount of time is not considered in "*tbRequest"* database. This preparation time includes material

handling time – loading, unloading; cleaning up the site; moving equipment; and etc.

- Some minor modifications, or some urgent scaffold requests may not have been updated in the *"tbRequest"* database;
- Some human errors may have been present.

Although this scaffold "*tbRequest"* database was designed to capture all the scaffold activities, and coordinate with the payroll sheet, the reality is a deviation from that.

3.3 Summary

All work and efforts related to the preparation of scaffold request data were presented in this chapter. This research focused on building a mathematical model using data mining method based on a historical scaffold request database. The key point for how well this mathematical model would perform significantly depends on the quality of the data. Thus a substantial amount of effort was devoted into data preparation and pre-process phase.

During this phase, a deep and thorough understanding of scaffold activities as well as the scaffold request database of the client company was obtained. Based on that, a business model was drawn to simplify and symbolize the scaffold request process. Meetings, interviews, and site visits were held to get first-hand information about this database form different experts.

After that, correcting errors, re-organizing the structure, grouping and re-formatting, under the supervision of scaffold coordinator, were done to clean up and prepare the database. Following that, a series of systematic data analysis using Excel was done to explore the data from different perspectives. To sum up, conclusions are listed below.

- First, scaffold planning is a short-term activity. Usually, foremen of an individual trade didn't consider the whole picture when they submitted scaffold requests; instead they only focused on the short-term benefit, and were constrained by their own trades.
- Second, on one hand different trades have different requirements for scaffolding; on the other, different construction areas contains different features showing different scaffold performance.
- Third, scaffold information from different sources show fairly good consistency. Bulk number of total scaffold man-hours spent in this project

from as-build schedule and pay roll sheet are very close, and total scaffold man-hours from scaffold request database showed about 80 percent of the numbers from the other two sources.

- Fourth, different sources could be connected to get a better data set for the following data mining investigation.

Data preparation in this research took the majority of the time and effort. Nevertheless, a solid foundation was built for the following data mining investigation. All the information gained in this phase of data analysis reveals the features of the scaffold request database involved in this research, which enabled the following data mining experiments.

In addition, some lessons learned for the client purpose are listed below:

- First, a better designed scaffold request database and increasing the consistency of recording data, for example a drop-down list which provides only appropriate options to choose from, would help reduce the human error;
- Second, all trade foremen should communicate with each other on a daily basis, which will help in multi-use of scaffolds. They should look at the big picture needs, not at those for a particular line/job. The scaffold coordinator and scaffold superintendent should review all the requests and organize scaffold requests from different trades as a unit system;
- Third, better connection among different documentation should be discussed. For example, if cost code is well tracked in scaffold request database, then scaffold request database will be easy to connect to payroll sheet, or if work breakdown system code is well documented in the scaffold request database, then each scaffold request could connect to the specific trade and location, which could correspond to the schedule.

Chapter 4 – Data Mining (Machine Learning) Investigation

The data mining of the scaffold data was done using an open source data mining tool – WEKA (Waikato Environment for knowledge Analysis), which provides an assembly of up-to-date data mining algorithms. WEKA also provides a group of tools for data pre-processing. This tool was developed in University of Waikato in New Zealand (Witten, Frank, & Hall, 2011; Bouckaert, et al. 2012).

This chapter elaborates the data mining process of this research. The data mining investigations followed a trial and error process. This process started from understanding the database, which was gained through data preparation. Then, with more experimental data generated from the data mining investigation, as well as discussion with field experts, changes and modifications were made to the data mining process. The whole data mining process can generally be divided into three phases. The first phase – initial data mining investigation – was based on the initial input table, which was built based on the understanding of the scaffold request database, as well as the experts' opinions. The experimental data from initial data mining investigation is contained in Section 4.1. Based on the results from the initial data mining investigation, and discussion with the experts, the second phase – data mining investigation on modification input table – was testing scaffold request data based on a modified input table, which reflected the changes done to the initial input table according to the results and experts' advice. The last data mining investigation phase – data mining investigation with scaffold man-hour as class was done based on the input table from phase two, with changed class value. All the experimental data from the second phase of the data mining investigation is presented in Section 4.2, followed by Section 4.3, containing the data mining results from phase three. Section 4.4 provides a summary of the whole data mining process.

4.1 Phase One – Initial Data Mining Investigation

The initial data mining investigation was based on an input table generated from the scaffold request database. This initial input table was built on understanding obtained from the data pre-process as well as meetings with the field experts. Then, using this input table, a set of data mining experiments were done. Section 4.1.1 explains the process of building the input table; the following Section 4.1.2 explains the experiments design, and Section 4.1.3 presents the experimental data from initial data mining

investigation including a best performance model and the summary for the data mining analysis at this phase.

4.1.1 Input Table Preparation

The target of this research was to build a mathematic model for scaffold estimation for future projects based on an existing scaffold database. Organization of input, selection of learning method (or algorithm), and format of output are vital decisions to the success of this research.

Due to the fast track feature of industrial construction projects, information ahead of a project is incomplete and limited. In addition, scaffolding is treated as indirect work. Thus, the information available for this scaffold estimation tool will be rough, and high level. Given this reality, information level of the existing scaffold database is too detailed, a model trained on the bottom information level is unnecessary as it would be useless at the beginning of a construction project.

Due to the fact that the scaffold estimates are done after the estimate and schedule of the direct work, to organize and represent the scaffold data at a higher level where each instance represents one trade in one construction area would be a proper choice for this research. Thus, the predicted class will be able to connect with the project schedules, which record the direct trade man-hours within each construction area. This was done through the following methods: grouping some information, aggregating some records, revealing information statistically, and converting the format. For example, as introduced in Section 3.2.1, values of attributes "Requested by Trade_Purpose" were messy and inconsistent. Grouping values of "Requested by Trade_Purpose" to "Trade," which only contains eight major disciplines, was done to better present the data. Another example, for volume and elevation records, a basic statistical analysis including average value, maximum value, minimum value, standard deviation, and mode, was done to present this type of information to trade in construction area level.

Other than scaffold request database "*tbRequest,*" as-built schedule and payroll sheet of this project were available for this research. Thus, information of direct man-hours for each trade in every construction area from as-built schedule was collected and integrated into the input table. However, payroll sheet was organized in a different format, which generally records the names of the employee, employee number, union code, job type, trade information, total hours, work data, and shift information. Directly based on this

information, as the payroll sheet failed to provide any information of construction area, or work package, it is not enough to generate an attribute to link to the input table, in which each instance represents one trade in one construction area.

As the result of preparation of the data, an initial input table was built with nineteen attributes plus a class, which is a quotient of scaffold man-hours by direct trade man-hours. A set of assumptions listed below, were made:

1. All scaffold information (scaffold Mhrs; scaffold count, and scaffold distribution) including height and volume came from day-shift records, because night-shift records in the database used a different ID number system, which cannot track construction areas or some other information needed to build this table.

2. Trade Mhrs, and Trade Mhr Distribution are from as-built schedule of the project. This information is not from payroll, because it's impossible to locate records in payroll into work packages and work areas. However, as presented in Section 3.2.2.2 the total man-hours for all the trades are very close.

3. Scaffold man-hour of night-shift is about 0.11 times that of day-shift; however, night-shift is not evenly distributed to each area. Thus, no effort was made trying to consider night-shift scaffold records in this phase of the research.

4. Winter efficiency decline is offset by choosing the class of this phase as a quotient of scaffold man-hours by direct trade man-hours. Since efficiency of both scaffold trade and direct trade would be affected by winter conditions, when choosing ratio as the class, it cancels out the influence of winter.

5. Three records have been excluded from the input table. These three records represent the special cases, where our client performed scaffold work for a sub-contractor who did the actual trade work. Due to the particularity of these cases, they would play no contribution to the data mining research, and would confuse the training process.

The initial input table contains 20 columns, which are: A – Construction Areas; B – Trades; C – Area; D – General; E – Elevation; F – Scaffold Type; G – H_Mean; H – H_StDev; I – H_Max; J – H_Min; K – H_Mode; L – V_Mean; M – V_StDev; N – V_Max; O – V_Min; P – V_Mode; Q – Count; R – Scaffold Mhrs; S – Trade Mhrs; T –

Scaffold Distribution; U – Trade Mhr Distribution; V – Scaffold Mhrs/Trade Mhrs. The following three figures are part of the input table showing its layout, due to the big size, the big table is broken into three parts.

Table 3 Part of the initial input table

A	B	C	D	E	F	G
Construction Areas	**Trades**	**Area**	**General**	**Elevation**	**Scaffold Type**	**H_Mean**
A Charge Pumps	CIV		Pumps		Access	84.8113866
A Charge Pumps	EL		Pumps		Access	84.8113866
A Charge Pumps	FP		Pumps		Access	84.8113866
A Charge Pumps	INSTR		Pumps		Access	84.8113866
A Charge Pumps	INSUL		Pumps		Access	84.8113866
A Charge Pumps	IW		Pumps		Access	84.8113866
A Charge Pumps	ME		Pumps		Access	84.8113866
A Charge Pumps	PF		Pumps		Access	84.8113866

H	I	J	K	L	M	N	O	P
H_StDev	**H_Max**	**H_Min**	**H_Mode**	**V_Mean**	**V_StDev**	**V_Max**	**V_Min**	**V_Mode**
6.0804346	88.56	6.75	84.78	4.1992589	7.3298312	91.935	0.135	1.62
6.0804346	88.56	6.75	84.78	4.1992589	7.3298312	91.935	0.135	1.62
6.0804346	88.56	6.75	84.78	4.1992589	7.3298312	91.935	0.135	1.62
6.0804346	88.56	6.75	84.78	4.1992589	7.3298312	91.935	0.135	1.62
6.0804346	88.56	6.75	84.78	4.1992589	7.3298312	91.935	0.135	1.62
6.0804346	88.56	6.75	84.78	4.1992589	7.3298312	91.935	0.135	1.62
6.08043	88.56	6.75	84.78	4.19925	7.32983	91.935	0.135	1.62

46				89	12			
6.0804346	88.56	6.75	84.78	4.1992589	7.3298312	91.935	0.135	1.62

Q	R	S	T	U	V
Count	**Scaffold Mhrs**	**Trade Mhrs**	**Scaffold Distribution**	**Trade Mhr Distribution**	**Scaffold Mhrs/Trade Mhrs**
0	15.1200	6131.751717	0.000562929	0.021505961	0.00033289
25	1059.615	10545.37269	0.039450288	0.036985903	0.013565004
5	293.6925	1369.871391	0.010934399	0.004804565	0.028943219
2	38.4075	3700.107236	0.001429941	0.012977427	0.001401314
6	160.8525	918.4421496	0.005988663	0.003221262	0.023643392
2	85.3200	1241.325000	0.00317653	0.004353713	0.009278956
2	61.0200	1071.770403	0.002271822	0.003759032	0.007686068
58	1912.005	13512.38392	0.071185428	0.047392135	0.019102527

Column A shows which construction area each instance belongs to; column B shows what trade this record marks. There are eighteen construction areas, and eight trades, which are CIV – Civil, EL – Electrical, FP – Fireproof, INSTR – Instrumentation, INSUL – Insulation, IW – Structural Steel, and PF – Pipe Fitting.

Columns C–F show the general features of each construction area; column C shows the description of area size; column D shows the general information; column E shows the general height description; column F shows the overall or most possible scaffold type. This is additional information added to the original database, trying to imitate the real information available at the starting point of a project.

Columns G–K show elevation information; these include mean, standard deviation, maximum, minimum, and mode of elevation, respectively, for each record in this table. This information was generated from records in the scaffold database, which represent the elevation distribution of each trade in each construction area. The unit for elevation is meters. Columns L to P show volume information; these include mean, standard

deviation, maximum, minimum, and mode of elevation, respectively, for each record in this table. The unit for volume is m^3.

Columns Q–S show scaffold information; scaffold information is collected from the database; number of scaffold requests, total man-hours of the scaffold requests, and the percentage of scaffold man-hours of each trade in the construction area to the total scaffold man-hours of all trades in the construction area are calculated.

Column T provides trade information for each trade in each construction area. Column U shows the percentage of the direct man-hours of each trade in the construction area to the total direct trade man-hours of all trades in the construction area; the way it was calculated and its meaning are quite similar to Column S.

The last column – column V is considered to be the class of this relation. For each instance, this is the quotient of column R divided by column T.

4.1.2 Experiments Design

The data mining experiments are designed on the scenario trying to exhaust the combinations of three parameters as much as possible: 1) learning algorithms, 2) settings of each learning algorithm, and 3) subset of attributes of the selected input table. However, to explore every possible combination is impossible if it is carried out manually. Thus, a systematic way of trying different possible combinations of these three parameters is carried out. This process follows the understanding of the scaffold request database, experts' advice, as well as the results from the existing experimental data through a trial and error method.

Specifically, the learning algorithm choosing process followed the rule "simplicity-first". As shown in the initial input table, the class of this research is a quotient of "Scaffold Mhrs" by "Trade Mhrs," which is numerical. When the class is numerical, and most of the attributes are numeric as well, linear regression is a natural technique to try first (Witten et al., 2011). Linear model is simple, concise, and frank, and shows good performance on lots of practical problems. Thus, it is chosen here as the major algorithm. Since WEKA provides collection of more sophisticated and state-of-art algorithms to treat numerical situation, other algorithms, for example Gaussian Process (Ebden, 2008; Rasmussen, & Williams 2006; Mackay, 1998), were tried a little bit in this research as a comparison.

For linear regression in WEKA data mining software, it contains three options for attribute selection method – M5 method, Greedy method, and no attribute selection (Kirkby, Frank, & Reutemann, 2008). The attributes selection concepts and knowledge were introduced in the Literature Review chapter, Section 2.4.5.2. Thus, at the beginning of this series of data mining experiments, for a large subset of attributes, different attribute selection methods were tested on the same data, to generate a computerized understanding of the different selection of attributes; after obtaining a certain amount of experimental data, some specific attribute subsets were chosen without any computer attribute selection, to train the model.

In WEKA, after loading the input table, the function – "visualize," which plots the data set into a two-dimensional area – can help researchers visualize the database itself. If the vertical axis presents the class, and the horizontal axis presents each attribute, then all the data points scattered in this two-dimensional area show correlation between each attribute with the class, which in this data mining investigation phase is ratio of scaffold man-hours over direct trade man-hours. The concept is that the closer all the data points are to the 45 degree line, the higher the correlation between this attribute to the class. For example if both axes present the class, the data points will follow the 45 degree line between two axes, like shown in figure 19.

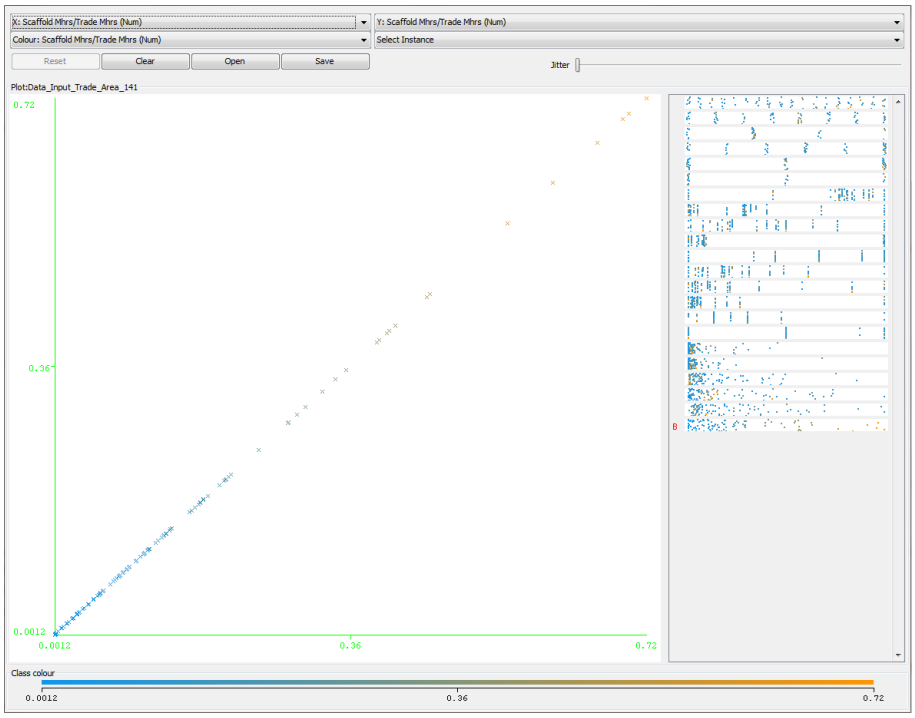


Figure 19 visualization of class value on both axes

The following two figures show the visualization of each attribute on class value (Y axis) – ratio of scaffold man-hours over direct trade man-hours. Please refer to the first line of blocks of each figure. From left to right, each block present one attribute – Construction Areas, Trades, Area, General, Elevation, Scaffold Type, H_Mean, H_StDev, H_Max, H_Min, H_Mode for figure 20, and V_Mean, V_StDev, V_Max, V_Min, V_Mode, Count, Scaffold Mhrs, Trade Mhrs, Scaffold Distribution, and Trade Mhr Distribution for figure 21.

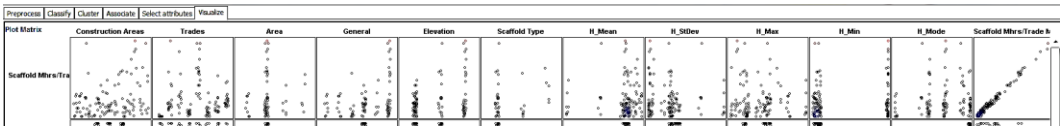


Figure 20 Visualization of correlation of first eleven attributes between the class

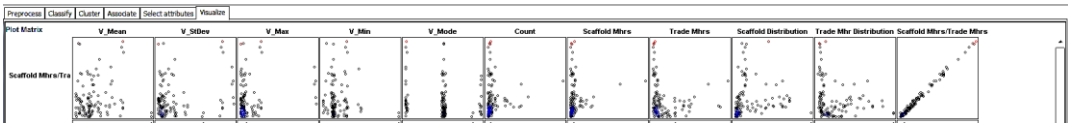


Figure 21 Visualization of correlation of last ten attributes between the class

It is clear that some attributes, for instance "Trade Mhrs," "Count," "Scaffold Mhrs," "Scaffold Distribution," "Trade Mhr Distribution," have higher correlation between the class value than others, for example "H_Min," "H_Mode," "V_Mode." However, some attributes, like "Count," "Scaffold Mhrs," and "Scaffold Distribution" have direct connection with the class; thus, they should be avoided as input attributes to train the model. As a consequence, according to the attributes visualization, a clearer idea of attribute quality could help the attribute selection process.

4.1.3 Experimental Data and Summary

4.1.3.1 Experimental Data

In this data mining investigation phase, all the experiments were done using 10-fold cross validation test mode, trying to get the best estimate of error on a set of unknown data (Witten, Frank, & Hall, 2011). The experimental data was documented in an Excel sheet, tracking four types of basic information of each data mining experiment. The first type is input, which refers to the input table and attribute selection; the second is information of experiment settings, which includes computer learning method, test mode, and parameter

choosing; the third one relates to output, which contains class as well as the model trained from the data mining process; the last part of the table shows the performance of the model, which includes correlation coefficient, mean absolute error, root mean squared error, relative absolute root error, and relative squared error, and it was followed by comments.

The complete data mining experimental approaches and the corresponding experimental results are attached as Appendix 1 to this thesis, due to its large size. Some simplified tables are shown here. The following table briefly shows the experimental results from part of the experiments that have been done.

Table 4 Simplified table of selected performance comparison of initial data mining

Input		Algorithm	Performance				
			Correlation Coefficient	Error			
Data set	Attributes			Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared
Initial Input	11 attributes	Linear Regression	0.7396	0.0804	0.1109	66.1564%	67.3091%
Initial Input	9 attributes	Linear Regression	0.7405	0.0796	0.1108	65.4841%	67.2064%
Initial Input	6 attributes	Linear Regression	0.7457	0.0793	0.1098	65.2455%	66.6172%
Initial Input	11 attributes	Gaussian Process	0.7614	0.0761	0.1082	62.6092%	65.6705%

Like some of the results shown in the above table, after a series of experiments, one simple conclusion is that no significant performance difference exists between linear regression and the more sophisticated algorithm (Gaussian process using normalized PolyKernel function with an exponent of 2). However, due to the limitation of random variation from dividing the folds in a single 10-fold cross validation test mode, the error estimate might not be reliable. In order to eliminate the random factor, a set of experiments repeating ten times have been run to get the accurate error estimate of the comparison between linear regression and Gaussian process. The table below shows the results of a set of experiments run at ten times 10-fold cross validation method by both linear regression and Gaussian process based on the same attribute selection.

Table 5 Results of statistical comparison between linear regression and gaussian process

Experiments Inputs	Algorithms	Result									
		Correlation Coefficient		Mean Absolute		Root Mean Squared		Relative Absolute		Root Relative Squared	
Initial Input_3	LinearRegression_NO attribute selection	0.73	Equal	0.08	Equal	0.11	Equal	63.87%	Equal	71.19%	Equal
	GaussianProcess_NormalizedPolyKernel_E2.0	0.71		0.08		0.11		64.54%		70.87%	
Initial Input_4	LinearRegression_NO attribute selection	0.74	Equal	0.08	Equal	0.11	Equal	64.33%	Equal	68.98%	Equal
	GaussianProcess_NormalizedPolyKernel_E2.0	0.72		0.08		0.11		63.42%		70.10%	
Initial Input_5	LinearRegression_NO attribute selection	0.74	Equal	0.08	Equal	0.11	Equal	65.20%	Equal	69.63%	Equal
	GaussianProcess_NormalizedPolyKernel_E2.0	0.7		0.08		0.11		65.66%		71.06%	

4.1.3.2 Best Performance Model of this Phase

Out of 29 data experiments based on different attribute selection, different algorithm, and different algorithm settings, the best performance linear model is shown below; the model is shifted for confidentiality purposes.

```
Scaffold Mhrs/Trade Mhrs =

      0.058  * Trades=ME,PF,EL,FP,INSUL +
      0.1179 * Trades=PF,EL,FP,INSUL +
      0.1631 * Trades=INSUL +
     -0.1438 * Area=Large, Small, Congestion, +
      0.0809 * Area=Small, Congestion +
      0.1517 * General=Heater, One big high structure, rest
low piperacks around +
     -0.1027 * Elevation=Mix +
     -0.0022 * H_Max +
      0.0006 * V_StDev +
     -0.0253 * V_Min +
     -0.2368 * Trade Mhr Distribution +
      1.6842
```

Figure 22 One linear model of the best performance experiment

This model is trained by Linear Regression with Greedy Selection, on an 11 attribute selection. These eleven attributes are: Trade, Area, General, Elevation, H_Mean, H_Max, H_Mode, V_Mean, V_StDev, V_Min, and Trade Mhr Distribution. While in the model, only Trade, Area, General, Elevation, H_Max, V_StDev, V_Min and Trade Mhr Distribution were chosen to contribute to this model. The performance of this linear model is shown in the table below.

Table 6 Performance of best linear regression model

Performance				
Correlation Coefficient	**Error**			
	Mean Absolute	**Root Mean Squared**	**Relative Absolute**	**Root Relative Squared**
0.7461	0.0777	0.1094	63.9310%	66.4006%

Similar to visualization of attributes, WEKA provides a function to plot all the results onto a two dimensional area with one axis presenting the actual value, the other axis

presenting the predicted value. If the predicted value is close to the actual one, then the value point shown in the figure will be a small cross, likewise, the bigger the difference between the actual class value and the predicted one, the bigger the cross shown in the plot. For example, the figure bellow shows the error visualization of the best performance linear model shown in figure 22. In both upper right hand corner and upper left corner, some big crosses were discovered, which presented the actual class value was big while the predicted value was small or the reverse correspondingly.

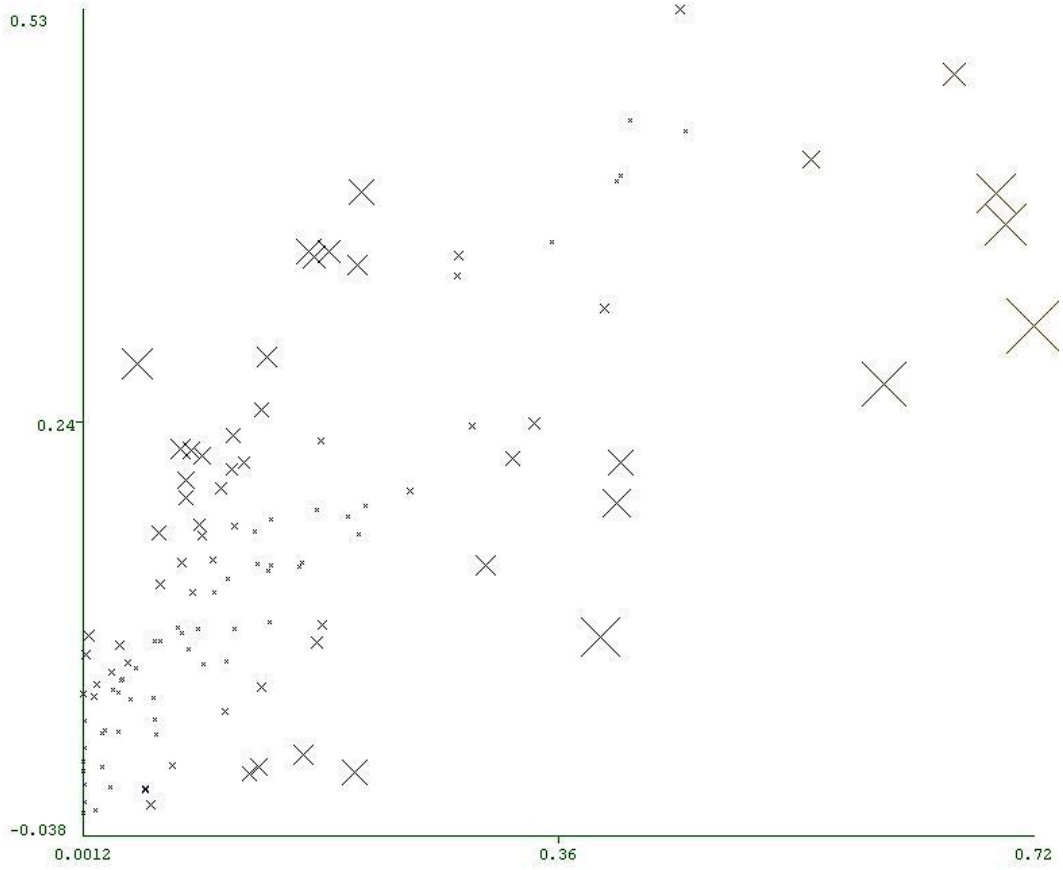


Figure 23 Error visualization of the best performance linear model from phase one

4.1.3.3 Summary

From all the data mining experimental data from this phase, several conclusions can be derived: first, it is clear that simple algorithm – linear regression -- shows good performance in this data mining phase, while the more advanced and sophisticated method – Gaussian process didn't surpass the simple linear regression; second, the performance of models trained on different attribute selection, different algorithms, or different algorithm settings vary, but no dramatic fluctuation was discovered among all

the data mining experiments; third, models on the same attribute subsets, while using different computerized attribute selection methods of linear regression, don't observe a significant difference of performance; forth, quality of data varies from attribute to attribute, some attributes contains quite a lot of noise data. When this type of data was involved in the modeling process, it won't help to improve the model structure or increase the results of performance. Thus, attributes contains too much noise data made no contribution to the building of the model.

4.2 Phase Two – Data Mining Investigation on Modified Input Table

Based on the agreement of the basic work of the first phase -- initial data mining investigation, experts from the client company proposed some suggestions to improve the initial input table. Section 4.2.1 provides the detailed explanation of all the modifications made to the initial input table; 4.2.2 presents the results of the data mining investigation using the modified input table based on experts' selected attributes subset; 4.2.3 presents the other experimental data based on the modified input table.

4.2.1 Input Table Change

The experts from the client company were content with the frame of this input table. Advice was proposed focusing on details regarding to data presenting format, and two more major factors which might significantly affect the structure and result of the model.

First, changes have been done to the four columns describing features of the construction areas. The four attributes are changed from "Area," "General," "Elevation," and "Scaffold Type" to "Area_Size," "Area_ Complexity," "Area_ Congestion Degree," and "Area_ Distance to Material Yard." Thus, features like location of specific construction area, which provides the information of how far each construction area is from the scaffold material yard is, is added to the input table. The distance between the actual scaffold activities location and the scaffold material storage yard is an important factor mentioned by scaffold superintendent on site.

Specifically, "Area_Size" describes the size of each construction area, the values of this attribute are "Large," "Medium," and "Small"; "Area_Complexity" describes the complex level of the scaffold work, the values for this attribute contain "Simple," which represents lower height, straightaway scaffold work, and "Complex" for high level scaffolds or scaffolds that need to go around vessels; "Area_Congestion Degree" describes the congestion level, the values for this attributes contain "Congested," "Less

Congested," and "Not Congested"; "Area_Distance to Material Yard" contains three values, "Far" for those construction areas that are furthest from the storage area, "Close" for construction areas located closest to the storage yard , and "Medium" for the construction areas in between.

These changes to the table make each column (attribute) specified to one feature of a construction area. In addition, the values of these added attributes are more general and easy to apply to future construction projects.

The two tables below show the comparison between the original table and the modified table based on these five attributes.

Table 7 Original table of the feature attributes

Construction Areas	Feature			
	Area	**General**	**Elevation**	**Scaffold Type**
A Charge Pumps		Pumps		Access
B Catalyst Handling	Large, Complex,	Piperack, Mixture of pumps	High level	
C S/D Cooler / Catalyst Slop Oil / Short Circ Flash Drum	Small, Congestion		Low level	Access
D Make-up Hydrogen Compressors		One big building		Heavy
E Electrical Substation		One big building		
F First LCF Reactors		Reactor	High level	
G Hydrotreater Reactor		Reactor	High level	
H LCF Feed/HDT Hydrogen Heaters		Heater	High level	
J Second LCF Reactors		Reactor	High level	
K Reactor Exchangers	Large		High level	
L Membrane	Small, Congestion		Low level	Access
M Amine	Small, Congestion		Low level	Access
N Heavy Oil Stripper	Small, Congestion	One big high structure, rest low piperacks around	Mix	
P Stabilizer	Small, Congestion	One big high structure, rest low piperacks around	Mix	
Q Flare Drum	Small, Congestion	One big high structure, rest low piperacks around	Mix	

R Piperack				decks
S Stripper O/H Sour Gas/Stabilizer O/H Compr.	Small, Congestion	One big high structure, rest low piperacks around	Mix	
T Depropanizer	Small, Congestion	One big high structure, rest low piperacks around	Mix	

Table 8 Modified table of the construction area feature

Construction Areas	**Feature**			
	Area_ Size	Area_ Complexity	Area_ Congestion Degree	Area_ Distance to Material Yard
A Charge Pumps	Medium	Simple	Less Congested	Close
B Catalyst Handling	Large	Complex	Less Congested	Close
C S/D Cooler / Catalyst Slop Oil / Short Circ Flash Drum	Small	Simple	Congested	Close
D Make-up Hydrogen Compressors	Small	Simple	Less Congested	Close
E Electrical Substation	Small	Simple	Not Congested	Medium
F First LCF Reactors	Large	Complex	Less Congested	Medium
G Hydrotreater Reactor	Large	Complex	Less Congested	Far
H LCF Feed/HDT Hydrogen Heaters	Large	Complex	Less Congested	Far
J Second LCF Reactors	Large	Complex	Less Congested	Medium
K Reactor Exchangers	Large	Complex	Less Congested	Far

L Membrane	Small	Simple	Congested	Medium
M Amine	Small	Simple	Congested	Medium
N Heavy Oil Stripper	Small	Complex	Congested	Far
P Stabilizer	Small	Complex	Congested	Far
Q Flare Drum	Small	Complex	Congested	Far
R Piperack	Large	Simple	Congested	Medium
S Stripper O/H Sour Gas/Stabilizer O/H Compr.	Small	Complex	Congested	Far
T Depropanizer	Small	Complex	Congested	Far

Second, changes were made to height and volume representation format. Originally, height information was kept in the database as sea level, which is the format used in the first phase data mining input table. After the meeting discussing the results from the initial data mining investigation, the experts from the client company asked to change that number into direct height, which will serve better in future projects. In addition, in order to show value of height and volume information from each scaffold request, a weight was added to the calculation of height and volume. Weight is calculated as below:

$$weight = \frac{Scaffold\ Mhrs\ of\ each\ scaffold\ request}{Total\ scaffold\ Mhrs\ in\ this\ construction\ area}$$

This weight is multiplied to height and volume attributes of each piece of scaffold request. Then as for height and volume, mean, standard deviation, maximum, minimum, and mode of elevation are calculated for each construction area.

Another factor that was mentioned by the experienced scaffold superintendent during the interview is winter factor. Generally speaking, winter in the northern part of Alberta may reduce productivity by 40% or even more. This percentage might be even higher as far as scaffolding trade is concerned, because scaffold trades are usually most exposed to the harsh conditions. Thus, "Winter factor" attribute was added to the input table to show the winter inefficiency. Assuming winter period starts from 15th November the previous year, and ends on 15th March next year, "Winter factor" is presented as a percentage, which is calculated as scaffold man-hours spent in winter period divided by the total scaffold man-hours for each trade in each construction area.

Another factor that might play a role in scaffold man-hour usage is day-shift and night-shift ratio. Due to the light conditions, weather, and internal biological clock of a human body, night-shift is considered to be less efficient than the day-shift. Thus a new attribute has been added to the input table to show the percentage of day-shift and night-shift. The shift information was abstracted from the pay roll sheet, connecting to trade and construction area by tracking the start and finish data of the scaffold request from "*tbRequest*" scaffold database.

$$ratio = \frac{day - shift\ scaffold\ man - hours}{all\ shifts\ scaffold\ man - hours}$$

Thus, all the modification to the input table was finished. The parts of the modified input table are shown below. Due to the size of the table, it is broken down to three parts.

Table 9 Part of the modified input table.

A	B	C	D	E	F	G	H
Construction Areas	**Trades**	**Area_Size**	**Area_Complexity**	**Area_Congestion Degree**	**Area_Distance to Material Yard**	**H_Mean**	**H_StDev**
A Charge Pumps	CIV	Medium	Simple	Less Congested	Close	0.014602205	0.017077935
A Charge Pumps	EL	Medium	Simple	Less Congested	Close	0.014602205	0.017077935
A Charge Pumps	FP	Medium	Simple	Less Congested	Close	0.014602205	0.017077935
A Charge Pumps	INSTR	Medium	Simple	Less Congested	Close	0.014602205	0.017077935
A Charge Pumps	INSUL	Medium	Simple	Less Congested	Close	0.014602205	0.017077935
A Charge Pumps	IW	Medium	Simple	Less Congested	Close	0.014602205	0.017077935
A Charge Pumps	ME	Medium	Simple	Less Congested	Close	0.014602205	0.017077935
A Charge Pumps	PF	Medium	Simple	Less Congested	Close	0.014602205	0.017077935

I	J	K	L	M	N	O	P	Q
H_Max	**H_Min**	**H_Mode**	**V_Mean**	**V_StDev**	**V_Max**	**V_Min**	**V_Mode**	**Count**
0.136099867	0.000795906	0.005969292	0.047660866	0.055741524	0.444223158	0.002597796	0.019483472	1
0.136099867	0.000795906	0.005969292	0.047660866	0.055741524	0.444223158	0.002597796	0.019483472	206

0.13609 9867	0.00079 5906	0.00596 9292	0.04766 0866	0.05574 1524	0.44422 3158	0.00259 7796	0.01948 3472	39
0.13609 9867	0.00079 5906	0.00596 9292	0.04766 0866	0.05574 1524	0.44422 3158	0.00259 7796	0.01948 3472	16
0.13609 9867	0.00079 5906	0.00596 9292	0.04766 0866	0.05574 1524	0.44422 3158	0.00259 7796	0.01948 3472	50
0.13609 9867	0.00079 5906	0.00596 9292	0.04766 0866	0.05574 1524	0.44422 3158	0.00259 7796	0.01948 3472	15
0.13609 9867	0.00079 5906	0.00596 9292	0.04766 0866	0.05574 1524	0.44422 3158	0.00259 7796	0.01948 3472	16
0.13609 9867	0.00079 5906	0.00596 9292	0.04766 0866	0.05574 1524	0.44422 3158	0.00259 7796	0.01948 3472	478

R	S	T	U	V	W	X
Scaffold Mhrs	**Trade Mhrs**	**Scaffold Distribution**	**Trade Mhr Distribution**	**Day/Night-shift Ratio**	**Winter Factor**	**Scaffold Mhrs/Trade Mhrs**
125.61808	50943. 04747	0.004676 858	0.17867311 9	1.0767264	0	0.0027656 76
8803.3599 1	87611. 73744	0.327755 912	0.30728162 5	0.9982151	0.3533853 92	0.1126990 6
2440.0190 45	11380. 99298	0.090843 8	0.03991668 4	0.9757833	0	0.2404624 06
319.09235 5	30740. 765	0.011880 056	0.10781742 8	0.9757833	0.0867310 37	0.0116422 21
1336.3744 85	7630.4 85411	0.049754 258	0.02676248 7	0.9757833	0.1496214 77	0.1964310 5
708.84488	10313. 02005	0.026390 844	0.03617097	0.9869992	0.3289530 73	0.0770902 53
506.95868	8904.3 479	0.018874 465	0.03123031 9	0.9869992	0.5377320 02	0.0638564 21
15885.079	112261	0.591413	0.39373736	0.9757833	0.2649011	0.1587052

17	.8866	808	8		2	07

Besides the input table modification, the experts also suggested two subsets of the attributes which correspondingly addressing two different situations. One situation sets at the starting point of a project, which fits the original objective of this research to estimate the scaffold man-hours; the second situation would be during the process of a project, the project manager periodically running the model to get a timely scaffold control and manager. The first scenario is focusing on estimation at the beginning, while the second scenario focuses on project management and controlling.

a. For estimation purpose, a subset of 14 attributes are listed below:
   Trades
   Area_Size
   Area_Complexity
   Area_Congestion Degree
   Area_Distance to Material Yard
   H_Mean
   H_StDev
   H_Max
   H_Min
   H_Mode
   Trade Mhrs
   Day/Night-shift Ratio
   Winter Factor
   Trade Mhr Distribution
b. For project controlling purpose, the 19 attributes selected are listed below:
   Trades
   Area_Size
   Area_Complexity
   Area_Congestion Degree
   Area_Distance to Material Yard
   H_Mean
   H_StDev
   H_Max

H_Min

H_Mode

V_Mean

V_StDev

V_Max

V_Min

V_Mode

Trade Mhrs

Day/Night-shift Ratio

Winter Factor

Trade Mhr Distribution

The difference between these attribute subsets is whether volume information is involved. Since volume information is a general measurement of existing scaffolding, at beginning of a project, this type of information is unavailable. However, during the project, this type of information could be an important factor to the results, and the class for both situations remained the same as a quotient – Scaffold Mhrs/Trade Mhrs.

4.2.2 Experimental Results of Experts Required Attributes Subsets

This section contains the experimental results for the specific attributes selection of the client expert. In order to get the performance of the trained model on the unknown data sets, 10-fold cross validation test mode was used in this phase of data mining experiment. Section 4.2.2.1 presents the results and linear models of these two experiments, which is followed by a brief summary in Section 4.2.2.2.

4.2.2.1 Experimental Results

Based on these two specified attribute selections, both linear regression with no attribute selection and Gaussian process normalized PolyKernel function with an exponent of 2 were tried. The table below shows the performance on the experts required attribute subsets.

Table 10 Performance of two experiments based on required attributes selection for estimating and controlling purpose

Input		Test Mode	Algorithm	Performance				
				Correlation Coefficient	Error			
Data set	Attributes				Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared
Modified Input	14 attributes	10-fold cross validation	Linear Regression	0.7361	0.0864	0.1132	71.1209%	68.6713%
Modified Input	14 attributes	10-fold cross validation	Gaussian Process	0.6868	0.0817	0.1268	67.2072%	76.9366%
Modified Input	19 attributes	10-fold cross validation	Linear Regression	0.7259	0.0874	0.1162	71.9636%	70.5224%
Modified Input	19 attributes	10-fold cross validation	Gaussian Process	0.736	0.0799	0.1142	65.7180%	69.2935%

According to the experimental data shown in the above table, Linear Regression models show no worse performance than models trained from Gaussian Process. In order to eliminate the stochastic factors, a series of experiments repeating ten times using 10-fold cross validation of the same attributes selection between Linear Regression and Gaussian Process have been done. The table below shows these experimental results of the two required attribute selections.

Table 11 Results of statistical comparison between linear regression and Gaussian process on modified input table based on selected attributes

Experiments Inputs	Algorithms	Result									
		Correlation Coefficient		Mean Absolute		Root Mean Squared		Relative Absolute		Root Relative Squared	
Modified_Attri14	GaussianProcess_NormalizedPolyKernel_E2.0	0.72	Equal	0.08	Equal	0.11	Equal	67.98%	Equal	69.15%	Equal
	LinearRegression_NO attribute selection	0.72		0.09		0.11		74.80%		74.39%	
Modified_Attri19	GaussianProcess_NormalizedPolyKernel_E2.0	0.72	Equal	0.08	Equal	0.11	Equal	67.98%	Equal	69.15%	Equal
	LinearRegression_NO attribute selection	0.73		0.09		0.11		74.80%		74.39%	

The following figures show the linear models trained from the experiments for estimate or controlling, as well as their corresponding error visualization plot from WEKA.

```
Scaffold Mhrs/Trade Mhrs =
0.1691 * Trades=INSTR,IW,ME,PF,EL,FP,INSUL +
-0.053  * Trades=IW,ME,PF,EL,FP,INSUL +
0.065  * Trades=ME,PF,EL,FP,INSUL +
0.1754 * Trades=PF,EL,FP,INSUL +
-0.0365 * Trades=EL,FP,INSUL +
-0.0143 * Trades=FP,INSUL +
0.1655 * Trades=INSUL +
0.0092 * Area_Size=Large,Small +
-0.0093 * Area_Size=Small +
0.0381 * Area_Complexity=Simple +
0.0501 * Area_Congestion Degree=Less Congested ,Congested +
0.1208 * Area_Congestion Degree=Congested +
-0.0287 * Area_Distance to Material Yard=Medium,Far +
0.0707 * Area_Distance to Material Yard=Far +
-2.5714 * H_Mean +
0.1982 * H_Max +
1.8249 * H_Min +
2.3154 * H_Mode +
0      * Trade Mhrs +
-0.4221 * Trade Mhr Distribution +
2.1084 * Day/Night-shift Ratio +
0      * Winter Factor +
-2.0775
```

Figure 24 Linear model for estimating purpose based on 14 attribute subset

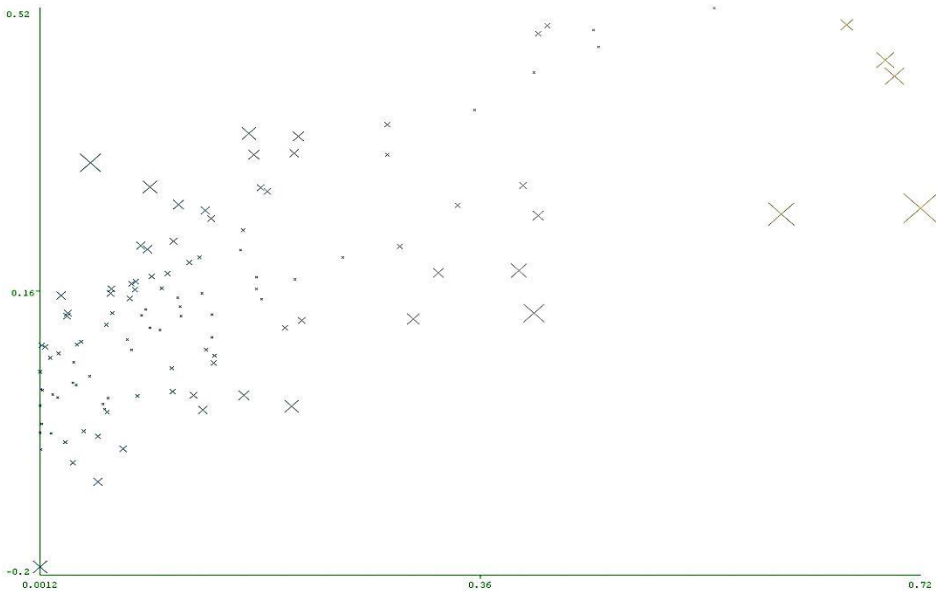


Figure 25 Error visualization of the linear model for estimation purpose

```
Scaffold Mhrs/Trade Mhrs =
      0.1513 * Trades=INSTR,IW,ME,PF,EL,FP,INSUL +
     -0.0437 * Trades=IW,ME,PF,EL,FP,INSUL +
      0.0591 * Trades=ME,PF,EL,FP,INSUL +
      0.1681 * Trades=PF,EL,FP,INSUL +
     -0.0331 * Trades=EL,FP,INSUL +
     -0.0182 * Trades=FP,INSUL +
      0.1733 * Trades=INSUL +
      0.0357 * Area_Size=Large,Small +
     -0.0391 * Area_Size=Small +
      0.0256 * Area_Complexity=Simple +
     -0.71   * Area_Congestion Degree=Less
Congested ,Congested +
      0.108  * Area_Congestion Degree=Congested +
     -0.0376 * Area_Distance to Material Yard=Medium,Far +
      0.0513 * Area_Distance to Material Yard=Far +
     -1.8835 * H_Mean +
      0.7076 * H_StDev +
     -0.2767 * H_Max +
     41.351  * H_Min +
      2.2717 * H_Mode +
      0.0775 * V_Max +
     -7.2439 * V_Min +
      0      * Trade Mhrs +
     -0.4303 * Trade Mhr Distribution +
      1.7142 * Day/Night-shift Ratio +
      0      * Winter Factor +
     -2.9484
```

Figure 26 The linear model for controlling purpose based on 19 attribute subset

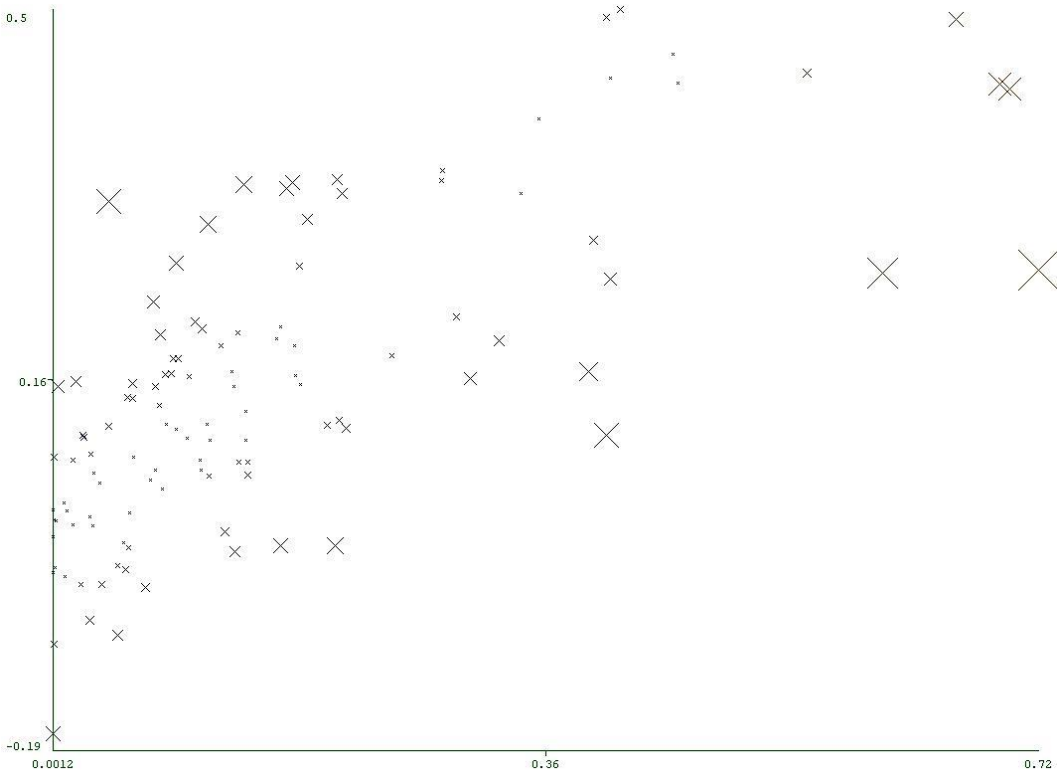


Figure 27 Error visualization of the linear model for controlling purpose

4.2.2.2 Summary

Though intention of the experts of choosing two sets of attributes to tackle different situations is beneficial, the data mining experimental result didn't observe an improvement. First, from above tables and figures, which accurately and graphically explained evaluation of these experiments based on experts' attributes selections, the performance of models based on the modified input table has no significant improvement. Second, comparing these two models trained by different attribute selections, no substantial difference of performance exists. Specifically, volume information didn't play an important role in the second model, which didn't help increase the accuracy of the performance of the previous model. Third, linear regression performs no worse than other more advanced algorithms (Gaussian process) in this modified input table. Thus, it is proved that the idea of training two different models for different phases in this project is not practical in this research due to the limitation of the data.

4.2.3 Other Experiments Based on Modified Input Table

Following the advice from the client experts, some other experiments were designed and performed systematically based on the modified input table. Section 4.2.3.1 experimental

data includes the general experiment approaches, settings, and results. Section 4.2.3.2 presents the best performance model.

4.2.3.1 Experimental Data

Like experiments done in data mining investigation phase one, the data mining investigation of this phase is aimed at comparing results from different combinations of learning algorithms, setting of algorithms, and subsets of attributes. The design of the experiments in this phase is based on a trial and error method, at the same time considering the experts' advice, and experimental results of the previous data mining experiments.

The experiments' design follows the same rules as in phase one. The most challenging part is attribute selection, which is based on the function of visualization of each attribute from WEKA, the results of existing experimental data, and experts' advice. Attributes visualization is shown in the below two figures to graphically explain the data itself, and help decision making in attribute selections. Please refer to the first line of blocks of each figure. From the left to right, each block present one attribute – Construction Areas, Trades, Area_Size, Area_Complexity, Area_Congestion Degree, Area_Distance to Material Yard, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, and V_StDev for figure 28, and V_Max, V_Min, V_Mode, Count, Scaffold Mhrs, Trade Mhrs, Scaffold Distribution, Day/Night-shift Ratio, Winter Factor, and Trade Mhr Distribution for figure 29.

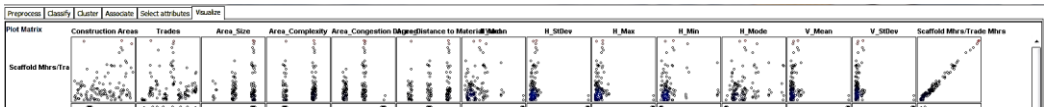


Figure 28 Visualization of correlation of first thirteen attributes and the class

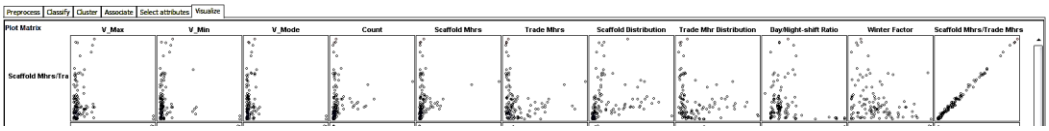


Figure 29 Visualization of correlation of last ten attributes and the class

From these two above figures, all the volume information are scattered all over in the two dimensional area, while cannot discover a certain trend following the 45 degree line. This proved that volume information is not a reliable data type, and won't make a big

contribution to the performance of the models. This proves the conclusions for data mining experiments from Section 4.2.2 that trying to generate two different models for both estimating and controlling purposes is not practical.

In this data mining investigation phase, a total of thirty-two experiments were conducted. However, due to the large size of the table, the full table is attached in Appendix 2. Part of the results is presented in Section 4.3.2.2.

Similar to data mining investigation phase one, some models were built on algorithm other than linear regression to make a comparison. The performance evaluations out of all these different algorithms are close. Thus, in order to eliminate the random factor, a set of separate experiments has been done repeating ten times using 10-fold cross validation test mode to arrive at an accurate comparison between linear regression and other algorithms (Gaussian Process). The table below shows the results of a set of experiments run at ten times 10-fold cross validation method using linear regression and Gaussian Process based on the attribute subsets of two best performance linear models.

Table 12 Results of statistical comparison between linear regression and gaussian process on selected attributes

Experiments Inputs	Algorithms	Result									
		Correlation Coefficient		Mean Absolute		Root Mean Squared		Relative Absolute		Root Relative Squared	
Experiment 4_14 attributes	GaussianProcess_NormalizedPolyKernel_E2.0	0.75	Equal	0.07	Equal	0.1	Equal	65.52%	Equal	67.39%	Equal
	LinearRegression_Greedy Method	0.76		0.08		0.1		61.12%		63.76%	
Experiment 15_4 attributes	GaussianProcess_NormalizedPolyKernel_E2.0	0.75	Equal	0.08	Equal	0.1	Equal	66.31%	Equal	67.81%	Equal
	LinearRegression_NO attribute selection	0.75		0.07		0.1		61.41%		68.89%	

4.3.2.2 Best Performance Model of the Phase

In this data mining investigation phase, a total of thirty-two experiments were conducted. Out of the thirty-two experiments, two stand out by their high performance and the simplicity of the model. These two experiments are all trained on linear regression. Experiment 4 is trained on a 14 attribute selection by linear regression with greedy attribute selection; the model trained out of this experiment contains 5 attributes – "Trade," "Area_Congestion Degree," "Trade Mhrs," "Trade Mhr Distribution," and "D/N Ratio." Experiment 15 is based on a 4 attribute subset and trained by linear regression with no attribute selection. The following table shows the experimental data of these two experiments, followed by two figures showing these two linear models, as well as two figures of error visualization of these two models.

Table 13 Performance of two best performance experiments

Item	Attributes	Test Mode	Algorithm	Performance				
				Correlation Coefficient	Error			
					Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared
4	14 attributes	10-fold-cross validation	Linear Regression	0.7527	0.0802	0.1086	65.9813%	65.9154%
15	4 attributes	10-fold-cross validation	Linear Regression	0.7541	0.0779	0.1085	64.1396%	65.8334%

```
Scaffold Mhrs/Trade Mhrs =

0.1285 * Trades=INSTR,IW,ME,PF,EL,FP,INSUL +
0.1548 * Trades=PF,EL,FP,INSUL +
0.1613 * Trades=INSUL +
0.1115 * Area_Congestion Degree=Congested +
0.0000 * Trade Mhrs +
-0.3419 * Trade Mhr Distribution +
1.6602 * Day/Night-shift Ratio +
-2.5602
```

Figure 30 The linear model from experiment 4

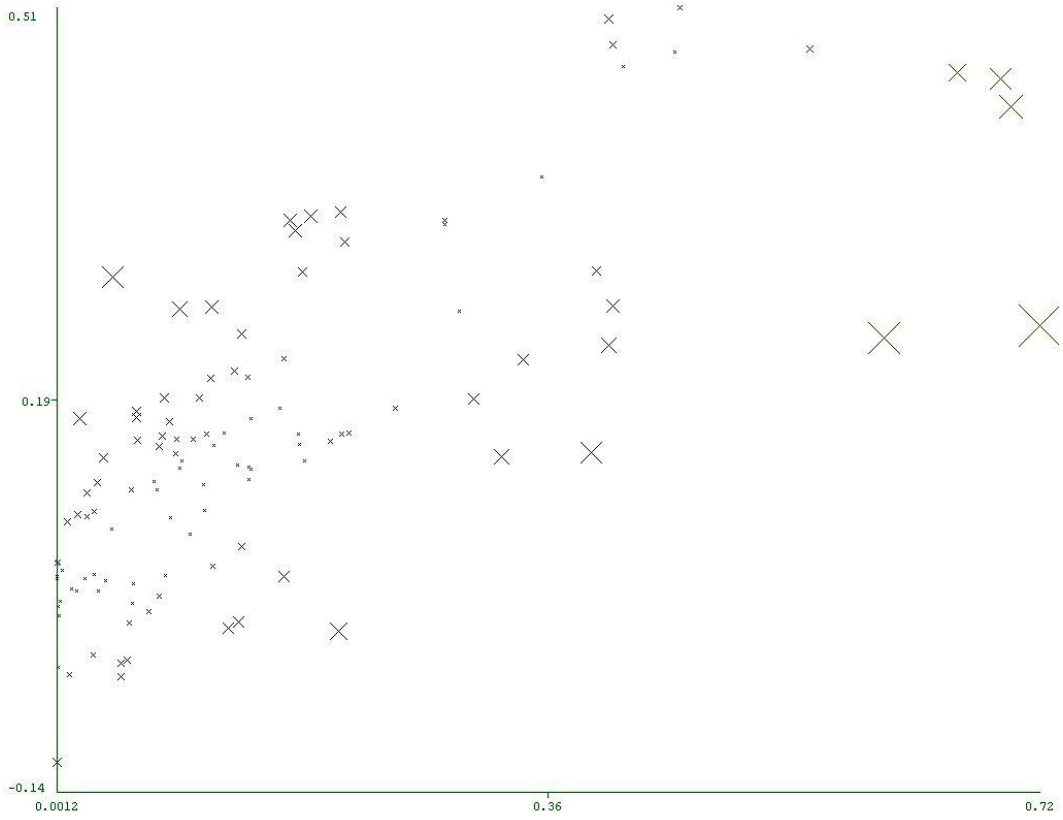


Figure 31 Error visualization of best performance linear model of experiment 4 using 14 attributes subset from phase two

```
Scaffold Mhrs/Trade Mhrs =

      0.1374 * Trades=INSTR,IW,ME,PF,EL,FP,INSUL +
     -0.0498 * Trades=IW,ME,PF,EL,FP,INSUL +
      0.0621 * Trades=ME,PF,EL,FP,INSUL +
      0.1955 * Trades=PF,EL,FP,INSUL +
     -0.0465 * Trades=EL,FP,INSUL +
     -0.0312 * Trades=FP,INSUL +
      0.1678 * Trades=INSUL +
     -0.0468 * Area_Congestion Degree=Less Congested,
Congested +
      0.097  * Area_Congestion Degree=Congested +
     -0.3643 * Trade Mhr Distribution +
      1.7706 * Day/Night-shift Ratio +
     -3.5897
```

Figure 32 the linear model from experiment 15

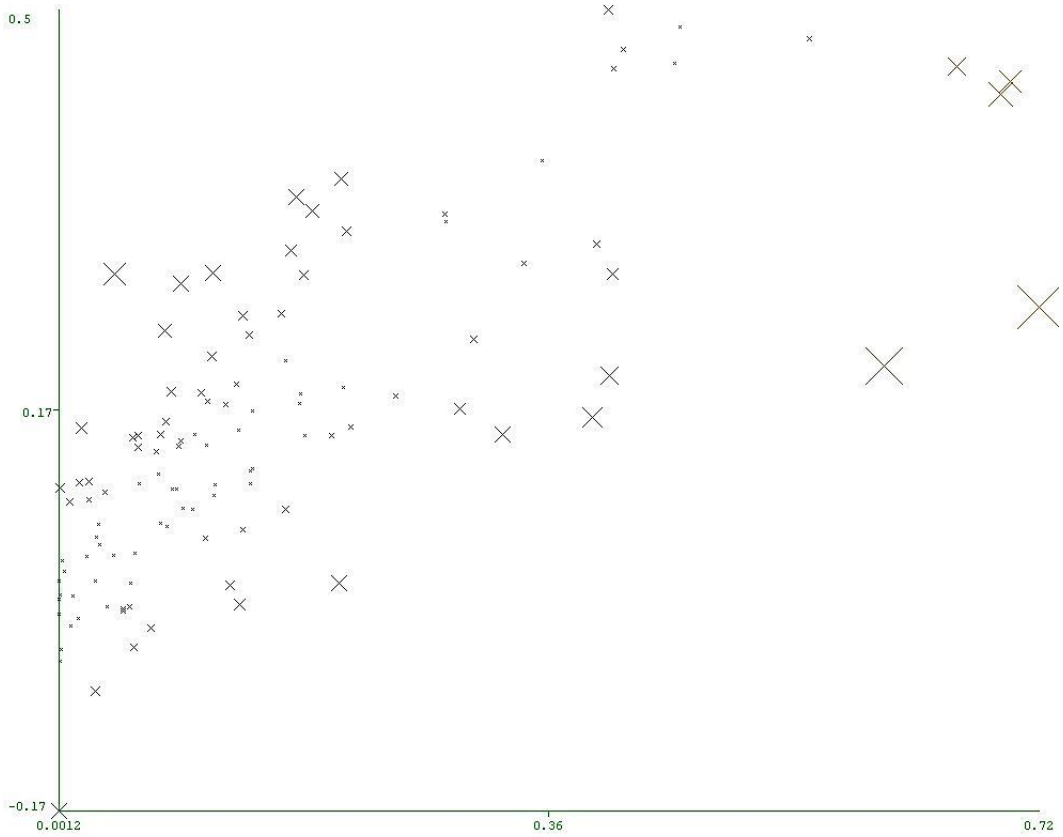


Figure 33 Error visualization of best performance linear model of experiment 15 using 4 attributes subset from phase two

It is clear that no matter the numerical measurement of the performance of these two models, or the figures showing error visualization, models trained from modified input table don't improve significantly compared to the models trained from initial data mining investigation phase.

4.2.3.3 Error Evaluation on the Best Performance Model

In order to show the performance of the model on the original training data, one of the best performance linear models from Experiment 4 is built in Excel to reveal the error evaluation. Then the predicted scaffold man-hours are grouped into each construction area as well as each trade to elaborate the error rate on a higher level. Table 15 below shows the error rate of the linear model trained from Experiment 4 on construction area level, followed by table 16 elaborating error rate of this model on trade level. The average relative error of this linear model for construction area level is 32.55%., for trade level it is 12.53%.

Table 14 Error rate on construction area level of linear model from experiment 4 based on the training data

Construction Areas	Actual Scaffold Mhrs	Scaffold Mhrs/Trade Mhrs	Predicted Scaffold Mhrs	Error	
				Absolute Difference	Relative
A Charge Pumps	23907.36	8.39%	25970.79	2063.44	8.63%
B Catalyst Handling	34198.57	10.91%	34114.25	84.32	0.25%
C S/D Cooler / Catalyst Slop Oil / Short Circ Flash Drum	7208.39	11.12%	10060.52	2852.13	39.57%
D Make-up Hydrogen Compressors	17114.82	4.84%	27354.99	10240.17	59.83%
E Electrical Substation	1558.99	2.26%	0.00	1558.99	100.00%
F First LCF Reactors	29952.62	10.14%	31571.48	1618.86	5.40%
G Hydrotreater	16010.04	10.38%	11601.42	4408.62	27.54%

Reactor					
H LCF Feed/HDT Hydrogen Heaters	19877.48	6.28%	20962.10	1084.63	5.46%
J Second LCF Reactors	24832.83	8.14%	33416.29	8583.46	34.56%
K Reactor Exchangers	13726.51	12.73%	3573.00	10153.52	73.97%
L Membrane	3433.07	15.16%	2802.03	631.05	18.38%
M Amine	6917.11	12.66%	5249.70	1667.40	24.11%
N Heavy Oil Stripper	13738.97	10.40%	18481.11	4742.13	34.52%
P Stabilizer	9543.98	15.00%	9884.65	340.67	3.57%
Q Flare Drum	6025.46	12.31%	5811.45	214.01	3.55%
R Piperack	102510.48	27.64%	39776.91	62733.57	61.20%
S Stripper O/H Sour Gas/Stabilizer O/H Compr.	2608.85	8.28%	4806.53	2197.68	84.24%
T Depropanizer	6803.84	10.10%	6878.08	74.24	1.09%

Statistics of Error Rates		
Items	**Abs Difference**	**Relative**
Max	2063.44	100.00%
Mean	84.32	32.55%
Min	2852.13	0.25%

Table 15 Error rate on trade level of linear model from experiment 4 on the training data

Trades	CIV	EL	FP	INSTR	INSUL	IW	ME	PF	Total
Actual Scaffold Mhrs	486.17	95568.66	12907.15	6261.37	23864.72	4681.31	12630.29	177309.33	333709.03
Actual Scaffold Mhrs/Trade Mhrs Ratio	0.18%	11.80%	14.98%	2.76%	28.34%	3.47%	6.08%	14.31%	10.92%
Predicted Scaffold Mhrs	2716.62	85621.70	11442.65	3108.81	20313.34	8169.85	6082.83	154436.03	291891.84
Predicted Scaffold Mhrs/Trade Mhrs Ratio	1.02%	10.57%	13.28%	1.37%	24.12%	6.06%	2.93%	12.46%	9.55%
Abs Scaffold Mhrs Error	2230.45	9946.96	1464.50	3152.57	3551.38	3488.54	6547.46	22873.30	41817.19
Relative Scaffold Mhrs Error	458.78%	10.41%	11.35%	50.35%	14.88%	74.52%	51.84%	12.90%	12.53%
Abs Scaffold/Trade Ratio Error	0.0083	0.0123	0.0170	0.0139	0.0422	0.0259	0.0315	0.0185	0.0137
Relative Scaffold/Trade Ratio Error	458.78%	10.41%	11.35%	50.35%	14.88%	74.52%	51.84%	12.90%	12.53%

4.2.4 Summary

The second phase of data mining investigation is based on an input table with several modifications upon the original input table from phase one according to experts' advice. After a series of data mining experiments on this modified input table, it is clear that: first, some of the experimental results from the modified input table have improved a little bit, but no significant improvement from the results of data mining investigation phase one have been observe; second, very similar to the phase one, simple algorithm linear regression performs as well as more sophisticated one (Gaussian Process), while providing a more simple model for the practical field; third, the quality of data is not good enough to realize experts' proposal of training two different models for different purposes; it is impractical, specifically, information of volume is unreliable and contains too much noise.

4.3 Phase Three – Data Mining Investigation with "Scaffold Man-hour" as Class

Instead of making direct changes to the input table, the third phase of data mining investigation is conducted forward from the results and discoveries of the previous two phases of data mining investigations. Specifically, based on the modified input table from phase two, data mining experiments simply replaced the class from "Scaffold Mhrs/Trade Mhrs Ratio" to directly predict "Scaffold Mhrs."

Section 4.3.1 reveals the experimental data, and Section 4.3.2 provides the error evaluation on the original training data from two of the best performance models. Section 4.3.3 presents a brief summary of this data mining investigation phase.

4.3.1 Experimental Data

Based on the experimental results from the previous two phases of data mining investigation, other than algorithms from function group, some algorithms out of decision trees have been tried in this data mining investigation phase. Out of all the decision trees, M5P tree suits this case most, thus it was selected in this phase to compare with linear regression. The design of experiments is on a trial and error method, trying possible combinations of learning algorithms, and subsets of attributes. These combinations are chosen based on previous experimental data, experts' advice, as well as visualization of the data set itself.

A total of fifteen experiments were done in this phase. The experimental data indicates that models that directly predict scaffold man-hours have better performance than the models that predict a ration of scaffold man-hours out of direct trade man-hours. Specifically, first, no matter if the model is trained by linear regression or decision tree, the correlation coefficient increased; second, the error evaluation decreased from the models trained by M5P trees, while the error stays at the same level from the models trained by linear regression. A table with the full results for this data mining investigation phase is attached in Appendix 3.

Out of all the experiments, two stand out by their better performance. One of these two experiments is trained on linear regression; the other is trained from M5P tree. The experimental data of these two experiments is shown in the below table 16 followed by figures 34 and 37 elaborating these two models and figures 35 and 38 showing error visualizations. In addition, figure 36 shows the structure of the tree model from M5P. Though both of them have an equal performance in correlation coefficient, the model trained by M5P tree shows a lower error rate than the model trained from linear regression. Thus, to get a better comparison between these two models without the influence of the random factor, a set of experiments repeating ten times of 10-fold cross validation on both attribute subsets have been conducted to get the accurate error estimate of the comparison between the linear model and the M5P tree model. The results are shown in table 17.

```
Scaffold Mhrs =
2716.5468 * Trades=IW,INSTR,ME,FP,INSUL,EL,PF +
   -381.1992 * Trades=INSTR,ME,FP,INSUL,EL,PF +
    542.6261 * Trades=ME,FP,INSUL,EL,PF +
    -20.0541 * Trades=FP,INSUL,EL,PF +
   1694.3034 * Trades=INSUL,EL,PF +
   -487.4319 * Trades=EL,PF +
   1963.579  * Trades=PF +
   4415.6657 * Area_Size=Medium, Large +
   1669.2234 * Area_Size=Large +
   1568.9912 * Area_Complexity=Simple +
   1216.2461 * Area_ Congestion Degree=Congested, Less
Congested +
  -8366.8615 * Area_ Congestion Degree=Less Congested  +
      0.1888 * Trade Mhrs +
 -10460.01   * Trade Mhr Distribution +
  -3922.2761
```

Figure 34 The linear model from experiment 12

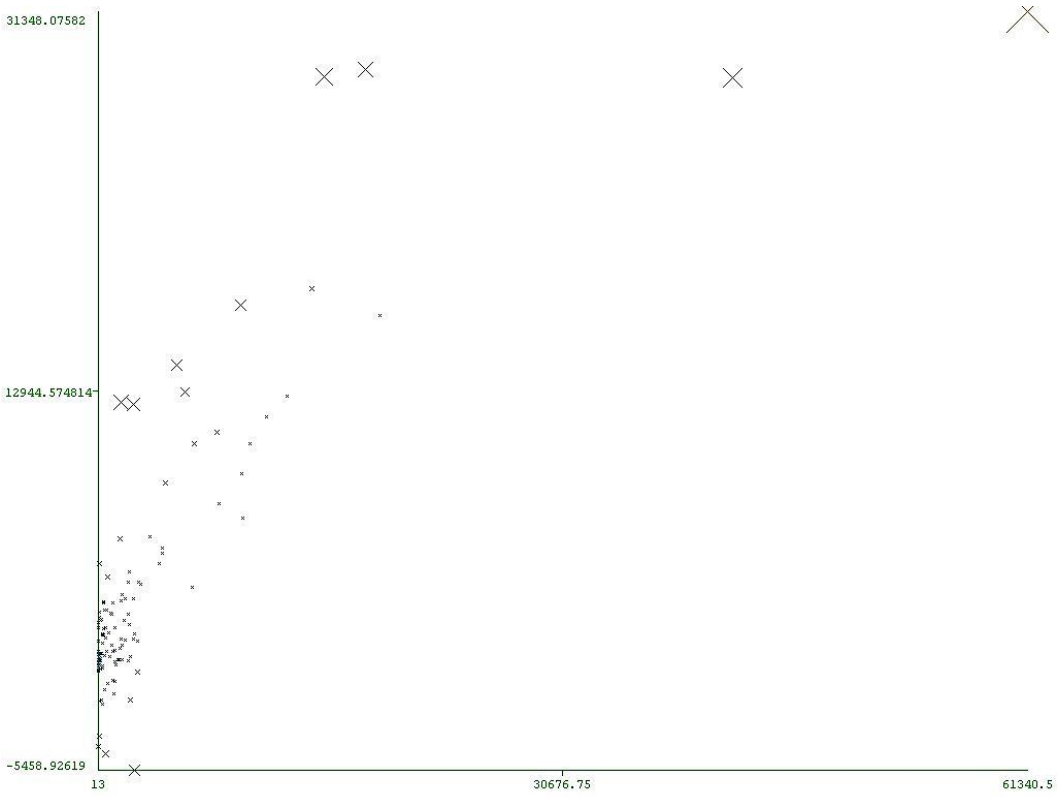


Figure 35 Error visualization of best performance linear model of experiment 12 using 6 attributes subset

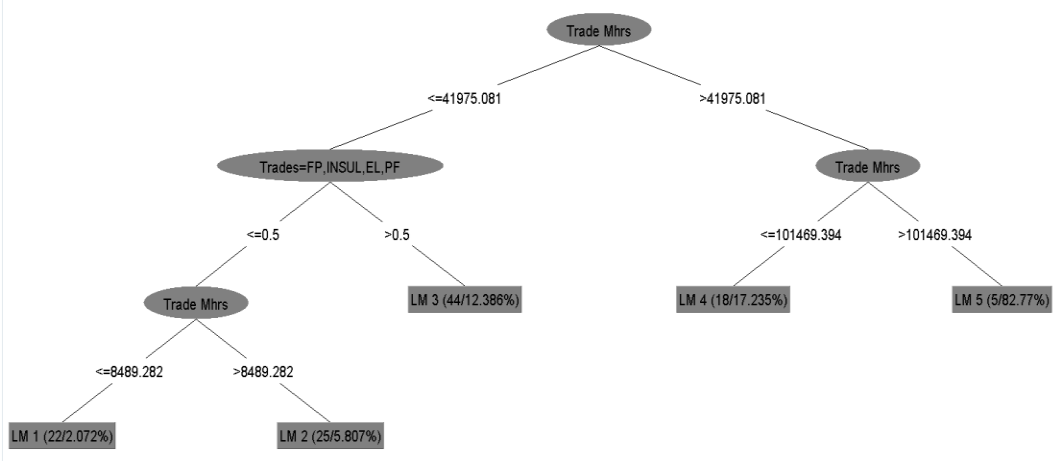


Figure 36 Structure of M5P tree model from experiment 14

```
Trade Mhrs <= 41975.081 :
|   Trades=FP,INSUL,EL,PF <= 0.5 :
|   |   Trade Mhrs <= 8499.282 : LM1 (22/2.072%)
|   |   Trade Mhrs >  8490.282 : LM2 (25/5.807%)
|   Trades=FP,INSUL,EL,PF >  0.5 : LM3 (44/12.386%)
Trade Mhrs >  41975.081 :
|   Trade Mhrs <= 101469.394 : LM4 (18/17.235%)
|   Trade Mhrs >  101469.394 : LM5 (5/82.77%)
LM num: 1
Scaffold Mhrs =
      530.7008 * Trades=FP,INSUL,EL,PF
      + 319.8098 * Trades=PF
      + 1011.3334 * Area_Size=Medium,Large
      - 1383.9984 * Area_Congestion Degree=Less Congested
      + 0.0421 * Trade Mhrs
      - 2411.0761 * Trade Mhr Distribution
      + 2216.8287
LM num: 2
Scaffold Mhrs =
      530.7008 * Trades=FP,INSUL,EL,PF
      + 319.8098 * Trades=PF
      + 1011.3334 * Area_Size=Medium,Large
      - 1383.9984 * Area_ Congestion Degree=Less Congested
      + 0.0418 * Trade Mhrs
      - 3970.0958 * Trade Mhr Distribution
      - 189.9289 * Winter Factor
      + 3609.5282
LM num: 3
Scaffold Mhrs =
      546.6437 * Trades=FP,INSUL,EL,PF
      + 319.8098 * Trades=PF
      + 1623.6375 * Area_Size=Medium,Large
      - 1988.3262 * Area_ Congestion Degree=Less Congested
      + 0.1016 * Trade Mhrs
      - 1721.4368 * Trade Mhr Distribution
      + 5559.571
LM num: 4
Scaf
Scaffold Mhrs =
      1955.0379 * Trades=ME,FP,INSUL,EL,PF
      + 605.7551 * Trades=FP,INSUL,EL,PF
      + 2065.8586 * Trades=PF
      + 2322.0497 * Area_Size=Medium,Large
      - 3383.4452 * Area_ Congestion n Degree=Less
Congested
      + 0.1795 * Trade Mhrs
      - 4801.9027 * Trade Mhr Distribution
      - 2846.0171 * Winter Factor
      - 3963.0914
LM num: 5
Scaffold Mhrs =
```

```
      605.7551 * Trades=FP,INSUL,EL,PF
      + 892.1009 * Trades=PF
      + 2322.0497 * Area_Size=Medium,Large
      + 5225.7349 * Area_Complexity=Simple
      - 3383.4452 * Area_ Congestion Degree=Less Congested
      + 0.2062 * Trade Mhrs
      - 4801.9027 * Trade Mhr Distribution
      - 3999.121
```

Figure 37 the M5P tree model from experiment 14

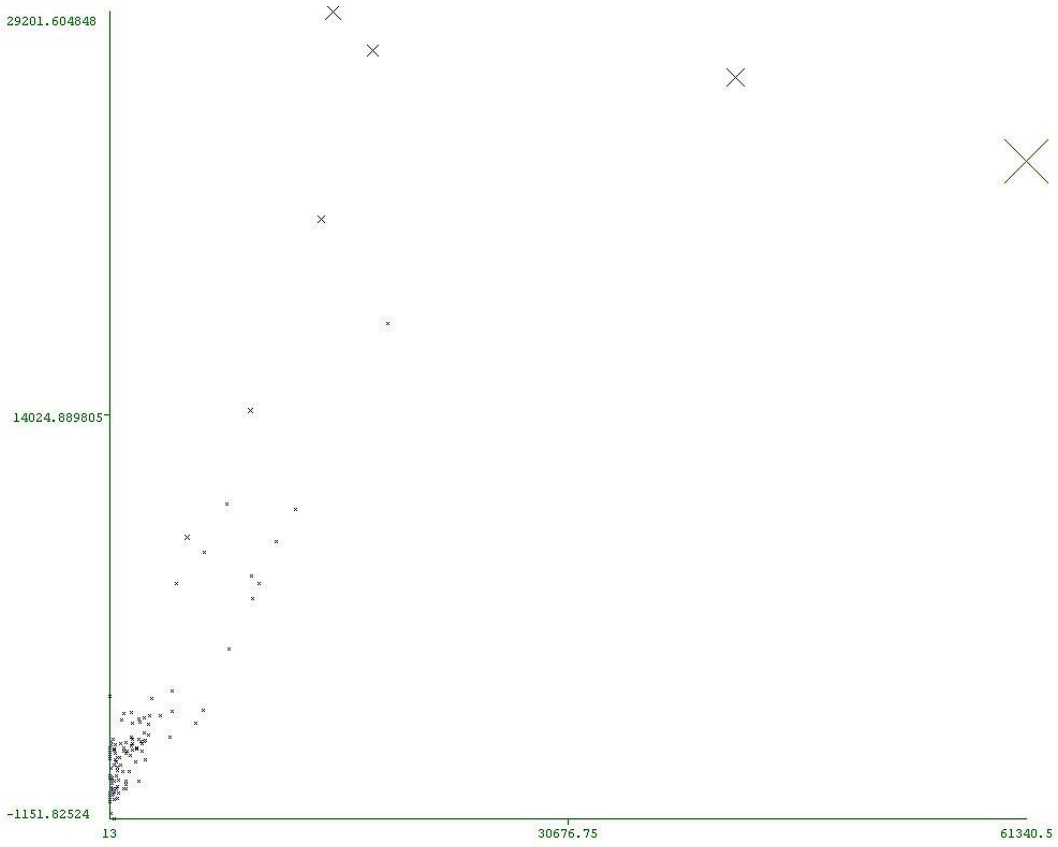


Figure 38 Error visualization of best performance M5P tree model of experiment 14 using 7 attributes subset

Table 16 Performance of two best performance experiments

Item	Attributes	Test Mode	Algorithm	Performance				
				Correlation Coefficient	Error			
					Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared
12	6 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_ Congestion Degree; Trade Mhrs; Trade Mhr Distribution;	10-fold-cross validation	Linear Regression	0.8153	2366.9857	4401.447	62.2972%	57.9321%
14	7 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_ Congestion Degree; Trade Mhrs; Trade Mhr Distribution; Winter Factor	10-fold-cross validation	M5P	0.8159	1634.9469	4397.5586	43.0305%	57.8809%

Table 17 Results of statistical comparison between linear regression and M5P trees on experiment 12 and 14

Experiments Inputs	Algorithms	Result									
		Correlation Coefficient		Mean Absolute		Root Mean Squared		Relative Absolute		Root Relative Squared	
Experiment 12: 7 ATTRITUBES	M5P	0.86	Equal	1754.09	Better	3291.05	Equal	44.53%	Better	57.55%	Equal
	Linear Regression NO attribute selection	0.81		2382.37	Worse	3784.64		66.63%	Worse	75.74%	
Experiment 14: 6 ATTRITUBES	M5P	0.86	Equal	1801.18	Better	3430.99	Equal	45.72%	Better	58.77%	Equal
	Linear Regression NO attribute selection	0.81		2402.46	Worse	3803.44		67.20%	Worse	76.20%	

4.3.2 Error Evaluation on the Best Two Performance Models

In order to clearly explain the performance of these two models trained by different algorithms, both linear model and M5P tree models are built in Excel. Then the predicted scaffold man-hours from these two models on the original training data are grouped into construction area level, as well as trade level to elaborate the error rate. Table 18 below shows the error rate of the linear model trained from Experiment 12 on construction area level, which is followed by table 19 which reveals the error rate on a trade level. The average relative error of this linear model for construction area level is 36.99%, for trade level it is 8.47%. Table 20 below shows the error rate of the M5P tree model trained from Experiment 14 on construction area level, which is followed by table 21 which reveals the error rate on a trade level. The average relative error of this linear model for construction area level is 21.06%, for trade level it is 5.26%. The error rates from the best performance M5P tree model witnessed a noticeable drop from the best linear model from phase two, where the best model provides an average 32.55% relative error on construction area level, and 12.53% relative error on trade level. Additional, this error rate from best M5P tree model is acceptable for the client company.

Table 18 Error rate on construction area level of linear model from experiment 12 based on the training data

Construction Areas	Actual Scaffold Mhrs	Scaffold Mhrs/Trade Mhrs	Predicted Scaffold Mhrs	Error	
				Absolute Difference	Relative
A Charge Pumps	23907.36	8.39%	26288.00	2380.65	9.96%
B Catalyst Handling	34198.57	10.91%	31528.31	2670.26	7.81%
C S/D Cooler / Catalyst Slop Oil / Short Circ Flash Drum	7208.39	11.12%	14645.16	7436.77	103.17%
D Make-up Hydrogen Compressors	17114.82	4.84%	18388.85	1274.03	7.44%
E Electrical Substation	1558.99	2.26%	2904.49	1345.50	86.31%
F First LCF Reactors	29952.62	10.14%	32653.62	2700.99	9.02%
G Hydrotreater Reactor	16010.04	10.38%	12857.49	3152.55	19.69%
H LCF Feed/HDT Hydrogen Heaters	19877.48	6.28%	35288.88	15411.41	77.53%
J Second LCF Reactors	24832.83	8.14%	34429.96	9597.14	38.65%
K Reactor Exchangers	13726.51	12.73%	6946.01	6780.51	49.40%
L Membrane	3433.07	15.16%	6490.60	3057.52	89.06%
M Amine	6917.11	12.66%	11765.27	4848.17	70.09%
N Heavy Oil Stripper	13738.97	10.40%	15104.24	1365.27	9.94%
P Stabilizer	9543.98	15.00%	6205.95	3338.04	34.98%

Q Flare Drum	6025.46	12.31%	4492.20	1533.26	25.45%
R Piperack	102510.48	27.64%	93249.93	9260.55	9.03%
S Stripper O/H Sour Gas/Stabilizer O/H Compr.	2608.85	8.28%	2218.30	390.55	14.97%
T Depropanizer	6803.84	10.10%	7027.49	223.65	3.29%

Statistics of Error Rates		
Items	**Abs Difference**	**Relative**
Max	15411.41	103.17%
Mean	4264.82	36.99%
Min	223.65	3.29%

Table 19 Error rate on trade level of linear model from experiment 12 on the training data

Trades	CIV	EL	FP	INSTR	INSUL	IW	ME	PF	Total
Actual Scaffold Mhrs	486.17	95568.66	12907.15	6261.37	23864.72	4681.31	12630.29	177309.33	333709.03
Actual Scaffold Mhrs/Trade Mhrs Ratio	0.18%	11.80%	14.98%	2.76%	28.34%	3.47%	6.08%	14.31%	10.92%
Predicted Scaffold Mhrs	6003.84	95958.76	14671.52	9709.51	27835.19	12927.04	17566.90	177286.90	361959.66
Predicted Scaffold Mhrs/Trade Mhrs Ratio	2.25%	11.85%	17.03%	4.28%	33.05%	9.59%	8.46%	14.31%	11.84%
Abs Scaffold Mhrs Error	5517.67	390.09	1764.37	3448.13	3970.46	8245.73	4936.61	22.43	28250.64
Relative Scaffold Mhrs Error	1134.92%	0.41%	13.67%	55.07%	16.64%	176.14%	39.09%	0.01%	8.47%
Abs Scaffold/Trade Ratio Error	2.06%	0.05%	2.05%	1.52%	4.71%	6.12%	2.38%	0.00%	0.92%
Relative Scaffold/Trade Ratio Error	1134.92%	0.41%	13.67%	55.07%	16.64%	176.14%	39.09%	0.01%	8.47%

Table 20 Error rate on construction area level of M5P tree model from experiment 14 based on the training data

Construction Areas	Actual Scaffold Mhrs	Scaffold Mhrs/Trade Mhrs	Predicted Scaffold Mhrs	Error	
				Absolute Difference	Relative
A Charge Pumps	23907.36	8.39%	27287.98	3380.62	14.14%
B Catalyst Handling	34198.57	10.91%	30976.65	3221.92	9.42%
C S/D Cooler / Catalyst Slop Oil / Short Circ Flash Drum	7208.39	11.12%	8331.50	1123.11	15.58%
D Make-up Hydrogen Compressors	17114.82	4.84%	19001.45	1886.63	11.02%
E Electrical Substation	1558.99	2.26%	1730.93	171.94	11.03%
F First LCF Reactors	29952.62	10.14%	32333.78	2381.16	7.95%
G Hydrotreater Reactor	16010.04	10.38%	15508.21	501.82	3.13%
H LCF Feed/HDT Hydrogen Heaters	19877.48	6.28%	29089.12	9211.64	46.34%
J Second LCF Reactors	24832.83	8.14%	34831.86	9999.04	40.27%
K Reactor Exchangers	13726.51	12.73%	10069.53	3656.99	26.64%
L Membrane	3433.07	15.16%	4856.99	1423.91	41.48%
M Amine	6917.11	12.66%	5715.82	1201.29	17.37%
N Heavy Oil Stripper	13738.97	10.40%	12835.62	903.36	6.58%

P Stabilizer	9543.98	15.00%	8138.56	1405.42	14.73%
Q Flare Drum	6025.46	12.31%	6071.66	46.20	0.77%
R Piperack	102510.48	27.64%	59045.21	43465.27	42.40%
S Stripper O/H Sour Gas/Stabilizer O/H Compr.	2608.85	8.28%	4317.73	1708.88	65.50%
T Depropanizer	6803.84	10.10%	6477.07	326.78	4.80%

Statistics of Error Rates		
Items	**Abs Difference**	**Relative**
Max	43465.27	65.50%
Mean	4778.66	21.06%
Min	46.20	0.77%

Table 21 Error rate on trade level of linear model from experiment 12 on the training data

Trades	CIV	EL	FP	INSTR	INSUL	IW	ME	PF	Total
Actual Scaffold Mhrs	486.17	95568.66	12907.15	6261.37	23864.72	4681.31	12630.29	177309.33	333709.03
Actual Scaffold Mhrs/Trade Mhrs Ratio	0.18%	11.80%	14.98%	2.76%	28.34%	3.47%	6.08%	14.31%	10.92%
Predicted Scaffold Mhrs	8834.19	82771.75	17812.21	5951.70	21225.71	5015.70	12849.69	161700.03	316160.99
Predicted Scaffold Mhrs/Trade Mhrs Ratio	3.30%	10.22%	20.67%	2.62%	25.20%	3.72%	6.19%	13.05%	10.35%
Abs Scaffold Mhrs Error	8348.02	12796.91	4905.06	309.67	2639.02	334.39	219.40	15609.30	17548.03
Relative Scaffold Mhrs Error	1717.09%	13.39%	38.00%	4.95%	11.06%	7.14%	1.74%	8.80%	5.26%
Abs Scaffold/Trade Ratio Error	3.12%	1.58%	5.69%	0.14%	3.13%	0.25%	0.11%	1.26%	0.57%
Relative Scaffold/Trade Ratio Error	1717.09%	13.39%	38.00%	4.95%	11.06%	7.14%	1.74%	8.80%	5.26%

4.3.3 Summary

Data mining experiments in phase three are trained in the modified input table, the same as experiments from phase two; the only difference is the change of the class. Also, a group of algorithms of decision trees are tested at this stage. The experimental data show that change of class from ratio of scaffold man-hours over direct trade man-hour to predict scaffold man-hours generally improved performance of the models. A noticeable increase of correlation coefficient and a drop of error rate were discovered. Especially from the M5P model, due to the different model for different ranges of the data, the error rate had a dramatic decrease.

4.4 Conclusion

Three different stages were contained in the data mining investigation of this chapter. Several conclusions could be drawn from the experimental results from these three phases, which are listed below:

- First, no matter in data mining investigation phase one, two or three, Linear Regression shows steady, good performance in this research; while more advanced function (Gaussian Process) could not surpass Linear model in performance but provides more complicated models; M5P decision tree adjusts models according to different data range, which provides better performance model in this research.
- Second, compared to the initial data mining investigation, no significant improvement was found in the second data mining phase after several modifications to the input table. However, the modified input table based on the advice from client's experts, contains more useful information than the original input table, for example winter factor, and day/night ratio. In addition, comparing with the first two phases of data mining investigation, models from the third data mining investigation observe a noticeable improvement of performance. All data mining experiments from third data mining investigation are based on the modified input table. This input table provides an example for the future scaffold tracking system.
- Third, the performance of models – correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error – is gradually increased from phase one to phase three. This improvement can also

be observed from two different sources: 1) graphically from the figures of error visualization; 2) the error evaluation – relative error rate – on construction area level and trade area of the original data, when the models were built in Excel.

- Forth, adding unnecessary attributes didn't necessarily improve the performance of the model, whether in first phase, second phase or third phase. At second data mining investigation phase, the experimental data shows that the model with best performance is built on 5 attributes; while the best performance M5P tree was built on a 6 attribute selection.
- Fifth, surprisingly, attributes like height or volume didn't affect the models in the way client's experts expected; nevertheless, this only concludes to the poor quality, inaccuracy and inconsistency of the database. Though efforts have been spent to make best use of this database, the results from solid data mining investigation show lots of the data recorded in the database is noise data, which doesn't contribute much to the improvement of the performance of the actual model. The results from these three phases of data mining investigation provide a lot of precious lessons learned for the future scaffold tracking system. Thus, a new database for scaffold request tracking was designed and built for the client, which is attached as Appendix 4. In addition, the proposal from the experts of training two different models for different phases in the project for different purposes is proved to be unattainable using this set of data.

Chapter 5 Evaluation of the Performance

5.1 Introduction

Though the model is trained on the historical scaffold database from one project, the objective of the research is to run this model on future projects to get the estimation for scaffold man-hours. Thus, all the data mining experiments presented in Chapter 4 were done using 10-fold cross validation, which aimed at obtaining the best performance prediction on a set of unknown data. 10-fold cross validation is a standard way of predicting the performance based on a limited data set. Nevertheless, to have a better idea of how these models were doing on this existing database is an interesting topic.

This chapter presents all the evaluation work of the models that were built in Section 4.2, which are trained on the experts' modified input tables. These evaluations are focused on different methods to break down the whole data set. The test mode used was 10-fold cross validation, which was shown in Chapter 4. However, comparing the model performance of more than one method of splitting the data set could provide error evaluated from different perspectives. The other two basic methods presented here are: first, treat the full data as a training set, which gives the error rate of the trained model on this existing database; second, holdout one third method – dividing the data set into three equal parts, train the model based on two parts, and obtain the evaluation information based on the third part, which hasn't been used in the training stage. The full evaluation investigation result is presented in Appendix 5, while parts of the evaluation comparison, as well as some charts showing the basic findings, are presented in Section 5.2.

5.2 Evaluation Investigation

This evaluation investigation is done based on the models of the second phase, which was trained using the experts' modified input table. Besides the original test model – 10-fold cross validation, full data set as training set, and one-third holdout for testing, two-third for training set are performed to build the comparison. Some typical samples of evaluation results are shown in the below table, and the table of full experimental data is attached as Appendix 5.

Table 22 Sample of evaluation investigation results

No.	Attributes	Algorithm			Output	Performance				
							Error			
		computer learning method	Test Mode	Parameter	Class	Correlation Coefficient	Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared
1	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_ Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8314	0.067	0.0911	55.3786 %	55.5726 %
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7361	0.0864	0.1132	71.1209 %	68.6713 %
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.734	0.0963	0.1221	81.6038 %	77.9371 %

	Factor									
6	11 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_ Congestion Degree; Area_Distance to Material Yard; H_Mean;H_Max; H_Mode;Trade Mhrs; Day/Night-shift Ratio; Winter Factor	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8014	0.0686	0.098	56.6582 %	59.8176 %
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.6939	0.0896	0.1209	73.7775 %	73.3233 %
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.6951	0.0959	0.1272	81.2690 %	81.1470 %
10	13 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_ Congestion Degree; Area_Distance to Material Yard;	Gaussian Process	100% Training data	Normalized PolyKerner-C250007 -E 2.0	Scaffold Mhrs/Trade Mhrs	0.9172	0.0385	0.0673	31.8168 %	41.0764 %
		Gaussian Process	10-fold-cross validation	Normalized PolyKerner-C250007 -E 2.0	Scaffold Mhrs/Trade Mhrs	0.6965	0.0797	0.1237	65.5898 %	75.0290 %

	H_Mean; H_StDev; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Gaussian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.7759	0.0795	0.101	67.3669 %	64.4734 %
13	5 ATTRITUBES:Trade; Area_Congestion Degree; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Gaussian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.8532	0.057	0.0858	47.0877 %	52.3155 %
		Gaussian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.4166	0.0862	0.1939	70.9223 %	117.6499 %
		Gaussian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.8232	0.0662	0.0904	56.0792 %	57.7084 %

Some box plots are drawn to compare the distribution of each measurement between different methods of breaking down the data set.

Figure 39 Box plot of correlation coefficient between three data breaking down methods

Figure 40 Plot of mean absolute error between three data breaking down methods

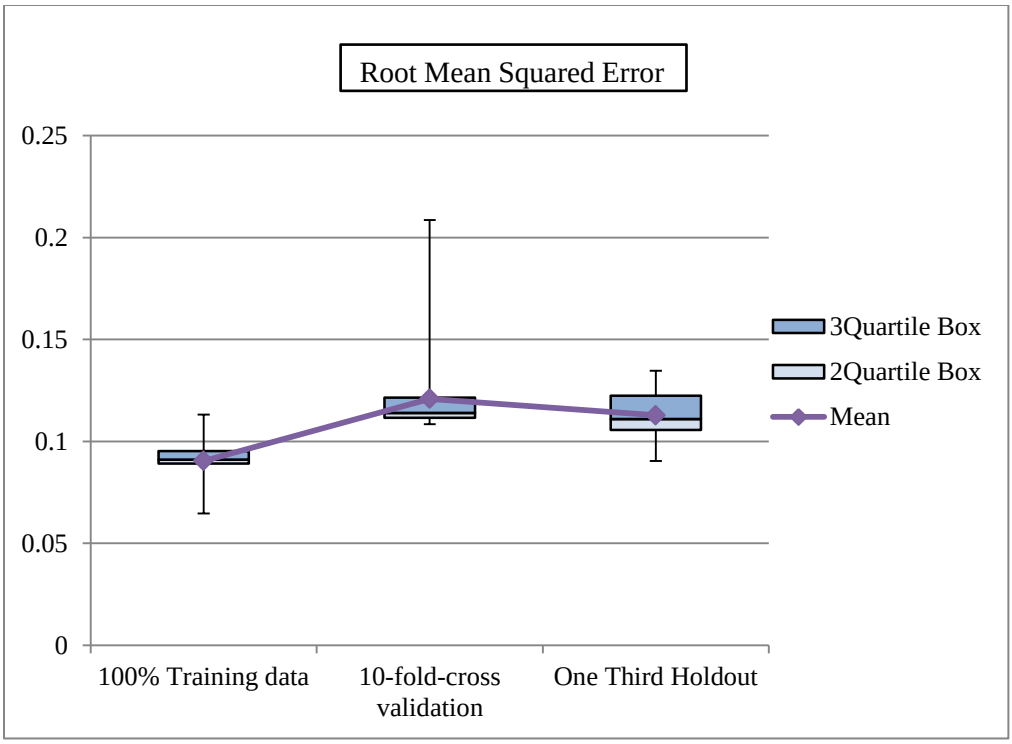


Figure 41 Plot of root mean squared error between three data breaking down methods

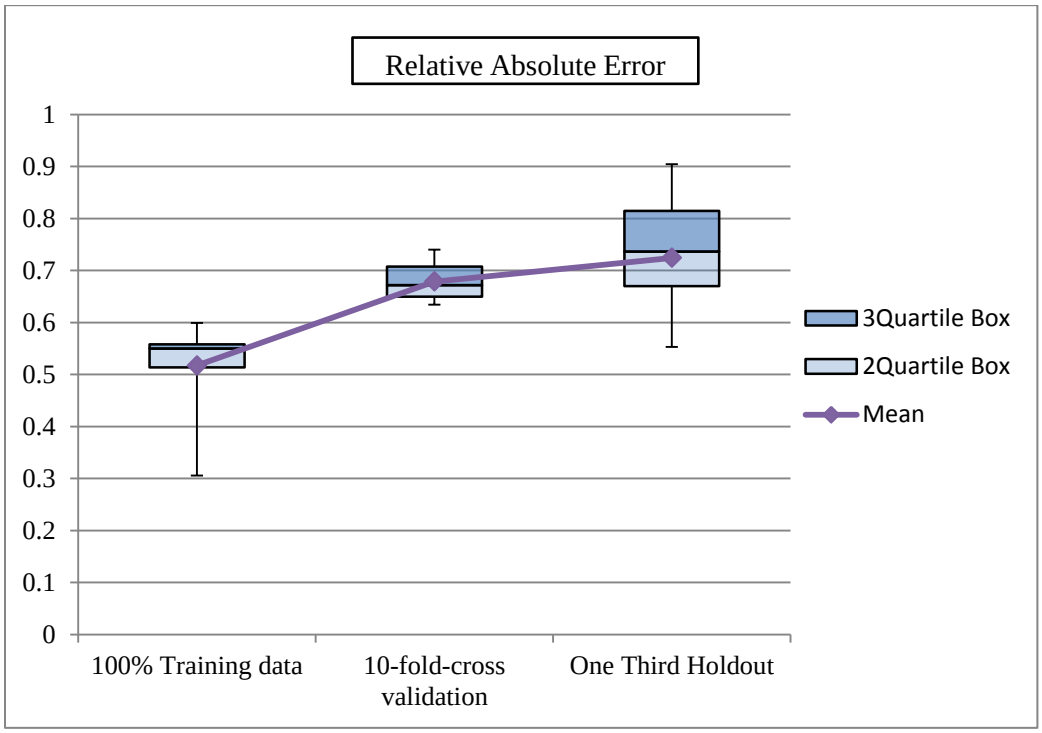


Figure 42 Plot of relative absolute error between three data breaking down methods

Figure 43 Plot of root relative squared error between three data breaking down methods

From the above five figures, it is clear that performance measurement using the full data set as training data is the most optimistic. It provides the lowest error rate, and highest correlation coefficient. Comparing to using the full data set as training data, one third holdout method and 10-fold cross validation method provides more reliable performance measure for the future unseen data set. In addition, if comparing 10-fold cross validation method and one third holdout method, the data spread wider in one third holdout method. This is because the two-third training data set or the one-third testing data was not representative in some cases. Thus, from another perspective, it proves that 10-fold cross validation is the best method to get a performance evaluation for the unknown data set. This is explained in Section 2.4.6.1.

5.3 Summary

Comparing the difference of error estimate among the same models with different data set break-down methods, as well as the box plots of different performance measurements, some conclusions are discovered.

- First, for each model, linear regression didn't show a dramatic difference between three different methods of evaluation. In comparison, Gaussian process model provides substantially better performance of error estimate on the 100 percent training data than 10-fold cross validation or one-third left out for testing

method. This indicates that this more sophisticated algorithm could fit a specific data set better. However, this feature could lead to an over-fit on the existing data set. This means a very high performance on the training data, but poor performance on an unknown data set.

- Second, for different data set break-down methods, it is clear that the full data set as training method provides over optimistic performance measure, which only gives the indication for the performance of this model on this specific data set. One-third holdout method and 10-fold cross validation gives more reliable performance measure. In addition, from the evaluation investigation result, it is clear that generally one-third holdout for testing method gives very close error estimate to that of 10-fold cross validation. But, due to the issue of whether training data or testing data is representative, one-third holdout method has an obvious limitation. Thus, 10-fold cross validation provides the most reliable performance measure for the unknown data set among these three methods.

Chapter 6 Conclusions

6.1 Research Summary

The research described in this thesis was motivated by the inaccuracy, as well as unspecific scaffolding estimation, and lack of effective control and management of scaffolding activities on industrial construction projects. This research focused on building a mathematical model based on data from a mega industrial construction scaffold database, to provide better and more specific scaffold estimation for future projects. The model can be continuously improved to perform more specifically and accurately by feeding on new scaffold request data.

This research began with understanding the process of the scaffold request process on site in this industrial construction project, as well as the structure and content kept in the scaffold request database. Interviews and meetings with site experts helped to build a business model on IDFE0 program to represent the actual scaffold request process on site. Clean up this mega industrial construction project's scaffold database was done under the supervision of the expert to correct the human error, inconsistency, and re-organize it. At the same time, a group of charts and tables were built to reveal the contents, features, and information within this database. All the work listed above was presented in Chapter 3 of this thesis.

The computer learning in this research used WEKA software based on the pre-processed data. Three phases of data mining investigation were conducted based on two slightly different input tables. The first phase of the data mining investigation was done using an original input table organized directly from the data preparation. Based on the results of the first phase of data mining investigation, the experts proposed a number of changes to the input table. The second phase of the data analysis was based on the expert modified input table. Then, a change of class was made in the third phase data mining investigation, on the modified input table. All the data mining experimental data from different phases was recorded in Chapter 4.

Followed by the data mining investigation, a group of evaluations of the models based on the second phase of data mining investigation was done. These evaluations tested the

model using different ways of splitting the data. The evaluation and the findings were shown in Chapter 5.

6.2 Research Contribution

The research contained in this thesis has a number of contributions to the academic world as well as the industrial construction field. The main contributions can be summarized as follows:

1. This research is one of the very first attempts to use data mining tool to improve the efficiency and the accuracy of scaffold estimation. Data mining has been involved in reduction of cost and increase of productivity and efficiency of a construction project for a long time. However, most of the effort was focused on the permanent structure and direct work. This research is one of the very first attempts to address scaffold estimates in industrial construction projects.
2. This research identified several important attributes which significantly affect the scaffold estimate according to the historical provided by the client company. Some of these attributes are numerical, others are norminal.
3. This research proposed a simple tree structural multivariate linear model for estimation of scaffold activities in an industrial construction project. This model provided a quantitative method for the project estimator and project manager to quickly come up with the scaffold man-hours needed for future projects based on the simple information available at the starting point of a project. This suits the fast track feature of industrial construction projects.
4. The process provided in this research could be followed and conducted by other industrial company and build their own model to provide a more accurate and detailed scaffold estimates.

6.3 Research Limitations

Besides the results and contribution of this research, several limitations include:

- First, the majority of the data mining experiments use Linear Regression. This is because this research was based on a real industrial project database, and aimed at producing a practical mathematical model for future projects. A simple algorithm is easy to implement into the real world.

- Second, the data mining experiment is heavily dependent on the preparation of the data set. In this research, two different input tables were built for the data mining process. However, based on the same database, different input tables can be set up for data mining purposes.
- Third, due to lack of original estimating information, no direct comparison between the estimate provided by the models and the original estimates can be made.

6.4 Recommendations for Future Work

This research was just a start to the long-term goal of improving efficiency and accuracy in scaffolding estimation, and reducing the cost of scaffolding activities. Future work can be done to extend and expand research in this area. Some of the recommended work includes:

- More effort should be devoted to keeping the scaffold information during the project. It is crucial that consistent and complete data is captured for data management processes to work properly. The weakest point for most of the scaffold management applications is inconsistent and poor quality data from the field (Ryan, 2009). A better designed data base should be recommended to track more important data in a consistent and clear way. On one side, better quality data should be tracked, on the other side, additional information, for example cost code and work package number, could be tracked during the project. Also, well explained rules and regulations are necessary to help field workers track good quality data. The better raw scaffold data resource is the key to the success of future research.
- More scaffold data from different projects could be tested in this model to optimize it. This research was based on a scaffold database from one mega industrial construction project, and was able to train a model out of it. However, due to the limitation of data, the model might be insufficient for future projects that contain different features. Thus, complete scaffold data from different projects is necessary for the improvement of this scaffold estimation model.
- In this research, only a limited number of algorithms have been tested on this data for training the model. However, other methods, which are more sophisticated, are available in data mining field. Another approach could be

expanding this research using different methods, for example neural network, to build the model other than linear regression.

References

Adeli H., Wu M. (1998). Regularization neural network for construction cost estimation. *Journal of Construction Engineering and Management*, 124, 18-24.

Alberta Construction Safety Association (ASCA). (n.d.). *Safe work procedures manual for scaffolds in Alberta.*

Bouckaert, R. R., Frank, E., Hall, M., Kerkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2012). *WEKA Manual for Version 3-6-8*. University of Waikato, Hamilton, New Zealand.

Bozdogan, H. (1987). Model selection and akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometerika*, 52(3), 345-370.

Carr, R. (1989). Cost-estimating principles. *Journal of Construction Engineering and Management*, 115(4), 545–551.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883.

Clough, R. H., Glenn, A. S., & Sears, S. K. (2000). *Construction project management* (4th ed.). New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: John Wiley & Sons, Inc.

Infrastructure Health & Safety Association (IHSA). (n.d.). *Scaffolds*, 1-24.

Ebden, M. (2008). Gaussian processes for regression: A quick introduction.

Fan, H. (2007). Leveraging operational data for intelligent deicision support in construction equipment management. PhD thesis, University of Alberta, Edmonton, AB.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 37-54.

Gonzalez-Villalobos, C. V. (2011). Analysis of industrial construction activities using knowledge discovery techniques. MSc thesis, University of Alberta, Edmonton, AB.

Gunaydin, H. M., & Dogan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. *International Journal of Project Management*, 22, 595-602

Hammad, A. M. (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, AB.

Hendrickson, C., & Au, T. (2008). *Project management for construction: fundamental concepts for owners, engineers, architects and builders.* Pittsburgh: Department of Civil and Environmental Engineering, Carnegie Mellon University.

Hill, H., Searer, G., Dethlefs, R., Lewis, J., & Paret, T. (2010). Designing suspended scaffold structural support elements and lifeline anchorages in conformance with federal OSHA requirements. *Practice Periodical on Structural Design and Construction* , 15(3), 186-193.

Hill, H., Searer, G., Dethlefs, R., Lewis, J., & Paret, T. (2010). Certifying that existing suspended scaffold structural support elements and lifeline anchorages comply

with federal OSHA requirements. *Practice Periodical on Structural Design and Construction* , 15(3), 194-200.

Illingworth, J. R. (1987). *Temporary works their role in construction.* London: Tomas Telford.

Kamara, J. M., Augenbroe, G., Anumba, C. J., & Carrillo, P. M. (2002). Knowledge management in the architecture, engineering and construction industry. *Construction Innovation: Information, Process, Management*, 2(1), 53-67.

Kim, G. H., An, A. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, 39(10), 1235-1242.

Kim, H. J., Seo, Y. C., & Hyun, C. T. (2012). A hybird conceptual cost estimation model for large building projects. *Automation in Construction*, 25, 72-81.

Kirkby, R., Frank, E. & Reutemann, P. (2008). *WEKA Explorer user guide for version 3-5-8*. University of Waikato, Hamilton, New Zealand.

Mackay, D. J. (1998). *Introduction to gaussian processes.* Cambridge, London: Cambridge University.

Moon, S. W., Kim, J. S., Kwon, K. N. (2007). Effectiveness of OLAP-based cost data management in construction cost estimate. *Automation in Construction*, 16, 336-344

Oregon OSHA (OR-OSHA). (n.d.). *Scaffolds temporary elevated work platforms guidelines for Oregon workers*.

Peng, J., Pan, A., Rosowsky, D., Chen, W., Yen, T., & Chan, S. (1996). High clearance scaffold systems during construction -- II. Structural analysis and development of design guidelines. *Engineering Structures*, 18(3), 258-267.

Piatetsky-Shapiro, G., Brachman, R., Khabaza, T., KloesGen, W., & Simoudis, E. (1996). An overview of issues in developing industrial data mining and knowledge discovery applications. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR: AAAI Press, 89-95.

Proverbs, D. P., Holt, G. D., & Olomolaiye, P. O. (1998). Scaffolding for high-rise concrete construction: A French, German and UK comparison. *Proc. Instn Civ. Engrs Structs and Bldgs*, 59-66.

Pyle, D. (1999), *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann.

Quinlan, J. (1992). Learning with continuous classes. *Proceedings AI'92,* Singapore: World Scientific, 343-348.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning.* Cambridge, MA: The MIT Press.

Rezgui, Y. (2001). Review of information and the state of the art of knowledge management practices in the construction industry. *The Knowledge Engineering Review*, 16(3), 241-254.

Ryan, G. (2009). *Schedule for sale workface planning for construction projects.* Bloomington, IN: AuthorHouse.

Son, K. S., & Park, J. J. (2010). Structural analysis of steel pipe scaffolding based on the tightening strength of clamps. *Journal of Asian Architecture and Building Engineering*, 9(2), 479-485.

Stine, R. A. (2003). *Model selection using information theory and the MDL principle.* Philadelphia, PA: The Wharton School of the University of Pennsylvania.

Tah, J. H. M., Thorpe, A., & McCaffer, R. (1994). A survey of indirect cost estimating in practice. *Construction management and economics*, 12(1), 31-36.

Wang, Y., & Witten, I. H. (1996). *Induction of model trees for predicting continuous classes.* Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining practical machine learning tools and techniques* (3rd ed.). Amsterdam, Boston, Heidelberg, London, New York, Ocford, Paris, San Diego, San Francisico, Singaporem, Sydney, Tokyo: Morgan Kaufmann Publishers.

Appendix 1

Table 23 Full Experimental Results from Data Mining Investigation Phase One

stages	objectives	Input		Algorithm		
		Data set	**Attributes**	**computer learning method**	**Test Mode**	**Parameter**
1	build an overall model for each trade in each construction area (141 records)	Initial Input Table	Full Set	Linear Regression	10-fold cross-Validation	Greedy method; 1.0E-8
2	build an overall model for each trade in each construction area (141 records)	Initial Input Table	19 attributes, remove Count, Scaffold Mhrs, and Scaffold Distribution	Linear Regression	10-fold cross-Validation	Greedy method; 1.0E-8
3	build an overall model for each trade in each construction area (141 records)	Initial Input Table	19 attributes, remove Count, Scaffold Mhrs, and Scaffold Distribution	Linear Regression	10-fold cross-Validation	M5 method; 1.0E-8

4	build an overall model for each trade in each construction area (141 records)	Initial Input Table	17 attributes (Trades, Area, General, Elevation, Scaffold Type, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Max, V_Min, V_Mode, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8
5	build an overall model for each trade in each construction area (141 records)	Initial Input Table	17 attributes (Trades, Area, General, Elevation, Scaffold Type, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Max, V_Min, V_Mode, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	M5 method; 1.0E-8
6	build an overall model for each trade in each construction area (141 records)	Initial Input Table	17 attributes (Trades, Area, General, Elevation, Scaffold Type, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Max, V_Min, V_Mode, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	No attributes selection method; 1.0E-8
7	build an overall model for each trade in each construction area (141 records)	Initial Input Table	15 attributes (Trades, Area, General, Elevation, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Min, V_Mode, Trade Mhr Distribution, Scaffold	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8

			Mhrs/Trade Mhrs)			
8	build an overall model for each trade in each construction area (141 records)	Initial Input Table	15 attributes (Trades, Area, General, Elevation, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Min, V_Mode, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	No attributes selection method; 1.0E-8
9	build an overall model for each trade in each construction area (141 records)	Initial Input Table	15 attributes (Trades, Area, General, Elevation, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Min, V_Mode, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	M5 method; 1.0E-8
10	build an overall model for each trade in each construction area (141 records)	Initial Input Table	12 attributes (Trades, Area, General, Elevation, H_Mean, H_Max, H_Mode, V_Mean, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8
11	build an overall model for each trade in each construction area (141 records)	Initial Input Table	12 attributes (Trades, Area, General, Elevation, H_Mean, H_Max, H_Mode, V_Mean, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	M5 method; 1.0E-8

12	build an overall model for each trade in each construction area (141 records)	Initial Input Table	9 attributes (Trades, Area, General, Elevation, H_Max, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8
13	build an overall model for each trade in each construction area (141 records)	Initial Input Table	11 attributes (Trades, Area, General, H_Mean, H_Max, H_Mode, V_Mean, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8
14	build an overall model for each trade in each construction area (141 records)	Initial Input Table	10 attributes (Trades, Area, General, H_Mean, H_Max, V_Mean, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8
15	build an overall model for each trade in each construction area (141 records)	Initial Input Table	9 attributes (Trades, Area, General, H_Mean, H_Max, V_Mean, V_StDev, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8
16	build an overall model for each trade in each construction area (141 records)	Initial Input Table	6 attributes (Trades, Area, General, V_StDev, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	Greedy Method; 1.0E-8

17	build an overall model for each trade in each construction area (141 records)	Initial Input Table	Full Set	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
18	build an overall model for each trade in each construction area (141 records)	Initial Input Table	19 attributes, remove Count, Scaffold Mhrs, and Scaffold Distribution	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
19	build an overall model for each trade in each construction area (141 records)	Initial Input Table	18 attributes (Trades, Area, General, Elevation, Scaffold Type, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Max, V_Min, V_Mode, Trade Mhrs, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
20	build an overall model for each trade in each construction area (141 records)	Initial Input Table	17 attributes (Trades, Area, General, Elevation, Scaffold Type, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Max, V_Min, V_Mode, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0

21	build an overall model for each trade in each construction area (141 records)	Initial Input Table	17 attributes (Trades, Area, General, Elevation, Scaffold Type, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Max, V_Min, V_Mode, Trade Mhr, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
22	build an overall model for each trade in each construction area (141 records)	Initial Input Table	15 attributes (Trades, Area, General, Elevation, H_Mean, H_StDev, H_Max, H_Min, H_Mode, V_Mean, V_StDev, V_Min, V_Mode, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
23	build an overall model for each trade in each construction area (141 records)	Initial Input Table	12 attributes (Trades, Area, General, Elevation, H_Mean, H_Max, H_Mode, V_Mean, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
24	build an overall model for each trade in each construction area (141 records)	Initial Input Table	11 attributes (Trades, Area, General, H_Mean, H_Max, H_Mode, V_Mean, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0

25	build an overall model for each trade in each construction area (141 records)	Initial Input Table	9 attributes (Trades, Area, General, H_Mean, H_Max, V_StDev, V_Min, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
26	build an overall model for each trade in each construction area (141 records)	Initial Input Table	6 attributes (Trades, Area, General, V_StDev, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Gaussian Process	10-fold cross-Validation	Normalized PolyKernel; -c 250007; Exponent 2.0
27	build an overall model for each trade in each construction area (141 records)	Initial Input Table	4 attributes (Trades, V_StDev, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	No attributes sellection; 1.0E-8
28	build an overall model for each trade in each construction area (141 records)	Initial Input Table	3 attributes (Trades, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	No attributes sellection; 1.0E-8
29	build an overall model for each trade in each construction area (141 records)	Initial Input Table	5 attributes (Trades, H_Max, V_StDev, Trade Mhr Distribution, Scaffold Mhrs/Trade Mhrs)	Linear Regression	10-fold cross-Validation	No attributes sellection; 1.0E-8

Output	**Performance**		**Comment**
	Correlation	**Error**	

Class	Model	Coefficient	Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared	
Scaffold Mhr/Trade Mhr	see result Exp-1	0.7272	0.0795	0.1152	65.4626%	69.9013%	contain Scaffold Mhrs and Scaffold Distribution in the model, which is impractical
Scaffold Mhr/Trade Mhr	see result Exp-2	0.6963	0.0869	0.1189	71.5290%	72.1187%	
Scaffold Mhr/Trade Mhr	see result Exp-3	0.7004	0.085	0.1189	69.9141%	72.1257%	
Scaffold Mhr/Trade Mhr	see result Exp-4	0.7163	0.0847	0.1161	69.6697%	70.4561%	
Scaffold Mhr/Trade Mhr	see result Exp-5	0.7278	0.0859	0.1142	70.7259%	69.2773%	

Scaffold Mhr/Trade Mhr	see result Exp-6	0.7312	0.0854	0.1137	70.2664%	68.9745%	different attributes selection method, different model; though no any attributes selection, the performance looks slightly better than the other two, the model is way more completed than the other two.
Scaffold Mhr/Trade Mhr	see result Exp-7	0.7349	0.081	0.1119	66.6694%	67.8713%	
Scaffold Mhr/Trade Mhr	see result Exp-8	0.704	0.0886	0.119	72.9359%	72.1793%	
Scaffold Mhr/Trade Mhr	see result Exp-9	0.733	0.0812	0.1125	66.8657%	68.2619%	

Scaffold Mhr/Trade Mhr	see result Exp-10	0.7461	0.0777	0.1094	63.9310%	66.4006%	
Scaffold Mhr/Trade Mhr	see result Exp-11	0.7339	0.081	0.1122	66.6983%	68.0833%	
Scaffold Mhr/Trade Mhr	see result Exp-12	0.7409	0.0798	0.1105	65.6562%	67.0339%	
Scaffold Mhr/Trade Mhr	see result Exp-13	0.7396	0.0804	0.1109	66.1564%	67.3091%	
Scaffold Mhr/Trade Mhr	see result Exp-14	0.7396	0.0803	0.111	66.0801%	67.3374%	

Scaffold Mhr/Trade Mhr	see result Exp-15	0.7405	0.0796	0.1108	65.4841%	67.2064%	
Scaffold Mhr/Trade Mhr	see result Exp-16	0.7457	0.0793	0.1098	65.2455%	66.6172%	
Scaffold Mhr/Trade Mhr	see result Exp-17	0.7187	0.0846	0.1189	69.5936%	72.1254%	
Scaffold Mhr/Trade Mhr	see result Exp-18	0.7056	0.0866	0.1204	71.2296%	73.0313%	
Scaffold Mhr/Trade Mhr	see result Exp-19	0.7162	0.0828	0.1174	68.1726%	71.2140%	
Scaffold Mhr/Trade Mhr	see result Exp-20	0.7161	0.0828	0.1174	68.1282%	71.2118%	

Scaffold Mhr/Trade Mhr	see result Exp-21	0.6929	0.084	0.1204	69.1164%	73.0548%	
Scaffold Mhr/Trade Mhr	see result Exp-22	0.7217	0.0817	0.1161	67.2202%	70.4454%	
Scaffold Mhr/Trade Mhr	see result Exp-23	0.7352	0.0799	0.1136	65.7919%	68.9492%	
Scaffold Mhr/Trade Mhr	see result Exp-24	0.7614	0.0761	0.1082	62.6092%	65.6705%	
Scaffold Mhr/Trade Mhr	see result Exp-25	0.7663	0.0756	0.1069	62.2376%	64.8849%	
Scaffold Mhr/Trade Mhr	see result Exp-26	0.765	0.0758	0.1066	62.3541%	64.6855%	

Scaffold Mhr/Trade Mhr	see result Exp-27	0.7258	0.0765	0.1133	62.9818%	68.7475%	Try use only two or three attributes to build the model.
Scaffold Mhr/Trade Mhr	see result Exp-28	0.6947	0.0762	0.1185	62.7267%	71.8688%	
Scaffold Mhr/Trade Mhr	see result Exp-29	0.7294	0.0771	0.1126	63.4128%	68.3349%	

Appendix 2

Table 24 Full Experimental Results from Data Mining Investigation Phase Two

No.	objectives	Input		Algorithm		
		Data set	**Attributes**	**computer learning method**	**Test Mode**	**Parameter**
1	For estimation purpose	Modified Input Table	15 ATTRITUBES: Construction Areas; Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
2	For estimation purpose	Modified Input Table	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection

3	For estimation purpose	Modified Input Table	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	M5 Method
4	For estimation purpose	Modified Input Table	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	Greedy Method
5	For estimation purpose	Modified Input Table	13 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
6	For estimation purpose	Modified Input Table	12 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection

7	For estimation purpose	Modified Input Table	11 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
8	For estimation purpose	Modified Input Table	6 ATTRITUBES:Trade; Area_Congestion Degree; H_Mode;Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
9	For estimation purpose	Modified Input Table	5 ATTRITUBES:Trade; Area_Congestion Degree; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
10	For estimation purpose	Modified Input Table	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0
11	For estimation purpose	Modified Input Table	13 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0

12	For estimation purpose	Modified Input Table	12 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0
13	For estimation purpose	Modified Input Table	6 ATTRITUBES:Trade; Area_Congestion Degree; H_Mode;Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0
14	For estimation purpose	Modified Input Table	5 ATTRITUBES:Trade; Area_Congestion Degree; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0
15	For estimation purpose	Modified Input Table	4 ATTRITUBES:Trade; Area_Congestion Degree; Trade Mhr Distribution; Day/Night-shift Ratio	Linear Regression	10-fold-cross validation	No attribute selection
16	For estimation purpose	Modified Input Table	7 ATTRITUBES:Trade; Area_Congestion Degree; H_Mean; H_Max, Trade Mhrs, Trade Mhr Distribution; Day/Night-shift Ratio	Linear Regression	10-fold-cross validation	No attribute selection

1	For control purpose	Modified Input Table	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	M5 Method
2	For control purpose	Modified Input Table	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
3	For control purpose	Modified Input Table	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	Greedy Method
4	For control purpose	Modified Input Table	17 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Min; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter	Linear Regression	10-fold-cross validation	No attribute selection

			Factor			
5	For control purpose	Modified Input Table	14 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Min; V_Mean; V_Min; Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
6	For control purpose	Modified Input Table	6 ATTRITUBES:Trade; Area_Congestion Degree; H_Mode;Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
7	For control purpose	Modified Input Table	5 ATTRITUBES:Trade; Area_Congestion Degree; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
8	For control purpose	Modified Input Table	15 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection
9	For control purpose	Modified Input Table	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0

10	For control purpose	Modified Input Table	17 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Min; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0
11	For control purpose	Modified Input Table	15 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0
12	For control purpose	Modified Input Table	12 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean;V_Min; Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner -C250007 -E 2.0
13	For control purpose	Modified Input Table	12 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean;V_Min; Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection

Output		Performance					Comment
		Correlation Coefficient	Error				
Class	Model		Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared	
Scaffold Mhrs/Trade Mhrs	see Result Est_1.2	0.6747	0.091	0.1267	74.9090 %	76.8696%	
Scaffold Mhrs/Trade Mhrs	see Result Est_2.2	0.7302	0.0874	0.1142	71.9153 %	69.2623%	Attributes Selection Required by PCL
Scaffold Mhrs/Trade Mhrs	see Result Est_3.2	0.7335	0.0843	0.1133	69.3507 %	68.7144%	Trades, Area_Congestion Degree and Area_Distance to Material Yard, H_Mean, and H_Max, Trade Mhrs, Trade Mhr Distribution, and D/N Ratio are selected by M5 Method
Scaffold Mhrs/Trade Mhrs	see Result Est_4.2	0.7527	0.0802	0.1086	65.9813 %	65.9154%	Trades, Area_Congestion Degree, H_Mean, and H_Max, Trade Mhrs, Trade Mhr Distribution, and D/N Ratio are selected by Greedy Method

Scaffold Mhrs/Trade Mhrs	see Result Est_5.2	0.7371	0.086	0.1126	70.8105 %	68.3029%	Removed "H_Min"
Scaffold Mhrs/Trade Mhrs	see Result Est_6.2	0.7473	0.0859	0.1111	70.6539 %	67.3791%	Removed "H_Min" and "H_StDev"
Scaffold Mhrs/Trade Mhrs	see Result Est_7.2	0.6939	0.0896	0.1209	73.7775 %	73.3233%	No "Trade Mhrs Distribution", can see a significant performance drop.
Scaffold Mhrs/Trade Mhrs	see Result Est_8.2	0.7514	0.0789	0.109	64.9833 %	66.1433%	according to attributes selection -- Evaluator: weka.attributeSelection.CfsSubsetEval and Search method: Greedy Stepwise
Scaffold Mhrs/Trade Mhrs	see Result Est_9.2	0.7517	0.0788	0.1089	64.8312 %	66.0787%	according to attributes selection -- Evaluator: weka.attributeSelection.CfsSubsetEval and Search method: Best First
Scaffold Mhrs/Trade Mhrs	see Result Est_10.2	0.6868	0.0817	0.1268	67.2072 %	76.9366%	

Scaffold Mhrs/Trade Mhrs	see Result Est_11.2	0.6965	0.0797	0.1237	65.5898%	75.0290%	Removed "H_Min"
Scaffold Mhrs/Trade Mhrs	see Result Est_12.2	0.6972	0.0796	0.1235	65.4754%	74.8988%	Removed "H_Min" and "H_StDev"
Scaffold Mhrs/Trade Mhrs	see Result Est_13.2	0.4113	0.0899	0.2086	74.0139%	126.5533%	
Scaffold Mhrs/Trade Mhrs	see Result Est_14	0.4166	0.0862	0.1939	70.9223%	117.6499%	
Scaffold Mhrs/Trade Mhrs	see Result Est_15.2	0.7541	0.0779	0.1085	64.1396%	65.8334%	Based on :attributes selection -- Evaluator: weka.attributeSelection.CfsSubsetEval and Search method: Best First, then delet Winter Factor

Scaffold Mhrs/Trade Mhrs	see Result Est_16.2	0.748	0.0798	0.1101	65.6629%	66.8027%	
Scaffold Mhrs/Trade Mhrs	See Result Con_1.2	0.7305	0.0838	0.1135	68.9905%	68.8579%	H_Mode, H_StDev are not selected by M5 Method
Scaffold Mhrs/Trade Mhrs	See Result Con_2.2	0.7139	0.0874	0.1162	71.9636%	70.5224%	Attributes Selection Required by PCL
Scaffold Mhrs/Trade Mhrs	See Result Con_3.2	0.726	0.0841	0.1138	69.2379%	69.0566%	Area_Congestion Degree, H_Max, H_StDev, H_Mean, V_Mode, V_Mean, V_StDev are not selected
Scaffold Mhrs/Trade Mhrs	See Result Con_4.2	0.7176	0.0865	0.1166	71.1986%	70.7356%	on the trained model, V_Max, V_Mode, and V_StDev.
Scaffold Mhrs/Trade Mhrs	See Result Con_5.2	0.7067	0.0888	0.1189	73.0442%	72.1629%	

Scaffold Mhrs/Trade Mhrs	see Result Est_8.2	0.7514	0.0789	0.109	64.9833 %	66.1433%	according to attributes selection -- Evaluator: weka.attributeSelection.CfsSubsetEval and Search method: Greedy Stepwise
Scaffold Mhrs/Trade Mhrs	see Result Est_9.2	0.7517	0.0788	0.1089	64.8312 %	66.0787%	according to attributes selection -- Evaluator: weka.attributeSelection.CfsSubsetEval and Search method: Best First
Scaffold Mhrs/Trade Mhrs	See Result Con_8.2	0.7399	0.0846	0.1121	69.6124 %	68.0005%	
Scaffold Mhrs/Trade Mhrs	See Result Con_9.2	0.736	0.0799	0.1142	65.7180 %	69.2935%	
Scaffold Mhrs/Trade Mhrs	See Result Con_10.2	0.7382	0.0791	0.1137	65.0714 %	68.9795%	
Scaffold Mhrs/Trade Mhrs	See Result Con_11.	0.7364	0.079	0.1139	64.9909 %	69.1056%	

	2						
Scaffold Mhrs/Trade Mhrs	See Result Con_12.2	0.7364	0.0788	0.1138	64.8691%	69.0399%	
Scaffold Mhrs/Trade Mhrs	See Result Con_13.2	0.7314	0.0843	0.1139	69.3690%	69.1162%	

Appendix 3

Table 25 Full Experimental Results from Data Mining Investigation Phase Three

stages	Input		Algorithm			Output		Performance				
									Error			
	Data set	Attributes	computer learning method	Test Mode	Parameter	Class	Model	Correlation Coefficient	Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared
1	Modified Input Table	10 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material	M5Rules	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5R_1	0.6948	1958.524	5608.9378	51.55 %	73.83 %

		Yard; H_Mean;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor										
2	Modified Input Table	10 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean;Tr	Linear Regression	10-fold cross validation	No attribute selection	Scaffold Mhrs	see Result Changeclass_LR_1	0.8073	2490.8697	4499.5888	65.5577%	59.2238%

		ade Mhrs; Trade Mhr Distribution ; Day/Night-shift Ratio; Winter Factor										
3	Modified Input Table	10 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean;Trade Mhrs; Trade Mhr	M5P	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5P_1	0.8007	1771.7311	4532.4162	46.6306%	59.6559%

		Distribution ; Day/Night-shift Ratio; Winter Factor										
4	Modified Input Table	9 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; Trade Mhrs; Trade Mhr Distribution ; Day/Night-	M5Rules	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5R_2	0.6953	2099.3217	5631.9032	55.2525%	74.1274%

		shift Ratio; Winter Factor										
5	Modified Input Table	9 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio;	Linear Regression	10-fold cross validation	No attribute selection	Scaffold Mhrs	see Result Changeclass_LR_2	0.8087	2477.3355	4483.0749	65.2015%	59.0064%

		Winter Factor										
6	Modified Input Table	9 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio;	M5P	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5P_2	0.7953	1842.0499	4584.7389	48.4813%	60.3446%

		Winter Factor										
7	Modified Input Table	8 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; Trade Mhrs; Trade Mhr Distribution; Winter Factor	M5Rules	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5R_3	0.6964	2085.7654	5625.1483	54.8957%	74.0385%

8	Modified Input Table	8 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; Trade Mhrs; Trade Mhr Distribution; Winter Factor	Linear Regression	10-fold cross validation	No attribute selection	Scaffold Mhrs	see Result Changeclass_LR_3	0.8115	2471.3932	4450.3348	65.0451%	58.5755%
9	Modified Input Table	8 ATTRITUBES: Trades; Area_Size; Area_Complexity;	M5P	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5P_3	0.803	1767.8729	4517.5977	46.5290%	59.4608%

		Area_Congestion Degree; Area_Distance to Material Yard; Trade Mhrs; Trade Mhr Distribution; Winter Factor										
10	Modified Input Table	10 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material	Linear Regression	10-fold cross validation	No attribute selection	Scaffold Mhrs	see Result Changeclass_LR_4	0.8064	2501.6398	4510.0808	65.8412%	59.3619%

		Yard; H_Mode;Trade Mhrs; Trade Mhr Distribution ; Day/Night-shift Ratio; Winter Factor										
11	Modified Input Table	7 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Trade Mhrs; Trade Mhr Distribution ; Winter	Linear Regression	10-fold cross validation	No attribute selection	Scaffold Mhrs	see Result Changeclass_LR_5	0.8138	2388.4041	4418.8519	62.8609%	58.1612%

		Factor										
12	Modified Input Table	6 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Trade Mhrs; Trade Mhr Distribution;	Linear Regression	10-fold cross validation	No attribute selection	Scaffold Mhrs	see Result Changeclass_LR_6	0.8153	2366.9857	4401.447	62.2972%	57.9321%
13	Modified Input Table	10 ATTRITUBES: Trades; Area_Size; Area_Complexity;	M5P	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5P_4	0.7982	1911.3709	4556.9843	50.3058%	59.9792%

		Area_Congestion Degree; Area_Distance to Material Yard; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor										
14	Modified Input Table	7 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion	M5P	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5P_5	0.8159	1634.9469	4397.5586	43.0305%	57.8809%

		Degree; Trade Mhrs; Trade Mhr Distribution ; Winter Factor										
15	Modified Input Table	6 ATTRITUBES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Trade Mhrs; Trade Mhr Distribution	M5P	10-fold cross validation	M 4.0	Scaffold Mhrs	see Result Changeclass_M5P_6	0.8025	1737.7345	4544.9006	45.7358%	59.8202%

Appendix 4

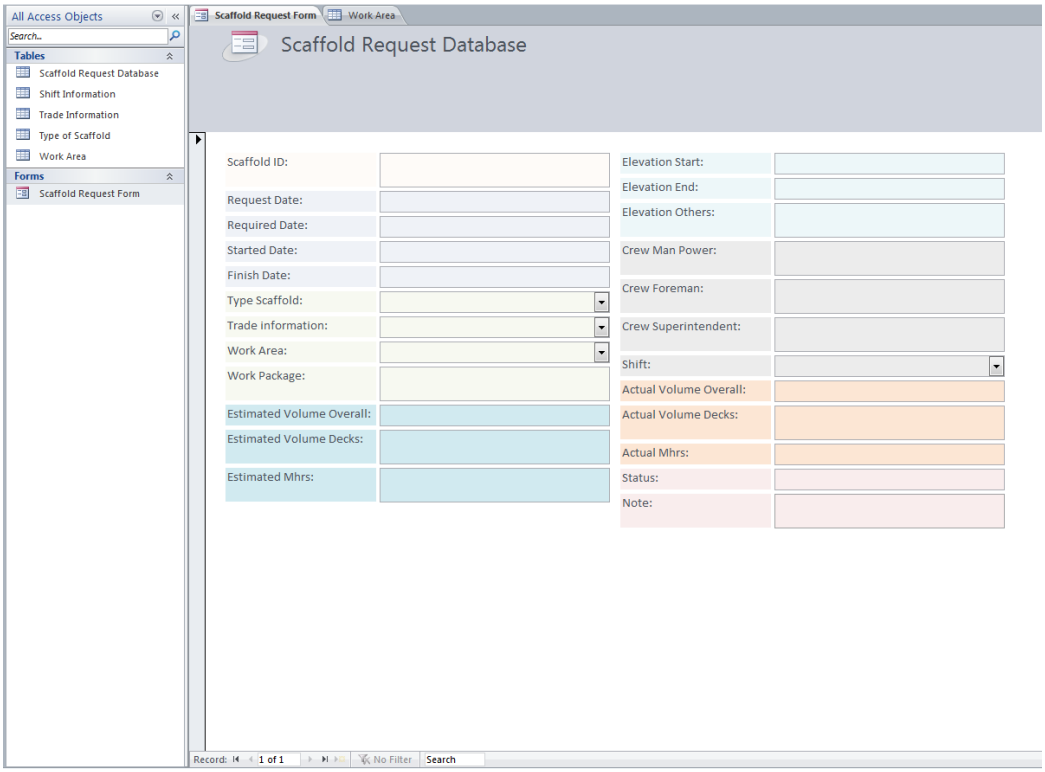


Figure 44 Form of Scaffold Request Database

Notes for the scaffold request database (form)

1. 4 dates will be kept in the new database, including request date, required date, start date and finish date. Three dates were tracked in Shell's project, which were request, required, and completion date.
2. Scaffold type is restricted to four, which are modification, erection, addition, and dismantle;
3. Trade information including trade and the foreman's name on a drop down list. A table called "Trade Information" is built and connected to the scaffold
4. d request database to record all the trades and their foreman's name. Thus, reducing the typing work and improving the consistency.
5. Work area contains a drop-down list of names of all the big construction areas in this project according to the blue print.
6. Work package, to keep which work package this scaffold request is needed for, if this level of information is possible to track.

7. Volume is recorded in two rows, one is an overall volume, which calculated as area times the total height form the ground to the highest level of working deck; the other is called “Volume Decks”, which records how many decks this scaffold request contains, and area size of each deck. The format of “Volume Decks” is locked as “00-0000”, to help the information tracking consistency.
8. Elevation is kept in three rows: if only one deck is required, then only “Elevation Start” will be filled; if two decks are included in this scaffold request, then “Elevation Start” records the lower deck elevation, “Elevation End” records the higher deck elevation; if three or more decks are required, then the middle decks’ elevations are kept in “Elevation Others”
9. Both estimated and actual volume and man-hours information is tracked.
10. Scaffold information tracking includes crew size, which is “Crew Man Power”, foreman’s name – “Crew Foreman”, superintendent’s name – “Crew Superintendent”, and which shift they are in.
11. Status is a percentage of the request, 0% means cancelled; 100% percentage means finished; other numbers need explanation, which is kept in “Note”.

Appendix 5

Table 26 Evaluation Investigation Results

No.	Attributes	Algorithm			Output	Performance				
						Correlation Coefficient	Error			
		computer learning method	Test Mode	Parameter	Class		Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared
1	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8314	0.067	0.0911	55.3786 %	55.5726%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7361	0.0864	0.1132	71.1209 %	68.6713%

	Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.734	0.0963	0.1221	81.6038 %	77.9371%
2	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr	Linear Regression	100% Training data	M5 Method	Scaffold Mhrs/Trade Mhrs	0.8252	0.0663	0.0926	54.8305 %	56.4772%
		Linear Regression	10-fold-cross validation	M5 Method	Scaffold Mhrs/Trade Mhrs	0.7335	0.0843	0.1133	69.3507 %	68.7144%
		Linear Regression	66%	M5 Method	Scaffold Mhrs/Trade Mhrs	0.7626	0.0857	0.1083	72.6329 %	69.1144%

	Distribution; Day/Night-shift Ratio; Winter Factor									
3	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	100% Training data	Greedy Method	Scaffold Mhrs/Trade Mhrs	0.8094	0.0697	0.0963	57.5667 %	58.7249%
		Linear Regression	10-fold-cross validation	Greedy Method	Scaffold Mhrs/Trade Mhrs	0.7527	0.0802	0.1086	65.9813 %	65.9154%
		Linear Regression	66%	Greedy Method	Scaffold Mhrs/Trade Mhrs	0.7626	0.0857	0.1083	72.6329 %	69.1144%

4	13 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8297	0.0664	0.0915	54.8985 %	55.8137%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7371	0.086	0.1126	70.8105 %	68.3029%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7592	0.0871	0.111	73.7672 %	70.8546%
5	12 ATTRITUBES:Trade; Area_Size;	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8297	0.0664	0.0915	54.8985 %	55.8137%

	Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7473	0.0859	0.1111	70.6539 %	67.3791%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7592	0.0871	0.111	73.7672 %	70.8546%
6	11 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean;H_Max;	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8014	0.0686	0.098	56.6582 %	59.8176%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.6939	0.0896	0.1209	73.7775 %	73.3233%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.6951	0.0959	0.1272	81.2690 %	81.1470%

	H_Mode;Trade Mhrs; Day/Night-shift Ratio; Winter Factor									
7	6 ATTRITUBES:Trade; Area_Congestion Degree; H_Mode;Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8145	0.0671	0.0951	55.4737 %	58.0183%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7514	0.0789	0.109	64.9833 %	66.1433%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7675	0.0869	0.1096	73.6387 %	69.9540%
8	5 ATTRITUBES:Trade; Area_Congestion	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8127	0.0675	0.0955	55.8197 %	58.2678%

	Degree; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7517	0.0788	0.1089	64.8312 %	66.0787%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7642	0.0874	0.1106	74.0024 %	70.5608%
9	14 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min;	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.9201	0.037	0.0647	30.5477 %	39.4653%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.6868	0.0817	0.1268	67.2072 %	76.9366%

	H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.772	0.0813	0.1029	68.9014 %	65.6907%
10	13 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.9172	0.0385	0.0673	31.8168 %	41.0764%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.6965	0.0797	0.1237	65.5898 %	75.0290%
		Garssian Process	66%	Normalized PolyKerner-C250007 -	Scaffold Mhrs/Trade Mhrs	0.7759	0.0795	0.101	67.3669 %	64.4734%

				E 2.0						
11	12 ATTRITUBES:Trade; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.9121	0.0386	0.0675	31.8690 %	41.1933%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.6972	0.0796	0.1235	65.4754 %	74.8988%
		Garssian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.7817	0.0786	0.0997	66.5702 %	63.6464%

12	6 ATTRITUBES:Trade; Area_Congestion Degree; H_Mode;Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.858	0.0562	0.0845	46.4755 %	51.5356%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.4113	0.0899	0.2086	74.0139 %	126.5533 %
		Garssian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.8202	0.0686	0.0911	58.1441 %	58.1568%
13	5 ATTRITUBES:Trade; Area_Congestion Degree; Trade Mhr Distribution;	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.8532	0.057	0.0858	47.0877 %	52.3155%

	Day/Night-shift Ratio; Winter Factor	Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 -E 2.0	Scaffold Mhrs/Trade Mhrs	0.4166	0.0862	0.1939	70.9223 %	117.6499 %
		Garssian Process	66%	Normalized PolyKerner-C250007 -E 2.0	Scaffold Mhrs/Trade Mhrs	0.8232	0.0662	0.0904	56.0792 %	57.7084%
14	4 ATTRITUBES:Trade; Area_Congestion Degree; Trade Mhr Distribution; Day/Night-shift Ratio	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8126	0.0676	0.0955	55.8892 %	58.2991%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7541	0.0779	0.1085	64.1396 %	65.8334%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8233	0.0659	0.0904	55.8527 %	57.6738%

15	7 ATTRITUBES:Trade; Area_Congestion Degree; H_Mean; H_Max, Trade Mhrs, Trade Mhr Distribution; Day/Night-shift Ratio	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8218	0.0666	0.0934	55.0149 %	56.9790%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.748	0.0798	0.1101	65.6629 %	66.8027%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7706	0.0728	0.1016	61.6386 %	64.8291%
16	2 ATTRIBUTES:Trade; Trade Mhr Distribution	Linear Regression	100% Training data	M5 Method	Scaffold Mhrs/Trade Mhrs	0.7461	0.071	0.1091	58.6975 %	66.5840%
		Linear Regression	10-fold-cross validation	M5 Method	Scaffold Mhrs/Trade Mhrs	0.6972	0.0771	0.1177	63.4461 %	71.3693%
		Linear Regression	66%	M5 Method	Scaffold Mhrs/Trade Mhrs	0.7744	0.0801	0.1105	58.5009 %	63.8627%

17	1 ATTRIBUTE: Trade	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7234	0.0725	0.1132	59.9066 %	69.0380%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.6694	0.0787	0.122	64.8199 %	74.0049%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7694	0.0758	0.1114	55.3313 %	64.3829%
18	2 ATTRIBUTES:Trade; Trade Mhrs	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7301	0.0713	0.112	58.9668 %	68.3339%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.6655	0.0789	0.1227	64.9471 %	74.4259%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7382	0.0812	0.1182	59.2760 %	68.3424%

19	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	100% Training data	M5 Method	Scaffold Mhrs/Trade Mhrs	0.833	0.0675	0.0907	55.7611 %	55.3266%
		Linear Regression	10-fold-cross validation	M5 Method	Scaffold Mhrs/Trade Mhrs	0.7305	0.0838	0.1135	68.9905 %	68.8579%
		Linear Regression	66%	M5 Method	Scaffold Mhrs/Trade Mhrs	0.6765	0.103	0.1295	87.2507 %	82.6510%
20	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity;	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8435	0.0652	0.0881	53.8563 %	53.7192%

	Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7259	0.0863	0.1152	71.0133 %	69.9070%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.6996	0.0995	0.1299	84.3085 %	82.8684%
21	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree;	Linear Regression	100% Training data	Greedy Method	Scaffold Mhrs/Trade Mhrs	0.8314	0.0672	0.0911	55.5646 %	55.5644%

	Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	Greedy Method	Scaffold Mhrs/Trade Mhrs	0.726	0.0841	0.1138	69.2379 %	69.0566%
		Linear Regression	66%	Greedy Method	Scaffold Mhrs/Trade Mhrs	0.7141	0.0966	0.1227	81.8269 %	78.3006%
22	17 ATTRIBUTES: Trades; Area_Size; Area_Complexity;	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8415	0.0652	0.0886	53.8539 %	54.0298%

	Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Min; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7176	0.0865	0.1166	71.1986 %	70.7356%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7113	0.1014	0.1264	85.8502 %	80.6821%
23	14 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard;	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8346	0.0664	0.0903	54.8644 %	55.0818%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7067	0.0888	0.1189	73.0442 %	72.1629%

	H_Mean; H_Max; H_Min; V_Mean; V_Min; Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7308	0.0971	0.121	82.2187 %	77.2342%
24	6 ATTRITUBES:Trade; Area_Congestion Degree; H_Mode;Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8145	0.0671	0.0951	55.4737 %	58.0183%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7514	0.0789	0.109	64.9833 %	66.1433%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7675	0.0869	0.1096	73.6387 %	69.9540%
25	5 ATTRITUBES:Trade; Area_Congestion	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8127	0.0675	0.0955	55.8197 %	58.2678%

	Degree; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7517	0.0788	0.1089	64.8312 %	66.0787%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7642	0.0874	0.1106	74.0024 %	70.5608%
26	15 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8372	0.0668	0.0896	55.2066 %	54.6823%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7399	0.0846	0.1121	69.6124 %	68.0005%

	Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7135	0.1006	0.1299	85.2360 %	82.8761%
27	19 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_StDev; H_Max; H_Min; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.8652	0.0623	0.0905	51.5270 %	55.2066%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.736	0.0799	0.1142	65.7180 %	69.2935%
		Garssian Process	66%	Normalized PolyKerner-C250007 -	Scaffold Mhrs/Trade Mhrs	0.685	0.0899	0.1235	76.1464 %	78.7980%

				E 2.0						
28	17 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mean; H_Max; H_Min; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.8663	0.0616	0.09	50.9450 %	54.8893%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.7382	0.0791	0.1137	65.0714 %	68.9795%
		Garssian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.6895	0.0878	0.1213	74.3844 %	77.4119%

29	15 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean; V_StDev; V_Max; V_Min; V_Mode;Trade Mhrs; Trade Mhr Distribution;	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.8636	0.0619	0.0905	51.1545 %	55.2219%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.7364	0.079	0.1139	64.9909 %	69.1056%

	Day/Night-shift Ratio; Winter Factor	Garssian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.7305	0.0837	0.1109	70.8779 %	70.7930%
30	12 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean;V_Min; Trade Mhrs; Trade	Garssian Process	100% Training data	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.9111	0.0388	0.0679	32.0944 %	41.4140%
		Garssian Process	10-fold-cross validation	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.6729	0.0826	0.13	67.9651 %	78.8560%

	Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Garssian Process	66%	Normalized PolyKerner-C250007 - E 2.0	Scaffold Mhrs/Trade Mhrs	0.7694	0.0799	0.1022	67.6615 %	65.2110%
31	12 ATTRIBUTES: Trades; Area_Size; Area_Complexity; Area_Congestion Degree; Area_Distance to Material Yard; H_Mode; V_Mean;V_Min; Trade Mhrs; Trade Mhr Distribution; Day/Night-shift Ratio; Winter Factor	Linear Regression	100% Training data	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.8273	0.0669	0.0921	55.2908 %	56.1828%
		Linear Regression	10-fold-cross validation	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7314	0.0843	0.1139	69.3690 %	69.1162%
		Linear Regression	66%	No attribute selection	Scaffold Mhrs/Trade Mhrs	0.7033	0.1068	0.1346	90.4555 %	85.9059%