BIM-Based Data Mining Approach to Estimating Job Man-Hour Requirements in
Structural Steel Fabrication

by

Xiaolin Hu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Construction Engineering and Management

Department of Civil and Environmental Engineering
University of Alberta

ABSTRACT

In a steel fabrication shop, jobs from different clients and projects are generally processed simultaneously in order to streamline production processes, improve resource utilization, and achieve cost-effectiveness in serving multiple concurrent steel-erection sites. Reliable quantity takeoff on each job and accurate estimation of shop fabrication man-hour requirements are crucial to plan and control fabrication operations and resource allocation on the shop floor. Building information modeling (BIM) is intended to integrate multifaceted characteristics of a building facility, but finds its application in structural steel fabrication largely limited to design and drafting. This research focuses on extending BIM's usage further to the planning and control phases in steel fabrication. Using data extracted from BIM-based models, a linear regression model is developed to provide the man-hour requirement estimate for a particular job. Actual data collected from a steel fabrication company was used to train and validate the model. Two Excel macro-enabled workbooks were also developed to provide decision-making support in fabrication planning.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere thanks to my two academic advisors, Dr. Ming Lu and Dr. Simaan M. AbouRizk for their vision, support, and guidance in the preparation of this thesis.

Thanks are extended to Darrell Mykitiuk, Robert Wright, Jim Kanerva and the staff at Waiward Steel Fabricators Ltd. (WSF) who shared their knowledge and experience and provided data for this research. Their cooperation and support are very much appreciated.

I am also grateful to all members at Construction Engineering and Management group for their support and assistance during my graduate program.

Finally I would like to thank my family and my boyfriend, Sheng Mao, for their continuous support and encouragement throughout my study at the University of Alberta.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. INTRODUCTION

1.1 BACKGROUND

Steel has long been the most important material to the construction sector for its strength, durability, flexibility, efficiency, sustainability, and versatility (SteelConstruction.info 2014). The production of steel pieces, which includes a variety of operations of detailing, fitting, welding, and surface processing, is a complex and critical process for a typical steel construction project. Most steel construction projects use off-site structural steel fabrication shops to support the erection sites in order to increase the productivity, gain better control over quality, and reduce the total cost of projects (Eastman and Sacks 2008). A steel fabrication shop usually makes use of shift work and serves multiple steel erection sites at the same time to keep the business economical. Efficient planning is substantial to steel fabrication to ensure a streamlined and delay-free production process.

Figure 1-1 shows the structure of a typical construction project (Dozzi and AbouRizk 1993). Personnel, materials, equipment, and management are consumed by the system as resources to produce the construction units. As the foundation of further planning and scheduling, estimating plays a critical role to every construction project. Quantity takeoff is the most time-consuming yet extremely important task in estimating. The following project scheduling and control would benefit a great deal if quantity takeoff could be done accurately and in a timely manner. For example, it can be used to foresee and plan the construction activities during the pre-construction stage; in the process of

construction, quantity takeoff can be used as a measurement of the project progress or for financial performance control of the project (Monteiro and Poças Martins 2013).
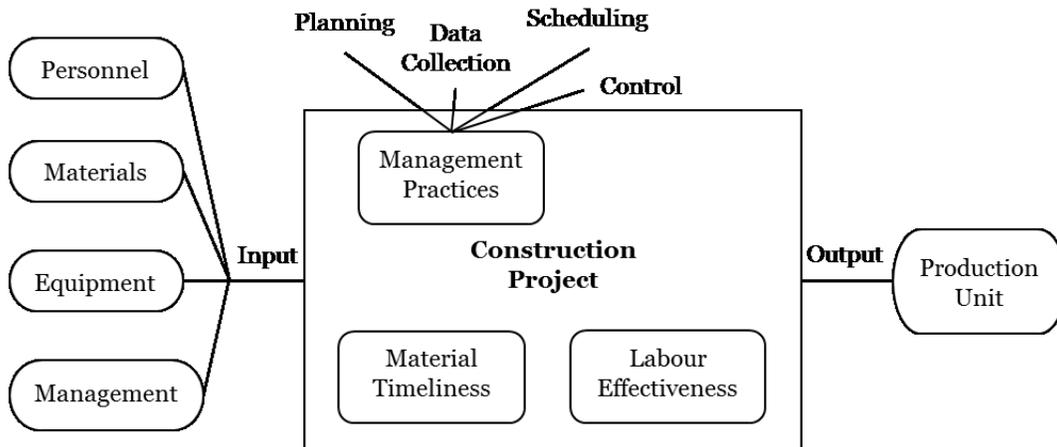
Figure 1-1: Frame for productivity improvement (Dozzi and AbouRizk 1993)

The measurement unit for workload for steel fabrication projects can be the number of steel pieces, weight of the final product, project duration, or monetary value. With the nature of steel fabrication being labour-intensive, man-hours are normally used as the major input for the steel fabrication processes (Dozzi and AbouRizk 1993). The other resources, such as labor, equipment, and overhead costs, are also closely correlated to man-hours. Therefore, it is most suitable to set the output of quantity takeoff as the man-hours needed to complete the project. In addition, the ratio of man-hours over the overall steel weight can be an excellent measure of production efficiency, i.e. productivity.

As defined by National Building Information Model Standard Project Committee (2014), BIM is "a shared knowledge resource for information about a facility forming a reliable basis for decisions during its life-cycle." The concept of

BIM has been rapidly gaining popularity and acceptance since Autodesk released the BIM white paper (Autodesk 2003). Ideally, the vitality of a BIM-based model spans the entire life-cycle of a project, from earliest conception to completion, supporting processes like planning, design, cost control, construction management, etc. This relatively new technology has also been adopted by the steel fabrication industry, but only to find its use limited mostly to design and drafting (Sattineni and Bradford 2011). Most of the advantages that BIM offers, such as increased coordination of documents and effective information communication and decision support for project management, are not exploited. BIM-based models are utilized solely as 3D visualization in most cases. The collaborating steel fabrication company for this research uses BIM software Tekla to build 3D models for structural visualization, and generate 2D drawings for the fabrication shop.

1.2   PROBLEM STATEMENT

A series of interviews with the estimators and project managers in the steel fabrication industry reveal that the current estimating practice followed by most steel fabricators is a manual process using spreadsheets and 2D drawings generated by computer aided design (CAD) software or exported from BIM-based models. Even with the availability of BIM, estimators use it as a visualization tool to help them with reading the 2D drawings. Estimators use their experiences to evaluate the project complexity and estimate the workload. The factor of human interpretation in the process determines the error-proneness of the process.

The collaborating company is a leader in the steel fabrication and construction services industries, offering services of procurement, engineering, 3D modeling, fabrication, coating, module assembly, erection, etc. Current

practice uses Tekla software (Tekla 2014a) to create 3D models from a customer's drawings, and further produce erection and fabrication drawings.

As shown in Figure 1-2, large projects are typically broken down in a hierarchical fashion from a project into one or more jobs, from jobs into one or more divisions, which is of the proper size to manage and to be processed in different shops. Shops are identified with different equipment and labor settings. For example, shop "A" is equipped with a 40-ton overhead crane, making it suitable to handle super assembly structures; shop "B" is set up to handle frames. A division is normally about 20 – 50 tons, consisting of multiple pieces. It is the basic unit for the estimators and project managers to manage projects. The estimators or fabrication shop managers use their experiences to evaluate the division complexity and come up with a labour productivity value measured by man-hours per tonne, which is to be multiplied by the overall weight of steel in order to get the man-hours budget needed to complete the work.
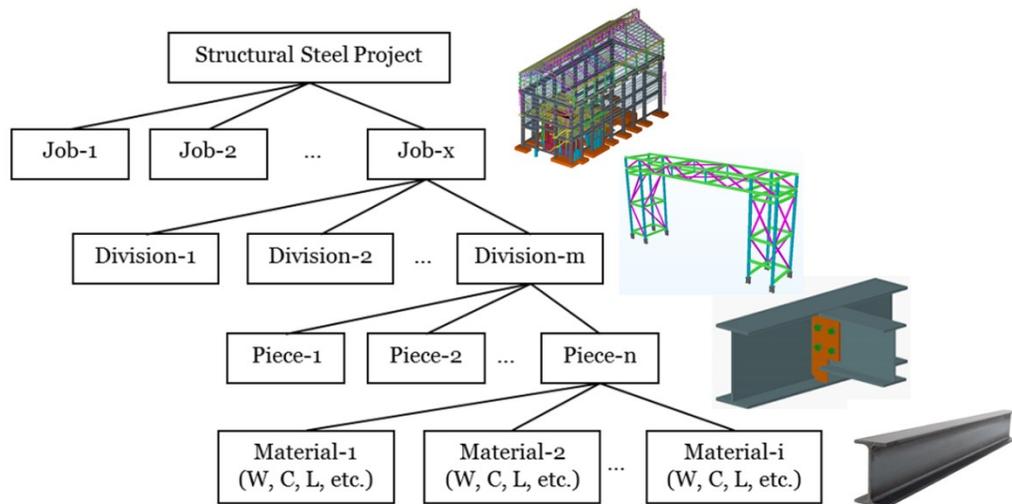
Figure 1-2: Hierarchy of a steel fabrication project

4

The effectiveness of this practice depends to a great extent on personal experience and knowledge, and may not always be consistent and reliable. The abundant information contained in BIM, such as design details and predefined or user-defined material properties, is not exploited properly.

Furthermore, job compositions of steel fabrication projects can vary greatly from one to another. Even within the same job or division, the labour requirements per unit weight of different material types are generally different. For example, a piece demanding extensive welding obviously requires more man-hours than a super-assembly connected by bolts.

1.3   RESEARCH OBJECTIVES

The objectives of the research presented in this thesis include:

- Investigating the common applications of BIM in structural steel fabrication;

- Understanding the current estimating practices in steel fabrication shops;

- Exploring the possibility of extending BIM's usage further to the fabrication planning and control phase;

- Providing decision support as to bid or not to bid after evaluating the current and future workload;

- Comparing three different modeling methods (linear regression, SVM regression, and RBF neural network) to model steel fabrication workload in terms of man-hours;

- Providing a quantitative approach to the prediction of fabrication man-hour requirements for structural steel projects by analyzing and learning from the historical schedules and cost information

stored in the company's central database for the benefits of detailed estimating.

1.4   METHODOLOGIES

This study makes use of the material parts report generated from Tekla. The essential attributes at the level of materials, as well as the summary level of divisions, are collected and analyzed for 298 jobs and 1605 divisions completed by the collaborating steel fabricator from 2009 to 2013. Only jobs that include "supply work" are considered because erection is a process almost completely separate from shop fabrication.

The first stage of this research is to design a meaningful data structure to sort out and organize the data at different levels, and to collect necessary information from the large database.

After historical data are collected, a regression model is developed. The basic attributes of different material types are defined as independent input variables. The man-hours needed to fabricate a division are defined as the output variable. An open-source software, WEKA (Hall et al. 2009), is chosen to complete the data mining task because of its wide collection of machine learning algorithms and various regression functions. The selection of contributing factors and the optimization of the variables through iterative experiments are all facilitated by using WEKA. Different modeling methods are tested and compared to find a suitable model for the workload of steel fabrication in terms of man-hours.

At the third stage, the developed model is verified through an independent dataset and the prediction results are compared with the forecast made by personal judgment.

1.5   THESIS ORGANIZATION

The thesis starts with an introductory chapter that presents an overview of the entire thesis, including the background, problem statement, research objectives, and methodologies used.

Chapter 2 provides a thorough review of previous studies related to construction estimating, data mining, application of regression analysis in the construction field, and implementation of data mining algorithms.

Chapter 3 explains the raw data structure in detail and describes how the data were prepared for modeling.

Chapter 4 presents the modeling process. A real case study from the collaborating steel fabrication company is conducted as an example to illustrate the validity, suitability and usefulness of the proposed method.

Chapter 5 demonstrates two automatic spreadsheet tools that can spread the workload in the shop on a weekly basis in order to facilitate shop operation planning.

Chapter 6 concludes the thesis with a summary of what has been achieved, and outlines a proposal for future enhancements.

Chapter 2.   BACKGROUND & LITERATURE REVIEW

2.1   QUANTITY TAKEOFF

Traditionally, a material takeoff (MTO) refers to the result or the process of generating a list of required materials with quantities and other specifications to accomplish a design by analyzing the drawings, blueprints, or other design documents (Whitt 2012). Takeoff is followed by the estimating process, which is to apply costs to the quantity measurements. Sometimes the terms *quantity takeoff* and *estimating* may be used interchangeably if the desired results use the same unit of measurements. As shown in the classic cost influence curve (Figure 2-1), the ability to influence the project outcome is the greatest and the cost is the lowest when the project is in the early stages. If quantity takeoff and estimating could be done accurately early in the project, the following project scheduling and control would benefit a great deal.
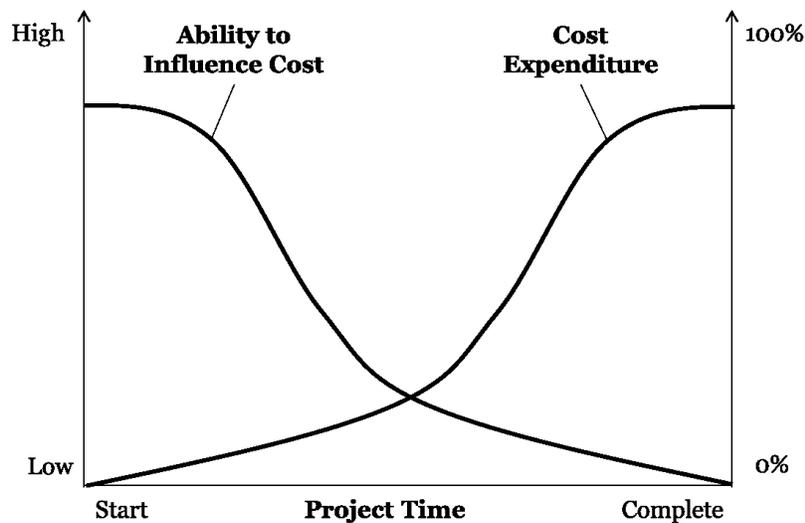
Figure 2-1: Cost influence curve for project lifecycle (CII 1995)

To perform quantity takeoff, several methods are available in the construction industry. Traditional estimators do their takeoffs manually with printed drawings. They would use colorful markers to keep track of different materials and enter relevant information onto leger sheets or spreadsheets for calculation. Figure 2-2 is a takeoff form template used by a company being investigated during the present research. The quantity, unit weight, section detail, etc. need be filled in manually for each line item. Some estimators adopt simple annotation software to view electronic drawings and perform color-coding, etc., but the process is still manual in essence (Vertigraph Inc. 2004). Special estimating software is another approach, but its input still relies heavily on human interpretation.

TAKEOFF FORM

			DESCRIPTION						Estimate No:				Date:	
													Pg.	of
Cat.	Line	Qty.	Section	unit wt	Length	item wt	Manhours	Line	Qty.	Section	unit wt	Length	item wt	Manhours
			main material				each / total			detail material				each / total
	0.01					0	0	0.36					0	0
	0.02					0	0	0.37					0	0
	0.03					0	0	0.38					0	0
	0.04					0	0	0.39					0	0
	0.05					0	0	0.40					0	0
	0.06					0	0	0.41					0	0
	0.07					0	0	0.42					0	0
	0.08					0	0	0.43					0	0
	0.09					0	0	0.44					0	0
	0.10					0	0	0.45					0	0
	0.11					0	0	0.46					0	0
	0.12					0	0	0.47					0	0
	0.13					0	0	0.48					0	0
	0.14					0	0	0.49					0	0

Figure 2-2: An example of takeoff form

As stated by Tiwari et al. (2009), "Model-based cost estimating is the process of integrating the object attributes from the 3D model of the designer with the cost information from the database of the estimator." Adopting BIM for

managing the design and construction process of projects has proven to be a shared understanding (Aranda-Mena et al. 2009). BIM-based estimating would assure the reduction of errors resulting from the repetitive manual entry of data, allow high accuracy and standardisation in estimate production, which improves estimators' productivity. As commented by Monteiro and Poças Martins (2013), BIM-based quantity takeoff is "one of the potentially most important and profitable applications for BIM." Yet, it is still generally underdeveloped and underutilized.

2.2 STRUCTURAL STEEL FABRICATION

Steel is a widely used building material throughout the construction industry because of its ability to suit different requirements of strength, weldability, corrosion resistance, etc. (Williams 2011). It works like a skeleton to hold the building structure up and together. When compared to other structural building materials steel has a great many advantages. Unlike wood, steel does not bend, twist, expand, or contract substantially because of the weather and temperature. Unlike concrete, steel does not have a curing process and is at full strength as soon as it is completed. Steel has more strength with less weight and durability. Steel structures require little maintenance, do not age or decay as fast as the other construction materials, and last longer (SteelConstruction.info 2014). Steel construction is cost-efficient and can take place in most weather conditions. Furthermore, steel is 100% recyclable and can be multi-cycled without losing quality, making it one of the most environmentally friendly building materials. There are generally a few stages in a typical steel construction project: design, procurement, steel fabrication, shipment, optional module assembly, and on-site erection, among which steel fabrication is a very critical part (Azimi et al. 2011).

Fabrication is defined by Berman (2014) as "the act of changing steel from the mill or warehouse into the exact configuration needed for assembly into a shipping piece or directly into a structural frame." It mostly takes place in an offsite fabrication shop that is highly regulated, controlled, confined, safe, and equipped with leading edge specialized fabrication systems. All structural steel components, such as columns, beams, channels, and plates, can be carefully designed and precisely fabricated before delivery to site to be assembled and erected.

The systematic fabrication process generally consists of a series of operations including cutting, grinding, drilling, burning, fitting, welding, and surface processing (painting, sand blasting, fireproofing etc.). The whole shop floor is divided into several main areas according to the specific functions, and each shop (for instance, cutting shop, fitting shop, welding shop, and painting shop) is equipped with specialized machines, tools, and skilled personnel. The inputs of a steel fabrication shop are raw steel materials and shop fabrication drawings, and the outputs are fabricated steel components that are ready to be assembled and shipped to site for erection (Figure 2-3).
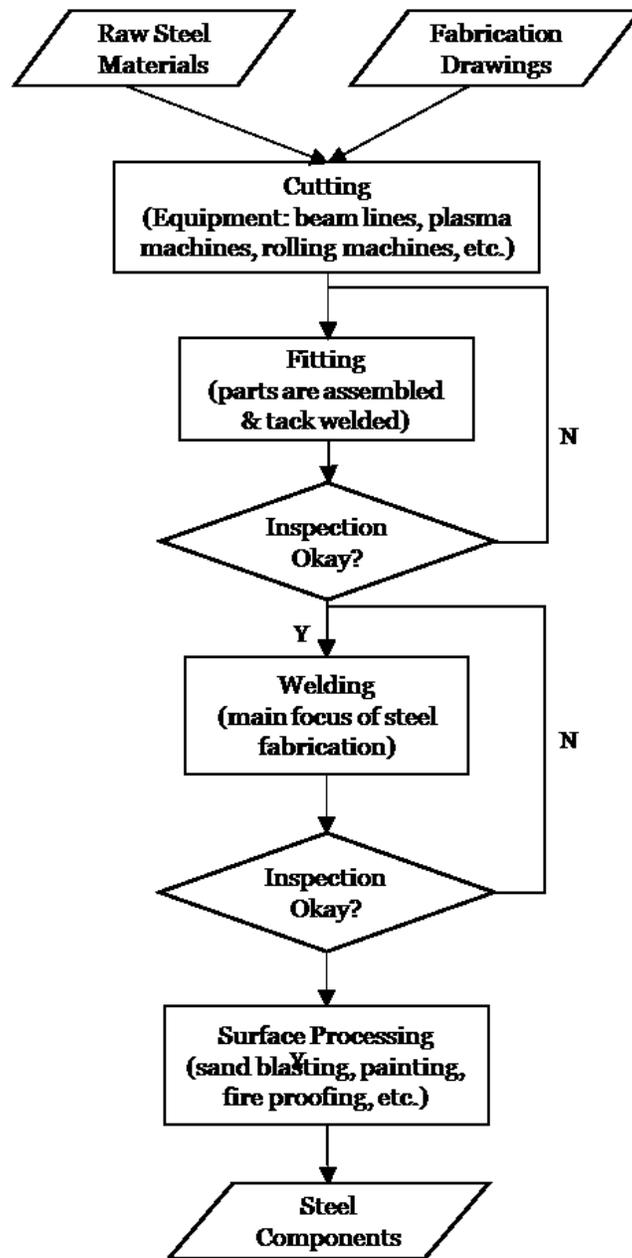
Figure 2-3: Steel fabrication processes

A Shop Operations Manager is responsible for spreading and scheduling jobs throughout the multiple shops on the shop floor, in collaboration with the respective Project Manager. In order to maintain a balanced workload, high

production rate, and streamlined operations, the manager needs to define the scope of a job well in terms of man-hours, which is what this study focuses on.

2.3   BUILDING INFORMATION MODELING (BIM)

According to the National Building Information Model Standard Project Committee (2014),

Building Information Modeling (BIM) is a digital representation of physical and functional characteristics of a facility. A BIM is a shared knowledge resource for information about a facility forming a reliable basis for decisions during its life-cycle; defined as existing from earliest conception to demolition.

A basic premise of BIM is collaboration by different stakeholders at different phases of the life cycle of a facility to insert, extract, update or modify information in the BIM to support and reflect the roles of that stakeholder.

The US National BIM Standard will promote the business requirements that BIM and BIM interchanges are based on:

- *a shared digital representation,*

- *that the information contained in the model be interoperable (i.e.: allow computer to computer exchanges), and*

- *the exchange be based on open standards,*

- *the requirements for exchange must be capable of defining in contract language.*

The concept of BIM has been rapidly gaining popularity and acceptance since Autodesk released the BIM white paper (Autodesk 2003). Ideally, the vitality of a BIM-based model spans the entire life-cycle of a project, from earliest conception to completion, supporting processes like planning, design, cost control, construction management etc. BIM solutions can be customised and applied to various areas, for instance, concrete construction, steel fabrication, steel erection, rebar fabrication, and structural design. Figure 2-4 (Tekla 2014b) shows the workflow and the integration of all the services in steel fabrication industry. Ideally, the bidding, preconstruction, construction, and post construction of a project can all be managed through BIM as a whole instead of jumping across multiple software and systems, avoiding having to deal with abundant document format transformations. It is a platform to share knowledge among different project stakeholders, providing consistent and coordinated representations of the digital model.
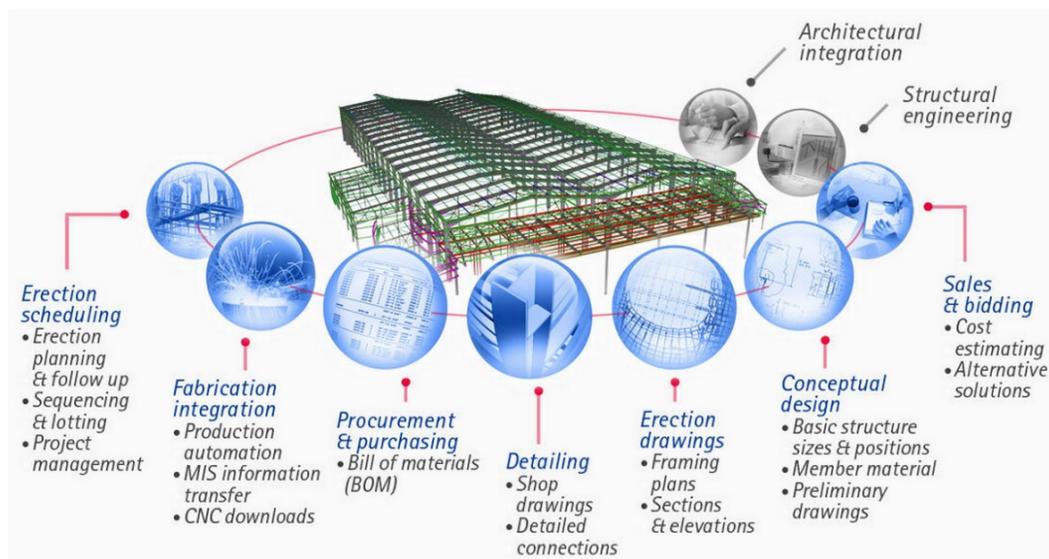
Figure 2-4: BIM workflow (Tekla 2014b)

Figure 2-5 (Tekla 2014b) shows a portion of a BIM-based model, along with a view of the component coding system the model uses. It represents a steel structure that consists of beams, columns, handrails, and is connected with bolts.
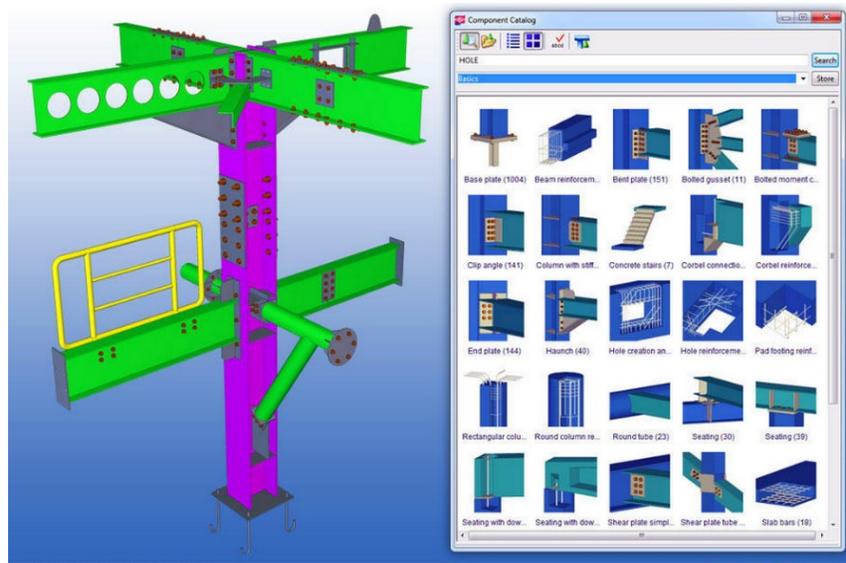
Figure 2-5: A BIM example of steel structures (Tekla 2014b)

To seize the full potential value of BIM, contractors cannot limit their exploration of BIM to 3D modeling and visualization only. 3D rendering is the basic use of BIM. BIM can also be used to detect clashes and conflicts. Detailed fabrication drawings can be generated for different trades. Change orders and addendums can easily be communicated between different parties. 3D models combined with other planning techniques and tools can provide powerful construction monitoring, which in turn helps with scheduling and updating 3D models (Hergunsel 2011).

The Industry Foundation Classes (IFC) data model is "a platform neutral, open file format specification" intended to provide a set of consistent data representations of building and construction industry information (Eastman

2006). It is developed by buildingSMART to facilitate interoperability in the Architecture, Engineering, and Construction (AEC) industry. Add-ons and extensions can be developed using IFC format to facilitate the communication between different BIM-related systems.

Significant research has been carried out in exploration of BIM. The study conducted by Howard and Björk (2008) is an overview of experts' views on BIM. It collects some pilot use cases and BIM user experience from a number of leading property owners in spite of the complexity of the formal standards such as the IFCs. Aranda-Mena et al. (2009) provided insights into BIM and illustrated its importance and potential applications in construction project management industry through case studies. Steel, Drogemuller, and Toth (2012) presented their experience with model-based interoperability issues, successes, and challenges in BIM exchange between various tools; a business case framework to facilitate the adoption of BIM was proposed. Jung and Joo (2011) developed a comprehensive BIM framework consisting of three dimensions and six categories, which provides a basis for the evaluation of practical BIM effectiveness. Nawari (2012) reviewed the importance of BIM standards in off-site construction and its role in data exchange. An Information Delivery Manual (IDM) was also proposed, which provides material description of building construction processes, information requirements, and expected process outputs in the study. Furthermore, BIM can provide support to teaching construction project management (Peterson et al. 2011). The introduction of BIM-based project management tools helped educators design more realistic project-based assignments and cases, and supported students with learning the integration and application of different project management functions.

Ikerd (2008) justified the importance of BIM in structural engineering. Shi (2009) proposed a framework of integrating Radio Frequency Identification (RFID) with BIM technologies for the decision-making processes in structural steel fabrication and erection. A portable RFID database scheme was developed to increase the efficiency and accuracy in steel fabrication and erection. Xie, Shi, and Issa (2010) further discussed the BIM/RFID implementation in computer-aided design, manufacturing, engineering, and installation processes. Tiwari et al. (2009) applied BIM tools for Target Value Design (TVD) on a large healthcare project in Northern California. A 4D simulation for a steel arch bridge was produced to illustrate the use of BIM tools in a design review and lifting plan study (Chiu et al. 2011). Lancaster and Tobin (2013) outlined their firms' extensive experience with BIM, providing strategies and new understandings of applying BIM to structural engineering projects aimed to accommodate Integrated Project Delivery (IPD). Kalavagunta (2012) presented an integrated structural modeling workflow for structural design. Sattineni and Bradford (2011) conducted a survey of construction practitioners in United States to determine the usage of BIM in various tasks, especially in construction cost estimating. Monteiro and Poças Martins (2013) also explored automatic BIM based quantity takeoff and a case study was presented. As one of its conclusions, the authors suggested that takeoff specifications such as manual measurements should be revised to account for BIM features in order to minimize the limitations in current practice.

2.4 MACHINE LEARNING

People have been looking for information in the sea of data ever since human beings became intelligent, and the identification of potentially useful

information or patterns hidden in the huge amount of data around us is what experience, knowledge, and intelligence are actually about. Machine learning, as a branch of artificial intelligence (AI), is about the study and construction of systems, other than human brains, that can solve problems by analyzing and learning from data. In 1959, Arthur Samuel defined machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed" (Simon 2013). It provides tools to make predictions automatically or help people make decisions about complex and scaled problems from data in a faster and more accurate way. A more formal and widely quoted definition of machine learning is provided by Mitchell (1997):

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

This fundamentally operational definition makes it clear that the class of tasks, the source of experience, and the measure of performance to be improved are the three features that have to be identified in order to have a well-defined learning problem (Mitchell 1997).

The term *machine learning* is commonly confused with *data mining*. These two areas overlap significantly in the methods they employ, focusing on slightly different goals. As mentioned before, machine learning relates to the study and development of learning algorithms and focuses on prediction in most cases. Data mining, on the other hand, can be defined as the process of trying to extract previously unknown knowledge, properties, or patterns from unstructured data. It focuses on the discovery aspect. Data mining may utilize

machine learning algorithms during the process, and may also drive the advancement of machine learning techniques (Cross Validated 2013; ResearchGate 2013).

2.4.1 Machine Learning Algorithms

A popular taxonomy of organizing machine learning algorithms is based on the learning styles algorithms can adopt (Brownlee 2013a):

- Supervised Learning: Algorithms are trained on input data that have a known label or desired result, such as sunny/rainy or spam/not-spam. Such an algorithm attempts to create a model to make predictions of the outputs according to the inputs. The model is like a function or mapping from the inputs to outputs. Once a desired level of accuracy is achieved (i.e. the predictions and the labels are close enough), the trained model is able to generate outputs for inputs that have not been used in the training process. Classification and regression problems fall into this category. Example algorithms are Decision Trees, Stepwise Regression and Back-Propagation Neural Networks.

- Unsupervised Learning: Training examples are not labelled and do not have a known result. Instead of generalising a function or mapping from inputs to outputs, a model is prepared by discovering structures present in the input data. Example algorithms are K-Means Clustering and Apriori Algorithm.

- Semi-Supervised Learning: Input data consist of both labelled and unlabelled examples. The desired model needs to be able to make predictions as well as deducing the structures in the data. An

example problem would be image classification where only few examples are labelled in a large dataset.

- Reinforcement Learning: Input data are provided as stimulus. A model attempts to gather knowledge in an environment through punishment or reward feedbacks about its actions. The goal is to maximize some cumulative reward. Example algorithms are Q-Learning and Temporal Difference Learning. This type of learning is more likely to be used in certain kinds of control system development.

Another grouping method is by algorithm similarity. For example, regression methods, decision tree methods, instance-based methods, associate rule learning, clustering methods, and artificial neural networks. In the following section, only the basic linear regression algorithm is introduced as it is the method used in this research.

2.4.2 Linear Regression

Linear regression is actually a fundamental method in statistics, suitable for situations where most or all the attributes are numeric. The basic idea is to express the model as a linear mapping from the attributes to the output class. The goal is to come up with "a function that approximates the training points well by minimizing the prediction error" (Witten, Frank, and Hall 2011). A model is represented as:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_k a_k \qquad (2\text{-}1)$$

where x is the outcome or class; $a_1, a_2, \ldots, a_k$ are the numeric attribute values; and $w_0, w_1, w_2, \ldots, w_k$ are weights for each attribute.

The weights are calculated from the training data. Each training instance has its own set of attribute values $a_1^{(i)}, a_2^{(i)}, \ldots, a_k^{(i)}$ and the outcome $x^{(i)}$. Assuming an extra attribute a_0 with a constant value of 1, then the predicted value for the class can be conveniently written as:

$$w_0 a_0^{(i)} + w_1 a_1^{(i)} + w_2 a_2^{(i)} + \cdots + w_k a_k^{(i)} = \sum_{j=0}^{k} w_j a_j^{(i)} \qquad (2\text{-}2)$$

The method of linear regression is to look for a set of numeric weights $w_0, w_1, w_2, \ldots, w_k$ to make the predicted values as close to the actual values as possible; in other words, to minimize the sum of the squares of the differences over all the training instances (Witten, Frank, and Hall 2011). In order to choose coefficients properly, the function shown in (2-3) is the target to be minimized. This is the classic least-squares linear regression method.

$$\sum_{i=1}^{n} \left(x^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)} \right)^2 \qquad (2\text{-}3)$$

Once a set of numeric weights has been calculated based on the training data, the prediction of the outcome of new instances can be accomplished using the formula.

Aside from the complete numeric cases, linear regression is able to handle nominal attributes as well. In contrast to the continuous nature of numeric attributes that measure real or integer numbers, nominal attributes handle a pre-defined set of values and are sometimes called categorical attributes. The finite set of values serve only as names or symbols (Witten, Frank, and Hall 2011). The trick of applying linear regression to nominal attributes is to view each possible value of the nominal attributes as a binary attribute, whose value is either 0 or 1. There are more advanced variations of the standard linear regression, such as logistic regression and multivariate linear regression, which is not covered in this research.

2.4.3 SVM Regression

Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms, belonging to a family of generalized linear classifiers. The current standard "soft margin" method was proposed by Cortes and Vapnik (1995) on the basis of the original algorithm invented by Vladimir N. Vapnik in 1979. SVM can be applied not only to classification problems but also to regression analysis for its ability of analyzing data and recognizing patterns. The basic idea of SVM is to find a hyperplane that divides data points into two classes with the largest separation or margin, which is defined by the distance from the hyperplane to the nearest data point of each class. As shown in Figure 2-6, only H_3 separates the classes with the maximum margin.
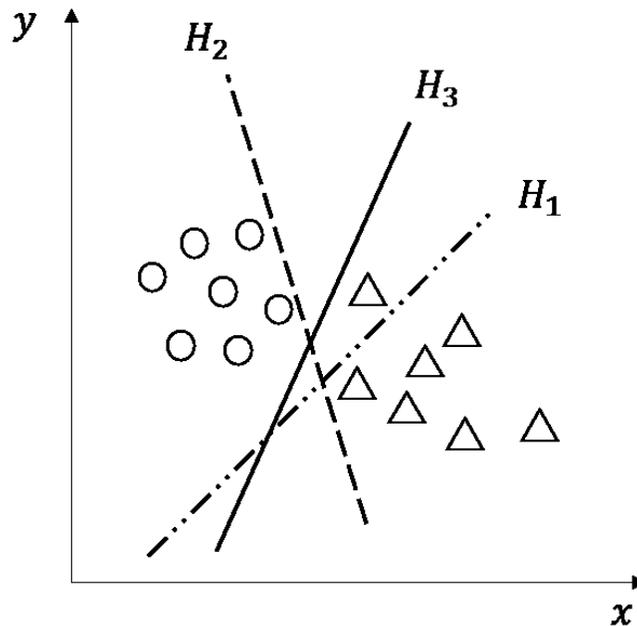
Figure 2-6: One-dimensional hyperplanes

The traditional support vector machine can only utilize complex and expensive quadratic programming (QP) solvers in optimization, until the introduction of Sequential Minimal Optimization (SMO) method, which greatly simplifies the optimization into two analytically solvable problems (Flake and Lawrence 2002).

SVM regression is a version of SVM proposed in 1996, which can also be called Support Vector Regression (SVR). The cost function for building the model ignores any training points that lie beyond the threshold. Therefore the model produced by SVR actually depends only on a subset of the training data. The basics of SVR is illustrated below (Smola and Schölkopf 2004; Cortes and Vapnik 1995; Cross Validated 2011; "Support Vector Machine Regression" 2014).

Suppose a training dataset consisting of n points is given as:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R} ,$$

where $\mathbb{R}$ is the set of real numbers.

The x_i is a d-dimensional real vector. The target is to find a function $f(x)$ such that the deviation from the actual target y_i for all the training data is within the threshold ε. At the same time, the flatness of the function needs to be maximized to minimize the sensitivity to errors in the data points.

The linear regression function $f(x)$ is in the form:

$$f(x) = \omega \cdot x + b, \omega \in \mathbb{R}^d, b \in \mathbb{R} \qquad (2\text{-}4)$$

The $\cdot$ denotes the dot product and ω the normal vector to the hyperplane. In ε-SV regression (Cortes and Vapnik 1995) where ε is a threshold, f needs to satisfy (for any $i = 1, \dots, n$):

$$|y_i - f(x_i)| \leq \varepsilon \qquad (2\text{-}5)$$

And maximization of the flatness of the function in (2-4) can be achieved by minimizing the norm of ω, i.e. $\|\omega\|$. Consequently the objective is to solve the following optimization problem (Smola and Schölkopf 2004):

$$\text{minimize } \|\omega\|$$

$$\text{subject to } \begin{cases} y_i - \omega \cdot x_i - b \leq \varepsilon \\ \omega \cdot x_i + b - y_i \leq \varepsilon \end{cases} \tag{2-6}$$

However, a function f that satisfies all pairs (x_i, y_i) with ε precision may not actually exist. Moreover some errors also need to be allowed for. Accordingly the infeasible constraints of the optimization problem (2-6) are loosened by introducing non-negative slack variables ξ_i^+, ξ_i^- that are used in the "soft margin" cost function in SVM. Hence the optimization problem is transformed to (Smola and Schölkopf 2004):

$$\text{minimize } \frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^{n}(\xi_i^+ + \xi_i^-), (C \text{ is a positive constant})$$

$$\text{subject to } \begin{cases} y_i - \omega \cdot x_i - b \leq \varepsilon + \xi_i^+ \\ \omega \cdot x_i + b - y_i \leq \varepsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \geq 0 \end{cases} \tag{2-7}$$

The solution to the problem above is to construct a Lagrange function from the objective function as (Smola and Schölkopf 2004):

$$L \triangleq \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) - \sum_{i=1}^{n}(\alpha_i\xi_i^+ + \beta_i\xi_i^-) - \sum_{i=1}^{n}\gamma_i(\varepsilon + \xi_i^+ - y_i + \omega \cdot x_i + b)$$

$$- \sum_{i=1}^{n}\delta_i(\varepsilon + \xi_i^- + y_i - \omega \cdot x_i - b) \tag{2-8}$$

L is the Lagrangian and $\alpha_i, \beta_i, \gamma_i, \delta_i$ are Lagrange multipliers. Having derived the Lagrange function, the Support Vector expansion is conducted as (Smola and Schölkopf 2004):

$$\omega = \sum_{i=1}^{n}(\gamma_i - \delta_i)x_i \tag{2-9}$$

Hence:

$$f(x) = \sum_{i=1}^{n}(\gamma_i - \delta_i)(x_i \cdot x) + b \hspace{2cm} (2\text{-}10)$$

According to (2-9), ω can be completely calculated by the linear combination of the training data points x_i.

The constant offset b can be computed via various methods such as exploiting the Karush-Kuhn-Tucker (KKT) conditions and as a by-product of the optimization process (Smola and Schölkopf 2004).

2.4.4 RBF Neural Network

A radial basis function (RBF) neural network is an artificial neural network widely used for functional approximation and prediction in areas such as time-series modeling, system control and pattern classification. The name comes from its use of radial basis functions as activation functions (Broomhead and Lowe 1988).

In a RBF network there are basically three layers with different roles as shown in Figure 2-7: an input layer, a hidden layer and an output layer. The first layer is simply a fan-out layer, acting as a connection between the network and the environment. No processing is done. The second layer, i.e. the hidden layer, transforms the nonlinear input space to the hidden space, which in most cases is higher dimensional. The last one, output layer, applies a linear transformation (Haykin 1999). The rationale is justified by Cover's theorem on the separability of patterns (Cover 1965). By using a nonlinear mapping to transform the input space in a higher dimensional space, the complex patterns can be more linearly separable. The nonlinear mapping is then followed by a linear mapping from the hidden space to the output space.
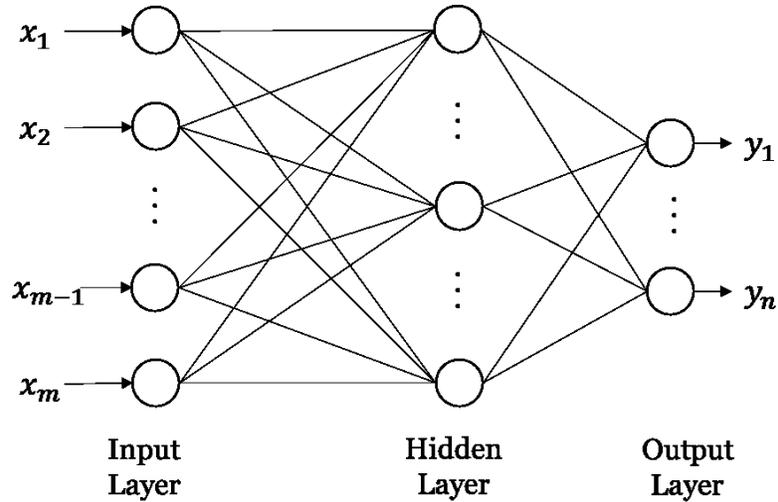
Figure 2-7: Architecture of RBF network (Haykin 1999)

Define a set of input vectors (or patterns) $X = \{x_1, x_2, \ldots, x_N\}$, in $\mathcal{R}^m$ space. Each of these vectors are assigned to one of two classes X_1 and X_2. Define a set of hidden functions $\{\varphi_i(x): \mathcal{R}^m \rightarrow \mathcal{R} | i = 1, 2, \ldots, p\}$. For each vector , define a vector $\varphi(x) = \left[\varphi_1(x), \varphi_2(x), \ldots, \varphi_p(x)\right]^T$. Then the vector $\varphi(x)$ maps patterns of dimension m into corresponding points in a p-dimentional space, which is referred to as the hidden space. If there exists a p-dimentional vector $\boldsymbol{w}$ such that (Cover 1965):

$$\begin{cases} w^T \varphi(x) > 0, x \in X_1 \\ w^T \varphi(x) < 0, x \in X_2 \end{cases} \tag{2-11}$$

Then X is φ-separable, and the separating surface in the hidden space is defined by the equation (Cover 1965):

$$w^T \varphi(x) = 0 \tag{2-12}$$

The inverse image of this hyperplane is the separating surface in the input space.

Given data pairs $(\boldsymbol{x_1}, d_1), (\boldsymbol{x_2}, d_2), \dots (\boldsymbol{x_N}, d_N) \in \mathcal{R}^m \times \mathcal{R}$, the interpolation problem is to find a function $F: \mathcal{R}^m \to \mathcal{R}$ that satisfies the interpolation condition:

$$F(\boldsymbol{x_i}) = d_i, i = 1, 2, \dots, N \tag{2-13}$$

The RBF technique is to choose a function F in the form (Haykin 1999):

$$F(\boldsymbol{x}) = \sum_{i=1}^{N} w_i \varphi(\|\boldsymbol{x} - \boldsymbol{x_i}\|) \tag{2-14}$$

where $w_i \in \mathcal{R}$ are weight factors. $\|\cdot\|$ denotes a norm between $\boldsymbol{x}$ and $\boldsymbol{x_i}$, which is usually Euclidean distance. $\{\varphi(\|\boldsymbol{x} - \boldsymbol{x_i}\|)|i = 1, 2, \dots, N\}$ is a set of radial basis functions, the value of which depends solely on the distance from the data point to the origin. Gaussian function is one of the popular choice and is in the following form:

$$\varphi(r) = e^{-\frac{r^2}{2\sigma^2}} \tag{2-15}$$

where σ defines the width of the bell-shape.

When choosing the center nodes of the RBF network in the hidden layer, aside from using K-means clustering, the centers can also be randomly sampled from the dataset. This step is unsupervised. If the RBF network is used for pattern classification, a hard-limiter or sigmoid function could be placed on the output neurons to generate categorical values.

2.4.5 Evaluation of Machine Learning Algorithms

Having defined the problem and prepared the data, machine learning algorithms will be applied to the data to solve the problem. Multiple tests are needed to run and tune the algorithms in order to discover whether there is a pattern or structure in the problem for the algorithm to learn, and decide which algorithms are effective for the problem.

The step before applying any algorithm is to prepare a training dataset and a test dataset out of the transformed dataset. The two datasets need to be representative of the problem. Generally the intersection of the two sets is empty, meaning that the training dataset and the test dataset are independent of one another. An algorithm will be trained on the training dataset and evaluated against the test dataset.

Other than using separate training and test datasets, another approach is to use the whole transformed dataset to train and test an algorithm, which is called cross validation. The first step of N-fold cross validation method is to separate the dataset into N groups of instances of the equal size M. Each group is called a fold. The model is trained on N-1 folds and then tested on the one fold that was left out. The process is repeated so that each of the N fold is left out and act as a test dataset. In the end, the average of the performance measures of the N folds is used to evaluate the performance of the algorithm on the problem. This method resolves the balance issue between the size and representation of training and test datasets. It is often used when the transformed dataset is not large enough to be split into a training and a test datasets of suitable size.

2.4.6 Performance Measure

The performance measure is the measurement of the performance or quality of solutions to a problem. It is the way to evaluate the success of different machine learning experiments. For numerical prediction, a few measures to interpret the performance of the predictions made by a trained model on the test dataset are listed in Table 2-1. Assume the actual values of the test instances are $a_1, a_2, \ldots, a_n$; the predicted values calculated by the model are $p_1, p_2, \ldots, p_n$; n is the total number of test instances; $\hat{a}$ is the average value from the training dataset; $\overline{a}$

is the mean value over the test dataset; $\overline{p}$ is the mean value of the predictions for the test data.

Table 2-1: Performance measures for numerical prediction (Witten, Frank, and Hall 2011)

Performance Measure	Expression
Correlation coefficient	$\dfrac{\sum_{i=1}^{n}(p_i - \overline{p})(a_i - \overline{a})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2 \times \sum_{i=1}^{n}(a_i - \overline{a})^2}}$
Mean absolute error	$\dfrac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$
Mean squared error	$\dfrac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}$
Root mean squared error	$\sqrt{\dfrac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$
Relative squared error	$\dfrac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \hat{a})^2 + \cdots + (a_n - \hat{a})^2}$
Root relative squared error	$\sqrt{\dfrac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \hat{a})^2 + \cdots + (a_n - \hat{a})^2}}$
Relative absolute error	$\dfrac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \hat{a}| + \cdots + |a_n - \hat{a}|}$

The first measure, correlation coefficient, is scale-independent and measures the statistical correlation between the actual $a_1, a_2, \ldots, a_n$ and the predicted $p_1, p_2, \ldots, p_n$. The larger the value, the better the performance. It ranges from 1 to -1. The value of 0 indicates that there is no correlation at all. Negative values indicate that the results are negatively correlated. For the other error measures, small values are the indications of good performance.

The appropriate choice of performance measures requires considerations of the specific problem and application. For example, the squared error measures and root squared error measures tend to amplify the large discrepancies of prediction errors, whereas the absolute error measures do not have this effect. Fortunately, all the performance measures are easy to calculate. In most situations, the measured results of a numerical prediction method is consistent no matter which mathematical performance measure is used.

2.4.7 Applications in Construction

Artificial intelligence has long been adopted by researchers for modeling and solving problems in the construction industry. Modeling techniques such as artificial neural network (ANN), regression models, and decision trees have been introduced to study the relationships between all kinds of factors in construction processes using historical data.

Song and AbouRizk (2008) used ANN to model the relationship of influencing factors and steel drafting and fabrication productivities. They proposed a systematic approach to make use of historical data, and applied the methodology to measuring and modeling steel drafting and fabrication tasks. Portas (1996) developed a back-propagation, feed-forward neural network system to provide support in the labor productivity estimation for concrete formwork. The inputs to the system are contributing factors to labour productivity, and the output is a set of binary scores representing certainty of occurrence in correspondence with the subset ranges of productivity values that can be used to predict performance of the labour productivity of future projects. ANN has also been used to model the relationship between influencing factors and construction productivity in trades like earthmoving equipment productivity (Karshenas and

Feng 1992), concrete construction productivity (Sonmez and Rowings 1998), and productivity of spool fabrication in the shop and pipe installation in the field (Lu 2001). These researches all proved the effectiveness of ANN in addressing the complexity in construction productivity modeling. At the preparation step, various methods of data collection and productivity measurement were also explored in different trades. Furthermore, instead of utilizing an existing ANN scheme, Lu (2001) developed a new ANN scheme, combining classification and prediction on the basis of Kohonen's LVQ concept and with a probabilistic method integrated, to suit the requirements in the problem domain. It is named the Probability Inference Neural Network (PINN). The new model was applied to predict labour production rates and was proven effective in solving high dimensional mapping of input and output with multiple influential factors.

Hu and Mohamed (2012) explored two different techniques, artificial intelligence planning and dynamic programming, to solve the automation problem in sequencing decision making in construction. More specifically, they applied Planning Domain Description Language (PDDL), which is a domain-independent artificial intelligence planning language.

Fayek and Oduba (2005) used fuzzy logic expert systems to predict productivity of pipe rigging and welding. Contributing factors that affect the productivity of each activity were identified. Fuzzy membership functions and expert rules were developed. Actual data collected from a construction project were used to validate the models, which were proved to have high accuracy of linguistic prediction.

Smith (1999) applied multiple regression-based models to study earthmoving productivity with focus on investigation of the relationships between earthmoving operating conditions and productivity and bunching. The models

developed suggested a strong linear relationship between the operating conditions and productivity. Lee et al. (2013) used regression analysis to develop a quantity prediction model for reinforced concrete and bricks in education facilities that were built as a result of the Build-Transfer-Lease (BTL) projects actively promoted by the Korean government. Linear regression is also used to develop condition prediction models of oil and gas pipelines in order to provide decision support to practitioners in planning for pipeline maintenance (El-Abbasy et al. 2014). Linear regression was explored to suit the numerical output type of the proposed pipeline condition assessment models. The influential factors that have a major impact on pipeline conditions were selected by presenting a questionnaire to experts and reviewing literature. Five condition prediction models were developed and a sensitivity analysis was conducted to learn about the impact degree of each factor on the model output individually.

2.5   WEKA

WEKA is an open source data mining software written in Java developed by the machine learning group at the University of Waikato, New Zealand. It is a modern platform and workbench for applied machine learning. The name WEKA is an acronym which stands for Waikato Environment for Knowledge Analysis. Incorporated into WEKA is a comprehensive collection of machine learning techniques and algorithms that can be applied directly to a dataset. Also included are tools for data pre-processing, classification, regression, clustering, association rules, evaluation methods, and functions that are suited for the development of new machine learning schemes (The University of Waikato 2014). It provides an environment to support and facilitate a range of machine learning activities. Furthermore, with the graphical user interfaces especially the data visualization

feature, a user can easily explore and apply machine learning algorithms and analyze and interpret the results. Figure 2-8 shows its graphical user interface (GUI). The latest stable version is 3.6.11, and that is the version utilised in this research.

Figure 2-8: WEKA GUI chooser

As shown in the figure above, WEKA consists of the following four major applications:

- Explorer: This application is an environment for exploring data with the various transformation schemes, algorithms, etc. Its interface is divided into 5 different tabs, preprocess, classify, cluster, associate, select attributes, and visualize.

- Experimenter: This environment is for designing controlled experiments with algorithm selections and datasets, conducting statistical tests, and analyzing and comparing results between different schemes over multiple runs.

- Knowledge Flow: This interface allows a user to design the iterative machine learning process graphically and run experiments for complex problems. Loading and preprocessing of data, application of algorithms can all be planned via simple drag-and-drop. It provides support of incremental learning.

- Simple CLI: This is a simple command-line interface (CLI) that provides access to all WEKA classes, allowing direct execution of commands for all WEKA features.

Chapter 3. DATA PREPARATION

In order to get solutions to a problem via machine learning, it is critical to feed the algorithms the right data, meaning that significant features need to be included, and that the data are in a useful format and scale (Brownlee 2013b). To prepare data for a machine learning algorithm, they need to be selected, preprocessed, and transformed.

3.1 DATA SOURCE

BIM software has the functionality to create all kinds of reports of the information included in the models. Tekla Structures, used by the collaborating company, creates reports in the format of "*.xsr" files. The reports include lists of drawings, bolts, parts, etc. (Tekla 2014a). Since the reports come directly from the model, the information is always accurate and reliable.

A customized report template (*.rpt) is used in Tekla to create reports containing necessary information from the BIM models. Figure 3-1 and Figure 3-2 demonstrate an example of a material parts report generated from Tekla, the original model of which is shown in Figure 3-3. In the report, essential material attributes, such as part number, description, quantity, length, unit weight, and drawing number, are listed.

Besides the BIM-based models and reports, the collaborating company's information management system (IMS) also includes an internal central database. There are over 400 tables in the SQL Server database maintaining data for shop fabrication, as well as drafting, accounting, quality control, shipping, etc. The

data of detailed materials need to be processed in combination with the production data in the database to be meaningful in achieving the modeling goal.

DRAWING NUMBER	PART NUMBER	REV	DESCRIPTION	QTY	WASTE FACTOR	UNIT WEIGHT {Kg's}	CAT	E-DRAWING NUMBER
2A-1PSA	2A-1PSA	0	W-BEAM-B	6	0.00	113.81	5	E-2A-S30-530
2A-2PSA	2A-2PSA	0	W-BEAM-B	4	0.00	93.02	5	E-2A-S30-531
2A-3PSA	2A-3PSA	0	W-BEAM-B	1	0.00	133.00	5	E-2A-S30-531
2A-3PSA	L56	0	W-BEAM-B	2	0.00	5.36	5	E-2A-S30-531
2A-4PSA	2A-4PSA	0	W-BEAM-B	1	0.00	133.00	5	E-2A-S30-531
2A-4PSA	L56	0	W-BEAM-B	2	0.00	5.36	5	E-2A-S30-531
2A-5PSA	2A-5PSA	0	HS-BRACE-B	4	0.00	70.80	5	E-2A-S30-530
2A-6PSA	2A-6PSA	0	HS-BRACE-B	2	0.00	68.50	5	E-2A-S30-531
2A-7PSA	2A-7PSA	0	HS-BRACE-B	2	0.00	62.24	5	E-2A-S30-531
2A-4000PSA	2A-4000PSA	0	FR-ASS-B	1	0.00	329.95	5	E-2A-S30-530
2A-4000PSA	L1	0	FR-ASS-B	1	0.00	188.62	5	E-2A-S30-530
2A-4000PSA	L15	0	FR-ASS-B	1	0.00	329.95	5	E-2A-S30-530
2A-4000PSA	L22	0	FR-ASS-B	2	0.00	165.82	5	E-2A-S30-530
2A-4000PSA	L31	0	FR-ASS-B	1	0.00	88.40	5	E-2A-S30-530
2A-4000PSA	P2	0	FR-ASS-B	2	0.00	7.55	5	E-2A-S30-530
2A-4001PSA	2A-4001PSA	0	FR-ASS-B	1	0.00	329.95	5	E-2A-S30-530

Figure 3-1: An example of Tekla report (part-1)

```
,                    ,JOB#: 11-770           ,              ,            ,  **v17.0.1**
,                    ,DATE: 20.10.2011  @13:54,              ,            ,
,                    ,BY:                     ,              ,            ,
,                    ,FOR:                    ,              ,            ,
,                    ,mm                      ,              ,            ,
CAT  ,  E-DRAWING NUMBER  , CUSTOMER DRG NUMBER  ,   PROFILE    ,  MATERIAL  , LENGTH
,                    ,  OR WORK TAG NUMBER    ,              ,   GRADE    ,  {M}
,                    ,                        ,              ,            ,
,                    ,                        ,              ,            ,
,                    ,                        ,              ,            ,
,E-2A-S30-530        ,RED-S30-530             ,W200X52       ,350W        ,2.190
,                    ,                        ,              ,            ,
,E-2A-S30-531        ,RED-S30-531             ,W200X52       ,350W        ,1.790
,                    ,                        ,              ,            ,
,E-2A-S30-531        ,RED-S30-531             ,W200X59       ,350W        ,2.250
,E-2A-S30-531        ,RED-S30-531             ,WT100X29.5    ,350W        ,0.182
,                    ,                        ,              ,            ,
,E-2A-S30-531        ,RED-S30-531             ,W200X59       ,350W        ,2.250
,E-2A-S30-531        ,RED-S30-531             ,WT100X29.5    ,350W        ,0.182
,                    ,                        ,              ,            ,
,E-2A-S30-530        ,RED-S30-530             ,HSS127X127X6.4 ,350W Class C ,3.047
,                    ,                        ,              ,            ,
,E-2A-S30-531        ,RED-S30-531             ,HSS127X127X6.4 ,350W Class C ,2.948
,                    ,                        ,              ,            ,
,E-2A-S30-531        ,RED-S30-531             ,HSS127X127X6.4 ,350W Class C ,2.679
,                    ,                        ,              ,            ,
,E-2A-S30-530        ,RED-S30-530             ,W200X59       ,350W        ,5.582
,E-2A-S30-530        ,RED-S30-530             ,W200X59       ,350W        ,3.191
,E-2A-S30-530        ,RED-S30-530             ,W200X59       ,350W        ,5.582
,E-2A-S30-530        ,RED-S30-530             ,W200X52       ,350W        ,3.191
,E-2A-S30-530        ,RED-S30-530             ,HSS127X127X6.4 ,350W Class C ,3.805
,E-2A-S30-530        ,RED-S30-530             ,PL19X225      ,300W        ,0.225
,                    ,                        ,              ,            ,
,E-2A-S30-530        ,RED-S30-530             ,W200X59       ,350W        ,5.582
```

Figure 3-2: An example of Tekla report (part-2)
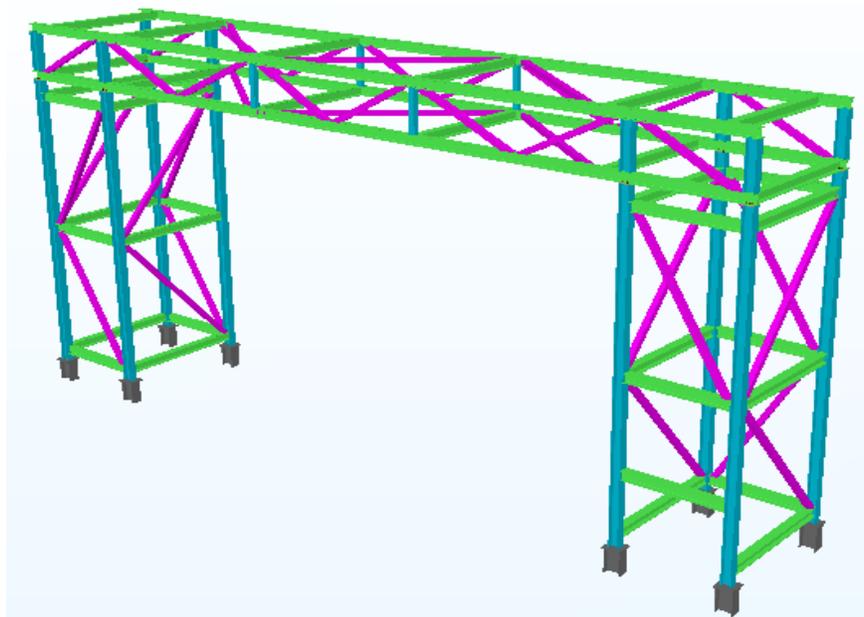
Figure 3-3: An example of Tekla model

3.2 DATA SELECTION

The targeted IMS contains data dating back over ten years. The productivity that can be achieved in the shop and the amount of resource required for fabricating the same amount of structural steel have both changed, compared with those ten years ago, on account of the technological evolution in fabrication methods and equipment, the growth of economy, as well as the development of company strategies. Therefore the input of the learning process must be recent enough to produce meaningful results that could benefit current practices. Furthermore, the dataset should be big enough to be representative of the trade, to contain useful features, and to be able to be split into training and test datasets.

In addition, the method of time-tracking and recording is always improved on the shop floor, but errors still exist in historical data for reasons such as assigning hours to the wrong division number, failure to keep track of time, and failure to digitalize physical timesheets properly. To reduce noise in data as much as possible, division records that have zero tonnage, zero budgeted hours or zero actual hours are excluded. The figure below is the paper timecard currently utilized at the collaborating company.

Figure 3-4: Labour tracking card (courtesy of WSF)

Because of the many services the company provides, jobs in the database are categorized into three different types based on the work scope: supply only, erection only, and supply and erection. Only jobs that include supply work are considered because on-site erection is a process almost completely separate from off-site shop fabrication. Drafting, accounting, quality, shipping, and other data that are irrelevant to the problem being addressed are also excluded.

As a result, the collected dataset accounts for 298 jobs and 1605 divisions in total that were completed by the collaborating company from 2009 to 2013. The data from 2009 to 2012 are used as training dataset, and 2013 data are reserved for testing the model built.

3.3 DATA PREPROCESSING

Having selected the data, preprocessing is necessary to get the data into a form suitable for machine learning. It is an important step that involves a lot of iterations, analysis, and exploration.

The selected data are in a relational database and flat files (*.xsr), and are not ready for application of machine learning algorithms. In the central database, the production-related data are scattered over several tables. A general illustration of the object relations is illustrated in Figure 3-5. The table columns in the figure are only partial. The physical steel materials are not directly associated with each division, but rather as parts of pieces and fabrication drawings. Divisions are assigned to different shops to be processed according to the characteristics of the division and the shops' capacities. Therefore the shop name is included as a nominal input of the model. A detailed description of the relational database structure can be found in Section 3.3.1 - Database Structure.

The database has evolved over the years, leaving misleading parameters and design problems in it. Without any well-written development logs or comments available, it took a lot of time to find the proper database tables and fields to be used for machine learning. A few lessons learnt from working on a production database are listed below.

- Read-only access is not enough. Ask the database administrator for write permission to allow the use of temporary tables, variables, and the viewing of stored procedures.

- Ensure nothing has changed to avoid affecting the functionality of the production database.

- Read all the SQL source code in detail, including data table structures, constraints, stored procedures, functions, etc.

- Comments in the source code may not be reliable. Always test the functions and keep adequate records.

- A same field (or attribute) name used in two tables may mean different things.

- Do not assume the data type of an attribute solely based on its name. For example, "ID" does not have to be *number*; it can also be *string*.

- Starting from a small amount of data makes it easier to verify the query or calculation results.

- Always check any constraints added to a table.

- When a foreign key constraint is included in a query, make sure all fields covered by the foreign key constraint are considered to avoid duplicate query results.

Data were collected via SQL queries and exported to comma-separated values (CSV) files. Figure 3-6 is what the raw query result looks like. Records are at the material level grouped by the division number and different material types.
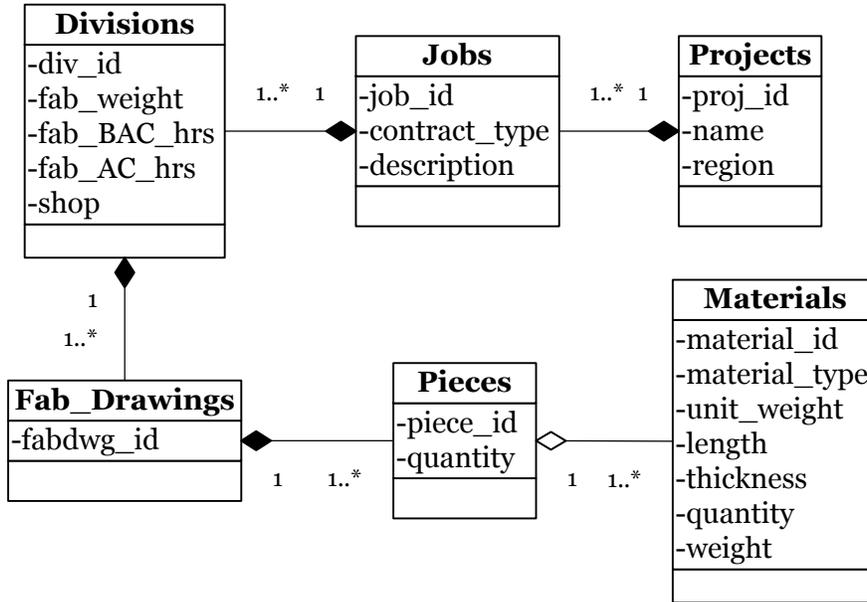
Figure 3-5: Fabrication information structure in database


```
 1   25181,1A,FLAT,162,269,38124.0000,7.921,96.6362,178.500000009194,0,D
 2   25181,1A,PLT,27,765,10962.0000,7.921,96.6362,178.500000009194,0,D
 3   25181,1A,W,54,6887,80074.0000,7.921,96.6362,178.500000009194,0,D
 4   25189,1A,PLT,50,459,24704.0000,3.938,115.7772,82.0000000195257,0,D
 5   25189,1A,W,18,3479,55758.0000,3.938,115.7772,82.0000000195257,0,D
 6   25208,1A,FLAT,208,548,858.5000,3.715,89.16,63.9999999192854,1,C
 7   25208,1A,HTB,336,192,700.0000,3.715,89.16,63.9999999192854,1,C
 8   25208,1A,MISC,112,28,0.0000,3.715,89.16,63.9999999192854,1,C
 9   25208,1A,PLT,28,840,308.0000,3.715,89.16,63.9999999192854,1,C
10   25208,1A,W,50,5822,5042.0000,3.715,89.16,63.9999999192854,1,C
11   25256,1A,C,20,1650,660.0000,1.5495,7.7475,41.5552000040609,1,C
12   25256,1A,FLAT,12,30,106.5000,1.5495,7.7475,41.5552000040609,1,C
13   25256,1A,PLT,8,1419,1198.0000,1.5495,7.7475,41.5552000040609,1,C
14   25402,1A,FLAT,36,60,8472.0000,1.846,46.15,47.8058479532164,0,D
15   25402,1A,PLT,6,170,2436.0000,1.846,46.15,47.8058479532164,0,D
16   25402,1A,W,12,1616,18792.0000,1.846,46.15,47.8058479532164,0,D
17   25403,1A,BV-WASHER,6,1,0.0000,4.662,134.40546,61.8670886102309,0,C
18   25403,1A,C,10,412,11875.0000,4.662,134.40546,61.8670886102309,0,C
19   25403,1A,FLAT,25,60,5270.0000,4.662,134.40546,61.8670886102309,0,C
20   25403,1A,GRAT,14,1293,40058.0000,4.662,134.40546,61.8670886102309,0,C
21   25403,1A,HTB,155,47,8147.0500,4.662,134.40546,61.8670886102309,0,C
22   25403,1A,MISC,120,11,0.0000,4.662,134.40546,61.8670886102309,0,C
23   25403,1A,PLT,3,116,29420.0000,4.662,134.40546,61.8670886102309,0,C
24   25403,1A,W,19,2722,45506.0000,4.662,134.40546,61.8670886102309,0,C
25   25404,2A,BV-WASHER,6,1,0.0000,4.663,134.43429,44.3291139259393,0,C
26   25404,2A,C,10,412,11875.0000,4.663,134.43429,44.3291139259393,0,C
27   25404,2A,FLAT,25,60,5270.0000,4.663,134.43429,44.3291139259393,0,C
```

Figure 3-6: Part of the query result

However, in order to study the productivity and schedule data at the division level, the detailed data of all the materials within the same division need to be collected and then aggregated to the division level. The basic attributes were collected at the level of each material type. Then the total quantity or length, and weight was summarized at the division level. A parser program was written in C# to do the summarization and transposition operations. Material data that belong to a same division were aggregated as one line item.

In addition, some of the data collected were in imperial units (feet, inch, and pound); others were in metric units (meter, centimeter, kilogram, and ton). The data need to have the same scale and unit. Scaling and unit conversion was also automated by the program.

3.3.1 Database Structure

The concept, *project*, is used for jobs that are too large to be managed as one. For example, in oil sands industry there are always multiple construction projects ongoing at the same time at one mining site. The name of the site can be used as the name of a *project*, and all the projects that take place at that site are considered as *jobs* under the same *project*, coded with the same project number. Once a contract is awarded, a job number is allocated following the naming convention "XX-YYY". "XX" is the last two digits of the year when the contract is awarded. "YYY" is an incremental number ranging from 001 to 999, which is tracked by a master logbook. Generally small job numbers are assigned for orders that are less than a certain amount of monetary value or man-hours; large jobs have large job numbers.

As mentioned in Section 1.2, Figure 1-2, jobs are divided into several divisions as the basic management unit. The assignment rules include location,

43

scope, structural type, etc. For example, the ladders and handrails within one job are generally separated from the other steel elements as one division. For a typical stair tower, each floor may be viewed as a division. Division numbers are coded in the format of "[0-9][0-9][A-Z]" customarily.

In the underlying database implementation, divisions are further divided and represented by subdivisions. Generally a division has only one subdivision, which is its main subdivision. Every division has one and only one main subdivision. In situations with handrails and ladders, a separate subdivision is created using the same division number. Subdivision tables contain all the schedule-related data, including all the milestones, planned dates, actual dates, detailer ID, fabrication requirements, etc. The division tables contain the general description, modification timestamps, information about cost categories, man-hour tracking data, etc. The relational diagram of the four tables (projects, jobs, divisions, and subdivisions) is in Figure 3-7.

projects
- proj_id
- name
- cust_id
- active
- color
- quote
- quantity
- budget_cost
- budget_unit_cost
- imperial
- accounting_email
- region_id
- hold_back

jobs
- proj_id
- job_id
- job_no
- description
- porder_no
- manager_id
- imperial
- revenue_type
- active
- quote
- porder_erect
- date_created
- job_types
- old_field_job
- active_timeentry
- rates_confirmed
- region_id

divisions
- proj_id
- job_id
- div_id
- div_no
- stick
- status
- category_set_id
- is_CCN
- last_updated
- last_processed
- draft_weight_for_BAC_hrs
- draft_BAC_hrs
- draft_EV_hrs
- draft_AC_hrs_direct
- draft_AC_hrs_indirect
- draft_weight_for_BAC_n...
- draft_BAC_nonlabor_hrs
- draft_EV_nonlabor_hrs
- fab_weight_for_BAC_hrs
- fab_BAC_hrs
- fab_EV_hrs
- fab_AC_hrs_direct

subdivisions
- proj_id
- job_id
- div_id
- sub_id
- sub_no
- description
- det_mhrs
- fab_mhrs
- weight
- weight_actual
- ifp_date
- ifp_actual
- ifc_date
- ifc_actual
- det_start_date
- det_start_actual
- det_end_date
- det_end_actual

Figure 3-7: Database diagram – jobs and divisions

45

The database records all the changes ever made to the schedule, which is achieved by a Subdivision History table as shown in Figure 3-8.

jobs
- proj_id
- job_id
- job_no
- description
- porder_no
- manager_id
- imperial
- revenue_type
- active
- quote
- porder_erect
- date_created
- job_types
- old_field_job
- active_timeentry
- rates_confirmed
- region_id

divisions
- proj_id
- job_id
- div_id
- div_no
- stick
- status
- category_set_id
- is_CCN
- last_updated
- last_processed
- draft_weight_for_BAC_hrs
- draft_BAC_hrs
- draft_EV_hrs
- draft_AC_hrs_direct
- draft_AC_hrs_indirect
- draft_weight_for_BAC_nonlabor_hrs
- draft_BAC_nonlabor_hrs
- draft_EV_nonlabor_hrs

projects
- proj_id
- name
- cust_id
- active
- color
- quote
- quantity
- budget_cost
- budget_unit_cost
- imperial
- accounting_email
- region_id
- hold_back

subdivisions
- proj_id
- job_id
- div_id
- sub_id
- sub_no
- description
- det_mhrs
- fab_mhrs
- weight
- weight_actual
- ifp_date
- ifp_actual
- ifc_date
- ifc_actual
- det_start_date
- det_start_actual
- det_end_date
- det_end_actual
- det_percent
- approved_date
- approved_actual
- issued_date
- issued_actual
- fab_start_date

subdivision_history
- proj_id
- job_id
- div_id
- sub_id
- history_id
- time_update
- ifp_date
- ifp_actual
- ifc_date
- ifc_actual
- det_start_date
- det_start_actual
- det_end_date
- det_end_actual
- det_percent
- approved_date
- approved_actual
- issued_dc_date
- issued_dc_actual
- issued_date
- issued_actual
- fab_start_date
- fab_start_actual

Figure 3-8: Database diagram – schedule history

Information of working hours are obtained by scanning workers' time cards. Data are associated with each employee. The database diagram related to timesheet data are shown in Figure 3-9 below. Besides division numbers, classification and cost codes are also identified when reporting working time.
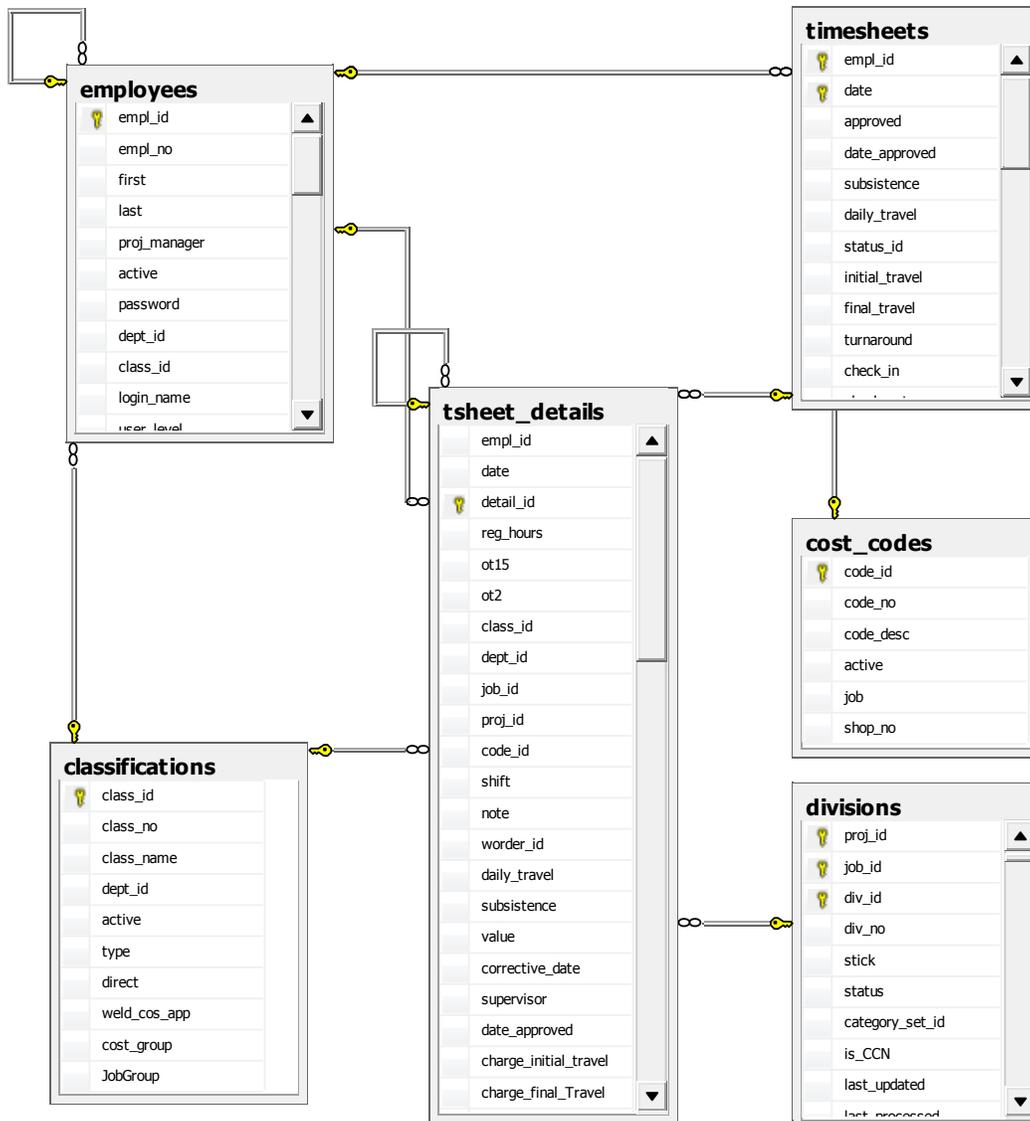
Figure 3-9: Database diagram – employee timesheets

Classification is the type of workforce, such as welders, fitters, labourers, and foremen. Cost coding is another mechanism to track project performance. Generally cost items include detailing, document control, engineering, fabrication labour, freight, galvanizing, grating, material, paint labour, profit, quality control, etc. Not only the fabrication crew in the shop, all the office employees are all recorded in the Employees table. Stored procedures and calculated columns are used to sum up the detailed timesheet information to the division level.

Figure 3-10 shows the connections between divisions, fabrication drawings, and pieces. All the drawings need to be approved before moving to the next phase. When the drafting department completes the initial engineering design, drawings are submitted to the customer for approval. This Issued for Approval (IFA) date is one of the several important milestones in the fabrication production cycle. The customer's reply may be one of the three options: proceed, revise and proceed, or revise and resubmit. The return date is called Return from Approval (RFA), another milestone. After that, drawings are issued to Document Control department for tracking, formatting, and printing to be distributed to shop for fabrication. A division has one or more than one fabrication drawings. A drawing has one or more than one steel pieces.
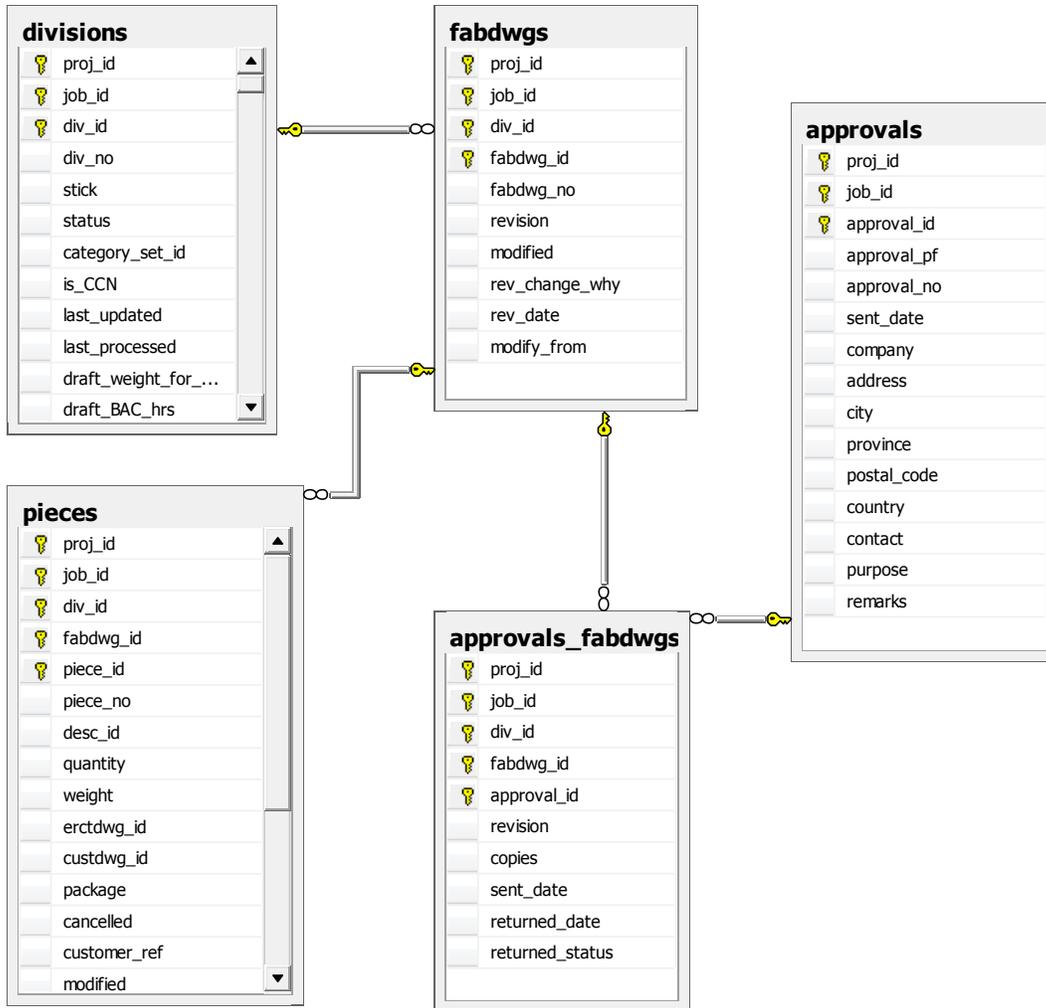
Figure 3-10: Database diagram – fabrication drawings

Steel pieces are composed of plain materials, and the complex relationships are shown in Figure 3-11. Pieces and divisions are linked by fabrication drawings. Pieces such as bolts are stored and tracked separately in the Packages table. The Materials table contains the detailed data about the plain materials that compose pieces, such as length, dimension, quantity, and material types. The material-related data are queried mainly from this Materials table.

49
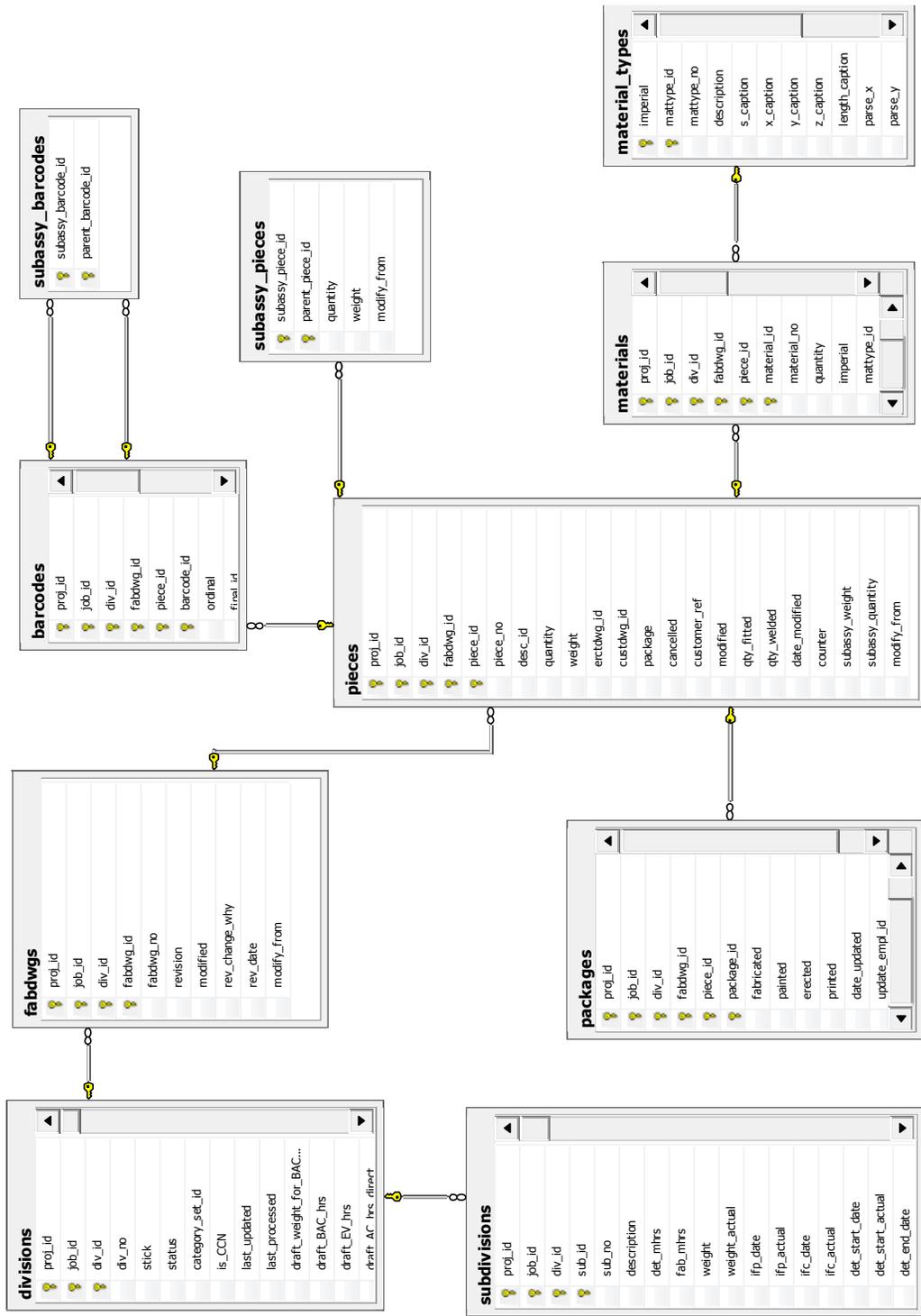
Figure 3-11: Database diagram – pieces and materials

3.4 DATA TRANSFORMATION

Weight and length are two dependent parameters of steel materials, and it would be more meaningful to the problem if they are aggregated into a single feature. The unit weight per unit length, for example "kg/m," and quantity are the two most basic attributes of steel materials. For major materials such as beams, columns, and channels, the fabrication man-hours required are positively correlated with the material length and weight, but for the various kinds of bolts and nuts used in the shop, quantity is a much more meaningful factor to be considered. The length of a bolt plays no role in determining the handling time of the piece it is attached to. Whether the bolt is long or short, it is the quantity that truly matters. The feature aggregation is done during the parsing process.

Table 3-1 lists the 45 material types examined in this study, according to the collaborating company's information library. The letters before the dash sign "-" are the abbreviations and the part after it are the complete description of the material type. For example, "W"-shaped steel actually refers to wide flange beams, and its key attribute is the length instead of quantity. Materials such as miscellaneous assemblies are excluded since their amounts and fabrication requirements are too small to make a difference.

Table 3-1: Material types and key attributes

Material Type	Key Attribute
W – Wide Flange Beams	Length
L – Equal Legs	Length
L – Unequal Legs	Length
C – Channels	Length
HS – Square Hollow Steel Sections	Length

HS – Rectangular Hollow Steel Sections	Length
STD.PIPE – Standard Pipe	Length
M – Miscellaneous Beams	Length
S – I Beams	Length
WT – Structural Tees from W Shapes	Length
MC – Miscellaneous Channels	Length
CFC – Tarpon Cold Formed Channels	Length
RD.HSS – Round Hollow Steel Sections	Length
RAIL – Crane Rails	Length
XS.PIPE – Extra Strong Pipe	Length
XXS.PIPE – Extra Extra Strong Pipe	Length
PLT – Plate	Length
GRAT – Grating	Length
CHECK.PL – Checker Plate	Length
FLAT – Flat Bar	Length
SQ.BAR – Square Bar	Length
RD.BAR – Round Bar	Length
HEX.BAR – Hex Bar	Length
NS.STUD	Length
REBAR	Length
Z – Tarpon Z Sections	Length
WSS – Beam	Length
CFC – Canam Cold Formed Channels	Length
PIPE	Length
CP_WELD – Complete Penetration Weld	Length
PP_WELD – Partial Penetration Weld	Length
BPL – Bent Plate	Length
M-BOLT – M Type Bolts	Quantity
H-NUT – Hex Nuts	Quantity
HTB – Hex Type Bolts	Quantity
FL.WASHER – Flat Washers	Quantity
BV.WASHER – Beveled Washers	Quantity
MBOLT.BTN – Button Head Machine Bolt	Quantity

MBOLT.HEX – Hex Head Machine Bolt	Quantity
HILTI.HAS – Threaded Anchor Rods	Quantity
HILTI-HVU – Adhesive Capsules	Quantity
HILTI-HIT – Adhesive Anchor Cartridges	Quantity
HILTI-KB3 – Expansion Anchor Bolts	Quantity
HILTI.HSL3 – Heavy Duty Expansion Anchor Bolts	Quantity
DTI.WASHER – Compressible Washers with DTI	Quantity

The outcome of the program is a CSV file ready for WEKA with each line representing one division. The format of the output file is demonstrated in Figure 3-12.

```
1   divWt,W-Length,M-Length,S-Length,WT-Length,C-Length,MC-Length,CFC-Length,L-Length,RAIL
    -Length,HS-Length,RD.HSS-Length,STD.PIPE-Length,XS.PIPE-Length,XXS.PIPE-Length,PLT-Len
    gth,GRAT-Length,CHECK.PL-Length,HTB-Quantity,M-BOLT-Quantity,H-NUT-Quantity,FL-WASHER-
    Quantity,BV-WASHER-Quantity,FLAT-Length,SQ.BAR-Length,RD.BAR-Length,HEX.BAR-Length,NS.
    STUD-Length,REBAR-Length,Z-Length,WSS-Length,CANM
    CFC-Length,PIPE-Length,CP_WELD-Length,PP_WELD-Length,MBOLT-BTN-Quantity,MBOLT-HEX-Quan
    tity,HILTI-HAS-Quantity,HILTI-HVU-Quantity,HILTI-HIT-Quantity,HILTI-KB3-Quantity,HILTI
    -HSL3-Quantity,DTI-WASHER-Quantity,BPL-Length,Fab,divAct
2   88841,103.32,0,0,0,0,0,0,0,0,0,0,0,0,0,594.02,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    0,0,0,0,0,0,0,0,A,6320.2
3   100652,125.591,0,0,0,4.984,0,0,201.35,0,0,0,0,0,0,80.534,2.612,0,0,0,0,0,0,7.16,0,0,0,
    0,0,0,137.024,0,0,0,0,0,0,0,0,0,0,0,0,0,A,4229.25170843036
4   57349,17.268,0,0,0,0,0,0,151.618,0,0,0,0,0,0,9.731,0,0,696,0,0,0,0,60.748,0,0,0,0,0,0,
    78.663,0,0,0,0,0,0,0,0,0,0,0,0,0,A,1373.36275626492
5   102885,134.475,0,0,0,4.984,0,0,181.427,0,0,0,0,0.675,0,0,77.525,0,0,0,0,0,0,5.604,0,0,
    0,0,0,0,142.742,0,0,0,0,0,0,0,0,0,0,0,0,0,A,3562.17539863504
6   23709,0,0,0,0,738.864,0,0,48.927,0,0,0,0,0,0,10.352,0,0,1076,0,0,0,0,6.277,0,0,0,0,0,0
    ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,A,407.724943052596
7   16684,285.785,0,0,26.198,0,0,0,0,0,0,0,0,0,0,37.348,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    0,0,0,0,0,0,0,0,0,0,A,728.544974094155
8   26021,353.056,0,0,0,70.368,0,0,381.494,0,0,0,0,0,0,53.806,0,0,672,0,0,0,0,29.436,0,4.7
    2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,A,529.403009650207
9   15690,192.328,0,0,0,46.365,0,0,237.992,0,1.6,0,0,0,0,35.604,0,0,795,0,0,0,0,31.958,0,3
    .776,0,0,0,0,0,0,3.328,0,0,0,0,0,0,0,0,0,0,0,A,483.288141204544
10  22896,26.966,0,0,0,216.939,0,0,33.37,0,0,0,0,0,0,143.99,0,0,865,0,0,0,0,157.718,0,0,0,
    0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,A,844.628115927895
11  23188,26.966,0,0,0,224.1,0,0,45.234,0,0,0,0.55,0,0,145.079,0,0,865,0,0,0,0,158.24,0,0,
    0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,A,924.375693752485
12  39163,53.932,0,0,0,302.718,0,0,96.64,0,0,0,0.275,0,0,272.673,0,0,1034,197,88,88,0,198.
    271,0,6.446,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,A,1708.56020902762
13  40662,53.932,0,0,0,348.229,0,0,236.338,0,0,0,0,0,0,301.041,0,0,896,0,0,0,0,167.021,0,0
    ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,A,1897.69699082588
14  51131,0,0,0,2.64,556.18,0,0,131.887,0,0,0,0,0,0,272.124,173.854,0,1900,0,0,0,0,244.096
    ,0,191.04,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,A,2907.72749038996
```

Figure 3-12: Formatted SQL data

To illustrate the transformed data structure better, Table 3-2 lists the characteristics of one of the divisions to be fed into WEKA. All zero-valued attributes were not included in the table.

Table 3-2: Sample data of a division

Characteristics	Division-*i*	Note
Division ID	18117	Input
Division Weight (kg)	28373	Input
Shop ("A", "B", "C", or "D")	"A"	Input
W Length (m)	13.89	Input
L Length (m)	9.27	Input
Plate Length (m)	78.06	Input
Flat Bar Length (m)	8.07	Input
Hex Type Bolts Quantity	381	Input
Fabrication Actual MHrs	875.57	Targeted Output
Fabrication Budget MHrs	692.30	Expert Judgment

In conclusion, the whole process of preparing the dataset for machine learning is summarized in Figure 3-13. A training dataset and a test dataset, each as a CSV file, are prepared for machine learning modeling, the detailed process of which is illustrated in the next chapter.
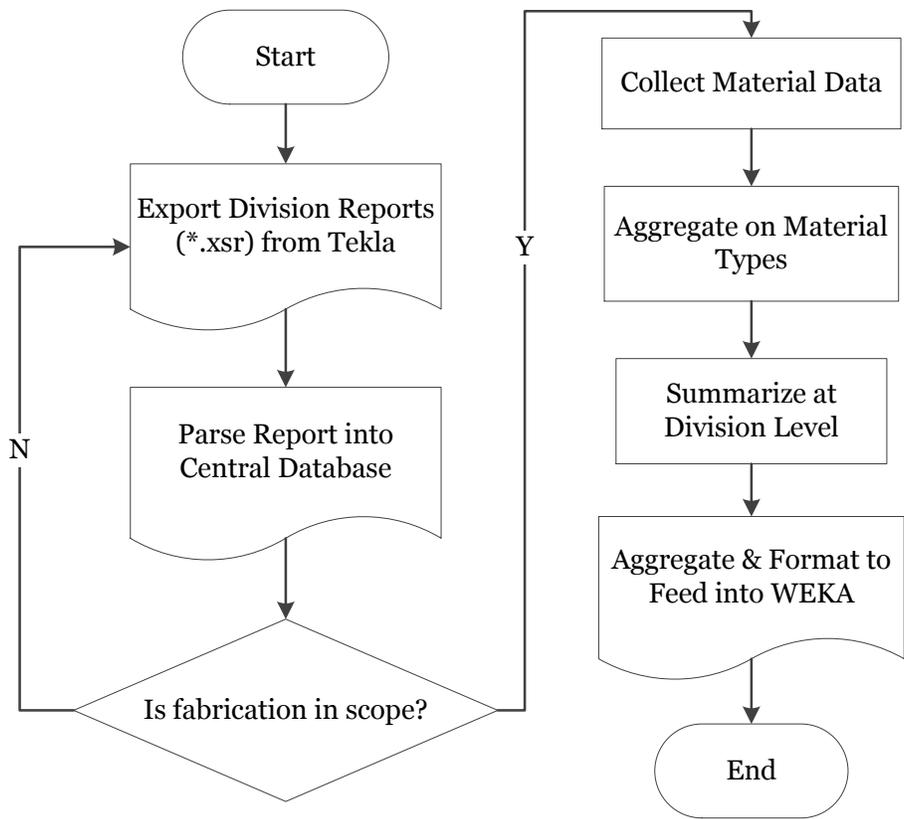
Figure 3-13: Data preparation framework

Chapter 4.   MODEL SELECTION & EVALUATION

4.1   CANDIDATE MODELS

Various types of models were investigated during the stage of model building and training. To obtain statistically meaningful results, 10 runs of 10-fold stratified cross-validation were performed on the training dataset, which is the production data collected from 2009 to 2012, using different schemes. 10 iterations of 10-fold cross-validation means 100 calls of each scheme with the same dataset (Bouckaert et al. 2014). Related knowledge about cross validation can be found in Section 2.4.5 - Evaluation of Machine Learning Algorithms.

It is generally believed that the more materials a job requires, the more man-hours it will cost. Accordingly, linear regression could be a suitable technique to use for quantitative man-hour prediction. The Linear Regression implementation in WEKA uses the Akaike information criterion (AIC) (Burnham and Anderson 2002) for model selection. AIC is a means for measuring the relative quality of a statistical model given a set of data. It rewards goodness of fit of the model, but penalizes increasing the number of parameters so as to lead to over-fitting. A statistical selection procedure is also incorporated in WEKA to determine the best combination of independent input variables. Details about the selected factors and the regression parameters are provided in Equation (1) in Section 4.3 - Evaluation on Test Dataset.

Although ANN models are generally popular in the construction industry, they are more suitable for non-linear problems. One of the most widely used ANN, Radial Basis Function (RBF) neural network, was tested in this study for comparison. The workflow of RBF can be found in Section 2.4.4 - RBF Neural

Network. One of the parameters, the number of clusters, was empirically set as the number of shops, which is 4 for the collaborating company.

Moreover, a Support Vector Machine with Sequential Minimal Optimization (SMO) algorithm (Platt 1998; Shevade et al. 2000) was also investigated. In Support Vector Regression (SVR), an objective function on the training set is defined with a constraint threshold, and the optimization target of regression is to find the best fit objective function while excluding the least outlying training data (Smola and Schölkopf 2004). The basic idea of SVR can be found in Section 2.4.3 - SVM Regression.

Some other models were also tested, such as:

- Least median squared linear regression

- M5P tree

- Multilayer perceptron

The modeling results were not good or typical enough to be elaborated in detail.

4.2   CROSS VALIDATION

The dataset from 2009 to 2012, which accounts for 248 jobs and 1343 divisions out of the total 298 jobs and 1605 divisions, was used to train the model. Data in 2013 were collected as an independent dataset and reserved for further testing. Various schemes were evaluated by performing 10-fold cross-validation on the 2009 – 2012 dataset.

The visualization of classifier errors of the results is shown in Figure 4-3, Figure 4-4, and Figure 4-5. The horizontal axis represents actual fabrication man-hours; the vertical axis represents the man-hour prediction calculated through the trained models. The closer the data points are aligned on the 45 degree line,

57

line $y = x$ to be exact, the closer the forecast is to the actual values. Figure 4-1 shows an example of perfect fitting. The red dashed line in the figure is line $y = x$ for reference. Each cross in the figures represents an instance. The further away a data point is from the reference line, the bigger the difference between the actual value and the predicted value, the bigger the cross sign.
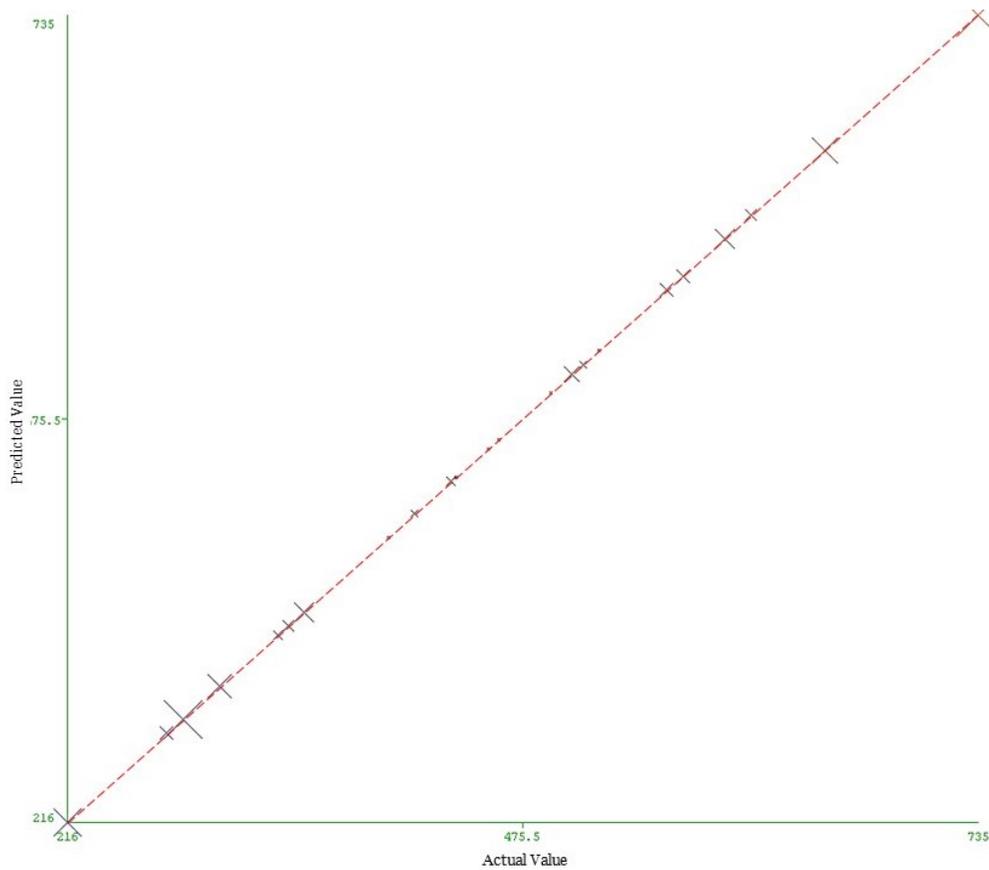
Figure 4-1: A perfect model

Generally speaking, training a model with the full set of instances without held-out evaluation instances and evaluating on the same training dataset bring about misleading results. The trained model would be too perfect to reflect the real performance of machine learning due to overfitting. Figure 4-2 shows the

testing result of RBF neural network using the training dataset itself. Even in this best possible test scenario, the points in Figure 4-2 are far from being a good fit as shown in Figure 4-1. It is clear that the trained model is not suitable for the defined problem.
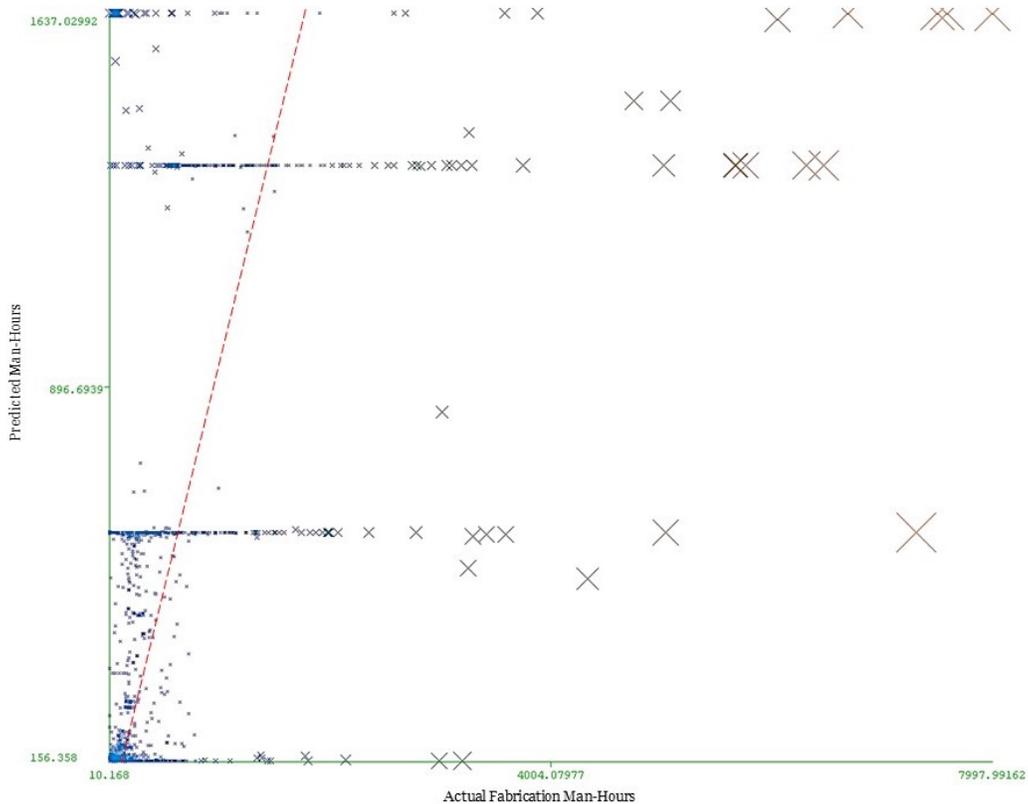
Figure 4-2: Evaluation of RBF network using training set

Figure 4-3 is the result of cross validation using the same dataset and model training. As expected, the results are unsatisfactory based on the current problem definition and dataset available. A summary of the evaluation statistics can be found in Table 4-1.
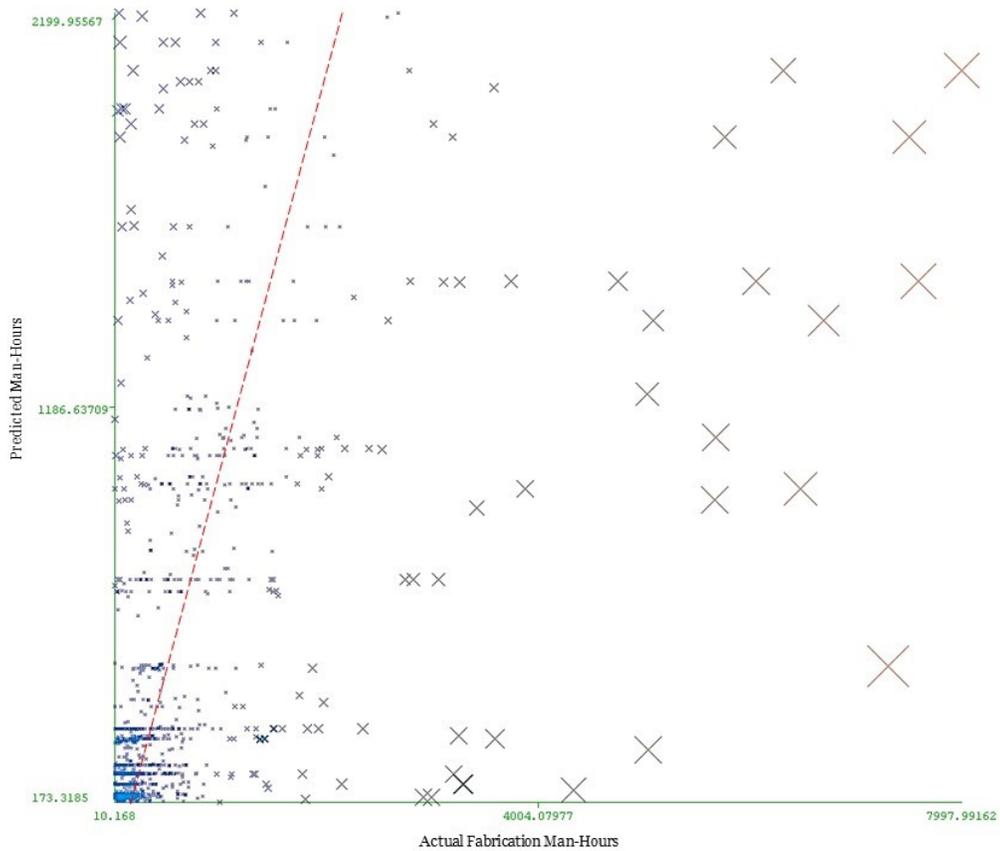
Figure 4-3: Cross validation result of RBF network

Table 4-1: Evaluation comparison of various models

Evaluation Parameter	RBF Network	SMO Regression	Linear Regression
Correlation coefficient	0.49	0.83	0.80
Mean absolute error	344.06	172.74	157.49
Root mean squared error	730.07	467.55	390.90
Relative absolute error	74.25%	37.28%	42.79%

Figure 4-4 and Figure 4-5 are the cross validation results of SMO regression and linear regression. The statistical results are also summarized in Table 4-1. As shown in the figures, both of the fitting results are promising. The overall convergence proves the validity of the trained models.
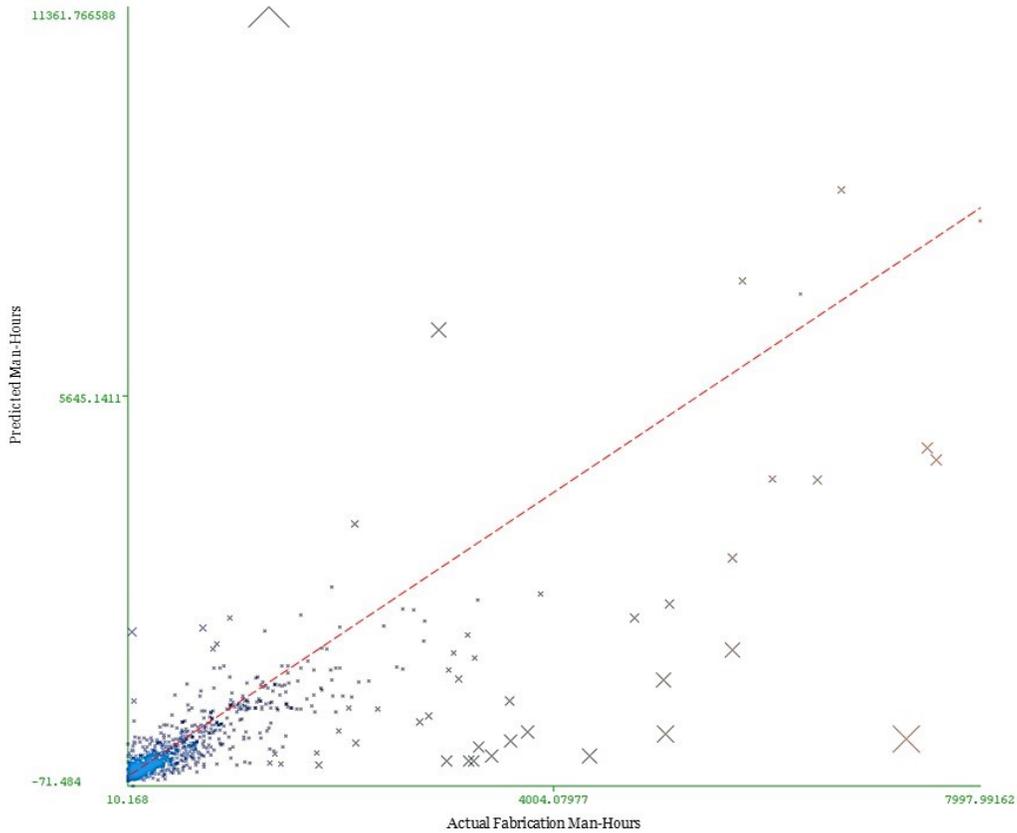
Figure 4-4: Cross validation result of SMO regression
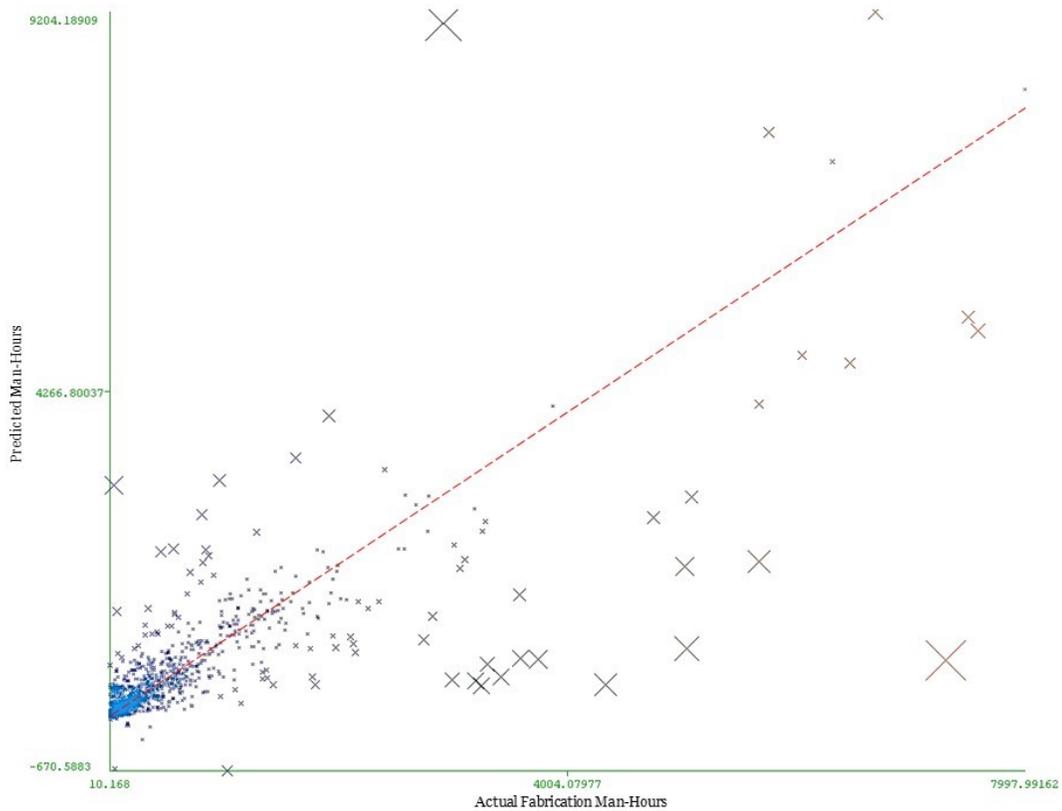
Figure 4-5: Cross validation result of linear regression

SMO regression's attempt to exclude outliers leads to a lower relative absolute error. Despite its complexity, SMO's performance can still be considered statistically as good as linear regression. However, the underlying algorithm of SMO is far more complicated than linear regression, making it function as a black box model, which is generally not desired or acceptable in engineering applications. It took 5.83 seconds and 0.08 second respectively for WEKA to perform the cross validation calculations on 1343 instances using SMO regression and linear regression. The complexity of SMO regression will certainly lead to higher implementation cost and lower user acceptance. Therefore, linear

regression is selected as the solution to the quantity take-off problem in the context of structural steel fabrication.

4.3   EVALUATION ON TEST DATASET

The best-fit model trained by linear regression is shown in Equation (4-1) below. The number of fabrication man-hours a division requires can be calculated based on some key properties that are obtained from BIM models. The complete list of attributes can be found in Table 3-1.

$$
\begin{aligned}
divAct = {}& 0.015 \times divWt + 0.2036 \times W - 0.9271 \times WT \\
& + 0.1708 \times C + 6.7115 \times MC - 0.2687 \times L + 0.7095 \times HS \\
& + 0.8187 \times RD.HSS - 2.4378 \times STD.PIPE + 50.4317 \times XS.PIPE \\
& + 11.5089 \times XXS.PIPE + 0.1645 \times PLT + 0.1334 \times HTB \\
& + 2.1164 \times CHECK.PL - 0.29 \times FL.WASHER - 0.6684 \times BV.WASHER \\
& - 1.7614 \times NS.STUD + 10.776 \times CP.WELD + 1.5694 \times PP.WELD \\
& + 0.4813 \times MBOLT.HEX + 340.4215 \times FabA + 15.054
\end{aligned}
\tag{4-1}
$$

Basic information about the input and output variables are shown in Table 4-2 below. The average values and ranges were based on the whole dataset, i.e. fabrication data from 2009 to 2013.

Table 4-2: Information about variables in Equation (4-1)

Variable	Description	Unit	Average	Range
divAct	Division actual MHrs	Man-Hour	454.16	[10.17, 7997.99]
divWt	Division weight	Kilogram	17877.76	[1001, 238169]
W	Length of wide flange beams	Meter	135.35	[0, 2050.66]

WT	Length of structural tees from W shapes	Meter	8.93	[0, 300.29]
C	Length of channels	Meter	29.27	[0, 3288.60]
MC	Length of miscellaneous channels	Meter	0.27	[0, 131.67]
L	Length of equal or unequal legs	Meter	64.15	[0, 3530.59]
HS	Length of square or rectangular hollow steel sections	Meter	18.93	[0, 1174.29]
RD.HSS	Length of round hollow steel sections	Meter	0.69	[0, 900.95]
STD.PIPE	Length of standard pipes	Meter	1.26	[0, 562.02]
XS.PIPE	Length of extra strong pipes	Meter	0.06	[0, 38.63]
XXS.PIPE	Length of extra extra strong pipes	Meter	0.10	[0, 112.77]
PLT	Length of plates	Meter	102.33	[0, 4202.98]
HTB	Quantity of hex type bolts	n/a	378.82	[0, 9235]
CHECK.PL	Length of checker plates	Meter	2.86	[0, 209.53]
NS.STUD	Length of Nelson S3L shear connectors and H4L headed concrete anchors	Meter	2.80	[0, 286.50]
CP.WELD	Length of complete penetration weld	Meter	3.84	[0, 589.31]
PP.WELD	Length of partial penetration weld	Meter	1.86	[0, 254.12]
FL.WASHER	Quantity of flat washers	n/a	8.92	[0, 1984]
BV.WASHER	Quantity of beveled washers	n/a	3.47	[0, 931]
MBOLT.HEX	Quantity of hex head machine bolts	n/a	4.39	[0, 1176]
FabA	Fabricated in shop "A"	n/a	1: 13%; 0: 87%	0 or 1

According to Equation (4-1), the coefficient of the variable "FabA" is much bigger than the others, meaning that whether a division is fabricated in shop "A" has a great influence on the predicted fabrication MHrs. This traces back to the different functions of the several shops described in Section 1.2 - Problem Statement. Shop "A" is the only one that can handle dimensional structures. Therefore those divisions require more MHrs.

As the next step, the 2013 data were used as an independent test set to evaluate the model performance. The number of fabrication man-hours of each division was calculated using Equation (4-1), and then compared to the actual values recorded in the database. Figure 4-6 shows the correlation between the actual fabrication man-hours and the values predicted by the model.
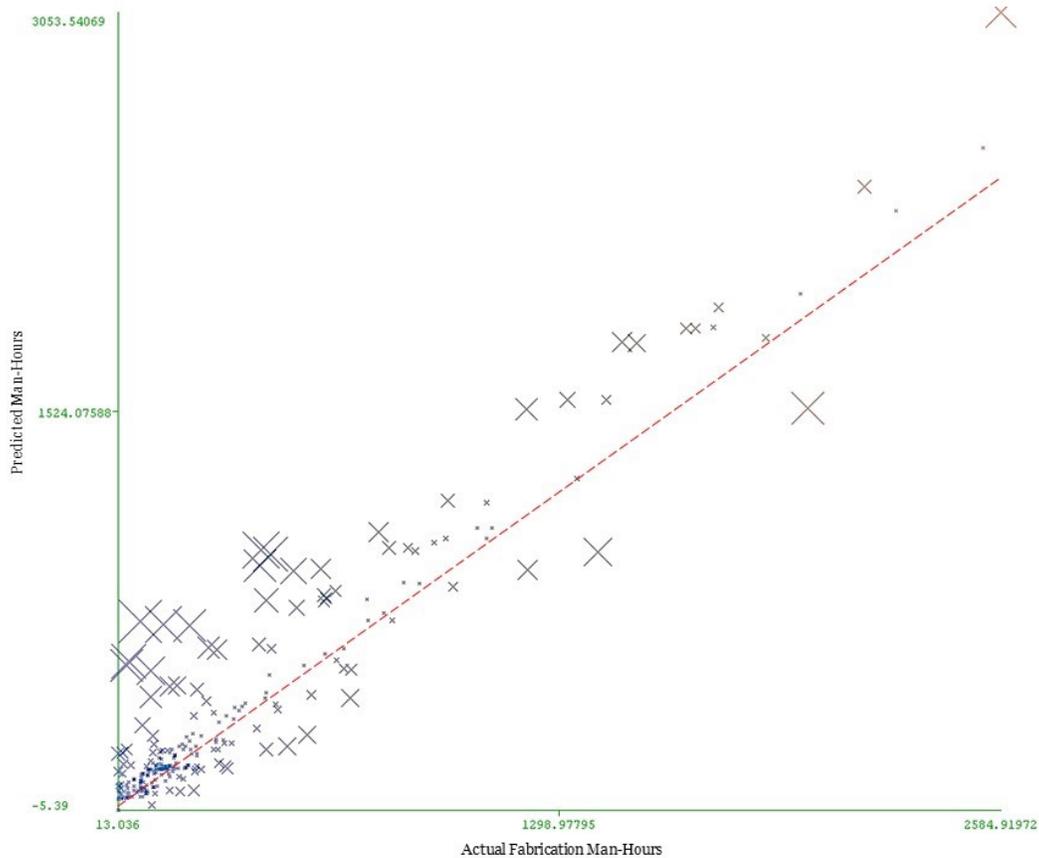

Figure 4-6: Evaluation on independent test set

The method of tracking and recording actual hours on the shop floor has been gradually enhanced. New working procedures of timesheet recording and division planning have been formalized and emphasized. Shop workers are more used to and better at recording their working time accurately over the past few years. The historical data can have errors due to inaccurate records, for instance, working hours assigned to the wrong division number. Compared to the visualization in Figure 4-5, which is the cross validation results on historical data , points of 2013 data in Figure 4-6 suggests better convergence. The reason for this can be partially due to the improved tracking and recording of actual hours.

Limitations in the data are also revealed in the figure. For some divisions, the model tends to predict the work to be more than the actual man-hours recorded. One reason may be that when a worker is working on multiple divisions in a day, it is very likely that he fails to precisely track the number of hours he has spent in each division. Nevertheless, the figure clearly demonstrates that the trained model can be considered satisfactory in terms of accuracy for rendering decision support for estimators to predict the fabrication man-hours.

To further evaluate the performance of the prediction model, the relative difference results were plotted in Figure 4-7.

$$Relative\ Difference = \frac{|p_i - a_i|}{a_i} \qquad (4\text{-}2)$$

where p_i is the value predicted by the model and a_i is the actual value.

For example, as shown in the figure, 68 of the 262 divisions in the test dataset, or 26% of the test dataset, have relative differences less than 10%. 44 divisions, or 17% of the test dataset, have relative differences that are between 10% and 20%. The "small" series in Figure 4-7 represents divisions whose actual MHrs, or workload, are less than 48 MHrs. It is clear that more than half of the large prediction errors (relative difference > 100%) are small divisions.

66
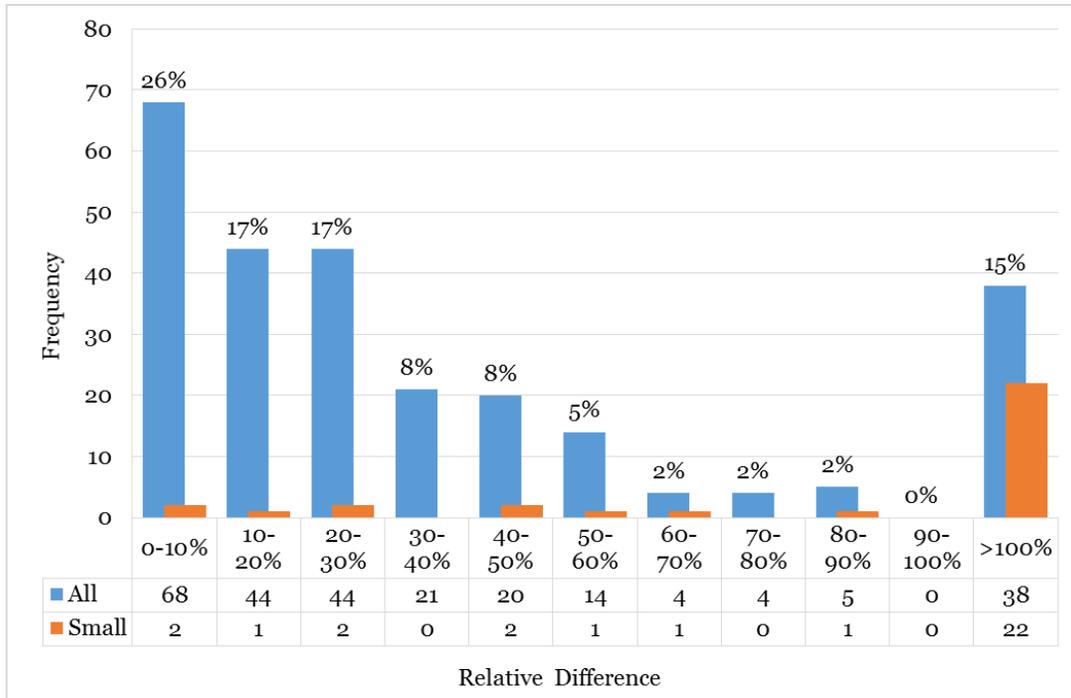
Figure 4-7: Prediction results of test dataset

Sensitivity analysis was conducted in order to test the sensitivity of the model output to changes in the input variables. Each of the input variables were modified while the values of the other variables were kept constant to determine their respective impact on the prediction result. Thus the variations in the prediction results would be under the exclusive influence of the one factor under study (El-Abbasy et al. 2014). The values of the variable under study were varied from $min + p$ to $max = min + 10p$, where p was calculated by Equation (4-3) below.

$$p = \frac{Max - Min}{10}$$ (4-3)

Max and *Min* are the maximum and minimum values in each factor's corresponding data space.

While changing the studied factor, the values of the other factors were set as their average values in their corresponding space. Variable "FabA" was not included in this analysis because there are only two possible values for this factor and its impact on the output is straightforward. The effects of each variable in Equation (4-1) on the predicted MHrs were plotted in Figure 4-8 below. The calculated values corresponding to each variable's changes were tabulated.

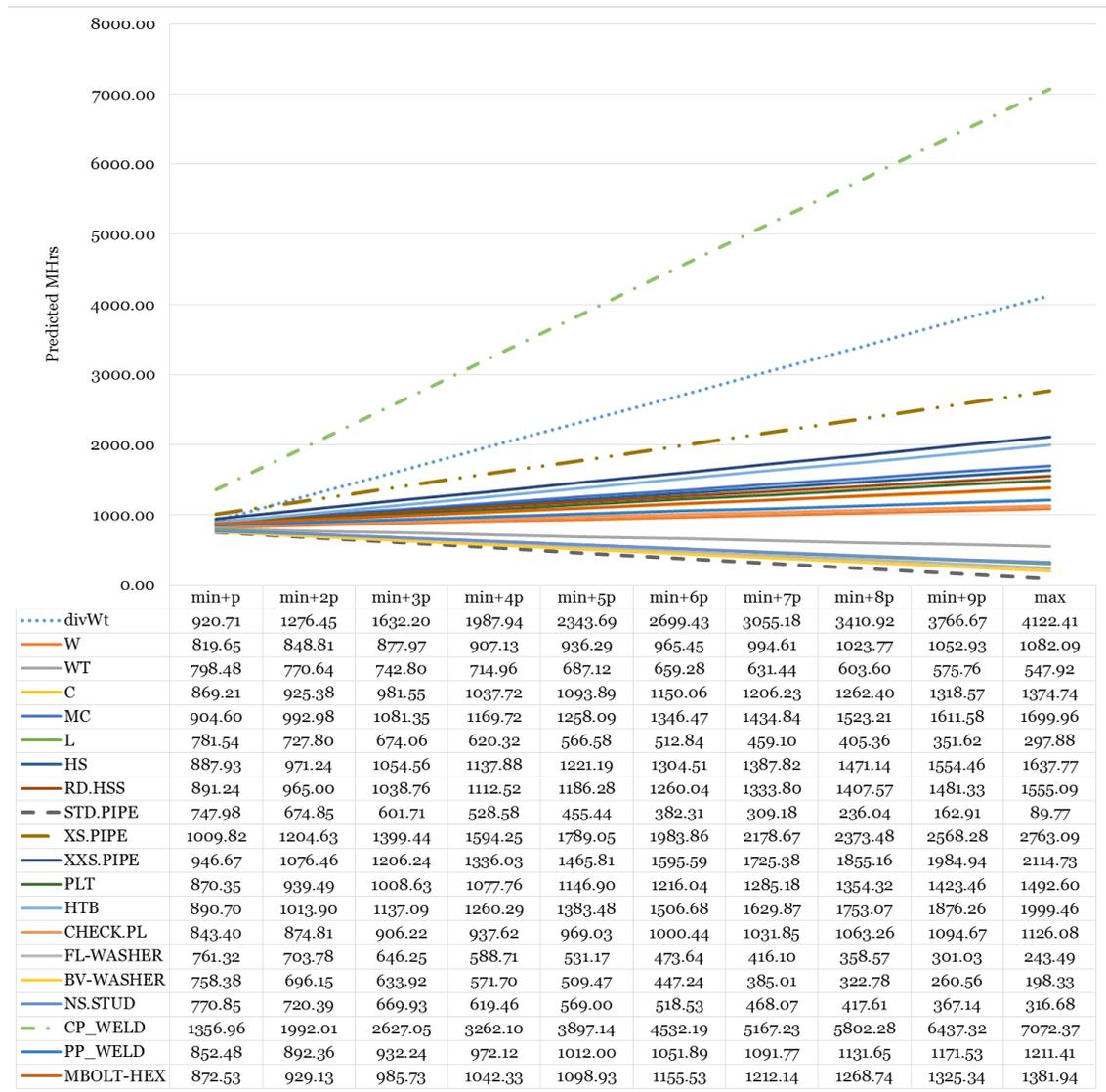|	min+p	min+2p	min+3p	min+4p	min+5p	min+6p	min+7p	min+8p	min+9p	max
divWt	920.71	1276.45	1632.20	1987.94	2343.69	2699.43	3055.18	3410.92	3766.67	4122.41
W	819.65	848.81	877.97	907.13	936.29	965.45	994.61	1023.77	1052.93	1082.09
WT	798.48	770.64	742.80	714.96	687.12	659.28	631.44	603.60	575.76	547.92
C	869.21	925.38	981.55	1037.72	1093.89	1150.06	1206.23	1262.40	1318.57	1374.74
MC	904.60	992.98	1081.35	1169.72	1258.09	1346.47	1434.84	1523.21	1611.58	1699.96
L	781.54	727.80	674.06	620.32	566.58	512.84	459.10	405.36	351.62	297.88
HS	887.93	971.24	1054.56	1137.88	1221.19	1304.51	1387.82	1471.14	1554.46	1637.77
RD.HSS	891.24	965.00	1038.76	1112.52	1186.28	1260.04	1333.80	1407.57	1481.33	1555.09
STD.PIPE	747.98	674.85	601.71	528.58	455.44	382.31	309.18	236.04	162.91	89.77
XS.PIPE	1009.82	1204.63	1399.44	1594.25	1789.05	1983.86	2178.67	2373.48	2568.28	2763.09
XXS.PIPE	946.67	1076.46	1206.24	1336.03	1465.81	1595.59	1725.38	1855.16	1984.94	2114.73
PLT	870.35	939.49	1008.63	1077.76	1146.90	1216.04	1285.18	1354.32	1423.46	1492.60
HTB	890.70	1013.90	1137.09	1260.29	1383.48	1506.68	1629.87	1753.07	1876.26	1999.46
CHECK.PL	843.40	874.81	906.22	937.62	969.03	1000.44	1031.85	1063.26	1094.67	1126.08
FL-WASHER	761.32	703.78	646.25	588.71	531.17	473.64	416.10	358.57	301.03	243.49
BV-WASHER	758.38	696.15	633.92	571.70	509.47	447.24	385.01	322.78	260.56	198.33
NS.STUD	770.85	720.39	669.93	619.46	569.00	518.53	468.07	417.61	367.14	316.68
CP_WELD	1356.96	1992.01	2627.05	3262.10	3897.14	4532.19	5167.23	5802.28	6437.32	7072.37
PP_WELD	852.48	892.36	932.24	972.12	1012.00	1051.89	1091.77	1131.65	1171.53	1211.41
MBOLT-HEX	872.53	929.13	985.73	1042.33	1098.93	1155.53	1212.14	1268.74	1325.34	1381.94

Figure 4-8: Prediction model sensitivity analysis

As shown in the figure, some variables are directly proportional while some are inversely proportional factors. As observed according to the curves in Figure 4-8, the length of CP weld and the division weight have the highest positive effect on the prediction value whereas the length of standard pipe has the highest negative effect.

The shop budgeted man-hours are also compared with the actual fabrication man-hours. Shop budgets are the numbers produced by estimators following the current practice, which relies on the overall steel weight and a man-hour per ton factor from experience. The evaluation results can be found in Table 4-3. As described in Section 2.4.5 - Evaluation of Machine Learning Algorithms, the closer the correlation coefficient is to 1, the better the fitting is. But for the other error evaluations, the smaller the value, the better the result.

Table 4-3: Evaluation summary of linear regression vs. experience

Evaluation Parameter	Cross Validation*	Linear Regression Prediction**	Experience***
Correlation coefficient	0.80	0.95	0.92
Output MHr Range	[10.17, 7997.99]	[13.04, 2584.92]	[13.04, 2584.92]
Mean absolute error	157.49	94.34	134.03
Root mean squared error	390.90	153.90	284.55
Relative absolute error	42.79%	25.01%	41.13%

* Cross validation was performed on the training dataset only.

** The prediction was for the test dataset using the model trained by training dataset.

*** Experience represents the expert estimation on the test dataset.

The forecasted results are closer to the actual values than the judgment made by the professional judgment and experience. The increased accuracy in quantity takeoff will help optimize the company's resource allocation and reduce the risk of cost and schedule overrun. More importantly, the model can be useful as decision support or guidance for someone with little or no experience, especially when no detailed estimating handbook or manual, except a procedure guideline, is available. The estimating process can be accelerated and managers better assured.

Chapter 5.   SHOP LOADING TOOLS

5.1   BACKGROUND

Resources in structural steel fabrication shops include specialized stationary and portable machines, equipment, and workers. Constrained by the available resources, the fabricator's strategies, and/or contractual requirements, the workload in a steel fabrication shop needs to be carefully scheduled. There are always multiple jobs taking place in a fabrication shop at the same time, with different schedules. The ideal situation is to maintain a stable or mildly fluctuating level in the total working man-hours in each particular shop.

In the collaborating company, when the Shop Operations Manager works with a Project Manager to come up with a schedule for a job, they need to enter the fabrication plans of each division into the company's IMS and use the software to spread the workload over time in order to see the overall fluctuation in workload of all the divisions. It takes time to enter the data into the system, make adjustments, and re-run the loading report.

In this research, an Excel Macro-Enabled Workbook file (*.xlsm) was developed to read the estimated man-hours and fabrication schedules of divisions and generate a summary loading chart for the managers to review. Schedule changes can be easily made in the spreadsheet and it is convenient to propagate the report. In addition, a second tool was developed producing a chart comparing the loading changes resulting from scheduling modification. Detailed Macro code can be found in Appendix A and Appendix B.

5.2 SHOP LOADING TOOL

Excel VBA (Visual Basic for Applications) is the name of the programming language of Excel, and the loading functions are implemented in VBA.

Figure 5-1 shows the user interface of the shop loading tool. Functional codes are associated with the two buttons "start" and "clear". To use the loading tool, enter division data right below the column headings, with no blank lines in between. The number of records will be calculated and shown in the corresponding field. User needs to ensure the number is correct. Then by clicking the "start" button, the numbers are read and calculated, populating a loading chart. When finished, the entire workbook can be cleared using the "clear" button.
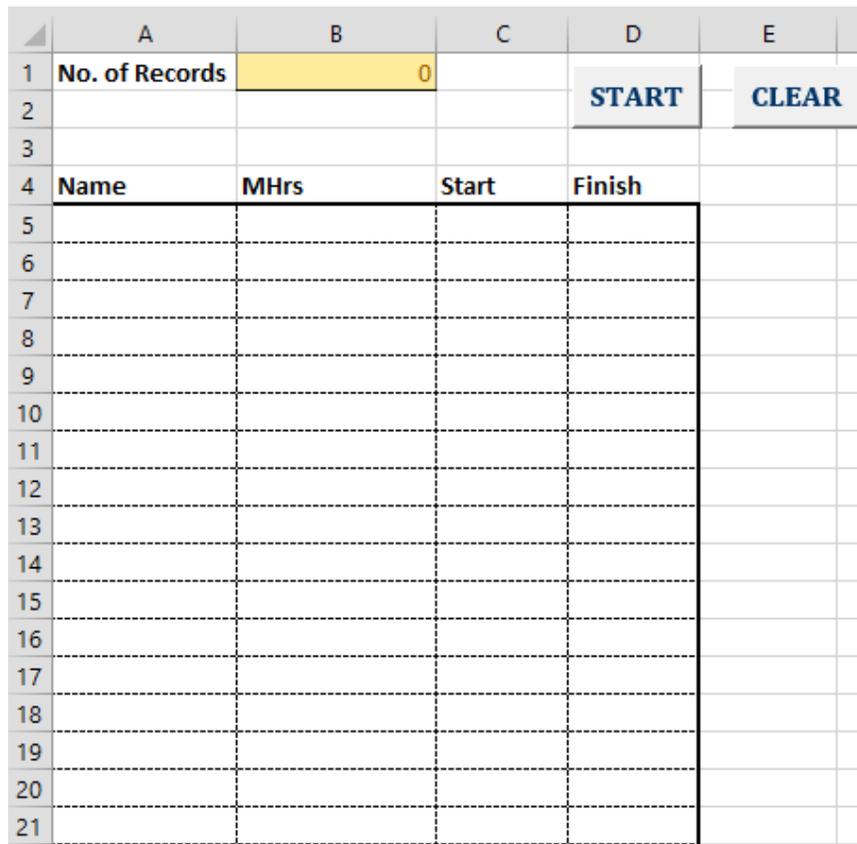
	A	B	C	D	E
1	No. of Records	0			
2				START	CLEAR
3					
4	Name	MHrs	Start	Finish	
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					

Figure 5-1: Shop loading tool user interface

The way this tool works is to evenly spread the workload from the start date to the finish date without considering any holiday breaks. Weekends or daily working hours do not affect the loading results either, due to the even-distribution rule. Having spread the workload of all divisions respectively, man-hours are summed up on a weekly basis and a column chart is plotted on the summary data. The workbook is suitable for planners to create preliminary shop loading chart based on division man-hours and fabrication dates. The detailed loading process is explained in Figure 5-2.
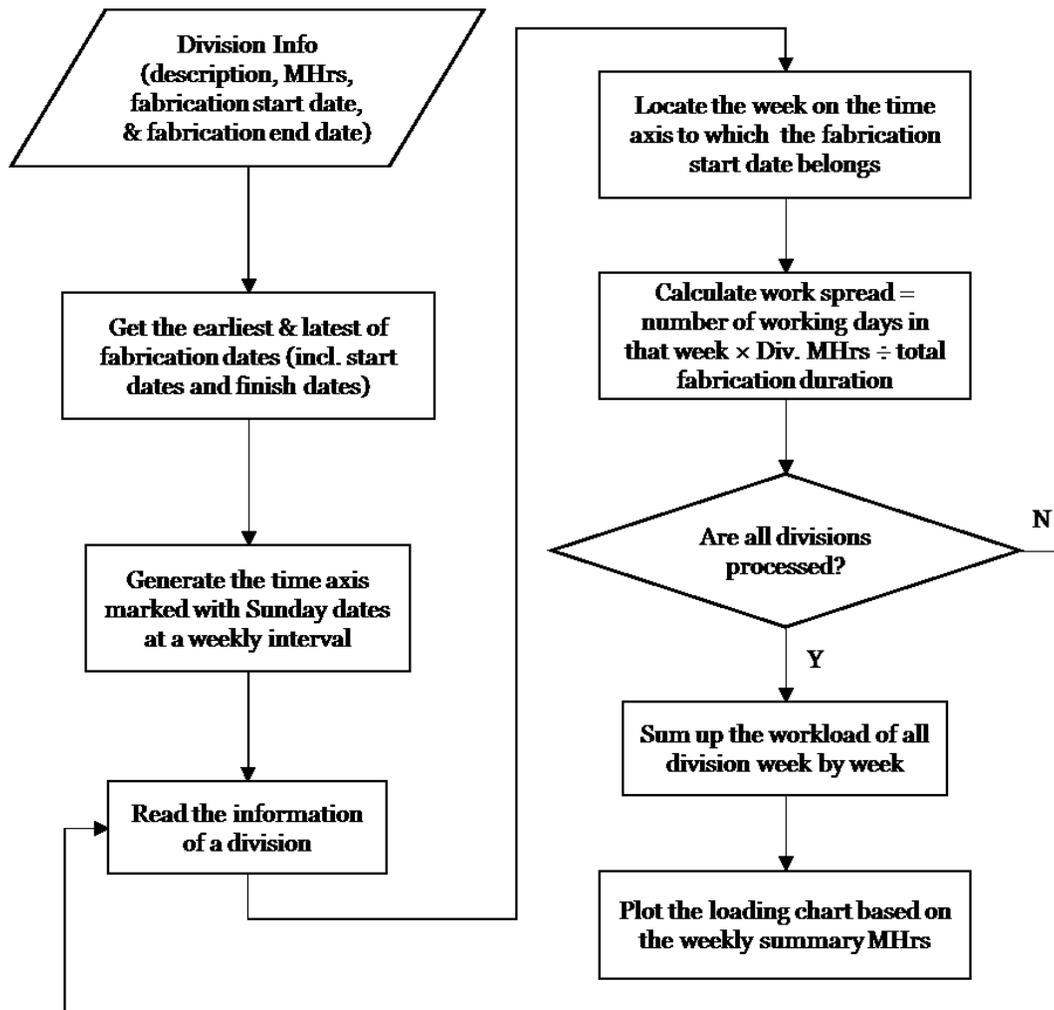

Figure 5-2: Workflow of generating shop loading chart

5.2.1 Demonstration

A job of 22 divisions was used as an example. The related information was filled in the workbook as shown in Figure 5-3. After clicking the button "start", a horizontal axis of time was generated and the detailed workload was spread on the time axis. A new row of data was added summarizing the weekly shop workload. Figure 5-4 is the loading chart plotted in the same worksheet. The trending of the shop workload is clearly revealed in the loading chart. The job planner can adjust the fabrication schedules accordingly.

	A	B	C	D	E	F	G	H	I	J	K	L
1	No. of Records	22										
2				START	CLEAR							
3												
4	Division	MHrs	Start	Finish	15-Sep-13	22-Sep-13	29-Sep-13	06-Oct-13	13-Oct-13	20-Oct-13	27-Oct-13	03-Nov-13
5	1A	1523.2	19-Sep-13	08-Jan-14	41.167568	96.057658	96.057658	96.057658	96.057658	96.057658	96.05766	96.057658
6	2A	1510.4	19-Sep-13	08-Jan-14	40.821622	95.25045	95.25045	95.25045	95.25045	95.25045	95.25045	95.25045
7	3A	2672	27-Sep-13	16-Jan-14		48.144144	168.5045	168.5045	168.5045	168.5045	168.5045	168.5045
8	4A	3014.4	27-Sep-13	16-Jan-14		54.313514	190.0973	190.0973	190.0973	190.0973	190.0973	190.0973
9	5A	1792	03-Oct-13	22-Jan-14			48.432432	113.00901	113.00901	113.00901	113.009	113.00901
10	6A	2220.8	28-Oct-13	14-Feb-14							122.2459	142.62018
11	7A	2108.8	17-Oct-13	05-Feb-14					56.994595	132.98739	132.9874	132.98739
12	8A	1568	03-Oct-13	22-Jan-14			42.378378	98.882883	98.882883	98.882883	98.88288	98.882883
13	9A	1148.8	17-Oct-13	05-Feb-14					31.048649	72.446847	72.44685	72.446847
14	10A	1814.4	07-Nov-13	26-Feb-14								49.037838
15	11A	1382.4	24-Oct-13	12-Feb-14						37.362162	87.17838	87.178378
16	12A	1116.8	01-Nov-13	20-Feb-14							20.12252	70.428829
17	13A	1132.8	01-Nov-13	20-Feb-14							20.41081	71.437838
18	14S	4601.6	14-Nov-13	23-May-14								
19	20A	3136	24-Jan-14	15-May-14								
20	21A	1760	24-Jan-14	15-May-14								
21	22A	2304	24-Jan-14	15-May-14								
22	23A	1664	24-Jan-14	15-May-14								
23	24A	1184	24-Jan-14	15-May-14								
24	25A	1312	24-Jan-14	15-May-14								
25	30A	1670.4	05-Jan-14	25-Apr-14								
26	31A	1027.2	05-Jan-14	24-Apr-14								
27				Sum	81.989189	293.76577	640.72072	761.8018	849.84505	1004.5982	1217.194	1387.9391

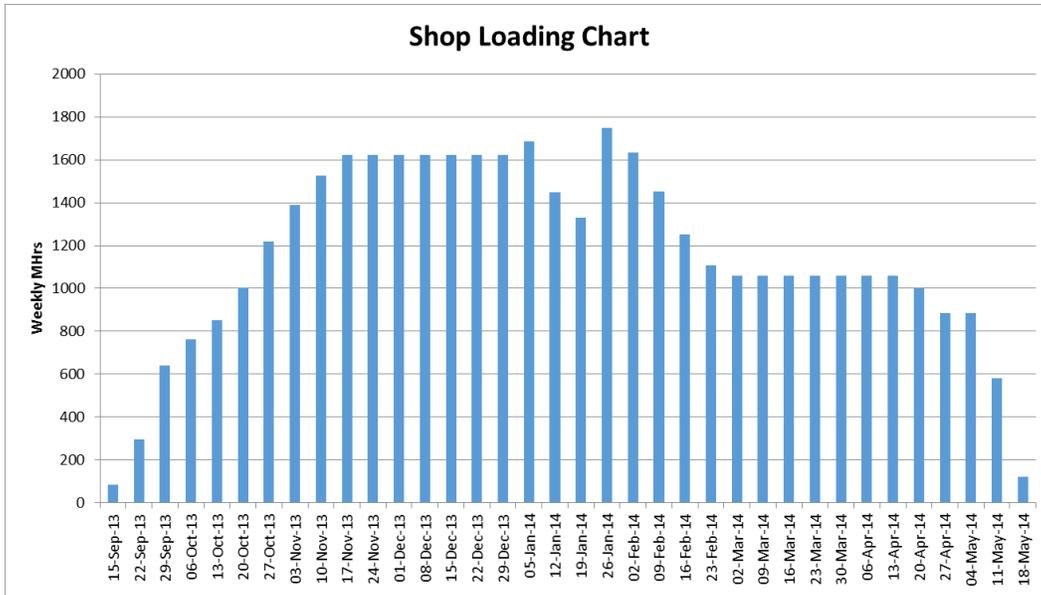**Figure 5-3: Shop loading tool example**

Figure 5-4: Shop loading chart example

5.3   LOADING COMPARISON TOOL

This tool was developed on the basis of the shop loading builder discussed in Section 5.2. The underlying calculation adopts the same even-distribution rule, meaning the workload is evenly spread from the start date to the finish date without considering any holiday breaks, weekends, or the number of daily working hours. Its main purpose is to show the difference in shop loading with different schedules. Figure 5-5 is the user interface of the loading comparison tool. The same simple design of two buttons ensures that it can be readily applied by users. Button "start" takes care of the calculation and chart plotting. Button "clear" resets the spreadsheet. Not only the fabrication dates but also the man-hours of each division can be changed, which is usually the case when more information becomes available to the planner.

Figure 5-5: Loading comparison tool user interface

5.3.1 Demonstration

A job of 41 divisions was used as an example. Two versions of the schedules, Plan A and Plan B, were filled in the workbook as shown in Figure 5-6. In the example, division man-hours are the same but the schedules differ. The original schedule (Plan A) was changed due to late issue of IFC (issued for construction) drawings.

	A	B	C	D	E	F	G	H	I	
1	No. of Records	41			No. of Records	41				START
2										
3	Plan Name	Plan A			Plan Name	Plan B				CLEAR
4	Name	MHrs	Start	Finish	Name	MHrs	Start	Finish		
5	1A	1523.2	19-Sep-13	08-Jan-14	1A	1523.2	19-Sep-13	19-Dec-13	25-Aug-13	01-S
6	2A	1510.4	19-Sep-13	08-Jan-14	2A	1510.4	19-Sep-13	07-Jan-14	18.104505	31.6
7	3A	2672	27-Sep-13	16-Jan-14	3A	2672	01-Oct-13	13-Jan-14	11.102762	19.4
8	4A	3014.4	27-Sep-13	16-Jan-14	4A	3014.4	01-Oct-13	31-Jan-14		
9	5A	1792	03-Oct-13	22-Jan-14	5A	1792	17-Oct-13	03-Feb-14		
10	6A	2220.8	28-Oct-13	14-Feb-14	6A	2220.8	18-Dec-13	31-May-14		
11	7A	2108.8	17-Oct-13	05-Feb-14	7A	2108.8	28-Oct-13	19-Feb-14		
12	8A	1568	03-Oct-13	22-Jan-14	8A	1568	10-Oct-13	11-Feb-14		
13	9A	1148.8	17-Oct-13	05-Feb-14	9A	1148.8	11-Oct-13	20-Feb-14		
14	10A	1814.4	07-Nov-13	26-Feb-14	10A	1814.4	10-Dec-13	04-Mar-14		
15	11A	1382.4	24-Oct-13	12-Feb-14	11A	1382.4	14-Nov-13	25-Feb-14		
16	12A	1116.8	01-Nov-13	20-Feb-14	12A	1116.8	23-Oct-13	12-Feb-14		
17	13A	1132.8	01-Nov-13	20-Feb-14	13A	1132.8	08-Nov-13	11-Feb-14		
18	14S	4601.6	14-Nov-13	23-May-14	14S	4601.6	16-Apr-14	27-Jun-14		
19	20A	3136	24-Jan-14	15-May-14	20A	3136	10-Apr-14	16-Jun-14		
20	21A	1760	24-Jan-14	15-May-14	21A	1760	10-Apr-14	23-Jun-14		
21	22A	2304	24-Jan-14	15-May-14	22A	2304	10-Apr-14	30-Jun-14		
22	23A	1664	24-Jan-14	15-May-14	23A	1664	10-Apr-14	07-Jul-14		
23	24A	1184	24-Jan-14	15-May-14	24A	1184	10-Apr-14	07-Jul-14		
24	25A	1312	24-Jan-14	15-May-14	25A	1312	10-Apr-14	17-Jul-14		
25	30A	1670.4	05-Jan-14	25-Apr-14	30A	1670.4	28-Feb-14	23-May-14		
26	31A	1027.2	05-Jan-14	24-Apr-14	31A	1027.2	28-Feb-14	25-May-14		
27	32A	1436.8	05-Jan-14	24-Apr-14	32A	1436.8	05-Feb-14	17-May-14		
28	40A	896	24-Oct-13	12-Feb-14	40A	896	21-Jan-14	14-Apr-14		
29	41A	128	24-Oct-13	12-Feb-14	41A	128	28-Nov-13	05-Mar-14		
30	42A	368	24-Oct-13	12-Feb-14	42A	368	03-Feb-14	21-Apr-14		
31	50A	904.8	30-Oct-13	18-Feb-14	50A	904.8	28-Nov-13	25-Mar-14		
32	51A	904.8	30-Oct-13	18-Feb-14	51A	904.8	28-Nov-13	26-Mar-14		
33	52A	904.8	30-Oct-13	18-Feb-14	52A	904.8	09-Dec-13	25-Mar-14		
34	53A	904.8	30-Oct-13	18-Feb-14	53A	904.8	09-Dec-13	25-Mar-14		
35	60A	544	30-Oct-13	18-Feb-14	60A	544	05-Nov-13	11-Feb-14		
36	61A	211.2	30-Oct-13	18-Feb-14	61A	211.2	05-Nov-13	18-Feb-14		
37	70A	502.4	28-Aug-13	17-Dec-13	70A	502.4	28-Aug-13	25-Feb-14		
38	71A	192	30-Sep-13	21-Feb-14	71A	192	11-Oct-13	11-Feb-14		

Figure 5-6: Loading comparison tool example

The loading comparison of these two plans is shown in Figure 5-7. The blue series represents the original plan, and the orange series shows the adjusted schedule. The chart exhibits a shift of the fabrication workload.
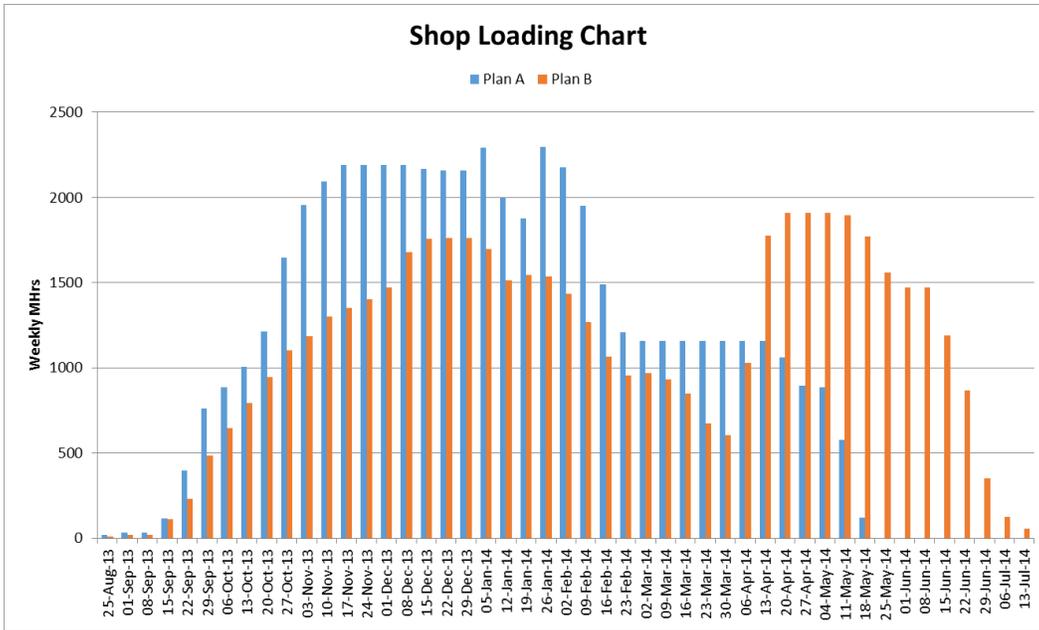
Figure 5-7: Loading comparison chart example

5.4   APPLICATION OF TOOLS

As proposed in the previous sections, division workload can be estimated using the BIM-based regression model, which is one of the input factors for the shop loading tools discussed in this section. These tools have been widely adopted and appraised by the project managers in the collaborating company. Combined with the quantity takeoff scheme, the shop loading builder and the loading comparison tool can provide much needed support in project managers' decision-making processes. A proposed process of future planning in structural steel fabrication industry is illustrated in Figure 5-8. With the automated facilitating tools integrated, this whole framework can help save time and improve accuracy in planning steel fabrication jobs. The process of scheduling is marked with dashed lines because it is handled manually at the current phase and is not the

focus of this research. Ideas about scheduling automation can be found in Section
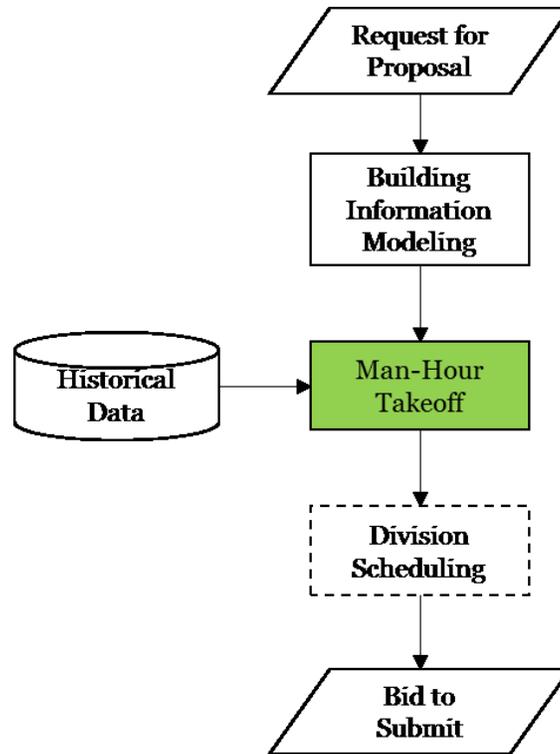
6.4 - Proposal for Future Research.

Figure 5-8: A proposed procedure for bid preparation

Chapter 6.   CONCLUSIONS

6.1   SUMMARY

Structural steel fabrication is an industry with characteristics that makes it different from traditional construction and manufacturing. The use of BIM is on the rise not only in general construction but also in structural steel fabrication. However, the functions and advantages of BIM-based models are limited to design and drafting in most cases. This research aims to develop an approach to extend BIM's usage further into estimating and planning phases. The performance information recorded in historical BIM data is important and can be predictive for the company's future projects. Expanding the scope of application of BIM is essential to improving the efficiency of steel fabrication work in terms of reducing implementation costs and satisfying project objectives such as schedule milestones.

This study develops a linear regression model to provide man-hour requirement estimate for steel fabrication projects in the planning phase. The proposed methodology is implemented and validated, proving the models to be both feasible and recommended to support project estimating and planning. The models were developed using the production data from the collaborating company, so that they were customized to the company's information management system (IMS). Another steel fabrication company may have different ways of tracking data and implementing IMS, but the methodology and framework of the study can still be used for the development of quantity take-off prediction models.

Furthermore, two Excel macro-enabled workbooks were developed for the purpose of facilitating shop operation scheduling. Used in combination with the BIM-based quantity takeoff solution, the shop loading tools provide decision-making support for project managers in the structural steel fabrication industry.

6.2 LIMITATIONS

Fitting and welding are generally the most complex and time-consuming steps in steel fabrication. Different designs of the connections make a big difference to the fitting and welding workload, affecting the overall fabrication duration. For example, a column with end plates may be designed to be connected using bolts only, the fabrication of which requires cutting and detailing. This issue is another contributing factor to the regression errors observed.

This research does not take into consideration the complexity of connections, such as the specific number of stiffeners to weld, the type of welding, and the number of plates to bolt. The reason for this is that most of the time, information on connections is not available during bidding. It is typically the contractor's duty to complement the customer's preliminary design with detailed connection designs, upgrading customer drawings to fabrication drawings. In general, the time before bid submission is not enough to complete connection design. Thus the information is normally not ready until the award of contract.

Currently structural steel fabrication shops are also sometimes responsible for assembling the steel elements to a certain extent before shipping to the erection site. Offsite assembly results in cost and time savings at the erection site, and improvement of construction quality. Those steel elements are called super-assembly or sub-assembly (SA) structures and are different from traditional beam line type fabrication. SA demands not only additional labour but

also takes more time in the shop. An SA structure is generally larger than a single piece so that extra labour and equipment is required to handle it. It also takes time to wait for the composition pieces to be fabricated before assembly can commence or continue. This research does not distinguish specifically the assembly of SA structures from traditional streamlined fabrication for that the problem definition and data collection are both on the division level. It is appropriate to let the modeling training process in WEKA resolve the issue by clustering structure types and factoring in extra labour hours by adjusting the coefficients for SA structures.

6.3   RECOMMENDATIONS TO WSF

To fully utilize the results of this research, a few recommendations are listed for WSF.

- It is recommended to keep adding new data to the training dataset and re-train the prediction model on a 3-month basis. Without proper update, the model trained a few years ago is not likely to work well for jobs now.
- The absolute error of the modeling results can be related to a margin estimate of the job in bidding. Questionnaires need to be designed to obtain expert opinions on this.
- To increase the prediction accuracy, jobs can be classified based on their dimensional characteristics, before the modeling process.
- The time tracking methods on the shop floor can be upgraded to improve the quality of the man-hour data collected.

- More data are to be collected to finalize the boundaries of the input parameters, and set the low and high limit for the output parameter, making the MHr prediction more reasonable.

6.4 PROPOSAL FOR FUTURE RESEARCH

The results of this study show much promise for advanced BIM in steel fabrication planning and control. The combination of BIM with the current scheduling and fabrication process can be investigated in future studies.

Starting from the current development, automatic scheduling on the shop floor can be investigated. A number of issues must be addressed, including how to determine the priority of various steel elements in scheduling, the complexity of the specific fabrication, and optimal allocation of the various resources on the shop floor. With the scheduling process handled by artificial intelligence, the whole process shown in Figure 5-8 would be complete and become an integrated system of estimating and scheduling, pushing the application of BIM further to the planning stage in steel fabrication shops.

REFERENCES

Aranda-Mena, Guillermo, John Crawford, Agustin Chevez, and Thomas Froese. 2009. "Building Information Modelling Demystified: Does It Make Business Sense to Adopt BIM?" *International Journal of Managing Projects in Business* 2 (3): 419–34. doi:10.1108/17538370910971063.

Autodesk. 2003. *Building Information Modeling*. San Rafael, CA: Autodesk, Inc. http://www.laiserin.com/features/bim/autodesk_bim.pdf.

Azimi, Reza, SangHyun Lee, Simaan M. AbouRizk, and Amin Alvanchi. 2011. "A Framework for an Automated and Integrated Project Monitoring and Control System for Steel Fabrication Projects." *Automation in Construction*, Global convergence in construction, 20 (1): 88–97. doi:10.1016/j.autcon.2010.07.001.

Berman, Gary. 2014. "STRUCTURAL STEEL DESIGN AND CONSTRUCTION - Greyhawk." Accessed June 18. https://www.yumpu.com/en/document/view/4478730/structural-steel-design-and-construction-greyhawk.

Bouckaert, Remco, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, and David Scuse. 2014. "WEKA Manual for Version 3-6-11." The University of Waikato.

Broomhead, D. S., and David Lowe. 1988. "Multivariable Functional Interpolation and Adaptive Networks." *Complex Systems* 2 (3): 321–55.

Brownlee, Jason. 2013a. "A Tour of Machine Learning Algorithms | Machine Learning Mastery." *Machine Learning Mastery*. November 25. http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms.

———. 2013b. "How to Prepare Data For Machine Learning | Machine Learning Mastery." *Machine Learning Mastery*. December 25. http://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/.

buildingSMART. 2014. "Industry Foundation Classes (IFC) Data Model — buildingSMART." Accessed June 24. http://www.buildingsmart.org/standards/ifc.

Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer.

Chiu, Chui-Te, Tseng-Hsing Hsu, Ming-Teh Wang, and Hsien-Yen Chiu. 2011. "Simulation for Steel Bridge Erection by Using BIM Tools." *Proceedings of the 28th ISARC, Seoul, Korea*, 560–63.

CII. 1995. *Pre-Project Planning Handbook.* Special Publication 39-2. University of Texas at Austin. https://www.construction-institute.org/scriptcontent/more/sp39_2_more.cfm.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97. doi:10.1007/BF00994018.

Cover, T.M. 1965. "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition." *IEEE Transactions on Electronic Computers* EC-14 (3): 326–34. doi:10.1109/PGEC.1965.264137.

Cross Validated. 2011. "Understanding SVM Regression: Objective Function and 'Flatness.'" Q&A site. *Cross Validated.* January. http://stats.stackexchange.com/questions/5945/understanding-svm-regression-objective-function-and-flatness.

———. 2013. "What Is the Difference between Data Mining, Statistics, Machine Learning and AI?" Q&A site. *Cross Validated.* http://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai.

Dozzi, S. P., and Simaan M. AbouRizk. 1993. *Productivity in Construction.* National Research Council Canada. http://archive.nrc-cnrc.gc.ca/obj/irc/doc/pubs/nrcc37001.pdf.

Eastman, Chuck. 2006. *IFC Overview.* Georgia Tech. https://open-acad-lib.googlecode.com/files/%E3%80%8ABIM%2BHandbook%E3%80%8B%E5%AF%B9IFC%E7%9A%84%E4%BB%8B%E7%BB%8D.pdf.

Eastman, Chuck, and R. Sacks. 2008. "Relative Productivity in the AEC Industries in the United States for On-Site and Off-Site Activities." *Journal of Construction Engineering and Management* 134 (7): 517–26. doi:10.1061/(ASCE)0733-9364(2008)134:7(517).

El-Abbasy, M., A. Senouci, T. Zayed, F. Mirahadi, and L. Parvizsedghy. 2014. "Condition Prediction Models for Oil and Gas Pipelines Using Regression

Analysis." *Journal of Construction Engineering and Management* 0 (0): 04014013. doi:10.1061/(ASCE)CO.1943-7862.0000838.

Fayek, A., and A. Oduba. 2005. "Predicting Industrial Construction Labor Productivity Using Fuzzy Expert Systems." *Journal of Construction Engineering and Management* 131 (8): 938–41. doi:10.1061/(ASCE)0733-9364(2005)131:8(938).

Flake, Gary William, and Steve Lawrence. 2002. "Efficient SVM Regression Training with SMO." *Machine Learning* 46 (1-3): 271–90. doi:10.1023/A:1012474916001.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. "The WEKA Data Mining Software: An Update." *SIGKDD Explorations* 11 (1).

Haykin, Simon S. 1999. *Neural Networks: A Comprehensive Foundation.* Prentice Hall.

Hergunsel, Mehmet F. 2011. "Benefits of Building Information Modeling for Construction Managers and BIM Based Scheduling." Worcester Polytechnic Institute. https://www.wpi.edu/Pubs/ETD/Available/etd-042011-135239/unrestricted/MHergunsel_Thesis_BIM.pdf.

Howard, Rob, and Bo-Christer Björk. 2008. "Building Information Modelling – Experts' Views on Standardisation and Industry Deployment." *Advanced Engineering Informatics*, Network methods in engineering, 22 (2): 271–80. doi:10.1016/j.aei.2007.03.001.

Hu, Di, and Yasser Mohamed. 2012. "Automating Fabrication Sequencing for Industrial Construction." *Gerontechnology* 11 (2). doi:10.4017/gt.2012.11.02.318.749.

Ikerd, W. F. 2008. "The Importance of BIM in Structural Engineering." *Structure Magazine.* http://www.structuremag.org/Archives/2008-10/C-Technology-Ikerd-Oct08.pdf.

Jung, Youngsoo, and Mihee Joo. 2011. "Building Information Modelling (BIM) Framework for Practical Implementation." *Automation in Construction* 20 (2): 126–33. doi:10.1016/j.autcon.2010.09.010.

Kalavagunta, Sreedhar. 2012. "Integrated Engineering Workflow (BIM) for Design of Steel Structures." In *Australasian Structural Engineering*

Conference 2012: The Past, Present and Future of Structural Engineering, 498. http://asec2012.conference.net.au/papers/063.pdf.

Karshenas, Saeed, and Xin Feng. 1992. "Application of Neural Networks in Earthmoving Equipment Production Estimating." In *Computing in Civil Engineering and Geographic Information Systems Symposium*, 841–47. ASCE. http://cedb.asce.org/cgi/WWWdisplay.cgi?9201884.

Lancaster, F., and J. Tobin. 2013. "Integrated Project Delivery: Next-Generation BIM for Structural Engineering." In *Structures Congress 2010*, 2809–18. American Society of Civil Engineers. Accessed November 18. http://ascelibrary.org/doi/abs/10.1061/41130%28369%29254.

Lee, Jong-Kyun, Boo-Young Kim, Jang-Young Kim, Tae-Hui Kim, and Kiyoung Son. 2013. "A Quantity Prediction Model for Reinforced Concrete and Bricks in Education Facilities Using Regression Analysis." *Journal of the Korea Institute of Building Construction* 13 (5): 506–12. doi:10.5345/JKIBC.2013.13.5.506.

Lu, Ming. 2001. "Productivity Studies Using Advanced ANN Models." Ph.D. thesis, Edmonton, Alberta.: University of Alberta. http://www.collectionscanada.gc.ca/obj/s4/f2/dsk1/tape4/PQDD_0012/NQ60322.pdf.

Mitchell, Tom M. 1997. *Machine Learning*. 1 edition. New York: McGraw-Hill Science/Engineering/Math.

Monteiro, André, and João Poças Martins. 2013. "A Survey on Modeling Guidelines for Quantity Takeoff-Oriented BIM-Based Design." *Automation in Construction* 35 (November): 238–53. doi:10.1016/j.autcon.2013.05.005.

National Building Information Model Standard Project Committee. 2014. "What Is a BIM?" *National BIM Standard - United States*. Accessed April 29. http://www.nationalbimstandard.org/faq.php#faq1.

Nawari, N. 2012. "BIM Standard in Off-Site Construction." *Journal of Architectural Engineering* 18 (2): 107–13. doi:10.1061/(ASCE)AE.1943-5568.0000056.

Peterson, Forest, Timo Hartmann, Renate Fruchter, and Martin Fischer. 2011. "Teaching Construction Project Management with BIM Support: Experience and Lessons Learned." *Automation in Construction* 20 (2): 115–25. doi:10.1016/j.autcon.2010.09.009.

Platt, John C. 1998. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Technical Report. MSR-TR-98-14. Microsoft Research.

Portas, Jason B. 1996. "Estimating Concrete Formwork Productivity." Master of Science, University of Alberta.

ResearchGate. 2013. "What Is the Difference between Machine Learning and Data Mining ?" Professional network. *ResearchGate*. http://www.researchgate.net/post/What_is_the_difference_between_machine_learning_and_data_mining.

Sattineni, Anoop, and R. Harrison Bradford. 2011. "Estimating with BIM: A Survey of US Construction Companies." *Proceedings of the 28th International Symposium on Automation and Robotics in Construction*, 564–69.

Shevade, S.K., S.S. Keerthi, C. Bhattacharyya, and K. R K Murthy. 2000. "Improvements to the SMO Algorithm for SVM Regression." *IEEE Transactions on Neural Networks* 11 (5): 1188–93. doi:10.1109/72.870050.

Shi, Wei. 2009. "Framework for Integration of BIM and RFID in Steel Construction." University of Florida. http://purl.fcla.edu/fcla/etd/UFE0041169.

Simon, Phil. 2013. *Too Big to Ignore: The Business Case for Big Data*. Hoboken, New Jersey: Wiley.

Smith, S. 1999. "Earthmoving Productivity Estimation Using Linear Regression Techniques." *Journal of Construction Engineering and Management* 125 (3): 133–41. doi:10.1061/(ASCE)0733-9364(1999)125:3(133).

Smola, Alex J., and Bernhard Schölkopf. 2004. "A Tutorial on Support Vector Regression." *Statistics and Computing* 14 (3): 199–222. doi:10.1023/B:STCO.0000035301.49549.88.

Song, Lingguang, and Simaan M. AbouRizk. 2008. "Measuring and Modeling Labor Productivity Using Historical Data." *Journal of Construction Engineering and Management* 134 (10): 786–94. doi:10.1061/(ASCE)0733-9364(2008)134:10(786).

Sonmez, R., and J. Rowings. 1998. "Construction Labor Productivity Modeling with Neural Networks." *Journal of Construction Engineering and*

Management 124 (6): 498–504. doi:10.1061/(ASCE)0733-9364(1998)124:6(498).

SteelConstruction.info. 2014. "The Case for Steel." *Steelconstruction.info*. Accessed April 27. http://www.steelconstruction.info/The_case_for_steel.

Steel, Jim, Robin Drogemuller, and Bianca Toth. 2012. "Model Interoperability in Building Information Modelling." *Software & Systems Modeling* 11 (1): 99–109. doi:10.1007/s10270-010-0178-4.

"Support Vector Machine Regression." 2014. *Support Vector Machine Regression*. Accessed July 22. http://kernelsvm.tripod.com/.

Tekla. 2014a. "Tekla." *Tekla*. Accessed April 30. http://www.tekla.com/us.

———. 2014b. "Structural Steel Detailing and Fabrication Software | Tekla." *Tekla*. Accessed June 24. http://www.tekla.com/solutions/steel-fabricators.

The University of Waikato. 2014. "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." *WEKA*. Accessed April 30. http://www.cs.waikato.ac.nz/ml/weka/.

Tiwari, Saurabh, Josh Odelson, Alan Watt, and Atul Khanzode. 2009. "Model Based Estimating to Inform Target Value Design." *AECbytes "Building the Future" Article*. http://www.aecbytes.com/buildingthefuture/2009/ModelBasedEstimating.html.

Vertigraph Inc. 2004. "Automating the Takeoff and Estimating Process." www.vertigraph.com.

Whitt, Michael. 2012. *Successful Instrumentation and Control Systems Design*. 2nd edition. Research Triangle Park, NC: ISA - The Instrumentation, Systems & Automation Society.

Williams, Alan. 2011. *Steel Structures Design: ASD/LRFD*. 1 edition. New York: McGraw-Hill Professional.

Witten, Ian H., Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. 3 edition. Burlington, MA: Morgan Kaufmann.

Xie, Haiyan, Wei Shi, and Raja RA Issa. 2010. "Implementation of BIM/RFID in Computer-Aided Design-Manufacturing-Installation Process." In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International*

Conference on, 2:107–11. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5563653.

APPENDIX A: SHOP LOADING TOOL MACRO CODE

Module 1:

```
Sub ButtonStart_Click()

    Dim shopSheet As Worksheet

    Set shopSheet = ThisWorkbook.ActiveSheet

    Call SpreadWork(shopSheet, shopSheet.Name)

End Sub


Sub ButtonClear_Click()

    Dim shopSheet As Worksheet

    Set shopSheet = ThisWorkbook.ActiveSheet

    Call ClearSheet(shopSheet)

End Sub
```

Module 2:

```
' Clear all the data of Row 4 and below.

Sub ClearSheet(ByRef mysh As Worksheet)

    mysh.Rows("4:" & mysh.Rows.Count).ClearContents

    ' Write the column titles back.

    mysh.Cells(4, 1).Value = "Name"

    mysh.Cells(4, 2).Value = "MHrs"

    mysh.Cells(4, 3).Value = "Start"

    mysh.Cells(4, 4).Value = "Finish"

    ' Reset the formula in cell B1

    mysh.Cells(1, 2).Formula = "=COUNTA(A:A)-2"

    Dim chartObj As ChartObject
```

```vba
    For Each chartObj In mysh.ChartObjects

        chartObj.Delete

    Next

End Sub


Sub SpreadWork(ByRef mysh As Worksheet, ByVal chtName As
String)

    If mysh.Range("B1").Value = "" Or mysh.Range("B1").Value
= 0 Then

        Call MsgBox("Please enter the Number of Records.",
vbOKOnly, "CAUTION!")

        Exit Sub

    End If


    Dim numOfRecords As Integer

    numOfRecords = mysh.Range("B1").Value


    Dim wkRow As Integer

    Dim wkCol As Integer

    wkRow = 4 ' Row number 4

    wkCol = 5 ' Column E


    Dim minStart As Date

    Dim maxFinish As Date

    minStart = WorksheetFunction.min(mysh.Range("C" & (wkRow
+ 1) & ":C" & (wkRow + numOfRecords)))
```

```vba
    maxFinish  =  WorksheetFunction.Max(mysh.Range("D"  &
(wkRow + 1) & ":D" & (wkRow + numOfRecords)))

    Dim minSun As Date

    Dim maxSun As Date

    minSun = FirstDayofWeek(minStart)

    maxSun = FirstDayofWeek(maxFinish)


    mysh.Cells(wkRow, wkCol).Select ' Select the start point
of the timeline axis.

    Dim axisDate As Date

    axisDate = minSun

    ' Create the horizontal time axis.

    Do

        Selection.Value = axisDate

        Selection.NumberFormat = "dd-mmm-yy"

        Selection.Offset(0, 1).Select ' Move right by one
cell.

        axisDate = axisDate + 7

    Loop While axisDate <= maxSun


    ' Clear all the cell contents.

    Dim colMax As String

    colMax = ColumnLetter(Selection.Column - 1)

    Dim wkColName As String

    wkColName = ColumnLetter(wkCol)
```

```
    mysh.Range(wkColName & (wkRow + 1) & ":" & colMax &
(wkRow + numOfRecords + 1)).ClearContents


    Dim counter As Integer

    For counter = 1 To numOfRecords

        Dim taskStart As Date

        Dim taskFinish As Date

        taskStart    =    mysh.Range("C"    &    (wkRow    +
counter)).Value

        taskFinish   =    mysh.Range("D"    &    (wkRow    +
counter)).Value

        ' Check

        If taskStart > taskFinish Then

            Call  MsgBox("Start  Date  is  later  than  Finish
Date  on  line  "  &  (wkRow  +  counter)  &  ".",  vbOKOnly,
"CAUTION!")

            Exit Sub

        End If


        Dim avgMh As Double

        avgMh = mysh.Range("B" & (wkRow + counter)).Value /
(taskFinish - taskStart)

        Dim firstWk As Date

        Dim lastWk As Date

        firstWk = FirstDayofWeek(taskStart)

        lastWk = FirstDayofWeek(taskFinish)
```

94

```
        mysh.Cells(wkRow, wkCol).Select

        ' Set workload of the first week

        Do While Selection.Value < firstWk

            Selection.Offset(0, 1).Select

        Loop

        mysh.Cells((wkRow          +          counter),
Selection.Column).Value = avgMh * (Selection.Value + 7 -
taskStart)


        ' Set workload in the middle week(s)

        Selection.Offset(0, 1).Select

        Do While Selection.Value < lastWk

            mysh.Cells((wkRow          +          counter),
Selection.Column).Value = avgMh * 7

            Selection.Offset(0, 1).Select

        Loop


        ' Set workload of the last week

        mysh.Cells((wkRow          +          counter),
Selection.Column).Value   =   avgMh   *   (taskFinish   -
Selection.Value)

    Next counter


    mysh.Cells(wkRow + numOfRecords + 1, wkCol - 1).Value =
"Sum"
```

```vba
    mysh.Cells(wkRow, wkCol).Select

    Dim colName As String

    Dim row As Integer

    row = wkRow + numOfRecords + 1

    Do While Selection.Value <> Empty And Selection.Value <=
maxSun

        colName = ColumnLetter(Selection.Column)

        Dim loadSum As Double

        loadSum = WorksheetFunction.sum(mysh.Range(colName &
(wkRow + 1) & ":" & colName & (wkRow + numOfRecords)))

        mysh.Cells(row, Selection.Column).Value = loadSum

        Selection.Offset(0, 1).Select

    Loop

    ' Add a bar chart.

    Dim objChart As ChartObject

    On Error Resume Next

    Set objChart = mysh.ChartObjects(chtName & " Chart")

    If objChart Is Nothing Then

        Set objChart = mysh.ChartObjects.Add(200, 500, 600,
300)

        With objChart

            .Name = chtName & " Chart"

            .Chart.HasTitle = True

            .Chart.ChartTitle.Text = chtName & " Chart"

            .Chart.HasLegend = False

            .Chart.Axes(xlValue).HasTitle = True
```

```vba
        .Chart.Axes(xlValue).AxisTitle.Text   =   "Weekly
MHrs"

        End With

    End If

    With mysh.ChartObjects(chtName & " Chart")

        .Chart.SetSourceData  Source:=mysh.Range(wkColName  &
row & ":" & colMax & row) ' Weekly MHrs

        .Chart.ChartType = xlColumnClustered ' Bar chart

        .Chart.SeriesCollection(1).XValues                =
mysh.Range(wkColName  &  wkRow  &  ":"  &  colMax  &  wkRow)  '
Sunday dates

        .Chart.Axes(xlCategory).CategoryType              =
xlCategoryScale

        .Chart.Axes(xlCategory).TickLabelSpacing = 1

    End With

    mysh.Cells(1, 1).Select

End Sub


Function FirstDayofWeek(ByVal target As Date) As Date

    FirstDayofWeek = target - Weekday(target, vbSunday) + 1

End Function


Function ColumnLetter(ByVal colNum As Integer) As String

    arr = Split(Cells(1, colNum).Address(True, False), "$")

    ColumnLetter = arr(0)

End Function
```

APPENDIX B: LOADING COMPARISON TOOL MACRO

CODE

Module 1:

```
Sub ButtonStart_Click()

    Dim shopSheet As Worksheet

    Set shopSheet = ThisWorkbook.ActiveSheet

    Call SpreadWork(shopSheet)

End Sub


Sub ButtonClear_Click()

    Dim shopSheet As Worksheet

    Set shopSheet = ThisWorkbook.ActiveSheet

    Call ClearSheet(shopSheet)

End Sub
```

Module 2:

```
Option Explicit

' Clear all the data below the column title at Row 4.

Sub ClearSheet(ByRef mysh As Worksheet)

    mysh.Rows("5:" & mysh.Rows.Count).ClearContents

    ' Reset the formula in cell B1 and F1

    mysh.Cells(1, 2).Formula = "=COUNTA(A:A)-3"

    mysh.Cells(1, 6).Formula = "=COUNTA(E:E)-3"
```

```vba
    Dim chartObj As ChartObject

    For Each chartObj In mysh.ChartObjects

        chartObj.Delete

    Next

End Sub


Sub SpreadWork(ByRef mysh As Worksheet)

    If mysh.Range("B1").value = "" Or mysh.Range("B1").value
= 0 Then

        Call MsgBox("Please enter the Number of Records and
MHrs per Tonne.", vbOKOnly, "CAUTION!")

        Exit Sub

    End If


    Dim ConsWkRow As Integer ' Two constants

    Dim ConsWkCol As Integer

    ConsWkRow = 5 ' Row number 5, the row below the column
titles.

    ConsWkCol = 9 ' Column I


    '''' Schedule A ''''

    Dim numOfRecords As Integer

    numOfRecords = mysh.Range("B1").value

    Dim minStart As Date

    Dim maxFinish As Date
```

```
    minStart    =    WorksheetFunction.min(mysh.Range("C"    &
ConsWkRow & ":C" & (ConsWkRow + numOfRecords - 1)))

    maxFinish    =    WorksheetFunction.Max(mysh.Range("D"    &
ConsWkRow & ":D" & (ConsWkRow + numOfRecords - 1)))

    Dim minSun As Date ' Start and end of the timeline

    Dim maxSun As Date

    minSun = FirstDayofWeek(minStart)

    maxSun = FirstDayofWeek(maxFinish)


    Dim wkLoading As Collection

    Set wkLoading = New Collection

    Dim axisDate As Date

    axisDate = minSun

    ' Create the horizontal time axis.

    Dim tmp As ClsWkLoad

    Do

        Set tmp = New ClsWkLoad

        tmp.Init axisDate

        wkLoading.Add tmp

        axisDate = axisDate + 7

    Loop While axisDate <= maxSun


    ''' Main Calculation '''

    Dim counter As Integer

    For counter = 1 To numOfRecords

        Dim taskStart As Date
```

```vba
        Dim taskFinish As Date

        taskStart = mysh.Range("C" & (ConsWkRow + counter -
1)).value

        taskFinish = mysh.Range("D" & (ConsWkRow + counter -
1)).value

        ' Check

        If taskStart > taskFinish Then

            Call MsgBox("Start Date is later than Finish
Date on line " & (ConsWkRow + counter - 1) & ".", vbOKOnly,
"CAUTION!")

            Exit Sub

        End If


        Dim avgMh As Double

        avgMh = mysh.Range("B" & (ConsWkRow + counter -
1)).value / (taskFinish - taskStart) ' Daily average Mhrs

        Dim firstWk As Date

        Dim lastWk As Date

        firstWk = FirstDayofWeek(taskStart)

        lastWk = FirstDayofWeek(taskFinish)


        Dim firstIdx As Integer ' Index in the collection

        Dim lastIdx As Integer

        firstIdx = (firstWk - wkLoading.Item(1).Week) / 7 +
1

        lastIdx = (lastWk - wkLoading.Item(1).Week) / 7 + 1
```

```vba
        wkLoading.Item(firstIdx).AddHr avgMh * (firstWk + 7
- taskStart), 1

        Dim index As Integer

        For index = firstIdx + 1 To lastIdx - 1

            wkLoading.Item(index).AddHr avgMh * 7, 1

        Next index

        wkLoading.Item(lastIdx).AddHr avgMh * (taskFinish -
lastWk), 1

    Next counter


    '''' Schedule B ''''

    numOfRecords = mysh.Range("F1").value


    If numOfRecords > 0 Then

        minStart = WorksheetFunction.min(mysh.Range("G" &
ConsWkRow & ":G" & (ConsWkRow + numOfRecords - 1)))

        maxFinish = WorksheetFunction.Max(mysh.Range("H" &
ConsWkRow & ":H" & (ConsWkRow + numOfRecords - 1)))

        minSun = FirstDayofWeek(minStart) ' Start and end of
the timeline

        maxSun = FirstDayofWeek(maxFinish)


        axisDate = wkLoading.Item(1).Week ' The first date
on the current time axis

        Do While axisDate > minSun ' The axis is not enough.
```

102

```
        axisDate = axisDate - 7

        Set tmp = New ClsWkLoad

        tmp.Init axisDate

        wkLoading.Add tmp, , 1 ' Insert in the beginning

    Loop

    axisDate = wkLoading.Item(wkLoading.Count).Week

    Do While axisDate < maxSun

        axisDate = axisDate + 7

        Set tmp = New ClsWkLoad

        tmp.Init axisDate

        wkLoading.Add tmp, , , wkLoading.Count ' Insert
to the end

    Loop


    ''' Main Calculation '''

    For counter = 1 To numOfRecords

        taskStart = mysh.Range("G" & (ConsWkRow +
counter - 1)).value

        taskFinish = mysh.Range("H" & (ConsWkRow +
counter - 1)).value

        ' Check

        If taskStart > taskFinish Then

            Call MsgBox("Start Date is later than Finish
Date on line " & (ConsWkRow + counter - 1) & ".", vbOKOnly,
"CAUTION!")

            Exit Sub
```

```
            End If


            avgMh = mysh.Range("F" & (ConsWkRow + counter -
1)).value / (taskFinish - taskStart) ' Daily average Mhrs
            firstWk = FirstDayofWeek(taskStart)
            lastWk = FirstDayofWeek(taskFinish)
            firstIdx = (firstWk - wkLoading.Item(1).Week) /
7 + 1
            lastIdx = (lastWk - wkLoading.Item(1).Week) / 7
+ 1


            wkLoading.Item(firstIdx).AddHr avgMh * (firstWk
+ 7 - taskStart), 2
            For index = firstIdx + 1 To lastIdx - 1
                wkLoading.Item(index).AddHr avgMh * 7, 2
            Next index
            wkLoading.Item(lastIdx).AddHr       avgMh      *
(taskFinish - lastWk), 2
        Next counter
    End If


    ''' Fill the weekly summary in the spreadsheet. '''
    mysh.Cells(ConsWkRow, ConsWkCol).Select ' Start of the
timeline axis
    For counter = 1 To wkLoading.Count
        Selection.value = wkLoading.Item(counter).Week
```

```vba
        Selection.NumberFormat = "dd-mmm-yy"

        Selection.Offset(1,          0).value          =
wkLoading.Item(counter).MhrA

        Selection.Offset(2,          0).value          =
wkLoading.Item(counter).MhrB

        Selection.Offset(0, 1).Select ' Move right.

    Next counter

    Dim colMax As String

    colMax = ColumnLetter(Selection.Column - 1)


    Dim xSeries As String

    Dim y1Series As String

    Dim y2Series As String

    Dim y1Name As String

    Dim y2Name As String

    xSeries = ColumnLetter(ConsWkCol) & ConsWkRow & ":" &
colMax & ConsWkRow

    y1Name = mysh.Range("B3").value

    y2Name = mysh.Range("F3").value

    y1Series = ColumnLetter(ConsWkCol) & (ConsWkRow + 1) &
":" & colMax & (ConsWkRow + 1)

    y2Series = ColumnLetter(ConsWkCol) & (ConsWkRow + 2) &
":" & colMax & (ConsWkRow + 2)

    Call  DrawChart(mysh,  mysh.Name,  y1Name,  xSeries,
y1Series)
```

```
    Call    DrawChart(mysh,    mysh.Name,    y2Name,    xSeries,
y2Series)


    Set wkLoading = Nothing

    mysh.Cells(1, 1).Select
End Sub


' Find the nearest Sunday before target, which represents
the first day of the week Target belongs to.
Function FirstDayofWeek(ByVal target As Date) As Date

    FirstDayofWeek = target - Weekday(target, vbSunday) + 1
End Function


' Conver a column to a letter.
Function ColumnLetter(ByVal colNum As Integer) As String

    Dim arr() As String

    arr = Split(Cells(1, colNum).Address(True, False), "$")

    ColumnLetter = arr(0)
End Function


' Create a bar chart.
Sub DrawChart(ByRef mysh As Worksheet, ByVal chtName As
String, ByVal seName As String, ByVal xSeries As String,
ByVal ySeries As String)

    Dim objChart As ChartObject

    On Error Resume Next
```

```
    Set objChart = mysh.ChartObjects(chtName & " Chart")

    If objChart Is Nothing Then ' Create if not exist

        Set objChart = mysh.ChartObjects.Add(200, 500, 600,
300)

        With objChart

            .Name = chtName & " Chart"

            .Chart.HasTitle = True

            .Chart.ChartTitle.Text = chtName & " Chart"

            .Chart.HasLegend = True

            .Chart.Legend.Position = xlLegendPositionTop

            .Chart.Axes(xlValue).HasTitle = True

            .Chart.Axes(xlValue).AxisTitle.Text  =  "Weekly
MHrs"

        End With

    End If


    With mysh.ChartObjects(chtName & " Chart")

        With .Chart.SeriesCollection.NewSeries

            .XValues = mysh.Range(xSeries) ' Sunday dates

            .Values = mysh.Range(ySeries) ' Weekly MHrs

            .Name = seName

        End With

        .Chart.ChartType = xlColumnClustered ' Bar chart

        .Chart.Axes(xlCategory).CategoryType             =
xlCategoryScale

        .Chart.Axes(xlCategory).TickLabelSpacing = 1
```

```
        End With

End Sub


        Class Module:

Option Explicit


Private pWeek As Date

Private pMhrA As Double

Private pMhrB As Double


Public Property Get Week() As Date

    Week = pWeek

End Property


Public Property Get MhrA() As Double

    MhrA = pMhrA

End Property

Public Property Let MhrA(value As Double)

    If value < 0 Then

        pMhrA = 0

    Else

        pMhrA = value

    End If

End Property


Public Property Get MhrB() As Double
```

```vba
        MhrB = pMhrB

End Property

Public Property Let MhrB(value As Double)

    If value < 0 Then

        pMhrB = 0

    Else

        pMhrB = value

    End If

End Property


Public Sub Init(value As Date)

    pWeek = value

    pMhrA = 0

    pMhrB = 0

End Sub


Public Sub AddHr(value As Double, setNo As Integer)

    If setNo = 1 Then

        pMhrA = pMhrA + value

    Else

        pMhrB = pMhrB + value

    End If

End Sub
```