Journal of Computer and Communications, 2014, 2, 1-27
Published Online March 2014 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2014.25001

Scientific Research

Knowledge Discovery in Data: A Case Study

Ahmed Hammad[1], Simaan AbouRizk[2]

[1]HMD Project & Knowledge Management Services, Edmonton, Canada
[2]Department of Civil and Environmental Engineering, Hole School of Construction Engineering and Management, University of Alberta, 3-014 Markin/CNRL Natural Resources Engineering Facility, Edmonton, Alberta, Canada
Email: ahmed.hammad@worleyparsons.com, abourizk@ualberta.ca

Received 13 December 2013; revised 10 January 2014; accepted 18 January 2014

Abstract

It is common in industrial construction projects for data to be collected and discarded without being analyzed to extract useful knowledge. A proposed integrated methodology based on a five-step Knowledge Discovery in Data (KDD) model was developed to address this issue. The framework transfers existing multidimensional historical data from completed projects into useful knowledge for future projects. The model starts by understanding the problem domain, industrial construction projects. The second step is analyzing the problem data and its multiple dimensions. The target dataset is the labour resources data generated while managing industrial construction projects. The next step is developing the data collection model and prototype data warehouse. The data warehouse stores collected data in a ready-for-mining format and produces dynamic On Line Analytical Processing (OLAP) reports and graphs. Data was collected from a large western-Canadian structural steel fabricator to prove the applicability of the developed methodology. The proposed framework was applied to three different case studies to validate the applicability of the developed framework to real projects data.

Keywords

Construction Management; Project Management; Knowledge Management; Data Warehousing; Data Mining; Knowledge Discovery in Data (KDD); Industrial Construction; Labour Resources

1. Introduction

Many industrial construction projects face delays and budget overruns, often caused by improper management of labour resources [1]. The nature of industrial construction projects makes them more complicated: a large number of

stakeholders with conflicting interests, sophisticated management tools, stricter safety and environmental concerns. In the changing environment, each involved contractor simultaneously manages multiple projects using one pool of resources. During this process, a large amount of data is generated, collected, and stored in different formats, but it is not analyzed to extract useful knowledge. The improvement of labour management practices could have a significant impact on reducing schedule delays and budget overruns. One solution to this problem is analysis of historical labour resources data from completed projects to extract useful knowledge that can be transferred and used to improve resource management practices.

Data warehouses are one method often used to extract useful knowledge. They are dedicated, read-only, and non-volatile databases that centrally store validated, multidimensional, historical data from Operation Support Systems (OSS) to be used by Decision Support Systems (DSS) [2]. Data warehouses are typically structured either on the star schema, consisting of a fact table that contains the data and dimension tables that contain the attributes of this data, for simple datasets, and on the snowflake schema, used either when multiple fact tables are needed or when dimension tables are hierarchical in nature [3], for complicated datasets. A data warehouse typically consists of three main components: the data acquisition systems (backend), the central database, and the knowledge extraction tools (frontend) [4]. On Line Analytical Processing (OLAP) techniques (roll-up and drill-down, slice and dice, and data pivoting) are typically used in the frontend of a data warehouse to present end-users with a dynamic tool to view and analyze stored data.

Data mining is "the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owners" [5]. Considering data mining, the knowledge discovered must be previously unknown, non-trivial, and useful to the data owners [6]. Data mining techniques rely on either supervised or unsupervised learning and are grouped into four categories [7]. Clustering methods minimize the distance between data points falling within a cluster, and maximize the distance between these clustered data points and other clusters [8]. Finding Association Rules highlights hidden patterns in large datasets. Classification techniques, including Decision Trees, Rule-Based Algorithms, Artificial Neural Networks (ANN), k-Nearest Neighbours (k-NN or lazy learning), Support Vector Machine (SVM), and many others, build a model using a training dataset to define data classes, evaluate the model, and then use the developed model to classify each new data point into the appropriate class [7]. Outliers' detection techniques focus on data points that are significantly different from the rest.

Data warehousing and mining techniques have been applied to solve problems in the construction industry over the last decade. However, none of the previous research applied these techniques to address management of multiple projects simultaneously using one common pool of labour resources; the problem is typically solved using other techniques (Heuristic rules, Numerical Optimization and Genetic Algorithms). Most previous research focused on leveling or allocating resources in a single project environment. Soibelman and Kim [9] analyzed schedule delays with a five-step KDD approach. Chau *et al*. [10] developed the Construction Management Decision Support System (CMDSS) by combining data warehousing, Decision Support Systems (DSS) and OLAP. Rujirayanyong and Shi [11] developed a Project-oriented Data Warehouse (PDW) for contractors, but it was limited to querying the warehouse without using data mining. Moon *et al*. [12] used a four-dimension cost data cube in their application of Cost Data Management System (CDMS), built using MS SQL Server-OLAP Analysis Services, to obtain more reliable estimates of construction costs. Fan *et al*. [13] used the Auto Regression Tree (ATR) data mining technique to predict the residual value of construction equipment.

In this research, the Cios *et al*. [7] hybrid model was modified and adapted to develop an integrated methodology for extracting useful knowledge from collected labour resources data in a multiple-project environment utilizing the concepts of KDD, data warehousing, and data mining. When the techniques are integrated, they combine quantitative and qualitative research approaches and facilitate working with large amounts of data impacted by a large number of unknown variables, which was integral to this research. Further information on the developed framework can be found in Hammad *et al*. [14]. The proposed integrated methodology based on a five-step Knowledge Discovery in Data (KDD) model is shown in **Figure 1**.

In this paper, the proposed modified hybrid KDD model is applied to three different case studies to test its ability to extract useful knowledge from datasets. Section 2 discusses discovering knowledge in the first dataset; Section 3 covers the second dataset and Section 4 the third dataset. The paper outlines the process of applying the model to extract data, the related procedures, and outlines the useful data collected.

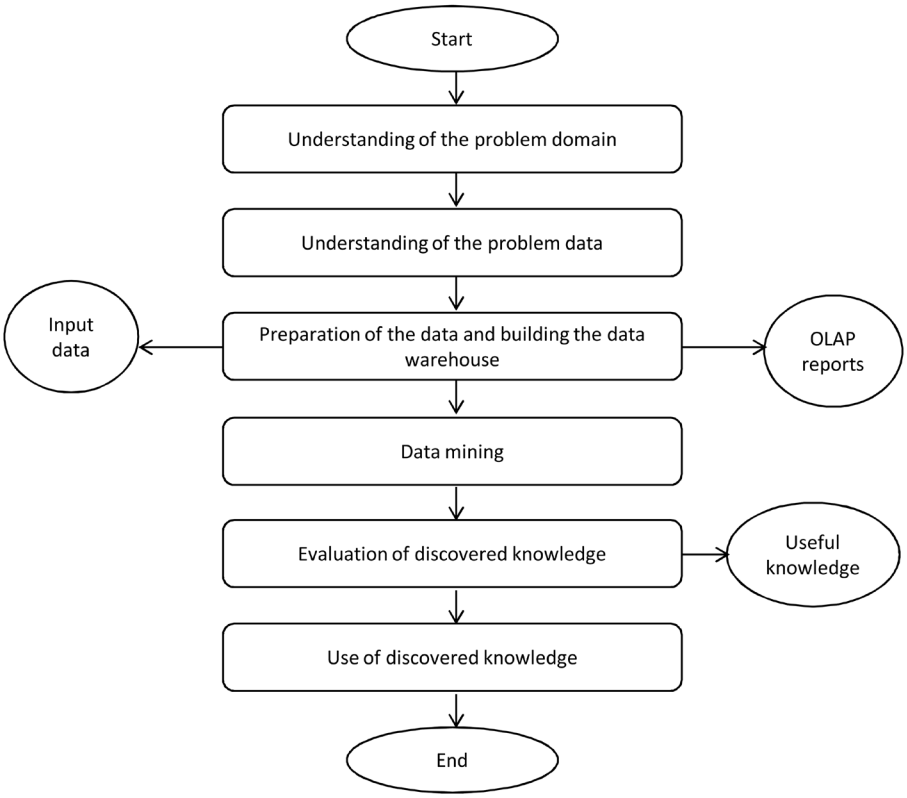

Figure 1. Modified hybrid KDD model.

2. Discovering Knowledge in the First Dataset

2.1. Data Cleaning and Preprocessing

This paper shows how to implement data mining techniques to extract useful knowledge from three datasets that were obtained from real projects. The purpose of this case study is to validate that data mining can be used to improve and increase the efficiency of labour estimating practices in contracting companies. Most of these companies rely on cost estimating units (norms) that are not based on historical data and are not updated to reflect changes in the industry. Applying the proposed approach that relies on data mining is expected to provide companies with knowledge-based probabilistic dynamic estimating units that always reflect the latest changes.

The first dataset contains data regarding the scope of a set of engineering work packages. This scope is represented as determinate amounts of key quantities per work package. The key quantity for engineering packages is the number of engineering deliverables. The dataset was obtained from the estimating system of the participating contractor. This estimating system is based on an old version of MS Access. The dataset contains the original and current baseline hours for five of the involved resources in this group of work packages. The current baseline values reflect the project scope after implementing all approved changes. The selected dataset to be analyzed in this case study contained data for more than one hundred projects, four project phases and five different resources.

The contractor did not track actual hours spent per work package; however, the same analysis can be easily applied if the data exists. The analysis was used to check the consistency of the estimating practices in this contracting company. The data was directly exported from the estimating system to MS Excel, where the cleaning and preprocessing took place prior to exporting the data to the data warehouse. **Figure 2** shows an example of the raw data. Data was found to be missing, and also, metadata (data about the data) was missing. The data lacked the values for two important control attributes: the internal program and the project phase, and had to be assigned manually.

This manual procedure required going back to the archived project documents to find the appropriate values

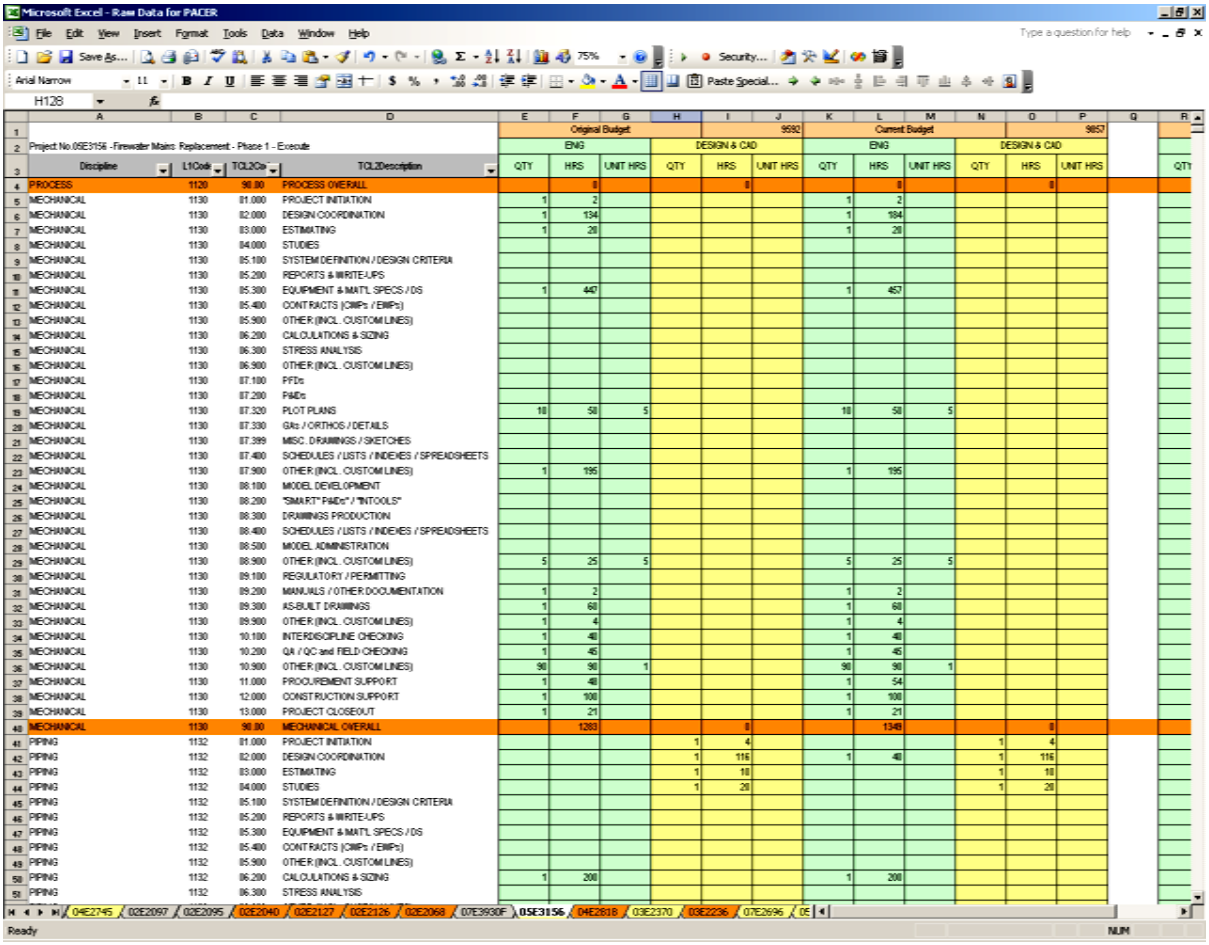

Figure 2. The raw data for the first data set.

to be assigned to each data point. Completing and verifying the dataset required ample time and effort. Furthermore, while working with the archived documents, some documents were not clear enough and assumptions had to be made to compensate for the missing data. Many of the staff members who were involved in these projects could not be located, and if they could be contacted, they could not provide meaningful input on the data as time has passed. All the effort and time spent searching, sorting, and cleaning in the archive would have been easily avoided if the data and its metadata were collected in the proposed integrated format.

Figure 3 shows the data from **Figure 2** after it was cleaned, pre-processed and ready for storage in the data warehouse. The objective of this analysis was to test if the resource unit cost per production package type could be extracted for future estimating of resource requirements in upcoming projects. For this analysis, fifteen standard production packages, three engineering phases and five engineering resources were selected. The data was modified by random numbers for confidentiality issues. This modified data was used in the analysis. Hence, all numbers shown here are used only for illustration.

Three control attributes were selected for this analysis. These control attributes were represented in this analysis with the independent variables: Package($_{1:15}$), Phase($_{1:3}$) and Resource($_{1:5}$). These are nominal variables with values assigned to them as discrete integers. These discrete integers were equal to the ID's used in the data warehouse for direct referencing. To test the significance of adding more attributes to the analysis, the internal program control attribute was selected. This attribute is represented in the analysis with the independent variable Program($_{1:3}$).

In this study, the term "class" refers to a unique combination of values of the three variables: package, phase and resource. For example, class 1 contains all the data points that have the value Pk($_1$) for the variable package, Ph($_1$) for the variable phase and R($_1$) for the variable resource. A class 2 contains all the data points that have the

Program	Project	Package	Phase	Resource	Original Unit Cost	Current Unit Cost
2	34	13	9	5	0.50	0.50
2	43	13	10	2	2.00	2.00
2	47	13	10	2	2.00	2.00
2	67	13	9	1	4.00	
2	118	13	9	1	4.00	4.00
2	39	13	10	1	4.00	4.00
2	69	13	9	2	5.00	
1	9	13	9	1	6.00	
1	13	13	9	1	6.00	
1	11	13	9	1	10.00	
1	27	13	9	1	10.00	
1	3	13	10	1	10.00	
1	6	13	10	1	10.00	
1	24	13	10	1	10.00	
2	36	13	9	6	15.00	15.00
1	18	13	9	1	20.00	
1	21	13	9	1	20.00	
2	36	13	9	1	20.00	20.00
1	17	13	8	1	25.00	
1	20	13	8	1	25.00	
2	63	13	9	1	40.00	40.00
1	26	13	8	1	330.00	
2	43	14	10	2	1.00	1.00
2	69	14	9	5	1.00	
1	34	14	9	6	1.00	1.00
2	67	14	9	6	1.00	
2	32	14	10	5	1.50	1.50
1	11	14	9	1	2.00	
2	120	14	9	1	2.00	2.00
2	47	14	10	1	2.00	2.00
1	11	14	9	2	2.00	
2	118	14	9	2	2.00	2.00
1	22	14	10	2	2.00	
2	39	14	10	2	2.00	2.00
2	45	14	9	5	2.00	2.00
2	67	14	9	5	2.00	

Figure 3. The dataset after cleaning and pre-processing.

value $Pk(_1)$ for the variable package, $Ph(_1)$ for the variable phase and $R(_2)$ for the variable resource, etc. The number of classes resulting from all the possible combinations is calculated using the formula:

$$\begin{aligned}\text{Number of Classes} &= \text{Number of Package} * \text{Number of Phases} * \text{Number of Resources} \\ &= 15 * 3 * 5 = 225 \text{ Classes}\end{aligned} \quad (1)$$

It is important to note that the dataset may not include data points for all the classes. Certain classes of the three main attributes do not exist in reality. For example, some packages are not needed in every phase, or some packages do not utilize all the five resources under investigation.

The key quantity for all the packages is the number of engineering deliverables. This analysis is implemented to the hourly portion of the collected data, since estimating of labour resource requirements relies on hourly units and not on cost. The dataset was normalized to eliminate the differences in project sizes by calculating three dependent variables: "Original Hourly Unit Cost," "Current Hourly Unit Cost," and "Actual Hourly Unit Cost." These variables are calculated using the following formulas:

$$\text{Original Hourly Unit Cost} = \text{Original Baseline Hours} / \text{Original Quantity} \quad (2)$$

$$\text{Current Hourly Unit Cost} = \text{Current Baseline Hours} / \text{Current Quantity} \quad (3)$$

$$\text{Actual Hourly Unit Cost} = \text{Actual Hours} / \text{Actual Quantity} \quad (4)$$

Because of the multidimensionality of this dataset, four new variables were formulated to represent the possible combinations of the three main attributes. These new variables were Package/Phase, Package/Resource,

Phase/Resource and Package/Phase/Resource. These variables were assigned unique values by combining ID's from the three main attributes.

After defining all the necessary variables, the dataset was then exported to the first analysis tool, SPSS-16 for Windows. SPSS was selected because of its ability to perform a wide range of statistical analysis tests, its ability to easily import and export data from databases and its user friendliness. It can be easily obtained by any contractor or industrial owner who needs to perform statistical analysis of the collected data in the data warehouse.

The objective of this analysis was to develop an estimating methodology to be implemented using unit costs and key quantities. First, the dataset was divided into clusters using stratification. Significant differences of means are used to establish these clusters. Within each cluster, unit cost and the characteristics of the most fitting distribution are obtained. Therefore, instead of relying solely on their intuitions, the estimators are presented with mined values for the unit costs that can be multiplied by the known determinate key quantities in order for these estimators to predict the resource requirements more accurately.

2.2. The Initial Investigation

Data mining models suggest starting any exercise with visual presentation of the available dataset. First, the frequency of data points within each independent variable is plotted. **Figure 4** graphically shows that phase Ph-03 had more data points than the other two phases. **Figure 5** shows that not every package utilized the five resources and that some packages only utilize a single resource.

Second, the data descriptive 'case summaries' test is performed to collect statistics on each class or data subset. Since the data is multidimensional, subsets can be generated using one attribute, a combination of any two attributes, or all three attributes combined. The following statistics are obtained: mean, standard deviation, number of data points, minimum value, maximum value and data range. **Figure 6** shows an excerpt of the test results.

Subsequently, statistical dispersion is measured using boxplots that are obtained for each of the data subsets. Boxplots show the Inter Quartile Range - IQR (the 25^{th} percentile, the median, 75^{th} percentile) minimum, maximum and extreme values. SPSS points to the raw number that contains data points that are out of the normal range.

The descriptive statistics as well as the boxplots showed very wide ranges and variance (**Figures 7** and **8**). They also showed that the dataset contains extreme outliers. As a result of this situation, it was necessary to implement an outlier detection procedure.

2.3. The Outliers Detection Procedure

Given that the boxplot results showed outliers in the dataset, detecting them was necessary. In this research, the technique implemented was based on Chebyshev Theorem [15]. This theorem can be used for single dimension (univariate) outliers analysis. Assuming the dataset follows a normal distribution, the mean and standard deviation of the distribution can be defined by calculating the mean (μ) and standard deviation (σ) of the dataset. Chebyshev stated that since most data points fall between ($\mu + 3\sigma$) and ($\mu - 3\sigma$), those that fall outside of this range can be considered outliers.

A four layer outlier analysis tool was developed based on the three-dimensional dataset.

- First layer = all data

  Second layer = each attribute
- Third layer = three possible combinations of paired attributes represented as three new category variables (package * phase provides 45 combinations, package * resource provides 75 combinations and phase * resource provides 15 combinations).
- Fourth layer = all attributes combined (provides 225 combinations) represented as new category variable.

A total of eight possible cases of outliers were calculated using the obtained means and standard deviations obtained from SPSS. Each data point was tested against the eight cases and was assigned a value of 1 if found to be an outlier in any case. A total outlier score is calculated by adding the number of cases where a data point was an outlier. An example of the output is shown in **Figure 9**. It is up to the user to go back and verify the outliers or eliminate them and perform the analysis. The procedure was repeated three times until the obtained standard deviations and ranges were found to be acceptable as shown in **Figures 10** and **11**.

Cases with less than three data points were eliminated from the analysis. The mean and standard deviation of

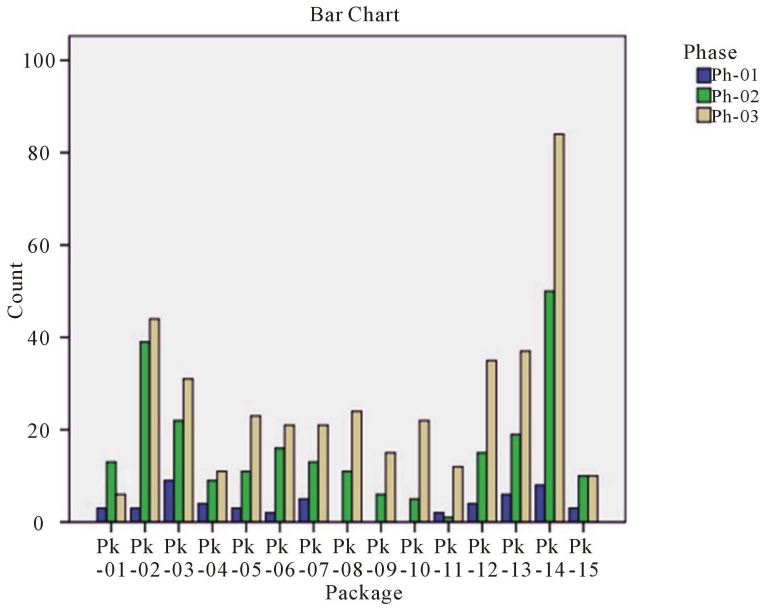

Figure 4. Frequency of data points within the three phases.

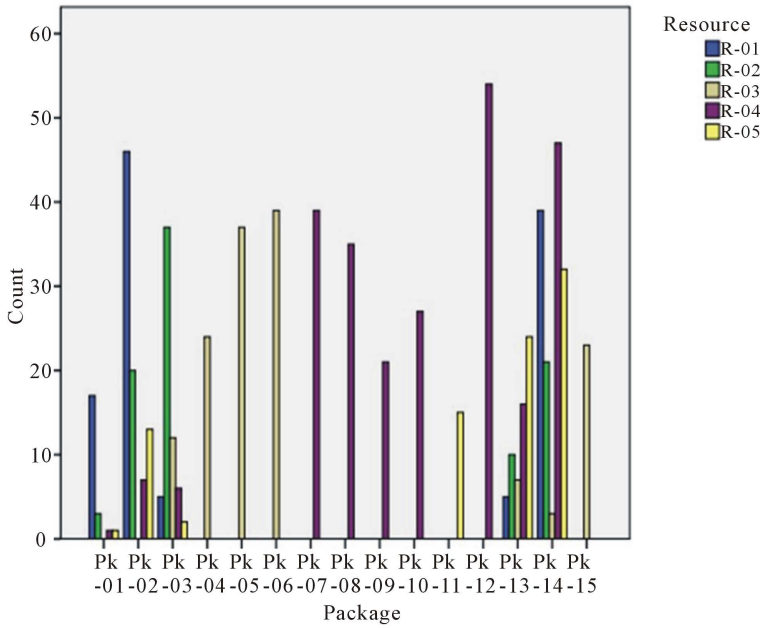

Figure 5. Frequency of data points within the five resources.

every class was calculated and summarized graphically on a tree, as shown in **Figure 12**. The user can now use the summary tree to find out the unit cost multiplier distributions to be used for estimating new projects in the future. For each layer, a new variable Select (*K*) is assigned to each data point, where k = layer number.

Instead of using the mean and standard deviation of the normal distribution, the user can also use fitting-distribution software such as @Risk to find the most fitting distribution for the data in a class. **Figure 13** shows an example of finding the most fitting distribution for one of the classes.

2.4. Clustering of Unit Cost using Statistical Methods

Building the unit cost tree shows a large number of classes, which can drastically increase if more variables are

Case Summaries

Original Multiplier

Pa	Ph	Re	Mean	Std Deviation	N	Minimum	Maximum	Range
Pk-01	Ph-01	R-01	126.6667	176.09183	3	25.00	330.00	305.00
		Total	126.6667	176.09183	3	25.00	330.00	305.00
	Ph-02	R-01	14.0000	11.27436	10	4.00	40.00	36.00
		R-02	5.0000	.	1	5.00	5.00	.00
		R-04	.5000	.	1	.50	.50	.00
		R-05	15.0000	.	1	15.00	15.00	.00
		Total	12.3462	10.69537	13	40.00	40.00	39.50
	Ph-03	R-01	8.5000	3.00000	4	10.00	10.00	6.00
		R-02	2.0000	.00000	2	2.00	2.00	.00
		Total	6.3333	4.08248	6	10.00	10.00	8.00
	Total	R-01	32.5882	77.06420	17	330.00	330.00	326.00
		R-02	3.0000	1.73205	3	5.00	5.00	3.00
		R-04	.5000	.	1	.50	.50	.00
		R-05	15.0000	.	1	15.00	15.00	.00
		Total	26.2955	68.52748	22	330.00	330.00	329.50
Pk-02	Ph-01	R-01	16.6667	5.77350	3	20.00	20.00	10.00
		Total	16.6667	5.77350	3	20.00	20.00	10.00
	Ph-02	R-01	50.3039	106.83125	18	465.00	465.00	463.00
		R-02	9.7000	8.52513	10	25.00	25.00	23.00
		R-04	2.8000	1.64317	5	5.00	5.00	4.00
		R-05	9.3333	12.75408	6	33.00	33.00	32.00
		Total	27.4992	74.87835	39	465.00	465.00	464.00

Figure 6. The descriptive data test.

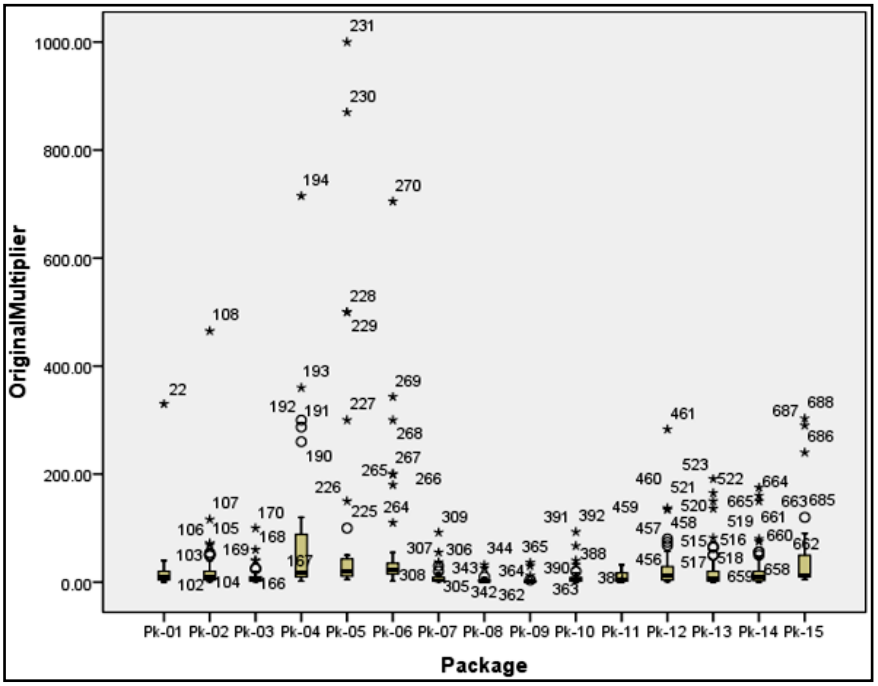

Figure 7. Boxplot of the unit cost showing outliers per package.

added to the dataset. To simplify the estimating procedure, classes that are not significantly different from each other are combined together in summary groups (clusters) with one distribution representing each cluster. The ANOVA test was implemented to the dataset to check the significance of mean differences within the seven data attributes and the results are shown in **Figure 14**.

The results for the Post Hoc tests for the three main attributes with $\alpha = 0.05$ are shown in **Figure 15**.

If the user decides to use only one attribute for dividing the dataset, test results in **Figure 15** show that pack-

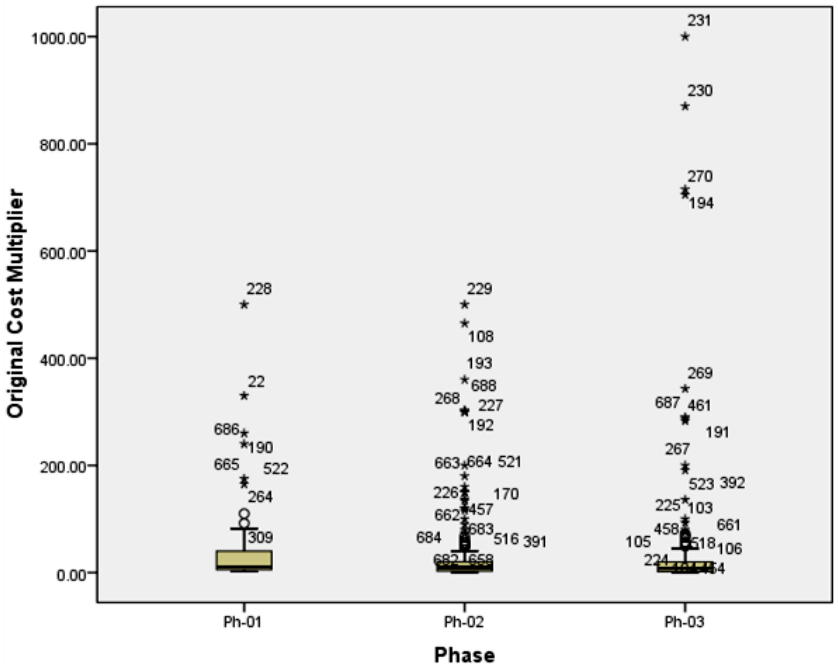

Figure 8. Boxplot of the unit cost showing outliers per phase.

Package PhaseResource	All	Package	Phase	Resource	Package / Phase	Package / Resource	Phase / Resource	Package / Phase / Resource	Score	Outlier
1381	1	1		1		1	1		5	1
1491	1	1	1	1	1	1	1	1	8	1
2014	1								1	1
2094	1		1						2	1
2094	1		1						2	1
2014	1	1	1	1		1	1		6	1
2194	1		1						2	1
2184	1		1				1		3	1
2194	1		1				1		3	1
2114	1	1	1	1		1	1		6	1
2114	1	1	1	1	1	1	1	1	8	1
2294	1		1						2	1
2214	1		1						2	1
2214	1	1	1	1	1	1	1	1	8	1
2415		1			1	1		1	4	1
2515					1			1	2	1
2615		1		1	1	1	1	1	6	1
2915	1	1		1	1	1	1	1	7	1
3015						1		1	2	1
3116					1			1	2	1
3191		1			1	1		1	4	1
3214	1								1	1
3294	1		1						2	1
	17	20	13	16	16	18	19	10		23

Figure 9. The output from the outlier detection tool.

ages can be grouped into four classes, phases can be grouped into two classes and resources can be grouped into two classes.

If the user decides to use the combination of the three main attributes (Package * Phase * Resource), **Figure 16** shows the Post Hoc test results for this combination. The test results are used to group the classes into eight clusters and a new variable $Cluster_{(1:8)}$ is assigned to each data point. The case summary and Boxplot tests were repeated and the results are shown in **Figure 17** and **Figure 18**.

Figure 19 shows the dataset in SPSS after assigning all the analysis variables. That dataset can be used for a lot more tests if more data and attributes are available. The simplicity of the analysis and the techniques used in it opens the door for the end user to continue searching for more patterns and hidden knowledge in the collected

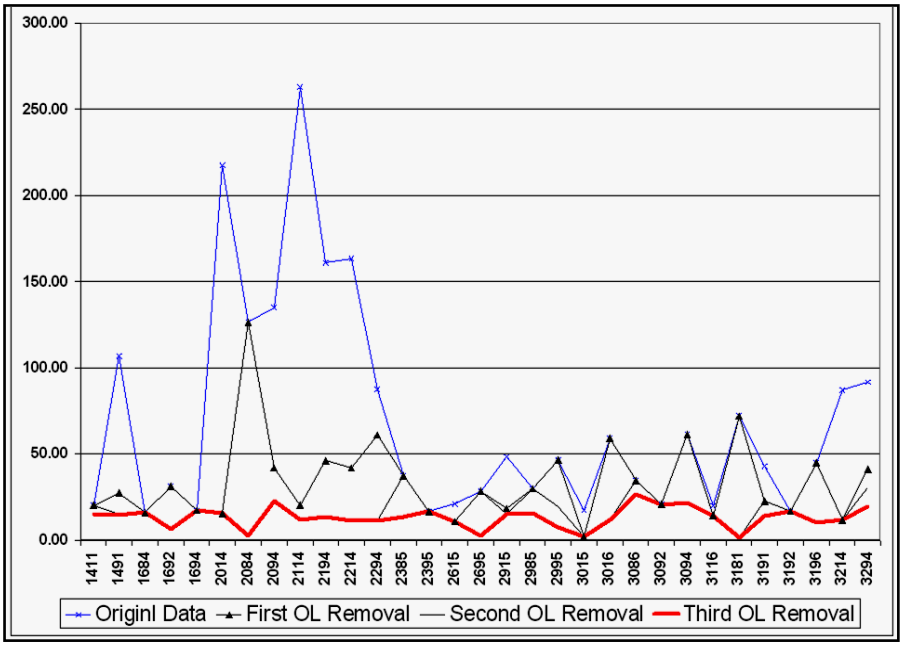

Figure 10. The decrease in standard deviations of the data classes.

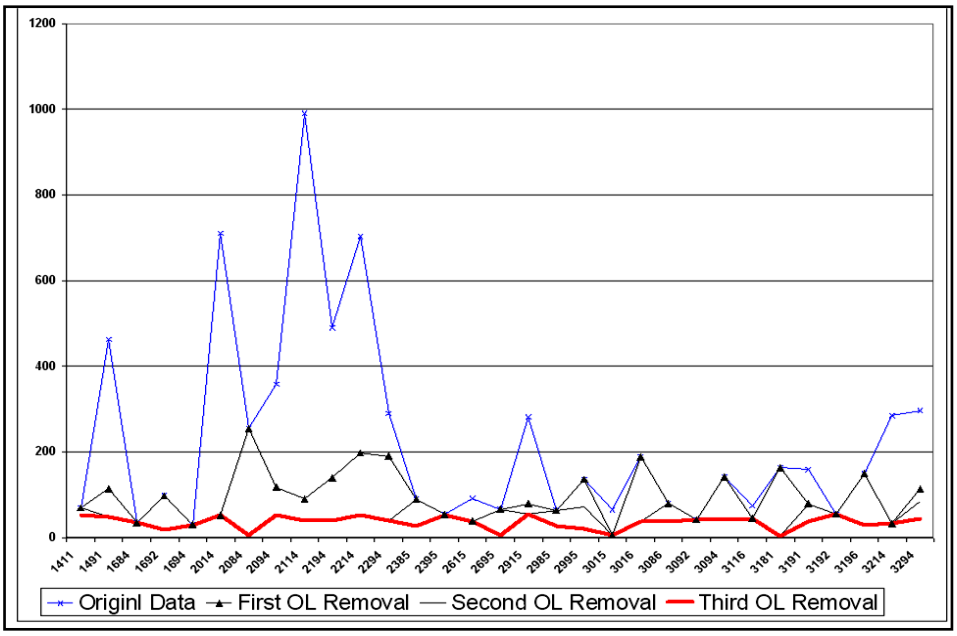

Figure 11. The decrease in ranges of the data classes.

data.

This case study presents the value of obtaining the unit costs from historical data using data mining. It shows that extracting useful knowledge from data can be maximized if all data elements are collected properly. Two major problems pertinent to the dataset were found. First, discrepancies were found among the different estimators' entries. Estimators are supposed to enter both the estimated quantity of a deliverable and the estimated amount of unit hours per quantity item. The system would then calculate the total estimated hours for a package. However, this was not the case for all data points. Some estimators did not provide estimated quantity; they only put the number T in the quantity field. This practice, hence, led to erroneous hourly unit estimation.

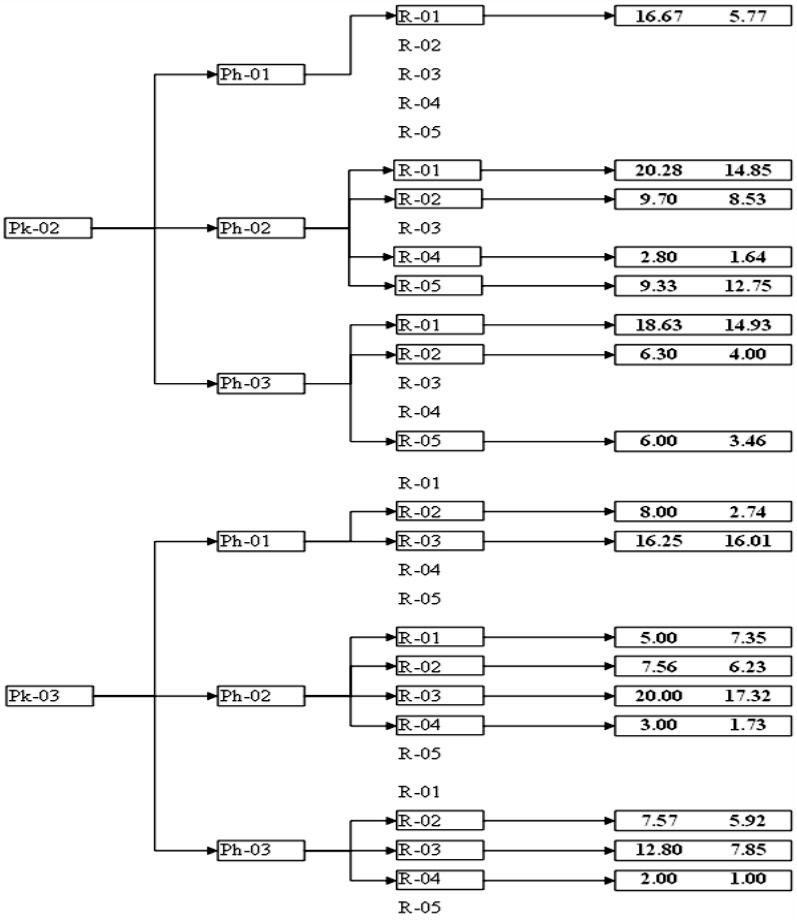

Figure 12. The output summary tree.

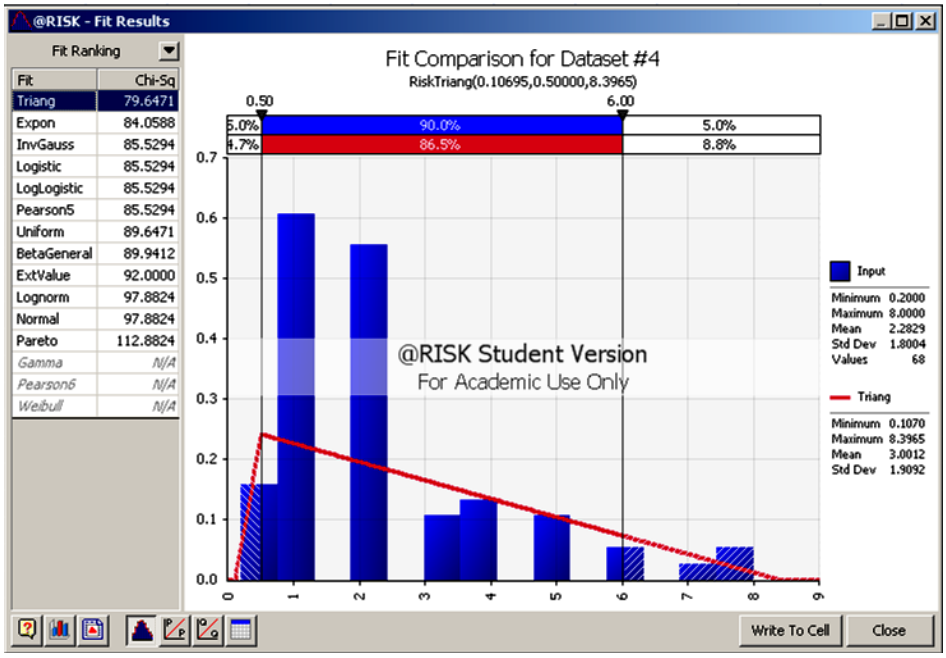

Figure 13. Fitting distribution to a class of data.

Tests of Between-Subjects Effects

Dependent Variable: OriginalMultiplier

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	683820.684[a]	79	8655.958	1.309	.045
Intercept	200551.553	1	200551.553	30.336	.000
Package	106060.040	14	7575.717	1.146	.314
Phase	9222.261	2	4611.130	.698	.498
Resource	6486.429	4	1621.607	.245	.913
Package * Phase	54561.285	24	2273.387	.344	.999
Package * Resource	14832.718	14	1059.480	.160	1.000
Phase * Resource	3390.225	7	484.318	.073	.999
Package * Phase * Resource	4903.522	13	377.194	.057	1.000
Error	4019430.959	608	6610.906		
Total	5279751.469	688			
Corrected Total	4703251.644	687			

a. R Squared = .145 (Adjusted R Squared = .034)

Figure 14. Univariate ANOVA test results for the three main attributes.

OriginalMultiplier

	Package	N	Subset 1	2	3	4
Tukey B[a]	Pk-08	35	4.2642			
	Pk-09	21	4.9286			
	Pk-07	39	9.7103			
	Pk-11	15	10.6167			
	Pk-03	62	10.7258			
	Pk-10	27	12.4575			
	Pk-14	142	16.8041			
	Pk-02	86	20.9814			
	Pk-13	62	22.3144			
	Pk-01	22	26.2955			
	Pk-12	54	27.0093			
	Pk-15	23	59.5676	59.5676		
	Pk-06	39	69.6022	69.6022		
	Pk-04	24		99.1034		
	Pk-05	37		109.4454		
Duncan[a]	Pk-08	35	4.2642			
	Pk-09	21	4.9286			
	Pk-07	39	9.7103			
	Pk-11	15	10.6167			
	Pk-03	62	10.7258			
	Pk-10	27	12.4575			
	Pk-14	142	16.8041	16.8041		
	Pk-02	86	20.9814	20.9814		
	Pk-13	62	22.3144	22.3144		
	Pk-01	22	26.2955	26.2955		
	Pk-12	54	27.0093	27.0093		
	Pk-15	23		59.5676	59.5676	
	Pk-06	39			69.6022	69.6022
	Pk-04	24			99.1034	99.1034
	Pk-05	37				109.4454
	Sig.		.361	.062	.063	.061

	Phase	N	Subset 1	2
Tukey B[a]	Ph-03	396	25.0976	
	Ph-02	240	30.6041	30.6041
	Ph-01	52		50.6154
Duncan[a]	Ph-03	396	25.0976	
	Ph-02	240	30.6041	30.6041
	Ph-01	52		50.6154
	Sig.		.607	.062

	Resource	N	Subset 1	2
Tukey B[a]	R-04	253	12.0119	
	R-02	91	12.2527	
	R-05	87	18.1609	
	R-01	112	26.3850	
	R-03	145		77.4241
Duncan[a]	R-04	253	12.0119	
	R-02	91	12.2527	
	R-05	87	18.1609	
	R-01	112	26.3850	
	R-03	145		77.4241
	Sig.		.220	1.000

Figure 15. Post hoc test results for the three main attributes.

Package Phase Resource	N	Subset 1	2	3	4	5	6	7	8	9	10	Group
2515	13	1.5769										1
2415	21	1.8213										1
1615	3	2.0000										1
3011	4	2.3050										1
1495	5	2.8000										1
3015	12	2.9392										1
1695	3	3.0000	3.0000									1
3095	3	3.1667	3.1667									1
2695	4	3.3750	3.3750									1
2315	21	4.2238	4.2238									2
1691	4	5.0000	5.0000									2
3195	13	5.0385	5.0385									2
2495	11	5.7273	5.7273									2
1416	7		6.0000	6.0000								2
3096	7		6.2857	6.2857								2
1412	10		6.3000	6.3000								2
2595	6		6.8333	6.8333								2
2084	3		7.0000	7.0000								3
1692	9		7.5556	7.5556								3
1612	21		7.5714	7.5714								3
2615	21		7.7549	7.7549								3
1682	5			8.0000	8.0000							3
1311	4			8.5000	8.5000							3
2995	12			8.7917	8.7917							3
3012	6			8.8333	8.8333							3
1496	6			9.3333	9.3333							3
1492	10			9.7000	9.7000							3
3111	19			10.9211	10.9211							4
2716	12			11.3542	11.3542							4
2395	13			11.3846	11.3846							4
3196	9			11.3889	11.3889							4
3116	19			11.5395	11.5395							4
3115	34			11.7479	11.7479							4
3016	11			11.8182	11.8182							4
2385	4				12.5000	12.5000						4
3191	13				12.7692	12.7692						4
1614	5				12.8000	12.8000						4
1391	10				14.0000	14.0000	14.0000					5
2985	3				14.0000	14.0000	14.0000					5
3181	4				14.0000	14.0000	14.0000					5
3112	10				15.1000	15.1000	15.1000					5
2915	33					16.2121	16.2121					5
1684	4					16.2500	16.2500					5
1481	3					16.6667	16.6667					5
3214	9					17.0556	17.0556					6
2194	8					17.0953	17.0953					6
1411	23					18.6273	18.6273	18.6273	18.6273	18.6273		6
3092	4					19.0000	19.0000	19.0000	19.0000	19.0000		6
2294	13					19.1606	19.1606	19.1606	19.1606	19.1606		6
2014	9						19.3194	19.3194	19.3194	19.3194		6
1694	3						20.0000	20.0000	20.0000	20.0000		7
1491	16						20.2794	20.2794	20.2794	20.2794		7
3086	2						21.0000			21.0000	21.0000	7
3192	10									21.9000	21.9000	7
3294	7									21.9364	21.9364	7
2214	18									23.4109	23.4109	7
2094	6									23.6009	23.6009	7
2114	20									23.9359	23.9359	7
3094	3										32.6667	8
Sig.		.055	.052	.051	.054	.051	.051	.054	.052	.052	.056	

Figure 16. Duncan test results.

Original Cost Multiplier

Duncan Group	N	Mean	Median	Std. Deviation	Minimum	Maximum	Range	Skewness	Kurtosis
G-01	68	2.28	2.00	1.81	0.20	8.00	7.80	1.56	2.00
G-02	79	5.41	3.00	6.08	0.50	36.00	35.50	2.50	8.53
G-03	97	8.21	6.00	7.75	1.00	40.00	39.00	1.74	3.44
G-04	139	11.68	7.00	11.91	0.25	55.00	54.75	1.28	0.98
G-05	67	15.51	12.00	12.64	1.00	57.00	56.00	1.17	1.10
G-06	66	18.45	14.01	13.64	2.00	56.88	54.88	1.41	1.16
G-07	82	22.45	20.00	14.53	1.00	56.00	55.00	0.60	-0.42
G-08	3	32.67	40.00	21.94	8.00	50.00	42.00	-1.34	0.00
Total	601	11.98	8.00	12.53	0.20	57.00	56.80	1.51	1.83

Figure 17. Statistical analysis for the eight data clusters.

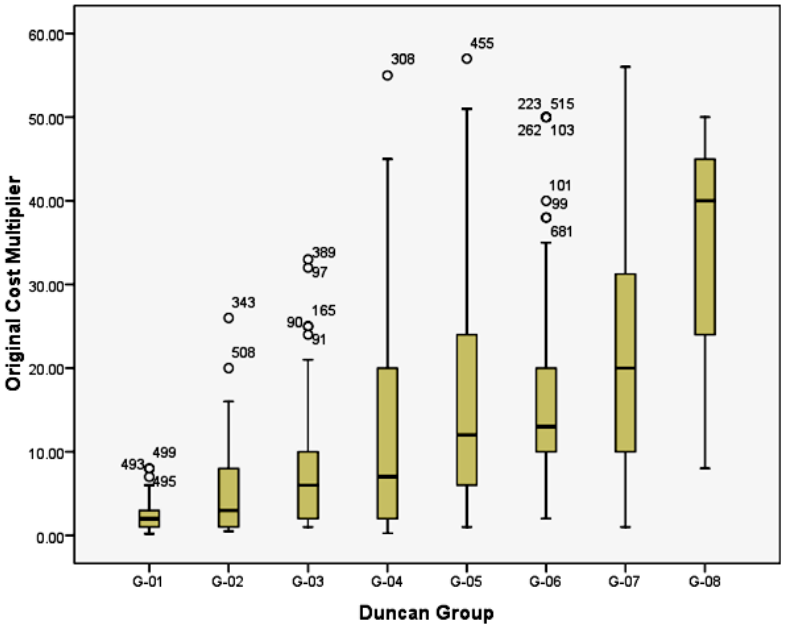

Figure 18. Box Plots for the eight data clusters.

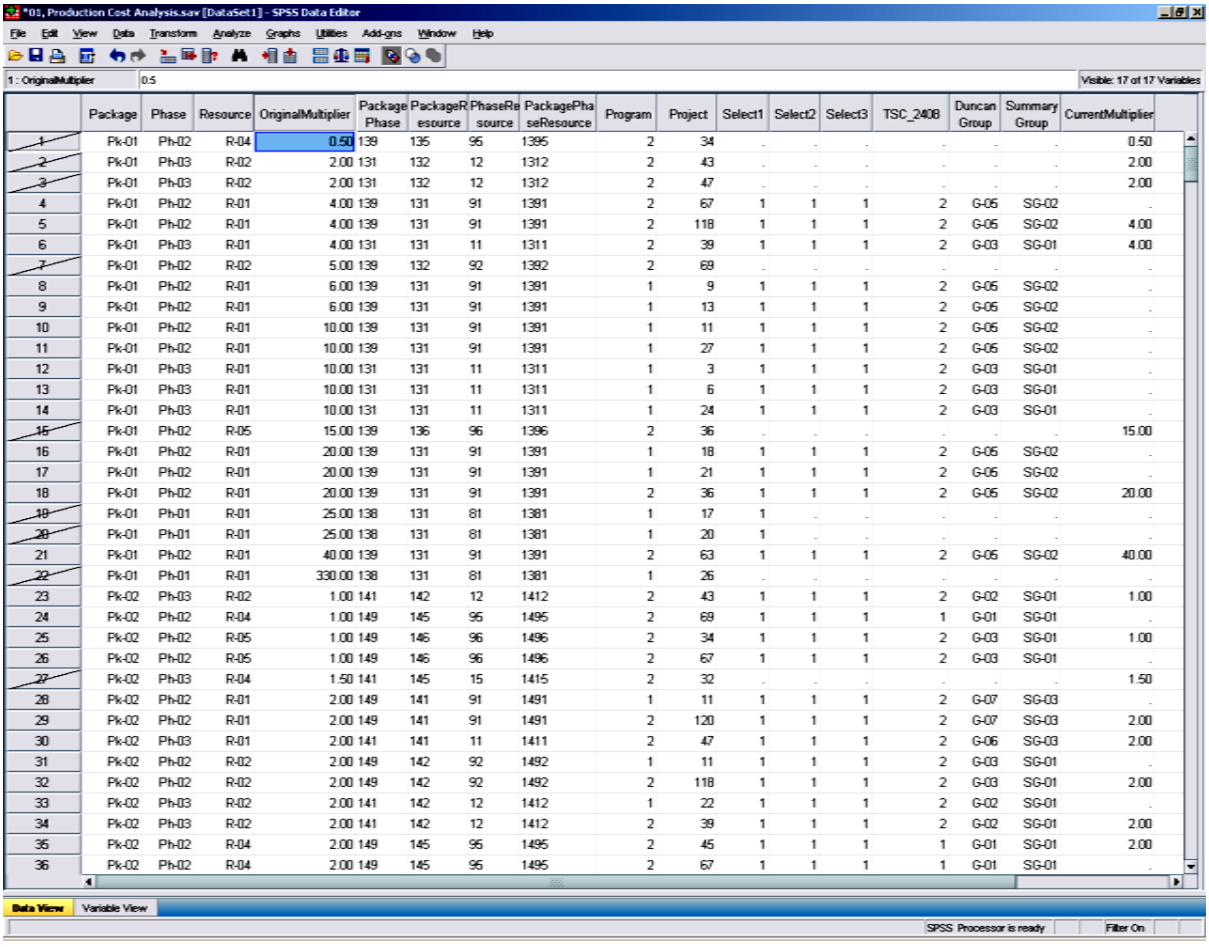

Figure 19. The final dataset with the select and cluster variables.

Another source of discrepancy was found in the estimating of hours required to complete work packages. Some estimators included all the support activities, such as meetings, site visits and quality inspections, in their production package estimates. Others estimated the requirement for the support activities independently from the production packages. Again, this led to erroneous hourly unit estimates of production packages.

In addition to the discrepancies found in the estimating entries, discrepancies were found in recording actual entries. The actual hours spent were collected at the project level, as opposed to the planning hours that were estimated at the work package level. Given the levels where the data was collected, there was no possibility to compare or analyze the variance between the estimated and the actual hours spent. Similar to the estimated dataset, the actual dataset should have been collected at the work package level.

These discrepancies caused inconsistencies in the data. When the dataset was analyzed, large amount of outliers caused significant disparity in the results. These outliers were highlighted using the outlier detection tool developed in this research and were presented to the data owner for corrective action. Two recommendations were made to the company about these issues, and were approved to be implemented: first, to issue estimating guidelines to ensure consistency among different estimators; second, to modify the timekeeping system in a way to collect actual hours spent at the work package level.

3. Discovering Knowledge in the Second Dataset

3.1. Data Gathering, Cleaning and Preprocessing

The purpose of this case study is to validate the concept that mining historical data enables contractors to better estimate the duration of their work packages. Current practices rely mostly on estimating the duration by dividing the total work hours by the daily number of hours or the scheduler experience. Both practices struggle to provide reliable estimates of package durations that utilize prior experience and current project conditions.

The second dataset used in this research contains actual duration and working hours for a large group of fabrication work packages. This dataset included 13,498 data points and was obtained from the second partner company. This company is a large EPC firm that specializes in fabricating structural steel for industrial construction projects. The data was obtained from the scheduling information system of this company, which is a SQL-Server database that was originally designed by the author and developed by the NSERC Industrial Research Chair in Construction Engineering and Management. The data was automatically extracted out of the SQL-Server data tables to MS Excel for cleaning and preprocessing.

The researcher helped the contractor to develop a predefined set of progress activities for their fabrication packages. The start and finish date for each one of these progress activities were collected over a long period of time. The actual steel weight and working hours to complete each fabrication package were also stored in the information system. The steel weight represents the key quantity for each of these work packages. However, the production package (work package type) was not assigned to the obtained dataset.

The cleaning procedure started by selecting the data point that represents the completed work packages, which means start-date and end-date were marked actual. After that, the obvious data entry errors, such as negative values, were also eliminated.

The data for handrails and miscellaneous very small fabrication packages were eliminated as well, because they are handled by a separate facility, and are not in the scope of this data-mining exercise. After the cleaning procedure, a large dataset with 5590 data points was still available to analyze. The duration $(D)_{(n)}$ in work weeks was calculated using the formula:

$$D_{(n)} = \text{NETWORKDAYS}(\text{Finish Date}, \text{Start Date})/5 \quad (5)$$

Figure 20 shows the data from **Figure 21** after it was cleaned, pre-processed and was ready for storage in the data warehouse.

3.2. Clustering of the Cost and Duration Units

This second dataset contained more than five thousand work packages for two standard phases: shop drawings and fabrication. The actual quantities of deliverables, hours and weeks spent on each package was recorded. The fabrication and shop drawings hourly unit cost and weekly unit duration are calculated for every work package in the dataset. This data was collected over a long period of time. This data had not been analyzed or used before

proj_id	job_id	drv_id	sub_id	description	det_mhrs	fab_mhrs	weight	fab_start_date	f_actual	fab_end_date	f_actual	ship_start_date	f_actual	ship_end_date	f_actual
1	512	1059	1059	25E-0054 Support Structure	2.50	20.00	10.00	07-Aug-01	TRUE	28-Aug-01	TRUE	28-Aug-01	TRUE	28-Aug-01	TRUE
1	512	1101	1101	Plant 25 Temp Pipe supports	2.50	20.00	1.00	07-Aug-01	TRUE	28-Aug-01	TRUE	28-Aug-01	TRUE	28-Aug-01	TRUE
1	514	1115	1115	Dow-Blk 250 T396 Platform	12.00	22.00	1.50	27-Aug-01	TRUE	07-Sep-01	TRUE	07-Sep-01	TRUE	07-Sep-01	TRUE
1	518	1184	1184	Husky Oil	2.00	16.00	26.00	09-Oct-01	TRUE	15-Oct-01	TRUE	15-Oct-01	TRUE	29-Oct-01	TRUE
1	518	1185	1185	Husky Oil	2.00	16.00	18.00	09-Oct-01	TRUE	23-Oct-01	TRUE	23-Oct-01	TRUE	29-Oct-01	TRUE
1	518	1186	1186	Husky Oil	2.00	16.00	9.00	21-Oct-01	TRUE	02-Nov-01	TRUE	02-Nov-01	TRUE	08-Nov-01	TRUE
1	521	1197	1197	ROM Bldg Hopper	5.00	65.00	36.80	09-Nov-01	TRUE	21-Dec-01	TRUE	23-Feb-02	TRUE	23-Feb-02	TRUE
1	521	1198	1198	Course Products Bin	5.00	65.00	20.90	10-Dec-01	TRUE	12-Feb-02	TRUE	12-Feb-02	TRUE	19-Feb-02	TRUE
1	521	1199	1199	Rejects Bin	5.00	65.00	10.60	10-Dec-01	TRUE	08-Feb-02	TRUE	12-Feb-02	TRUE	12-Feb-02	TRUE
1	532	1269	1269	Suncor Breaker Beam	1.00	62.00	4.20	09-Oct-01	TRUE	24-Oct-01	TRUE	24-Oct-01	TRUE	29-Oct-01	TRUE
1	535	1338	1338	Jib Beams	0.00	12.00	3.00	16-Oct-01	TRUE	20-Oct-01	TRUE	20-Oct-01	TRUE	20-Oct-01	TRUE
1	536	1337	1337	Dow Chemical	0.00	54.00	2.70	15-Oct-01	TRUE	29-Oct-01	TRUE	29-Oct-01	TRUE	31-Oct-01	TRUE
1	537	1336	1336	Dow Chemical	0.00	30.00	4.00	15-Oct-01	TRUE	29-Oct-01	TRUE	29-Oct-01	TRUE	31-Oct-01	TRUE
1	538	1361	1361	Gunther Construction - Exterior	8.00	36.00	1.10	31-Oct-01	TRUE	07-Nov-01	TRUE	07-Nov-01	TRUE	14-Nov-01	TRUE
1	538	1405	1405	Gunther Construction - Exterior	8.00	36.00	0.20	05-Nov-01	TRUE	07-Nov-01	TRUE	07-Nov-01	TRUE	14-Nov-01	TRUE
1	539	1357	1357	Slave lake Pulp	0.00	0.00	0.80	31-Oct-01	TRUE	06-Nov-01	TRUE	06-Nov-01	TRUE	20-Nov-01	TRUE
1	539	1358	1358	Slave lake Pulp	0.00	0.00	34.00	31-Oct-01	TRUE	06-Nov-01	TRUE	06-Nov-01	TRUE	20-Nov-01	TRUE
1	540	1359	1359	Slavelake Pulp Flash Dryer	0.00	33.00	0.80	27-Nov-01	TRUE	05-Dec-01	TRUE	05-Dec-01	TRUE	12-Dec-01	TRUE
1	541	1360	1360	Slavelake Pulp Stair Landing	0.00	36.00	0.80	11-Jan-02	TRUE	16-Jan-02	TRUE	21-Jan-02	TRUE	21-Jan-02	TRUE
1	542	1363	1363	Blanchett Neon Limited	0.00	0.00	0.75	24-Oct-01	TRUE	30-Oct-01	TRUE	30-Oct-01	TRUE	30-Oct-01	TRUE
1	544	1364	1364	(5) 2000mm Dp Plate Girders	1.50	8.00	302.70	14-Jan-02	TRUE	12-Mar-02	TRUE	12-Mar-02	TRUE	14-Mar-02	TRUE
1	544	1365	1365	Floor Beams c/w Stiffners	1.50	24.00	171.50	25-Jan-02	TRUE	20-Mar-02	TRUE	20-Mar-02	TRUE	20-Mar-02	TRUE
1	544	1366	1366	(2) End Wall Assemblies	1.50	40.00	59.30	24-Jan-02	TRUE	20-Feb-02	TRUE	15-Mar-02	TRUE	22-Mar-02	TRUE
1	544	1367	1367	Bridge Rails /Grating (WT of	1.50	22.00	15.40	25-Feb-02	TRUE	08-Mar-02	TRUE	15-Mar-02	TRUE	15-Mar-02	TRUE
1	545	1371	1371	Dow Chemical	0.00	33.00	3.00	29-Oct-01	TRUE	05-Nov-01	TRUE	05-Nov-01	TRUE	05-Nov-01	TRUE
1	546	1406	1406	Blanchett Neon	0.00	5.00	0.59	29-Oct-01	TRUE	31-Oct-01	TRUE	31-Oct-01	TRUE	31-Oct-01	TRUE
1	547	1407	1407	6 x 3-1/2 angles	0.00	4.00	4.40	29-Oct-01	TRUE	05-Nov-01	TRUE	05-Nov-01	TRUE	05-Nov-01	TRUE
1	547	7561	7701		0.00	0.00	0.00		FALSE		FALSE		FALSE		FALSE
1	548	1408	1408	2 x 2 angles	0.00	1.50	2.98		FALSE	31-Oct-01	TRUE	31-Oct-01	TRUE	05-Nov-01	TRUE
1	549	1427	1427	Dow Site Outfall Upgrade	10.40	28.60	1.20	19-Nov-01	TRUE	04-Dec-01	TRUE	04-Dec-01	TRUE	04-Dec-01	TRUE
1	549	1436	1436	Dow Site Outfall Upgrade	10.40	28.60	2.50	07-Jan-02	TRUE	14-Jan-02	TRUE	14-Jan-02	TRUE	15-Jan-02	TRUE
1	550	1428	1428	10 Channels for MRC	0.00	15.00	1.30	13-Nov-01	TRUE	15-Nov-01	TRUE	15-Nov-01	TRUE	15-Nov-01	TRUE
1	550	1433	1433	Deaerator Pipeway 240 FP-2-	0.00	20.00	1.30	28-Nov-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1434	1434	Platform for Turbidity Probes	0.00	30.00	0.70	03-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1435	1435	Tailings Pumphouse S/S Utility	0.00	15.00	6.50	03-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1450	1450	Utility Water Heater Area FP-	0.00	15.00	0.20	28-Nov-01	TRUE	30-Nov-01	TRUE	30-Nov-01	TRUE	30-Nov-01	TRUE
1	550	1470	1470	Impact Cushions FP-2C-5370	0.00	34.00	0.90	07-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE	14-Dec-01	TRUE
1	550	1471	1471	HVAC Supports FP-2-5371	0.00	12.00	4.80	12-Dec-01	TRUE	18-Dec-01	TRUE	18-Dec-01	TRUE	18-Dec-01	TRUE
1	550	1472	1472	TSRU Train Pipe Supports FP-2-	0.00	18.00	24.00	07-Jan-02	TRUE	14-Jan-02	TRUE	14-Jan-02	TRUE	14-Jan-02	TRUE
1	550	1510	1510	Tailings Pumphouse platforms	0.00	15.00	2.20	27-Dec-01	TRUE	16-Jan-02	TRUE	18-Jan-02	TRUE	18-Jan-02	TRUE
1	551	1431	1431	Skid and Loading Frames	0.00	110.00	4.40	19-Dec-01	TRUE	04-Feb-02	TRUE	31-Jan-02	TRUE	08-Feb-02	TRUE
1	552	1442	1442	Tower Crane Base For Potain	0.00	15.00	0.50	16-Nov-01	TRUE	26-Nov-01	TRUE	26-Nov-01	TRUE	27-Nov-01	TRUE
1	553	1443	1443	(2) HSS Cols	0.00	15.00	0.63	26-Nov-01	TRUE	27-Nov-01	TRUE	27-Nov-01	TRUE	27-Nov-01	TRUE
1	554	1448	1448	33 Low W-Beam Brackets	0.00	57.00	0.89	03-Dec-01	TRUE	07-Dec-01	TRUE	12-Dec-01	TRUE	08-Dec-01	TRUE
1	555	1451	1451	Atco - Ruth Lake power station	0.00	14.00	9.80	29-Nov-01	TRUE	10-Dec-01	TRUE	07-Dec-01	TRUE	11-Dec-01	TRUE
1	556	1462	1462	Rotolift Project	17.00	35.00	4.00	18-Dec-01	TRUE	03-Jan-02	TRUE	07-Jan-02	TRUE	07-Jan-02	TRUE

Figure 20. Raw dataset for the second analysis.

for data mining or knowledge discovery.

The purpose of the analysis of this data was to use historical data to develop realistic, reliable and more accurate estimating units for both resource requirement and expected duration. These estimating units were then multiplied by the known quantities to estimate the total duration and resource requirement of a work package.

Since this data is based on actual values, the dataset has been used to validate the developed estimating methodology in this research. The dataset was divided into two parts. The first part, consisting of 85% of the data points, was selected randomly and used for calculating the estimating units. The second part, the remaining %15 of the data points, was used for testing purposes.

The software selected to perform the analysis is called *Weka* (Waikato Environment for Knowledge Analysis), which is a powerful and user friendly data mining and machine learning tool. *Weka* was developed at the University of Waikato in Hamilton, New Zealand [16]. The software was selected because of its powerful data mining capabilities. The software is also easy to obtain, and doesn't require any special hardware; therefore, it would be accessible to any contractor seeking to perform data mining without incurring major cost. Minimizing the cost of implementing data mining in industrial construction makes it more appealing to decision makers and also maximizes the return on investment of the increased efficiency.

Weka is able to read data from different types of data files. The first 85% of the dataset was exported from the data warehouse to a Comma Separated Values (CSV) file. Then, it was transferred to *Weka* in order to perform the analysis. The data contained a unique ID for each data point, two control variables: program and project, the actual amount of key quantity, and total hours and weeks for two resources. One resource is utilized during the fabrication phase and the other one is utilized during the shop drawings phase. The unit cost was calculated by dividing total hours by the key quantity. The unit duration was calculated by dividing the total number of weeks by the key quantity. An excerpt of the CSV data file for the fabrication resource is shown in **Figure 22**.

Unlike the first dataset where several resources in multiple phases with different package type were analyzed

ID	Program	Project	Package	Type	Weight	FabHours	FabCostMultiplier	ShopDuration	PaintDuration	FPDuration	ShipDuration	FabDuration	FabDurationMultiplier
640	1	1569	8640	2	1.10	3.60	3.27	1.40	0.20		0.20	1.80	1.64
680	1	1597	7932	1	8.87	30.00	3.38	1.40			0.20	1.60	0.18
692	1	1607	8003	1	1.67	11.00	6.59	1.20			0.40	1.60	0.96
695	1	1609	8022	2	21.86	48.00	2.20	6.00	0.80		0.20	7.00	0.32
696	1	1609	8023	2	26.08	48.00	1.84	5.00	0.60		0.20	5.80	0.22
697	1	1609	8024	2	35.84	48.00	1.34	4.40	0.60		0.20	5.20	0.15
698	1	1609	8025	2	7.45	48.00	6.44	4.80	0.40		0.20	5.40	0.72
703	1	1611	8035	2	1.44	52.80	36.74	1.00	0.60		0.20	1.80	1.25
715	1	1626	8137	2	13.80	33.00	2.39	5.40	1.20		1.20	7.80	0.57
732	1	1632	8197	2	4.20	38.00	9.05	1.40	0.40		0.20	2.00	0.48
739	1	1640	8239	1	1.49	20.82	13.97	1.00			0.20	1.20	0.81
741	1	1640	8239	1	11.97	20.80	1.74	2.60			0.20	2.80	0.23
742	1	1640	8240	1	35.30	16.27	0.46	2.00			0.20	2.20	0.06
746	1	1640	8243	1	5.31	36.40	6.85	1.40			0.60	2.00	0.38
747	1	1640	8244	1	6.20	53.22	8.58	1.60			0.80	2.40	0.39
749	1	1640	8245	1	4.80	35.32	7.36	2.20			0.20	2.40	0.50
754	1	1641	8253	2	1.50	29.60	19.73	1.40	0.60		0.20	2.20	1.47
756	1	1642	8252	1	35.37	18.00	0.51	6.20			0.20	6.40	0.18
776	1	1679	8476	2	2.00	14.83	7.42	7.80	1.40		0.20	9.40	4.70
777	1	1679	8476	1	1.30	14.80	11.38	1.20			1.80	3.00	2.31
778	1	1679	8477	2	18.00	14.80	0.82	5.20	1.40		0.20	6.80	0.38
779	1	1679	8478	2	24.10	14.80	0.61	4.80	1.40		0.40	6.60	0.27
780	1	1679	8479	2	15.70	14.80	0.94	7.40	1.40		0.40	9.20	0.59
781	1	1679	8482	2	2.46	14.80	6.02	5.80	1.40		0.40	7.60	3.09
790	1	1683	8530	4	63.33	18.00	0.28	2.00	0.20	0.20	1.20	3.60	0.06
791	1	1683	8531	4	69.00	14.68	0.21	2.80	0.20	0.20	4.20	7.40	0.11
793	1	1683	8532	4	32.04	20.40	0.64	4.00	0.20	0.20	0.20	4.60	0.14
794	1	1683	8533	4	50.74	20.00	0.39	3.00	0.20	0.20	1.00	4.40	0.09
795	1	1683	8534	4	15.00	21.22	1.41	3.80	0.20	0.20	4.00	8.20	0.55
797	1	1683	8535	4	9.30	57.15	6.15	5.20	0.20	0.20	0.20	5.80	0.62
822	1	1691	8551	1	15.60	28.80	1.85	1.60			5.00	6.60	0.42
829	1	1696	8571	1	14.38	24.00	1.67	8.40			0.20	8.60	0.60
831	1	1696	8714	1	5.13	24.00	4.68	8.40			0.20	8.60	1.68
840	1	1699	8591	1	1.50	20.66	13.77	3.20			0.20	3.40	2.27
863	1	1729	8729	1	1.45	47.05	32.45	1.40			0.20	1.60	1.10
875	1	1752	8787	2	57.78	20.00	0.35	3.00	1.20		0.20	4.40	0.08

Figure 21. Calculating the total fabrication duration.

ID	Program	Project	Package	Weight	FabHours	FabHourePerUnit	FabWeeks	FabWeeksPerUnit
1	1	512	1059	10	20	2	3.2	0.32
3	1	514	1115	1.5	22	14.67	2	1.33
4	1	518	1184	26	16	0.62	1	0.04
5	1	518	1185	18	16	0.89	2.2	0.12
6	1	518	1186	9	16	1.78	2	0.22
7	1	521	1197	36.8	65	1.77	6.2	0.17
8	1	521	1198	20.9	65	3.11	9.4	0.45
9	1	521	1199	10.6	65	6.13	9	0.85
10	1	532	1269	4.2	62	14.76	2.4	0.57
12	1	536	1337	2.7	54	20	2.2	0.81
13	1	537	1336	4	30	7.5	2.2	0.55
14	1	538	1361	1.1	36	32.73	1.2	1.09
22	1	544	1365	171.5	24	0.14	7.8	0.05
27	1	547	1407	4.4	4	0.91	1.2	0.27
30	1	549	1427	1.2	28.6	23.83	2.4	2
33	1	550	1433	1.3	20	15.38	2.6	2
35	1	550	1435	6.5	15	2.31	2	0.31
38	1	550	1471	4.8	12	2.5	1	0.21
39	1	550	1472	24	18	0.75	1.2	0.05
40	1	550	1510	2.2	15	6.82	3	1.36
41	1	551	1431	4.4	110	25	6.8	1.55
46	1	556	1462	4	35	8.75	2.6	0.65
47	1	557	1467	2.7	15	5.56	1.6	0.59
55	1	565	1507	1.5	15	10	2.2	1.47
56	1	566	1506	3.75	15	4	2.2	0.59
63	1	573	1548	5.32	51	9.59	2	0.38
65	1	574	1551	11.7	51	4.36	9.4	0.8
70	1	579	1582	18.9	16	0.85	1.6	0.08
71	1	579	1646	24	16	0.67	2	0.08

Figure 22. An excerpt of the CSV data file for the fabrication phase.

simultaneously, for this dataset, the analysis is done on one single resource per phase. Since there is no data collected regarding production package type, the data was analyzed with the assumption that it is all under one production package type. For this analysis, clustering, which is an unsupervised learning technique, was selected. Among the several clustering techniques available in *Weka* that were tested, the Expectation Maximization (EM) technique was found to be the most efficient one. The software developers highly recommend this technique for clustering large sets of data and it is the default technique to be used.

The EM clustering technique is applied to the dataset and the results are summarized in **Figures 23-26**. In order to ensure the stability of the clustering results, each clustering analysis was repeated three times, with each run taking about two and half hours of processing time on an Intel Pentium(R) personal computer. The results were as follows: nineteen clusters were obtained for the fabrication hourly unit cost (**Figure 23**), thirteen clusters for the fabrication weekly unit duration (**Figure 24**), five clusters for the shop drawings hourly unit cost (**Figure 25**) and six clusters for the shop drawings weekly unit duration (**Figure 26**). For each cluster, the number of data points, mean, standard deviation, and prior probability are obtained from *Weka*.

Initial results of the clustering exercise demonstrate trends that would benefit the contractor. Clusters with a large number of data points are expected to represent common cases of packages in the contracting company, while clusters with a small number of data points represent either rare types of work packages or outliers that have to be further investigated.

For instance, results in **Figure 23** show that almost a quarter of the work packages fall in cluster 13, with a mean of 0.6 hours per unit. In the same table, packages in cluster 7 represent a case of outliers that should be investigated. When a contractor needs to investigate the clustering analysis results, they can easily find out which data point belongs to which cluster, since *Weka* assigns the results of the clustering to every data point in the dataset and automatically draws the frequency histograms as shown in **Figures 27** and **28**. Assigning clusters to every data point makes it easy for contractors to go back to their files and find out the reasons behind the variation in actual package cost and durations.

The *Weka* analysis supports the claim of this research model that data, which up to now was not used, can be transferred into useful knowledge that ultimately provides meaningful insights into the work of contractors. When data is collected, stored and pre-processed in a proper way, as proposed in this research, an endless wealth of knowledge can be harvested from this data. After assigning the clusters, a fitting distribution can be found for each cluster.

3.3. Case Study Results Validation

The second part of the data, the remaining 15% was used for validation, as mentioned earlier. The obtained unit costs and durations from the clustering analysis were used to estimate the resource requirement and duration of each work package in the validation dataset. Each package was assigned a duration unit cluster and a cost unit cluster (**Figure 29**). The means of these two clusters were used to estimate the total resource requirement and duration for each package.

Both the cost and duration variances, accompanied with error percentages, were calculated for each package as well.

The validation test showed that, when comparing the estimated values using the obtained unit based on historical data with the actual values that were recorded for these packages, more than 80% of the tested data points had an estimating error of below 25%. These results demonstrate a significant increase in the accuracy of estimating practices when relying on historical data that existed already in the contractor's management systems.

The work package types were not identified when the data was recorded. When the data mining analysis was conducted, data clusters were identified. Consequently, it was left to the estimator to decide which cluster to use for estimating future projects. The partner company did not record its planned data in a structured way as it did with the actual data. Thus, performance evaluation using EVM was not possible.

4. Discovering Knowledge in the Third Dataset

4.1. Data Gathering, Cleaning and Preprocessing

The purpose of this case study was to validate the concept that data mining can be used to provide reliable probabilistic resource utilization graphs (resource baseline histograms) that can be used for proper staffing of

Fabrication Hours per Unit					
Cluster	N	%	Mean	StdDev	Prior Probability
0	233	4.94%	3.92	0.5080	0.0520
1	39	0.83%	46.12	13.2537	0.0123
2	73	1.55%	28.58	3.9670	0.0159
3	265	5.62%	2.41	0.2544	0.0562
4	142	3.01%	8.81	0.9488	0.0305
5	77	1.63%	19.48	1.7999	0.0170
6	227	4.81%	3.08	0.2724	0.0433
7	2	0.04%	112.09	2.0850	0.0004
8	30	0.64%	37.96	1.2437	0.0044
9	62	1.31%	13.23	0.8030	0.0129
10	32	0.68%	23.05	0.6121	0.0047
11	376	7.97%	1.83	0.2362	0.0814
12	95	2.01%	15.53	0.9313	0.0190
13	1151	24.39%	0.64	0.2078	0.2432
14	171	3.62%	6.81	0.7811	0.0362
15	593	12.57%	0.26	0.1254	0.1210
16	123	2.61%	11.12	0.9338	0.0274
17	212	4.49%	5.27	0.7573	0.0468
18	816	17.29%	1.17	0.2478	0.1753
Total	4,719.00	100.00%	4.22	7.8950	1.00

Log likelihood: -2.17993

Figure 23. Clustering of fabrication hourly unit cost.

Fabrication Weeks per Unit					
Cluster	N	%	Mean	StdDev	Prior Probability
0	184	3.90%	1.12	0.1774	0.0376
1	396	8.39%	0.33	0.0589	0.0933
2	38	0.81%	3.06	1.0114	0.0113
3	275	5.83%	0.65	0.1006	0.0600
4	4	0.08%	5.55	2.5024	0.0015
5	143	3.03%	1.61	0.2966	0.0287
6	754	15.98%	0.14	0.0305	0.1411
7	70	1.48%	2.19	0.4809	0.0184
8	202	4.28%	0.88	0.1452	0.0441
9	774	16.40%	0.04	0.0165	0.1409
10	436	9.24%	0.47	0.0754	0.0862
11	622	13.18%	0.22	0.0458	0.1407
12	821	17.40%	0.09	0.0252	0.1962
Total	4,719.00	100.00%	0.40	0.5820	1.00

Log likelihood: 0.05683

Figure 24. Clustering of fabrication weekly unit duration.

projects. These graphs show the required weekly hours of a specific resource within the duration of a project or work package. Data mining provides a set of various graphs based on different combinations of control attributes; hence, it provides contractors with the ability to utilize the most suitable graph. The current practices mostly rely on using uniform or predefined distribution graphs that do not rely on historical data and are not customized to reflect current conditions.

The third dataset to be used in this research was obtained from the same partner company that provided the first dataset. This third dataset contains the actual weekly hours for a set of resources per project phase. The current practice in the company is to collect actual hours by project phase instead of work packages. Although this data was not collected at the work package level as proposed in this research, this data is still very useful for providing analysis on the project level for providing Initial Planned Values (IPV) of project resource require-

Shop Drawings Hours per Unit					
Cluster	**N**	**%**	**Mean**	**StdDev**	**Prior Probability**
0	17	0.69%	10.07	8.3876	0.0087
1	360	14.64%	0.82	0.3455	0.1615
2	641	26.07%	0.32	0.1335	0.2887
3	186	7.56%	2.39	1.1511	0.0888
4	1,255	51.04%	0.12	0.0595	0.4524
Total	**2,459.00**	**100.00%**			**1.00**

Log likelihood: -0.24503

Figure 25. Clustering of shop drawings' hourly unit cost.

Shop Drawings Weeks per Unit					
Cluster	**N**	**%**	**Mean**	**StdDev**	**Prior Probability**
0	1,293	31.24%	0.42	0.1781	0.329
1	621	15.00%	1.05	0.4051	0.1673
2	199	4.81%	2.45	0.8836	0.0555
3	14	0.34%	12.64	5.2183	0.0041
4	1,966	47.50%	0.15	0.0742	0.4296
5	46	1.11%	4.78	1.5437	0.0145
Total	**4,139.00**	**100.00%**			**1.00**

Log likelihood: -0.41406

Figure 26. Clustering of shop drawings' weekly unit duration.

Viewer

Relation: 02, Production Duration Analysis_clustered

No.	Instance_number Numeric	ID Numeric	Program Numeric	Project Numeric	Package Numeric	Type Numeric	Weight Numeric	FabHours Numeric	FabCostMultiplier Numeric	FabDuration Numeric	FabDurationMultiplier Numeric	Cluster Nominal
1	0.0	1.0	1.0	512.0	1059.0	1.0	10.0	20.0	2.0	3.4	0.34	cluster12
2	1.0	3.0	1.0	514.0	1115.0	1.0	1.5	22.0	14.67	2.2	1.47	cluster10
3	2.0	4.0	1.0	518.0	1184.0	1.0	26.0	16.0	0.62	3.2	0.12	cluster14
4	3.0	5.0	1.0	518.0	1185.0	1.0	18.0	16.0	0.89	3.2	0.18	cluster15
5	4.0	6.0	1.0	518.0	1186.0	1.0	9.0	16.0	1.78	3.0	0.33	cluster12
6	5.0	7.0	1.0	521.0	1197.0	1.0	36.8	65.0	1.77	6.4	0.17	cluster8
7	6.0	8.0	1.0	521.0	1198.0	1.0	20.9	65.0	3.11	10.6	0.51	cluster17
8	7.0	9.0	1.0	521.0	1199.0	1.0	10.6	65.0	6.13	9.2	0.87	cluster11
9	8.0	10.0	1.0	532.0	1269.0	1.0	4.2	62.0	14.76	3.2	0.76	cluster5
10	9.0	12.0	1.0	536.0	1337.0	1.0	2.7	54.0	20.0	2.8	1.04	cluster16
11	10.0	13.0	1.0	537.0	1336.0	1.0	4.0	30.0	7.5	2.8	0.7	cluster11
12	11.0	14.0	1.0	538.0	1361.0	1.0	1.1	36.0	32.73	2.4	2.18	cluster9
13	12.0	21.0	1.0	544.0	1364.0	1.0	302.7	8.0	0.03	9.0	0.03	cluster14
14	13.0	22.0	1.0	544.0	1365.0	1.0	171.5	24.0	0.14	8.0	0.05	cluster14
15	14.0	23.0	1.0	544.0	1366.0	1.0	59.3	40.0	0.67	5.2	0.09	cluster14
16	15.0	25.0	1.0	545.0	1371.0	1.0	3.0	33.0	11.0	1.4	0.47	cluster11
17	16.0	27.0	1.0	547.0	1407.0	1.0	4.4	4.0	0.91	1.4	0.32	cluster15
18	17.0	30.0	1.0	549.0	1427.0	1.0	1.2	28.6	23.83	2.6	2.17	cluster10
19	18.0	31.0	1.0	549.0	1436.0	1.0	2.5	28.6	11.44	1.6	0.64	cluster11
20	19.0	33.0	1.0	550.0	1433.0	1.0	1.3	20.0	15.38	2.8	2.15	cluster10
21	20.0	35.0	1.0	550.0	1435.0	1.0	6.5	15.0	2.31	2.2	0.34	cluster17
22	21.0	38.0	1.0	550.0	1471.0	1.0	4.8	12.0	2.5	1.2	0.25	cluster8
23	22.0	39.0	1.0	550.0	1472.0	1.0	24.0	18.0	0.75	1.4	0.06	cluster14
24	23.0	40.0	1.0	550.0	1510.0	1.0	2.2	15.0	6.82	3.2	1.45	cluster0

Undo OK Cancel

Figure 27. *Weka* results viewer.

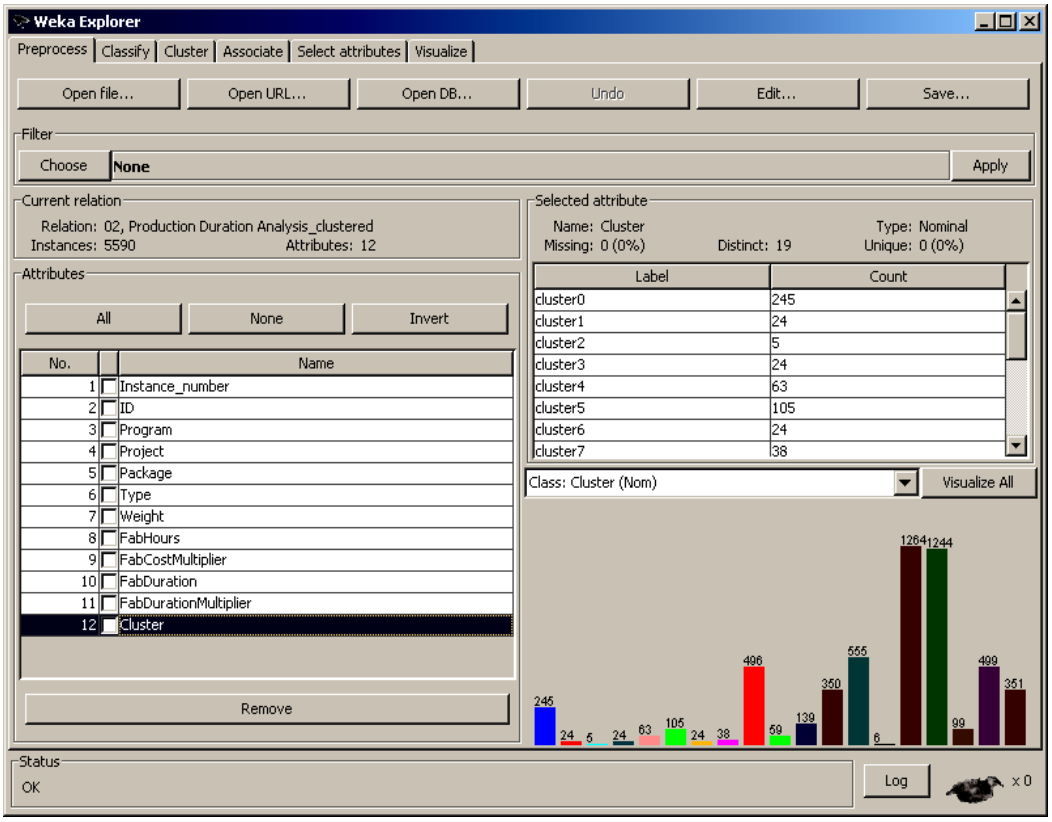

Figure 28. The frequency histogram for the obtained clusters.

ID	Weight	Actual Fab Weeks	FabWeeks Per Unit Cluster	Estimated Fab Weeks	Weeks Variance	Weeks Error
25	3	1.2	cluster13	1.4	0.21	14.77%
31	2.5	1.2	cluster13	1.2	-0.03	2.28%
57	4	1.2	cluster7	1.3	0.11	8.31%
75	9.7	5.6	cluster9	6.3	0.72	11.43%
91	6	3.6	cluster9	3.9	0.31	7.95%
109	7.9	3.2	cluster13	3.7	0.51	13.69%
161	1.5	1.6	cluster6	1.7	0.08	5.00%
171	1.6	4.6	cluster2	4.9	0.30	6.12%
177	5.3	9	cluster5	8.5	-0.49	5.72%
185	2.2	1	cluster13	1.0	0.03	3.14%
201	5.6	2	cluster7	1.8	-0.17	9.15%
235	49.9	3.8	cluster10	4.3	0.52	11.96%
247	2.6	1.8	cluster1	1.7	-0.11	6.21%
251	2.1	3.6	cluster5	3.4	-0.23	6.72%
266	11.4	2.8	cluster8	2.5	-0.27	10.54%
279	2.8	1.6	cluster9	1.8	0.23	12.33%
294	5.4	3.6	cluster1	3.5	-0.08	2.28%
316	2.7	1.6	cluster9	1.8	0.16	9.08%
336	117	5.4	cluster12	5.1	-0.25	4.90%
353	3.4	2.4	cluster1	2.2	-0.18	8.30%
364	2.2	1	cluster13	1.0	0.03	3.14%
366	13.2	1.8	cluster15	1.9	0.07	3.63%
422	5.3	2.2	cluster13	2.5	0.29	11.55%
459	30.3	14	cluster13	14.2	0.22	1.55%
505	2.8	14.2	cluster14	15.5	1.33	8.54%
559	1.7	1	cluster9	1.1	0.11	9.75%
583	8.5	3	cluster7	2.8	-0.22	7.87%

Figure 29. An excerpt of the validation tool for the fabrication phase.

ments. The same methodology can be applied to obtain resource utilization curves per work package for estimating resource requirements during the detailed planning stage of any project.

The procedure for obtaining the third dataset started with getting a list of all completed projects between the years 2004 and 2007, as shown in **Figure 30**. This list was obtained from the timekeeping system of the company, which is an in-house developed SQL server application. The list contained more than 1500 projects that vary in duration, cost and complexity. The data was automatically extracted out of the SQL-Server data tables to MS Excel for cleaning and preprocessing.

Project phase is an important control attribute for the data mining exercise. However, the company did not clearly assign project phases to the data points in their timekeeping system. As a result, it was necessary in this research to go back to the archives in order to assign the proper phase to each project. This process again consumed lots of time and effort.

Since the construction support phase is mostly responding to requests from sites and is not performed based on clearly defined scope, projects that were assigned to the "construction support" phase were eliminated from the dataset. Projects that were cancelled or put on hold prior to delivering their scope were eliminated from the list as well. At the end, there were more than 350 projects in the dataset. For each of these projects, a SQL statement was run to query the weekly working hours per resource type as shown in **Figure 31**.

The company did not store the original planned, current planned, earned hours on a weekly basis. Therefore, the missing data was simulated using random numbers in order to populate the data warehouse according to the proposed structure. The complete dataset was used to calculate the performance measures (CPI and SPI) on a weekly basis for all the data points.

The period end-date was used to calculate the week, month, quarter, and year numbers for each data point to expedite the procedure of running OLAP reports and queries. The formula used to calculate the year is:

$$\text{Year Number} = \text{Year}\left(\text{Period End Date}\right) \tag{6}$$

The formula used to calculate the month number is:

$$\text{Month Number} = \text{Month}\left(\text{Period End Date}\right) \tag{7}$$

The formula to calculate the week number is:

$$\text{Week Number} = \text{Weeknum}\left(\text{Period End Date}\right) \tag{8}$$

The three-point sliding moving average was used to reduce the noise in the dataset [17]. After that, the duration data was normalized by dividing the week number by the total number of weeks. The cost data was also normalized by dividing the weekly hours by the total number of hours. The normalized data is shown in **Figure 32**.

Nassar [18] stated that dividing project progress to twenty equal periods with 5% increments is a very good method to measure project performance. Based on that, the dataset was normalized using the interpolation function of the R software.

As shown in **Figure 33**, each resource is now presents as an array $R_{(1, 20)}$. Each array is assigned to a single class. Each class represents a unique combination of a project phase, resource, size cluster and duration cluster. To obtain the size and duration clusters, the M-means clustering technique from *Weka* is used to classify the total resource of hours and project durations into groups. The clustering results are shown in **Figures 34** and **35**.

A dynamic program that allows using the polynomial regression to develop a function that represents the variation of resource utilization per week is developed in R. Polynomial regression is used when a relation between a dependent variable Y and independent variable X cannot be fit to a linear or curvilinear such as logarithmic (Log(X)), power (X^b) or exponential (b^x) relationships, where b is a constant. As shown in the code below, the program reads the data from a Comma Separated Values (CSV) file and checks for the number of classes in the file.

After that, a cycle is used to transpose the data of each group and assign it in an array that can be recognized by the R software. For each array, the “Fit” function is used to obtain a polynomial regression function of the third degree that represents the data in each group. The function is in the format:

$$Y = b_0 + b_1 * \left(X\right) + b_2 * \left(X^2\right) + b_3 * \left(X^3\right) \tag{9}$$

JobNum	JobGroup	Company	Description	EIC	ProjSponsor	ProjManager	Comments
00E1276	PACER Alliance	PETRO-CANADA PACER	Motor Protective Relay Enhancements		GAM	F NOLTE	
00E1299	PACER Alliance	PETRO-CANADA PACER	2000 / 2001 Heavy Oils Platforms		GAM	F NOLTE	
00E1450	PACER Alliance	PETRO-CANADA PACER	Replacement of HF Detection / H2O Monitor PLC				
00E1450c	PACER Alliance	PETRO-CANADA PACER	Replacement of HF Detection / H2O Monitor PLC	CKS	G MACMILLAN	A LENUIK	
01E1465		COLT ENGINEERING CORP - EDMONTON					
01E1476	CORE PROJECTS	BP CANADA CHEMICAL	Set Up Project Management Files & System	RJT	B Bowhay	B TURCOT	
01E1505	CORE PROJECTS	BP CANADA CHEMICAL	MOC #62 - Rail Loading Pumps Shutdown Control	RJT	B Bowhay	B TURCOT	
01E1506	CORE PROJECTS	BP CANADA CHEMICAL	MOC #79 - T-5906 A/B Valve for Rail Loading2643	RJT	B Bowhay	B TURCOT	
01E1510	CORE PROJECTS	BP CANADA CHEMICAL	MOC #89 Portable Nitrogen Heater Cart Construction	RJT	B Bowhay	B TURCOT	
01E1511	CORE PROJECTS	BP CANADA CHEMICAL	MOC #49 K-5201 A/B Suction Line Drain	RJT	B Bowhay	B TURCOT	
01E1512	CORE PROJECTS	BP CANADA CHEMICAL	MOC #10 VP-5452 Vapour Bypass	RJT	B Bowhay	B TURCOT	
01E1514	CORE PROJECTS	BP CANADA CHEMICAL	MOC #104 T-5802 Rework Tank Transfer to D-5615	RJT	B Bowhay	B TURCOT	
01E1534	CORE PROJECTS	BP CANADA CHEMICAL	MOC #15 D-5475, D-5480 Hotwell Sample Port	RJT	B Bowhay	B TURCOT	
01E1535	CORE PROJECTS	BP CANADA CHEMICAL	MOC #102 C-5430 Butane Sample Points	RJT	B Bowhay	B TURCOT	
01E1537	CORE PROJECTS	BP CANADA CHEMICAL	MOC #082 Hot Oil Pump Isolation Valve Controls	RJT	B Bowhay	B TURCOT	
01E1560	CORE PROJECTS	SHELL CHEMICALS CANAD	Alliance Procedures		GAM / ASN	W MATTER	
01E1562		SUNCOR ENERGY INC 200	Millennium Extraction Wood Removals	RA	B Bowhay	M EWANCHUK	
01E1571	CORE PROJECTS	BP CANADA CHEMICAL	MOC #02 Zone Store in LAO Plant	RJT	R Karren	T Kucher	
01E1572	CORE PROJECTS	BP CANADA CHEMICAL	MOC #072 Control Building Lab. Modifications	RJT	B Bowhay	B TURCOT	
01E1572A	CORE PROJECTS	BP CANADA CHEMICAL	MOC 72 - Contro Buiding Lab. Modifications	RJT	B BOWHAY	B TURCOT	
01E1578	CORE PROJECTS	BP CANADA CHEMICAL	MOC #132 Decene for Seal Liquid D-5982 and D-5984	RJT	B Bowhay	B TURCOT	
01E1582	CORE PROJECTS	BP CANADA CHEMICAL	Redlines and As-Builts	RJT	B Bowhay	B TURCOT	
01E1582A	CORE PROJECTS	BP CANADA CHEMICAL	Redlines & As-Builts - NON COLT MOC's REV 6B	DGL	B Bowhay	B TURCOT	
01E1582B	CORE PROJECTS	BP CANADA CHEMICAL	Redlines & As-Builts - NON COLT MOC's after April	DGL	R Karren	T Kucher	
01E1582D	CORE PROJECTS	BP CANADA CHEMICAL	Structural As-Builts	Don Le	B BOWHAY	B TURCOT	
01E1597	CORE PROJECTS	BP CANADA CHEMICAL	MOC #135 - Installation of Maintenance Access Door	RJT	B Bowhay	B TURCOT	
01E1599	CORE PROJECTS	BP CANADA CHEMICAL	Document Control	RJT	B Bowhay	B TURCOT	
01E1608	CORE PROJECTS	BP CANADA CHEMICAL	MOC #150 Drainage for Control Valves Outside	RJT	B Bowhay	B TURCOT	
01E1614	CORE PROJECTS	BP CANADA CHEMICAL	MOC #164 - Install tie-ins to reroute SF-5210A/B O	RJT	B Bowhay	B TURCOT	
01E1617	CORE PROJECTS	LAO CANADA CHEMICAL P	MOC 086 - Maintenance Small Equipment Decontami	RJT	R Karren	B Turcot	
01E1643	CORE PROJECTS	BP CANADA CHEMICAL	AA - Line Design Pressure Change for Start-Up Mode				
01E1657	CORE PROJECTS	BP CANADA CHEMICAL	Inst. of Pad and Utilities for Skid Fuel Tanks	DM	B Bowhay	B TURCOT	
01E1662	CORE PROJECTS	BP CANADA CHEMICAL	Betz BFW Treatment Skid Shelter (MOC #84)	DL	B Bowhay	B TURCOT	
01E1663	CORE PROJECTS	BP CANADA CHEMICAL	Cylinder / Packing Lubrication Reservoir Upgrade (	DL	B Bowhay	B TURCOT	

Figure 30. List of completed projects between 2004 and 2007.

JOBNUM	SUBJOB	L1CODE	JobGroup	SUM(Hours)	Period End
04E2583	891	1130	CORE PROJECTS	1.00	07-Jan-05
04E2583	891	1130	CORE PROJECTS	5.50	21-Jan-05
04E2583	891	1130	CORE PROJECTS	16.50	11-Feb-05
04E2583	891	1130	CORE PROJECTS	0.50	25-Feb-05
04E2583	896	1130	CORE PROJECTS	2.50	03-Sep-04
04E2583	903	1130	CORE PROJECTS	0.50	25-Feb-05
04E2583	904	1130	CORE PROJECTS	0.50	25-Feb-05
04E2583	904	1130	CORE PROJECTS	1.00	04-Mar-05
04E2583	905	1130	CORE PROJECTS	7.50	10-Dec-04
04E2583	908	1130	CORE PROJECTS	1.00	26-Nov-04
04E2583	908	1130	CORE PROJECTS	0.50	10-Dec-04
04E2583	908	1130	CORE PROJECTS	4.50	21-Jan-05
04E2583	908	1130	CORE PROJECTS	7.00	04-Feb-05
04E2583	908	1130	CORE PROJECTS	16.25	25-Feb-05
04E2583	914	1130	CORE PROJECTS	10.00	26-Nov-04
04E2583	914	1130	CORE PROJECTS	6.50	17-Dec-04
04E2583	914	1130	CORE PROJECTS	1.00	07-Jan-05
04E2583	914	1130	CORE PROJECTS	0.50	21-Jan-05
04E2583	920	1130	CORE PROJECTS	1.00	12-Nov-04
04E2583	920	1130	CORE PROJECTS	1.50	26-Nov-04
04E2583	920	1130	CORE PROJECTS	1.00	03-Dec-04
04E2583	920	1130	CORE PROJECTS	0.50	17-Dec-04
04E2583	920	1130	CORE PROJECTS	1.00	07-Jan-05
04E2583	920	1130	CORE PROJECTS	2.50	18-Feb-05
04E2583	920	1130	CORE PROJECTS	23.25	04-Mar-05
04E2583	921	1130	CORE PROJECTS	1.00	03-Dec-04
04E2583	949	1130	CORE PROJECTS	1.00	12-Nov-04
04E2553	0	1130	SHELL ALLIANCE	5.00	26-Mar-04

Figure 31. Weekly actual working hours per resource.

Period	Hours				Normalized	
Number	Raw		Smoothe	Project	Duratio	Hours
4	33.00	27.50	28.86	04E2510	0.36364	0.12401
5	2.00	17.00	18.36	04E2510	0.45455	0.07890
6	16.00	11.58	12.95	04E2510	0.54545	0.05563
7	16.75	12.17	13.53	04E2510	0.63636	0.05813
8	3.75	10.00	11.36	04E2510	0.72727	0.04882
9	9.50	5.42	6.78	04E2510	0.81818	0.02913
10	3.00	5.00	6.36	04E2510	0.90909	0.02734
11	2.50	1.83	3.20	04E2510	1.00000	0.01374
1	1.00	4.17	4.37	04E2518	0.05556	0.03776
2	11.50	10.00	10.20	04E2518	0.11111	0.08815
3	17.50	13.17	13.37	04E2518	0.16667	0.11551
4	10.50	11.00	11.20	04E2518	0.22222	0.09679
5	5.00	7.33	7.54	04E2518	0.27778	0.06511
6	6.50	7.42	7.62	04E2518	0.33333	0.06583
7	10.75	6.67	6.87	04E2518	0.38889	0.05936
8	2.75	8.75	8.95	04E2518	0.44444	0.07735
9	12.75	6.83	7.04	04E2518	0.50000	0.06080
10	5.00	8.33	8.54	04E2518	0.55556	0.07375
11	7.25	4.75	4.95	04E2518	0.61111	0.04280
12	2.00	4.58	4.79	04E2518	0.66667	0.04136
13	4.50	2.33	2.54	04E2518	0.72222	0.02192
14	0.50	1.83	2.04	04E2518	0.77778	0.01760
15	0.50	0.50	0.70	04E2518	0.83333	0.00608
16	0.50	2.75	2.95	04E2518	0.88889	0.02552
17	7.25	5.92	6.12	04E2518	0.94444	0.05288
18	10.00	5.75	5.95	04E2518	1.00000	0.05144

Figure 32. The normalized dataset.

Group	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
61	0.01859	0.02289	0.02719	0.03903	0.05181	0.07207	0.09446	0.12672	0.16392	0.1924	0.2139	0.21933	0.20465	0.18332	0.14865	0.11441	0.08164	0.05089	0.03634	0.02179
61	0.01995	0.07725	0.11094	0.11779	0.10512	0.10697	0.09117	0.05608	0.04095	0.0595	0.06176	0.03778	0.0117	0.00559	0.00567	0.00588	0.00588	0.00868	0.01049	0.00979
61	0.02356	0.06549	0.06856	0.08981	0.06422	0.05598	0.03317	0.03209	0.04361	0.08195	0.09916	0.07918	0.04058	0.02804	0.03306	0.03437	0.03302	0.02071	0.01593	0.0065
61	0.04786	0.06744	0.07782	0.08053	0.07686	0.07584	0.07539	0.07488	0.07311	0.07044	0.06952	0.07117	0.07674	0.08202	0.08556	0.08553	0.06789	0.04823	0.02819	0.00888
62	0.10235	0.11396	0.06453	0.03201	0.04584	0.03713	0.03239	0.02761	0.02808	0.02949	0.02898	0.028	0.02385	0.01907	0.01997	0.02202	0.03247	0.03811	0.04391	0.03504
62	0.01623	0.06654	0.07814	0.05078	0.03472	0.03697	0.03313	0.01664	0.01469	0.01502	0.01675	0.0188	0.01989	0.01255	0.01099	0.01726	0.04354	0.05835	0.02588	0.01264
62	0.07212	0.09203	0.10451	0.11304	0.12194	0.11577	0.09677	0.06834	0.03744	0.06714	0.08914	0.09273	0.09273	0.05487	0.02689	0.02543	0.01863	0.01817	0.01784	0.01312
62	0.0326	0.04334	0.04427	0.04209	0.0394	0.03789	0.04062	0.05405	0.06983	0.07717	0.07885	0.07507	0.06484	0.06752	0.07524	0.07625	0.07566	0.07138	0.0601	0.04935
62	0.01968	0.01574	0.03836	0.08971	0.07553	0.05062	0.03374	0.04031	0.02409	0.01349	0.01261	0.00909	0.00853	0.00915	0.01303	0.0228	0.01869	0.01617	0.01285	0.00605
62	0.06871	0.06296	0.02233	0.01239	0.01293	0.01333	0.00713	0.0063	0.01412	0.02501	0.02642	0.02136	0.01937	0.02623	0.03702	0.07638	0.08154	0.04518	0.0367	0.02732
62	0.14623	0.1613	0.15377	0.05207	0.05395	0.03983	0.02571	0.01911	0.03889	0.04171	0.04548	0.02571	0.02665	0.02006	0.02006	0.01629	0.01629	0.02476	0.03983	0.0325
62	0.10074	0.11945	0.10345	0.07748	0.05526	0.04021	0.02757	0.01228	0.0232	0.03914	0.06518	0.08846	0.08332	0.06855	0.06536	0.06201	0.06019	0.04842	0.02146	0.01148
62	0.04551	0.066	0.05578	0.05076	0.07815	0.08364	0.05865	0.04745	0.03553	0.02654	0.01848	0.01245	0.00983	0.01153	0.0107	0.00647	0.00708	0.00514	0.00384	0.00434
62	0.09602	0.12268	0.11389	0.09043	0.08726	0.06083	0.05065	0.03169	0.02296	0.01256	0.01223	0.01055	0.00955	0.01431	0.01907	0.02601	0.03076	0.05158	0.05158	0.04015
62	0.0492	0.03532	0.02403	0.05303	0.07449	0.06198	0.04523	0.02948	0.01003	0.01645	0.01979	0.01419	0.01455	0.03837	0.05535	0.03507	0.01796	0.01776	0.03841	0.03714
62	0.00473	0.03866	0.02601	0.01466	0.02131	0.02185	0.03203	0.0367	0.03206	0.03272	0.02357	0.02541	0.03922	0.03245	0.02915	0.01669	0.01443	0.04119	0.03652	0.02002
62	0.02429	0.0349	0.03324	0.02859	0.02243	0.02896	0.03083	0.02716	0.02486	0.05468	0.08358	0.07939	0.04651	0.03138	0.05715	0.07279	0.0388	0.02375	0.02856	0.02593
62	0.06597	0.09953	0.0936	0.09454	0.08653	0.07754	0.06537	0.04497	0.03625	0.03894	0.03988	0.04168	0.04201	0.03938	0.03675	0.03445	0.03264	0.0379	0.05632	0.04514
62	0.09371	0.11381	0.1339	0.14769	0.14567	0.14366	0.14331	0.14421	0.1451	0.14064	0.13527	0.13046	0.12889	0.12733	0.12644	0.12644	0.12644	0.11591	0.10118	0.08644
62	0.01884	0.03198	0.0421	0.04837	0.05125	0.05312	0.05708	0.06101	0.0714	0.09075	0.10459	0.11497	0.11498	0.10548	0.09016	0.07685	0.05582	0.03824	0.02461	0.00888
64	0.09938	0.12642	0.12385	0.09674	0.07962	0.05666	0.04242	0.09605	0.10975	0.10975	0.12327	0.09068	0.05715	0.05264	0.04432	0.0353	0.02629	0.02511	0.01936	0.0144
64	0.15785	0.25616	0.20546	0.09575	0.02066	0.02937	0.03789	0.03059	0.0195	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00957	0.00712
64	0.06106	0.06313	0.04826	0.02068	0.01489	0.02183	0.0352	0.04017	0.09357	0.1575	0.18482	0.19526	0.1776	0.10806	0.06828	0.04412	0.03625	0.03144	0.02441	0.01448
64	0.02371	0.02109	0.0128	0.05771	0.08689	0.0733	0.02612	0.04898	0.06482	0.07795	0.05974	0.07428	0.04957	0.02614	0.0239	0.0216	0.03612	0.03064	0.01184	0.0039
64	0.00284	0.01255	0.0785	0.14162	0.1128	0.04629	0.02198	0.0403	0.051	0.05442	0.03735	0.01442	0.0109	0.00521	0.00413	0.00454	0.00874	0.00928	0.01131	0.01185
64	0.00221	0.0031	0.00399	0.00965	0.03999	0.0831	0.12392	0.14896	0.14574	0.1485	0.13848	0.10853	0.07038	0.03671	0.01968	0.01297	0.00775	0.00562	0.00688	0.0042
64	0.1053	0.14783	0.18566	0.18123	0.17007	0.13197	0.09919	0.0788	0.06817	0.07216	0.07637	0.0808	0.07725	0.06839	0.06635	0.06724	0.05891	0.04827	0.03525	0.02196
64	0.08889	0.13292	0.14024	0.12296	0.09349	0.06728	0.06154	0.06965	0.07092	0.05089	0.03279	0.02792	0.02824	0.03288	0.03317	0.03447	0.03001	0.02422	0.02178	0.00874
64	0.0702	0.08209	0.08414	0.07985	0.0734	0.07649	0.07836	0.07872	0.07892	0.07809	0.06651	0.04627	0.03761	0.03591	0.04353	0.06043	0.06909	0.07359	0.06856	0.04942
64	0.05332	0.09461	0.09371	0.08929	0.05732	0.02126	0.01253	0.01015	0.01344	0.01168	0.01963	0.01951	0.00939	0.00814	0.01833	0.01932	0.0083	0.00738	0.01706	0.01648
64	0.02551	0.02743	0.01311	0.01157	0.01701	0.03446	0.0715	0.09031	0.04855	0.04007	0.0665	0.03428	0.01832	0.00877	0.00934	0.01527	0.01256	0.01404	0.01259	0.00621
64	0.07831	0.109	0.13969	0.1518	0.16159	0.13885	0.10683	0.10327	0.11394	0.12338	0.13184	0.1389	0.14423	0.1346	0.09502	0.06117	0.04738	0.03413	0.02524	0.01634
64	0.00232	0.00582	0.02755	0.05636	0.06334	0.04084	0.01961	0.01852	0.04414	0.03799	0.03813	0.05846	0.08723	0.08978	0.08671	0.04851	0.00767	0.00277	0.00452	0.00394

Figure 33. The normalized dataset after interpolation.

Hours - All Data					
Cluster	N	%	Mean	StdDev	Prior Probability
Cluster0	221	40.48%	128.58	36.3162	0.3554
Cluster2	220	40.29%	291.38	111.7130	0.4201
Cluster1	89	16.30%	968.03	484.1862	0.1880
Cluster3	16	2.93%	2841.21	1,774.2157	0.0365
Total	546.00	100.00%	453.93	687.3120	1.00

Figure 34. Clustering of total resource hours.

Weeks - All Data					
Cluster	N	%	Mean	StdDev	Prior Probability
Cluster0	360	65.93%	15.04	5.6524	0.5951
Cluster2	148	27.11%	29.85	9.6972	0.3083
Cluster1	38	6.96%	56.24	20.3446	0.0966
Total	546.00	100.00%	25.58	15.7160	1.00

Figure 35. Clustering of total duration weeks.

The third degree polynomial, sometimes referred to as cubic function, provides an S-curve, which fits reasonably well to the distribution of resource utilization over the project percent complete. The output of the developed code is a list of the coefficients: b_0, b_1, b_2 and b_3. The user can easily change the degree of the polynomial to any other degree using the function "PolDgr". The goodness of fit is measured using the least square errors (R^2) and the user can try different functions to find the one that fits best for the dataset under investigation.

The output of the code is written to another CSV file and an example of it is shown in **Figure 36**. The goodness of fit is tested using the R^2 function and graphically. The output for each class is plotted accompanied with the original values of any class to visually test the goodness of fit. At the beginning of any internal project, the project can decide on the size and duration class for each resource, use the characteristics of these classes accompanied with the polynomial function for the distribution of these resource utilization over project percent complete to predict the initial planned values for each resource. These predicted values are based on PM judgment and historical data.

Another approach is to connect the averages of each percent complete (PI, P2 to P20). It is up to the user to decide on which methodology fits better for the existing data. This case study was used to provide the user with the Initial Planned Values (IPV) needed prior to the detailed planning of any project.

In the third dataset, project attributes were not clearly identified when data was collected. Moreover, some of the projects were not broken into clearly defined phases as proposed in this research. When data was analyzed, discrepancies were found among the resource utilization graphs. These discrepancies were highlighted and recommendations were made to the partner company.

5. Conclusion

5.1. Research Summary

The aim of this research was to improve resources management practices by using existing historical data from completed projects to forecast needs of future projects. During the process of managing labour resources in a multiple-project environment, a large amount of multidimensional data is generated, collected and stored in scattered formats. Currently, there is no consistent methodology to manage this wealth of data. Most of this data gets lost and is never viewed, analyzed or transferred to useful knowledge that could be an asset in improving resource management practices. This research developed an integrated framework for managing resources data in multiple-project environment. The framework is built on a KDD model to transfer the collected multidimensional historical data from completed projects to useful knowledge for new projects.

Three case studies were performed to validate the applicability of the developed framework to real projects data. The first dataset was obtained from a partner company and was utilized to define the distribution parame-

					0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Grou	b0	b1	b2	b3	P00	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10
61	0.03646	0.44515	-0.9392	0.47088	0	0.05642	0.07205	0.08369	0.09168	0.0964	0.09819	0.09739	0.09438	0.08949	0.08309
62	0.05595	-0.0827	-0.0071	0.07006	0	0.05181	0.04768	0.04362	0.03969	0.03593	0.0324	0.02915	0.02622	0.02369	0.02159
64	0.22147	-0.9332	1.33796	-0.6206	0	0.17808	0.14091	0.1095	0.08338	0.0621	0.04517	0.03214	0.02254	0.01591	0.01178
65	0.12167	0.0576	-0.0232	-0.157	0	0.12447	0.12704	0.12925	0.131	0.13216	0.13262	0.13225	0.13095	0.12858	0.12505
66	0.00571	0.3515	-0.6502	0.30994	0	0.0217	0.03466	0.04485	0.05248	0.05779	0.06101	0.06237	0.06211	0.06046	0.05765
71	0.03005	0.02464	-0.0711	0.03333	0	0.0311	0.03183	0.03225	0.0324	0.03228	0.03194	0.03139	0.03066	0.02978	0.02876
72	0.12914	0.23537	-1.0044	0.67833	0	0.13848	0.14331	0.14414	0.14146	0.1358	0.12767	0.11756	0.10599	0.09347	0.08051
74	0.013	0.10858	0.33491	-0.2978	0	0.01923	0.02691	0.03582	0.04573	0.05643	0.06768	0.07926	0.09096	0.10255	0.1138
75	0.02104	0.00439	0.20323	-0.1888	0	0.02174	0.02332	0.02563	0.02854	0.03189	0.03555	0.03938	0.04323	0.04697	0.05044
76	0.0367	0.216	-0.2163	0.0314	0	0.04697	0.05617	0.06434	0.0715	0.07768	0.08289	0.08716	0.09051	0.09297	0.09456
81	0.00592	0.39438	-0.6569	0.29545	0	0.02404	0.03909	0.0513	0.06089	0.06808	0.0731	0.07616	0.07748	0.0773	0.07582
82	0.03564	-0.0609	0.54407	-0.4564	0	0.0339	0.03453	0.0372	0.04156	0.04728	0.054	0.06139	0.06911	0.0768	0.08414
84	0.03982	-0.1486	0.65913	-0.5018	0	0.03397	0.03104	0.03066	0.03244	0.03601	0.041	0.04703	0.05371	0.06068	0.06756
85	0.01738	0.1342	0.09617	-0.2218	0	0.02431	0.03154	0.03893	0.0463	0.05348	0.06031	0.06663	0.07226	0.07704	0.0808
86	0.03438	-0.0103	0.43175	-0.454	0	0.03489	0.03721	0.04102	0.04596	0.05169	0.05789	0.06419	0.07028	0.0758	0.08041
91	0.02108	0.15613	-0.2557	0.09549	0	0.02826	0.03423	0.03907	0.04284	0.04562	0.04748	0.04849	0.04872	0.04825	0.04714
92	0.0212	0.04282	0.17614	-0.2099	0	0.02376	0.02704	0.03088	0.03513	0.03964	0.04423	0.04877	0.05308	0.05701	0.06041
94	0.01992	0.26625	-0.3084	0.05304	0	0.03247	0.04351	0.0531	0.06126	0.06804	0.07347	0.07761	0.08048	0.08212	0.08258
95	0.0244	0.19691	-0.2891	0.07442	0	0.03353	0.04127	0.04768	0.05281	0.05672	0.05947	0.0611	0.06167	0.06125	0.05989
96	0.02095	0.0163	0.23396	-0.2686	0	0.02232	0.02465	0.02775	0.03142	0.03545	0.03964	0.0438	0.04771	0.05118	0.05401
101	0.01586	0.52732	-1.3385	0.8201	0	0.03898	0.05603	0.06761	0.07434	0.07685	0.07573	0.07161	0.06511	0.05683	0.0474
102	0.0205	0.19671	-0.4563	0.24289	0	0.02922	0.03585	0.04056	0.04353	0.04495	0.045	0.04386	0.04172	0.03875	0.03514
104	-0.0002	0.56418	-1.3467	0.79482	0	0.02472	0.04352	0.05679	0.06511	0.06908	0.06929	0.06635	0.06085	0.05339	0.04455
105	0.02886	-0.0122	0.22436	-0.2233	0	0.02878	0.02966	0.03133	0.03361	0.03635	0.03937	0.04251	0.0456	0.04847	0.05095
106	-0.0261	0.72905	-1.2376	0.54602	0	0.00735	0.035	0.05728	0.0746	0.08737	0.096	0.10089	0.10247	0.10113	0.09729
221	0.04911	0.20791	-0.4086	0.19002	0	0.05851	0.066	0.07174	0.07587	0.07852	0.07984	0.07997	0.07906	0.07724	0.07467
222	0.15758	-0.0957	-0.3868	0.37041	0	0.15187	0.14451	0.13577	0.12593	0.11526	0.10405	0.09257	0.08111	0.06993	0.05932

Figure 36. An example of the coefficients output.

ters of estimating unit costs. An anomaly detection methodology was developed to highlight the inconsistent data points for the end-user. A unit cost tree with branches was obtained. PostHoc tests and the One-way ANOVA were used to classify the cost units into a smaller number of groups. The second dataset was obtained from another partner company and was used to define the distribution parameters of estimating unit durations within different data clusters. The dataset was randomly divided into training set and testing set for validation purposes. More than 85% of the testing data points had an estimating error of less than 25%. The third dataset was used to analyze various resource utilization patterns over time units and to find the most fitting resources utilization curve per cluster.

By studying the original dataset, several problems were identified. These problems are mainly pertaining to the lack of a proper definition of data dimensions, objects and attributes and to the lack of a systematic consistent integrated approach to data collection and storage. There is a perception in the industry that each project is unique and its data is unique as well, and therefore, data from projects are not easily aggregated nor transferred to useful knowledge.

By implementing the data collection integrated framework to the original dataset, this research demonstrated that data can be collected in a systematic and consisted manner, which then could be analysed in a variety of ways, and then leads to extracting useful knowledge that would improve labour resources management practices and forecasts. As a result of this framework, productivity and efficiency would increase. As well, a continuous knowledge cycle and a self-learning loop would be established between completed and future projects.

5.2. Recommendations for Future Research

The developed KDD model was implemented into the management of labour resources data in industrial construction project domain. Further research can be carried out to investigate the feasibility of applying this model to other non-labour resources types. In addition, other researchers can investigate extending the application of this model to other domains such as infrastructure or commercial construction.

Clustering and anomaly detection data mining techniques were used to extract knowledge from the available datasets. Future research can apply other data mining techniques or knowledge discovery techniques such as classification, finding association rules, simulation, artificial neural networks, and fuzzy sets. The data warehouse would provide a systematic methodology to model projects, their objects and projects' data for analysis by these sophisticated research methods.

Once populated with enough data, the data warehouse along with advanced research techniques can be used to identify the main factors impacting labour resources performance and overall project performance.

Acknowledgements

The authors would like to thank WorleyParsons Canada—Edmonton Division and Waiward Steel Fabricators Ltd. for providing the necessary data for the case study. This research was supported by the NSERC Industrial Research Chair in Construction Engineering and Management, IRCPJ 195558-10.

References

[1] Jergeas, G. (2008) Analysis of the Front-End Loading of Alberta Mega Oil Sands Projects. *Project Management Journal*, **39**, 95-104. http://dx.doi.org/10.1002/pmj.20080

[2] Inmon, W.H. (2005) Building the Data Warehouse. Wiley, Indianapolis.

[3] Giovinazzo, W.A. (2000) Object-Oriented Data Warehouse Design: Building a Star Schema. Prentice Hall, Upper Saddle River.

[4] Ahmad, I., Azhar, S. and Lukauskis, P. (2004) Development of a Decision Support System Using Data Warehousing to Assist Builders/Developers in Site Selection. *Automation in Construction*, **13**, 525-542. http://dx.doi.org/10.1016/j.autcon.2004.03.001

[5] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann, Elsevier Science Distributor, San Francisco.

[6] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, **17**, 37.

[7] Cios, K.J. (2007) Data Mining: A Knowledge Discovery Approach. Springer, New York.

[8] Zaiane, O.R., Foss, A., Lee, C.H. and Wang, W. (2002) On Data Clustering Analysis: Scalability, Constraints, and Validation. *Proceedings of the* 6*th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer-Verlag, Berlin, 28-39.

[9] Soibelman, L. and Kim, H. (2002) Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases. *Journal of Computing in Civil Engineering*, **16**, 39-48. http://dx.doi.org/10.1061/(ASCE)0887-3801(2002)16:1(39)

[10] Chau, K.W., Cao, Y., Anson, M. and Zhang, J. (2002) Application of Data Warehouse and Decision Support System in Construction Management. *Automation in Construction*, **12**, 213-224. http://dx.doi.org/10.1016/S0926-5805(02)00087-0

[11] Rujirayanyong, T. and Shi, J.J. (2006) A Project-Oriented Data Warehouse for Construction. *Automation in Construction*, **15**, 800-807. http://dx.doi.org/10.1016/j.autcon.2005.11.001

[12] Moon, S.W., Kim, J.S. and Kwon, K.N. (2007) Effectiveness of OLAP-Based Cost Data Management in Construction Cost Estimate. *Automation in Construction*, **16**, 336-344. http://dx.doi.org/10.1016/j.autcon.2006.07.008

[13] Fan, H., AbouRizk, S., Kim, H. and Zaiane, O. (2008) Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model. *Journal of Computing in Civil Engineering*, **22**, 181-191. http://dx.doi.org/10.1061/(ASCE)0887-3801(2008)22:3(181)

[14] Hammad, A., AbouRizk, S. and Mohamed, Y. (2013) Application of Knowledge Discovery in Data (KDD) Techniques to Extract Useful Knowledge from Labour Resources Data in Industrial Construction Projects. *Journal of Management in Engineering*. http://dx.doi.org/10.1061/(ASCE)ME.1943-5479.0000280

[15] Zaiane, O.R. (2006) Principles of Knowledge Discovery in Data. Lecture at University of Alberta. http://webdocs.cs.ualberta.ca/~zaiane/courses/cau/slides/cau-Lecture7.pdf

[16] Witten, I.H. and Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman, Amsterdam, Boston.

[17] Teicholz, P. (1993) Forecasting Final Cost and Budget of Construction Projects. *Journal of Computing in Civil Engineering*, **7**, 511-529. http://dx.doi.org/10.1061/(ASCE)0887-3801(1993)7:4(511)

[18] Nassar, N.K. (2005) An Integrated Framework for Evaluation, Forecasting and Optimization of Performance of Construction Projects. PhD Thesis, University of Alberta (Canada), Canada.