Data Acquisition Model for Analyzing Cost Overrun in Construction Projects using KDD

Mai Monir Ghazal
Department of Industrial Engineering and Engineering Management
University of Sharjah
Sharjah, U.A.E
u15104212@sharjah.ac.ae

Ahmed Mohamed Hammad
Assistant Professor
Department of Industrial Engineering and Engineering Management
University of Sharjah
Sharjah, U.A.E
ahammad@sharjah.ac.ae

Abstract

Projects are considered successful when completed on time as per baseline schedule and within allocated target budget. Cost overrun is a worldwide challenge to successful completion of construction projects. To overcome this problem, earlier studies were conducted to investigate the main causes of cost overrun. Knowledge Discovery in Data (KDD) and data mining techniques have been implemented successfully in other research areas to extract new and useful knowledge from historical data. These techniques can be also applied to projects' historical data if this data is captured in an organized and consistent manner. First section of this paper applies a comprehensive literature review on previous research to detect the major factors causing cost overrun. This analysis resulted in selecting twelve major factors that can be easily measured and analyzed at construction projects. After that, a data acquisition model is developed to capture the relevant historical data and metadata from completed construction projects in a reliable data warehouse. The developed data warehouse would enable the implementation of KDD and data mining techniques to tackle cost overrun problem.

Keywords

Construction management, Construction projects, Cost overrun, Cost performance, Data mining, Data warehousing, Knowledge discovery, Knowledge management.

1. Study Background

Introduction

In a construction project, usually the contractors and their project team focus their efforts towards handing over the project within an acceptable budget and schedule. While on the other hand most of the clients care firstly about the project's cost then quality. A successful project means that the project has accomplished its technical performance, maintained its schedule, and remained within budgetary costs (Hammad, AbouRizk and Mohamed, 2014). Therefore, it is very important for both the contractor and the client to prepare the project estimates at the initial stages to provide the contractor by the ability to form the basis of cost-benefit analysis and to enable the client to take the "to-build-or-not-to-build" decision. Moreover, it is well known that taking decisions during the pre-execution stage is less expensive compared to the massive cost of correction actions during the execution process. Effective cost estimation is, therefore, so vital, it can seal a project's financial fate, Nicholas (2004) notes.

In construction industry, cost overrun is a severe problem due to limited availability of information during the initiation and planning stages and the huge cost of rectifying faults during execution stage if applicable, thus resulting in the failure of numerous projects. As Hegazy (2002) stated that despite the importance of cost estimation, it is undeniably neither simple nor straightforward because of the lack of information in the early stages of the project. Whereas researchers tried to identify the reason behind the failure of many projects to meet initially set cost limits they found that this failure might be due to a number of causes ranging from the inability to accurately identify and quantify risk (Akintoye, 2000), error in estimation (Jennings, 2012), project schedule changes, engineering and construction complexities, scope changes, local concerns and requirements, effects of inflation and market conditions (Shane, Molenaar, Anderson and Schexnayder, 2009). Moreover, schedule delay, project location and poor project management can also cause an obvious cost overrun.

Nowadays, it is well known that huge amount of data is continuously produced every day during all phases of Project Life Cycle (initiation, planning, implementation, and closure), this huge amount of data is typically being generated, collected and archived without proper analysis and advantage. For instance, an extensive data and documents are produced during the initiation, and planning phase such as project-scoping documents, written specifications, cost estimations, drawings and plans, through using data mining (Which is just a step of Knowledge Discovery in Database) a beneficial knowledge can be extracted from these documents to support the construction company in making timely knowledge-based decisions for future projects.

The knowledge pyramid consists of data, information, knowledge, and wisdom (Liebowitz and Megbolugbe, 2003). Knowledge discovery is a process that seeks new knowledge about an application domain. It consists of many steps, one of which is data mining (DM), each aiming to complete a discovery task, and accomplished by the application of a discovery method (Cios, Pedrycz, Swiniarski and Kurgan, 2007). Knowledge discovery process models consist of several steps that are performed in a sequence. The most common steps are understanding the project domain and data, preparing the data and analyzing the generated knowledge.

After understanding the project's domain and data, it is necessary to provide well-organized data retrieval and summarization capabilities and this can be done by structuring consistent data warehouses. A Database is "any collection of organized, related data which is used as a source of information to answer user queries or to facilitate other data processing activities" (UM Libraries, 1986). For instant, stored objects could be reports, tables and queries. While a data warehouse is a repository of data collected in different locations (relational databases) and stored using a unified schema. Data warehouses are usually created by applying a set of processing steps to data coming from multiple databases. The steps usually include data cleaning, data transformation, data integration, data loading, and periodical data update (Cios, Pedrycz, Swiniarski, Kurgan, 2007). The difference between the database and data warehouse is that the main purpose of database is storing the data while the main purpose of a data warehouse is analyzing the data. Also, in data warehouse the data is organized around a group of subjects of interest to the data user. Moreover, data warehouses store a summary of the data not the full detailed content as the source databases.

At the end of data warehousing process, the data set will be ready for applying a data mining technique to extract hidden knowledge. Han and Kamber (2006) defined data mining as "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". The aim of data mining is to make sense of large amounts of mostly unsupervised data, in some domain (Cios, Pedrycz, Swiniarski and Kurgan, 2007). Data Mining was also defined as "A process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large databases" (Turban, Aronson, Liang, and Sharda, 2007). The discovered knowledge by an effective data mining process must be previously unknown, nontrivial, and beneficial to the data owners Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996), emphasized.

Data mining techniques are categorized into two categories: supervised and unsupervised learning. Whereas, Data Mining Algorithms can be grouped into five categories: Clustering, Association Rules, Outliers Detection, Classification and Prediction (Cios, Pedrycz, Swiniarski and Kurgan, 2007). Clustering methods main aim is minimizing the distance between data points falling within the same cluster, and maximizing the distance between these clustered data points and the data points in other clusters (Zaiane, Foss, Lee, and Wang, 2002). Association Rules usually find interesting associations (relationships, dependencies) in large sets of data items. Outliers' detection

techniques detect data points that are significantly unlike from others in the data set. Classification techniques start building the model using a training dataset to define data classes, assess the model, and then use the developed model to categorize each new data point into the appropriate category. Classification techniques include Artificial Neural Networks (ANN), Decision Trees, Support Vector Machines (SVM), Bayesian Networks, Case - Based Reasoning (CBR), Instance (Memory) - Based Learning, K - Nearest Neighbors (Lazy learning), Rule - Based Induction (Cios, Pedrycz and Swiniarski Kurgan, 2007).

The objective of this research is to investigate the factors of cost overrun in construction projects then identify the measurable factors that can form the basis for developing a data acquisition model to enable capturing the relevant historical data and meta data in a reliable data warehouse. Moreover, this research forms a base which will be extended towards implementing a prediction KDD model that involves the selected cost overrun factors and developed data warehouse to forecast project's cost using a suitable data mining technique to overcome the cost overrun problem.

2. Literature Review and Contribution

2.1 Causes of Cost Overrun in Construction Industry

Among both qualitative and quantitative descriptors to evaluate the success of a construction project, it was found that cost performance of a construction project remains one of the main measures of success even though there were other emerging qualitative measures like health and safety and environmental performance (Chan and Chan, 2004). Researchers Nassar et al. (2005) have discussed in their survey that cost overrun is a common problem in construction industry through performing an intense analysis on a set of data which was collected by survey answers. The analysis results indicated that the average of cost overrun was 4% above the bid price. Researchers Yun and Caldas (2009) stated in their research that as a result of the time gap between the project's planning phase and the actual start of execution phase the possibility of cost and schedule overrun may increase. While Olawale and Sun (2010) indicated that although causes of cost overrun have common features in projects worldwide, they are also affected by the specific conditions of each country which may not be valid in other countries. According to Art Chaovalitwongse et al. (2011) the main causes of cost overrun are inflationary increases in material cost, inaccurate material estimating, and project complexity, those causes were highlighted based on their observed importance and frequencies of occurrence. So, to obtain a successful construction project, it is crucial to prepare a reliable cost estimate through taking into consideration some important factors such as the type of project, likely design and scope changes, risk and uncertainty, effect of policy and regulatory conditions, duration of project, type of client, ground conditions or tendering method (Ahiaga-Dagbui and Smith, 2014). Other researchers intensively studied the factors of cost overrun and tried to categorize and list the most important factors that affect the construction projects cost performance, their results are highlighted below:

Jahren and Ashe (1990) defined the cost overrun as "the percent difference in cost (positive or negative) between the final contract cost and the contract award amount". They deemed that both cost overrun and change orders are affected by the same factors as: project size, difference between the selected bid and the government estimate, type of construction, level of competition. Where on the other hand they also found that non-quantifiable factors such as the quality of the contract document, the nature of interpersonal relations on the project and the policies of the contractor, could also have a significant impact on cost overrun. It was found that cost overrun rate was more likely to occur on larger projects than smaller ones, so they tried to investigate the reason behind it and explained that when projects become larger the execution phase will become more complex so more cost overrun may occur.

In the same field of interest, Kaming, Olomolaiye, Holt and Harris (1997) performed a questionnaire survey among project managers working on high-rise construction projects in two Indonesian cities: Jakarta and Yogyakarta, to investigate factors that impact construction time and cost overrun. According to their results it seemed that cost overrun occurred more frequently and therefore considered more severe problem than time overrun. The result of ranking the factors influencing time overrun/delays showed that the highest ranked factors were design changes, poor labor productivity, inadequate planning and resource shortages. While emphasizing on cost overrun only, results showed that the most important factors are material cost increases due to inflation, inaccurate material estimating and degree of project complexity.

Similarly, a study was done in Kuwait by researchers Koushki, Al-Rashid and Kartam (2005) to investigate the main causes of time delay and cost overrun in construction project, they performed a person-interview survey among 450

private residential projects' owners and developers. They stated that the occurrence of schedule delay and cost overrun usually increased with the increase in the total cost of a project. Moreover, their survey results had shown that owners who spent more time on the pre-planning phase had experienced less schedule delay and cost overrun during the execution phase of their projects. Also, they found that the main causes of cost overrun in construction projects are: 1- Contractor-Related Problems, 2- Material-Related Problems, 3- Owners' Financial Constraints.

Other researchers such as Le-Hoai, Lee and Lee (2008) had also performed a deep investigation among previous studies to identify causes of cost overrun, they identified 21 of cost overrun causes and categorized them into 6 groups: 1- Owner Related Factors, 2- Contractor Related Factors, 3- Consultant Related Factors, 4- Project Related Factors, 5- Material and Labor Related Factors, 6- External Factors. They used the 21 causes of delay and cost overrun which were identified previously to prepare a questionnaire survey, 87 Vietnamese construction experts were interviewed and participated in the questionnaire survey by ranking the causes in terms of degree of occurrence and level of severity. After analyzing the results, they found that the most frequent, severe and important causes of delay and cost overrun are poor site management and supervision, poor project management assistance, financial difficulties of owner, financial difficulties of contractor and design changes.

Shane, Molenaar, Anderson and Schexnayder (2009) performed an extensive literature review to identify the cost escalation factors in transportation construction industry, they identified eighteen cost escalation factors. Them they categorized them into two categories: internal and external. Internal factors are the factors that are being controlled by the agency/owner such as: bias, delivery/procurement approach, project schedule changes, engineering and construction complexities, scope changes, scope creep, poor estimating, inconsistent application of contingencies, faulty execution, ambiguous contract provisions and contract document conflicts. While on the other hand, external factors are the factors that are outside the direct control of the agency/owner such as: local concerns and requirements, effects of inflation, scope changes, scope creep, market conditions, unforeseen events and unforeseen conditions.

Memon, Abdul Rahman, Abdullah, and Abdul Azis (2010) also performed a deep literature review and identified 24 most frequent cost overrun factors of large construction projects, these factors were used to prepare a questionnaire survey. The results showed that the most significant 15 factors affecting the project's construction cost are: 1- Cash flow and financial difficulties faced by contractors, 2- Contractor's poor site management and supervision, 3- Inadequate contractor experience, 4- Shortage of site workers, 5- Incorrect planning and scheduling by contractors, 6- Fluctuation in prices of material, 7- Practice of assigning contract to lowest bidder, 8- Lack of communication among parties, 9- Underestimate project duration resulting schedule Delay, 10- Delay in material procurement, 11- Incompetent project team (designers and contractors), 12- Unforeseen ground conditions, 13- Low speed of decisions making, 14- Change in the scope of the project, 15- Frequent design changes.

Similarly, an intense literature review was done by researchers Park & Papadopoulou (2012), they reviewed 15 research papers to identify the causes of cost overrun. As per their review they found 27 causes of cost overrun, then they identified and analyzed the most common ten causes which are: 1- Inaccurate estimates, 2- Shortage of material, 3- Price fluctuations (inflation), 4- Inappropriate procurement route/contract type, 5- Delayed payment of completed works, 6- Unforeseen site conditions, 7- Poor site management by contractor and poor planning, 8- Change orders by client (variations), 9- Lack of communication between parties, 10- Inadequate duration of contract period. After that they prepared a survey and asked the professionals from construction projects to rank the causes of cost overrun according to their frequency, severity and significance. Their survey results showed that the most significant and frequent cause of cost overrun in transport infrastructure projects in Asia is the practice of awarding contracts to the lowest bidder.

Abdul Rahman, Memon, Abdul Azis and Abdullah (2013), had reviewed several previous researches in order to identify the common causes of cost overrun in construction projects, they identified 35 of cost overrun causes and categorized them into 7 groups: 1- Contractor's Site Management Related Factors (CSM), 2- Design and Documentation Related Factors (DDF), 3- Financial Management Related Factors (FIN), 4- Information and Communication Related Factors (ICT), 5- Human Resource (Workforce) Related Factors (LAB), 6- Non Human Resource Related Factors (MMF), 7- Project Management and Contract Administration Related Factors (PMCA).

2.2 Data Mining and Cost Overrun in Construction Projects

Since 1996, data mining was widely applied in different industries to investigate improvement programs, strategic priorities, environmental factors, manufacturing performance dimensions and their interactions (Hajirezaie, Husseini, Barfourosh, and Karimi, 2010). Utilization and implementation of Knowledge Discovery in Databases, Data Warehousing and Data Mining techniques in construction field had appeared recently. These techniques provided good contributions to other industries such as marketing and healthcare and applying them to the construction industry nowadays is showing promise for improving construction and project management practices.

The amount of data generated by construction firms presents both a challenge and an opportunity: a challenge to traditional methods of data analysis since the data are often complex, and usually, voluminous. Also, an opportunity since construction firms may have the chance of gaining competitive edge and performance improvement by making their data work for them using detailed data mining (Ahiaga-Dagbui and Smith, 2014). Controlling costs in the construction industry, is continuously a concern for contractors. By using data mining process and utilizing the knowledge extracted from the previous completed projects, projects that have possibility of cost overrun can be identified during the early stages, then sufficient funds can be considered during the estimation phase to complete the project within the budget and avoid project failure. To deeply recognize the advantage of using data mining process in dealing with cost overrun issue in construction field, a second section of literature review has been carried out. This literature review includes an overview of several related research papers where for each of which the applied model procedure will be explained, type and condition of the collected data will be mentioned and finally research results will be discussed.

The researchers Soibelman & Kim (2002) used the KDD process to identify the causes of construction activity delays. During their study, they did the data mining process in three steps, the first step was feature subset selection which was used to calculate the relevance of features. Then, they used the decision tree to extract rules from the data sets. At last, the rules that were extracted from the decision tree were used as inputs to neural networks which was their third data mining technique used to predict the future trends in a construction project. The selected nine input variables to neural network were: inaccurate site survey, number of workers, incomplete drawing, change order, shortage of equipment, duration, season, weekends and rain/snow. Whereas the output value was the delay in days in the installation of drainage pipeline. To validate their results, they conducted a comparison between the results from the KDD model and other project control softwares such as RSMeans and Monte Carlo Simulation, it was found that the obtained results from the implemented KDD model were more realistic, since it considered all possible causes that may cause activity delay. Finally, the researchers suggested that their study can be extended and applied in different areas of construction projects such as identifying the causes of cost overrun, or quality control/assurance.

Other researchers such as Art Chaovalitwongse et al. (2011) addressed the issue of cost overrun in construction projects from another point of view by studying the relation between the bid selection policy and the occurrence of cost overrun using data mining techniques. Towards describing the special patterns of the submitted bid, they referred to five bidding ratios (Second-Lowest-Bid Ratio, Mean-Bid Ratio, Median-Bid Ratio, Maximum-Bid Ratio and Coefficient of Variation) which were proposed by Williams (2005). They suggested that these ratios might indicate the likelihood of cost overrun since they captured the relationships between different bids for a project. The calculated five ratios for each project were the input layer for the two types of neural networks which were used in Art Chaovalitwongse et al. (2011) study, the first neural network is the Probabilistic Neural Network (PNN) which was used to classify projects into two groups whether their actual cost is closer to the lowest-bid or the second lowest bid. Whereas the second neural network is the Generalized Regression Neural Network (GRNN) which was used to approximately estimate the position ratio of the submitted bid that is closest to the actual project cost. Their results showed that the implemented neural network models were superior to the traditional policy (Lowest Bid Policy), for instance using the suggested neural network classification model and neural network regression model in selecting the best bid presented the percentages of cost overrun projects equal to 54.81% and 54.12% consequently, comparing to the percentage of 68.49% for using the traditional policy of selecting the lowest bid.

Researchers Lee, Kim, Park, Son, and Kim (2011) have used support vector machine (SVM) which is a data mining classification technique to generate a predictive model that can forecast project's cost performance, depending on the project definition level throughout the initial stages of the project before starting the detailed design. They have collected their data from 77 completed construction projects in Korea and performed an interview using a questionnaire with each project participant. The aim of the conducted questionnaire was to measure the influence of various project definition elements on the project performance. The questionnaire contained questions about project

cost and duration to measure project performance and used five-point Likert scale to measure the degree of project definition elements. They used Project Definition Rating Index (PDRI) to collect project definition information on project planning (Construction Industry Institute (CII), 1999; Cho and Gibson, 2001). The performance of multiple predictive models' other than SVM was studied and compared to emphasize the reason behind selecting SVM method in the research. Their results showed that using SVM for implementing the prediction model to forecast project cost performance provided the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) and the highest correlation coefficient.

Similarly, researchers such as Chi, Suk, Kang, & Mulva (2012), presented a data mining analysis framework to analyze multi-attribute data and predict project cost performance. In their research, they used two types of data mining algorithms which were ranking and classifying using the data mining software called WEKA (Waikato Environment for Knowledge Analysis) (Witten and Frank 2005). Then to validate their proposed model, a case study was conducted using data collected from 139 projects from the research unit within the University of Texas at Austin in the United States. Their aim was to study whether the increase use of information technology (IT) will affect the project cost performance or not. The proposed decision support framework used ranking algorithm to identify the key work functions used for project performance prediction. Then the designed decision tree eliminated the less correlated work functions and the final tree contained only 8 functions to be the predictors for cost performance. Their model showed good results since the selected decision tree produced 80% predication accuracy.

While other researchers such as Ahiaga-Dagbui & Smith (2014), used data for 1600 projects collected from the major water infrastructure client in the UK to validate their proposed predictive model. The predictive model was developed using artificial neural network (ANN) and had five significant input factors which were: purpose of project (wastewater, water or general), scope of project (new-build, upgrade or replacement), ground condition (contaminated or non-contaminated ground), delivery partner (anonymized as X, Y, Z) and estimated project duration. The developed model was used to estimate the final construction cost of water infrastructure projects. The results of Ahiaga-Dagbui & Smith (2014) proposed model were impressive as it was found that 92% of the 100 validation predictions were within ±10% of the actual final cost of the project with 77% within ±5% of actual final cost.

Another research was done by Williams and Gong (2014) that discussed how to combine text data which describes the construction project with numerical data to produce a prediction of the percentage of project cost overrun using data mining classification algorithms. The text data used in their research was a brief description of the project which consisted of two to three sentences. The researchers collected their data from 1221 competitively bid highway projects from California Department of Transportation websites which contained information about the low bid, the completed project cost, and the numbers of bidders. Two powerful softwares were used for mining the data, at the first stage they used WEKA data mining software then the output data was ported to the Rapid Miner system which is widely used as a data and text mining system. The average accuracy of the model was 43.72%. To investigate whether the text data contained useful information that can contribute to improve predictions of cost overrun, they have tested the model using numeric inputs only, the results showed poorer recall and precision results for each level of cost overrun which indicates that the tested text contained useful information as expected.

Hammad, AbouRizk and Mohamed (2014) stated that inappropriate management of labor resources in industrial construction projects is one of the main causes of cost overrun and schedule delays. So, to improve estimating practices for future utilization of labor resources they developed an integrated methodology through referring to the five-step knowledge discovery in data (KDD) model and using the concepts of data warehousing and data mining to extract useful knowledge from collected labor resources data in a multiple-project environment. They obtained their data for the case study from one of the largest structural steel fabricators in Canada, their dataset included 13,498 data points that contained actual durations and working hours for fabrication work packages. To perform the knowledge extraction, they used WEKA software and applied clustering technique on the testing data points. The results of the validation test presented that more than 80% of the tested data points had an estimating error below 25%, when comparing the results attained using the proposed model with the actual values which were collected from the company. Since the actual error in estimation at the company was exceeding 25% using their current estimation techniques and according to the obtained results by Hammad, AbouRizk and Mohamed (2014), it was proven that the proposed model provided a reasonably accurate method for estimating working hours and durations based on knowledge obtained from the historical data.

The literature which was reviewed previously have some limitations, since only two research papers (Soibelman and Kim, 2002); (Hammad, AbouRizk and Mohamed, 2014) presented a complete KDD model, whereas both researches

didn't address the cost overrun problem as they used KDD model to predict activity delays and estimate the utilization of labor resources, respectively. The other five researches didn't prepare their data sets by structuring a reliable data warehouse, as they used data mining algorithms (which is only a step of KDD model) to address the problem of cost overrun in construction projects. Also, their models didn't involve most of cost overrun factors which were identified in the section of cost overrun factors literature. For instance, Art Chaovalitwongse et al. (2011) prediction model included one factor only which was bid selection policy. Lee, Kim, Park, Son, and Kim (2011) prediction model included one factor only which was Project Definition Rating Index (PDRI). Chi, Suk, Kang, & Mulva (2012) prediction model included eight key work functions not cost overrun factors to predict whether the project's cost performance will improve when the use of information technology (IT) is increased. Ahiaga-Dagbui & Smith (2014) perdition model included five cost overrun factors. Williams and Gong (2014) prediction model included two cost overrun factors which were project cost and number of bidders.

3. Research Methodology

3.1 Analyzing factors of cost overrun

Many researchers such as Jahren and Ashe (1990); Kaming, Olomolaiye, Holt and Harris (1997); Koushki, Al-Rashid and Kartam (2005); Le-Hoai, Lee and Lee (2008); Shane, Molenaar, Anderson and Schexnayder (2009); Memon, Abdul Rahman, Abdullah, and Abdul Azis (2010); Jennings (2012); Park & Papadopoulou (2012); Abdul Rahman, Memon, Abdul Azis and Abdullah (2013); and Brunesa & Lind (2015) have identified the quantitative and qualitative factors of cost overrun in construction projects. In this research, we selected twelve factors that are measurable, their related data sets are usually collected by contractors and available in the historical data of their projects. These factors were nominated from the studies mentioned previously to inspect their effect on project's cost overrun. The data for the first nine factors is collected during the project's initiation stage where on the other hand the data for the remaining three factors is collected during both initiation and closeout stages. The twelve factors which will be used as the predictors for the proposed model are: 1- project type, 2- owner type, 3- contractor type, 4- bid selection policy, 5- contract type, 6- project location, 7- project complexity, 8- risk assessment value, 9- Project Definition Rating Index (PDRI), 10- project's duration, 11- project's cost, 12- project's variations.

Starting by the data that is collected during project's initiation phase, the first factor is the **project type**, types such as (Offshore/Onshore) or (Greenfield/Brownfield) might play a significant role in project's cost overrun (Ahiaga-Dagbui and Smith, 2014). Through applying a suitable data mining technique to the case study data set, the results may show the most popular project type that may have cost overrun.

Then the second factor is the **type of owner**, whether it is a government or private sector, so we will investigate if the projects that are related to government sectors usually experience cost overrun more likely than private ones or vice versa, also the level of familiarity of the owner with projects and the presence of owner representative might be good sub factors that may affect the project's cost overrun.

Third factor is the **type of contractor,** such as publicly traded company, privately owned company or government sector, by studying this factor we will try to explore whether there is a correlation between one of these types and project's cost overrun.

Then the forth factor which will be investigated is the **bid selection policy**, as the results of the survey which was conducted by Park and Papadopoulou (2012) indicated that the most significant and frequent cause of cost overrun in transport infrastructure projects in Asia is the practice of awarding contracts to the lowest bidder.

Also, according to the results of Park and Papadopoulou (2012) conducted survey, 64 percent of the survey participants reported that they faced cost overrun in lump-sum contracts, whereas 33 percent reported that they faced cost overrun in measurement contracts and 3 per cent in cost plus contracts. So, the fifth selected factor is **contract type** (lump sum/measurement/ cost plus) to highlight -if possible- the most popular contract type that may experience cost overrun.

Williams and Gong (2014) recommended in their future work section that **project location** might be a significant factor that require examination to differentiate between rural and urban projects and pinpoint if distance from the nearest urban area and easy access might affect cost overrun, so it is the sixth selected factor in our study. The historical data for all the above-mentioned six factors are simply found in the project initiation form (PIF).

Shane, Molenaar, Anderson and Schexnayder (2009) and Kaming, Olomolaiye, Holt and Harris (1997) concluded in their researches that **project complexity** is considered as an important cost overrun factor, accordingly it is selected as our seventh factor, since companies usually predict the complexity level during the estimation or initiation stage as per the available information and documents, hence it might be important to find if there is a correlation between high or low project complexity level value and cost overrun.

Also, to prepare a reliable estimate of final cost, one of the important factors that shall be taking into consideration is risk and uncertainty level (Ahiaga-Dagbui and Smith, 2014). So, the eighth factor which will be investigated is the **risk assessment value** for each project since it is required to calculate the expected risk rate and implement a management plan during the initiation phase to mitigate risk. As Akintoye (2000) stated that the inability to accurately identify and quantify risk might be a reason behind the failure of many projects to meet initially set cost limits.

The next investigated factor is a document which might be available in the project's historical data, it is called the **Project Definition Rating Index (PDRI)**, PDRI presents a method that is used to define the status of early planning, a previous research which was done by Lee, Kim, Park, Son, and Kim (2011) used this index along with other data as an input to their model to predict the rate of cost overrun for the project during the planning phase.

After concluding the analysis for the data that is collected during initiation stage as mentioned earlier, the data that is related to the remaining three factors for cost overrun which is collected during both initiation and closeout stages will be analyzed. Researchers Abdul Rahman, Memon, Abdul Azis and Abdullah (2010) (2013), concluded in their both researches that schedule delay is one of the significant factors that affects projects cost overrun. Also, Shane, Molenaar, Anderson and Schexnayder (2009) found that changes in the project schedules is causing cost overrun too. Consequently; we selected **project duration** as the tenth factor, which will be investigated in two levels, first level will be done by classifying the project's duration into 3 classes and explore whether the long duration of the project increases the rate of cost overrun or vice versa. The second level will be done by studying the effect of the delay of project's commencement date, to check if there is a correlation between this delay and cost overrun.

According to Kaming, Olomolaiye, Holt and Harris (1997) cost increases due to inflation, and inaccurate material estimating. Similarly, Koushki, Al-Rashid and Kartam (2005) stated that cost overrun occurs at projects due to material-related problems. Later, researchers Le-Hoai, Lee and Lee (2008) also declared that material and labor related factors are responsible for project's cost overrun. In year 2010, Memon, Abdul Rahman, Abdullah, and Abdul Azis (2010) explained in their research that fluctuation in prices of material and shortage of site workers are causing cost overrun while in year 2013 they added other factors such as high cost of labors. To conclude the five researches mentioned above it was found that **project's cost** – our eleventh's inspected factor - is the most important and main factor in this research, this factor will be investigated in two levels: project level and work package level and in two stages: current cost during initiation and actual cost at the project's closeout. To deeply study the project's cost factor, it will be categorized into three categorize: material, labor and equipment cost, as the existence of cost overrun might be due to changes in prices and/or quantities of material, labor and equipment.

Finally, the last factor is **project variations**, as per researchers Kaming, Olomolaiye, Holt and Harris (1997), Le-Hoai, Lee and Lee (2008), Memon, Abdul Rahman, Abdullah, and Abdul Azis (2010) and Brunesa & Lind (2015) design changes is considered as an important factor which have impact on project's cost overrun. So, to study the effect of this factor we need to classify the variation amount into three classes (Low - Medium - High), calculate the percentage of variations to the actual cost of the project and analyze the relation between the defined classes, percentages and the project's cost overrun.

Identifying the twelve factors of cost overrun was done for a significant reason since our research will be extended to enable implementing a prediction model that uses the data related to these twelve factors as decision variable inputs whereas the output of the model will be a prediction for project's cost performance, the proposed prediction model will analyze the input data using a suitable data mining technique. Moreover, the model is supposed to support the decision support system in construction companies that usually carry out the detailed work in Engineering, Procurement and Construction (EPC). Accordingly, the investigated factors are entirely related to Engineering, Procurement and Construction (EPC) work categories, each factor may belong to one or more work category/ categories. Table 1 provides a summary of the twelve cost overrun factors, the related category/categories and the reason behind investigating these factors in the proposed model.

3.2 Building the cost overrun data warehouse

The twelve-selected cost overrun factors provided the main keys that form the schema of the data warehouse. This data warehouse will be used to organize, prepare and summarize cost overrun data collected from construction projects and will keep the data ready for the next step which is the analysis process using a suitable data mining technique. The proposed data warehouse is recommended to be used by construction companies as it offers a simple and user-friendly interface since it is designed using Microsoft Access which is widely used as an efficient database management system.

The data in our research will be collected in two levels: 1- Project Level and 2- Work Package Level. Also, it will be collected in two stages: during project initiation and closeout. The first step in building the data warehouse is defining the data and its related meta data. Since the star schema is selected to describe the relationships between the collected data, one fact table and seven dimension tables were defined to form the data warehouse. Star schema is considered a simple database type as the number of joins are less, so retrieving the data will be easier and faster which means higher efficiency when compared with the other two major schemas snowflake and galaxy (Cios, Pedrycz, Swiniarski and Kurgan, 2007). The dimension tables are representing our meta data such as project type, project (current & actual cost), project location and project duration, etc. Figure 1 represents data warehouse design and details.

Table 1. Cost overrun factors description

No.	Cost Overrun Factor	Category	Selection Reason
1	Project type	Engineering & Construction	To find the most popular project type (Offshore/Onshore) or (Greenfield/Brownfield) that encounters cost overrun.
2	Owner type	Engineering	To find the most popular owner type (Public/ Private sector) that encounters cost overrun.
3	Contractor type	Engineering & Construction	To find the most popular contractor type (publicly traded company, privately owned company or government sector) that encounters cost overrun.
4	Bid selection policy	Procurement	To find if the projects which their contracts are awarded to the lowest bidder encounter cost overrun more than other bidding policies.
5	Contract type	Procurement	To find the most popular contract type (lump sum/ measurement/ cost plus) that encounters cost overrun.
6	Project location	Engineering	To find the most popular project location (rural/ urban) that encounters cost overrun.
7	Project complexity	Engineering	To check if there is a relation between project complexity and cost overrun.
8	Risk assessment value	Engineering	To check if there is a relation between risk and uncertainty level value and cost overrun.
9	Project definition rating index (PDRI)	Engineering	To check if there is a relation between PDRI value and cost overrun.
10	Project duration	Engineering	To check if there is a relation between project duration category (short /medium/ long) or schedule delay with cost overrun.
11	Project cost	Engineering & Procurement	To check if there is a relation between project cost category (low/ medium/ high) with cost overrun. Also, to find the most popular project cost category (low/ medium/ high) or type (material/ labor/ equipment) that encounters cost overrun.
12	Project variations	Engineering, Procurement & Construction	To check if there is a relation between the percentage of variations to the actual project cost (low/ medium/ high) with cost overrun.

Conclusion

In this research, factors of cost overrun in construction projects which is a current worldwide problem were intensively studied, then twelve measurable factors were investigated and selected to form the basis for developing a data

acquisition model that enable capturing the relevant historical data and meta data in a data warehouse. The developed data warehouse would enable the implementation of KDD model and facilitate the process of extracting the hidden knowledge using a suitable data mining technique. This KDD model would be applied in modern construction management as a key business tool to utilize the huge amount of daily generated data, extract the embedded knowledge and transform it into decision support systems. This model may supplement the traditional estimation methods and provide more reliable final cost forecasting to overcome the cost overrun problem. For future work, this research will be extended through implementing a prediction KDD model that uses a suitable data mining technique to analyze the cost overrun real historical data from construction projects at U.A.E., the tested data shall contain projects' cost performance and the relevant data and meta data from the twelve suggested quantitative factors responsible for cost overrun.

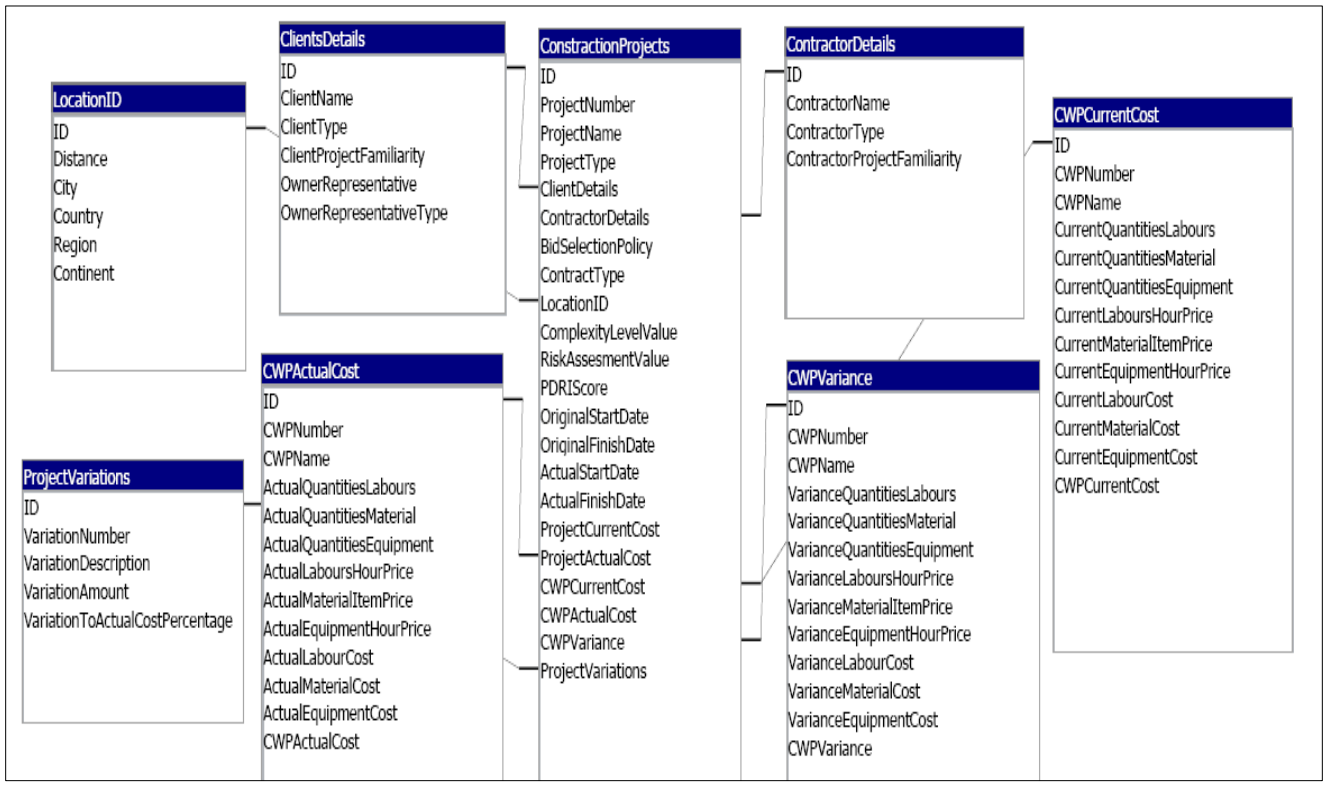

Figure 1. Data warehouse for data related to cost overrun factors

References

Abdul Rahman, I, Memon, A.H, Abdul Azis, A. A, and Abdullah, N.H. (2013). Modeling Causes of Cost Overrun in Large Construction Projects with Partial Least Square-SEM Approach: Contractor's Perspective. Research Journal of Applied Sciences, Engineering and Technology 5(6): 1963-1972.

Ahiaga-Dagbui, D. D., and Smith, S. D., (2014). Dealing with construction cost overrun using data mining. Construction Management and Economics, 32 (7-8), 682-694.

Akintoye, A. (2000). Analysis of factors influencing project cost estimating practice. Construction Management and Economics, 18(1), 77–89.

Art Chaovalitwongse, W., Wang, W., Williams, T. P., & Chaovalitwongse, P. (2011). Data mining framework to optimize the bid selection policy for competitively bid highway construction projects. Journal of Construction Engineering and Management, 138(2), 277-286.

Brunesa, F. & Lind, H. (2015) Explaining cost overrun in infrastructure projects: a new framework with applications to Sweden. Journal of Construction Management and Economics.

Chan, A.P. & Chan, A.P. (2004). Key performance indicators for measuring construction success. Benchmarking: An International Journal, 11(2), 203–21.

Chi, S., Suk, S. J., Kang, Y., & Mulva, S. P. (2012). Development of a data mining-based analysis framework for multi-attribute construction project information. Advanced Engineering Informatics, 26(3), 574-581.

Cho, C.S. & Gibson, G.E. (2001). "Building Project Scope Definition Using Project Definition Rating Index", Journal of Architectural Engineering, (4), 115–125.

Cios, K.J., Pedrycz, W., Swiniarski, R.W., Kurgan, L. (2007). Data Mining a Knowledge Discovery Approach. Springer US.

Construction Industry Institute (CII). (1999), Pre-Project Planning Tool: PDRI for Buildings, Research Summary, 155–1, Austin, TX.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996). Advances in Knowledge Discovery and Data Mining. AAAI Press.

Hajirezaie, M., Husseini, S., Barfourosh, A. & Karimi, B. (2010). Modeling and evaluating the strategic effects of improvement programs on the manufacturing performance using neural networks. African Journal of Business Management, 4(4), 414–424.

Hammad, A., AbouRizk, S., Mohamed, Y. (2014). Application of KDD Techniques to Extract Useful Knowledge from Labor Resources Data in Industrial Construction Projects. Journal of Management in Engineering, 30(6).

Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques, Morgan Kaufmann, Elsevier Science distributor, San Francisco, Oxford, U.K.

Hegazy, T. (2002). Computer-Based Construction Project Management, Prentice Hall, Upper Saddle River, NJ.

Jahren, C. T. & Ashe, A. M. (1990). Predictors of cost-overrun rates. Journal of Construction Engineering and Management, 116(3),548-552.

Jennings. (2012). Why costs overrun: risk, optimism and uncertainty in budgeting for the London 2012 Olympic Games. Construction Management and Economics, 30(6), 455–462.

Kaming, P. F., Olomolaiye, P. O., Holt, G. D. & Harris, F. C. (1997). Factors influencing construction time and cost overrun on high-rise projects in Indonesia. Journal of Construction Management and Economics, 15, 83-94.

Koushki, P. A., Al-Rashid, K. and Kartam, N. (2005) Delays and cost increases in the construction of private residential projects in Kuwait, Construction Management and Economics, 23:3, 285-294.

Lee, S., Kim, C., Park, Y., Son, H., & Kim, C. (2011). Data Mining-Based Predictive Model to Determine Project Financial Success Using Project Definition Parameters. International Association for Automation and Robotics in Construction (IAARC), 473-478.

Le-Hoai, L., Lee, Y. D., and Lee, J. Y. (2008) Delay and Cost Overrun in Vietnam Large Construction Projects: A Comparison with Other Selected Countries. Journal of Civil Engineering, 12(6):367-377.

Liebowitz, J., and Megbolugbe, I. (2003). "A set of frameworks to aid the project manager in conceptualizing and implementing knowledge management initiatives." Int. J. Proj. Manage., 21(3), 189–198.

Memon, A.H., Abdul Rahman, I., Abdullah, M.R, and Abdul Azis, A.A. (2010). Factors affecting construction cost in Mara large construction project: Perspective of project management consultant. Int. J. Sustain. Constr. Eng. Technol., 1(2).

Nassar, K. M., Nassar, W. M., and Hegab, M. Y. (2005). "Evaluating cost overrun of asphalt paving project using statistical process control methods." J. Constr. Eng. Manage., 131(11), 1173–1178.

Nicholas, J.M. (2004) Project Management for Business and Engineering: Principles and Practice, 2nd edn, Elsevier Butterworth-Heinemann, Oxford.

Olawale, Y.A. and Sun, M. (2010), "Cost and time control of construction projects: inhibiting factors and mitigating measures in practice", Construction Management and Economics, Vol. 28 No. 5, pp. 509-26.

Park, Y. I., & Papadopoulou, T. C. (2012). Causes of cost overrun in transport infrastructure projects in Asia: their significance and relationship with project size. Built Environment Project and Asset Management, 2(2), 195-216.

Shane, J., Molenaar, K., Anderson, S., & Schexnayder, C. (2009). "Construction Project Cost Escalation Factors." Journal of Management in Engineering, 25, 221-229.

Turban, E., Aronson, J.E., Liang, T., and Sharda, R. (2007) Decision Support and Business Intelligence Systems (8th Edition). Prentice Hall Press Upper Saddle River, NJ, USA.

UM Libraries (1986), Getting Organized: Introduction to Database Management Systems on MTS, University of Michigan Computing Center. Retrieved Nov. 5, 2016 from the World Wide Web: https://babel.hathitrust.org/cgi/pt?id=mdp.39015014920717;view=1up;seq=7

Williams, T. P. (2005). "Bidding ratios to predict highway project costs." Eng., Constr., Archit. Manage., 12(1), 38–51.

Williams, T. P., & Gong, J. (2014) Predicting construction cost overrun using text mining, Automation in Construction, 43, 23–29.

Witten, I.H. & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco.

Yun, S., & Caldas, C. H. (2009). Analyzing decision variables that influence preliminary feasibility studies using data mining techniques. Construction Management and Economics, 27(1), 73-87.

Zaiane, O. R., Foss, A., Lee, C-H, and Wang, W. (2002). On data clustering analysis: Scalability, constraints, and validation, Springer, Berlin, 28–39.

Soibelman, L., & Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. Journal of Computing in Civil Engineering, 16(1), 39-48.

Biographies

Mai Ghazal is currently a Senior Electrical Engineer in a reputed construction company at Dubai - United Arab Emirates (UAE). She has 11 years of experience in designing and managing high rise residential and commercial projects executed at U.A.E. Eng. Mai holds a bachelor of Science degree in Electrical and Electronics Engineering from University Of Sharjah, at present she is a master student at the Engineering Management and Innovation Program at University Of Sharjah and she is preparing her research which involves multi topics such as Application of Data Warehousing, Data Mining, Knowledge Discovery in Data (KDD) and Decision Support Systems (DSS) in construction project management.

Ahmed Hammad is currently an Assistant Professor in the Department of Industrial Engineering and Engineering Management (IEEM) at University of Sharjah, United Arab Emirates (UAE). Prior to joining Academia, Dr. Ahmed has 25 years of industrial experience in managing mega multi-billion dollars Engineering, Procurement & Construction (EPC) projects. These projects were planned and executed between: Egypt, Canada, United States of America (USA), Australia, South Korea and UAE. His experience covers infrastructure, commercial and industrial EPC projects. His research interests include: Application of Data Warehousing, Data Mining, Knowledge Discovery in Data (KDD) and Artificial Intelligence (AI) in construction project management. Other research interests include Multi-Criteria Decision Making (MCDM), Decision Support Systems (DSS), Building Information Modeling (BIM), Environmental and Social aspects in project management. Dr. Ahmed is a Professional Engineer (P.Eng.) and Project Management Professional (PMP).