EPiC Series in Built Environment
Volume 3, 2022, Pages 192–199
ASC2022. 58th Annual Associated Schools of Construction International Conference

Using Historical Bid Data for Enhanced Conceptual Estimating

Ahmed Abdelaty, PhD
University of Wyoming
Laramie, Wyoming

Kevin Nesselhauf
L. Keeley Construction
St. Louis, Missouri

Nathan Munie
L. Keeley Construction
St. Louis, Missouri

Several small and medium-size contractors store bid day data regarding potential projects creating large datasets of bid day information without meaningful utilization. Many of these companies fail to leverage the archived bid day data because of their format or lack of effort to use historical data. Thus, most conceptual estimates are done using personal judgment and experience with little to no historical data support. Because of this approach, many small and medium-sized companies lack a data-driven approach to develop conceptual estimates. As such, this study aims at leveraging historical bid data to build a data-driven approach for creating conceptual estimates. This objective is achieved by presenting a framework for one company's historical bid day data to develop a conceptual cost estimating model. The framework uses bid day data for the past 45 years to build a data-driven conceptual estimating model using a case-based reasoning approach. The model allows estimators to retrieve the most similar projects from a historical database to create an informed conceptual estimate for potential projects. It is expected that this research will help many small and medium-size contractors leverage their historical bid data by utilizing it.
Key Words: Conceptual estimating, project comparison, data analytics, preconstruction services

Introduction

The American Association of Cost Engineers (AACE) classifies cost estimates into five distinct classes in which class 5 estimate is used for concept screening, and class 1 is used for bidding purposes. The level of accuracy for each cost estimate class is expected to increase with more information regarding the project definition and deliverables (AACE RP 18R-97 2020). When project estimators approach a potential project, they would typically develop a conceptual estimate before creating a detailed cost estimate for screening purposes. This process will help project estimators determine whether a project is feasible for bidding. As such, developing a reliable conceptual cost estimate to evaluate projects from a bidding feasibility perspective is crucial. Most construction firms store a vast amount of historical bid data, which can generate conceptual estimates for future projects. Although many companies realize the importance of leveraging historical data for estimating purposes, many find it challenging because of their business scale, the way the data is stored, or their willingness to spend resources on development efforts. Therefore, many small and medium-sized firms have not fully utilized their historical cost data continuing to rely on their personnel experience to develop conceptual estimates for projects. This study aims to demonstrate how small and medium-sized construction firms can utilize historical bid data to assist project estimators in developing conceptual estimates. This objective is achieved by creating a user-friendly tool to help establish conceptual estimates using the project comparison method. The tool is developed based on historical bid day data for hundreds of projects.

First, the authors compiled historical bid day data estimates to form one database that contains historical bid data. Most projects' bid data were stored in a spreadsheet containing the cost for each division and general information about the project (i.e., size, duration, profit, and contingency). Second, the authors compiled the bid day data for all projects using Python to create one unified spreadsheet. The spreadsheet contains all relevant information such as bid date, project location, bid total, amount of general condition, profit, contingency, project size in square feet, construction duration, market type, and project type. After compiling all bid day data in one spreadsheet, the authors used the project comparison method to develop the conceptual estimating tool. The project comparison method compares the attributes of a new project against the attributes of historical projects to produce a similarity score. The similarity score is calculated for each project in the database and considers four matching attributes: market type, project type, project size, and project duration. The similarity score is then used to rank all historical projects based on their similarity to a new project. Then, project estimators can easily retrieve the data for the most similar historical project to develop an accurate conceptual estimate using the project comparison methods. Finally, two domain experts validated the developed tool by retrieving similar projects for multiple hypothetical projects. It is expected that this paper will benefit a significant sector of small and medium-sized contractors to utilize historical data to develop cost estimating models actively.

Literature Review

Several research studies have focused on estimating construction costs, whether conceptually or detailed. For example, Kim et al. (2012) developed a hybrid conceptual cost estimating model to estimate the construction costs for mixed-used buildings. The conceptual estimating model presented uses two methods: assembly-based and historical data-based estimating to increase the accuracy of the estimating model. The study focuses mainly on improving conceptual cost estimates for mixed-used buildings since not much historical data is available on this type of project. The model also considers the estimator's role in the estimating process by giving them the flexibility to select the most similar historical projects to develop the cost estimate. However, there are some limitations of the model created, such as the number of assemblies included in the model and historical data available.

Choi et al. (2014) also developed a conceptual cost estimating model for public road planning using case-based reasoning, rough set theory, and genetic algorithms. The goal of the model developed is to help government officials estimate the cost of public road work at early stages during the planning phase for budget allocation and investment decisions. The study used data for 207 projects with 17 attributes (e.g., type of project, road length, number of lanes, etc.) to create the conceptual estimating model. Similar to other conceptual estimating models, the authors stressed the lack of historical data and limited project definition to develop a conceptual cost estimate. Thus, the authors aimed at developing conceptual cost estimating to estimate the cost of road projects accurately. It is worth noting that conceptual cost estimating models are inherently inaccurate because of the level of definition of the project. Thus, the model developed aimed at improving the accuracy of conceptual estimates to better estimate project contingencies.

Mahamid (2011) also developed a conceptual estimating model to estimate the cost of road construction using multiple regression techniques. The model was developed using data from 131 road construction projects. Intuitively, the author indicated that models that use bid quantities generate more accurate results than models that use project attributes such as road length, number of lanes. Zima (2015) also developed a conceptual estimating model using fuzzy case-based reasoning to estimate the costs of sports facilities. The model uses criteria such as field type, the quantity of work, other attributes related to the sports facility.

Abdelaty et al. (2020) developed multiple estimating models to predict the cost of preconstruction services for bridges. The authors used artificial neural networks, regression analysis, and case-based reasoning to predict the engineering hours and consultant's fees for preconstruction services. The authors looked at possible 33 bridge design attributes but only used 15 design attributes because they were determined to be well-known during the planning phase. The prediction model is built based on historical data for 67 projects. The study concludes that statistical methods such as neural networks and regression analysis provide practitioners meaningful insights. However, conceptual estimates are inherently inaccurate, and historical data may be inconsistent with generating reliably statistical prediction models. As such, the authors suggested that a case-based reasoning model may be more effective in helping project estimators develop conceptual estimates rather than using statistical methods with a high margin of error.

Data Collection

Bid day data for almost 500 projects spanning between 1975 and 2019 were collected from one construction company. Most projects included in the development of the model were successfully awarded. However, because the data ranged for more than 45 years, it was difficult to determine if specific projects were awarded. Each bid day datasheet contains information regarding the project as follows:

- Project name
- Location (i.e., city and state)
- Type of the project (i.e., new construction, tenant improvement, tenant finish, addition)
- Total area in square foot
- Owner
- Bid date
- Total bid amount
- Percentage of general conditions
- Profit
- Risk
- Duration in months
- Award status (i.e., whether the project was awarded)

Since the data spanned approximately 45 years, the bid day datasheet had different formats, which is challenging to compile all this together in one database. Therefore, a Python script was developed to retrieve project parameters from bid day datasheets and store them in one spreadsheet. The script loops through all the historical bid day data files. Afterward, the script reads each file to extract the attributes described earlier. Finally, it writes the project attributes in a separate spreadsheet to compile the historical bid data.

The final compiled spreadsheet contains bid date, project location, bid total, amount of general condition, profit, contingency, project size in square feet, construction duration, market type, and project type. The market type attribute is a new parameter introduced to classify projects into nine categories: banking, commercial, education, gas station, healthcare, industrial, municipality, restaurants, and retail. The project type attribute classifies projects into three main categories: new construction, addition, and tenant improvement/finish. Finally, the award status attribute indicates whether the contractor was awarded the project. This attribute provides the estimator with more confidence when developing a conceptual estimate.

Methodology

The research methodology used in this study is case-based reasoning. In case-based reasoning, every new project is compared against the historical projects based on a similarity score. The similarity score consists of four terms representing four matching attributes which are listed as follows:

- Market type
- Project type
- Project area
- Project duration

Domain experts selected these four attributes because they are readily available information for any new project and can be easily retrieved with little effort. For each matching attribute, a numeric similarity score is calculated. For example, the market type similarity score is binary, meaning that every historical project will receive a score of one point if the project's market type matches the market type of the project under study. The same concept applies to the project type attribute. However, project size and duration are continuous variables. Because of that, the matching scores of project size and construction duration are calculated using multiple steps. First, the difference between the duration or size of the project under study and the historical project's duration or size is calculated, respectively. Then, based on the difference in project size and duration, historical projects are ranked in ascending order, in which projects with the lowest difference are ranked first. Finally, based on the rank for each historical project, a normalized score with a value between zero and one is computed. Projects ranked first with receiving the highest normalized score, while projects ranked last will receive no points.

A weighted total matching score is computed after the scores are calculated for each matching attribute. The total matching score is calculated by summing up the attribute weight of importance multiplied by its matching score. Finally, historical projects are ranked based on the total matching score, and the most similar historical projects are retrieved to develop conceptual estimates. Users can then use the retrieved historical data to develop a more accurate conceptual estimate after adjusting the bid value for time and location. The remainder of this section explains the detailed procedure set to calculate the project similarity score.

The first term is the market type score (MTS_i) compares the market type of a new project to the market type of every historical project in the database. Equation 1 calculates the market type similarity score as follows:

$$MTS_i = \begin{cases} 1, & MT_v = MT_{iv} \\ 0, & MT_v \neq MT_{iv} \end{cases} \quad (1)$$

Where MTS_i is the market type score for project i in the database and MT_v is the market type value for the project under study and MT_{iv} is the market type value for project i in the database. As explained earlier in this section, the MTS_i is a binary variable. For example, if two projects share the same market type, then a score of one will be assigned. On the other hand, if two projects have two different market types, a score of zero will be given.

Similarly, the second term of the project similarity score is the project type score (PTS_i) which is calculated according to equation two as follows:

$$PTS_i = \begin{cases} 1, & PT_v = PT_{iv} \\ 0, & PT_v \neq PT_{iv} \end{cases} \qquad (2)$$

Where PTS_i is the project type score for project i in the database and PT_v is the project type value for the project under study and PT_{iv} is the project type value for project i in the database. The PTS_i is also a binary variable with possible values of zero or one.

The third term of the similarity score is the project size score (PS_i), which is calculated according to a two-step process as shown in equations three and four. The first step calculates the absolute difference between the square footage of a new project and every single project in the database. The square footage difference is computed according to equation three as follows:

$$PS_{di} = |PS - PS_{\mathrm{i}}| \qquad (3)$$

Where PS_{di} is the absolute value of the project size difference for project i, PS is the project size in square feet for the project under study, and PS_{i} is the project size in square feet for the project i in the database. The second step includes ranking projects using the PS_{di} in ascending order. For example, the lowest project size difference is ranked first, and the highest is ranked last. This is done in ascending order to projects with the least difference receive the highest similarity score. The rankings are then normalized to calculate a project size score which is a continuous variable with any value between zero and one. The project size score is calculated as follows:

$$PS_i = \frac{RS_i - \min(RS_1, \ldots, RS_i)}{\max(RS_1, \ldots, RS_i) - \min(RS_1, \ldots, RS_i)} \qquad (4)$$

Where PS_i is the project size match score for project i in the database, RS_i is the project rank based on PS_{di}. Projects with the lowest size difference receive the first rank. Similarly, a project duration score is calculated using a two-step process, as shown in equations five and six.

$$PD_{di} = |PD - PD_{\mathrm{i}}| \qquad (5)$$

Where PD_{di} is the absolute value of the project duration difference for project i, PD is the project duration in months for the project under study, and PD_{i} is the project duration in months for the project i in the database

$$PD_i = \frac{RD_i - \min(RD_1, \ldots, RD_i)}{\max(RD_1, \ldots, RD_i) - \min(RD_1, \ldots, RD_i)} \qquad (6)$$

Where PD_i is the project duration match score for project i in the database, RD_i is the project rank based on PD_{di}. Projects with the lowest duration difference receive the first rank. If the difference PD_{di} has the same value for multiple projects, the same rank is assigned to these projects

Finally, the model calculates a total weighted similarity score (SS_i) based on the four scores and the weights of importance of each score. The SS_i is calculated as follows:

$$SS_i = MTS_i \times W_{MT} + PTS_i \times W_{PT} + PS_i \times W_{PS} + PD_i \times W_{PD} \quad (7)$$

Where SS_i is the total weighted matching score for project i in the database;
W_{MT} is the weight of importance of market type;
W_{PT} is the weight of importance for project type;
W_{PS} is the weight of importance for project size; and
W_{PD} is the weight of importance of project duration.
The total score of W_{MT}, W_{PT}, W_{PS}, and W_{PD} must equal 100%.

Conceptual Estimating Model

The methodology described in the previous section is implemented by using data from one construction company. Historical bid day data were cleaned and compiled using Python script, and the estimating model was built using Microsoft Excel. Figure 2 shows the user interface once they open the conceptual estimating model.

	Value	Weight of Importance
Market Type*	Retail	25%
Project Type*	New Construction	25%
Project size (sf)	100000	25%
Project duration (mo.)	5	25%

** All dollar values in this workbook are not adjusted for inflation. Bid value and other monetary figures corresponds to the value of money on the project's bid date.

*Unspecified market or project type means that type was not recorded in LKC bid day worksheet

Figure 2. Conceptual estimating model user interface

Estimators are required to enter the values for all four parameters: 1) market type, 2) project type, 3) project size in square feet, and 4) project duration in months. Afterward, estimators can change the default weight of importance for each parameter to increase or decrease the influence of one parameter on the project retrieval process. The estimating model retrieves the most similar 20 projects based on the matching attributes and their weight of importance for each historical project. The following project attributes are provided to assist project estimators in building their conceptual estimates.

- Project rank
- Project number
- Filename
- Year
- Project name
- Location
- Project and market type
- Owner
- Bid Date
- Bid Total
- General conditions amount
- Markup
- Gross profit
- Project size in square foot

- Construction duration in month
- Total bid per square foot
- Percentage of general conditions
- Markup and profit percentage
- The risk or contingency percentage
- General conditions amount per month
- Markup or profit per month

The retrieved attributes assist project estimators in developing a quick conceptual cost estimate using the project comparison method. In case project estimators need more data about one of the similar historical projects, they can retrieve the complete information of the project (i.e., estimate per division) using the filename attribute.

Conclusions

This paper presents a framework for small and medium-sized construction firms to utilize historical cost estimating data to develop an efficient and reliable way for conceptual estimating. The paper uses historical data from one construction company to create a conceptual estimating model using project comparison. First, bid day data for almost 500 projects were combined using a Python script in one database. Second, the authors developed an algorithm that retrieves similar projects based on market type, project type, project size, and construction duration. Third, the algorithm calculates a similarity score between the new project and every project in the historical database for a new project under study. Fourth, historical projects are ranked from the most similar to the least similar projects using the computed similarity scores. Finally, the algorithm retrieves the most similar 20 projects, which can then be used to develop a conceptual estimate by the project estimators. It is worth noting that project costs stored in the spreadsheet are not adjusted to present value. As such, project estimators should adjust the actual costs after retrieving the most similar projects.

Two domain experts validated the model by using the spreadsheet several times to retrieve historical data. For every spreadsheet run, the domain experts compared the historical data retrieved by the model to what they thought would be similar projects. The validation process was done to try different project types, market types, construction duration, and project size. It is observed that the retrieval process was significantly beneficial to the contractor because of the following:

- Significant time saving on retrieving similar projects.
- Having a methodological approach to retrieve projects instead of relying on personal judgment.
- The ability to see how retrieved historical projects are different compared to the project under study.
- The ability to see the project bid information quickly, such as percentage of general conditions, contingency, and profit.

The research presented in this paper paves the way for many small and medium-sized construction companies to utilize historical data to develop data-driven conceptual estimates instead of relying on personnel experience. The framework and model presented significantly boost the productivity of project estimators by automating the historical project retrieval process instead of relying on personnel memory and experience. Additionally, the project retrieval process presented in this paper is more consistent and reliable when compared to manual retrieval. There is a lot of potentials to improve the efficiency and reliability of preconstruction tasks performed by small and medium-sized companies. Future research should focus on increasing the efficiency of these construction firms that represent a significant portion of the construction industry. Many of these companies can move from personnel experience-based decisions to data-driven decisions. This will only be achieved through the utilization of historical data.

References

AACE Technical Board, AACE Recommended Practice No. 18R-97, Cost Estimate Classification System. AACE International.

Abdelaty, A., Shrestha, K. J., & Jeong, H. D. (2020). Estimating Preconstruction Services for Bridge Design Projects. Journal of Management in Engineering, 36(4), 04020034. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000793

Choi, S., Kim, D. Y., Han, S. H., & Kwak, Y. H. (2014). Conceptual Cost-Prediction Model for Public Road Planning via Rough Set Theory and Case-Based Reasoning. Journal of Construction Engineering and Management, 140(1), 04013026. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000743

Kim, H.-J., Seo, Y.-C., & Hyun, C.-T. (2012). A hybrid conceptual cost estimating model for large building projects. Automation in Construction, 25, 72–81. https://doi.org/10.1016/j.autcon.2012.04.006

Mahamid, I. (2011). Early cost estimating for road construction projects using multiple regression techniques. The Australasian Journal of Construction Economics and Building, 11(4), 87–101. https://doi.org/10.3316/informit.772369219553131

Zima, K. (2016). Fuzzy case-based reasoning in sports facilities unit cost estimating. AIP Conference Proceedings, 1738(1), 200008. https://doi.org/10.1063/1.4951980