Machine Learning-based Life-cycle Cost Analysis for Educational Facilities

Xinghua Gao, Pardis Pishdad-Bozorgi, Ph.D., Dennis R. Shelden, Ph.D., AIA, and Shu Tang
Georgia Institute of Technology
Atlanta, GA, USA

A large amount of resources are spent on constructing new facilities and maintaining the existing ones. The total cost of facility ownership can be minimized by focusing on reducing the facilities life-cycle costs (LCCs) rather than the initial design and construction costs. In recent years, with the developments of machine learning in predictive analytics and the building systems that provide ubiquitous sensing and metering devices, new opportunities have emerged for Architecture, Engineering, Construction, and Owner-operated (AECO) professionals to obtain a deeper level of knowledge on buildings’ and their systems’ LCCs. This study investigates the feasibility of obtaining an accurate forecast of facilities’ LCCs during the programming phase by implementing machine learning on historical facility data of similar projects. We propose a generalizable LCC analysis framework that specifies the data needs, the data acquisition process, and the machine learning-enabled cost components derivation procedure. The data on 121 educational facilities have been collected and analyzed to demonstrate the viability of this framework.

Keywords: Machine learning, Life-cycle Cost Analysis (LCCA), data mining, cost prediction

Introduction

Because of the long life spans of buildings, robust decisions regarding the economic efficiency of alternative materials, components, and systems demand a full lifecycle perspective that goes beyond the initial cost and regular maintenance and repair (Noshadravan et al., 2017). The Life Cycle Cost Analysis (LCCA) has become increasingly important in new building design and existing building retrofitting, refurbishment, and renovations. However, the real service lives and costs of many buildings and their systems are difficult to predict for multiple reasons. One is that there is always a mismatch between the predicted energy performance of buildings and actual measured performance, typically addressed as “the performance gap” (De Wilde, 2014). Another reason is that many building systems and components, with proper maintenance and repair, can function beyond the warranty, which makes their true costs difficult to predict because the facility owners typically do not know how much money and labor is needed to repair them when malfunction after the warranty expires. Moreover, even the same type of systems used in different buildings may have different LCCs because the monetary and labor costs vary depending on each facility manager’s operational profile.

Machine learning is an automated process that extracts patterns from data (Kelleher et al., 2015). In the field of predictive data analytics, machine learning is a method used to devise complex prediction algorithms and models (Mitchell, 1997; Kelleher et al., 2015). These analytical models enable data analysts to uncover hidden insights, predict future values, and produce reliable, repeatable decisions through learning from historical relationships and trends in the data (SAS, 2018). With the networks of sophisticated sensors and devices, building systems – Computerized Maintenance Management System (CMMS), Building Automation System (BAS), etc. – are generating extensive data, such as utility consumption, maintenance work order history, etc., a portion of which is potentially valuable for facility LCCA. The developments of machine learning techniques and more advanced building systems provide building experts with new opportunities to achieve more accurate predictions of facility-related costs.

The scope of this study is to investigate the feasibility of utilizing the historical data housed in heterogeneous building systems of a multi-facility entity to predict the LCCs of its facilities through machine learning. We propose a LCCA approach that collects the data generated by different building systems and analyzes the data with machine learning.
learning techniques to predict the future costs of the organization’s new and existing facilities, and thus to achieve better decision-making in building design, retrofitting, refurbishment, and renovations. The specific research objectives involve: 1) proposing a generalizable LCCA framework that specifies the data needs, the data acquisition process, and the machine learning-enabled cost components derivation procedure; and 2) conducting a case study to validate this framework and demonstrate the machine learning implementation process.

**Method**

We first discuss the components of building LCC based on the literature review and then propose a process to acquire LCC-related data – design and construction costs, utility consumptions, maintenance work orders, etc. – from multiple building systems. After that, each component of LCC is derived from the historical data using machine learning. We conducted a case study on a university campus to demonstrate the proposed LCCA framework. The commonly used descriptive features of each prediction model – the initial design and construction cost model, the utility consumption model, and the operation and maintenance model – are preliminarily identified from the literature and then, in the case study, the feature selection process is conducted based on the historical data to determine the features that significantly affect the prediction results. Multiple machine learning methods are tested to determine the best ones for developing the LCC regression model, including multi-linear regression, support vector machine (SVM), k-Nearest Neighbors (kNN), decision trees, multi-output SVM, multi-output regression trees, and multilayer perceptron.

**A Historical Data-based Facility LCCA Framework**

**The Components of LCC**

Numerous costs are associated with the design, construction, installation, operating, maintaining, and disposing of a building or building system. According to (Fuller, 2010), building-related costs usually fall into the following categories:

- Initial costs – purchase, acquisition, and construction costs.
- Utility costs – electricity, water, gas, and garbage costs.
- Operation, maintenance, and repair (O&M) costs.
- Replacement costs – capital replacements of building systems that have different service lives.
- Residual values – resale or salvage values, or disposal costs.
- Finance charges – loan interest payments.
- Non-monetary benefits or costs – such as the benefit derived from a quiet HVAC system or improved lighting.

In this paper, we implement machine learning on the historical data to forecast the costs of a new building– initial costs, utility costs, O&M costs, and replacement costs. The prediction of residual values, finance charges, and non-monetary benefits or costs are excluded from the scope of this research.

**Data acquisition**

The raw data, which can be used to derive each LCC component, are distributed in separate building systems or archives. The relevant building LCC components and their potential data sources are listed in Table 1.

<table>
<thead>
<tr>
<th>LCC Component</th>
<th>Potential Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial costs</td>
<td>The capital planning and investment control system</td>
</tr>
<tr>
<td></td>
<td>Some digital spreadsheets that record the design and construction costs</td>
</tr>
<tr>
<td></td>
<td>Some physical spreadsheets that record the design and construction</td>
</tr>
<tr>
<td></td>
<td>Some physical documents related to building design and construction</td>
</tr>
</tbody>
</table>
Utility costs  
The Building Automation System (BAS) / Building Management System (BMS)

The Building Energy Management Systems (BEMS)

Operation, maintenance, and repair (O&M) costs  
The Building Energy Management Systems (BEMS)

Computerized Maintenance Management System (CMMS)

Replacement costs  
The same source as the initial costs
CMMS

Figure 1 shows a general data acquisition process we proposed for LCCA.

The design and construction documentation refer to the construction drawings, estimation reports, project schedule, manuals, and specifications. The Building Information Model (BIM) is the “digital twin” of a building (Eastman et al., 2011). The required building data can be automatically extracted from BIM if it is properly developed and relevant information included (Gao & Pishdad-Bozorgi, 2018; Pishdad-Bozorgi et al., 2018).

Data derivation through machine learning

The first and most challenging task of conducting LCCA for a building is to determine the economic effects of alternatives and to quantify these effects and express them in monetary amounts (Fuller, 2010). Our hypothesis is that by extracting and formatting the LCC-related data generated by and housed in different building systems, and applying appropriate machine learning techniques, we can forecast each LCC component of a new building as early as the programming phase. A literature review is conducted to determine the most commonly used independent variables affecting building LCC (Tables 2).

Table 2: Commonly used independent variables in the LCC prediction model

<table>
<thead>
<tr>
<th>Building function/type</th>
<th>Total building area</th>
<th>Building age</th>
<th>Structural type</th>
<th>Number of rooms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of floors</td>
<td>Floor height</td>
<td>Façade type</td>
<td>Number of people</td>
<td>Building volume</td>
</tr>
</tbody>
</table>

Table 2: Commonly used independent variables in the LCC prediction model

(Kim, An, et al., 2004; Kim, Yoon, et al., 2004; An et al., 2007; Tu et al., 2007; Sonmez, 2008; Cheng et al., 2009; Hong et al., 2011; Ji et al., 2011; Koo et al., 2011; Li & Guo, 2012b; Tu & Huang, 2013; Jin et al., 2016; Robinson et al., 2017; Deng et al., 2018; Zhang et al., 2018)
Machine learning methods that have been proven valid in predicting building-related costs involve: regression with SVM (Idowu et al., 2016), decision trees (Dogan et al., 2008), random forests (Deng et al., 2018), Artificial Neural Network (Li & Guo, 2012a; Tu & Huang, 2013), and multistep ahead approach (Dursun & Stoy, 2016).

To compare the LCCs of facilities built in different years, their costs need to be discounted to the present value of a certain year. The present value of the initial cost is calculated according to the following equation:

\[ PV_{IC} = IC \times \prod_{i=1}^{t} (1 + r_i) \]  \hspace{1cm} (1)

Where:
- \( PV_{IC} \) is the present value of the initial cost.
- \( IC \) is the amount of initial cost.
- \( t \) is the building age.
- \( r_i \) is the annual inflation rate of \( i \) years ago.

The present value of the utility cost is calculated according to the following equation:

\[ PV_U = \sum_{j=1}^{n} (UC_j \times UP_j) \times \prod_{i=1}^{t} (1 + r_i) \]  \hspace{1cm} (2)

Where:
- \( PV_U \) is the present value of utility cost, which can be electricity cost, water cost, gas cost, etc.
- \( UC_j \) is the annual utility consumption of \( j \) years ago.
- \( UP_j \) is the utility price of \( j \) years ago.
- \( n \) is the length of the study period in years.
- \( r_i \) is the annual inflation rate of \( i \) years ago.

The present value of the O&M cost is calculated according to the following equation:

\[ PV_{OM} = \sum_{j=1}^{n} ((LH_j \times LP_j + OMC_j) \times \prod_{i=1}^{t} (1 + r_i)) \]  \hspace{1cm} (3)

Where:
- \( PV_{OM} \) is the present value of O&M cost.
- \( LH_j \) is the annual labor hours spent on O&M \( j \) years ago.
- \( LP_j \) is the O&M labor rate \( j \) years ago.
- \( OMC_j \) is the annual O&M monetary cost \( j \) years ago.
- \( r_i \) is the annual inflation rate of \( i \) years ago.

**Validation Experiments and Findings**

We conducted a series of experiments using the proposed LCCA framework in a university campus. 121 buildings were studied, and their basic statistics information is shown in Table 3. The building types involve residential buildings, libraries, dining halls, athletic facilities, parking decks, and educational complexes that consist of laboratories, classrooms, and offices.
The overall LCCs of these buildings were analyzed over a 20-year study period. To compare the buildings’ overall costs during the past 20 years, several assumptions were made. First, all buildings’ initial costs were converted to the “present values” in 1998. We assume that the changes in each building’s initial cost over time are proportional to the inflation rate. Second, for the years in which we do not have the actual data, because the building was newly built or sensors were not deployed, we use the simulated data generated by time series backcasting. The sums of annual utility and O&M costs were also converted to present values of 1998. The inflation rates used were provided by U.S. Bureau of Labor Statistics (US Inflation Calculator, 2019).

The initial cost data were housed in the Capital Planning and Space Management system established by the university’s facilities management (FM) department. The initial costs are discounted to present values of 1998, using Equation (1). The utility consumption data are generated by the university’s BAS – Metasys® (Johnson Controls Inc., 2018). We collected the data by downloading comma-separated value (CSV) files from the Ion data grabber system (Ntrepid Corporation, 2018). The utility data included the consumption of electricity, water, and gas. The data were available from October 2012 to present (January 2019) and the interval was 15 minutes. We first calculated each building’s weekly consumptions and found that the utility consumptions of studied buildings show repeating trends. Then, we used the machine learning software tool Microsoft R (Microsoft, 2019) to apply time series backcasting to simulate the past utility consumptions, thereby to obtain the estimated annual utility consumption from 1999 to 2012. Figure 2 shows an example of the electricity consumption raw data (left), and the historical consumption and the simulated historical data plot (right). The annual monetary costs were calculated with Equation (2). The historical utility price was provided by the U.S. Bureau of Labor Statistics (Bureau of Labor Statistics, 2018).

In this case study, all building costs other than the initial costs and utility costs are considered O&M costs, including replacement costs, preventive and reactive maintenance costs, janitorial costs, decoration and renovation costs, etc. The data related to O&M costs and replacement costs were acquired from the university’s CMMS – AiM system (AssetWorks, 2018). More than 750,000 lines of O&M work order records from 2006 to present (September 2018) were exported from AiM in CSV format. We used OpenRefine (openrefine.org, 2018) to clean the data and obtained the annual O&M cost of each individual building. We also applied time series analysis on the O&M cost data with Microsoft R and simulated the historical O&M costs before 2006.
We developed two kinds of machine learning models for LCC prediction – the single-target regression model and the multi-target regression model. The former assumes the LCC components (the target features) are independent of each other, while the latter considers their intercorrelations. To develop the single-target regression model, we tested multiple algorithms, including multilinear regression, kNN, random forest, SVM, and multilayer perceptron. To develop the multi-target regression model, we tested multi-output support vector regression, multi-output regression trees, and multilayer perceptron.

The input parameters (independent variables) involve gross square footage (GSF), building age, architect, contractor, owner (college), number of floors, LEED certifications, centralized heating/cooling, and building space allocations. The building space allocations are the percentages of building space usage, which involve building service, circulation, mechanical, classroom, laboratory, office, study facility, special usage, general usage, support facility, healthcare facility, residential, and other.

Because the data set used in this research was a small (121 instances), we implemented leave-one-out cross validation during the model training process (Kelleher et al., 2015). The evaluation results indicated that the most suitable model was the multi-target regression model using multilayer perceptron – with relative mean absolute errors (RMAE) of 22.211%, 24.575%, and 34.852% for the predictions of initial cost, utility cost, and O&M cost, respectively.

The open-source software library TensorFlow (Google Brain Team, 2019) was used to develop the multilayer perceptron model. The developed multi-layer perceptron involves three hidden layers with 10, 8, and 5 neurons in each layer, respectively. The activation function used for the hidden layer is rectified linear unit. The structure of the developed multilayer perceptron model is shown in Figure 3.

**Figure 3:** The structure of the developed multilayer perceptron model for facility LCC prediction

We found that the utility cost and O&M cost are positively correlated to the initial cost, with correlation coefficients of 0.650 and 0.512, respectively. The most relevant independent variables affecting the initial cost involve the GSF, the number of floors, the building age, and the building owner (used by each college or department). The architect firm and the general contractor have some influence on the initial cost but not significant enough to identify the trend.

In a 20-year time span, the ratio of the utility cost and the O&M cost to the initial cost show an approximately normal distribution. On average, the overall utility cost is equal to 31.60% of the initial cost (standard deviation = 22.31%), while the O&M cost 75.26% (standard deviation = 49.22%), both with inflation considered.
facilities and computing centers consume much higher energy than other types of facilities. For example, the electricity costs of a football stadium and a computing center in this study are higher than their initial design and construction costs. The residential buildings and parking lots required the least maintenance cost. In addition, the building age have a negative correlation (-0.416) with the initial cost, which indicates that the design and construction of buildings become increasingly expensive over time. On the other hand, the older buildings tend to cost more on O&M and are less energy efficient.

Conclusions and Future Research

This research contributes to the body of knowledge by investigating the feasibility of obtaining an accurate estimation of facilities’ life-cycle costs (LCCs) by implementing machine learning on historical data. Even though the experiment case study was conducted in a university campus and the buildings studied are all associated with education, the proposed LCC analysis framework is applicable to any kind of organization that owns multiple facilities. By exploring the new possibility for better prediction of a facility’ LCC through leveraging historical data housed in heterogeneous building systems across a continuous network of buildings, this research has a greater impact than simply studying the LCC of an individual project in the programming or design phase. The impact involves data-based LCC inputs in future facility cost benchmarking and informed project developments by incorporating the data pertaining to the total cost of ownership. Using existing available data to benchmark facility costs can assist decision making, and new data can be incorporated as they become available. It is an iterative knowledge accumulation of facility costs that could not only identify performance trends but also identify the best practices of facility design, construction, and operation from a cost efficiency perspective.

References:


