

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 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 1: The LCC components and their potential data sources

LCC Component	Potential Data Source
Initial costs	The capital planning and investment control system
	Some digital spreadsheets that record the design and construction costs
	Some physical documents related to building design and construction

Utility costs	The Building Automation System (BAS) / Building Management System (BMS)
	The Building Energy Management Systems (BEMS)
Operation, maintenance, and repair (O&M) costs	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.

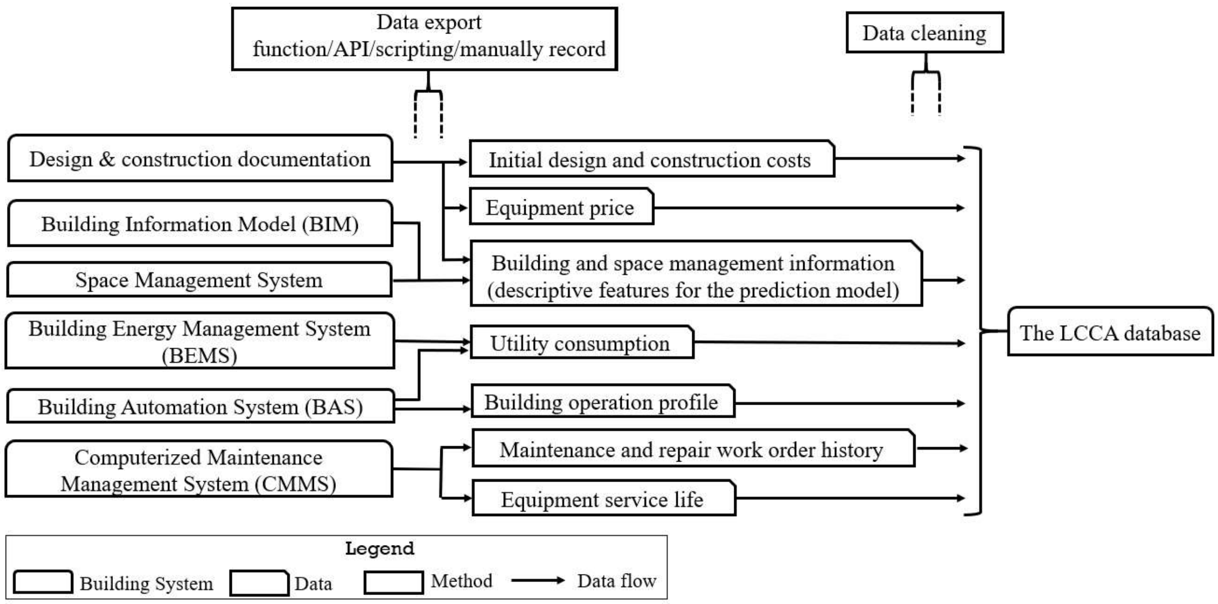


Figure 1: The proposed LCCA data acquisition process

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
(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)

Building function/type	Total building area	Building age	Structural type	Number of rooms
Number of floors	Floor height	Façade type	Number of people	Building volume

Roof type	Foundation type	Number of elevators	Mechanical installations	Total operating hours open per week

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:

$$PV_{IC} = IC \times \prod_{i=1}^{t}(1 + r_i) \quad (1)$$

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:

$$PV_U = \sum_{j=1}^{n}(UC_j \times UP_j \times \prod_{i=1}^{j}(1 + r_i)) \quad (2)$$

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:

$$PV_{OM} = \sum_{j=1}^{n}((LH_j \times LP_j + OMC_j) \times \prod_{i=1}^{j}(1 + r_i)) \quad (3)$$

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.

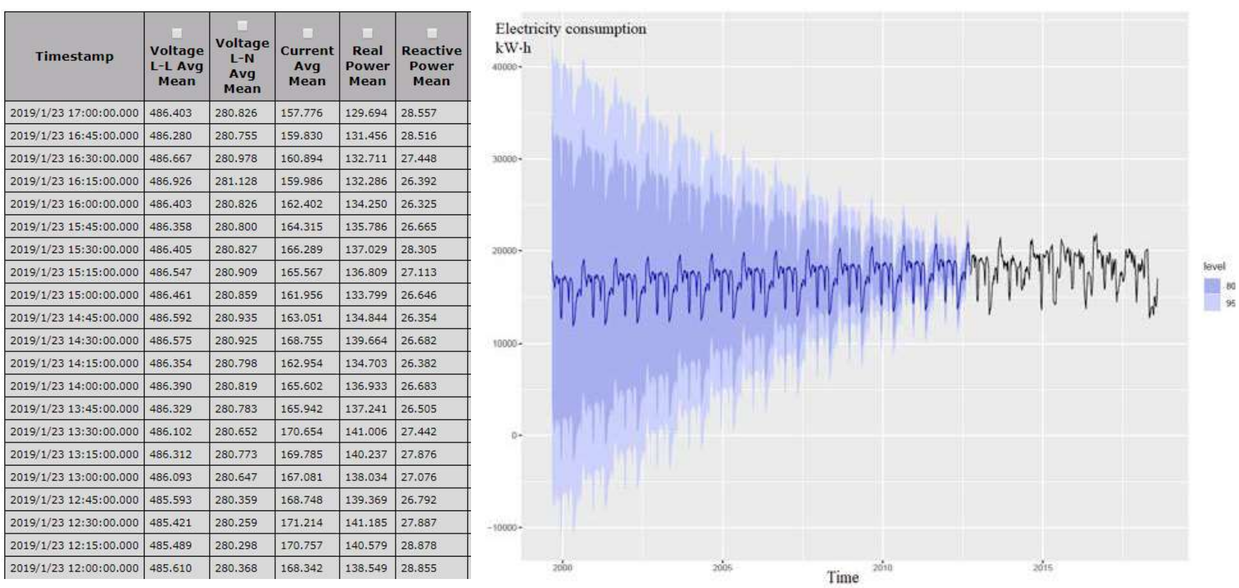
Table 3: The basic statistics information of the buildings in the case study

	Building age	Gross Square Footage (GSF)	Number of Floors	Initial Cost (Present Value in 1998)
Maximum	99	966,203	13	$113,216,000
Minimum	2	384	1	$280,000
Mean	39.37	96,871	3.9	$18,107,000

Median	33	48,666	4	$9,560,000

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).



Timestamp	Voltage L-L Avg Mean	Voltage L-N Avg Mean	Current Avg Mean	Real Power Mean	Reactive Power Mean
2019/1/23 17:00:00.000	486.403	280.826	157.776	129.694	28.557
2019/1/23 16:45:00.000	486.280	280.755	159.830	131.456	28.516
2019/1/23 16:30:00.000	486.667	280.978	160.894	132.711	27.448
2019/1/23 16:15:00.000	486.926	281.128	159.986	132.286	26.392
2019/1/23 16:00:00.000	486.403	280.826	162.402	134.250	26.325
2019/1/23 15:45:00.000	486.358	280.800	164.315	135.786	26.665
2019/1/23 15:30:00.000	486.405	280.827	166.289	137.029	28.305
2019/1/23 15:15:00.000	486.547	280.909	165.567	136.809	27.113
2019/1/23 15:00:00.000	486.461	280.859	161.956	133.799	26.646
2019/1/23 14:45:00.000	486.592	280.935	163.051	134.844	26.354
2019/1/23 14:30:00.000	486.575	280.925	168.755	139.664	26.682
2019/1/23 14:15:00.000	486.354	280.798	162.954	134.703	26.382
2019/1/23 14:00:00.000	486.390	280.819	165.602	136.933	26.683
2019/1/23 13:45:00.000	486.329	280.783	165.942	137.241	26.505
2019/1/23 13:30:00.000	486.102	280.652	170.654	141.006	27.442
2019/1/23 13:15:00.000	486.312	280.773	169.785	140.237	27.876
2019/1/23 13:00:00.000	486.093	280.647	167.081	138.034	27.076
2019/1/23 12:45:00.000	485.593	280.359	168.748	139.369	26.792
2019/1/23 12:30:00.000	485.421	280.259	171.214	141.185	27.887
2019/1/23 12:15:00.000	485.489	280.298	170.757	140.579	28.878
2019/1/23 12:00:00.000	485.610	280.368	168.342	138.549	28.855

Figure 2: An example of the electricity consumption raw data (left), and the historical consumption and the simulated historical data plot (right)

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 multiplayer 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.

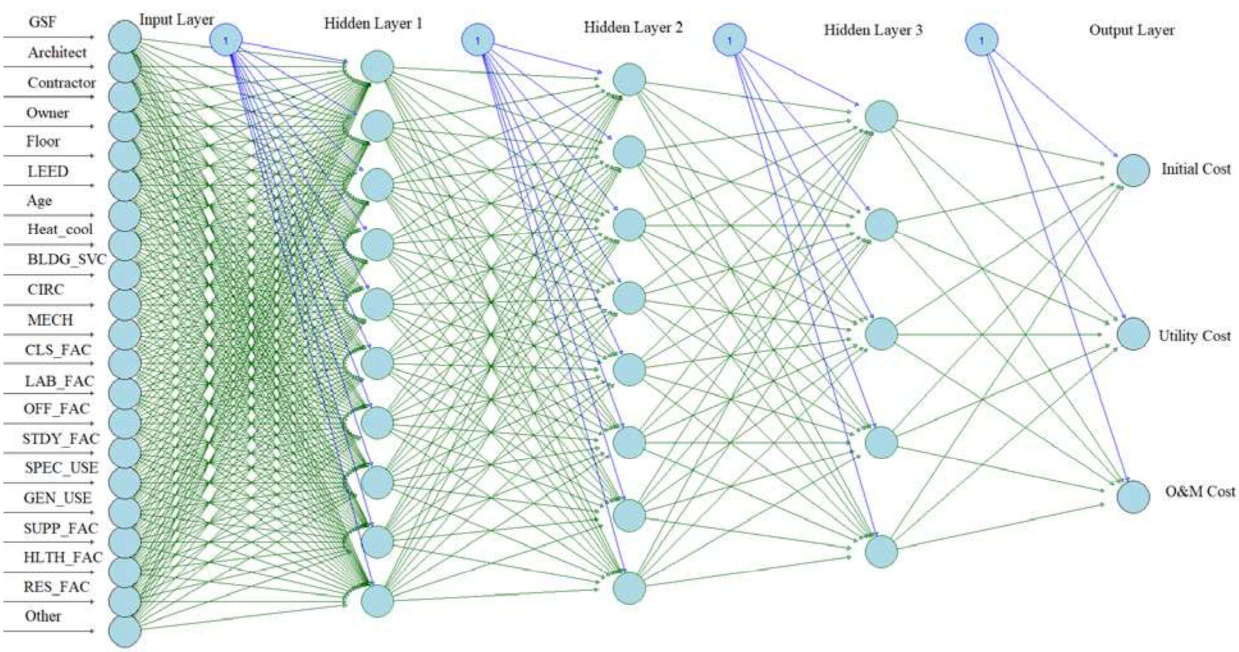


Figure 3: The structure of the developed multilayer perceptron model for facility LCC prediction

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, respectably. 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. Athletic 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:

An, S.-H., Kim, G.-H., & Kang, K.-I. (2007). A case-based reasoning cost estimating model using experience by analytic hierarchy process. *Building and Environment, 42*(7), 2573-2579.

AssetWorks. (2018). AiM Operations & Maintenance (O&M). Accessed Jan 24, 2019, from https://www.assetworks.com/iwms/aim-oandm/

Bureau of Labor Statistics. (2018). Average Energy Prices, Atlanta-Sandy Springs-Roswell – September 2018. Accessed Jan 24, 2019, from www.bls.gov/regions/southeast/news-release/averageenergyprices_atlanta.htm

Cheng, M.-Y., Tsai, H.-C., & Hsieh, W.-S. (2009). Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model. *Automation in Construction, 18*(2), 164-172.

De Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in Construction, 41*, 40-49.

Deng, H. F., Fannon, D., & Eckelman, M. J. (2018). Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. *Energy and Buildings, 163*, 34-43.

Dogan, S. Z., Arditi, D., & Gunaydin, H. M. (2008). Using decision trees for determining attribute weights in a case-based model of early cost prediction. *Journal of Construction Engineering and Management-Asce, 134*(2), 146-152.

Dursun, O., & Stoy, C. (2016). Conceptual Estimation of Construction Costs Using the Multistep Ahead Approach. *Journal of Construction Engineering and Management, 142*(9).

Eastman, C., Teicholz, P., Sacks, R., & Liston, K. (2011). *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors (2nd Edition)*: John Wiley & Sons.

Fuller, S. (2010). Life-cycle cost analysis (LCCA). *National Institute of Standards and Technology (NIST)*.

Gao, X., & Pishdad-Bozorgi, P. (2018). *Past, Present, and Future of BIM-Enabled Facilities Operation and Maintenance.* Paper presented at the Construction Research Congress 2018.

Google Brain Team. (2019). TensorFlow. Accessed Jan 24, 2019, from www.tensorflow.org

Hong, T., Hyun, C., & Moon, H. (2011). CBR-based cost prediction model-II of the design phase for multi-family housing projects. *Expert Systems with Applications, 38*(3), 2797-2808.

Idowu, S., Saguna, S., Ahlund, C., & Schelen, O. (2016). Applied machine learning: Forecasting heat load in district heating system. *Energy and Buildings, 133*, 478-488.

Ji, S.-H., Park, M., & Lee, H.-S. (2011). Cost estimation model for building projects using case-based reasoning. *Canadian Journal of Civil Engineering, 38*(5), 570-581.

Jin, R., Han, S., Hyun, C., & Cha, Y. (2016). Application of Case-Based Reasoning for Estimating Preliminary Duration of Building Projects. *Journal of Construction Engineering and Management, 142*(2).

Johnson Controls Inc. (2018). Metasys® Building Automation System. Retrieved Oct.13, 2017, Accessed Jan 24, 2019, from http://www.johnsoncontrols.com/buildings/building-management/building-automation-systems-bas

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*: MIT Press.

Kim, G.-H., An, S.-H., & Kang, K.-I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and environment, 39*(10), 1235-1242.

Kim, G.-H., Yoon, J.-E., An, S.-H., Cho, H.-H., & Kang, K.-I. (2004). Neural network model incorporating a genetic algorithm in estimating construction costs. *Building and Environment, 39*(11), 1333-1340.

Koo, C., Hong, T., & Hyun, C. (2011). The development of a construction cost prediction model with improved prediction capacity using the advanced CBR approach. *Expert Systems with Applications, 38*(7), 8597-8606.

Li, C. S., & Guo, S. J. (2012a). Development of a Cost Predicting Model for Maintenance of University Buildings. In F. L. Gaol & Q. V. Nguyen (Eds.), *Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science, Vol 1* (Vol. 144, pp. 215-221).

Li, C. S., & Guo, S. J. (2012b). Life cycle cost analysis of maintenance costs and budgets for university buildings in Taiwan. *Journal of Asian Architecture and Building Engineering, - 11*(- 1), - 94.

Microsoft. (2019). Microsoft R. Accessed Jan 24, 2019, from mran.microsoft.com

Mitchell, T. M. (1997). *Machine Learning*: McGraw-Hill.

Noshadravan, A., Miller, T. R., & Gregory, J. G. (2017). A Lifecycle Cost Analysis of Residential Buildings Including Natural Hazard Risk. [Article]. *Journal of Construction Engineering and Management, 143*(7), 10.

Ntrepid Corporation. (2018). Ion data grabber. Accessed Jan 24, 2019, from ion.ntrepidcorp.com

openrefine.org. (2018). OpenRefine. Accessed Jan 24, 2019, from http://openrefine.org/

Pishdad-Bozorgi, P., Gao, X., Eastman, C., & Self, A. P. (2018). Planning and developing facility management-enabled building information model (FM-enabled BIM). *Automation in Construction, 87*, 22-38.

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W. W., Guhathakurta, S., Brown, M. A., et al. (2017). Machine learning approaches for estimating commercial building energy consumption. *Applied Energy, 208*, 889-904.

SAS. (2018). Machine Learning: What it is & why it matters. Accessed Jan 24, 2019, from https://www.sas.com/it_it/insights/analytics/machine-learning.html

Sonmez, R. (2008). Parametric range estimating of building costs using regression models and bootstrap. *Journal of construction Engineering and Management, 134*(12), 1011-1016.

Tu, K. J., & Huang, Y. W. (2013). Predicting the operation and maintenance costs of condominium properties in the project planning phase: An artificial neural network approach. *International Journal of Civil Engineering, 11*(4A), 242-250.

Tu, K. J., Huang, Y. W., Lu, C. L., & Chu, K. H. (2007). *Predicting the operation and maintenance costs of apartment buildings at preliminary design stage: Comparing statistical regression and artificial neural network methods*. Paper presented at the CME 25 Conference Construction Management and Economics. Retrieved from - https://www.scopus.com/inward/record.uri?eid=2-s2.0-84877614694&partnerID=40&md5=6373edd2ee68e87e1891ed7ffd61e00e

US Inflation Calculator. (2019). Historical Inflation Rates: 1914-2018. Accessed Jan 24, 2019, from https://www.usinflationcalculator.com/inflation/historical-inflation-rates/

Zhang, C., Cao, L. W., & Romagnoli, A. (2018). On the feature engineering of building energy data mining. *Sustainable Cities and Society, 39*, 508-518.