



ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/rbri20>

## A framework of developing machine learning models for facility life-cycle cost analysis

Xinghua Gao & Pardis Pishdad-Bozorgi

**To cite this article:** Xinghua Gao & Pardis Pishdad-Bozorgi (2020) A framework of developing machine learning models for facility life-cycle cost analysis, *Building Research & Information*, 48:5, 501-525, DOI: [10.1080/09613218.2019.1691488](https://doi.org/10.1080/09613218.2019.1691488)

**To link to this article:** <https://doi.org/10.1080/09613218.2019.1691488>



Published online: 22 Nov 2019.



Submit your article to this journal 



Article views: 1049



View related articles 



View Crossmark data 



Citing articles: 13 View citing articles 



## A framework of developing machine learning models for facility life-cycle cost analysis

Xinghua Gao <sup>a</sup> and Pardis Pishdad-Bozorgi  <sup>b</sup>

<sup>a</sup>Myers-Lawson School of Construction, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA; <sup>b</sup>School of Building Construction, Georgia Institute of Technology, Atlanta, GA, USA

### ABSTRACT

Machine learning techniques have been used for predicting facility-related costs but there is a lack of research on developing machine learning models for the complete life-cycle cost (LCC) analysis of facilities. This research aims to systematically investigate the feasibility of forecasting facilities' LCC by implementing machine learning on historical data. The authors propose a comprehensive and generalizable framework for developing facility LCC analysis machine learning models. This framework specifies the data requirements, methods, and expected results in each step of the model development process. First, a literature review and a questionnaire survey were conducted to determine the independent variables affecting facility LCC and to identify the potential data sources. The process of using raw data to derive LCC components is then discussed. Finally, a proof-of-concept case study was conducted on a university campus to demonstrate the application of the proposed framework. This research concludes that current building systems already contain the data for LCC analysis and that the proposed framework is effective in facility LCC prediction.

### ARTICLE HISTORY

Received 30 May 2019

Accepted 1 November 2019

### KEYWORDS

Data availability; machine learning; life-cycle cost (LCC); facility management

## Introduction

A large quantity 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 facility's life cycle costs (LCC) rather than the initial design and construction costs. LCC analysis has become increasingly important in the design of new buildings and in the retrofitting, refurbishment, and renovations of existing buildings. Despite the importance of LCC analysis, however, researchers and industry professionals are facing challenges when practicing LCC analysis in the Architecture, Engineering, Construction, and Owner-operated (AECO) industry. Currently, most LCC analysis methods, such as the ones introduced in ASTM (2017), assume that one can estimate a building component's LCC if one knows its price, life expectancy, and the cost of all the operating and maintenance activities associated with it. The real service lives and costs of many buildings and their systems, however, are difficult to predict for multiple reasons. One is that there is always a mismatch between the predicted energy performance of a building and the actual measured performance, typically addressed as 'the performance gap' (De Wilde, 2014). Another reason

may be that, according to the authors' experience, many building systems and components, with appropriate 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 labour will be needed to repair them when they malfunction after the warranty expires. Moreover, even the same types of systems used in different buildings may have different LCC because monetary and labour costs vary depending on each facility manager's operational profile on building systems.

In recent years, with the development of machine learning technologies, new opportunities have been emerging for data analyst and cost estimators to predict building-related costs more precisely. Machine learning is an automated process that extracts patterns from data (Kelleher, Namee, & D'Arcy, 2015). In the field of predictive data analytics, machine learning is a method used to devise complex prediction algorithms and models (Kelleher et al., 2015; Mitchell, 1997). With sufficient data, it is possible to quickly and automatically produce machine learning models that can analyse a large amount of complex data and deliver fast and accurate results (Pantic, 2019). By building precise models, an

organization can uncover hidden insights, predict future values, and produce reliable, repeatable decisions through learning from historical relationships and trends in the data (SAS, 2018). As a viable alternative to simulation tools, machine learning techniques can give an accurate quantitative estimation of energy demand for different building systems (Deng, Fannon, & Eckelman, 2018) and predict facility-related costs (Gao, Pishdad-Bozorgi, Shelden, & Hu, 2019).

The development of valid and robust machine learning models requires extensive data (Alpaydin, 2014). The hypothesis is that the evolving building systems, such as Building Automation Systems (BAS), Computerized Maintenance Management Systems (CMMS), and Building Energy Management Systems (BEMS), already contain many valuable data but have not been sufficiently used for developing LCC machine learning models. Data such as building features, utility consumptions, and maintenance work orders can be extracted from these systems. By implementing machine learning on this data, it is possible to achieve a better understanding of a facility's LCC and overcome multiple barriers associated with current LCC analysis methods. Hence, more informed decisions in building design, construction, and facility management can be achieved. More research is needed to elucidate how to utilize the existing data housed in separate building systems for LCC machine learning model development.

For this purpose, in this research, the authors systematically investigate the feasibility of forecasting facilities' LCC by implementing machine learning on historical data. Specific research questions include the following: (1) whether an organization that operates multiple buildings can predict its new and existing buildings' overall LCC by utilizing machine learning models trained from historical data; (2) whether a generalized framework can be developed that is applicable to different organizations with various types of facilities; (3) whether the existing building systems already contain the data required to establish the predictive models for LCC.

To answer these questions, the authors propose a comprehensive and generalizable framework for developing facility LCC analysis machine learning models. This framework specifies the data requirements, methods, and expected results in each step of the model development process. It offers guidance for formalizing knowledge in facility LCC analysis by capturing necessary information from diverse data sources and reasoning about the captured data with machine learning techniques. It is envisioned that, through the capture and analysis of historical data relevant to facility costs, tacit knowledge of LCC analysis can be semi-automatically formalized through the proposed framework, which

will reduce reliance on individual researchers for knowledge formalization.

## Literature review

In recent years, the development of machine learning techniques has been providing building experts with new opportunities to achieve more accurate predictions of facility-related costs in the early design phase or even the programming phase. This section introduces the machine learning algorithms that can be used for facility-related cost prediction, before presenting a summary of current research on the development of machine learning techniques for the prediction of facilities' initial costs, utility costs, and operation and maintenance (O&M) costs. Research gaps are identified and discussed at the end of this section.

To identify related publications involving machine learning applications in the facility LCC prediction field, a keyword search is performed in academic databases, including Elsevier, Emerald Insight, EBSCO, Wiley, ASCE, CIB, Springer, Taylor & Francis, and ISPRS. Articles with abstracts containing 'machine learning' or 'prediction' and the keywords 'building cost', 'energy consumption', 'operation cost', and 'maintenance cost' are identified and reviewed. The following aspects of each reviewed paper are examined: (1) research methodology, (2) algorithm used, (3) applicable facility type, (4) what kind(s) of costs are considered, (5) what descriptive attributes are used in the prediction model, (6) whether a case study/experiment has been conducted, and (7) the size of the dataset. The reviewed research studies are summarized in a table that was published on the open data platform OSF (Gao & Pishdad-Bozorgi, 2019b).

## *Machine learning methods for facility cost prediction*

### *Linear regression and gradient descent*

Regression analysis is a technique for modelling the relationship between variables (Montgomery, Peck, & Vining, 2012). If the relationship between the independent variables (descriptive attributes) and the dependent variable (target attribute) is linear, then the model is called a linear regression model. A model that involves only one independent variable is called a simple linear regression (SLR) model; a model that involves multiple independent variables is called a multiple linear regression (MLR) model. If the relationship between the independent variable(s),  $x$ , and the dependent variable,  $y$ , is modelled as an  $n$ th degree polynomial in  $x$ , it is called a polynomial regression model. Although the

polynomial regression model is nonlinear from the data perspective, it is considered a linear machine learning model. This is because the regression function is linear in the unknown parameters that are derived from the data. Therefore, polynomial regression is considered to be a special case of MLR (Montgomery et al., 2012).

Gradient descent is a commonly employed iterative optimization algorithm to find the values of parameters (coefficients) of a function that minimizes a cost function (Alpaydin, 2014). It can be used to solve both a linear and a nonlinear system. In predictive analytics, MLR with gradient descent is the most common approach to error-based machine learning, the goal of which is to find the set of parameters for a model that minimizes the total error across the predictions made by the model (Kelleher et al., 2015).

### **K-nearest neighbours regression**

The K-nearest neighbours (KNN) algorithm is a non-parametric method that can be used for regression analysis (Altman, 1992). It is a type of instance-based learning, or lazy learning, and is considered one of the simplest machine learning algorithms (Alpaydin, 2014). The output of a KNN regression is the object's property value, which is the average of the values of the object's  $k$  nearest neighbours (Altman, 1992). The KNN regression model is a composition of each local model with the prediction made, which is a function of the target feature value of the instance in the dataset closest to the query; hence, it is sensitive to noise in the target feature (Kelleher et al., 2015). In addition, the KNN regression model uses the full set of descriptive features when making a prediction, which renders it particularly sensitive to the occurrence of missing descriptive feature values (Kelleher et al., 2015). The KNN is a similarity-based approach to machine learning, which originates from the idea of making predictions based on what has proved effective in the past (Kelleher et al., 2015).

### **Regression trees and random forest**

A decision tree is a hierarchical tree-like model composed of a root node, interior nodes, and leaf nodes (Alpaydin, 2014). The decision tree machine learning model uses a decision tree to progress from the descriptions of an item (represented in the root node and interior nodes) to conclusions of the item's target value (represented in leaf nodes) (Kelleher et al., 2015). Decision trees where the target variable can take continuous values are called regression trees. The decision tree is the fundamental structure used in information-based machine learning, which adopts information theory (Gleick, 2011) as a method of determining the shortest sequence of descriptive feature tests required

to generate a prediction (Kelleher et al., 2015). In regression analysis, Random Forest is a method that constructs many decision trees during training and outputs the class that is the mean prediction of the individual trees (Barandiaran, 1998).

### **Support vector machines regression**

The Support vector machines (SVM) regression is another commonly used method of error-based machine learning for predictive analytics. The Support Vector algorithm is a nonlinear generalization of the Generalized Portrait algorithm (Smola & Schölkopf, 2004). It is grounded in the framework of statistical learning theory, characterizing properties of learning machines that enable them to generalize well to unseen data (Smola & Schölkopf, 2004). SVMs are a specific class of algorithms that are characterized by 'usage of kernels, absence of local minima, sparseness of the solution, and capacity control obtained by acting on the margins, or on the number of support vectors' (Gelfusa et al., 2015). SVMs map input vectors into a high dimensional feature space, where a maximal margin hyperplane is constructed (Chapelle & Vapnik, 2000). It is possible to apply SVMs to regression problems by introducing an alternative loss function that is modified to include a distance measure (Dibike, Velickov, & Solomatine, 2000; Smola, 1996; Smola & Schölkopf, 2004).

### **Artificial neural network and multilayer perceptron**

Artificial neural networks (ANNs) are computing systems inspired by the biological neural networks that constitute animal brains (Haykin, 2009). The ANN itself is not an algorithm but rather a framework to federate different machine learning algorithms for complex analysis (Haykin, 2009). Multilayer perceptron (MLP) is an ANN structure and is a non-parametric estimator that can be used for regression (Alpaydin, 2014). MLP utilizes a supervised learning technique called backpropagation for training and consists of at least three layers of nodes: an input layer, one or multiple hidden layers, and an output layer (Goodfellow, Bengio, & Courville, 2016; Rosenblatt, 1961; Rumelhart, Hinton, & Williams, 1985). One strength of MLP concerns the capability to distinguish data that are not linearly separable (Cybenko, 1989). MLP is also a method of error-based machine learning for predictive analytics (Kelleher et al., 2015).

### **Initial costs prediction**

Accurate estimation in the early design stage is vital for the successful execution of a construction project. Using machine learning techniques, research studies have provided practitioners with decision-support tools

for estimating construction duration and costs before the completion of a project's design stage, or even during the programming phase (Hong, Hyun, & Moon, 2011; Jin, Han, Hyun, & Cha, 2016; Koo, Hong, Hyun, Park, & Seo, 2010). Construction costs prediction studies can be categorized into three major groups based on the methods used, which are (1) regression analysis (Alshamrani, 2017; Jafarzadeh, Wilkinson, Gonzalez, Ingham, & Amiri, 2014; Li, Shen, & Love, 2005; Lowe, Emsley, & Harding, 2006; Sonmez, 2008; Trost & Oberlender, 2003; Zayed & Halpin, 2005) (2) case-based reasoning (CBR; Dogan, Ardit, & Gunaydin, 2006, 2008; Hong et al., 2011; Jin et al., 2016; Koo et al., 2010), and (3) ANN (Bala, Ahmad Bustani, & Shehu Waziri, 2014; Cheng, Tsai, & Sudjono, 2010; Dursun & Stoy, 2016; Kim, Yoon, An, Cho, & Kang, 2004; Shi & Li, 2008).

Studies have been conducted to compare the cost prediction performance of models based on different machine learning methods. For example, Kim, An, and Kang (2004) compared the accuracy of MLR, ANN, and CBR by experimenting on 530 residential buildings' construction costs. The results indicated that, although the ANN model yields more accurate results than the MLR and CBR models, the CBR model performs better than the ANN model in terms of ease of updating and consistency in the variables stored for long-term use. Researchers have also studied the performance of machine learning methods in specific cost prediction cases. Based on 71 projects conducted by a medium-sized electrical contractor, Aibinu, Dassanayake, Chan, and Thangaraj (2015) concluded that cost forecasting models based on ANN outperform regression models in predicting the costs of light wiring, power wiring, and cable pathways. Sajadfar and Ma (2015) compared the prediction accuracies of the models based on linear regression, MLR, KNN regression, decision tree regression, and ANN. They found that the ANN model exhibits the highest accuracy for welding operations.

### **Utility costs prediction**

Understanding the underlying dynamics of building utility consumption (energy, water, and gas) and predicting the consumption are essential for building resource planning, management, and conservation (Amasaly & El-Gohary, 2018; Zhang, Cao, & Romagnoli, 2018). Energy (electricity) consumption prediction is the most extensively studied topic in the facility LCC prediction field. This is probably because the electricity meters and sensors distributed in facilities provide sufficient high-resolution data, hourly or even quarter-hourly, for researchers to investigate utility costs in detail (Moon,

Park, Hwang, & Jun, 2018; Park, Choi, Hong, Lee, & Moon, 2018; Sala-Cardoso, Delgado-Prieto, Kampouropoulos, & Romeral, 2018). The most commonly used machine learning methods for energy forecasting involve (1) ANNs (Mocanu, Nguyen, Kling, & Gibescu, 2016; Park et al., 2018; Sala-Cardoso et al., 2018), (2) SVM regression (Chou & Ngo, 2016; Jain, Smith, Culligan, & Taylor, 2014), and (3) CBR (An, Kim, & Kang, 2007; Ji, Hong, Jeong, & Leigh, 2014).

Most of the reviewed studies in the utility consumption prediction field developed multiple machine learning models and compared their performance (Bouktif, Fiaz, Ouni, & Serhani, 2018; Robinson et al., 2017). For example, Geysen, De Somer, Johansson, Brage, and Vanhoudt (2018) developed a thermal load forecasting system that incorporated a collection of machine learning methods: linear regression, extremely randomized trees regression (ETR), ANN, and SVM regression. The experiment results indicated that linear regression performs the worst while ANN and ETR perform slightly better than SVM. The study conducted by Moon et al. (2018) also showed that the ANN-based model outperforms the SVM regression-based model in electric load forecasting. The study of Idowu, Saguna, Ahlund, and Schelen (2016), however, demonstrated that SVM offers better prediction performance than ANN and MLR in forecasting the thermal load in district heating substations.

### **O&M costs prediction**

Studies on using machine learning to predict O&M costs are relatively rare. This is probably because obtaining accurate maintenance data is challenging (Weerasinghe, Ramachandra, & Rotimi, 2016). The most commonly used machine learning methods in O&M costs forecasting are multiple regression (Au-Yong, Ali, & Ahmad, 2014; Krstić & Marenjak, 2017; Li & Guo, 2012a; Weerasinghe et al., 2016) and ANN (Li & Guo, 2012a; Tu & Huang, 2013). Au-Yong et al. (2014) found that the characteristics of condition-based maintenance of office buildings directly influence the cost performance. Based on these relationships, they developed a regression model for maintenance planning and prediction. Krstić and Marenjak (2017) developed a multiple regression model to predict the O&M costs for university buildings during the initial design phase. Li & Guo (2012a, 2012b) developed maintenance cost prediction models for university buildings using SLR, multiple regression, and back-propagation ANN. The results indicated that the back-propagation ANN model outperforms the other two models. Li and Guo (2012a, 2012b) also found that the first peak of renovation for university buildings will

be at around 20 years of age and that the second peak will occur at around 35 years of age; for a building with more than five floors, meanwhile, they found that the first and second peak of renovation will be at 15 and 30 years of age, respectively.

### **Research gaps**

Machine learning techniques can provide an accurate quantitative estimation of energy demand for different building systems (Deng et al., 2018) and can predict facility-related costs (Gao et al., 2019). There are, however, gaps in research regarding the development of machine learning models for LCC analysis. These are enumerated as follows.

Although machine learning techniques have been implemented in forecasting construction costs, utility consumption, and O&M costs, respectively, their application in predicting a building's total LCC is rarely found in the literature. There is a need for more studies that utilizing machine learning to predict a building's overall LCC and illuminate the underlying relationships between each cost components (such as initial design and construction costs, utility costs, and O&M costs).

Many challenges discussed in this field can be attributed to data insufficiency, including a lack of sufficient metering and accessibility, and poor data quality (Gallagher, Leahy, O'Donovan, Bruton, & O'Sullivan, 2018). The machine learning models used in many studies were established based on a very limited dataset (Shi & Li, 2008; Sonmez, 2008). As Milion, Paliari, and Liboni (2016) point out, 'data survey is the most difficult challenge in estimation studies'. Limited and uncertain information hinder accurate prediction of construction-related costs (Koo, Hong, & Hyun, 2011). The lack of reliable and consistent data also limits the application of LCC analysis in the early design stage (Weerasinghe et al., 2016). What data to record and how organizations should record the facility data for machine learning-based LCC analysis are seldom discussed in the literature.

Most of the developed machine learning models are only applicable to one type of building projects, such as housing (Hong et al., 2011; Jin et al., 2016), educational buildings (Li & Guo, 2012a), and office buildings (Koo et al., 2010). By nature, predictive models involve assumptions and simplifications based on similarities between the studied subjects. Therefore, the uniqueness of different building projects precludes the use of one model to predict different types of buildings' LCC (Bala et al., 2014; Banihashemi, Ding, & Wang, 2017; Hong et al., 2011).

The authors believe that it is possible to establish generalizable frameworks for developing facility LCC analysis machine learning models. These frameworks would specify the means and processes for the following: (1) identifying potential descriptive attributes (the input to machine learning models), such as by conducting a literature review or a survey; (2) data acquisition, such as by exporting data from BAS or CMMS, by finding the records in drawings and specifications, and by conducting surveys; (3) attributes selection; (4) machine learning algorithm selection; (5) applying the algorithms to the data; and (6) model evaluation. Currently, such a framework is yet to be developed. This motivated the authors to conduct the present study in an attempt to bridge the identified research gaps.

### **Research method**

To examine the feasibility of developing LCC machine learning models based on the historical data housed in different building systems, the authors designed this research as a mixed-methods study, which consists of a literature review, a questionnaire survey, and a case study. A literature review was conducted first to preliminarily identify the independent variables affecting building LCC. All variables that were used for building-related cost prediction in the reviewed research studies were collected. A questionnaire survey was then designed to collect experts' opinions on the preliminary variable list to supplement the results derived from the literature review and to ensure the comprehensiveness of the independent variable pool of the machine learning models. The participating experts were grouped into three categories – experts in initial costs, experts in utility costs, and experts in O&M costs – and each expert was asked to fill out a corresponding questionnaire. The initial cost experts consisted of eight individuals who are construction cost estimators and project managers with experience ranging from 2 years to 34 years. The utility cost experts consisted of six individuals who are college professors, postdoctoral researchers, and doctoral students working in this area, with experience ranging from 2 years to 42 years. The O&M cost experts consisted of five individuals who are facility managers and building data analysts with experience ranging from 12 years to 26 years. The potential influential factors (independent variables) were listed in the questionnaire, and the experts were asked to assess their impact on facility initial design and construction costs, utility costs, or O&M costs. In addition, the experts were asked to specify any additional influential factors that were not listed in the questionnaire.

Based on the identified independent variable list, the authors propose a framework that specifies the data

sources, the data integration process, the method for using the raw data to derive LCC components, and the overall machine learning model development and evaluation process. Finally, a proof-of-concept case study was conducted on a university campus to demonstrate the application of the proposed framework. The proposed framework is used to develop LCC prediction models for the university's budget planning and administration, and facilities management departments. These prediction models were designed to forecast facilities' LCC during the programming phase using very limited input such as gross square footage (GSF), the owner of the building (which college), the number of floors, and the space allocations. The goal was to provide these departments with a tool to quickly estimate and forecast the LCC of existing and future buildings (during the programming phase when building design is not yet available) and without the input of building cost experts.

### **Identifying the independent variables through literature review and questionnaire survey**

The influential factors (independent variables for the machine learning models) that affect the overall building LCC were identified through the literature review and questionnaire survey. The developed questionnaires were published on the open data platform OSF (Gao & Pishdad-Bozorgi, 2019d). The full list of independent variables is summarized in a table that was also published on OSF (Gao & Pishdad-Bozorgi, 2019c). Several factors that have not been mentioned by others were also proposed and listed in this table.

In the case study, the proposed framework was used to develop LCC prediction models during the programming phase when the available model inputs are limited to variables such as GSF, the owner, the total number of floors, and the space allocation (the percentage of classroom, laboratory, office, etc.). Hence, the variable pool for the case study was narrowed down to 21 variables that are available during the programming phase. These variables are shown in Table 1.

### **The overall framework for developing machine learning models for facility LCC analysis**

Figure 1 presents the overall framework for developing machine learning models for facility LCC analysis, which consists of four major modules: (1) obtaining the descriptive attributes, (2) obtaining the target attributes, (3) training the machine learning models, and

**Table 1.** The descriptive attributes of the machine learning models.

<b>Gross square footage</b>	<b>The Architect Company (architect)</b>	<b>The General Contractor (contractor)</b>
The college (owner)	Number of floors (floor)	LEED Certification (LEED)
Centralized heating/ cooling? (heat_cool)	Building Service Area % (BLDG_SVC)	Circulation Area % (CIRC)
Mechanical Area % (MECH)	Laboratory Facilities % (LAB_FAC)	Classroom Facilities % (CLS_FAC)
Office Facilities % (OFF_FAC)	Study Facilities % (STDY_FAC)	Residential Facilities % (RES_FAC)
Special Use Facilities % (SPEC_USE)	General Use Facilities % (GEN_USE)	Support Facilities % (SUPP_FAC)
Health Care Facilities % (HLTH_FAC)	Other Usage % (other)	Building Age (age)

\* The attribute names are in the brackets.

- (4) evaluating the models and selecting the most suitable one.

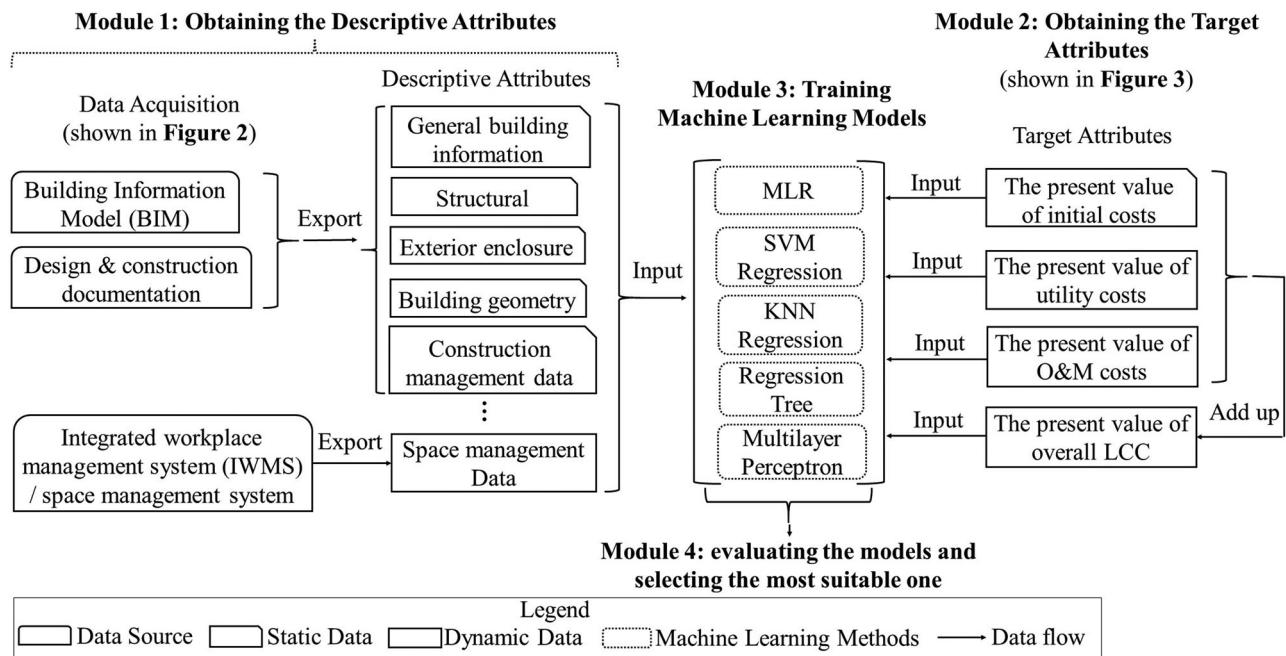
### **Assumptions**

In this research, there are three underlying assumptions for utilizing the historical data to predict facility LCC:

- (1) All historical data are correct: All the meter reading are accurate, and the records in each building system, whether automatically saved or manually inputted, are correct, with outliers in the data identified and excluded.
- (2) The simulated data can reflect the actual costs: The missing data (because the sensors were not deployed or malfunctioning), such as utility consumptions or O&M costs, can be estimated by forecasting or back-casting based on the historical data.
- (3) The inflation rate related to building costs is the same as the general inflation rate: It is assumed that the discount rate for building costs in the United States can be represented by the general inflation rate provided by the United States Bureau of Labor Statistics (Bureau of Labor Statistics, 2018b).

### **Module 1: obtaining the descriptive attributes (the independent variables)**

This module answers two questions: What are the factors significantly influencing facility LCC, and where can the related data be found? The results derived from the literature review and the questionnaire survey regarding the descriptive attributes to be used in the machine learning models are summarized in the table published on the open data platform OSF (Gao & Pishdad-Bozorgi, 2019c).



**Figure 1.** The overall LCC machine learning model development.

Data availability is the most significant challenge for facility LCC analysis (Gao & Pishdad-Bozorgi, 2019a; Milion et al., 2016). Many building managers and operators are using building management and control systems in their daily work. These systems, such as the BAS, the CMMS, and the BEMS, are constantly collecting or generating facility – and human-activity-related data, a portion of which can serve as the raw data for LCC analysis. Potential data sources that can be used to derive each LCC component are itemized in Table 2.

**Table 2.** The LCC components and their potential data sources

LCC component	Potential data source
Initial design and construction costs	<ul style="list-style-type: none"> <li>IWMS – the capital planning and investment control module.</li> <li>The construction cost estimation report that records the detailed construction costs.</li> <li>The design contract that records the design costs.</li> </ul>
Utility costs (utility consumptions)	<ul style="list-style-type: none"> <li>The BAS / Building Management System (BMS)</li> <li>The BEMS</li> </ul>
O&M costs	<ul style="list-style-type: none"> <li>CMMS</li> </ul>
Replacement costs	<ul style="list-style-type: none"> <li>The same source as the initial costs</li> <li>CMMS</li> </ul>

Figure 2 illustrates a high-level data acquisition process for LCC analysis. The design and construction documentation refers to construction drawings, estimation reports, scheduling, manuals, and specifications. The required building data can be automatically extracted from building information models (BIM) if they are appropriately developed and include relevant information (Gao & Pishdad-Bozorgi, 2018; Pishdad-Bozorgi, 2017; Pishdad-Bozorgi, Gao, Eastman, & Self, 2018).

BIM, which refer to the ‘digital twin’ of a building (Eastman, Teicholz, Sacks, & Liston, 2011), can provide the data related to potential descriptive attributes, such as structure type, building geometry, and foundation, as Figure 1 shows (Gao & Pishdad-Bozorgi, 2019a; Pishdad-Bozorgi et al., 2018). For organizations that do not have well-developed BIM (e.g. BIM with a level of development of 400) for all facilities, the required data can be found in the design and construction documentation. For example, design drawings contain building geometry and structural, foundational, and general building information, such as building age and function, while the construction documents contain construction management-related information, such as the delivery method and construction duration (which may influence the initial cost). After operation, a building’s space allocations may change over time, and this kind of change may not be timely reflected in the BIM. In this case, the up-to-date space allocation data can be found in the integrated workplace management system (IWMS) or other space management system.

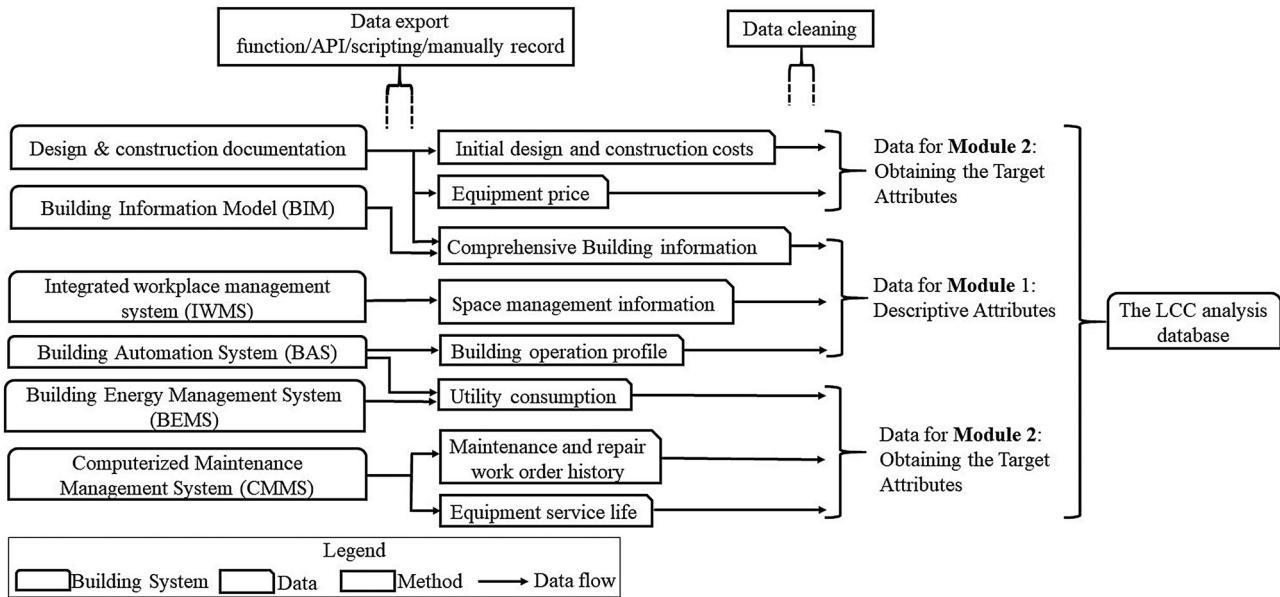


Figure 2. The overall LCC data acquisition process.

### Module 2: obtaining the target attributes (the present value of cost components)

The first and most challenging task of an LCC analysis for a building is to determine the economic effects of alternatives and to quantify these effects and express them in monetary amounts (Fuller, 2010). After the cost-related data are extracted from the building systems

and stored in one database, machine learning techniques can be implemented on them to forecast each LCC component of a building. The overall process of deriving LCC components is illustrated in Figure 3. The raw data used for deriving the initial design and construction costs, utility costs, and O&M costs are extracted from multiple building systems (discussed in the section entitled 'Data Requirements and Data Sources'). The data

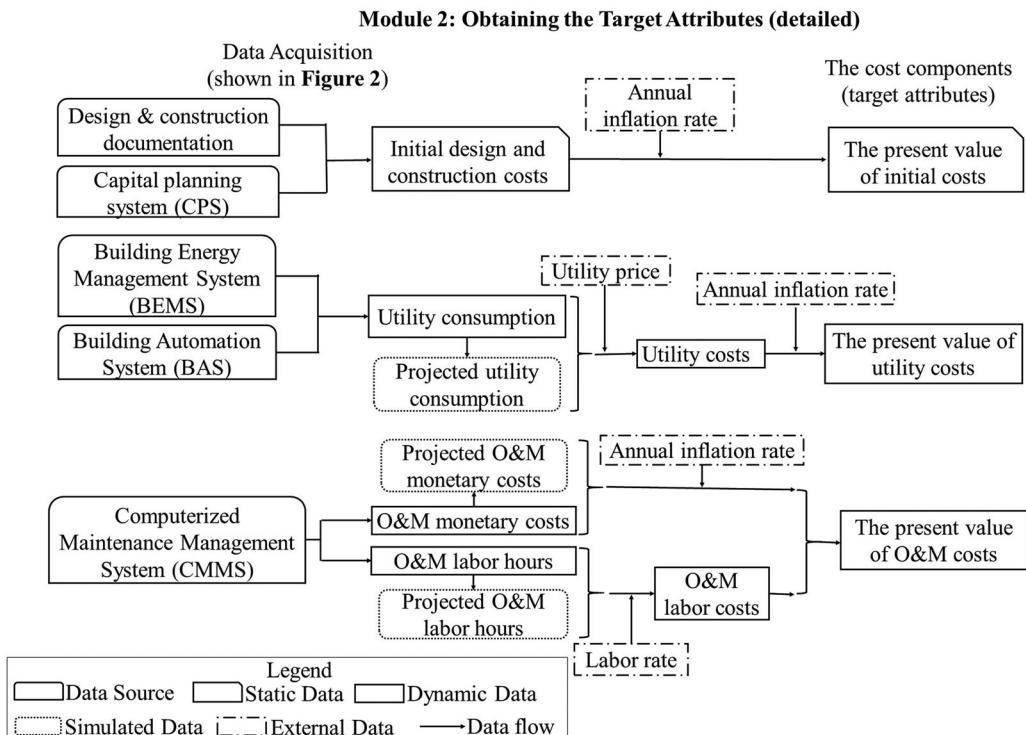


Figure 3. Module 2: obtaining the target attributes.

indexed in time order, utility consumption and O&M costs, are analysed using time series methods, and projections are made when necessary. For example, projections are made when there are missing values because sensors were not deployed in the past. Public statistics such as the historical inflation rate, utility price, and labour rate are incorporated into the analysis to calculate the monetary costs and to convert the costs to their present values. These present values of the LCC components are the target attributes of the LCC analysis machine learning models.

To compare the LCC 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 by multiplying the amount of the initial cost by the cumulative interest rate, as the following equation shows:

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

where  $PV_{IC}$  is the present value of the initial cost,  $IC$  is the amount of the initial cost,  $t$  is the building age and  $r_i$  is the annual inflation rate of  $i$  years ago.

The present value of the utility cost is calculated as follows: (1) First, each year's utility consumption is multiplied by the utility price of that year to determine the cost amount of that year, which is then multiplied by the cumulative interest rate to yield the present value of that year's utility cost; and (2) Finally, the present values of each year's utility cost are added up. The following equation presents the calculation:

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

where  $PV_U$  is the present value of the utility cost, which can be composed of 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 and  $r_i$  is the annual inflation rate of  $i$  years ago.

The present value of the O&M cost is calculated by (1) multiplying each year's labour hours expended by the average O&M labour rate of that year, which produces the labour cost amount of that year; (2) adding each year's the labour cost amount to the O&M monetary cost of that year, which yields the total O&M cost amount of that year; (3) multiplying each year's total O&M cost amount by the cumulative interest rate to derive the present value of that year's O&M cost; and (4) calculating the sum of the present values of each year's O&M cost. The following equation represents

the calculation:

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

where  $PV_{OM}$  is the present value of the O&M cost,  $LH_j$  is the annual labour hours expended on O&M  $j$  years ago,  $LP_j$  is the O&M labour rate  $j$  years ago,  $OMC_j$  is the annual O&M monetary cost  $j$  years ago and  $r_i$  is the annual inflation rate of  $i$  years ago.

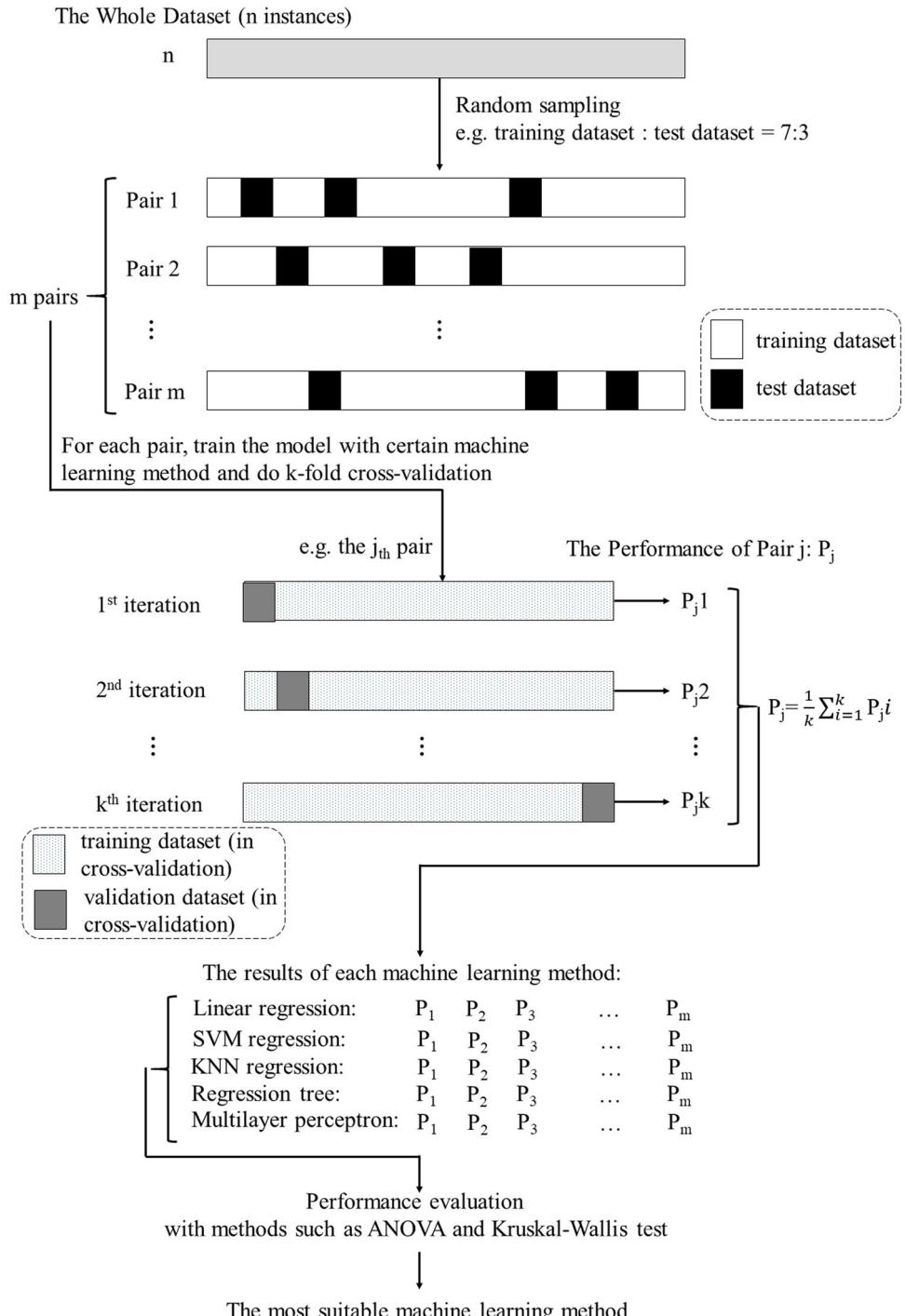
The present values of the initial costs, utility costs, and O&M costs are the target attributes of the LCC analysis machine learning models. In contrast to the descriptive attributes, which remain relatively static over a certain time period (such as 3 months), the target attributes are dynamic and can vary with real-time utility consumption and O&M costs. Therefore, a framework to acquire and integrate the dynamic facility data in an automated fashion is desirable for the overall facility LCC analysis machine learning model development process.

### **Module 3: training machine learning models**

With the descriptive attributes and target attributes ready, the next step is to train the machine learning models based on these data. In this research, the machine learning methods tested involve linear regression, SVM regression, KNN regression, regression trees, and MLP. According to the literature review, these machine learning methods have proved effective in building-related cost prediction. In the proposed framework, the method pool of training regression models for facility LCC analysis is expandable: As new machine learning techniques in predictive data analytics are developed, more methods can be adopted and implemented within the framework.

### **Module 4: evaluating the models**

Evaluating the models and selecting the best performing machine learning algorithm for facility LCC prediction can be accomplished by comparing the models' performance through repeated random sampling and cross-validation, as demonstrated in Figure 4, which is inspired by Hu and Castro-Lacouture (2019). First, the dataset is randomly split into the training set and test set at a ratio of, say, 7:3. Multiple pairs are generated by randomly sampling data points in the training set and data points in the test set (e.g.  $m$  pairs). For each pair (e.g. the pair  $j$ ), its training data point is used for training machine learning models implementing each algorithm, and then  $k$ -fold cross-validation is conducted to yield a series of evaluation results on each of these models



**Figure 4.** Machine learning algorithm evaluation and selection process.

$(P_1, P_2, \dots, P_k)$ . The average performance,  $P_j$ , is used to represent the performance of the corresponding algorithm. After repeating the training and cross-validation process for each of the randomly generated data pairs ( $m$  in total), each of the algorithms has  $m$  performance evaluation outcomes  $(P_1, P_2, \dots, P_j, \dots, P_m)$ . These outcomes are then analysed by evaluation methods such as an analysis of variance (ANOVA) or Kruskal-Wallis test to determine which algorithm outperforms the others

(Hu & Castro-Lacouture, 2019). ANOVA and the Kruskal-Wallis test are two of the most commonly used methods to compare the performance of algorithms and evaluate if there are significant differences. ANOVA is commonly used to assess differences between groups on a continuous measurement, but it requires that the data follow normal distributions (Tabachnick, Fidell, & Ullman, 2007). Unlike ANOVA, the Kruskal-Wallis test does not assume a normal distribution of

the residuals; this makes it a suitable assessment method for use when the data do not follow normal distributions (Kvam & Vidakovic, 2007). The most suitable machine learning method would possibly be different from one case to another, depending on the length of the time span studied, the attributes used, and the size and quality of the dataset.

## A proof-of-concept case study

### Overview

One of the challenges usually faced by an organization's capital planning department and/or facility management department is that they lack an effective means to quickly estimate a new facility's LCC during the programming phase, when no building design is available. Typically, during this phase, the decision makers have to determine the budget (estimated initial building cost) based on very limited information: the owner, the building function (user requirements), total building area, the number of floors, and, probably, the space distribution. It is already a challenging task without the consideration of the life-cycle utility and operation costs. Moreover, the predicted LCC data provided by survey and consulting companies, such as the White-stone facility operation cost reference (Romani, Abate, Miller, & Dotz, 2014) and maintenance and repair cost reference (Abate, Towers, Dotz, Romani, & Miller, 2014), may be overgeneralized and cannot reflect the organization's facility operation profile.

### The goal

The proposed machine learning-based LCC analysis approach in this research provides organizations that own multiple facilities with a solution to the LCC prediction issue. A series of experiments were conducted on a university campus (hereafter referred to as 'the university'). In these experiments, the proposed machine learning-enabled LCC analysis framework was used to develop LCC prediction models for the university's budget planning and administration, and facilities management departments. These prediction models were designed to forecast facilities' LCC during the programming phase using very limited input, such as GSF, the owner of the building (which college), the number of floors, and the space allocations. The goal was to provide these departments with a tool to quickly estimate and forecast the LCC of existing and future buildings (during the programming phase when building design is not yet available) without the input of building cost experts.

### About the university

The university has been established for over 130 years and owns more than 250 buildings, half of which are well metered with networks of sophisticated sensors and devices. These networks of devices embedded in the buildings' systems are generating the data needed for developing the LCC prediction machine learning models. The building systems operated by the university involve BAS Metasys (Johnson Controls Inc., 2018), the CMMS AiM System (AssetWorks, 2018), and the Capital Planning & Space Management System (CPSMS) INSITE (The INSITE Consortium, 2019).

Machine learning models were developed based on the historical data of 123 buildings on campus. The basic statistics of these buildings are presented in **Table 3**. The building types include residential buildings, libraries, dining halls, athletic facilities, parking decks, and educational complexes that consist of laboratories, classrooms, and offices.

### Data acquisition

#### The initial cost and space allocation

The university's CPSMS publishes the space management data of each campus building through a web system (Georgia Institute of Technology, 2019), which is based on Tableau (Tableau, 2019). This website also contains each building's initial cost. The raw data of initial cost and space allocation were downloaded from this website. **Figure 5** shows the website's interface.

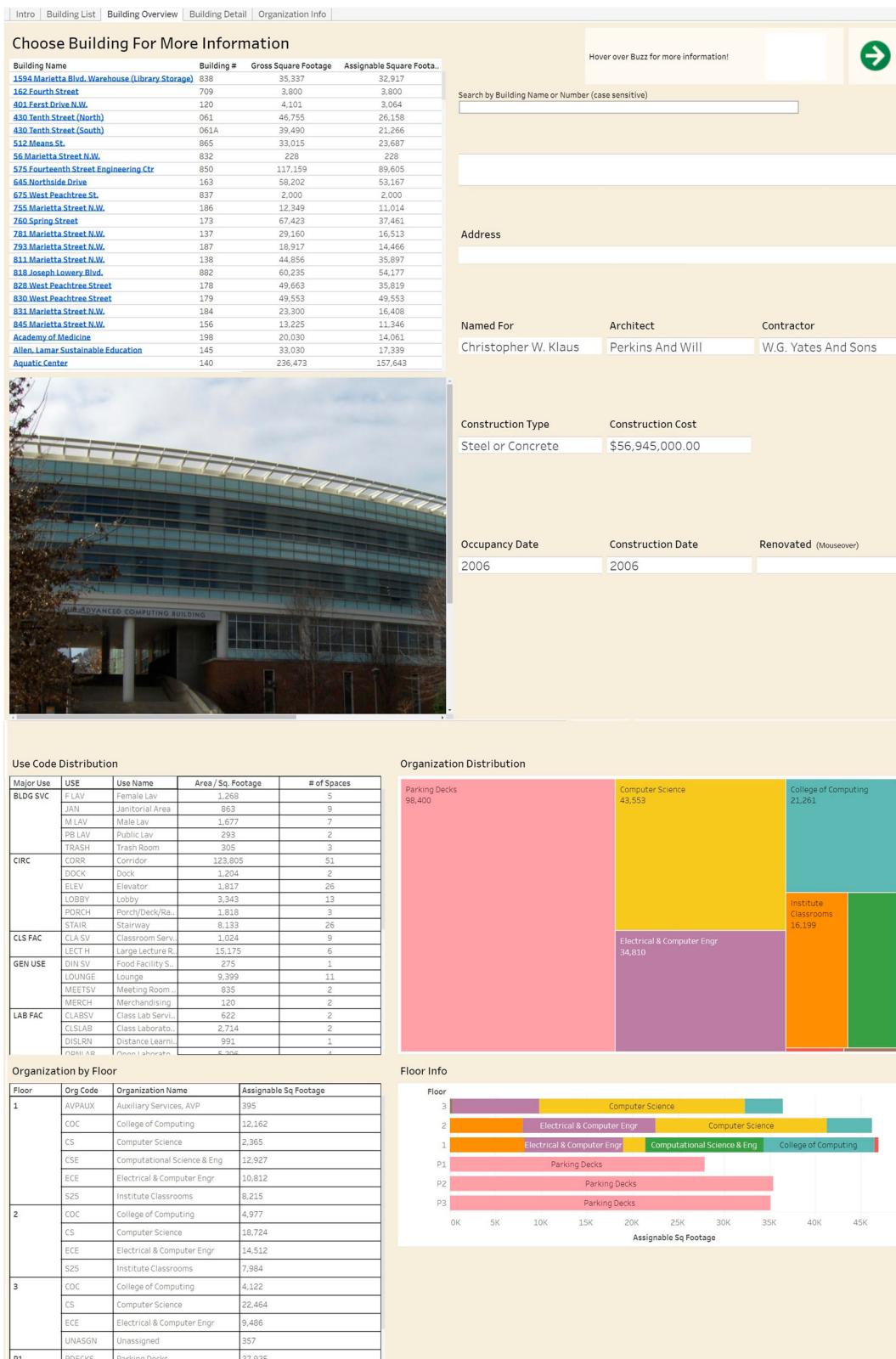
#### The utility consumption

The utility consumption data – electricity, water, and gas – were generated by the university's BAS and published on a website (shown in **Figure 6**) (Georgia Institute of Technology, 2018).

For most buildings of the university, the utility consumption data are available since 1 October 2012, with an interval of 15 min. In this research, the utility data used were from 1 October 2012 to 1 September 2018. The data (CSV files) were downloaded using Ion Data Grabber (Ntrepid Corporation, 2018) from the

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

	Building age	Gross square footage (GSF/GSM)	Number of floors	initial cost (present value in 1999)
Maximum	99	966,203/89,763	13	\$113,216,000
Minimum	2	384/36	1	\$280,000
Mean	39.37	96,871/9,000	3.9	\$18,107,000
Median	33	48,666/4,521	4	\$9,560,000

**Figure 5.** The web-based building information dashboard.

EnergyWatch system developed by the university's Aerospace Systems Design Laboratory (Aerospace System Design Laboratory, 2019).

### The O&M costs

The university's O&M work order records in the CMMS, AiM System, are available since 2006. Up to 1 September

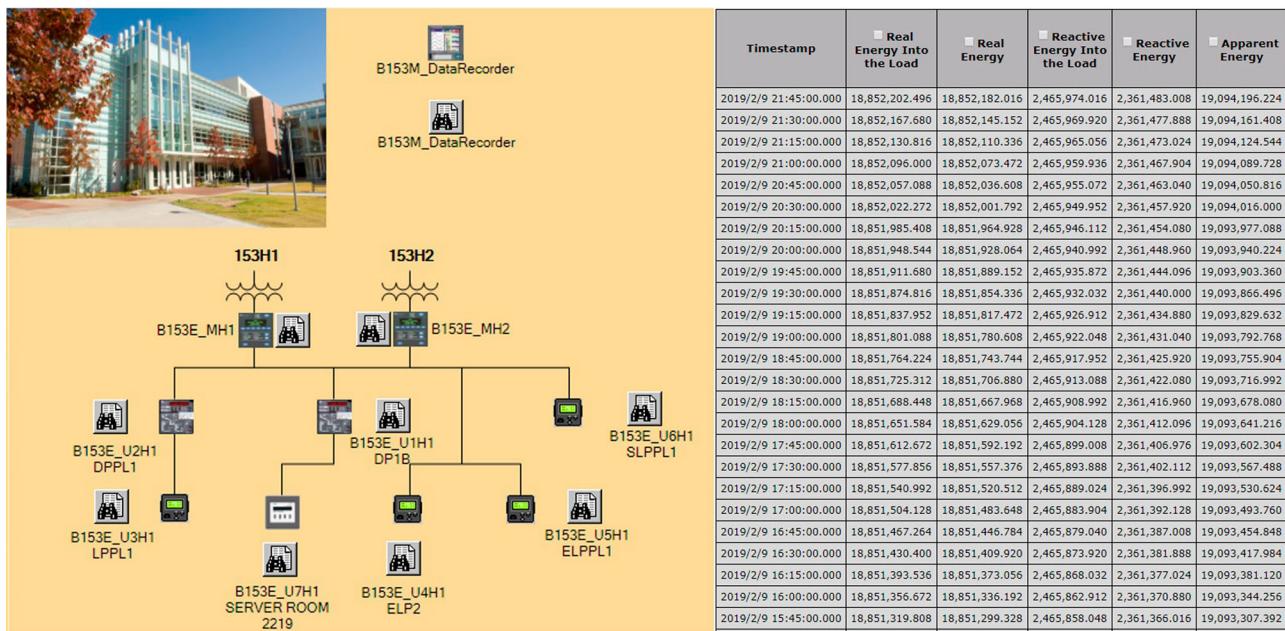


Figure 6. The website that publishes the utility consumption data.

2018, there were over 750,000 lines of records. These records were exported from the AiM System as a CSV file.

## Data processing

### Data cleaning

Based on the raw data, each building's weekly and monthly utility consumption was calculated, and outliers removed. Most of the studied buildings' monthly utility

consumption exhibits a repetitive pattern every year. Figure 7 provides two examples.

OpenRefine ([openrefine.org](http://openrefine.org), 2018) and MATLAB were used to clean the O&M work order records and thus to obtain the annual O&M cost of each building.

### Data simulation: time series backcasting

In the experiments, the building costs are studied within a 20-year time frame – from 1999 to 2018. During this time frame, for the buildings that lack historical data before a certain year (because the building was newly

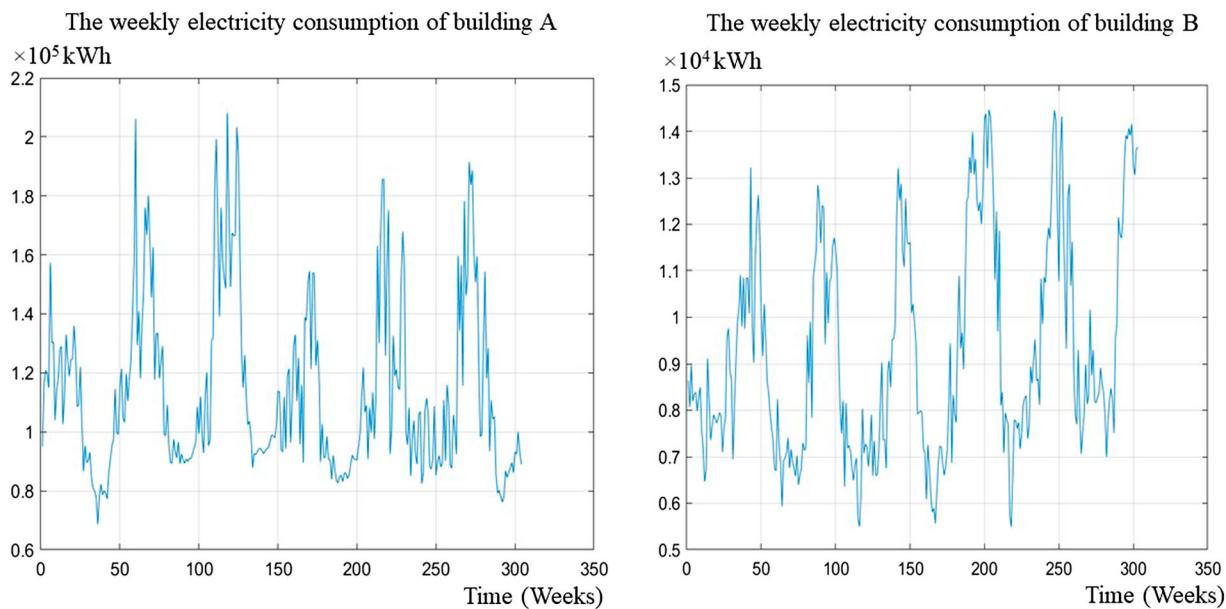
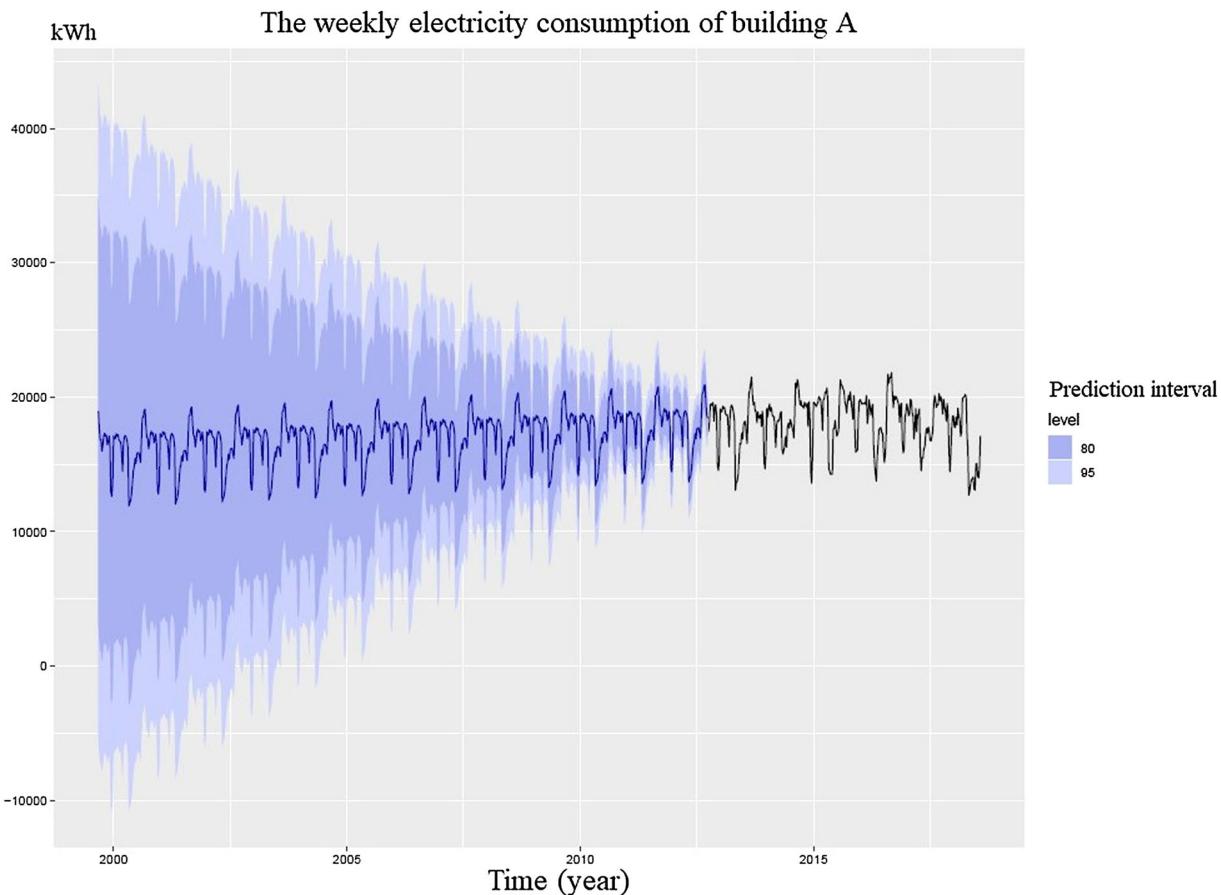


Figure 7. Examples of building electricity consumption trends.



**Figure 8.** An example of building electricity consumption data backcasting.

built or sensors were not deployed), time series backcasting was used to simulate the data. The machine learning software tool R Studio (Microsoft, 2019) was used to perform time series backcasting to simulate the past utility consumption and O&M cost. Figure 8 presents an example of building electricity consumption backcasting. The electricity consumption figures reveal repeating patterns. In the figure, the polygonal lines outside the blue area represent the actual consumption, while the ones in the blue area are the backcast expectations of actual consumption (simulated historical data). The darker blue area is the 95% prediction interval; the lighter blue area is the 80% prediction interval. For the buildings whose utility consumption or O&M cost did not exhibit repeating patterns, the mean of the actual annual cost was used as the simulated data.

#### **Discounting to present value**

The present value (1999) of the initial cost, utility cost, and O&M cost were calculated according to Equations (1), (2), and (3), respectively. The historical annual inflation rate used was based on the Consumer Price Index (CPI) statistics provided by the Bureau of Labor Statistics (BLS) (Bureau of Labor Statistics, 2018b; US Inflation

Calculator, 2019). The utility prices used were the average energy prices provided by BLS (Bureau of Labor Statistics, 2018a). The O&M labour rate used was based on the Current Employment Statistics (CES National) provided by BLS (Bureau of Labor Statistics, 2019).

#### **Model development**

Because the prediction models were designed to forecast facilities' LCC during the programming phase, the descriptive attributes (model inputs) used are the ones that can be known in this phase. The descriptive attributes involved are listed in Table 1. Most of the descriptive attributes are related to space allocation, such as the percentage of building service area, classroom facilities, and laboratory facilities.

Two kinds of machine learning models were developed 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, the authors tested multiple regression algorithms, including (1) MLR, (2) KNN, (3) random

forest, (4) SVM, and (5) MLP. To develop the multi-target regression model, the authors tested multi-output random forest and MLP.

To determine the best performing algorithm, the experiment was repeated 100 times (loops). In each loop, the full dataset was randomly split into the training set and the test set at a ratio of 8:2. The training set was then used to train the machine learning models with each algorithm. The trained models were then tested using the test set and the mean absolute error (MAE) was used to evaluate their effectiveness. Finally, the MAEs of each model calculated in all the loops were averaged to produce a final evaluation result, which was used to compare the performance of each model.

The MLR models were developed using the R package stats version 3.5.1 (R-project.org, 2019). The method used for fitting was QR decomposition.

The KNN regression models were developed using the R package FNN version 1.1 (Beygelzimer et al., 2019). In general, a large  $k$  value (number of neighbours) is more precise as it reduces the overall noise. Given limited data and the number of different campus building types, however, after a series of tests, the authors had to set the  $k$  value to 3, rather than 10 or more, to achieve the best possible results. The nearest neighbour search algorithm used was KD Tree.

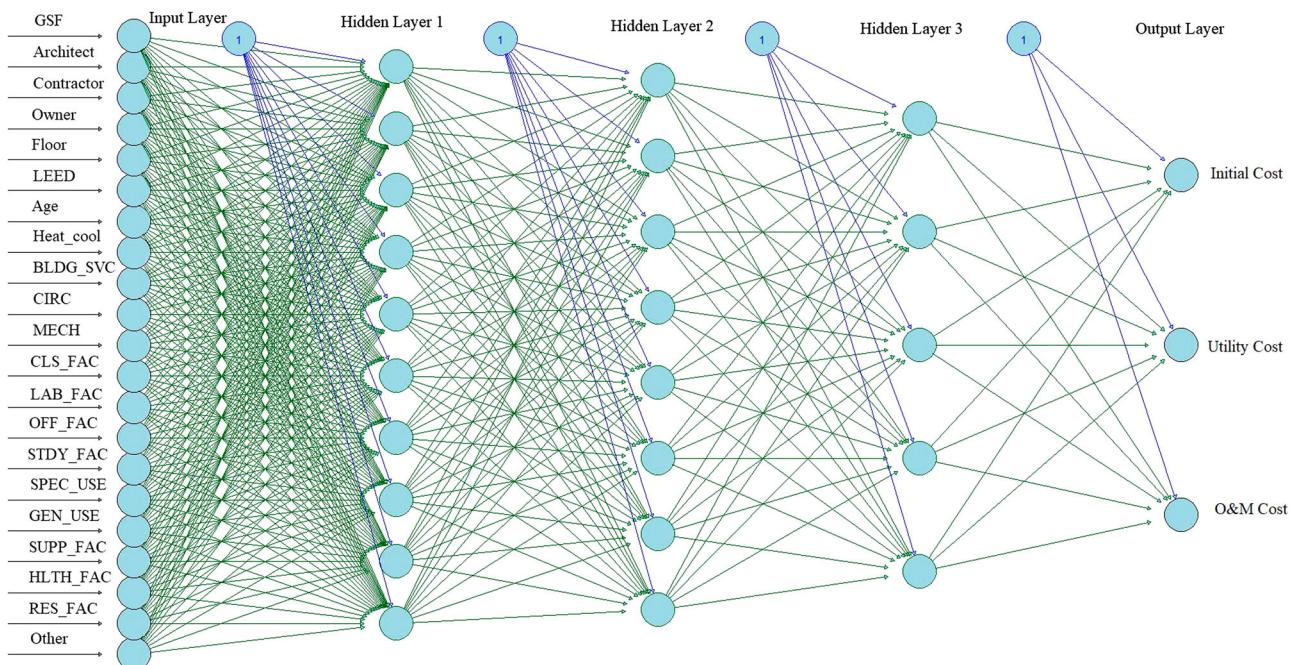
The random forest regression models were developed with the R package randomForest version 4.6-14 (Breiman, Cutler, Liaw, & Wiener, 2019). Five variables were randomly sampled as candidates at each split. The number of trees to grow was set to 500.

The SVM regression models were developed with the R package e1071 version 1.7-0 (Probability Theory Group, 2019). The kernel function used was polynomial. The gamma parameter, which defines how far the influence of a single training example reaches, was set to 1/ (the number of descriptive attributes). The coef0 parameter was set to 0 and the degree was set to 3.

The MLP single-target regression models were developed with the R package keras version 2.1.6 (Allaire et al., 2019). The MLP models contained three hidden layers with 10, 8, and 5 nodes, respectively. The batch size (the number of samples per gradient update) was set to 80. The number of epochs used to train the model was 100. In each epoch, 95% of the training set instances were used to train the model and used the other 5% for validation.

The multi-output random forest models were developed with the R Package MultivariateRandomForest version 1.1.5 (Rahman, 2019). The number of trees in the forest was set to 100, and the number of randomly selected descriptive attributes considered for a split in each regression tree node was set to 10. The minimum number of samples in the leaf node was 40.

The MLP multi-target regression models were developed using the R Package keras version 2.1.6. The models contained three hidden layers with 10, 8, and 5 nodes, respectively. The batch size was set to 90. The number of epochs to train the model was, again, 100. In each epoch, 90% of the training set instances were used to train the model and the other 10% for validation. Figure 9 shows the structure of the developed MLP multi-target



**Figure 9.** The structure of the MLP multi-target regression model for facility LCC prediction.

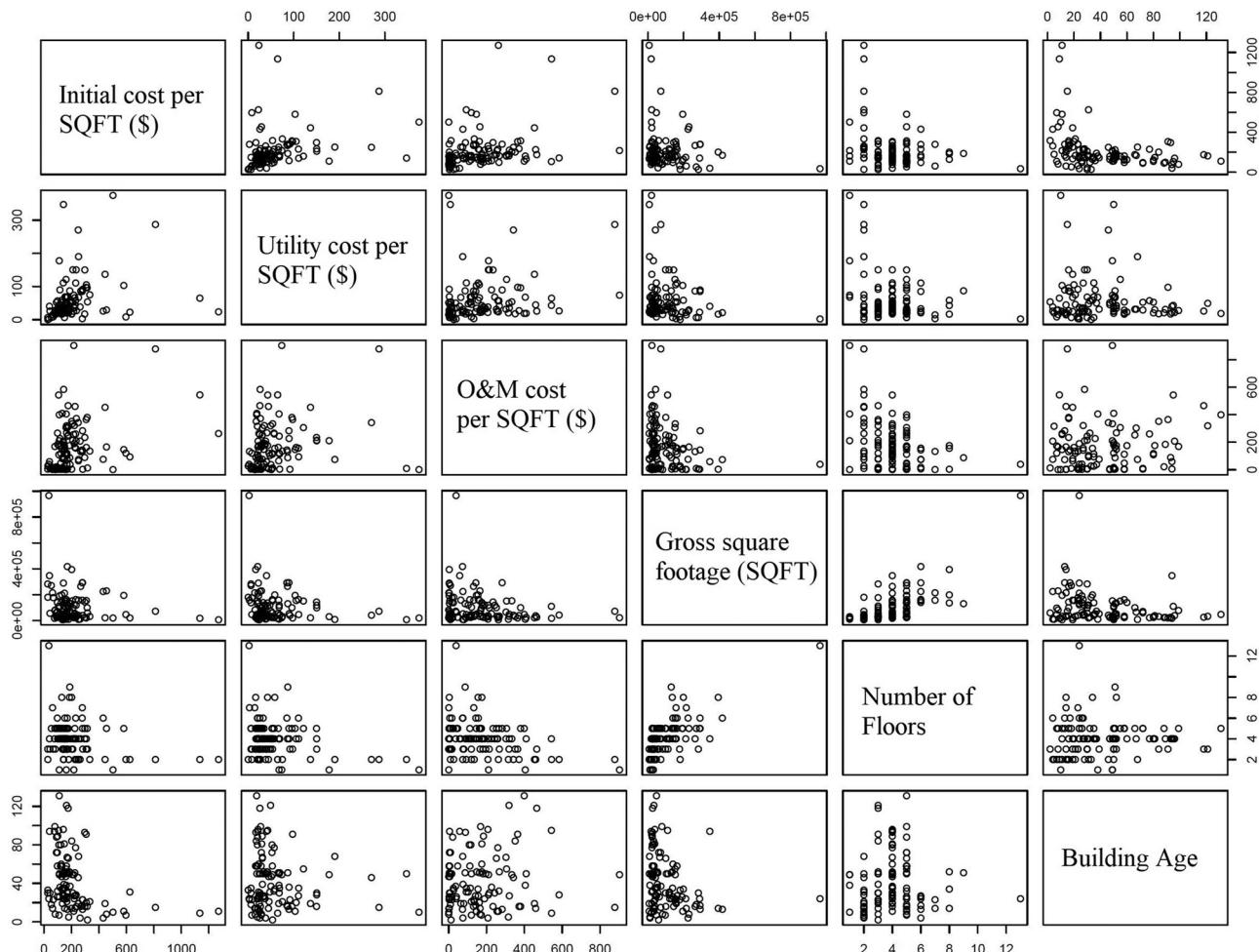
regression model. This optimal MLP structure for the case study was devised after a series of experiments were conducted to test MLP structures containing one to six hidden layers with nodes ranging from 3 to 12 in number.

## Results

This experiment incorporated the data of 123 buildings and used 21 descriptive attributes (as Table 1 indicates). By studying the linear correlation of each numerical descriptive attributes and target attribute, it is found that, in terms of cost per square foot (SF), the initial cost exhibits weak positive correlations with the utility cost (0.27), the O&M cost (0.34), and the percentage of special use facilities (0.33). This indicates that more expensive buildings (per SF) tend to cost more in terms of utility cost and O&M cost. Buildings with a higher percentage of special use facilities are also more expensive. The utility cost shows weak negative correlations with the number of floors ( $-0.30$ ) and circulation area ( $-0.31$ ). The O&M cost displays a weak positive

correlation with office facilities (0.40) and negative correlations with the number of floors ( $-0.35$ ) and the percentage of residential facilities ( $-0.41$ ). Figure 10 presents the scatter matrix of linear correlations between the studied buildings' initial cost per SF, utility cost per SF, O&M cost per SF, GSF, number of floors, and building age. These results imply that buildings with more floors tend to cost less in terms of utility cost and O&M cost. The buildings with a greater percentage of office facilities tend to be more expensive to operate and maintain, while residential buildings tend to cost less in terms of O&M cost.

It is also found that the building owner (the college or department that uses the building) is one of the most relevant independent variables affecting the studied buildings' LCC. Figure 11 shows the number of buildings by the college/owner and the initial cost per SF. The College of Engineering and the Department of Housing (residential buildings) own the most buildings. Buildings owned by the College of Engineering are, on average, more expensive. The residential buildings and parking decks are the least expensive facilities in terms of cost per SF.



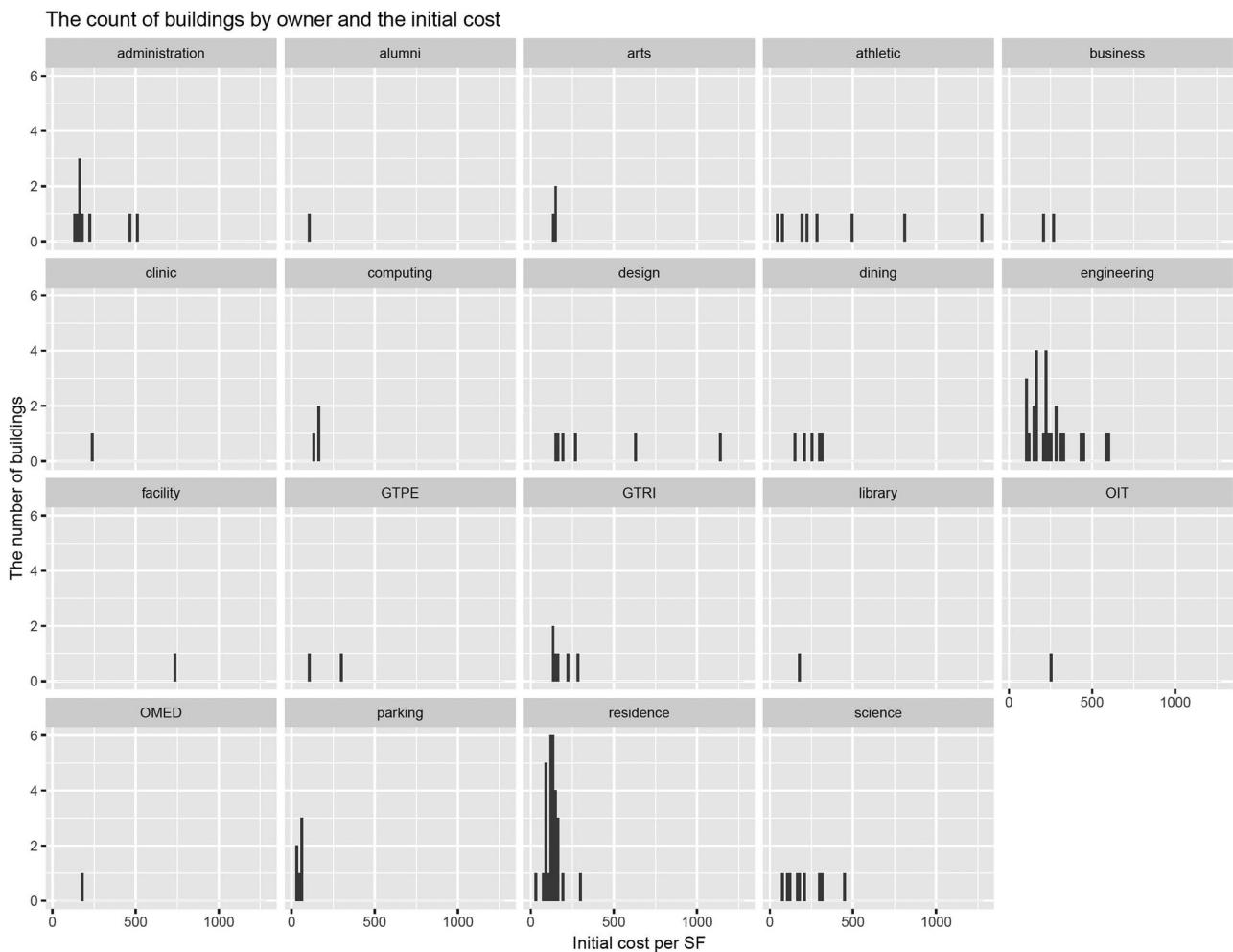
**Figure 10.** The linear correlations between variables.

Figure 12 presents the number of buildings by the college/owner and the utility cost per SF. Figure 13 provides the number of buildings by the college/owner and the O&M cost per SF. These figures indicate that the buildings owned by the College of Engineering generally cost more in terms of utilities and O&M. There is an athletic facility and an administrative building that cost over \$300 dollars per SF (present value in 1999) in utilities during the 20-year study period. The residential buildings and parking decks are also the most energy-efficient facilities and demand low maintenance cost.

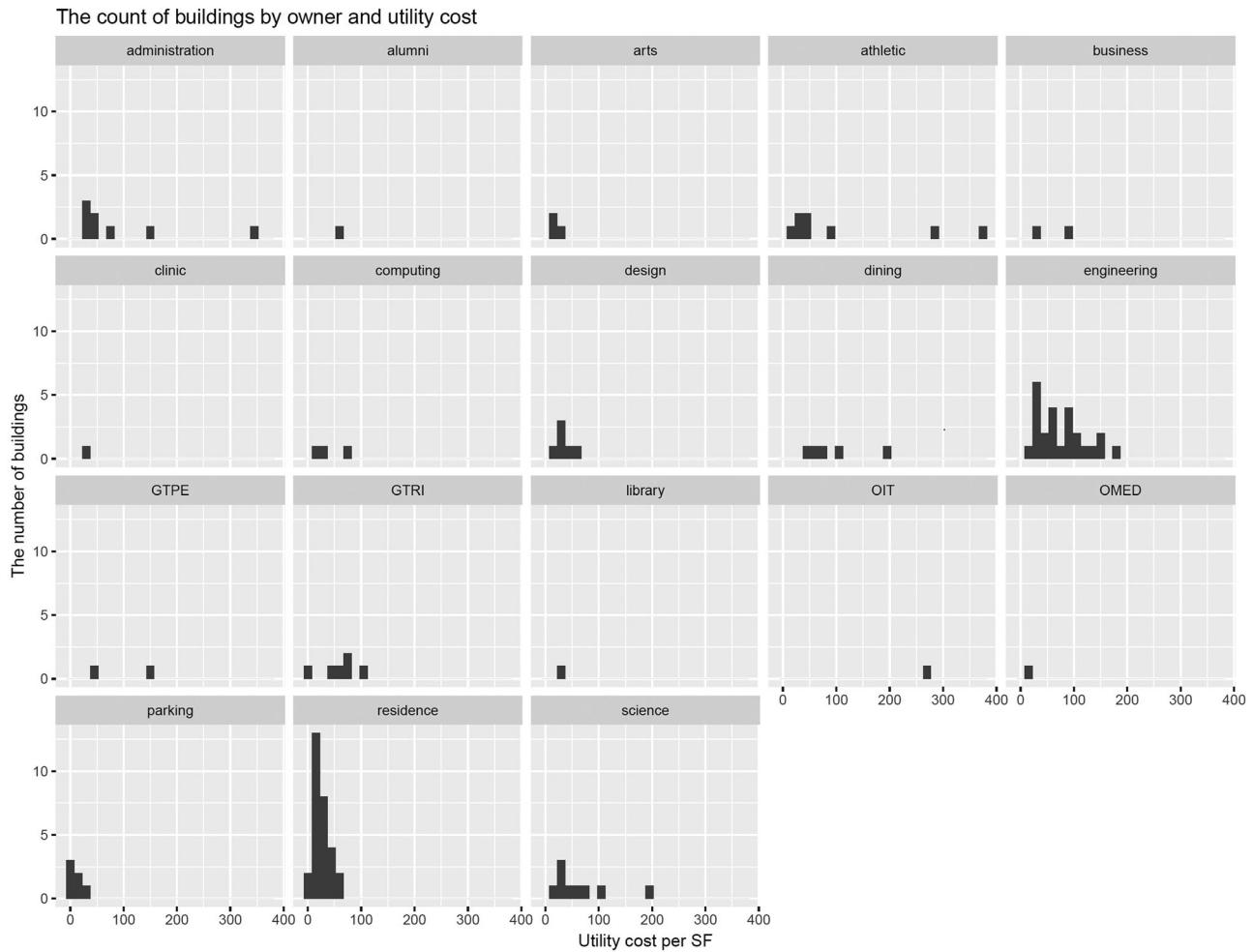
The experiment results for each machine learning algorithm (MAE) are presented in Table 4. When developing the model, the target attributes – initial cost, utility cost, and O&M cost – were normalized based on the mean and standard deviation of the overall data. Hence, the values of these target attributes range from 0 to 1. As a result of this, more intuitive results were derived to interpret and to compare the accuracy of each model.

## Discussions

In the case study, the MLR models serve as a control group because MLR is one of the most straightforward machine learning algorithms, which is based on the following assumptions: (1) the dependent variable and the independent variables are linearly correlated, (2) the independent variables are not highly correlated with each other, and (3) the residuals are normally distributed with a mean of 0 (Montgomery et al., 2012). After analysing the linear correlation of each independent variable in the experiment, it is found that some of the independent variables exhibit correlations with one another, such as the GSF and the number of floors (0.66), the office facilities percentage and the building service area percentage (0.47), and the residential facilities percentage and the office facilities percentage (-0.54). Moreover, there are too many independent variables (21) compared to the size of the dataset (123). Therefore, MLR models are expected to yield inaccurate prediction results, and the case study results confirmed the conclusions made



**Figure 11.** The count of buildings by the owner and initial cost.



**Figure 12.** The count of buildings by owner and utility cost.

by Kim, Yoon, et al. (2004) and Sajadfar and Ma (2015) in terms of the accuracy of MLR models.

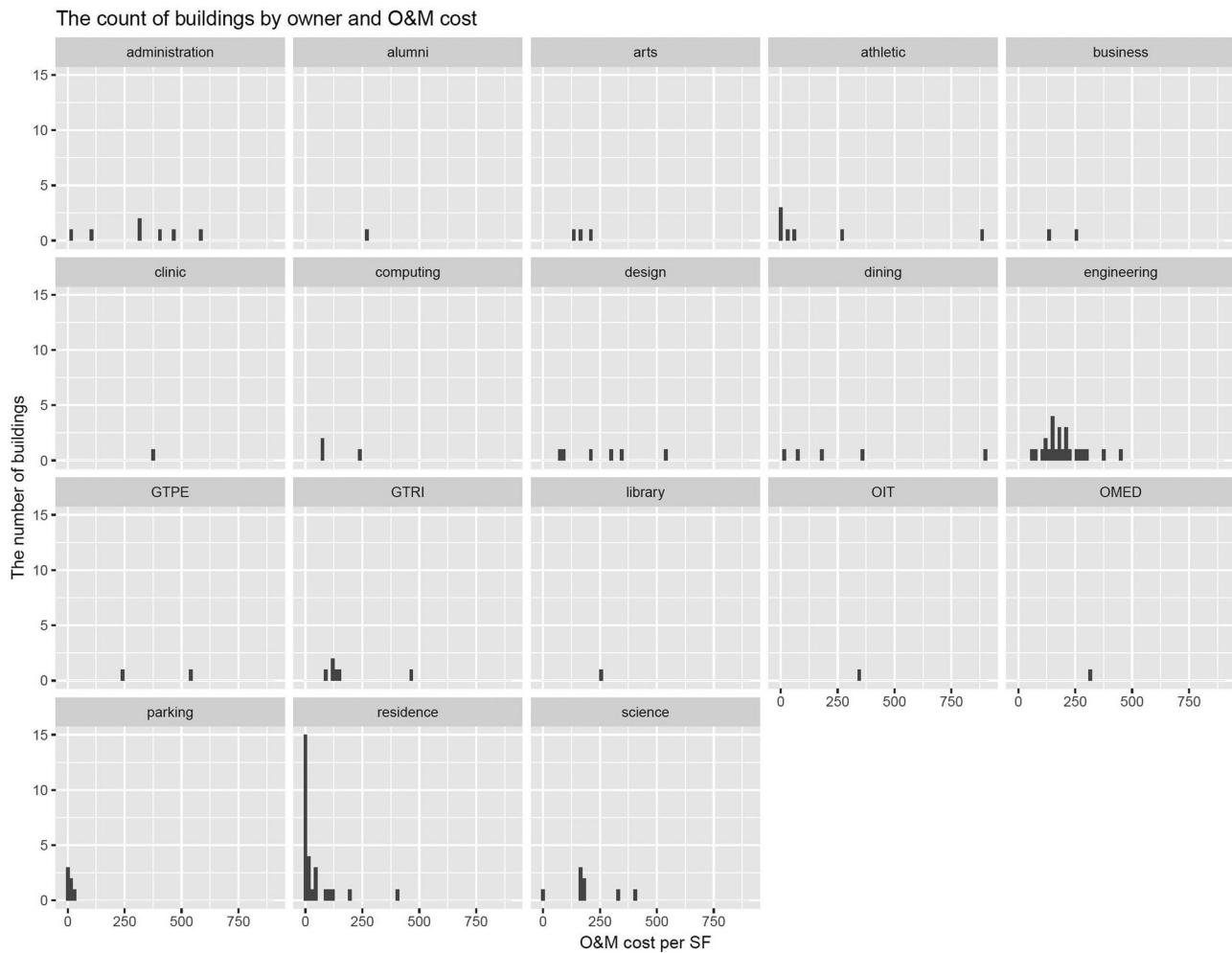
KNN is also a simple algorithm, which predicts the numerical target based on a similarity measure. Because it is a non-parametric lazy learning algorithm, no assumptions on the underlying data distribution are made; this means that KNN does not require the data to obey the typical theoretical assumptions, such as normal distribution. This makes KNN a better fit than MLR because most of the data in this experiment do not follow a typical distribution. The results confirmed this expectation: The KNN models significantly improved on the prediction accuracies of the MLR models, with MAE lowered by 12.66%, 16.90%, and 21.04% for initial cost, utility cost, and O&M cost, respectively.

A single regression tree tends to overfit the data (Kelleher et al., 2015). Random forest models combine the results of different decision trees, and can thus overcome the problem of overfitting, to a certain extent. Random forest models also have less variance than a single decision tree, which means that they perform better for a large

range of data items than single decision trees. The results of this experiment indicated that the random forest models slightly outperformed the KNN models, with MAE lowered by approximately 3% for each cost component. The performance of the random forest models, however, may have been compromised by the non-parametric nature of most of the data in this experiment.

The L2 regularization feature of SVM provides the generalization capability that can prevent it from overfitting (Awad & Khanna, 2015). Using kernel trick, SVM can also efficiently handle non-linear data (Hofmann, 2006), which makes it ideal for this experiment's dataset. The evaluation results indicated that the SVM regression models generated the most accurate predictions. Based on the results of 100 experiments, the single-target regression models using SVM had average MAEs of 27.97%, 37.08%, and 39.40% for the predictions of initial cost, utility cost, and O&M cost, respectively.

The MLP models' poor performance was not anticipated. During the testing phase to determine the optimal MLP structure, some of the MLP models achieved an



**Figure 13.** The count of buildings by the owner and O&M cost.

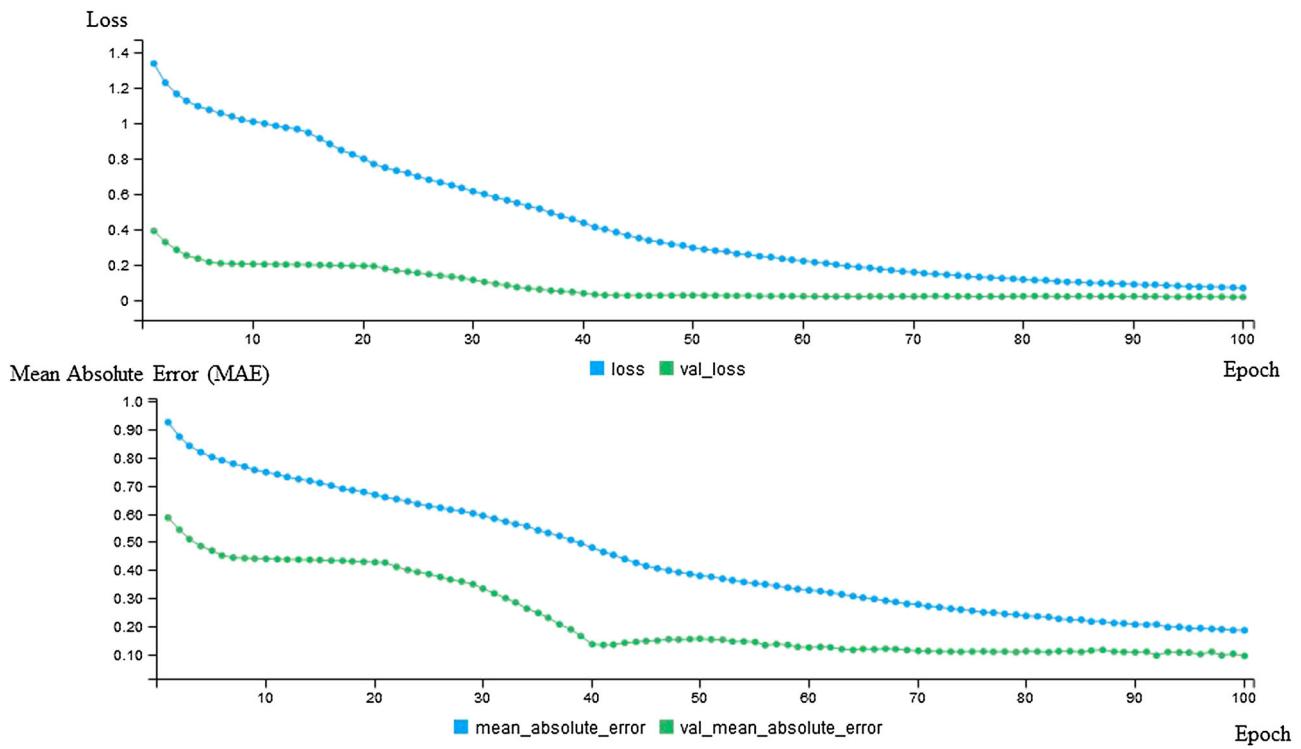
MAE as low as 12% to 18%. Figure 14 shows an example of the MLP model training process (100 epochs, batch size 32, validation split 0.1). In the experiment, however, because the training set and testing set were randomly selected in each loop, there were cases in which the instances (buildings) in the testing set had characters so different from those of the instances in the training set (such as in the case where all athletic facilities were in the testing set) that the trained MLP models

performed extremely poorly, with an MAE over 200%. This resulted in the unexpectedly high overall MAE of the MLP models. Moreover, because MLP models produce problem solutions without explaining the reason why the network behaves in this way, it is difficult to pinpoint other problems. Consequently, it is challenging to further improve MLP models.

It was expected that the single-target regression models would outperform their multi-target counterparts in terms of single cost component predictions. This is because single-target regression models are optimized for the single target rather than all the targets together. Each set of the LCC prediction results was generated by three single-target regression models, while one multi-target regression model can accomplish the same. In the case of MLP, however, the multi-target model outperformed the single-target model in the prediction of initial cost and utility cost, with the MAE lowered by 5.64% and 2.10%, respectively. Using the relationships between the target variables, MLP showed potential for being used to establish a complicated

**Table 4.** The evaluation results of each machine learning model in MAE (normalized)

		Initial cost (%)	Utility cost (%)	O&M cost (%)
Single-target regression model	MLR	46.01%	60.99%	62.49%
	KNN regression	33.35%	44.09%	41.45%
	Random forest	30.23%	41.07%	38.80%
	SVM regression	27.97%	37.08%	39.40%
	MLP	57.32%	59.81%	52.48%
Multi-target regression model	Multi-target random forest	46.44%	55.88%	58.84%
	Multi-target MLP	51.68%	57.71%	57.35%



**Figure 14.** An example of MLP training process, and the losses and MAE.

multi-target regression model to predict all the components of LCC.

Data scarcity is the major barrier that compromised the accuracy of the prediction models, as pointed out by Gallagher et al. (2018). Although a large amount of data were collected from multiple building systems, 123 instances (buildings) are still a small dataset for machine learning methods. Moreover, the building types of the studied university are diverse, and there are very few instances of some of the building types that were studied in this experiment. For example, the college campus had only one healthcare facility, one library, and three College of Computing buildings. The diversity of the studied buildings' characteristics rendered it difficult for the machine learning models to make highly accurate predictions based on the 21 descriptive attributes that were available during the programming phase.

Despite their limitations, the machine learning models developed in this experiment demonstrated their usefulness in facility LCC prediction. The SVM models had an average MAE of 27.97% for initial cost predictions. This level of accuracy is on a par with a rough order of magnitude estimate provided by an expert (with 28 years' experience in construction cost estimation) during the programming phase. The advantage of the tool developed in the case study, however, lies in the fact that it can provide an LCC estimate, whereas an estimator typically only provides an estimate for

design and construction costs. Unlike with commercial or residential buildings, there is typically no effective method of predicting a college campus building's future utility and O&M costs during the programming phase. With the models developed in this experiment, the university's management now has the capability to predict a campus building's future utility and O&M costs.

The proposed framework offers guidance for formalizing knowledge in facility LCC analysis by capturing necessary information from diverse data sources and reasoning about the captured data with machine learning techniques. Because the framework was validated through a case study conducted on a university campus, this research offers the potential for developing a universally applicable machine learning-enabled LCC analysis framework for other types of organizations with multiple facilities. College campus buildings are diverse in terms of functions and costs. If the proposed framework is effective for universities, there is a strong possibility that it could be applicable for infrastructure, government buildings, military bases, commercial buildings, and residential buildings. Further case studies are needed, however, to demonstrate the generalizability of the proposed framework.

From a practitioner's perspective, by exploring the new possibility for a more accurate prediction of a facility's LCC through leveraging machine learning and historical data housed in heterogeneous building systems, the

authors have achieved more than simply studying the LCC of an individual project in the programming phase. This achievement involves demonstrating an example of the use of machine learning-enabled automation to facilitate decision making in facilities management. By utilizing current machine learning approaches, the present work transformed historical building data into actionable knowledge: 'This is how much you will spend over the next 20 years on this building.' This research provides a knowledge base for decision makers to access whenever they need to take a building's LCC into account. One of the advantages of this knowledge base is that it can evolve over time: Existing available data are used to predict facility LCC, and new data can be incorporated as they become available. It is an iterative knowledge accumulation of facility costs that can not only identify performance trends and the 'hot spots' of utility and O&M expense, but also contribute to help identifying the best practices in facility design, construction, and operation from a cost-efficiency perspective.

This research has two major limitations. First, the case study used for validating the proposed LCC framework was limited to developing machine learning models for overall LCC predictions during the programming phase. This framework is applicable to all building design, construction, and facilities management phases, and machine learning models for analysis of a building's LCC can be developed as soon as the relevant data become available. The models developed in this experiment can only predict the lump sums of the initial cost, utility cost, and O&M cost, respectively. With more detailed building cost data, such as the cost breakdown according to CSI MasterFormat structure or UniFormat structure, machine learning models for more detailed cost estimation could be developed based on the proposed framework. Second, no benchmarking tool (the baseline) was available to evaluate the improvements in prediction accuracy provided by the developed models. The studied university did not have a prediction tool to use during the programming phase. The university's budget planning and administration department, and facilities management departments had not used the historical data for building LCC predictions before. Typically, cost estimators are hired to perform the initial cost prediction, but these data were not available to compare with the predictions produced by the developed machine learning models. Moreover, the utility and O&M cost estimation did not have a comparison base, because the estimation of these costs, if any, is typically conducted after the design has become available. The stakeholders in the university lacked a viable tool to conduct LCC analysis of utility and O&M costs during the programming phase.

## Conclusion

In this research, the authors have demonstrated that many current building systems already contain the data needed for facility LCC analysis. The utility and O&M costs can be derived from the raw data generated by and housed in the corresponding systems, such as BAS, BEMS, and CMMS. Most of the descriptive attributes needed for machine learning can be found in BIM and building management systems such as IWMS and space management system. Although some organizations have many buildings that are not equipped with all the above-mentioned systems, the general trend is that building managers and operators are increasingly relying on building management and control systems in their daily work. It can be expected that most large organizations operating multiple facilities will, if they have not already done so, adopt these building systems in the near future.

This research contributes to the body of knowledge through the authors' innovative implementation of machine learning to predict the total LCC of facilities, using historical data stored in building systems. The proposed framework minimizes human involvement to the greatest possible extent. People make mistakes: The higher the number of people involved in the data processing and analysis process, the greater the risk that the analysis will be exposed to human errors. In addition, some stakeholders tend to be very protective of their money-related data, which makes collecting historical data extremely difficult (Weerasinghe et al., 2016). In this research, the authors bypassed some of the existing barriers in cost analysis by using data directly derived from building systems. This more transparent approach provides reliable insights into facility LCC patterns.

There are some opportunities for further applications of the proposed LCC analysis framework. First, BIM could serve as the platform for both acquiring building data and presenting LCC knowledge. In this research, the LCC analysis results are presented in a 'one-dimensional' form, in tables containing numbers. Future studies could develop a BIM-based presentation platform to allow for visualization of the LCC analysis results in a multi-dimensional form that would be comprehensible and intuitive for stakeholders. For example, future research could involve the study of using BIM platforms to provide data visualizations of the analysis results.

## Acknowledgments

This work was supported by the Digital Building Laboratory (DBL) at Georgia Institute of Technology and the Centre for the Development and Application of Internet of Things Technologies (CDAIT) at Georgia Institute of Technology (project

No. 4906638). The authors appreciate all the members for their gracious support and input.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Data availability statement

The data that support the findings of this study are available from the Georgia Institute of Technology. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available from the authors with the permission of the Georgia Institute of Technology.

## ORCID

Xinghua Gao  <http://orcid.org/0000-0002-3531-8137>  
Pardis Pishdad-Bozorgi  <http://orcid.org/0000-0003-4208-9755>

## References

- Abate, D., Towers, M., Dotz, R., Romani, L., & Miller, J. (2014). *The Whitestone facility maintenance and repair cost reference 2014–2015*. Santa Barbara, CA: Whitestone Research.
- Aerospace System Design Laboratory. (2019). EnergyWatch for Georgia Institute of Technology. Retrieved from [energywatch.gatech.edu](http://energywatch.gatech.edu)
- Aibinu, A. A., Dassanayake, D., Chan, T.-K., & Thangaraj, R. (2015). Cost estimation for electric light and power elements during building design: A neural network approach. *Engineering, Construction and Architectural Management*, 22(2), 190–213. doi:10.1108/ECAM-01-2014-0010
- Allaire, J., Chollet, F., Tang, Y., Falbel, D., Van Der Bijl, W., Studer, M., & Keydana, S. (2019). keras: R Interface to ‘Keras’. Retrieved from [CRAN.R-project.org/package=keras](https://CRAN.R-project.org/package=keras)
- Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge, MA: MIT Press.
- Alshamrani, O. S. (2017). Construction cost prediction model for conventional and sustainable college buildings in North America. *Journal of Taibah University for Science*, 11(2), 315–323. doi:10.1016/j.jtusci.2016.01.004
- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185.
- Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable & Sustainable Energy Reviews*, 81, 1192–1205. doi:10.1016/j.rser.2017.04.095
- 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. doi.org/10.1016/j.buildenv.2006.06.007
- AssetWorks. (2018). AiM Operations & Maintenance (O&M). Retrieved from <https://www.assetworks.com/iwms/aim-oandm/>
- ASTM. (2017). *Standard practice for measuring life-cycle costs of buildings and building systems*. West Conshohocken, PA: Author.
- Au-Yong, C. P., Ali, A. S., & Ahmad, F. (2014). Prediction cost maintenance model of office building based on condition-based maintenance. *Maintenance and Reliability*, 16(2), 324.
- Awad, M., & Khanna, R. (2015). *Support vector regression efficient learning machines* (pp. 67–80). Berkeley, CA: Apress.
- Bala, K., Ahmad Bustani, S., & Shehu Waziri, B. (2014). A computer-based cost prediction model for institutional building projects in Nigeria: An artificial neural network approach. *Journal of Engineering, Design and Technology*, 12(4), 519–530. doi:10.1108/JEDT-06-2012-0026
- Banihashemi, S., Ding, G., & Wang, J. (2017). Developing a hybrid model of prediction and classification algorithms for building energy consumption. In F. Alam, R. Jazar, & H. Chowdhury (Eds.), *1st international conference on energy and power, ICEP2016* (Vol. 110), pp. 371–376.
- Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 1–22.
- Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., & Li, S. (2019). FNN: Fast Nearest Neighbor search algorithms and applications. Retrieved from [CRAN.R-project.org/package=FNN](https://CRAN.R-project.org/package=FNN)
- Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. (2018). Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: comparison with machine learning approaches. *Energies*, 11(7), doi:10.3390/en11071636
- Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2019). RandomForest: Breiman and Cutler’s random forests for classification and regression. Retrieved from [CRAN.R-project.org/package=randomForest](https://CRAN.R-project.org/package=randomForest)
- Bureau of Labor Statistics. (2018a). Average energy prices, Atlanta-Sandy Springs-Roswell – September 2018. Retrieved from [www.bls.gov/regions/southeast/news-release/averageenergyprices\\_atlanta.htm](http://www.bls.gov/regions/southeast/news-release/averageenergyprices_atlanta.htm)
- Bureau of Labor Statistics. (2018b). Consumer price index. Retrieved from [www.bls.gov/cpi](http://www.bls.gov/cpi)
- Bureau of Labor Statistics. (2019). Current Employment Statistics – CES (National) Retrieved from [www.bls.gov/ces/](http://www.bls.gov/ces/)
- Chapelle, O., & Vapnik, V. (2000). Model selection for support vector machines. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in neural information processing systems*. Cambridge, MA: MIT Press.
- Cheng, M.-Y., Tsai, H.-C., & Sudjono, E. (2010). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*, 37(6), 4224–4231. doi.org/10.1016/j.eswa.2009.11.080
- Chou, J.-S., & Ngo, N.-T. (2016). Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. *Applied Energy*, 177, 751–770. doi:10.1016/j.apenergy.2016.05.074
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.

- 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., 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. doi:10.1016/j.enbuild.2017.12.031
- Dibike, Y. B., Velickov, S., & Solomatine, D. (2000). *Support vector machines: Review and applications in civil engineering*. Proceedings of the joint workshop on applications of AI in civil engineering, Cottbus-2000, Germany.
- Dogan, S. Z., Ardit, D., & Gunaydin, H. M. (2006). Determining attribute weights in a CBR model for early cost prediction of structural systems. *Journal of Construction Engineering and Management*, 132(10), 1092–1098. doi:10.1061/(asce)0733-9364(2006)132:10(1092)
- Dogan, S. Z., Ardit, 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*, 134(2), 146–152. doi:10.1061/(asce)0733-9364(2008)134:2(146)
- Dursun, O., & Stoy, C. (2016). Conceptual estimation of construction costs using the multistep ahead approach. *Journal of Construction Engineering and Management*, 142(9), doi:10.1061/(asce)co.1943-7862.0001150.
- 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 ed.). Hoboken, NJ: Wiley.
- Fuller, S. (2010). *Life-cycle cost analysis (LCCA)*. Retrieved from [www.wbdg.org/resources/life-cycle-cost-analysis-lcca](http://www.wbdg.org/resources/life-cycle-cost-analysis-lcca)
- Gallagher, C. V., Leahy, K., O'Donovan, P., Bruton, K., & O'Sullivan, D. T. J. (2018). Development and application of a machine learning supported methodology for measurement and verification (M&V) 2.0. *Energy and Buildings*, 167, 8–22. doi:10.1016/j.enbuild.2018.02.023
- Gao, X., & Pishdad-Bozorgi, P. (2018). *Past, present, and future of BIM-enabled facilities operation and maintenance*. Construction Research Congress 2018.
- Gao, X., & Pishdad-Bozorgi, P. (2019a). BIM-enabled facilities operation and maintenance: A review. *Advanced Engineering Informatics*, 39, 227–247. doi:10.1016/j.aei.2019.01.005
- Gao, X., & Pishdad-Bozorgi, P. (2019b). Current research developments of machine learning-based facility cost prediction techniques. Retrieved from [https://osf.io/7cb6q/?view\\_only=7b5af8b3455941e88c46faafaaf9d3bb](https://osf.io/7cb6q/?view_only=7b5af8b3455941e88c46faafaaf9d3bb)
- Gao, X., & Pishdad-Bozorgi, P. (2019c). Independent variables significantly affecting facility initial costs, utility costs, and O&M costs. Retrieved from [https://osf.io/waj3e/?view\\_only=7b5af8b3455941e88c46faafaaf9d3bb](https://osf.io/waj3e/?view_only=7b5af8b3455941e88c46faafaaf9d3bb)
- Gao, X., & Pishdad-Bozorgi, P. (2019d). Questionnaire for estimators and project managers. Retrieved from [https://osf.io/x8w6y/?view\\_only=7b5af8b3455941e88c46faafaaf9d3bb](https://osf.io/x8w6y/?view_only=7b5af8b3455941e88c46faafaaf9d3bb)
- Gao, X., Pishdad-Bozorgi, P., Shelden, D., & Hu, Y. (2019). *Machine learning applications in facility life-cycle cost analysis: A review*. Paper presented at the 2019 ASCE international conference on computing in civil engineering, Atlanta, GA.
- Gelfusa, M., Malizia, A., Murari, A., Parracino, S., Lungaroni, M., Peluso, E., ... Gaudio, P. (2015). First attempts at measuring widespread smoke with a mobile LiDAR system.
- Georgia Institute of Technology. (2018). The facility data of Georgia Tech. Retrieved from <http://ionsrv2.fac.gatech.edu/ion/default.aspx?dgm=/IONSVR2/ION-Ent/config/diagrams/ud/network.dgm&node=&logServerName=QUERYSERVER.IONSVR2&logServerHandle=327952>
- Georgia Institute of Technology. (2019). Georgia Tech Capital Planning & Space Management System. Retrieved from [https://tableau.gatech.edu/t/CPSM/views/BuildingInformationDashboard/Intro?3Aembed=y&3AshowShareOptions=true&3Adisplay\\_count=no&3AshowVizHome=no](https://tableau.gatech.edu/t/CPSM/views/BuildingInformationDashboard/Intro?3Aembed=y&3AshowShareOptions=true&3Adisplay_count=no&3AshowVizHome=no)
- Geysen, D., De Somer, O., Johansson, C., Brage, J., & Vanhoudt, D. (2018). Operational thermal load forecasting in district heating networks using machine learning and expert advice. *Energy and Buildings*, 162, 144–153. doi:10.1016/j.enbuild.2017.12.042
- Gleick, J. (2011). *The information: A history, a theory, a flood*. New York, NY: HarperCollins.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.
- Haykin, S. S. (2009). *Neural networks and learning machines*. Upper Saddle River, NJ: Pearson.
- Hofmann, M. (2006). Support vector machines-kernels and the kernel trick. *Notes*, 26, 3.
- 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. doi:10.1016/j.eswa.2010.08.071
- Hu, Y., & Castro-Lacouture, D. (2019). Clash relevance prediction based on machine learning. *Journal of Computing in Civil Engineering*, 33(2), 04018060.
- 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. doi:10.1016/j.enbuild.2016.09.068
- Jafarzadeh, R., Wilkinson, S., Gonzalez, V., Ingham, J. M., & Amiri, G. G. (2014). Predicting seismic retrofit construction cost for buildings with framed structures using multilinear regression analysis. *Journal of Construction Engineering and Management*, 140(3). doi:10.1061/(asce)co.1943-7862.0000750
- Jain, R. K., Smith, K. M., Culligan, P. J., & Taylor, J. E. (2014). Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy*, 123, 168–178. doi:10.1016/j.apenergy.2014.02.057
- Ji, C., Hong, T., Jeong, K., & Leigh, S.-B. (2014). A model for evaluating the environmental benefits of elementary school facilities. *Journal of Environmental Management*, 132, 220–229. doi:10.1016/j.jenvman.2013.11.022
- 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). doi:10.1061/(asce)co.1943-7862.0001072
- Johnson Controls Inc. (2018). Metasys® Building Automation System. Retrieved 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*. London, UK: 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. doi:10.1016/j.buildenv.2004.02.013
- 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. doi:10.1016/j.buildenv.2004.03.009
- 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. doi:10.1016/j.eswa.2011.01.063
- Koo, C.-W., Hong, T., Hyun, C.-T., Park, S. H., & Seo, J. (2010). A study on the development of a cost model based on the owner's decision making at the early stages of a construction project. *International Journal of Strategic Property Management*, 14(2), 121–137. doi:10.3846/ijspm.2010.10
- Krstić, H., & Marenjak, S. (2017). Maintenance and operation costs model for university buildings. *Technical Gazette*, 24, 200. doi:10.17559/TV-20140606093626
- Kvam, P. H., & Vidakovic, B. (2007). *Nonparametric statistics with applications to science and engineering*. Hoboken, NJ: Wiley.
- 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. doi:10.3130/jaabe.11.87
- Li, H., Shen, Q. P., & Love, P. E. D. (2005). Cost modelling of office buildings in Hong Kong: An exploratory study. *Facilities*, 23(9/10), 438–452. doi:10.1108/02632770510602379
- Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management*, 132(7), 750–758. doi:10.1061/(ASCE)0733-9364(2006)132:7(750)
- Microsoft. (2019). Microsoft R. Retrieved from [mran.microsoft.com](http://mran.microsoft.com)
- Milion, R. N., Paliari, J. C., & Liboni, L. H. B. (2016). Improving consumption estimation of electrical materials in residential building construction. *Automation in Construction*, 72, 93–101. doi:10.1016/j.autcon.2016.08.042
- Mitchell, T. M. (1997). *Machine learning*. Norwell, MA: McGraw-Hill.
- Mocanu, E., Nguyen, P. H., Kling, W. L., & Gibescu, M. (2016). Unsupervised energy prediction in a smart grid context using reinforcement cross-building transfer learning. *Energy and Buildings*, 116, 646–655. doi:10.1016/j.enbuild.2016.01.030
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*. Hoboken, NJ: Wiley.
- Moon, J., Park, J., Hwang, E., & Jun, S. (2018). Forecasting power consumption for higher educational institutions based on machine learning. *Journal of Supercomputing*, 74(8), 3778–3800. doi:10.1007/s11227-017-2022-x
- Ntrepid Corporation. (2018). Ion data grabber. Retrieved from [ion.ntrepidcorp.com](http://ion.ntrepidcorp.com)
- openrefine.org. (2018). OpenRefine. Retrieved from <http://openrefine.org/>
- Pantic, M. (2019). Introduction to machine learning & case-based reasoning. Retrieved from <https://ibug.doc.ic.ac.uk/media/uploads/documents/courses/syllabus-CBR.pdf>
- Park, B. R., Choi, E. J., Hong, J., Lee, J. H., & Moon, J. W. (2018). Development of an energy cost prediction model for a VRF heating system. *Applied Thermal Engineering*, 140, 476–486. doi:10.1016/j.applthermaleng.2018.05.068
- Pishdad-Bozorgi, P. (2017). Future smart facilities: State-of-the-art BIM-enabled facility management. *Journal of Construction Engineering and Management*, 143(9), 02517006.
- 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. doi:10.1016/j.autcon.2017.12.004
- Probability Theory Group. (2019). e1071: Misc functions of the department of statistics. Retrieved from [cran.r-project.org/web/packages/e1071/index.html](http://cran.r-project.org/web/packages/e1071/index.html)
- R-project.org. (2019). Documentation for package 'stats' version 3.6.0. Retrieved from [stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html](http://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html)
- Rahman, R. (2019). MultivariateRandomForest: Models multivariate cases using random forests. Retrieved from [CRAN.R-project.org/package=MultivariateRandomForest](http://CRAN.R-project.org/package=MultivariateRandomForest)
- Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017). Machine learning approaches for estimating commercial building energy consumption. *Applied Energy*, 208, 889–904. doi:10.1016/j.apenergy.2017.09.060
- Romani, L., Abate, D., Miller, J., & Dotz, R. (2014). *The Whitestone facility operation cost reference 2014–2015*. Santa Barbara, CA: Whitestone Research.
- Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptions and the theory of brain mechanism*. Washington, DC: Spartan Books.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (No. ICS-8506). San Diego: California University; La Jolla Institute for Cognitive Science. Retrieved from <https://apps.dtic.mil/dtic/tr/fulltext/u2/a164453.pdf>
- Sajadfar, N., & Ma, Y. (2015). A hybrid cost estimation framework based on feature-oriented data mining approach. *Advanced Engineering Informatics*, 29(3), 633–647. doi:10.1016/j.aei.2015.06.001
- Sala-Cardoso, E., Delgado-Prieto, M., Kampouropoulos, K., & Romeral, L. (2018). Activity-aware HVAC power demand forecasting. *Energy and Buildings*, 170, 15–24. doi:10.1016/j.enbuild.2018.03.087
- SAS. (2018). Machine learning: What it is & why it matters. Retrieved from [https://www.sas.com/it\\_it/insights/analytics/machine-learning.html](https://www.sas.com/it_it/insights/analytics/machine-learning.html)
- Shi, H. W., & Li, W. Q. (2008). *The integrated methodology of rough set theory and artificial neural-network for construction project cost prediction*. In *2008 second international symposium on intelligent information technology application* (Vol. 2, pp. 60–64). IEEE.

- Smola, A. J. (1996). *Regression estimation with support vector learning machines* (Master's thesis). Technische Universität München.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
- Sonmez, R. (2008). Parametric range estimating of building costs using regression models and bootstrap. *Journal of Construction Engineering and Management*, 134(12), 1011–1016. doi:10.1061/(ASCE)0733-9364(2008)134:12(1011)
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics*. Boston, MA: Pearson.
- Tableau. (2019). Tableau. Retrieved from [www.tableau.com](http://www.tableau.com)
- The INSITE Consortium. (2019). INSITE Retrieved from [www.insite.org](http://www.insite.org)
- Trost, S. M., & Oberlender, G. D. (2003). Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *Journal of Construction Engineering and Management*, 129(2), 198–204.
- 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.
- US Inflation Calculator. (2019). Historical inflation rates: 1914–2018. Retrieved from <https://www.usinflationcalculator.com/inflation/historical-inflation-rates/>
- Weerasinghe, A., Ramachandra, T., & Rotimi, J. O. B. (2016). *A simplified model for predicting running cost of office buildings in Sri Lanka*. Paper presented at the proceedings of the 32nd annual ARCOM conference.
- Zayed, T. M., & Halpin, D. W. (2005). Productivity and cost regression models for pile construction. *Journal of Construction Engineering and Management*, 131(7), 779–789.
- Zhang, C., Cao, L., & Romagnoli, A. (2018). On the feature engineering of building energy data mining. *Sustainable Cities and Society*, 39, 508–518. doi:10.1016/j.scs.2018.02.016