A FRAMEWORK FOR DEVELOPING MACHINE LEARNING MODELS FOR FACILITY LIFE CYCLE COST ANALYSIS THROUGH BIM AND IOT

A Dissertation
Presented to
The Academic Faculty

by

Xinghua Gao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Design

Georgia Institute of Technology
May 2019

A FRAMEWORK FOR DEVELOPING MACHINE LEARNING MODELS FOR FACILITY LIFE CYCLE COST ANALYSIS THROUGH BIM AND IOT

Approved by:

Dr. Pardis Pishdad-Bozorgi, Advisor and Chair
School of Building Construction
College of Design
Georgia Institute of Technology

Dr. Dennis R. Shelden
School of Architecture
College of Design
Georgia Institute of Technology

Dr. Javier Irizarry
School of Building Construction
College of Design
Georgia Institute of Technology

Dr. Duen Horng Chau
School of Computational Science & Engineering
College of Computing
Georgia Institute of Technology

Andreas Kristanto
Vice President
RIB U.S. COST

Date Approved: March 29, 2019

To My Loving Parents and Wonderful Wife

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Pardis Pishdad-Bozorgi for all the help and guidance that she has given me over the past four years. She is honest, strict, inspiring, and always has the student's best interest in mind. She has set an example of excellence as a researcher, a teacher, and a mentor. I consider myself extremely lucky to have been one of her Ph.D. students. I especially would like to extend my sincere gratitude to Professor Dennis R. Shelden who supported me on this journey and for his guidance. He has been a role model in both research and leadership. I would also like to thank Professor Javier Irizarry for his valuable discussions and helpful recommendations. I would like to express my gratitude to other members of my doctoral committee, Professor Duen Horng Chau and Mr. Andreas Kristanto, for their time and guidance through this process.

I am truly grateful for the love and constant encouragement of my parents and parents-in-law and for the support they have given me in all my pursuits. My heartfelt thanks and appreciations to my four loving parents.

I would like to thank the faculty and staff, and all my friends in the College of Design and the School of Building Construction, especially Professor Daniel Castro-Lacouture, Ms. Lorie Wooten, Ms. Laura Alger, Dr. Jianli Chen, Yuqing Hu, Dr. Yang Cao, Hao Lu, and Yuping Liang for their support and friendship through the years.

Finally, many warm thanks for her and just for her, my wife Chen Wu, for her daily care, concern, support, patience, tolerance, and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AECO	Architecture, Engineering, Construction, and Owner-operated
AI	Artificial Intelligence
API	Application Programing Interface
BIM	Building Information Modeling
BAS	Building Automation System
BFA	Building floor area
BFDP	Basic Facility Data Package
CIP	Cast-in-place
CMAR/CM at risk	Construction Manager at Risk
CMMS	Computerized Maintenance Management System
CPSMS	Capital Planning & Space Management System
CSV	Comma-separated Values
GA	Genetic Algorithm
GFA	Gross floor area
GIA	Gross internal area
GSF	Gross square footage
GRV	Gross building volume
HVAC	Heating, ventilation, air-conditioning
ICT	Information and communication technology
IFC	Industry Foundation Classes
IoT	The Internet of Things
IWMS	Integrated Workspace Management System

JSON	JavaScript Object Notation
KPI	Key performance indicator
LCC	Life-cycle Cost
LCCA	Life-cycle Cost Analysis
LEED	Leadership in Energy and Environmental Design
MVD	Model View Definition
O&M	Operation and maintenance
SLRM	Simple Linear Regression Model
SMS	Space Management System
UFA	Usable floor area
XML	Extensible Markup Language

SUMMARY

A large amount of resources are spent on constructing new facilities and maintaining the existing ones. The total cost of facility ownership can be minimized by focusing on reducing the facilities life-cycle costs (LCC) rather than the initial design and construction costs. This thesis presents a research project that developed a machine learning-enabled facility life-cycle cost analysis (LCCA) framework using data provided by Building Information Models (BIM) and the Internet of Things (IoT).

First, a literature review and a questionnaire survey were conducted to determine the independent variables affecting the facility LCC. The potential data sources were summarized, and a data integration process introduced. Then, the framework for developing machine learning models for facility LCCA was proposed. A domain ontology for machine learning-enabled LCCA (LCCA-Onto) was developed to encapsulate knowledge about LCC components and their roles in relation to sibling ontologies that conceptualize the LCCA process. After that, a series of experiments were conducted on a university campus to demonstrate the application of the proposed machine learning-enabled LCCA framework. Finally, the author's vision of the future smart built environment was discussed.

This research contributes to the body of knowledge by investigating the feasibility of forecasting facilities' LCCs by implementing machine learning on historical data. By exploring the new possibility for better prediction of a facility' LCC through leveraging historical data housed in heterogeneous building systems across a continuous network of buildings, this research has a greater impact than simply studying the LCC of an individual

project in the design phase. The impact involves data-based LCC inputs in future facilities thus enabling cost benchmarking and informing project developments based on owned historical data. Using existing available data to benchmark facility costs can assist decision making, and new data can be incorporated as they become available. It is an iterative knowledge accumulation of facility costs that could not only identify performance trends and operation and maintenance expense “hot spots”, but also identify the best practices of facility design, construction, and operation from a cost efficiency perspective.

CHAPTER 1 INTRODUCTION

This chapter gives an overview of the study and briefly explains the opportunities and challenges that machine learning has introduced to the domain of facility Life-cycle Cost Analysis (LCCA). Building Information Modeling (BIM) and the Internet of Things (IoT) are proposed as solutions to the challenges of developing LCCA machine learning models, such as data insufficiency. The significance of developing a comprehensive and generalizable framework for BIM and IoT-enabled LCCA machine learning model developments is also discussed. In addition, this chapter outlines the research objectives, hypotheses, research scope, and contributions.

1.1 Research Motivation and Problem Statement

Because of the long life spans of buildings, robust decisions regarding the economic efficiency of alternative materials, components, and systems demand a full lifecycle perspective that goes beyond the initial cost and regular maintenance and repair [1]. The LCCA has become increasingly important in new building design and existing building retrofitting, refurbishment, and renovations. However, despite its importance, researchers and industry professionals are facing challenges when practicing LCCA in the Architecture, Engineering, Construction, and Owner-operated (AECO) industry. Two of the main barriers are the shortage of life-cycle cost (LCC) data [2,3] and the complexity of predicting real future costs [3,4].

Currently, most LCCA methods, such as the ones introduced in [5,6], assume that we can estimate a building component's LCC by knowing its price, life expectancy, and

the cost of all the operating and maintenance activities associated with it [2]. However, the real service lives and costs of many buildings and their systems are difficult to predict for multiple reasons. One is that there is always a mismatch between the predicted energy performance of buildings and actual measured performance, typically addressed as "the performance gap" [7]. Another reason is that many building systems and components, with proper maintenance and repair, can function beyond the warranty, which makes their true costs are difficult to predict because the facility owners typically do not know how much money and labor is needed to repair them when malfunction after the warranty expires. Moreover, even the same type of systems used in different buildings may have different LCCs because the monetary and labor costs vary depending on each facility manager's operational profile on building systems.

Machine learning is an automated process that extracts patterns from data [8]. In the field of predictive data analytics, machine learning is a method used to devise complex prediction algorithms and models [8,9]. These analytical models enable data analysts to uncover hidden insights, predict future values, and produce reliable, repeatable decisions through learning from historical relationships and trends in the data [10]. As a viable alternative to simulation tools, machine learning techniques can give an accurate quantitative estimation of energy demand for different building systems [11] and predict facility related costs [12]. However, there are gaps in research regarding the development of machine learning models for LCCA. They are listed as follows:

Overall facility LCC prediction. Although machine learning techniques have been implemented in forecasting construction costs, utility consumption, and Operation and maintenance (O&M) costs, respectively, its application in predicting a building's whole

LCC is rarely found in the literature. More studies that utilize machine learning to predict a building's overall LCC and shed light on the underlying relationships between each cost components (initial design and construction costs, O&M costs, utility costs, etc.) are needed.

Generalizable machine learning frameworks for facility LCCA. Most of the developed machine learning models are only applicable to one type of building projects, such as housing [13,14], educational buildings [15], and office buildings [16]. The nature of predictive models involves assumptions and simplifications based on the similarities of the studied subjects. The uniqueness of different building projects basically makes it impossible to use one model to predict more than one type of building's LCC [14,17,18]. However, it is possible to establish generalizable frameworks for developing facility LCCA machine learning models. These frameworks would specify the means and process of 1) identifying potential descriptive attributes (the input of machine learning models), such as by literature review and survey, 2) data acquisition, such as by exporting from BIM, the Building Automation System (BAS), and the Computerized Maintenance Management System (CMMS), by finding the records in drawings and specifications, and by survey, 3) attributes selection, 4) machine learning algorithm selection, and 5) model evaluation. Currently, this kind of frameworks is yet to be developed.

Data availability, accessibility, and quality. Many research challenges discussed in this field can be attributed to data insufficiency, including a lack of sufficient metering and accessibility, and poor data quality [19]. The machine learning models in many studies were established based on a very limited data set [20,21]. As Milion et al. [22] pointed out, "data survey is the most difficult challenge in estimation studies". Limited and uncertain

information make the accurate prediction of construction-related costs difficult [23]. The lack of reliable and consistent data also limits the application of LCCA in the early design stage [24]. What data to record and how organizations should record the facility data for machine learning-based LCCA are seldom discussed in the literature.

1.2 Hypothesis, Research Objectives, and Research Questions

An IoT is a network that connects uniquely identifiable “things” that have sensing/actuation and potential programmability capabilities [25]. “Through the exploitation of unique identification and sensing, information about the ‘Thing’ can be collected and the state of the ‘Thing’ can be changed from anywhere, anytime, by anything” [25]. IoT envisions a future in which digital and physical entities can be linked through embedded identification, sensing, and/or actuation capabilities to enable various innovative applications and services that improve the quality of human life.

Building Information Modeling (BIM) is "an improved planning, design, construction, operation, and maintenance process using a standardized machine-readable information model for each facility, new or old, which contains all appropriate information created or gathered about that facility in a format useable by all throughout its lifecycle” [26,27]. The referred building information model is “the shared digital representation of physical and functional characteristics of any built object” [26,28,29]. BIM is a maturing paradigm for the development of higher-level semantic information assets for facilities [26]. For the past two decades, the AECO industry has been evolving from two-dimensional symbolic drawing documents to BIM: three-dimensional, metadata rich, object models representing building spaces, components, and their relations. With the advancement of

cloud data stores and now IoT, BIM models offer a clear potential as the "digital twin" of the built environment. In recent years, the proliferation of BIM has provided designers and builders with new opportunities to achieve better quality buildings at lower cost and shorter project duration [26,30-32]. BIM technology has the potential to provide value to the owners and operators by offering them a powerful means to retrieve information from a virtual model of the facility [33].

The author's hypothesis is that BIM and the IoT network embedded in evolving building systems – such as BAS, CMMS, and Building Energy Management Systems (BEMS) – already contain many valuable data for LCCA but not being used because they are not connected, available to analysts in a consumable way. By extracting relevant data from BIM and IoT, integrating them on a comprehensive building data platform, and implementing machine learning on the data, we can have a better understanding of the facility's LCC and overcome multiple barriers of current LCCA methods, and thus to achieve more informed decisions in building design, construction, and facility management.

This research systematically investigates the feasibility of forecasting facilities' LCCs by implementing machine learning on historical data. It proposes a comprehensive and generalizable framework for developing facility LCCA machine learning models. This framework specifies the data requirements, methods, and expected results in each step of the model development process. It is a guidance for formalizing knowledge in facility LCCA by capturing necessary information from diverse data sources and reasoning about the captured data with machine learning techniques. It is envisioned that by capturing and analyzing historical data relevant to facility costs, tacit knowledge of LCCA can be semi-automatically extracted and formalized, which will reduce the reliance on individual

researchers for knowledge formalization. A series of experiments were conducted to validate the proposed framework and to demonstrate its implementation process.

The primary objective of this research is to develop a data-driven knowledge formalization approach for facility LCCA in new and renovation construction projects. Secondary objectives are to use the formalized knowledge to improve LCCA efficiency, facilitate structured learning from historical data, provide guidance for facility cost-related data management, and support the decision making of capital planning and facility management. The following research questions have been developed in support of the research objectives.

Question 1 – Data requirements. What are the factors that have a significant influence on facility LCC and can be explicitly captured?

Question 2 – Data acquisition. Where to find the data and how to efficiently extract the data from the data source(s)?

Question 3 – LCC component derivation. How to derive the LCC components (initial cost, utility cost, and O&M costs) from the data collected?

Question 4 – Analysis and evaluation. Which machine learning method(s) yields the best prediction results? How effective is the developed machine learning models?

1.3 Research Scope and Contributions

Numerous costs are associated with the design, construction, installation, operating, maintaining, and disposing of a building or building system. According to [6], Building-related costs usually fall into the following categories:

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

This research investigates how to implement machine learning on the historical data to forecast the costs of the first four categories – initial costs, utility costs, O&M costs, and replacement costs. The prediction of residual values, finance charges, and non-monetary benefits or costs are out of this research's scope. In addition, this research only considers monetary costs. Other cost factors such as environmental impacts and human welfares are not studied in this research.

This research contributes to the body of knowledge by investigating the feasibility of forecasting facilities' LCCs by implementing machine learning on historical data. By exploring the new possibility for better prediction of a facility' LCC through leveraging historical data housed in heterogeneous building systems across a continuous network of

buildings, this research has a greater impact than simply studying the LCC of an individual project in the design phase. The impact involves data-based LCC inputs in future facilities thus enabling cost benchmarking and informing project developments based on owned historical data. Using existing available data to benchmark facility costs can assist decision making, and new data can be incorporated as they become available. It is an iterative knowledge accumulation of facility costs that could not only identify performance trends and operation and maintenance expense "hot spots" [34], but also identify the best practices of facility design, construction, and operation from a cost efficiency perspective.

The proposed LCCA framework 1) specifies the potentially influential factors pertaining to the whole LCC of a facility, 2) utilizes BIM and IoT (embedded in heterogeneous building systems) as the data sources to provide robust data stream for analysis, and 3) implements multiple machine learning algorithms to forecast each critical LCC component and analyze their interrelationships. Compared with conventional LCCA methods, the proposed approach has improvements in the following aspects:

Implementing multiple machine learning algorithms. This research discusses multiple applicable machine learning algorithms for facility LCC prediction. In the data analysis field, there is no one solution or one approach that fits all. Different algorithms have different requirements and applicable scenarios. In the experiments of this study, with the LCC-related data in the building systems extracted, cleaned, and stored in one database, multiple machine learning algorithms were implemented to develop LCC forecasting models and the results are comparatively analyzed. The proposed framework can serve as a comprehensive guideline that directs researchers and practitioners to implement machine learning in facility LCCA.

Data source. Many challenges for machine learning implementations in the AECO sector discussed in the literature can be attributed to data insufficiency, including a lack of sufficient metering and accessibility, and poor data quality [12]. With the developments of building systems, an increasing amount of facility-related data are being generated, such as the progressively detailed energy consumption data in the BAS and maintenance work order history in CMMS. However, this kind of historical data has not yet been widely used to forecast facility costs. This research develops innovative methods to use the data housed in BIM and IoT (ubiquitous building systems) to solve the data insufficiency issue.

Limited human intervention and improved transparency. The data inquiry process of LCCA can be challenging for many researchers because some stakeholders tend to be very protective of the money-related data. Moreover, people make mistakes – more people involved in the data processing and analysis process usually means more human errors. Using BIM, IoT, and relevant software applications, the proposed framework minimizes human involvements to the greatest extent possible. Only the generation of maintenance work order records still rely on human input; other data are extracted and processed in an automated or semi-automated fashion without any human intervention. The knowledge related to LCC is developed directly based on the raw data generated by the building systems rather than depending on human-made reports. This more transparent approach provides reliable insights into facility LCC patterns than conventional LCCA methods.

Value and Applicability in the whole life cycle. Many building LCCA research studies are focusing on LCCA for design decisions but it is also important during the facility operation phase. The recursive LCCA can serve as both a cost prediction for facility changes (retrofitting/refurbishment/renovation) and an indicator to identify areas to

improve – find out which buildings, systems, or devices cost more resources than they should. The knowledge developed through the proposed framework is valuable and applicable in the whole life cycle of a facility.

1.4 Research Methodology and The Organization of the Thesis

The scope of this research is to study the feasibility of utilizing data housed in BIM and facility IoT network to predict the facility LCC through machine learning. The overall research framework is shown in Figure 1.

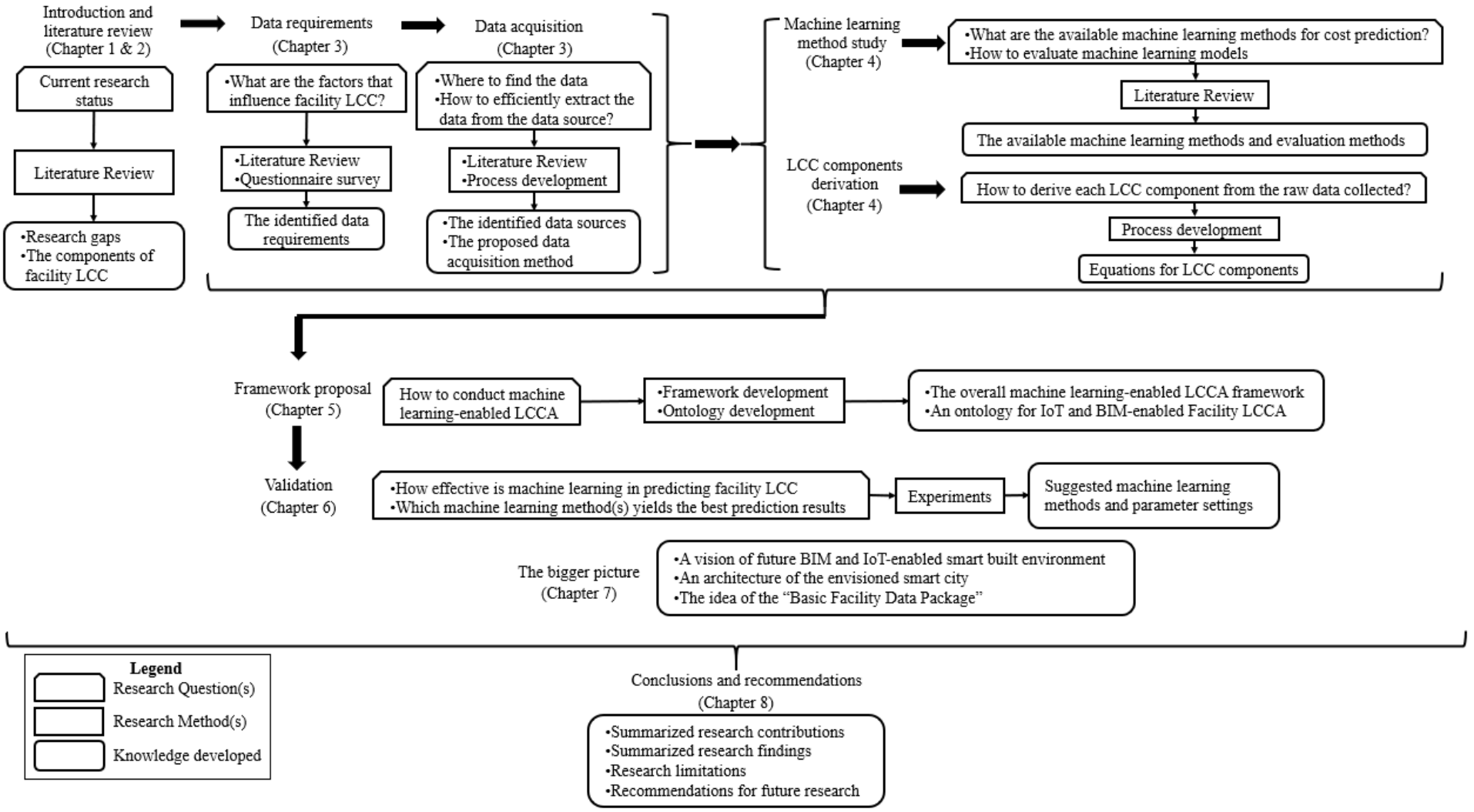


Figure 1. 1 The overall research framework

This thesis is organized in a sequence that highlights a step-by-step progression of the development of the proposed machine learning LCCA framework and its realization by means of the proof-of-concept. There are eight chapters that form the thesis and they are organized as followed:

Chapter 2 – Literature Review

Chapter 2 investigates how LCCA is carried out in the current building cost estimation practice and summarizes the machine learning applications in facility cost prediction. It identifies the challenges faced in the current LCCA practice and the research gaps in machine learning for LCCA in the AECO industry. The combination of BIM and IoT is proposed as a viable solution to these challenges and provides a means to partially address the research gaps. In addition, the BIM applications in facilities management and LCCA are also summarized and discussed in this chapter.

Chapter 3 – Building Data Collection and Integration

Chapter 3 answers the first two research questions: what are the factors that have a significant influence on facility LCC and can be explicitly captured, and where to find the data and how to efficiently extract the data from the data source(s)?

The data requirements for facility LCCA, the potential data sources, and data acquisition and integration methods are discussed in this chapter. A literature review and a questionnaire survey were conducted to determine the independent variables affecting the initial construction costs, utility consumption costs, and O&M costs. Major building systems – the BAS, CMMS, BEMS – are summarized and the data they can provide are discussed. This chapter also proposes a data

integration framework that enables the utilization of the data housed in separate building systems for facility LCCA.

Chapter 4 – Deriving LCC Components through Machine Learning

Chapter 4 answers the third research question: how to derive the LCC components (initial cost, utility cost, and O&M costs) from the data collected?

This chapter first proposes the framework for machine learning-based facility LCC component prediction, and then summarizes the most commonly used machine learning methods for facility cost prediction. It also discusses the attribute selection process and the applicability of these methods in the prediction of each LCC component – initial cost, utility cost, and O&M cost. The validation methods and performance measures are also discussed.

Chapter 5 – The Overall LCCA Process

A framework for developing machine learning models for facility LCCA is presented in Chapter 5. A domain ontology for machine learning-enabled LCCA (LCCA-Onto) is developed to encapsulate knowledge about LCC components and their roles in relation to sibling ontologies that conceptualize the LCCA process. This domain ontology is then tailored to be the cornerstone (the knowledge base) that will enable the automated LCCA data collection from building systems. The contents presented in this section can be used as an institutional guideline for facility LCC monitoring and prediction.

Chapter 6 – Proof-of-Concept Validation

Chapter 6, in a case specific context, answers the last research questions: which machine learning method(s) yields the best prediction results? How effective is the developed machine learning models?

This chapter presents a series of experiments conducted on a university campus using the proposed LCCA framework. It also discusses the implementation process of the framework in detail. The overview of the experiments is presented first. Then, the data acquisition, processing, and integration works are demonstrated. After that, Machine learning model developments, evaluation, and comparison are discussed. The findings of the experiments are discussed at the end of this chapter.

Chapter 7 – The BIM and IoT-enabled Smart Built Environment

Chapter 7 extends the discussion to the author's vision of the future BIM and IoT-enabled smart built environment. First, the background of IoT, Cyber-physical System (CPS), and the Smart Built Environment are introduced. Then, an architecture of the envisioned smart city is presented and the idea of the "Basic Facility Data Package" (BFDP), which is the foundation of the data infrastructure for the envisioned smart city, is proposed and discussed. This research is one of the use cases enabled by the BFDP and a proof of concept for the envisioned future BIM and IoT-enabled smart built environment.

Chapter 8 – Conclusions

Chapter 8 summarizes the contributions, findings, and limitations of this research and gives recommendations for future research.

1.5 Additional Notes

Because “Building Information Modeling” and “Building Information Model” are often used interchangeably [26], in this thesis, the abbreviation “BIM” is used to represent both the Building Information Modeling (the related process and technology) and the Building Information Model (the digital representation of built objects).

CHAPTER 2 LITERATURE REVIEW

This chapter investigates how LCCA is carried out in the current building cost estimation practice and summarizes the machine learning applications in facility cost prediction. It identifies the challenges faced in the current LCCA practice and the research gaps in machine learning for LCCA in the AECO industry. The combination of BIM and IoT is proposed as a viable solution to these challenges and provides a means to partially address the research gaps. In addition, the BIM applications in facilities management and LCCA are also summarized and discussed in this chapter.

2.1 The Life-cycle Cost Analysis in the AECO Industry

LCCA is an analysis technique that "encompasses the total cost of a system over a specified period in its lifetime" [35]. The concept of LCC was first applied by the US Department of Defense (DoD) [36]. Its importance received the attention of DoD because of the findings that operation and support costs for typical weapon systems accounted for approximately 75% of the total cost [35,36]. In the AECO industry, LCCA refers to a method for assessing the total cost of the facility ownership, involving all costs of acquiring, owning, and disposing of a building or building system [6]. It is used as "an economic evaluation tool for choosing among alternative building investments and operating strategies by comparing all of the significant differential costs of ownership over a given time period in equivalent economic terms" [37]. This subsection discusses 1) the typical process of LCCA, 2) the components of building LCC, 3) common assumptions, 4) LCC calculation models, 5) guidelines and data sources, and 5) the challenges.

2.1.1 The Typical Process of LCCA

Currently, the main purpose of building LCCA is to consider whether increased initial costs – such as more durable materials or equipment that require less operation and maintenance costs – can be compensated by the savings in the future [37]. Although there are no universal procedures that assure all analyses reflect similar definitions or assumptions, several guidelines have been developed for LCCA. A standard LCCA practice established by American Society for Testing and Materials (ASTM International) specified the steps to calculate the LCC for a building or building system [5]:

1) Identify objectives, alternatives, and constraints.
2) Establish basic assumptions for the analysis.
3) Compile cost data.
4) Compute the LCC for each alternative.
5) Compare LCCs of each alternative to determine the one with the minimum LCC.
6) Make the final decision, based on LCC results as well as consideration of risk and uncertainty, unquantifiable effects, and funding constraints (if any).

The estimators have to forecast the LCC of each alternative option for each building element. The LCC of each option may consist of several cost items, whose performance and cost data should be acquired because the cost data are expressed in unit rates and the estimator has to predict each cost based on the option's physical characteristics [2]. In addition, the building-wide costs, such as energy consumption, insurance, and janitorial costs, are also estimated. Then, all costs are discounted, added up, and projected over the building's life cycle" [2].

2.1.2 The Components of Building LCC

Numerous costs are associated with the design, construction, installation, operating, maintaining, and disposing of a building or building system. According to [6], Building-related costs usually fall into the following categories:

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

2.1.3 Common Assumptions in LCCA

The National Research Council [37] summarized the common assumptions in LCCA, as follows:

1) All costs are measurable in monetary terms. In practice, LCCA is often restricted to financial costs alone – the expense of purchasing building-related goods and services.

2) All alternatives in LCCA deliver the same performance throughout their service lives. This assumption is violated, for example, when decisions made to reduce future maintenance costs resulted in indoor air quality problems.

3) There is an easy interchangeability of present and future costs. For example, “spending more initially to purchase durable materials or systems to achieve future savings in a building's maintenance costs, or choosing less costly and durable materials and systems that can be maintained at higher cost (i.e., by painting or lubrication) to yield the same service” [37].

Typical LCCA does not consider many uncertainties in cost estimates, such as a system or material may fail prematurely, or maintenance efforts may be ineffective [37]. These uncertainties may invalidate the results of LCCA.

2.1.4 LCC calculation models

In the industry, LCCA refers to the utilization of fundamental economic evaluation approaches – such as the annual worth method, the net present value method, and the savings/investments ratio method – to evaluate the various cash flows of facility’s LCC [38].

There are three possible ways for modeling LCC computation, involving the deterministic calculation model, the stochastic calculation model, and the fuzzy calculation model [39].

The deterministic calculation model computes the present value of the LCC from the time series of projected cash flows; all future costs and benefits (LCC components) are discounted to the net present value (NPV) using a certain discount rate, and then added up [39]. Sensitivity analysis can be carried out to test the LCC variation with the changes of input parameters. The following is a generic present value formula for the LCC deterministic calculation model [39]:

$$LCC = C_p + \sum_{t=0}^{n} \frac{C_t}{(1+d)^t}$$

Where:

LCC is the present value of a facility's total ownership cost.

C_t is the sum of facility LCC.

n is the number of years of the studied period.

d is the discount rate.

C_p is the initial capital costs.

The stochastic calculation model assumes each cost element, the discount rate, and the study period are randomly distributed according to one of the probability distribution forms, such as the normal distribution and the gamma distribution [39]. Each cost element is treated stochastically and the present value of cash flow in each year is described as probability density functions (PDF) or uncertain cash flow profiles. If the PDF or cash flow profile of each LCC component and the discount parameter are known or can be simulated, the total LCC can be estimated using the following equation [39]:

$$f(PV) = f\left(C_p\right) + \sum_{t=0}^{n} \frac{f(C_{ti})}{(1 + f(d))^t}$$

Where:

$f(PV)$ is the probability distribution function of total LCC in present value.

$f(C_{ti})$ is the probability distribution of LCC with a center of i in period t.

n is the number of years of the studied period.

$f(d)$ is the probability distribution of discount rate.

$f(C_p)$ is the probability distribution of initial capital costs.

The fuzzy calculation model, considering human and processes subjectivity, estimates LCC and present value parameters by using expert judgment and statistical techniques [39]. Fuzzy set theory is proposed as a device for modeling fuzzy variables in a mathematical domain [40]. The term 'fuzzy' refers to the situation where ambiguity and vagueness exist [41]. Estimators use fuzzy numbers to calculate present values of LCC and thus overcome the difficulty of uncertainty. For example, the following is a formula for computing fuzzy present value [42]:

$$\widetilde{PV} = \left(\sum_{t=0}^{n} \left(\frac{\max\left(P_t^{l(y)}, 0\right)}{\prod_{t'=0}^{t}\left(1 + r_{t'}^{r(y)}\right)} + \frac{\min\left(P_t^{l(y)}, 0\right)}{\prod_{t'=0}^{t}\left(1 + r_{t'}^{l(y)}\right)} \right), \sum_{t=0}^{n} \left(\frac{\max(P_t^{r(y)}, 0)}{\prod_{t'=0}^{t}(1 + r_{t'}^{l(y)})} + \frac{\min(P_t^{r(y)}, 0)}{\prod_{t'=0}^{t}(1 + r_{t'}^{r(y)})} \right) \right),$$

Where:

$P_t^{l(y)}$ is the left representation of the cash at time t.

$P_t^{1(y)}$ is the right representation of the cash at time t.

$r_t^{l(y)}$ is the left representation of the interest rate at time t.

$r_t^{r(y)}$ is the right representation of the interest rate at time t.

There are formulas for the analyses of fuzzy present value, fuzzy equivalent uniform annual value, fuzzy future value, fuzzy benefit-cost ratio, and fuzzy payback period [43,44].

2.1.5 Guidelines and data sources

Several organizations have established standard or recommended procedures for evaluating the LCC of a building or building system and for comparing the LCC of alternative building designs that satisfy the same functional requirements. These organizations involve the National Institute of Standards and Technology (NIST) [6,45,46], U.S. Department of Energy (DOE) Federal Energy Management Program (FEMP) [47], American Society of Civil Engineers (ASCE) [48], National Research Council [37], and ASTM International [5].

Organizations such as NIST [49], RSMeans data [50], and Whitestone [51,52], publish annual statistics for the energy price [49], the discount factor [49], and the expected maintenance, repair, and replacement cost [50,51], and the operation cost [52], basing on facility characteristics such as the facility type, location, age, etc. These data sources provide estimators with the statistic-based LCC component reference, which answers questions such as "How much does it cost to maintain a facility over its service lifetime? What is the historic inflation rate of maintenance and repair construction costs? How do maintenance and repair costs vary across different areas? What

is the lifetime of a specific asset component ?" [51]. In the industry, LCCA is typically based on these statistic-based data.

2.1.6 Challenges

There are several important challenges in the application of LCCA in the AECO industry. These challenges can be grouped into two major categories: the technical challenges and the management challenges.

The three major technical challenges involve:

1) Data availability and data acquisition frameworks: the basis of LCCA is the historical data of each LCC component, without which the application of LCCA is not possible. However, the availability of LCC data is rather limited [2,37]. The main reason for this is the lack of any frameworks or mechanisms for collecting and storing the data [53]. For example, [37] pointed out that "the accounting systems used by building managers and contractors seldom make it possible to identify accurately the costs of maintenance and repair of specific components".
2) The difficulty in predicting real future costs: many factors that influence the future costs of building components are uncertain. These factors involve external factors, such as discount rates and energy prices [4], and internal factors, such as the future behavior of materials and mechanical and electrical systems [37]. Forecasting these factors over a long period of time is challenging. In addition, the assumptions reflected in LCCA parameter selections – such as building service life and discount rate – may or may not reflect well the conditions that actually occur in the future [37]. The results of LCCA

depend on these assumptions and different assumptions may require different courses of action [37].

3) The complexity of application: LCCA is a rather complex exercise. Because improvements in one area may have negative effects in others, complex iterations between alternative building components may make it difficult to select the best option [2]. As the number and range of alternative increase, the level of effort required in LCCA increases rapidly, while the ability of an estimator to find the best option is always limited by the data and time available [37].

The three major management challenges involve:

1) Incentives for builders to reduce LCC: capital costs and operating expenses are usually met by different parties, hence there is no incentive for builders to reduce the operation costs [54]. In addition, many decision makers tend to minimize the initial expenditures to increase the return on investment or to meet budgetary restrictions [37].
2) Short term versus long term: owners or operators with the short-term responsibility for a building may fail to consider effectively the longer-term impact of their decisions on the building's O&M requirements [37]. The National Research Council found that “the most difficult obstacles to controlling total costs of ownership are those raised by administrative procedures and managerial or political decisions driven by short-term gains” [37]. For example, budgeting processes that divorce capital costs and operating expenses make it difficult to identify and manage total costs of ownership [37].
3) Restrictions: many restrictions are hindering building managers and operators of government agencies from selecting the best option to achieve the lowest LCC [37]. These restrictions involve legislative budget procedures, procurement regulations that

limit design specificity to enhance competition, and administrative separation of responsibilities for design, construction, and maintenance [37].

2.2 Machine Learning Applications on Facility Cost Prediction

"Machine learning" is a broad term that refers to multiple techniques that give computer systems the ability to "learn" with data [55]. Depending on whether the input and output are available to a learning system, these techniques can be categorized as following types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [56]. Supervised learning is suitable for building the models used in predictive data analytics applications [8]. In this research, the author focus exclusively on supervised learning and use the terms "supervised learning" and "machine learning" interchangeably. The purpose of supervised learning techniques is to establish a relationship model between a set of descriptive attributes (also referred as descriptive features) and a target attribute (also referred as target features) based on a set of historical instances, and then this model can be used to make predictions for new instances [8]. For example, the two hundred buildings' electricity consumption data in the past decade can be used to predict the following five years' electricity consumption of a new campus building, or an existing one. In this case, the descriptive attributes can be the building's floor area, purpose, age, hours of operation, occupant density, vendors of building systems, etc.; and the target attribute is the electricity consumption. The relationship between them may be found by supervised learning, and if it is found, the relationship model can be established and be used to predict any existing or new building's future electricity consumption.

The basic process of machine learning consists of three phases – the training phase, the validation phase, and the application phase [57]. In the training phase, the model can be trained by

mapping the input descriptive attributes with expected output target attribute. The dataset used to train the model is called the “training set”. Then, in the validation phase, the quality of the model is evaluated by testing the validation data (validation set) and observing the outputs. The quality of the model depends on the size of data, the algorithm used, the assumptions, etc. In the application phase, the established model is used to predict future values.

In recent years, the developments of machine learning techniques provide building experts with new opportunities to achieve more accurate predictions of facility-related costs in the early design phase or even programming phase. This section synthesizes and presents a summary of current research developments of machine learning techniques for LCC. The identified research gaps are discussed in Section 1.1 Research Motivation and Problem Statement.

2.2.1 Initial Construction 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 [13,14,16]. The construction costs prediction studies can be categorized into three major groups based on the methods used, which are 1) regression analysis [20,58-63] 2) Case-based Reasoning [13,14,16,64,65], and 3) Artificial Neural Network [17,21,66-68].

Studies have been conducted to compare the cost prediction performance of models based on different machine learning methods. For example, Kim et al. [69] compared the accuracy of Multiple Regression Analysis (MRA), Artificial Neural Network (ANN), and Case-based Reasoning (CBR) by experimenting on 530 residential buildings’ construction costs. The results

indicated that although the ANN model yields more accurate results than the MRA and the CBR models, the CBR model performed 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 et al. [70] concluded that the cost forecasting models based on ANN outperform regression models in predicting the costs of light wiring, power wiring, and cable pathways. Sajadfar & Ma [71] compared the prediction accuracies of the models based on Linear Regression, Multilinear Regression, K-Nearest Neighbors (KNN) Regression, Decision Tree Regression, and ANN. They found that the ANN model shows to be the most accurate for welding operations.

The most commonly used descriptive attributes for developing the construction cost prediction models in above-mentioned references involve 1) building floor area, 2) number of floors, 3) structure type, 4) number of rooms, 5) roof type, 6) foundation type, 7) topography & soil condition, and 8) construction duration.

2.2.2 Utility Consumption 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 [72,73]. 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 the utility costs in detail [74-76]. The most commonly used

machine learning methods for energy forecasting involve: 1) Artificial Neural Network [74,75,77], 2) Support Vector Machines (SVM) Regression [78,79], and 3) Case-based Reasoning [80,81].

Most of the reviewed studies in the utility consumption prediction field developed multiple machine learning models and compared their performance [82,83]. For example, Geysen et al. [84] developed a thermal load forecasting system that incorporates a collection of machine learning methods – linear regression (LR), extremely randomized trees regression (ETR), ANN, and SVM regression. The experiment results indicated that the LR performs worst while the ANN and ETR are slightly better than the SVM. The study conducted by Moon et al. [76] also showed that the ANN-based model outperforms the SVM regression-based model in electric load forecasting. However, Idowu et al. [85]'s study showed SVM gave the best prediction performance compared to ANN and multiple linear regression in forecasting the thermal load in district heating substations.

Although attribute importance (weight) depends on the specific machine learning model, yet certain attributes will always dominate the attribute space [72]. The most commonly used descriptive attributes for utility consumption models are 1) building age, 2) building function/type, 3) building floor area, and 4) number of floors. Advances in machine learning techniques enabled researchers to develop prediction models without a large quantity of data. Li et al. [86] proposed an extreme deep learning approach that can extract most influential building energy consumption attributes and improve the prediction accuracy.

2.2.3 Operation and Maintenance Costs Prediction

Studies on using machine learning to predict O&M costs are relatively rare. This is probably because obtaining accurate maintenance data is a challenging [87]. The most commonly used machine learning methods in O&M costs forecasting are multiple regression [15,24,88,89] and

ANN [15,90]. Au-Yong et al. [88] found that the characteristics of condition-based maintenance of the office buildings directly influence the cost performance. Based on these relationships, they developed a regression model for maintenance planning and prediction. Krstić & Marenjak [89] developed a multiple regression model to predict the O&M costs for university buildings during the initial design phase. Li & Guo [15,91] developed maintenance cost prediction models for university buildings using simple linear regression, multiple regression, and back-propagation ANN. The results indicated that the back-propagation ANN model outperforms the other two models. They also found that, for university buildings, the first peak of renovation will be around 20 years of age and second peak 35; for a building with more than five floors, the first and second peak of renovation will be 15 years and 30 years, respectively.

The most commonly used descriptive attributes for developing the O&M costs prediction models involve 1) building age, 2) number of rooms, 3) building floor area, and 4) number of floors.

2.3 BIM Applications in Facilities Management and LCCA

Building Information Modeling (BIM) is one major progress in the Architecture, Engineering, Construction, and Operations (AECO) industry [26]. BIM technology involves the creation and utilization of digital building models. It has various applications in building design, construction, and facilities management, including 3D visualization, design checking, various building performance analysis, constructability checking, improved facility operation and maintenance, etc. BIM, as the digital representation of buildings, can serve as a comprehensive database that provides the building component information for the LCCA. In this research, BIM

serves as the building data source in the proposed framework. Descriptive attributes (model inputs) related to buildings are extracted from BIM in an automated fashion.

2.3.1 BIM Applications in Facility Management

Effective facility LCC management requires the operator to take proper O&M actions – operations, maintenance, repair, and renewal programs, because inadequate O&M efforts may raise the costs of ownership well above the levels anticipated in the LCCA [37]. For example, spending less on routine maintenance may substantially decrease the service life of the equipment and increase costs of repair and replacement. BIM has the potential to advance and transform facility management by providing a platform for owners and operators to retrieve, analyze, and process building information in a digitalized 3D environment, and thus to improve the efficiency and effectiveness of facility O&M. This subsection presents a review of recent publications on BIM applications on facilities management.

The most common problem facility managers face is the information accessibility issues [92]. During the facility operation phase, they usually do not have easy and quick access to the needed information to process work orders [93-95]. To address this issue, BIM is used to integrate the fragmented FM information that is housed within different building management systems and to provide an intuitive information access interface [93-103].

BIM-enabled date exchange and the integration of separate systems are extensively discussed. Industry foundation classes (IFC) has been used as the data exchange schema between BIM and CMMS [103,104], EIC [105], GIS [106], and BAS [107]. Researchers also used commercial software applications to perform the data exchange, such as AutoCAD Civil 3D as a means of data exchange between BIM and GIS [97] and Revit DB Link to enable exchanging data

between BIM and CMMS [108]. The concept of Central Facility Repository (CFR) proposed by the General Services Administration (GSA) describes the comprehensive information system built around a central facility database, which serves as the data foundation for FM-related software applications [109]. Although CFR is a promising development direction, current BIM-O&M studies seldom show the effort of establishing the CFR. Rather than creating a comprehensive CFR, loosely coupled system integration solutions are more commonly used approaches [94,103,104,107]. Shen et al. [110] claimed that the most promising system integration approach for the construction industry is the "distributed loosely coupled integration solution using intelligent agents and Web services technologies" because using a single central repository to store all the information is not a viable option due to "the fragmented nature and adversarial behavior that characterizes the industry" [94,111]. Using this approach, Shen et al. [94] developed a system framework for providing decision support to facility management and maintenance. One advantage of this approach is that it has better generalizability than establishing a centralized comprehensive FM database. In this research, the proposed approach can theoretically integrate BIM with multiple building systems, such as HVAC control system, local weather station, building façade monitoring system, equipment and people tracking system, equipment condition monitoring system, fire response, and evacuation simulation system. However, the article did not show much evidence of the developed system prototype and the authors did not conduct experiments to validate its effectiveness. In another example, Motamedi et al. [104] integrated the data housed in CMMS, Condition Assessment System (CAS), Computer Aided Facilities Management (CAFM), and the data in Construction Operations Building Information Exchange (COBie) format. They linked these systems' databases by identifying a unique ID for each building element and using it in all related applications. This kind of loosely coupled integration lies the

interoperability foundation for automated data transmission among the existing and proposed building systems. However, to fully realize the automation in data exchange still requires extensive work.

Once the required information is ready-to-use, the next question is how to make the information available and accessible for facility managers when needed and how to present the information in an intuitional fashion such that maintenance personnel can easily comprehend it. To achieve these goals, researchers have adopted barcodes, RFID, and Augmented Reality (AR) together with BIM to facilitate maintenance and repair activities [95,96,99,100,112].

Barcodes and RFID. Barcodes and RFID tags serve as identifications of building items to access the relevant information which is linked with the corresponding objects in the database. By scanning the barcode or RFID tag of the item, the mobile device will present the corresponding 3D BIM component and its information, such as instruction manuals, photos, videos of operations, maintenance history, and manufacturer information, etc. [100,113]. The RFID has some advantages over the 2D barcode. RFID tags can be scanned from a distance and do not require line-of-sight or clean environments, which are necessary for 2D barcodes [113]. Additionally, each RFID tag has a chip that can store some modifiable data, which gives the RFID tag some flexibilities when used in an environment not connected to the remote information server (ibid.). The RFID also has some shortcomings, including the interference among each tagged component and the interference between the tagged component and some materials [113,114].

Augmented Reality (AR). By providing the superimposed geometric representation on the physical space along with the relevant BIM-based facility information in real-time, AR provides a suitable interface for O&M fieldwork support [95,96,99]. Similar to barcode and RFID-based

systems, AR systems also require installing identification tags on facility items to identify them. As pointed out by Lee & Akin in [96], the computer vision-based AR technology can only recognize pre-defined physical markers to identify building components. Hence, the proposed AR-based BIM-maintenance system will not function properly if the markers cannot be seen clearly. Another challenge is that deploying physical markers to all the maintainable building components is neither economical nor realizable. These issues were partially addressed by a later study conducted by Koch et al. [99], who used natural markers (such as exit signs, position marks of fire extinguishers, and signs with textual information hints) as the defined visual markers that can be captured by the BIM-AR system. The relative position between the visual marker and the user, together with the camera's orientation, are used to locate the user's position; hence the maintenance related information can be accurately displayed on the screen, overlaying on top of the identified equipment. This system still has some limitations. It fails to locate the user's position and orientation when 1) no pre-defined natural marker is in the camera view, 2) the distance between two markers is larger than 10 meters, and 3) the same marker appears at multiple locations (ibid.).

2.3.2 BIM Applications related to Facility LCCA

Among the many publications on building cost prediction and analysis in the past decade, the ones that have discussed BIM applications only account for a little proportion. BIM-based frameworks have been developed for building economic assessments [115-119]. For example, Marzouk et al. [120] proposed a framework that integrates BIM with the green building rating system LEED (Leadership in Energy and Environmental Design) and calculates building LCC using Genetic Algorithms (GA) optimization. This framework acts as a decision-making tool to select optimum building materials by expanding the materials library in BIM software and using

Application Programming Interface (API) to integrate BIM with the developed GA model. This integration allows extracting data from the BIM to the GA model, importing the optimization results to BIM after analysis, and modifying BIM based on these results.

Some preliminary attempts of using BIM for facility LCCA have been made since 2003 [121]. Fu et al. presented a prototype of an IFC-based modeling tool that can predict building LCCs [121]. However, this article only described high-level ideas and did not demonstrate the detailed methodology of using IFC for LCCA nor the experiment to test the proposed prototype. Dawood et al. [122] proposed a life-cycle energy assessment framework that uses Revit models for visualization. Chen et al. [123] proposed a BIM-based framework for selection of cost-effective green building design. Hosny & Elhakeem [124] proposed a design analysis framework for buildings in the desert environment that aims to achieve environmental friendly designs with minimum LCC. This framework uses BIM to serve as an interactive database that houses the data useful for evaluating design success measures, such as the LCC and embodied energy based on a material breakdown.

Researchers also used BIM in their environmental impact analyses. Marzouk et al. [118,120] proposed two frameworks that use BIM to select optimum sustainable building materials and to calculate the building LCCs. Jalaei et al. [116] proposed a method that integrates BIM with decision-making approaches to optimize the selection of sustainable building components during the conceptual design phase. Liu et al. [117] proposed a BIM-based building design optimization method to optimize building designs and improve buildings' sustainability. They applied a revised particle swarm optimization (PSO) algorithm to calculate the trade-off between LCC and life cycle carbon emissions (LCCE) of building designs. Nour et al. [125] implemented a Genetic Algorithm (GA) with IFC, an energy simulation software program, and a LCC estimation model to achieve

“an optimal allocation of energy saving elements to buildings' external envelopes, the use of which allows for a positive return on additional investment in energy saving elements”. Shin & Cho [126] mapped the required information of facility LCCA and the information that could be obtained using the BIM, and conducted a case study in which three building facade alternatives were analyzed.

CHAPTER 3 BUILDING DATA COLLECTION AND INTEGRATION

This chapter answers the first two research questions: what are the factors that have a significant influence on facility LCC and can be explicitly captured, and where to find the data and how to efficiently extract the data from the data source(s)?

The data requirements for facility LCCA, the potential data sources, and data acquisition and integration methods are discussed in this chapter. A literature review and a questionnaire survey were conducted to determine the independent variables affecting the initial construction costs, utility consumption costs, and O&M costs. Major building systems are summarized and the data they can provide discussed. This chapter also proposes a data integration framework that enables the utilization of the data housed in separate building systems for facility LCCA.

3.1 The Data Requirements of Facility LCCA

The first and most challenging task of conducting LCCA for a buildings is to determine the economic effects of alternatives and to quantify these effects and express them in monetary amounts [6]. The author's hypothesis is that by extracting and formatting the LCC-related data generated by and housed in different building and computerized systems, and applying appropriate machine learning techniques, we can forecast each LCC component of a new building as early as the programming phase. A literature review and a questionnaire survey were conducted to determine the potentially influential factors that affect the overall LCC, which are also the potential descriptive attributes of the LCCA machine learning models.

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, EBSCO, WILEY, ASCE, CIB, SPRINGER, T&F, and ISPRS. Articles with abstracts contain “machine learning” or “prediction” and the keywords “building cost”, “energy consumption”, “operation cost”, and “maintenance cost” are identified and reviewed. 88 related publications are identified and reviewed. Each reviewed paper is examined in the following aspects: 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) has a case study/experiment or not, and 7) the size of the data set. The results and findings regarding the descriptive attributes used in the machine learning models are summarized in Table 3.2 to 3.4. Some influential factors not mentioned in the literature but tested in this research are incorporated in these tables.

A set of questionnaires were designed to collect experts' opinion on the influential factors of facility LCC, and thus to supplement the results derived from the literature review and to ensure the comprehensiveness of the model attribute pool. The participate experts were grouped into three categories – experts of initial costs, utility costs, and O&M costs – and each expert was asked to fill out a corresponding questionnaire. These questionnaires were developed based on the literature review. The potential influential factors (descriptive attributes) were listed in the questionnaire and the experts were asked to rate on their influence on facility initial design and construction costs, utility costs, or O&M costs. In addition, the experts thought there were influential factors not listed in the questionnaire, they were also asked to specify them. The questionnaires are shown in Appendix A.

The information of participants is shown in Table 3.1. The influential factors not mentioned in the literature but identified in the questionnaire survey are also incorporated in Table 3.2 to 3.4.

Table 3. 1 The information of survey participants

Expert in	Title	Year of experience
Initial design and construction cost	Cost estimator	5
	Cost estimator	3
	Cost estimator	2
	Cost estimator	11
	Registered architect	24
	Vice president	25
	Senior vice president	31
	Assistant vice president	34
Energy consumption	Postdoctoral researcher	5
	Ph.D. candidate	4
	Ph.D. candidate	4
	Ph.D. student	3
	Ph.D. student	3
	Ph.D. student	2
O&M cost	Communications manager	16
	Assistant vice president	26
	Associate director	14
	Project coordinator	12
	Associate director	15

Table 3. 2 Independent variables affecting initial design and construction cost to be incorporated in the LCC prediction model

Group	Attribute Name	Description	Reference
General	Building function/type	e.g. commercial building, medical building, residential building, educational building.	[16,60]
	Project type	New construction, adaptive reuse, historical preservation, etc.	Identified in the survey
	LEED	The green building rating system Leadership in Energy and Environmental Design (LEED). LEED certifications involve certified, silver, gold, and platinum.	Proposed in this study
Structural	Structure type	e.g. concrete, steel, masonry, and timber structure.	[14,16,21,58,59,64,81,127]
	Type of floor structure	e.g. Cast-in-place (CIP) concrete and precast concrete.	[65]
Building Geometry	Building floor area (BFA)/ built-up area	The total floor area inside the building envelope, including the external walls.	[16,20,21,23,58,59,64-66,81,128]
	Floor area	Including Gross floor area (GFA), gross internal area (GIA), usable floor area (UFA), etc.	[13,16,17,23,60,62,64-69,80,81,127,129]

Number of floors	Including floors aboveground and underground.	[13,14,16,17,20,23,58,59,64-70,80,81,127,129]
Floor height	The average floor to ceiling height.	[58,66,127]
Total height	The total building height.	[17,21,62]
External wall area	External wall area is the difference between the external and internal gross areas.	[17]
Internal perimeter length	The perimeter of a building measured on the internal face of the enclosing structural walls.	[70]
Footprint area	The gross area of the ground floor.	[65]
Gross building volume (GRV)	The total volume of all interior spaces in a building over the gross floor area.	[66]
The location of the core of the building	The location of the vertical circulation system including stairs, elevators, and the service ducts. It can be at the sides or in the middle. “A central location requiring less cost than a side location which necessitates extra curtain walls to counteract torsion effects” [65].	[65]
Fully enclosed covered area	The sum of all fully enclosed and covered building areas at all floor levels, such as basements, garages, floored roof spaces, and attics.	[70]

	Unenclosed covered area	The sum of all such areas at all building floor levels, including roofed balconies, open verandahs, porches, and porticos, etc.	[70]
Space Distribution	Number of rooms	e.g. the number of apartment units, classrooms, and households.	[13,14,20,23,68,69,80,127,129]
	Percentage of usage	e.g. 30% classroom, 50% laboratory, and 20% office.	[11]
Roof	Roof type	e.g. gable roof, flat roof, and hip roof.	[13,68,69,80,127]
Exterior Enclosure	Façade type	Different systems and materials used in the exterior of a building.	[14,16,128]
	The surface area of exterior wall	The value of external surface area subtracting the external window area.	[67]
	Window type	e.g. fixed windows, casement windows, and sliding windows.	[16]
	The proportion of opening on external walls	Area of external doors and windows divided by external wall area ×100%.	[17,128]
Interior	Finishing	Finishing type/grade (e.g. luxurious, medium, and simple).	[68,69,80]
Occupancy	Number of people	Designed/predicted number of occupants.	[81]
Foundation	Foundation type	e.g. foundation system (pier, wall, slab) and basement system (crawlspace, full, none, walkout).	[14,60,64,65,67-69,80]

	Foundation area	The total area of the foundation.	[21]
Civil and landscape	Topography & soil condition	e.g. plan irregularity, soil condition (hard, medium, or soft).	[14,59,66,67,129]
	Parking area	The total area of the parking lot.	[13,20]
	Landscape area	The total area of the landscape.	[16]
Construction Management	Construction duration	The period of time between the date of the construction contract start on site and the date of practical completion.	[17,21,23,60,68,69]
	The general contractor	Self-explanatory.	Proposed in this study
	Construction site area	The total area of the construction site.	[20,129]
	Delivery method	e.g. CM at risk, design-build, and design-bid-build.	[23]
	Construction site access	e.g. free, semi-restricted, and restricted access.	[66]
Mechanical, Electrical,	Mechanical installations	The type of mechanical systems and the vendors/subcontractors.	[60]

and Plumbing (MEP)	Number of elevators	Self-explanatory.	[127]
Other	Security	The security requirements	Identified in the survey
	Information technology systems	Communication/audio-visual/special Systems	Identified in the survey

Table 3. 3 Independent variables affecting utility consumption to be incorporated in the LCC prediction model

Group	Attribute Name	Description	Reference
General	Building age	Self-explanatory.	[11,72]
	Building function/type	e.g. commercial building, medical building, residential building, educational building.	[72,83]
	LEED	The green building rating system Leadership in Energy and Environmental Design (LEED). LEED certifications involve certified, silver, gold, and platinum.	Proposed in this study
Building Geometry	Building floor area (BFA)/ built-up area	The total floor area inside the building envelope, including the external walls.	[11,72,83]

	Number of floors	Including floors aboveground and underground.	[11,83]
	Floor height	The average floor to ceiling height.	[11]
	Conditioned floor area (CFA)	The total floor area of enclosed conditioned space on all floors of a building.	[72]
	Gross building volume (GRV)	The total volume of all interior spaces in a building over the gross floor area.	[72]
Space Distribution	Number of rooms	e.g. the number of apartment units, classrooms, and households.	[72]
	Percentage of usage	e.g. 30% classroom, 50% laboratory, and 20% office.	Proposed in this study
Occupancy	Number of computers and televisions	e.g. desktop PC, laptops, TVs	[11]
	Number of printers/ photocopiers	Self-explanatory.	[11]
	Number of kitchen electrical products	e.g. conventional ovens and microwave ovens.	[72]
	Occupants' average time spent in the building	e.g. none, medium, and long	[72]

	Electric vehicle number	Self-explanatory.	[72]
	Occupancy percentage	The proportion of rooms occupied to the number of rooms available for the selected date or period.	[11]
	Reduce energy cost willingness	Yes/No	[72]
	Number of regular occupants	e.g. the number of employees or students.	[11]
	Total hours open per week	The total hours of operation each week.	[11]
Heating/cooling	HVAC system used	The HVAC system types and its corresponding control strategies	Identified in the survey
	Heating percentage	The percentage of the total floor space within the facility that is served by mechanical heating equipment.	[11]
	Cooling percentage	The percentage of the total floor space within the facility that is served by mechanical cooling equipment.	[11]
	Programmable thermostat	Have programmable thermostat? Yes/No.	[72]
	Apparent temperature	The temperature equivalent perceived by humans, caused by the combined effects of air temperature, relative humidity, and wind speed.	[72]

Electrical	Photovoltaic (PV) system	Have PV system? Yes/No.	[72]
	Percent lit when open	The percentage of lit area to total building area when open.	[11]
	Percent lit off hours	The percentage of lit area to total building area when close.	[11]
Weather	Heating degree day (HDD)	The number of degrees that a day's average temperature is below the degree which buildings need to be heated.	[11,83]
	Cooling degree day (CDD)	The number of degrees that a day's average temperature is above the degree which buildings need to be cooled.	[11,83]
	Temperature	The temperature at the building location throughout the year	[72]
	Dew point	The temperature to which air must be cooled to become saturated with water vapor.	[72]
	Humidity	The amount of water vapor present in air.	[72]
	Daily average sky cover (Cloud cover)	The daily average fraction of the sky obscured by opaque clouds when observed from the building.	[130]
Firefighting System	The number of sprinkler heads.	Self-explanatory.	[72]

Exterior Enclosure	Total window area footage	The total area of exterior windows.	[72]

Table 3. 4 Independent variables affecting O&M costs to be incorporated in the LCC prediction

	Attribute Name	Description	Reference
General	Building age	Self-explanatory.	[15,89-91,131-135]
	Building function/type	e.g. commercial building, medical building, residential building, educational building.	[135]
	LEED	The green building rating system Leadership in Energy and Environmental Design (LEED). LEED certifications involve certified, silver, gold, and platinum.	Proposed in this study
Structural	Structure type	e.g. concrete, steel, masonry, and timber structure.	[91]
Space distribution	Number of rooms	e.g. the number of apartment units, classrooms, and households.	[89]
	Percentage of usage	e.g. 30% classroom, 50% laboratory, and 20% office.	[15,90,91,134]
Occupancy	Number of people	Designed/predicted number of occupants.	[89]

	Number of shifts	Shifts of operation and maintenance team.	[89]
Financial	Average sale price	The average price at which the building is sold in the market.	[90,134]
	Budget constraint	The budget constraint of operation and maintenance.	[132]
Building geometry	Building floor area (BFA)/ built-up area	The total floor area inside the building envelope, including the external walls.	[90,91,134]
	Floor area	Including Gross floor area (GFA), gross internal area (GIA), usable floor area (UFA), etc.	[24,90,131,134]
	Number of floors	Including floors aboveground and underground.	[15,24,89-91,134]
	Building geometry in general	Footprint area, building volume, etc.	[135]
	Total height	The total building height.	[24]
	Floor height	The average floor to ceiling height.	[24]
Performance	Building performance index (BPI)	BPI is a high-level indicator of building performance. It sometimes refers to energy use intensity and sometimes used to combine several factors (energy, environmental, economic, etc.) into one term.	[131]

MEP	Number of elevators	Self-explanatory.	[15,91]
	Technical equipment	Number of certain equipment that needs special maintenance.	[135]

3.2 The Data Sources

Many building managers and operators are using building management and control systems in their daily work. These systems, such as the Building Automation System (BAS), the Computerized Maintenance Management Information System (CMMS), and the Building Energy Management Systems (BEMS), are constantly collecting or generating facility and human activity-related data, a portion of which can serve as the raw data for LCCA. This section briefly introduces the major building systems and discusses the LCCA-related data they can provide. In addition, the BIM-based Central Facility Repository (CFR) is also introduced.

3.2.1 Building systems

Each building system is designed to provide specific functionalities for facilities management and control but, with the advancement of Information and Communication Technology (ICT), the vendors of modern building systems are expanding their products' functionalities, thus there may be some overlaps among them. For example, some BAS may have certain energy management functions [136]. Regardless of the existing and potential overlaps, the mainstream building systems are summarized as follows:

Building Automation System (BAS). The BAS is also known as the Building Management System (BMS). It is a computer-based system that monitors and controls the building's mechanical and electrical equipment such as heating, ventilation, air-conditioning (HVAC) systems, fire systems, and security systems [137,138]. A BAS uses sensors to collect data pertaining to the building conditions and uses actuators to conduct physical control. A large amount of records pertaining to temperature, power, flow rate and pressure, control signals, states of equipment, etc., are collected by the BAS stored in its database [139].

Building Energy Management System (BEMS). A BEMS is a comprehensive approach to monitor and control the building's energy needs by collecting the building's energy-related data, analyzing the performance status, and controlling corresponding equipment [140]. Similar to the BAS, the BEMS are generally applied to the control of active systems such as the HVAC system and lighting system. The main difference between a BEMS and other building systems is the characteristic of energy-related data collection, processing, and centralized and/or automated building control [141]. The role of the BEMS is known and significant because BEMS can contribute to the continuous energy management and therefore to achieve energy and cost savings and the comfort of the building occupants [142]. When the goal of energy saving is expanded to reducing environmental impacts, the name Environmental Management System (EMS) is used instead of BEMS.

Computerized Maintenance Management Information System (CMMS). The CMMS, also known as the Computerized Maintenance Management Information System (CMMIS), is utilized by facilities maintenance organizations to record, manage, and communicate their daily operations [143]. The core function of a CMMS is "to manage information related to maintenance, including but not limited to work orders, asset histories, parts inventories, maintenance personnel management and the calculation of maintenance metrics" [144]. The CMMS is an essential tool for modern facility operation and maintenance work because it can 1) provide building component information for maintenance and repair work, 2) generate reports for managing resources, 3) prepare facilities key performance indicators (KPIs) metrics for evaluating the effectiveness of the current operations, and thus 4) support organizational decision makings [143].

Integrated workplace management system (IWMS). An IWMS is a software platform designed to optimize an organization's deployment of workplace resources, including real estate

management, capital project management, asset management, and space management [145]. It sometimes includes the functionality of a CMMS and/or an enterprise resource planning system (ERP) [144].

3.2.2 The potential data sources for each LCC component

Data availability is the biggest challenge for facility LCCA [22,87]. The building systems introduced in the previous section are constantly collecting or generating facility and human activity-related data, a portion of which can serve as the raw data for LCCA. However, these data are rarely interpreted and utilized for LCC-related analysis. This research solves the data availability issue of facility LCCA by creating a framework for utilizing the data in building systems directly. The potential sources of the data that can be used to derive each LCC component are listed in Table 3.5.

Table 3. 5 The LCC components and their potential data sources

LCC Component	Potential Data Source
Initial design and construction costs	IWMS – the capital planning and investment control module.
	The construction cost estimation report that records the detailed construction costs.
	The design contract that records the design costs.
Utility costs (utility consumptions)	The BAS / BMS
	The BEMS
O&M costs	CMMS

Replacement costs	The same source as the initial costs
	CMMS

3.2.3 The BIM-based Central Facility Repository (CFR)

The concept of Central Facility Repository (CFR) proposed by the General Services Administration (GSA) describes the comprehensive information system built around a central facility database, which serves as the data foundation for FM-related software applications [109]. BIM can serve as the platform of a CFR by integrating 3D object parametric data, facility management data, building drawings and specifications, real-time sensor data, etc. [146,147]. The BIM-based CFR offers an efficient way to extract building data, and thus to obtain the descriptive attributes (listed in Table 3.2 to 3.4) for the LCCA machine learning models.

3.3 Building Data Acquisition and Integration

Currently, the inter-system data interoperability among the building systems is limited and the data formats vary based on different vendors. The data housed in these systems are not connected, available to analysts and developers in a consumable way. Therefore, these data are rarely utilized for LCC-related analysis or any other analysis in an integrated fashion. In this research, the universal set of all the data needed to establish the LCCA machine learning models is termed as “the LCCA data package”. This section proposes a building data acquisition and integration framework to establish the LCCA data package.

Figure 3.1 shows a high-level data acquisition process for LCCA. The design and construction documentation refer to the construction drawings, estimation reports, scheduling, manuals, and specifications. The BIM refers to the as-built building information model, which is the "digital twin" of a building [26]. The required building data can be automatically extracted from BIM if it is properly developed and include relevant information [148,149].

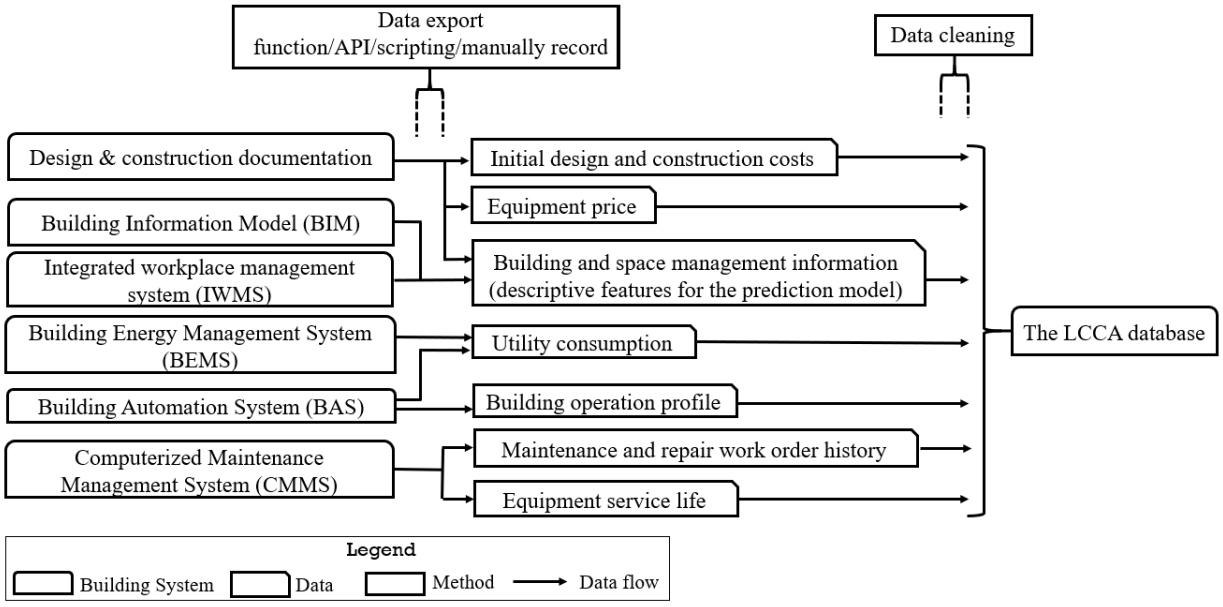
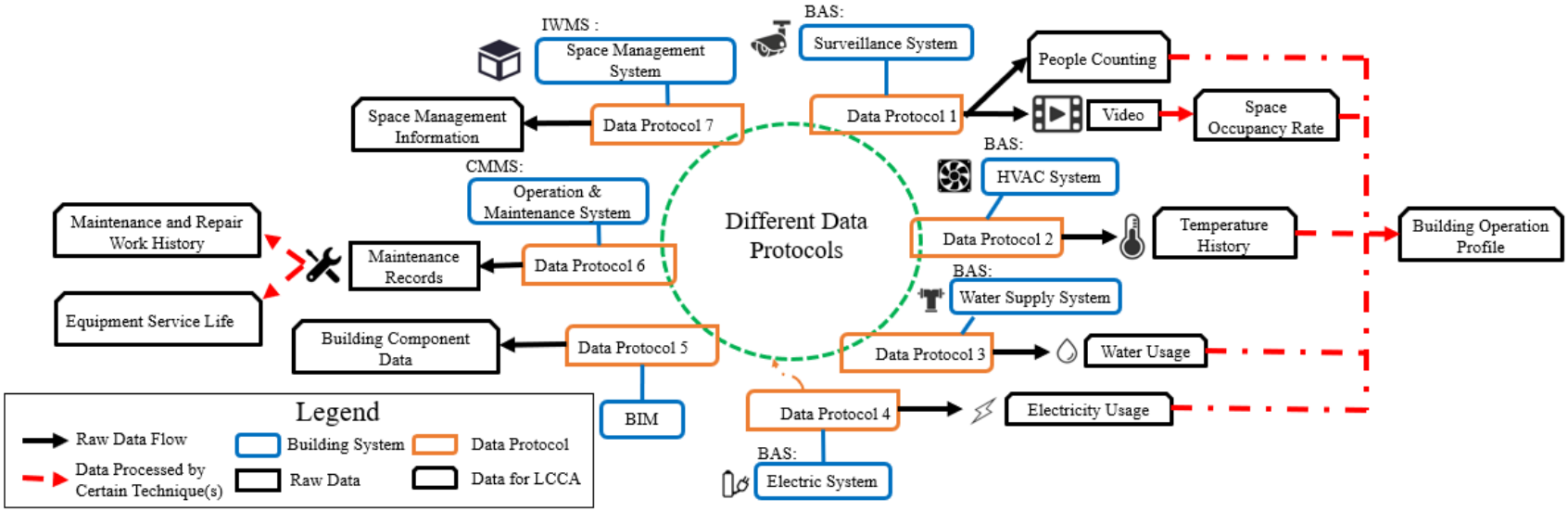


Figure 3. 1 The overall LCCA data acquisition process

3.3.1 The framework to establish the LCCA data package

The evolving building systems already contain the data needed for machine learning-based LCCA. By extracting relevant data from each building system, storing them in a federated database which consists of multiple connected individual databases, we can establish the LCCA data

package. The database connections can be established by using data exchange schemas, such as customized Extensible Markup Language (XML) schemas, which are derived from the data mapping between different data standards.


IWMS :
Space Management System
Space Management Information
Data Protocol 7
CMMS:
Operation & Maintenance System
Maintenance and Repair Work History
Maintenance Records
Data Protocol 6
Equipment Service Life
Building Component Data
Data Protocol 5
BIM
Different Data Protocols
BAS:
Surveillance System
Data Protocol 1
People Counting
Video
Space Occupancy Rate
BAS:
HVAC System
Data Protocol 2
Temperature History
Building Operation Profile
BAS:
Water Supply System
Data Protocol 3
Water Usage
Data Protocol 4
Electricity Usage
BAS:
Electric System
Legend
Raw Data Flow
Data Processed by Certain Technique(s)
Building System
Raw Data
Data Protocol
Data for LCCA

(a)

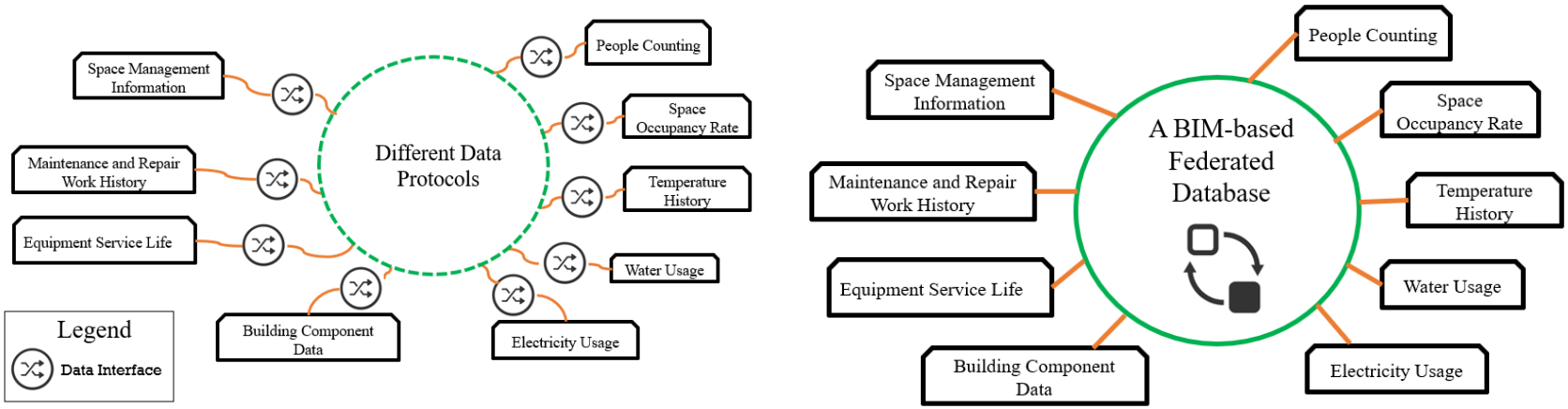


Figure 3. 2 Establishing the LCCA data package by linking the databases of building systems

Figure 3.2 illustrates a conceptual framework to establish the LCCA data package. Separate building systems with different data protocols generate various types of raw data, such as the electric system generates energy usage data and the water supply system generates water usage data. Some of the raw data need to be further processed by certain techniques, such as data mining and machine learning, to produce the data ready for LCCA (Figure 3.2a). The lack of means to integrate these building data is hindering the machine learning-based LCCA applications. A federated database network that can provide integrated and comprehensive building data for LCCA – the LCCA data package – by connecting multiple databases, which are based on various standards and protocols, is the basis of the efficient implementation of machine learning-based LCCA.

If we examine the building systems from a perspective of IoT, they are already embedded with the IoT network which contains sophisticated sensing and actuation devices. BIM models offer a clear potential as the "digital twin" of the built environment – one that can provide significantly enhanced spatial context for the building IoT network. A strategy for connecting the data protocols used by the building systems with the BIM data schemas can provide a critical layer of spatial semantics to these IoT systems, such as device geo-positioning and metadata tagging, and can harmonize these data sources with various data protocols.

Extensive work is needed to enable the data connections between different building systems and to establish the proposed federated database network and hence to acquire the LCCA data package. One of the prerequisites is a thorough investigation of the data standards and protocols of the IoT devices – sensors, cameras, actuators, etc. – and the BIM data schemas. The most commonly-used data protocols of building automation and control are summarized in Table 3.6.

Table 3. 6 Data protocols of building automation and control

Data standards	Data standards	Data standards
1-Wire	Dynet	VSCP
BACnet	EnOcean	xAP
C-Bus	KNX	X10
CC-Link	LonTalk	Z-Wave
DALI	Modbus	ZigBee
DSI	oBIX	INSTEON

Industry Foundation Classes (IFC) specification is the leading neutral BIM data schema to describe, exchange and share building information [150]. Most of the BIM software applications support IFC. Currently, 653 entities (geometry, properties, and relationships of building components) can be defined in IFC and the capabilities of IFC as a data standard keep expanding [151]. gbXML [152] is another commonly used building data schema. CityGML is a city-level open data standard and exchange format to store digital 3D models of cities and landscapes [153].

3.3.2 The method for establishing the linkage among different databases

The connections between the IoT data protocols with the BIM data schema can be established by identifying the overlaps between them and creating a federated data framework that enables the data collection, query, and exchange. Figure 3.3 shows an example in which the

overlaps of the data protocol in each level – device level, building level (BIM), and city/community level – are identified. Then, as Figure 3.4 shows, the overlaps can be used to establish the linkage among the databases of devices, buildings, and the city/community, thereby a federated database network can be established to provide real-time LCCA data package.

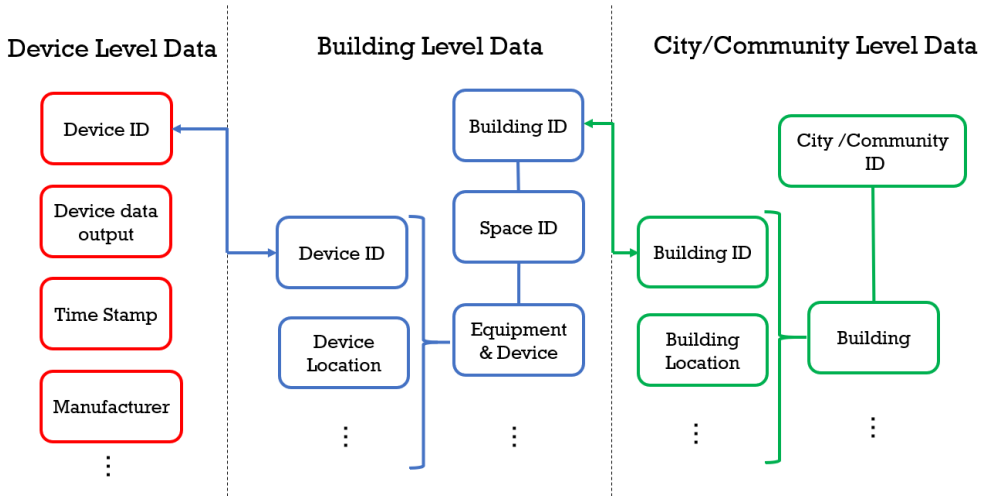


Figure 3. 3 Identifying the overlaps between data protocols

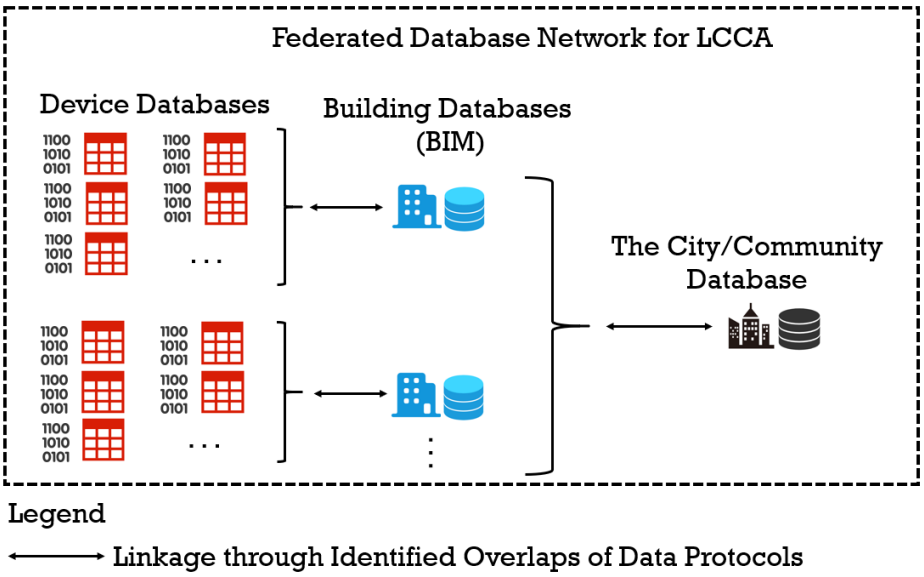


Figure 3. 4 Federated database network for LCCA

The challenge of establishing the linkage among the databases lies in how to align the "common" data – such as device ID and building ID – in each database. For example, the device ID is "abcd123" in the device database, but it may be "123-abcd" in the building database. XML schemas can be created to address this issue. They enable automatically editing and concatenating the values of the key data fields and thus to align the common data.

3.3.3 An example experiment: connecting BACnet, IFC, and CityGML based databases

An example experiment is designed to demonstrate the proposed method for establishing the federated data framework and prove its feasibility. In this experiment, BACnet, IFC, and CityGML are adopted as the data protocol of device level, building level, and city level, respectively.

BACnet (A Data Communication Protocol for Building Automation and Control Networks) is the dominant protocol in the Building Automation Industry [154]. The BAS used in the

experiment – Johnson Controls' Metasys system – is based on BACnet and its devices are BACnet compatible [155].

Figure 3.5 shows the experiment framework. The data generated by the BAS's sensor network deployed in multiple campus buildings are collected and stored in a MySQL database. The database of Metasys is not used directly due to the lack of privilege and changes need to be made in the database without disturbing the BAS' normal operations. An open-source tool named BCVTB (Building Controls Virtual Test Bed) [156] is used to read the data generated by BACnet devices and write them to the database. This approach can extract any type of building data from BACnet supported device networks. In this experiment, the electricity consumption data are used as the device level data to connect with the IFC-based building database.

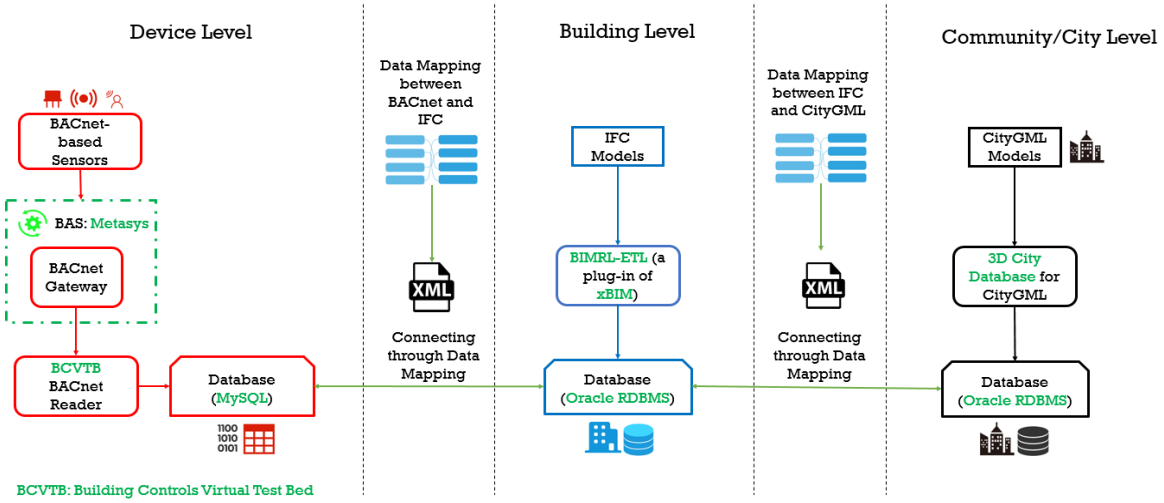


Figure 3. 5 The experiment framework: connecting BACnet, IFC, and CityGML based databases

In the building level, to provide an efficient query-ability into BIM data, which is traditionally difficult and slow, two IFC models are transformed into a simplified RDBMS (Relational Database Management System) data format and stored in an Oracle Database Server (Oracle RDBMS). The tool the author used in this step is called BIMRL-ETL (BIM Rule Language - Extract, Transform and Load) [157], which is a plug-in of xBIM [158]. The xBIM project aims to provide developers the codebase for innovative BIM applications.

In the community/city level, a CityGML city model is also transformed and stored in an Oracle Database Server. A tool named 3D City Database (3DCityDB) [159] is used to automate this process. The 3DCityDB is an open-source package consisting of a database schema and a set of software tools to import, manage, analyze, visualize, and export CityGML city models. "The database schema results from a mapping of the object-oriented data model of CityGML 2.0 to the relational structure of a spatially-enhanced relational database management system" (SRDBMS) [160].

To federate the data in each level's database by using the common data fields, the author have identified the overlaps (not exhaustively) between BACnet XML [161] and ifcXML (IFC in the XML form) [162], that between BACnet XML and gbXML,and that between ifcXML and CityGML XML [163]. Please see Appendix B for the data mapping detail. The common data fields used for database federation in this experiment are shown in Table 3.7 and 3.8.

Table 3. 7 The common data fields between BACnet XML and IFC XML

BACnet XML (DR-034A-28)	IFC XML (IFC4 ADD2)
▼Object	▼ifcBuilding
propertyIdentifier	id
name	Name
type	ObjectType
description	Description
contextTag	Tag

Table 3. 8 The common data fields between IFC XML and CityGML

IFC XML (IFC4 ADD2)	CityGML 2.0 XML
▼ifcBuilding	▼Building
id	gmlid
Name	gmlname
Description	gmldescription
ElevationOfRefHeight	measuredHeight
ifcBuildingAddresss	address

Based on the identified overlaps of these data schemas, XML files are created in MapForce [164] to automatically edit and concatenate the values of the identified data fields, thus to align the common data in each database.

This experiment demonstrates the process of establishing the federated databased network, which is: 1) studying different building IoT data schemas, 2) identifying common data fields between these schemas and BIM data schemas, 3) creating the tools (may be as simple as an XML schema) to establish the linkage between databases, and thus 4) establishing the federated data network. The proposed framework is scalable, which means if one type of building data can be extracted and stored in an SQL database, and this database can be linked to the building level database and then the community/city level database using the proposed methods, other types of data can also be processed by similar methods according to the framework, although the specific tools may be different in each case.

CHAPTER 4 DERIVING LCC COMPONENTS THROUGH MACHINE LEARNING

The first and most challenging task of an LCCA for buildings or building systems is to determine the economic effects of alternatives and to quantify these effects and express them in monetary amounts [6]. 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. This chapter first proposes the framework for machine learning-based facility LCC component prediction, and then summarizes the most commonly used machine learning methods for facility cost prediction. It also discusses the attribute selection process and the applicability of these methods in the prediction of each LCC component – initial cost, utility cost, and O&M cost. The validation methods and performance measures are also discussed.

4.1 The Process of Machine Learning-based Facility LCC Component Derivation

The overall process of deriving LCC components is shown in Figure 4.1. 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 Chapter 3). The data indexed in time order – utility consumption and O&M costs – are analyzed by time series methods, and projections are made when necessary, such as there are missing values because of sensors were not deployed before a certain time. The public statistics, such as the historical inflation rate, utility price, and labor rate, are incorporated into the analysis to calculate the monetary costs and to convert the costs to their present values, which are the LCC components (target attributes) for the LCCA machine learning model development.

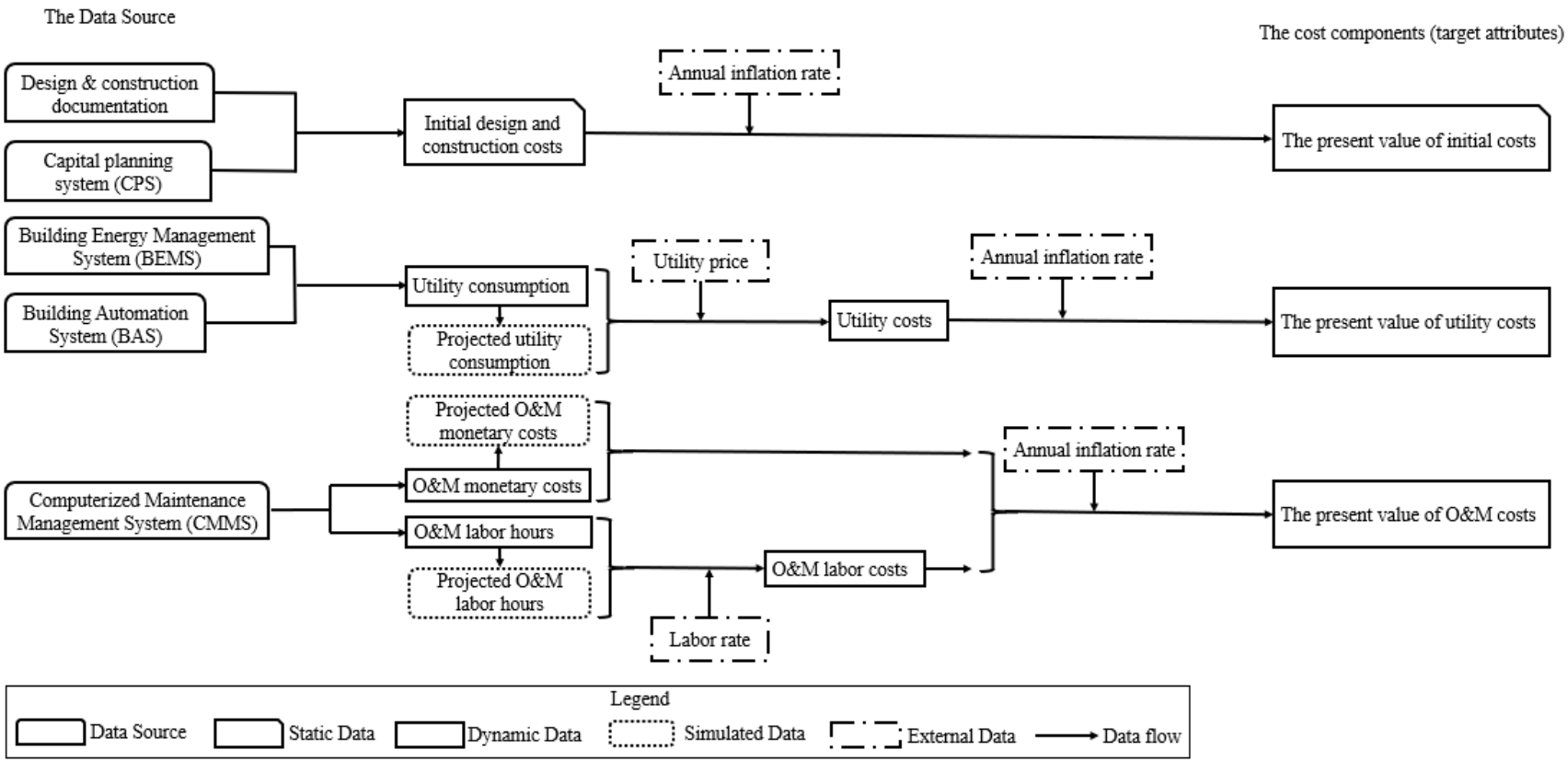


Figure 4. 1 LCC component derivation

4.2 Machine Learning Methods for Facility Cost Prediction

The literature review indicates machine learning methods, including linear regression, K-Nearest Neighbors (KNN) regression, Support Vector Machines (SVM) regression, regression trees, and Artificial Neural Networks (ANN), have been implemented for building cost prediction [12]. This section introduces these machine learning methods, which were implemented in the experiments (Chapter 6).

4.2.1 Linear regression and gradient descent

Regression analysis is a technique for modeling the relationship between variables [165]. 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. The model involves only one independent variable is called a Simple Linear Regression (SLR) model; the one 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 are modeled 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 as 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 [165].

Gradient descent is a commonly employed iterative optimization algorithm to find the values of parameters (coefficients) of a function that minimizes a cost function [166]. 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, whose goal is to find the set of parameters for a model that minimizes the total error across the prediction made by the model [8].

4.2.2 Support Vector Machines (SVM) 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 [167]. It grounded in the framework of statistical learning theory – characterizing properties of learning machines which enable them to generalize well to unseen data [167]. Support Vector Machines (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" [168]. SVMs map input vectors into a high dimensional feature space, where a maximal margin hyperplane is constructed [169]. SVMs can be applied to regression problems by introducing an alternative loss function that is modified to include a distance measure [167,170,171].

4.2.3 K-Nearest Neighbors (KNN) regression

The K-Nearest Neighbors (KNN) algorithm is a non-parametric method that can be used for regression analysis [172]. It is a type of instance-based learning, or lazy learning, and considered as one of the simplest o machine learning algorithms [166]. 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 neighbors [172]. 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 [8]. In addition, the KNN regression model uses the full

set of descriptive features when making a prediction, which makes it particularly sensitive to the occurrence of missing descriptive feature values [8]. The KNN is a similarity-based approach to machine learning, which comes from the idea that making predictions based on what has worked well in the past [8].

4.2.4 Regression Tree

A decision tree is a hierarchical tree-like model composed of a root node, interior nodes, and leaf nodes [166]. The decision tree machine learning model uses a decision tree to start 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) [8]. Decision trees with 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 [173] as a method of determining the shortest sequence of descriptive feature tests required to make a prediction [8].

4.2.5 Time Series Forecasting and Backcasting

A time series is a series of data points indexed in time order. One of the fundamental functions of the time series analysis is forecasting. Time series forecasting is the use of a model to predict future as yet unobserved values based on historical data [174]. The deterministic trend model is one of the most commonly used time series models for handling trends and seasonality in economic and business data [175]. It can be established by the method of ordinary least squares, thus can be easily interpreted [175]. An autoregressive integrated moving average (ARIMA) model is another commonly used model for predicting future time series data [174].

While time series forecasting involves predicting the future based on the analysis of existing data, time series backcasting works in the opposite direction – using the existing data to estimate the unknown historical data, i.e. forecast in reverse time [176]. In this research, the future utility consumptions and O&M costs are predicted through time series forecasting approaches, while the unknown historical consumptions and costs data are estimated through time series backcasting approaches.

4.2.6 Artificial Neural Network (ANN) and Multilayer Perceptron (MLP)

Artificial Neural Networks (ANN) are computing systems inspired by the biological neural networks that constitute animal brains [177]. The ANN itself is not an algorithm but rather a framework to federate different machine learning algorithms for complex analysis [177]. The Multilayer Perceptron (MLP) is an ANN structure and is a nonparametric estimator that can be used for regression [166]. The 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 [178-180]. One strength of MLP is the capability of distinguishing data that is not linearly separable [181]. MLP is also a method of error-based machine learning for predictive analytics [8].

4.3 Attribute Selection

In machine learning, attribute selection, also known as feature selection or variable selection, is the process of selecting a subset of the relevant attribute (features, variables) for use in model construction. Attribute selection is an essential process of machine learning model development because it simplifies the model for easier interpretation, shortens the model training time, and enhances generalization by reducing overfitting [182]. In this research, the preliminary attribute

selection is conducted through a literature review and questionnaire survey (Chapter 3). This section introduces three general classes of attribute selection algorithms: filter methods, wrapper methods, and embedded methods.

4.3.1 Filter Methods

Filter methods use variable ranking techniques and act as a pre-processing step to rank the features wherein the highly ranked features are selected and applied to a predictor [183]. The predictor is one or more variables that are used to determine the target attribute (feature, or variable). In filter methods, the general characteristics of the training data, such as distances between classes or statistical dependencies, are used to select features with the independence of any predictor [184]. The commonly used filter methods involve the Relief algorithm [185], correlation based feature selection (CFS) [186], fast correlated-based filter (FCBF) method [187], and the INTERACT algorithm [188].

4.3.2 Wrapper Methods

Wrapper methods search through subsets of variables and use the predictor performance as the objective function to evaluate the variable subset [183]. Wrapper methods are considered a superior alternative in supervised learning problems, but they generally have a high computational cost than filters methods because the search process requires a large number of executions [189]. The algorithms that can be used in wrapper methods can be broadly classified into Sequential Selection Algorithms and Heuristic Search Algorithms [183]. Sequential selection algorithms involve Sequential Feature Selection (SFS) algorithm [190,191], Sequential Backward Selection (SBS) algorithm [190], Sequential Floating Forward Selection (SFFS) algorithm [190,191], and Adaptive Sequential Forward Floating Selection (ASFFS) algorithm [192,193]. Heuristic Search

Algorithms involve the Genetic Algorithm (GA) [194], the CHCGA (a modified GA) [195,196], and the binary PSO algorithm [197].

4.3.3 Embedded Methods

Embedded methods aim to reduce the computation time of reclassifying the subsets – a process done in wrapper methods – mainly by incorporating the feature selection as part of the training process [183]. In contrast to the filter methods and the wrapper methods, in embedded methods, the learning part and the feature selection part cannot be separated [198]. Embedded methods may be more efficient than the wrapper methods in several respects: the available data are fully utilized because they do not need to be split into the training set and the validation set; moreover, the solution can be reached faster because embedded methods do not need to retrain a predictor from scratch for every variable subset investigated [199].

4.4 LCC Components Derivation

To build the dataset for training machine learning models, each of the LCC component – the initial cost, the utility cost, and the O&M cost – needs to be derived from the raw data (Figure 4.1). This section presents the general derivative formulas for the LCC components.

4.4.1 The Initial Cost

The present value of the initial cost is calculated by the following equation:

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

Where:

PV_{IC} is the present value of initial cost.

IC is the amount of initial cost.

t is the building age.

r_i is the annual inflation rate of *i* years ago.

4.4.2 The Utility Cost

The utility consumptions of a facility are time-series data. To convert these data into roll-up present values, the studied time frame must be specified first, such as 20 years, 50 years, or 70 years. The present value of the utility cost – electricity costs, water (and sewer) costs, and gas costs – can be calculated by the following equation:

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

Where:

PV_U is the present value of utility cost, which can be electricity cost, water (and sewer) cost, gas cost, etc.

UC_j is the annual utility consumption of *j* years ago.

UP_j is the utility price of j years ago.

n is the length of the study period in years.

r_i is the annual inflation rate of i years ago.

4.4.3 The Operation, Maintenance, and Repair Cost

The O&M cost typically involves monetary costs – such as material costs, equipment costs, and labor costs when hiring a vendor to do the O&M work – and the labor hours of the organization's employees spent on the O&M work. A specific time frame is also needed here since the O&M cost is also time-series data. The present value of the O&M cost can be calculated by the following equation:

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

Where:

PV_{OM} is the present value of O&M cost.

LH_j is the annual labor hours spent on O&M j years ago.

LP_j is the O&M labor rate j years ago.

OMC_j is the annual O&M monetary cost j years ago.

r_i is the annual inflation rate of i years ago.

4.5 Evaluation Methods

This section summarizes four commonly used model performance evaluation methods and discusses when each is most appropriate.

4.5.1 Hold-out Sampling

The hold-out sampling method splits the dataset into a training set to train the model and a test set to evaluate it. Implementing this method requires that the training set is no more representative for the test set than for the population as a whole [200]. Hold-out sampling is a simple form of sampling that is suitable when a large dataset is available to ensure that the model is accurate and fully evaluated [8]. In some cases, a third dataset, the validation set, is needed for hold-out sampling when data outside the training set and test set are needed to tune particular aspects of a model [8].

4.5.2 k-Fold Cross Validation

Cross-validation is an evaluation technique for assessing a machine learning model's ability to predict new data that were not used in training it, and to give an insight on how the model will generalize to an independent dataset [201]. The k-fold cross-validation method randomly partitions the original sample dataset into k equal sized subsamples and uses one of these subsamples as the test set and uses the remaining k − 1 subsamples as the training set. The cross-validation process

repeats k times, with each of the subsamples used exactly once as the test set [202]. The k results can then be averaged to produce a single analysis accuracy. The advantage of this method is that all instances of the dataset are used for both training and validation, and each instance is used for validation exactly once [8].

4.5.3 Leave-one-out Cross-validation

Leave-one-out cross-validation is an extreme form of k-fold cross-validation in which the number of folds is the same as the number of training instances [8]. In this method, each fold of the test set contains only one instance, and the training set contains all the remaindering data. Leave-one-out cross-validation is suitable when the amount of data is limited such that no big enough training sets can be constructed for k-fold cross-validation [8].

4.5.4 Bootstrapping

When the dataset is very small (for example, fewer than 300 instances), bootstrapping approaches are preferred over cross-validation approaches [8]. Similar to the k-fold cross-validation, the iteratively performs evaluation experiments using slightly different training and test sets each time to evaluate the performance of a model. For each iteration, a random selection of some instances is taken from the original dataset to generate a test set and the remaining ones are used as the training set, then a performance measure of the trained model is calculated for each iteration. The main difference between bootstrapping and cross-validation is that bootstrapping is resampling with replacement while k-fold cross-validation without, which means bootstrapping can have duplicate data while k-fold cross-validation cannot. The bootstrapping process is repeated for k iterations and the average of the individual performance measures gives the overall

performance of the model. The k here is typically greater than or equal to 200, which is much larger than that of the k-fold cross-validation method [8].

4.6 Performance Measures

Multiple performance measures can be used to evaluate the effectiveness of machine learning models. This section summarizes four commonly used measures of error for regression machine learning models.

4.6.1 Basic Measures of Error

The basis of most common performance measures for regression machine learning models is the sum of squared errors of prediction (SSE), which is the sum of the squares of the errors – the deviations of predicted from observed data [8]. SSE is a measure of the discrepancy between the actual observations and the machine learning predictions. A small SSE indicates a tight fit of the model to the data [166]. SSE is given by [8]:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Where:

n is the size of the dataset.

y_i is a set of n expected target values.

$f(x_i)$ is a set of n predictions for a set of test instances, $x_1, x_2 \ldots, x_n$.

Mean Squared Error (MSE) is the average of the squares of the errors, i.e. the average squared difference between the estimated values and what is estimated [8]. The MSE is a commonly used measure of the quality of an estimator; it is always non-negative, and values closer to zero are better [57]. MSE is given by [8]:

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{n}$$

Where the notations are the same as the equation above.

Root Mean Squared Error (RMSE) is the square root of MSE. It has the same units as the quantity being estimated.

Because including the squared term, the RMSE tends to overestimate error as it overemphasizes individual large errors [8]. An alternative measure that addresses this problem is the Mean Absolute Error (MAE), which is given by [8]:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - f(x_i)|}{n}$$

Where the terms in the equation have the same meaning as before. Same as SSE, MSE, and RMSE, the smaller values of MAE indicate better model performance.

4.6.2 Domain Independent Measures of Error

The basic measures of error introduced above can give intuitive measures of the performance of a machine learning model. However, these measures by themselves are not sufficient for an analyst to judge whether a model is making accurate predictions without deep knowledge of a domain [8]. For example, without understanding the domain of construction and some information

of the project (such as the project size), one cannot judge whether a construction cost prediction model that has an RMSE of $500,000 is actually making accurate predictions or not.

The coefficient of determination, R^2, is a domain independent measure of model performance that is frequently used for predictive machine learning models. It is the proportion of the variance in the target attribute that is predictable from the descriptive attribute(s) [166]. R^2 is calculated as [8]:

$$R^2 = 1 - \frac{SSE}{SST}$$

Where SSE is the sum of squared errors of prediction introduced above, and SST is the performance of an imaginary model that always predicts the average values from the test set – the Total Sum of Squares, which is given by [8]:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The values of R2 fall in the range of [0, 1) and larger values indicate better model performance. R2 is used as the model performance measure in the experiment demonstrated in Chapter 6.

CHAPTER 5 THE OVERALL PROCESS AND ONTOLOGY OF LCCA

This chapter presents a framework for developing machine learning models for facility LCCA. A domain ontology for machine learning-enabled LCCA (LCCA-Onto) is developed to encapsulate knowledge about LCC components and their roles in relation to sibling ontologies that conceptualize the LCCA process. This domain ontology is then tailored to be the cornerstone (the knowledge base) that will enable the automated LCCA data collection from building systems. The contents presented in this section can be used as an institutional guideline for facility LCC monitoring and prediction.

5.1 The Overall LCCA Framework

5.1.1 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 readings of meters are accurate and the records in each building system, no matter automatically saved or manually inputted, are correct, with outliers of the data identified and processed.

(2) The simulated data can reflect the actual costs – the missing data (because the sensors were not deployed or malfunctioning), such as utility consumptions and O&M costs, can be estimated by forecasting or backcasting based on the historical data.

(3) The inflation rate related to building costs is the same as the general inflation rate – the author assume the discount rate for building costs in the U.S. can be represented by the general inflation rate provided by the U.S. Bureau of Labor Statistics [203].

5.1.2 A framework for developing machine learning models for facility LCCA

The proposed framework for developing machine learning models for facility LCCA is shown in Figure 5.1. It consists of four major modules: 1) obtaining the descriptive attributes, 2) obtaining the target attributes, 3) training machine learning models, and 4) evaluating the models and selecting the most suitable one.

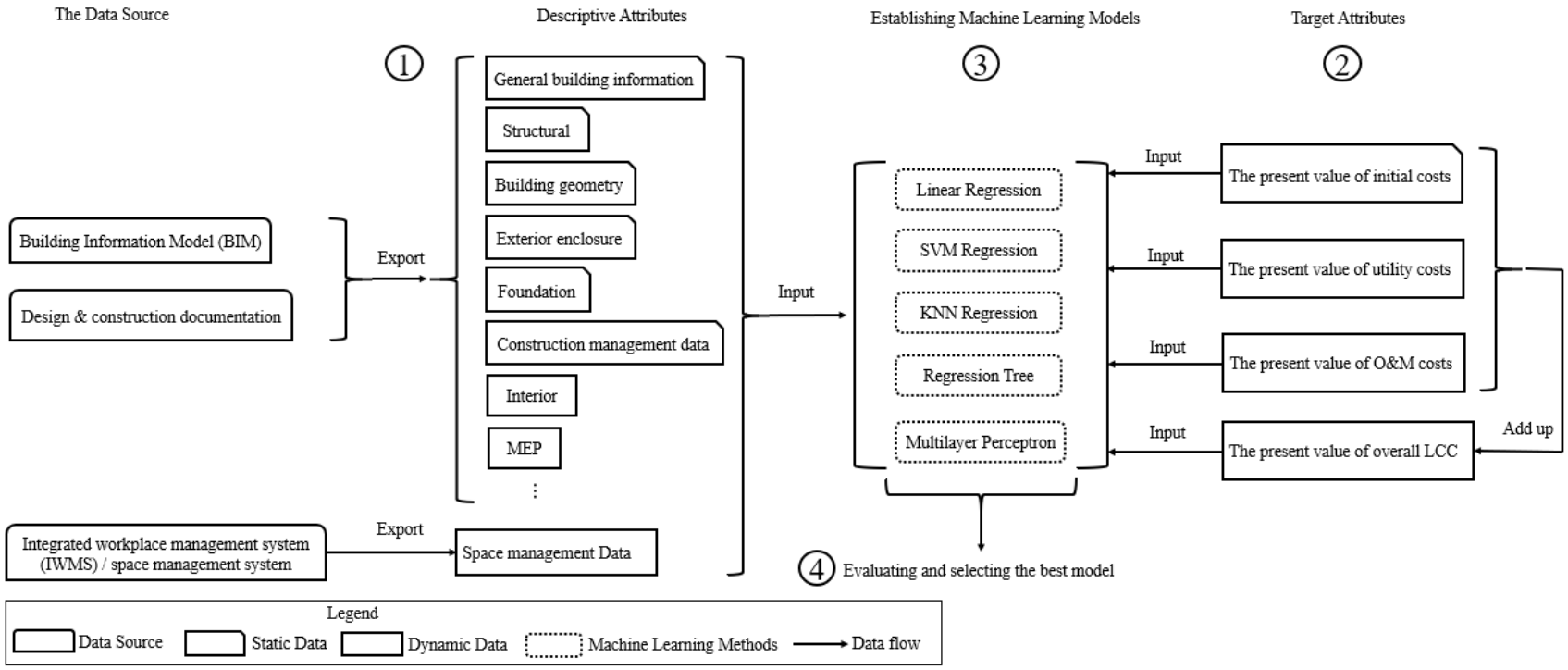


Figure 5. 1 The overall LCC machine learning model development

The potential descriptive attributes in the LCCA machine learning model are discussed in Chapter 3, listed in Table 3.2 to 3.4. The data related to those attributes can be provided by BIM [147,148]. However, for the organizations that do not have well-developed BIM (e.g. BIM with a level of development 400) for all facilities, the required data can be found in the design and construction documentation. For example, design drawings contain building geometry, structural, foundation, 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. After operation, the buildings' 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.

The derivation of target attributes – the present value of initial costs, utility costs, and O&M costs – is discussed in Chapter 4. The process of facility LCC component derivation is shown in Figure 4.1. In contrast to the descriptive attributes, which are relatively static in a certain time period (such as three months), the target attributes are dynamic and can vary with the real-time utility consumption and O&M costs. Therefore, a framework to acquire and integrate the dynamic facility data in an automated fashion, as described in Section 3.3, is desirable for overall facility LCCA machine learning model development process.

With the descriptive attributes and target attributes ready, the next step is to train the machine learning models based on these data. The machine learning methods listed here in the proposed framework are the ones that have been proved to be effective in building-

related costs prediction [12]. The method pool of training regression models for facility LCCA is expandable. With the development of machine learning techniques in predictive data analytics, more methods can be adopted and implemented within the framework.

Evaluating the models and selecting the outperformed machine learning algorithm for facility LCC prediction can be done by repeated random sampling and cross-validation, and then comparing their performance, as shown in Figure 5.2 [204]. First, the dataset is randomly split into the training set and test set with a proportion of, for example, 7:3. Multiple pairs of training set and test set are generated by the random sampling method (e.g. *m* pairs). For each pair (e.g. the pair *j*), its training set 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 (P_j1, P_j2, ..., P_jk). 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, …, P_j, …, P_m). These outcomes are then analyzed by the evaluation methods such as ANOVA or Kruskal-Wallis test to determine which algorithm outperforms others [204]. The most suitable machine learning method would possibly be different from case to case, depending on the length of studied time span, attributed used, and data size and quality.

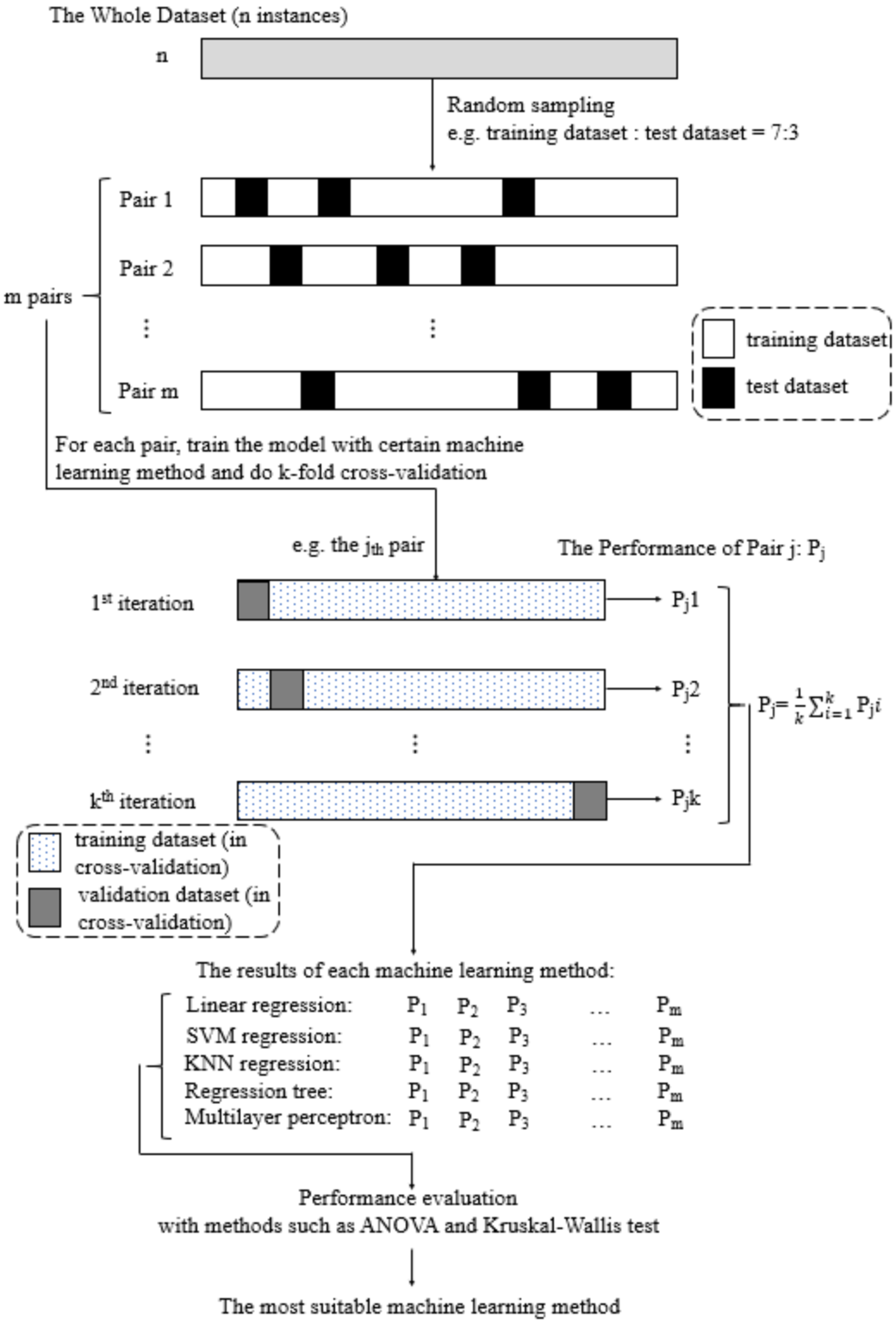


Figure 5. 2 Algorithm evaluation and selection process (inspired by [204]).

5.1.3 Usage of the developed models

The developed machine learning models can be used by the organization's facility management and capital planning department to have a preliminary estimate of a building's LCC, new or existing. Without any input of professional estimators nor the detailed building design, these models are expected to yield a reasonable prediction (e.g. with an error of 20%) based on limited input parameters, such as gross square footage, number of floors, structural type, and space allocation (such as 30% residential, 20% office, 20% general usage, etc.).

The organization's management can have a better understanding of each facility's operation profile based on the LCCA results. Representing facility design, construction, operation, and maintenance related activities in monetary terms makes it easier for the management to detect abnormal patterns of facilities management. For example, the management may find the predicted lifecycle energy cost of a new building with LEED Platinum certification is much higher than that of a similar-sized old building with the same function. This may indicate that the new building is not properly operated as the design intent. The machine learning-based LCCA tool provides the management with a means of monitoring to effectively detect the "hot spot" of facilities management and thus support the management to adjust strategies accordingly. Moreover, because the data used to train the machine learning models are directly acquired from the IoT network that embedded in the heterogeneous building systems, human intervention is minimized in the whole analysis process, which makes the analysis results more reliable and trustworthy.

The design and construction department of the organization can benefit from the knowledge developed from the LCCA results, and thus to achieve more informed decision makings. During the whole LCCA process, new knowledge can be developed based on the analysis of the historical cost and the projected future cost. For example, with the vendors of the major building systems inputted as descriptive attributes of the machine learning models, if these attributes have a significant influence on a certain LCC component (initial cost, utility cost, or O&M cost) or the overall LCC, the analysis can detect which vendor's products consumes more energy than the others' or require more labor and money to maintain. With this kind of knowledge, the design and construction team can have a better understanding of the trade-offs during the design and construction process. For example, based on the experience and projected future costs, this vendor's products are inexpensive and energy efficient but usually require more maintenance, while the other one's are more expensive and energy intensive but more robust and durable. The LCC prediction models developed based on the framework proposed in this research can provide the designers and builders quick estimates of different alternatives, given sufficient historical data.

5.2 An Ontology for IoT and BIM-enabled Facility LCCA

Conceptual analysis and knowledge representation often require ontological support, therefore, developing a domain ontology is one of the fundamental steps when developing a shared model of knowledge [205]. This section involves the development of a domain-level ontology for machine learning-enabled LCCA (LCCA-Onto). The LCCA-Onto provides a formal, specific conceptualization of the involved entities (participants, data, and methods) and their relationships within the facility LCCA domain.

5.2.1 The Ontology methodology

An ontology defines a common vocabulary for sharing information in a domain [206,207]. It includes both human-understandable and machine-interpretable definitions of concepts in the domain and relations among them [206,207]. Typically, an ontology involves the formal, explicit description of 1) concepts in a domain of discourse – usually referred as "classes" or "concepts", 2) properties of each concept, describing various features and attributes of the concept – usually referred as "slots" or "properties", and 3) restrictions on slots – usually referred as "facets" or "role restrictions" [206]. An ontology can be used 1) to share the structure of information among people or software agents, 2) to make domain assumptions explicit, and 3) to analyze and reuse domain knowledge [206]. Formal ontology is "the systematic, formal, axiomatic development of the logic of all forms and modes of being" [208]. Formal ontologies are a popular research topic in many fields, such as knowledge management, knowledge engineering, and artificial intelligence [209].

The major steps of developing an ontology involve [206]:

1) Determine the domain and scope of the ontology;
2) Check the availability of existing ontologies in this domain, and consider reusing it, if there is any;
3) Enumerate important terms in the ontology and define the classes (concepts) and the class hierarchy;
4) Define the properties (slots) of classes;
5) Define the facets of the properties, such as the value type, allowed values, and the number of values;

6) Create instances.

5.2.2 LCCA-Onto: scope and language

The domain ontology developed in this research, LCCA-Onto, is focused on machine learning-enabled facility LCCA using the data provided by BIM and IoT. Currently, there is no ontology can fulfil this purpose in the facility LCCA domain. Existing ontologies, such as IFC [150], UniFormat [210], and MasterFormat [211], are adopted to represent the classes (concepts) in the corresponding domain (e.g. IFC represents the building components, UniFormat and MasterFormat represent the construction work breakdown structure).

LCCA-Onto is developed with the W3C Web Ontology Language (OWL). OWL is a Semantic Web language designed to represent ontologies [212]. The tool used to develop LCCA-Onto is protégé (version 5.5) [213]. The major classes involved in the LCCA-Onto are *Community*, *Building*, *Building_system*, *Data*, *Data_standard*, and *MACH_LRN_tool* (machine learning tool). A high-level representation of LCCA-Onto is shown in Figure 5.3.

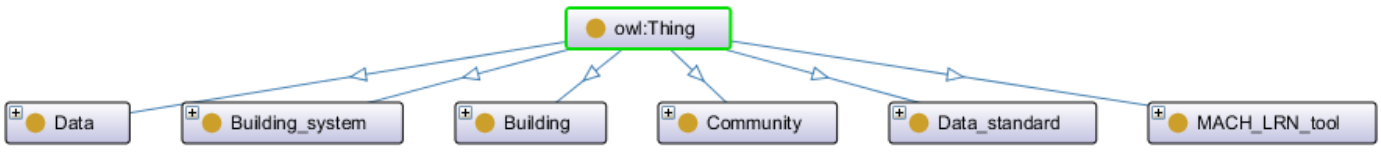


Figure 5. 3 A high-level representation of LCCA-Onto

5.2.3 LCCA-Onto: definition of class

In LCCA-Onto, the classes *Community* and *Building* do not have any subclasses. The class *Building_system* contains five subclasses: *BAS* (Building Automation System), *BEMS* (Building Energy Management System), *CMMS* (Computerized Maintenance Management System), *IWMS* (Integrated Workplace Management System), and *BIM* (Building Information Model).

The subclasses of *MACH_LRN_tool* represent the machine learning tools for predictive data analytics (non-exhaustive), which are listed in Table 5.1.

Table 5. 1 Definition of *MACH_LRN_tool*'s subclasses

Subclass	Description
Linear_regression	Linear regression, discussed in Section 4.2.1
SVM_regression	Support Vector Machines regression, discussed in Section 4.2.2
KNN_regression	K-Nearest Neighbors regression, discussed in Section 4.2.3
Regression_tree	Regression tree, discussed in Section 4.2.4
Time_series_regression	Time series forecasting and backcasting, discussed in Section 4.2.5
MLP_regression	Multilayer perceptron regression, discussed in Section 4.2.6

Data_standard contains subclasses: *City_data_standard*, *Building_data_standard*, and *IoT_data_protocol*. An instance of *City_data_standard* is CityGML, a city-level open data standard and exchange format to store digital 3D models of cities and landscapes [153]. Instances of *Building_data_standard* can be IFC [151] and gbXML [214], which are commonly used building data standard. Instances of *IoT_data_protocol* can be the data protocols of building automation and control, such as BACnet, Modbus, and Zigbee, which are listed in Table 3.5.

The hierarchy of class *Data* is shown in Figure 5.4 and the corresponding descriptions are presented in Table 5.2.

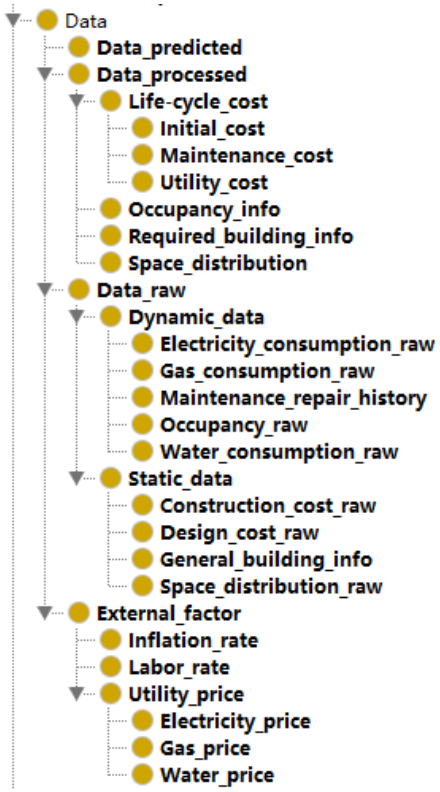


Figure 5. 4 The hierarchy of class *Data*

Table 5. 2 Definition of *Data*'s subclasses

Subclass		Description
Data_raw	Dynamic_data	Data streams that update within a short period of time (such as less than a month)
	Static_data	Data that do not change within a certain period of time (such as over six months)
External_factor	Inflation_rate	The overall increase in the Consumer Price Index (CPI), which is a weighted average of prices for different goods
	Labor_rate	The historical cost of labor that is used to deriving O&M related costs
	Utility_price	The historical price of utilities, including electricity, gas, and water prices.
Data_processed	Life-cycle_cost	The calculated LCC in present value, including the initial design and construction cost, the utility cost, and the O&M cost
	Occupancy_info	The occupancy related data in a format that can be used in the machine learning model development
	Required_building_info	The building data that are used as the descriptive attributes in the machine learning model
	Space_distribution	The organization's space management data used as the descriptive attributes in the machine learning model

Data_predicted	The data predicted with the developed machine learning model, such as a building's LCC or the LCC components

The subclasses in LCCA-Onto are not exhaustive and can be further expanded. For example, when a new machine learning method is proved to be effective in predicting facility-related costs, it can then be added as a subclass of *MACH_LRN_tool*. Another example is that when there is a new type of building system is invented and can provide useful data for facility LCCA, it can be a new subclass of *Building_system*.

5.2.4 LCCA-Onto: definition of property

In OWL, a property is a characteristic of a class – "a directed binary relation that specifies some attribute which is true for instances of that class" [212]. Properties may have domains and ranges. The domain is the subject of a relation while the range is the object of that relation. For example, in LCCA-Onto, the property *contains* has a domain of *Community* and a range of *Building*, which means community contains buildings, as Figure 5.5 shows.

Figure 5. 5 An example of OWL object property

There are two types of OWL property: Object property (*owl:ObjectProperty*) and Datatype property (*owl:DatatypeProperty*) [212]. The object property is used to represent relations between instances of two classes. The relation between the *Community* class and the *Building* class described above is an example for object property. Datatype property (owl: DatatypeProperty) is used to represent relations between object and data values, such as numerical values (e.g. integer, double and, float), boolean, and string. For example, the *Building* class has a datatype property *ID*, whose data type is string. A building instance whose *ID*=060A means the identification number of this building is 060A. Object properties of the LCCA-Onto are described in Table 5.3. Main object properties are shown in Figure 5.6, 5.7, and 5.8.

Table 5. 3 Definition of object properties in LCCA-Onto

Object property	Domain	Range	Description
contains	Community	Building	Community contains buildings
isEquippedWith	Building	Building_system	Buildings are equipped with building systems
exports	Building_system	Data_raw	Building systems export the raw data
isProcessedBy	Data_raw	MACH_LRN_tool	The raw data is processed by machine learning tools

derives	MACH_LRN_tool	Data_processed	Machine learning tools are used to derive the processed data
predicts	Data_processed	Data_predicted	The processed data is used to predict the facility LCC and its components
describes	Data_standard	Building Community Data	Data standards are used to describe buildings (e.g. IFC and gbXML), communities (e.g. CityGML), and data (e.g. BACnet)

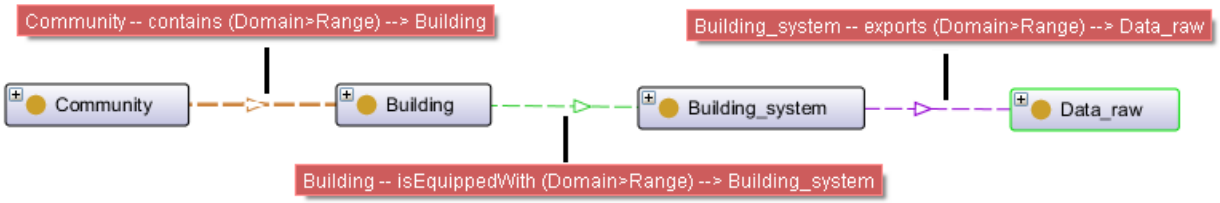


Figure 5. 6 Object properties *contains*, *isEquippedWith*, and *exports*

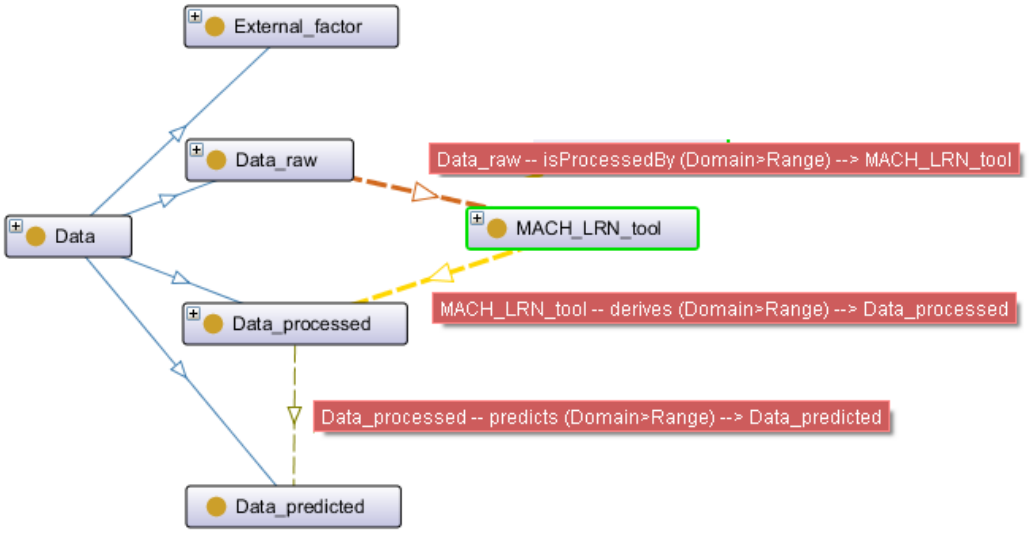


Figure 5. 7 Object properties *isProcessedBy*, *derives*, and *predicts*

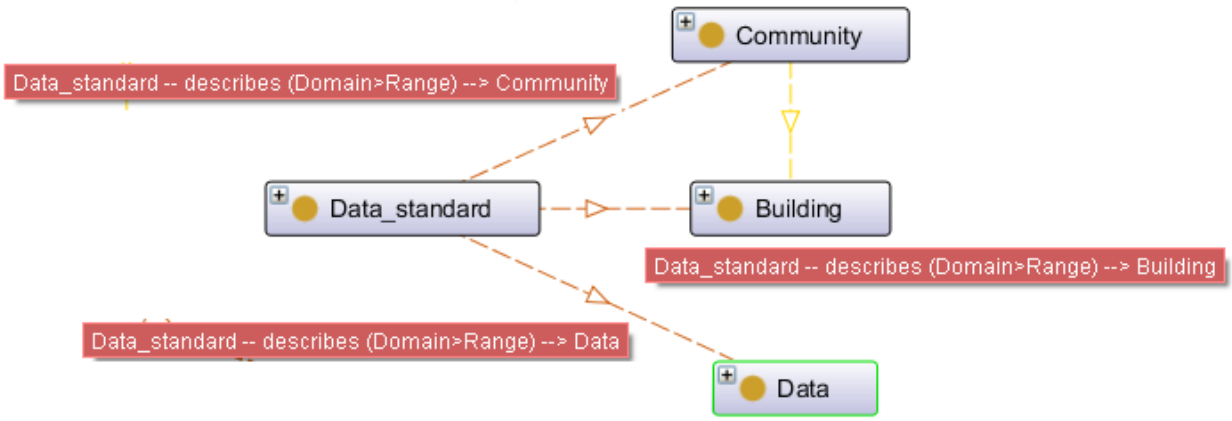


Figure 5. 8 Object properties *describes*

5.2.5 LCCA-Onto: incorporating existing ontologies

There are many ontologies applied in the AECO domain, including BIM open standard IFC (ifcOWL [215]), cost estimation work breakdown structure UniFormat II [210] and MasterFormat [211], the construction classification system for electronic databases – OmniClass [216], and the open data format for virtual 3D city models – CityGML [153], etc. Many of these ontologies were not developed with strict ontology development standards but they have become the de facto standards for concept communication in the AECO Industry [217].

Buildings systems already incorporate proprietary networks of sophisticated sensors and devices in the form of BAS, BEMS, CMMS, IWMS, etc. These building systems are built based on different data protocols, such as BACnet [154], Modbus [218], and LonWorks [219]. These data protocols used for building automation are actually IoT data protocols applied in the building domain.

In the LCCA-Onto, the existing AECO and IoT ontologies concerned with facility LCCA are incorporated as instances of the subclasses of *Data_standard* – *City_data_standard*, *Building_data_standard,* and *IoT_data_protocol,* as Figure 5.9 shows. These data standards and protocols are used to describe the community, the building, or the data generated by building systems. Different standards/protocols can be used in different cases. For example, the BAS used by an organization may adopt BACnet while that of another organization may be LonWorks. The LCCA-Onto does not specifically integrate with certain existing ontologies but uses the subclasses of *Data_standard* to represent data standard/protocol ontologies in general, thus to be flexible for different LCCA scenarios.

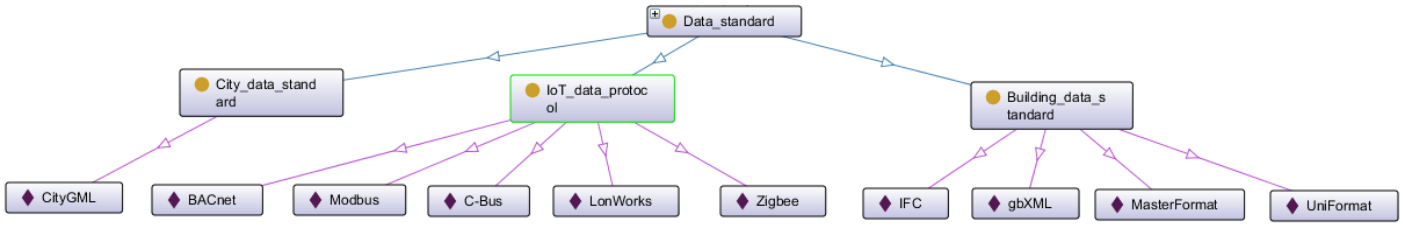


Figure 5. 9 Existing AECO ontologies and corresponding instances in LCCA-Onto

5.2.6 LCCA-Onto: the overall framework

The LCCA-Onto is expandable to include any facility cost related building system, data standard, data, and machine learning tools. Figure 5.11 shows the overall framework of LCCA-Onto, which summarizes the main classes and their relationships (object properties). The LCCA-Onto provides a foundation for the machine learning-based facility LCCA domain knowledge developed in this research.

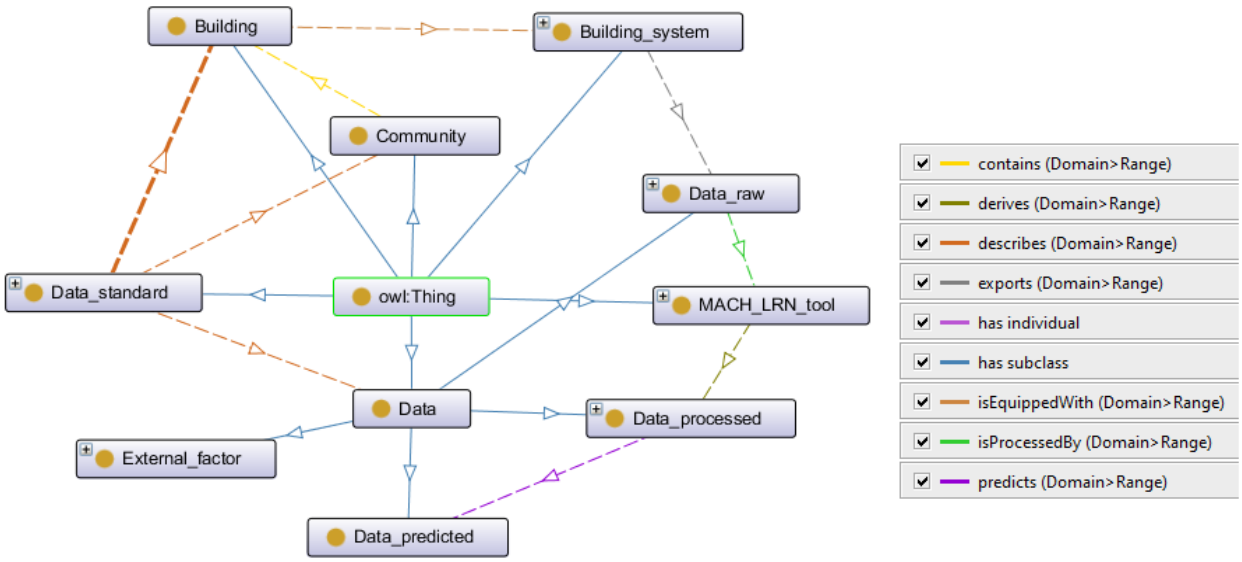


Figure 5.11 - The overall framework of LCCA-Onto

5.3 A Use Case Scenario: Facility LCC Prediction During the Programming Phase

One of the challenges usually faced by an organization's capital planning department and/or facility management department is that they do not have 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 Whitestone facility operation cost reference

[52] and maintenance and repair cost reference [51], may be overgeneralized and cannot reflect the organization's facility operation profile.

The proposed machine learning-based LCCA approach in this research provides organizations who own multiple facilities with a solution to the LCC prediction issue. The next chapter uses a case study on multiple facilities of a university to demonstrate the implementation process of the proposed approach for facility LCCA during the programming phase.

CHAPTER 6 PROOF-OF-CONCEPT VALIDATION

A series of experiments have been conducted on a university campus (hereafter referred as “the university”) using the proposed LCCA framework. This chapter presents these experiments and discusses the implementation process of the framework in detail. The overview of the experiments is presented first. Then, the data acquisition, processing, and integration works are demonstrated. After that, Machine learning model developments, evaluation, and comparison are discussed. The findings of the experiments are discussed at the end of this chapter.

6.1 The Overview

6.1.1 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 device networks embedded in the building systems are generating the data for developing the LCC prediction machine learning models. The building systems operated by the university involve BAS Metasys [155], CMMS AiM System [220], and the Capital Planning & Space Management System (CPSMS) INSITE [221].

6.1.2 The goal

In the experiments, the proposed machine learning-enabled LCCA framework was used to develop LCC prediction models for the university’s Budget Planning and Administration Department, and Facilities Management departments. These prediction

models were designed to forecast of facilities' LCCs during the programming phase using very limited input, such as Gross Square Footage (GSF), the owner of the building (which college), number of floors, and the space allocations. The goal is to provide these facility related departments a tool to quickly estimate the LCCs of the existing and to-be-built buildings (when building design is not available) without the input of building cost experts.

6.1.3 The buildings studied

Machine learning models were developed based on the historical data of 123 buildings on campus. The basic statistics of these buildings are shown in Table 6.1. The building types include residential buildings, libraries, dining halls, athletic facilities, parking decks, and educational complexes that consist of laboratories, classrooms, and offices.

Table 6. 1 The basic statistics information of the buildings in the case study

	Building age	**Gross Square Footage (GSF)**	**Number of Floors**	**Initial Cost (Present Value in 1999)**
Maximum	99	966,203	13	$113,216,000
Minimum	2	384	1	$280,000
Mean	39.37	96,871	3.9	$18,107,000
Median	33	48,666	4	$9,560,000

6.1.4 The programming phase

The programming phase is the first step in the building development process. During this phase, an estimation of a realistic project cost (initial design and construction costs) is typically needed for budget planning. The challenge in estimating the project cost at this point is that no design is available yet. The available information for cost estimation in the programming phase involves: 1) the building functions, 2) approximate building geometry, such as the GSF, the footprint, total height, and 3) space allocations, which are determined by user requirements.

6.2 Data acquisition

This section discusses the data acquisition from each building system – CPSMS, BAS, and CMMS.

6.2.1 The initial cost and space allocation

The university's CPSMS publishes the space management data of each campus building through a web system [222], which is based on Tableau [223]. This website also contains each building's initial cost. The raw data of initial cost and space allocation were downloaded from this website. Figure 6.1 shows the website's interface.

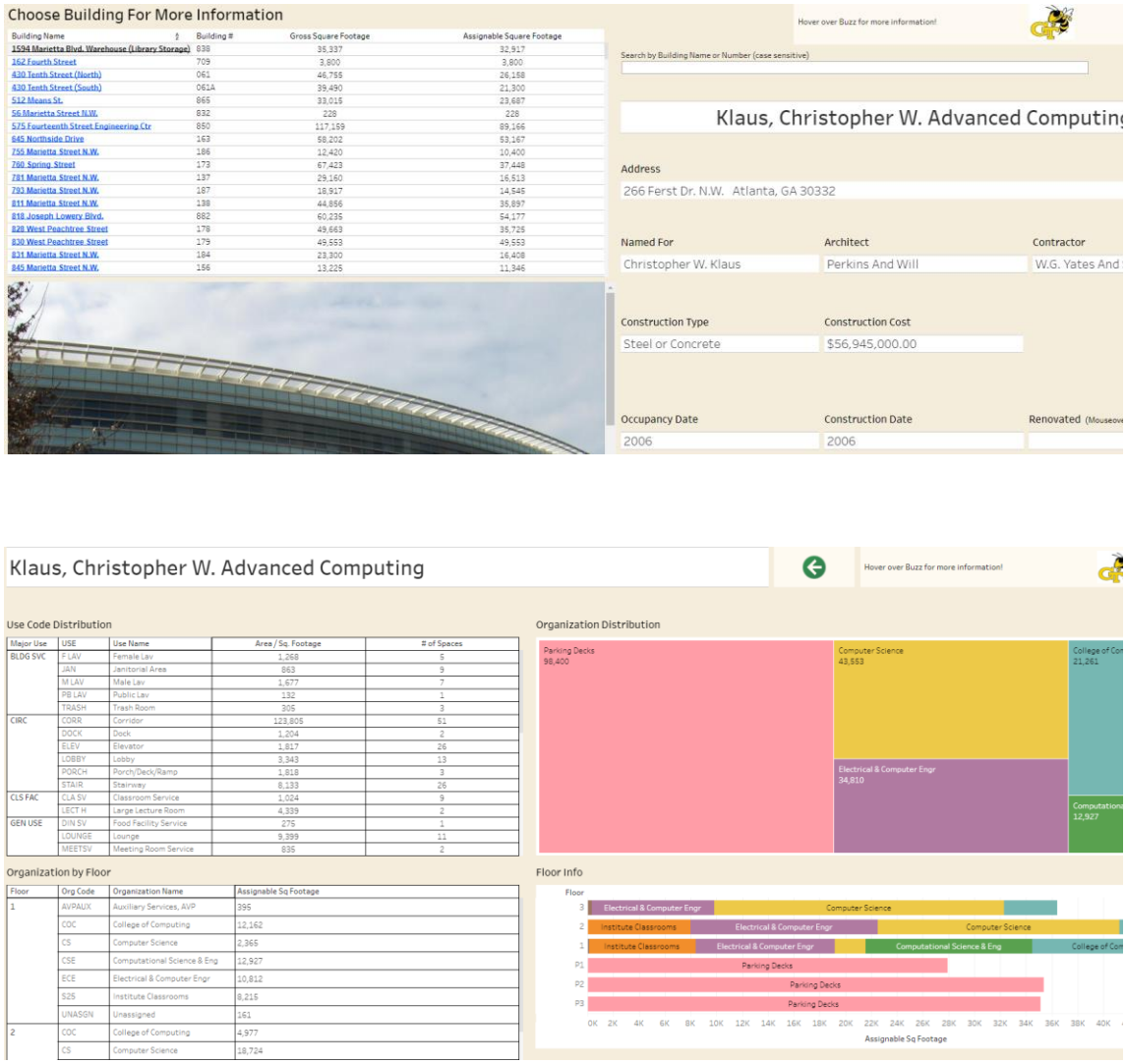
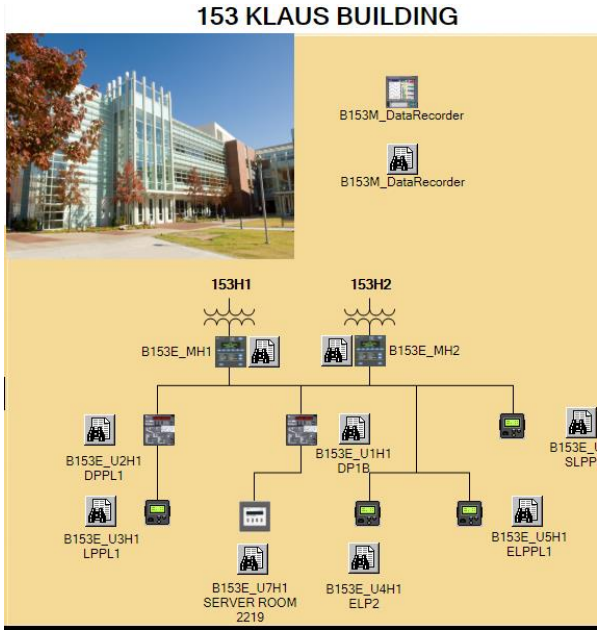


Figure 6. 1 The web-based building information dashboard

6.2.2 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 6.2) [224].



Timestamp	Real Energy Into the Load	Real Energy	Reactive Energy Into the Load	Reactive Energy	App Ener
2019/2/9 21:45:00.000	18,852,202.496	18,852,182.016	2,465,974.016	2,361,483.008	19,094,1
2019/2/9 21:30:00.000	18,852,167.680	18,852,145.152	2,465,969.920	2,361,477.888	19,094,1
2019/2/9 21:15:00.000	18,852,130.816	18,852,110.336	2,465,965.056	2,361,473.024	19,094,1
2019/2/9 21:00:00.000	18,852,096.000	18,852,073.472	2,465,959.936	2,361,467.904	19,094,0
2019/2/9 20:45:00.000	18,852,057.088	18,852,036.608	2,465,955.072	2,361,463.040	19,094,0
2019/2/9 20:30:00.000	18,852,022.272	18,852,001.792	2,465,949.952	2,361,457.920	19,094,0
2019/2/9 20:15:00.000	18,851,985.408	18,851,964.928	2,465,946.112	2,361,454.080	19,093,9
2019/2/9 20:00:00.000	18,851,948.544	18,851,928.064	2,465,940.992	2,361,448.960	19,093,9
2019/2/9 19:45:00.000	18,851,911.680	18,851,889.152	2,465,935.872	2,361,444.096	19,093,9
2019/2/9 19:30:00.000	18,851,874.816	18,851,854.336	2,465,932.032	2,361,440.000	19,093,8
2019/2/9 19:15:00.000	18,851,837.952	18,851,817.472	2,465,926.912	2,361,434.880	19,093,8
2019/2/9 19:00:00.000	18,851,801.088	18,851,780.608	2,465,922.048	2,361,431.040	19,093,7
2019/2/9 18:45:00.000	18,851,764.224	18,851,743.744	2,465,917.952	2,361,425.920	19,093,7
2019/2/9 18:30:00.000	18,851,725.312	18,851,706.880	2,465,913.088	2,361,422.080	19,093,7
2019/2/9 18:15:00.000	18,851,688.448	18,851,667.968	2,465,908.992	2,361,416.960	19,093,6
2019/2/9 18:00:00.000	18,851,651.584	18,851,629.056	2,465,904.128	2,361,412.096	19,093,6
2019/2/9 17:45:00.000	18,851,612.672	18,851,592.192	2,465,899.008	2,361,406.976	19,093,6
2019/2/9 17:30:00.000	18,851,577.856	18,851,557.376	2,465,893.888	2,361,402.112	19,093,5
2019/2/9 17:15:00.000	18,851,540.992	18,851,520.512	2,465,889.024	2,361,396.992	19,093,5
2019/2/9 17:00:00.000	18,851,504.128	18,851,483.648	2,465,883.904	2,361,392.128	19,093,4
2019/2/9 16:45:00.000	18,851,467.264	18,851,446.784	2,465,879.040	2,361,387.008	19,093,4
2019/2/9 16:30:00.000	18,851,430.400	18,851,409.920	2,465,873.920	2,361,381.888	19,093,4
2019/2/9 16:15:00.000	18,851,393.536	18,851,373.056	2,465,868.032	2,361,377.024	19,093,3
2019/2/9 16:00:00.000	18,851,356.672	18,851,336.192	2,465,862.912	2,361,370.880	19,093,3
2019/2/9 15:45:00.000	18,851,319.808	18,851,299.328	2,465,858.048	2,361,366.016	19,093,3

Figure 6. 2 The website that publishes the utility consumption data

For most buildings of the university, the utility consumption data are available since October 1st, 2012, with an interval of 15 minutes. In this research, the utility data used were from October 1st, 2012 to September 1st, 2018. The data (CSV files) were downloaded through Ion Data Grabber [225] from the EnergyWatch system developed by the university's Aerospace Systems Design Laboratory [226]. An example of the electricity consumption raw data (a small portion) is shown in Appendix C.

6.2.3 The O&M costs

The university's O&M work order records in the CMMS, AiM System, are available since 2006. Up to September 1st, 2018, there were over 750,000 lines of records. These

records were exported from the AiM System as a CSV file. Appendix D shows a small portion of the raw data of O&M work order records.

6.3 Data processing

6.3.1 Data cleaning

Based on the raw data, each building's utility weekly consumption and monthly consumption were calculated, and outliers removed. The MATLAB code used for this processing are presented in Appendix E. Most of the studied buildings' monthly utility consumptions show repetitive patterns every year. Figure 6.3 shows two examples.

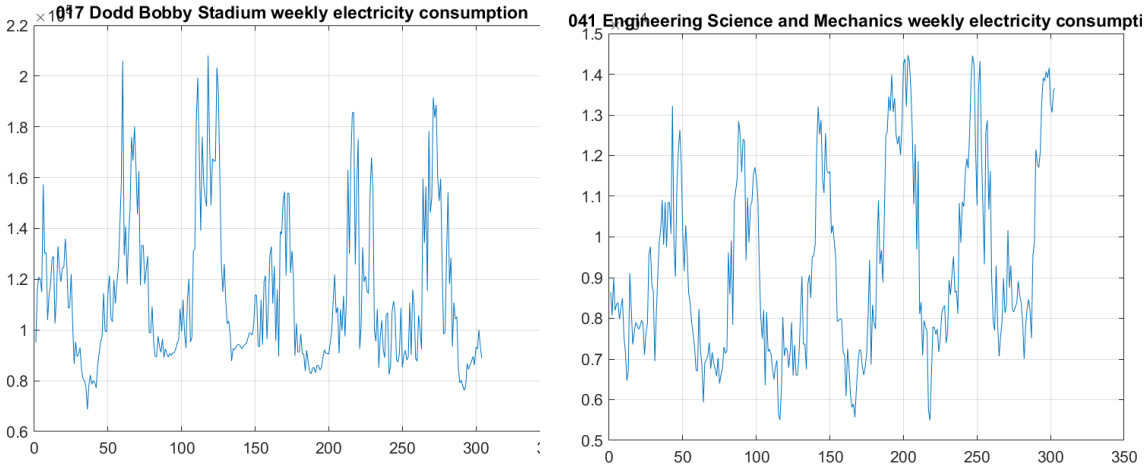
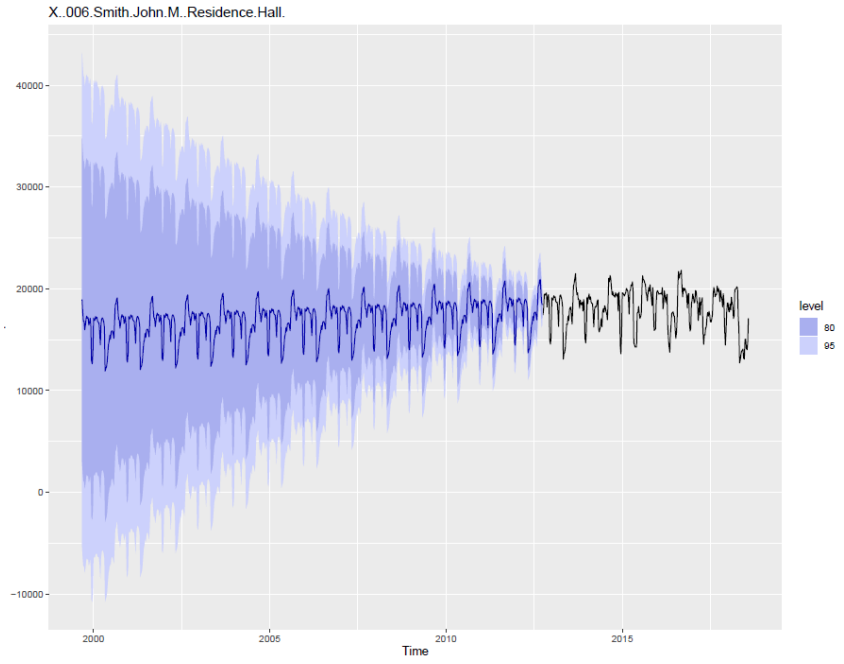


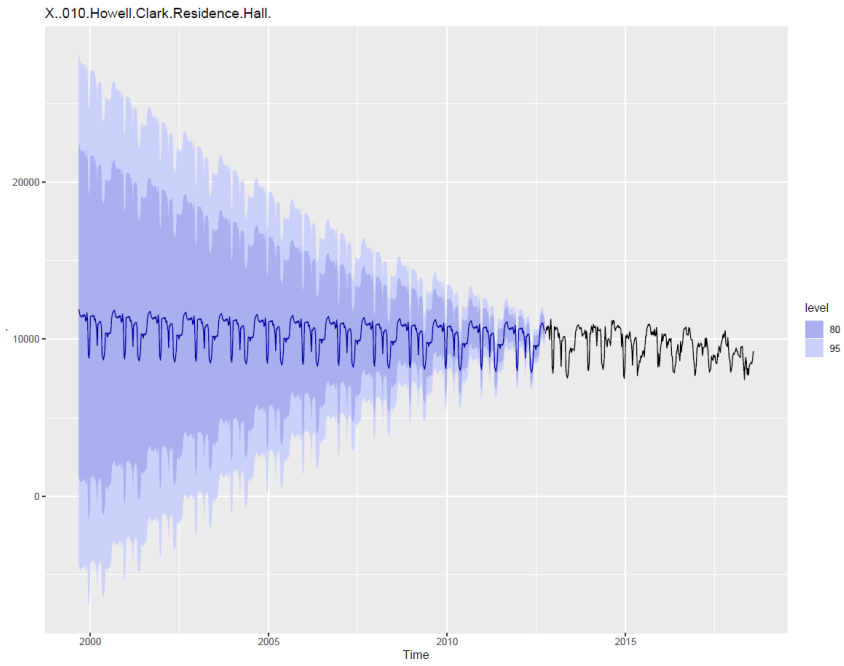
Figure 6. 3 Examples of building electricity consumption trends

OpenRefine [227] and MATLAB were used to clean the O&M work order records and thus to obtain the annual O&M cost of each building. The OpenRefine operation history and MATLAB code for annual O&M cost calculation are presented in Appendix F and Appendix G, respectively.

6.3.2 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 do not have some of the historical data before a certain year (because the building was newly built or sensors were not deployed), the author used time series backcasting to simulate the data. The machine learning software tool Microsoft R [228] to perform time series backcasting to simulate the past utility consumptions and O&M cost. The R code are presented in Appendix H. Figure 6.4 shows three examples of building electricity consumption backcasting. The electricity consumptions revealed repeating patterns. In the figure, the darker blue area is the prediction interval of 95%; the lighter blue area is the prediction interval of 80%. For the buildings whose utility consumption or O&M cost did not show a repeating pattern, the mean of actual annual cost was used as the simulated data.


X..006.Smith.John.M..Residence.Hall.
40000
30000
20000
10000
0
−10000
2000
2005
2010
2015
Time
level
80
95


X..010.Howell.Clark.Residence.Hall.
20000
10000
0
2000
2005
2010
2015
Time
level
80
95

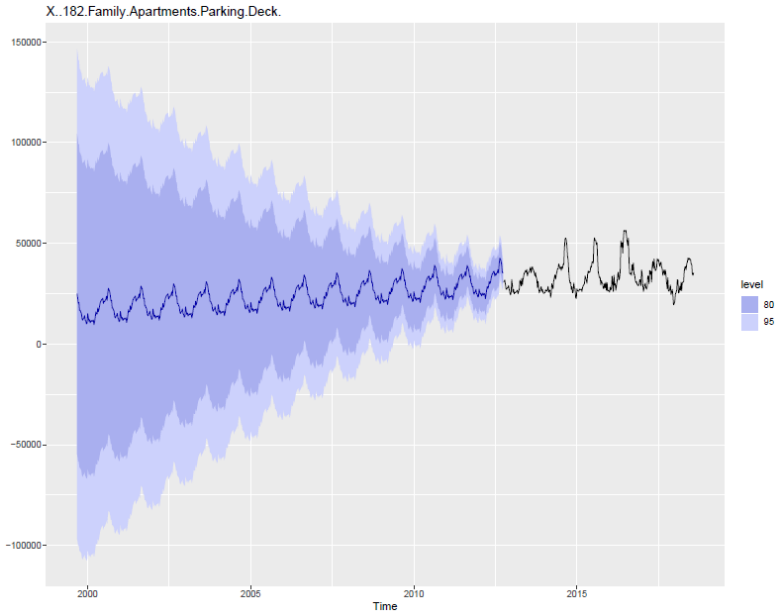


Figure 6. 4 Three examples of building electricity consumption backcasting

6.3.3 Discounting to present value

The present value (in 1999) of the initial cost, utility cost, and O&M cost were calculated according to the equation (1), (2), and (3) in Section 4.4, correspondingly. The historical annual inflation rate used was based on the statistics of Consumer Price Index (CPI) provided by the Bureau of Labor Statistics (BLS) [203,229]. The utility price used was the Average Energy Prices provided by BLS [230]. The O&M labor rate used was based on the Current Employment Statistics (CES National) provided by BLS [231].

6.4 Model development

6.4.1 Descriptive attributes

Because the prediction models were designed to forecast of facilities' LCCs during the programming phase, hence the descriptive attributes (model inputs) used are the ones that can be determined in this phase. The descriptive attributes involved are listed in Table 6.2.

Table 6. 2 The descriptive attributes of the machine learning models

Gross Square Footage (gsf)	The Architect Company (architect)	The General Contractor (contractor)
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.

Most of descriptive attributes are related to space allocation, such as the percentage of building service area, classroom facilities, and laboratory facilities.

6.4.2 Data analysis

This experiment involves the data of 123 buildings and 20 descriptive attributes. The linear correlation of each numerical attributes is presented in Appendix I. According to Appendix I.1, the total utility cost shows a strong positive correlation with the initial cost and O&M cost (0.68 and 0.70, respectively); the initial cost and O&M cost also show a moderate positive correlation of 0.58. The initial cost also shows a moderate positive correlation with GSF (0.56). These results are intuitive – in terms of the total cost, larger buildings are more expensive to build and to operate. The scatterplot matrix of the correlation of selected attributes is shown in Figure 6.5.

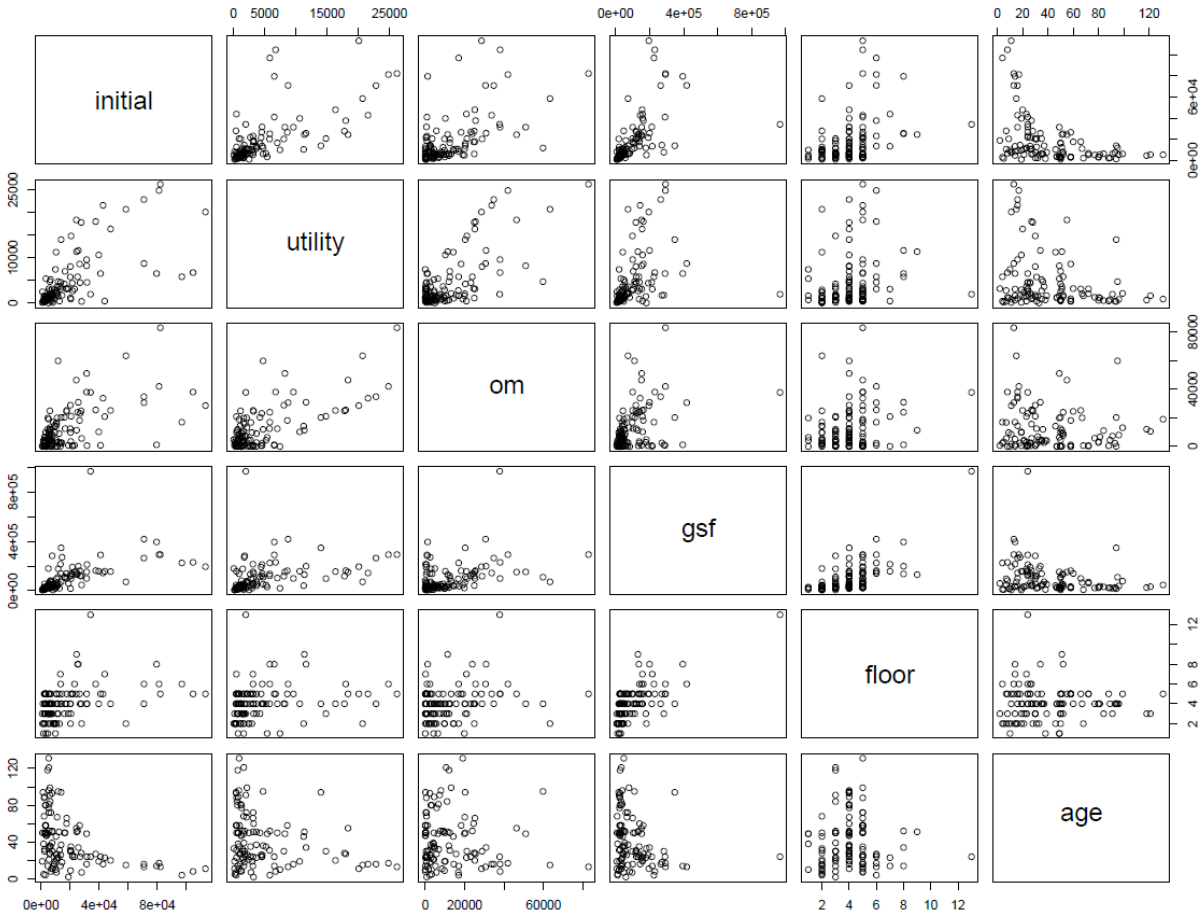


Figure 6. 5 The scatterplot matrix of the correlation of selected attributes

Appendix I.2 presents the linear correlation of each numerical attributes, with the initial cost, utility cost, and O&M cost are changed to cost per square footage (SF). In terms of cost per SF, the initial cost shows weak positive correlations with the utility cost (0.27), O&M cost (0.34), special use facilities (0.33). This indicates that more expensive buildings (per SF) tend to cost more in utility and O&M. The buildings with more percentage of special use facilities are 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 shows a weak positive correlation with office facilities (0.40) and negative correlations with the number of floors (-0.35), residential facilities (-0.41). These results imply that the buildings with more floors tend to cost less in utility and O&M. The buildings with more percentage of office facilities tend to be more expensive to operate and maintain, while residential buildings cost less in O&M. The scatterplot matrix of the correlation of selected attributes (cost per SF) is shown in Figure 6.6.

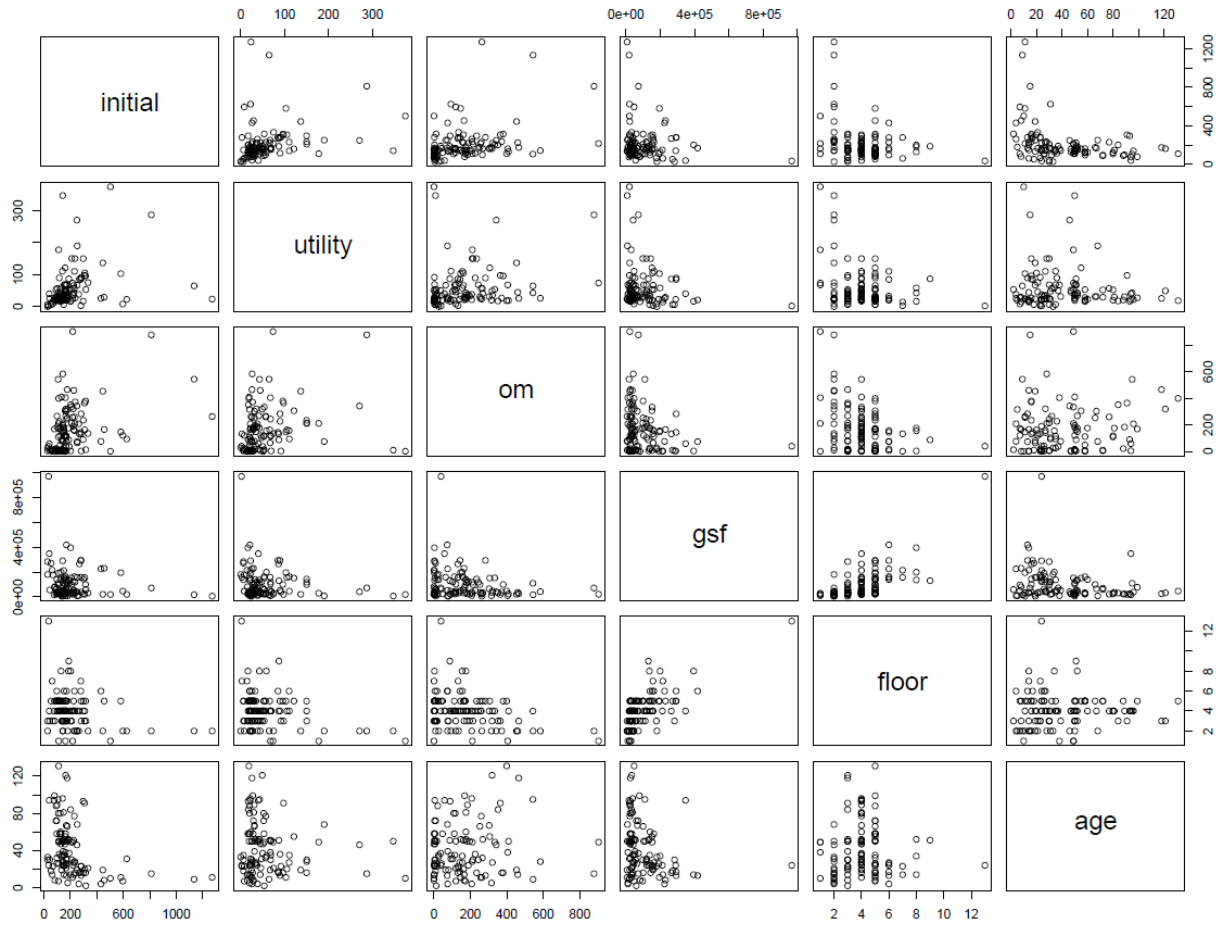


Figure 6. 6 The scatterplot matrix of the correlation of selected attributes (cost per SF)

Figure 6.7 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 most buildings. Buildings owned by the College of Engineering are more expensive on average. The residential buildings and parking decks are the least expensive facilities in terms of cost per SF.

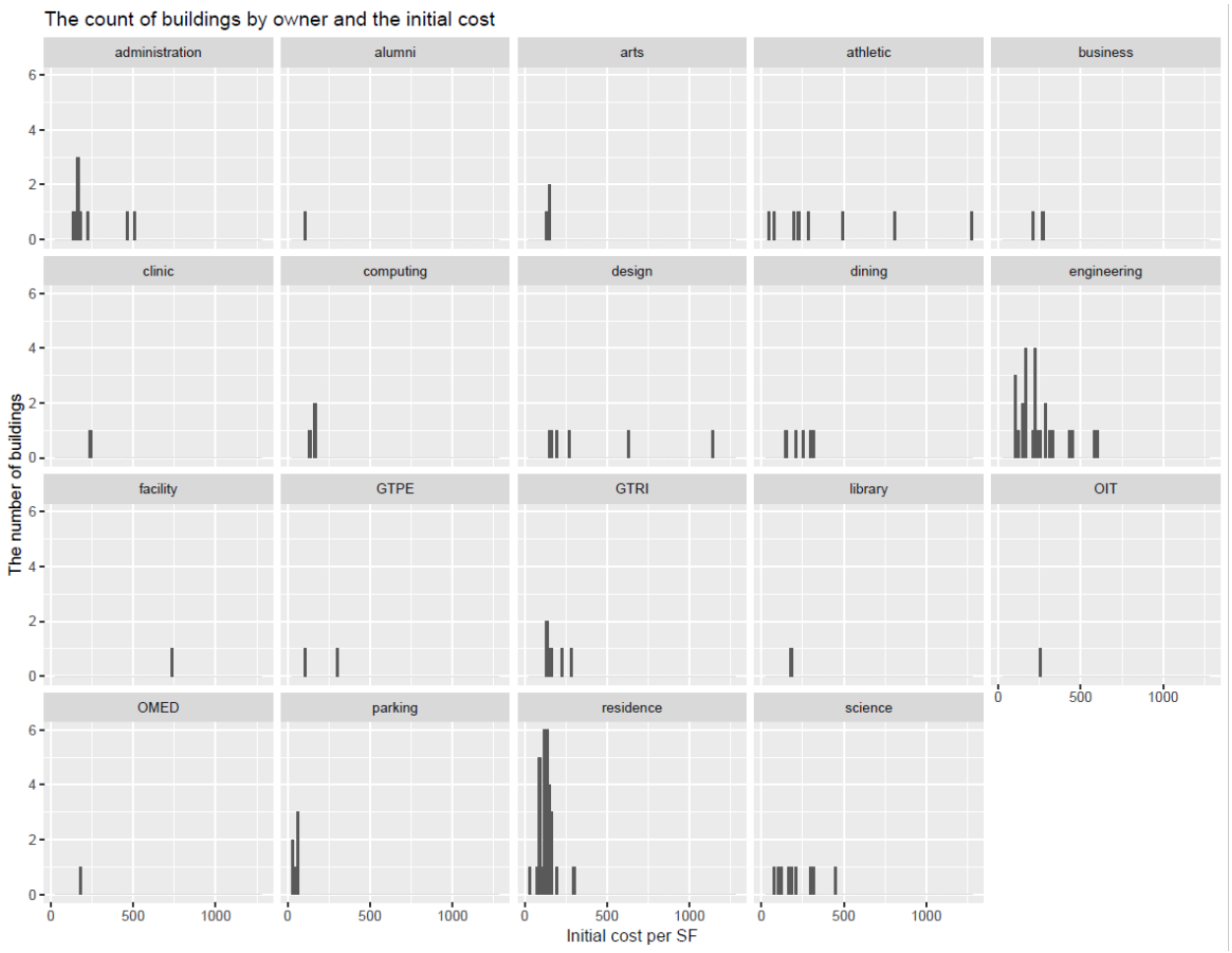


Figure 6. 7 The count of buildings by the college/owner and the initial cost per SF

Figure 6.8 shows the number of buildings by the college/owner and the utility cost per SF. Figure 6.9 shows the number of buildings by the college/owner and the O&M cost per SF. They indicate that the buildings owned by the College of Engineering are generally more energy intensive and cost more on O&M. There is an athletic facility and an administrative building cost more than $300 dollars per SF (present value in 1999) on 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.

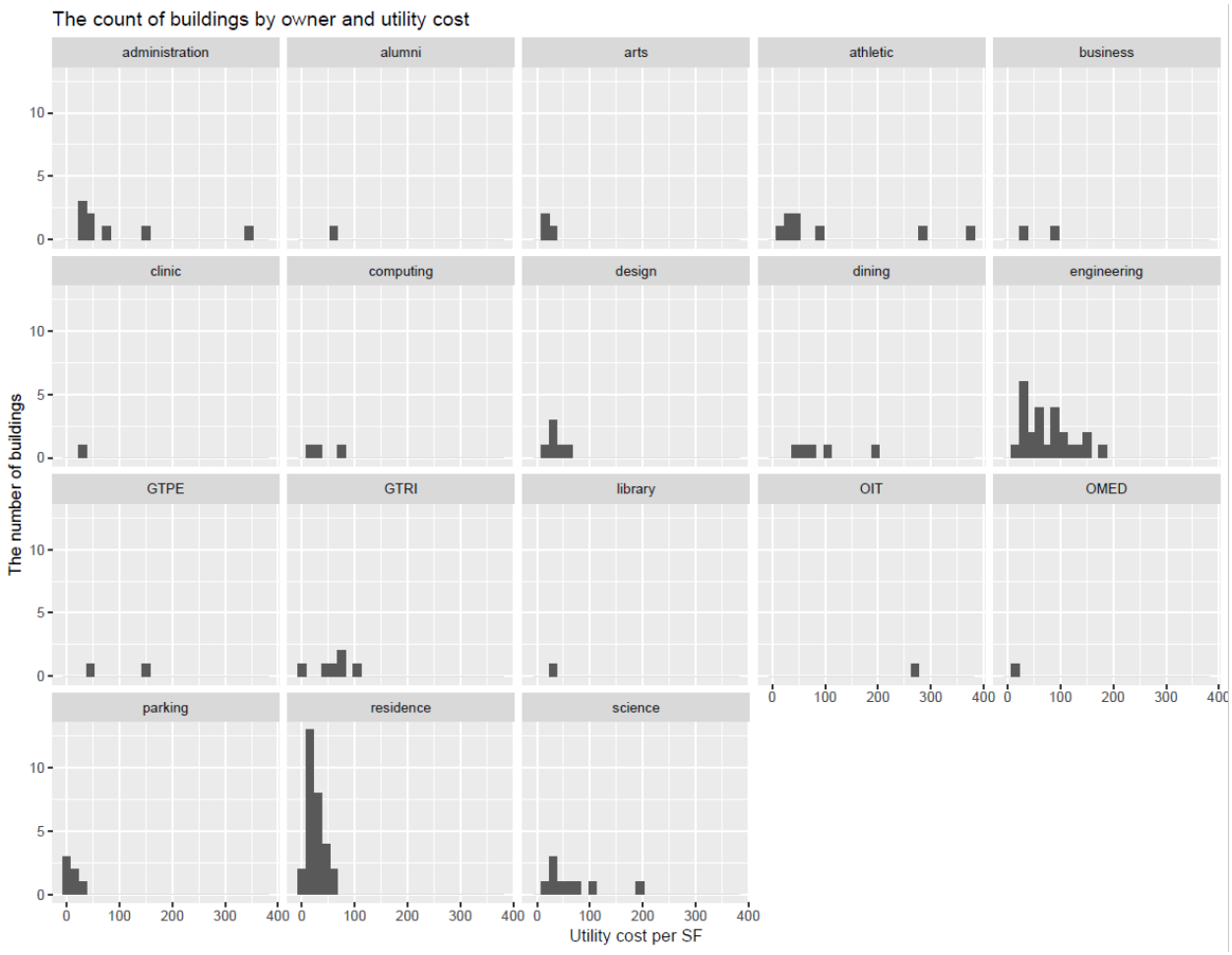


Figure 6. 8 The count of buildings by the college/owner and the utility cost per SF

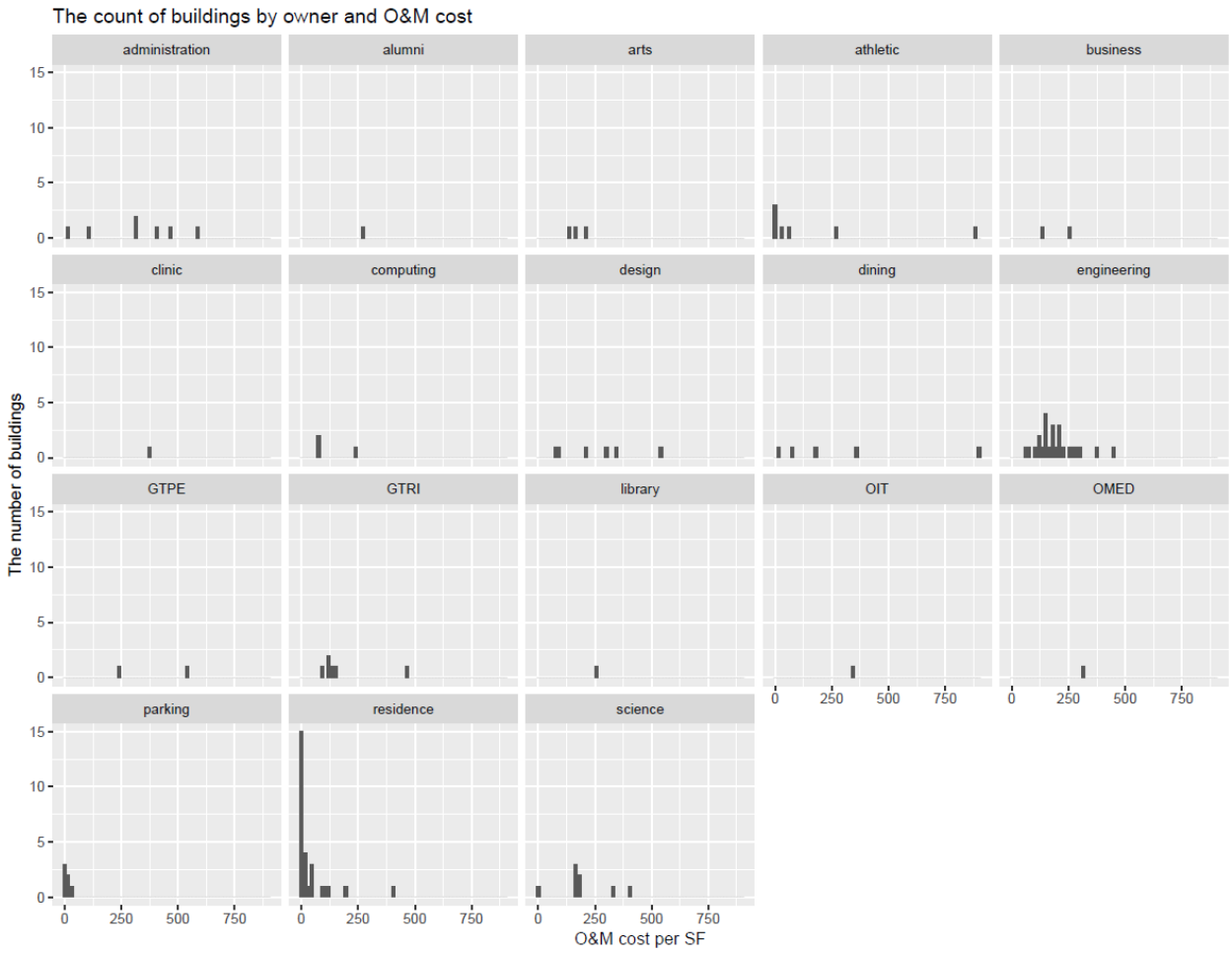


Figure 6. 9 The count of buildings by the college/owner and the O&M cost per SF

The R code for the basic data analysis presented in this subsection is shown in Appendix J.

6.4.3 Model training and validation

The author developed two kinds of machine learning models for LCC prediction – the single-target regression model and the multi-target regression model. The former assumes the LCC components (the target features) are independent of each other, while the

latter considers their intercorrelations. To develop the single-target regression model, the author tested multiple regression algorithms, involving 1) multilinear regression, 2) kNN, 3) random forest, 4) SVM, and 5) multilayer perceptron. To develop the multi-target regression model, the author tested multi-output random forest and multilayer perceptron.

To determine the best performing algorithm, the experiment repeated 100 times (loops). In each loop, the full dataset was randomly split into the training set and the test set with a proportion of 8:2. Then, the training set was used to train the machine learning models with each algorithm. The trained models were then tested with the test set and the mean absolute error (MAE) was used to evaluate the 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 multilinear regression models were developed with the R package *stats* version 3.5.1 [232]. The method used for fitting was QR decomposition.

The KNN regression models were developed with the R Package *FNN* version 1.1 [233]. The experiment indicated that the most suitable number of neighbors considered (k) was 3, which yields the most accurate predictions. The nearest neighbor search algorithm used was KD Tree.

The random forest regression models were developed with the R Package *randomForest* version 4.6-14 [234]. 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 [235]. 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 3.

MLP single-target regression models were developed with the R Package *keras* version 2.1.6 [236]. The MLP models contained three hidden layers with 10 nodes, 8 nodes, and 5 nodes, correspondingly. The batch size (the number of samples per gradient update) was set to 80. The number of epochs to train the model was 100. 95% of the training set instances were used to train the model and 5% were used for validation in each epoch.

The multi-output random forest model was developed with the R Package *MultivariateRandomForest* version 1.1.5 [237]. 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.

MLP multi-target regression model was also developed with the R Package *keras* version 2.1.6 and contained three hidden layers with 10 nodes, 8 nodes, and 5 nodes, correspondingly. The batch size was set to 90. The number of epochs to train the model was also 100. 90% of the training set instances were used to train the model and 10% were used for validation in each epoch. Figure 6.10 shows the structure of the developed MLP multi-target regression model.

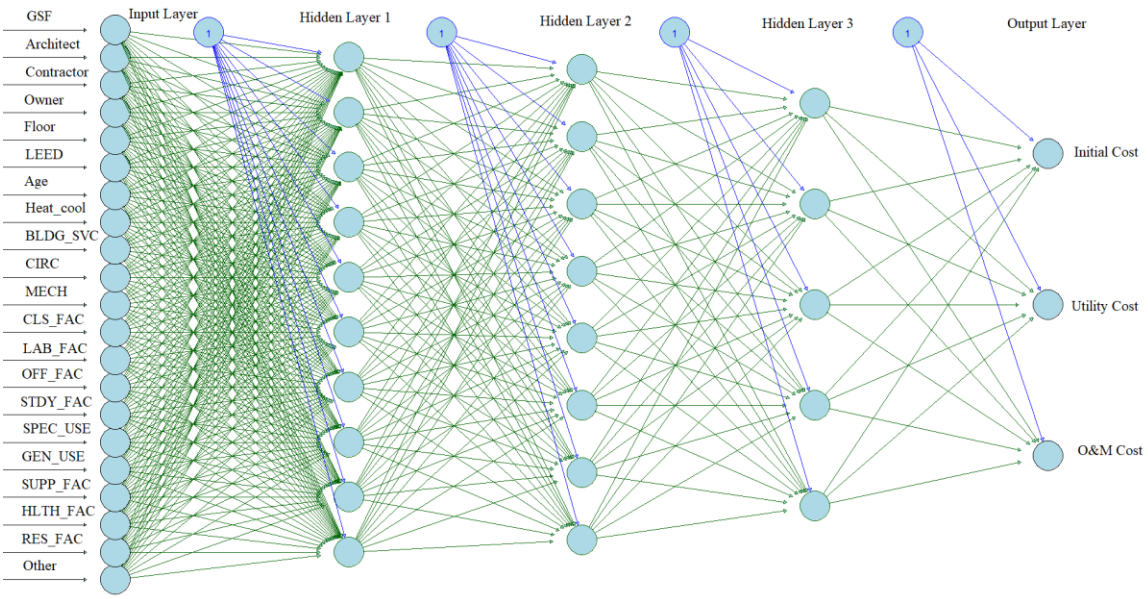


Figure 6. 10 The structure of the MLP multi-target regression model for facility LCC prediction

The R code for model training and validation is shown in Appendix K.

6.5 Results and discussions

The prediction results (MAE) of each model developed are shown in Table 6.3. When developing the model, the target attributes – initial, utility, and O&M costs – were normalized based on the mean and standard deviation of the overall data. Hence, the values of these target attributes are ranging from 0 to 1. This gives more intuitive results to interpret and to compare the accuracy of each model.

Table 6. 3 The evaluation results of each machine learning model in MAE (normalized)

		Initial Cost (%)	Utility Cost (%)	O&M Cost (%)
Single-target regression model	Multilinear regression	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%
	Multilayer perceptron	57.32%	59.81%	52.48%
Multi-target regression model	Multi-output regression Random forest	46.44%	55.88%	58.84%
	Multilayer perceptron	51.68%	57.71%	57.35%

The evaluation results indicated that the SVM regression models give the most accurate predictions and the single-target models perform better than the multi-target models. Based on the results of 100 experiments, the single-target regression models using SVM have average MAE of 27.97%, 37.08%, and 39.40% for the predictions of initial cost, utility cost, and O&M cost, respectively. The developed machine learning models provides decision makers a tool to quickly estimate the LCC of a facility.

The experiments presented in this Chapter demonstrated the implementation of proposed LCCA framework in detail. The data acquisition and processing, machine learning model developments, evaluation, and comparison are discussed. Even though these experiments were conducted on a university campus and the buildings studied are all associated with education, the proposed LCCA approach is applicable to any kind of organization that owns multiple facilities.

In the envisioned future smart city (presented in Chapter 7), the facility LCCA can be one of the services provided by the city-level IoT network. As the historical facility cost-related data are collected by the network, the LCCA can be conducted by some computing providers and the results could be shared with interested stakeholders.

CHAPTER 7 THE BIM AND IOT-ENABLED SMART BUILT ENVIRONMENT

So far, this dissertation has presented a literature review on facility related cost prediction, proposed a framework for developing machine learning models for facility LCCA, demonstrated how to implement the framework to obtain a quick estimation of facilities' LCCs based on the historical data generated in a "smart-like" built environment that contains sensor network and BIM. The entire study is a use case identified and realized under an initiative of BIM and IoT-enabled smart built environment innovations [238].

This chapter extends the discussion to the author's vision of the future BIM and IoT-enabled smart built environment. First, the background of IoT, Cyber-physical System (CPS), and the Smart Built Environment are introduced. Then, an architecture of the envisioned smart city is presented and the idea of the "Basic Facility Data Package" (BFDP), which is the foundation of the data infrastructure for the envisioned smart city, is proposed and discussed. This research is a proof of concept for the envisioned future BIM and IoT-enabled smart built environment.

7.1 The Cyber-physical Systems and the Smart Built Environment

According to the United Nations, the world's urban population is projected to grow by 2.5 billion from 2014 to 2050, and will account for 66 percent of the total global population by then [239]. The growing population in cities increases the demand for the fundamental needs of people living there, such as housing, utilities, medical care, welfare, education and employment [240]. To deal with challenges faced during the growth of cities,

the concept of Smart City has been envisioned, which denotes “the effective integration of physical, digital and human systems in the built environment to deliver a sustainable, prosperous and inclusive future for its citizens” [241]. As the cells of smart cities, Smart Buildings are buildings which integrate intelligence, enterprise, control, and materials and construction as an entire building system to meet the drivers for building progression: energy efficiency, longevity, comfort, and satisfaction [242]. In both the contexts of smart cities and smart buildings, the “smart” refers to the development, integration, and utilization of intelligent systems based on Information and Communication Technologies (ICT) and, more specifically, the CPS [243].

7.1.1 The definition of CPS

CPS refers to smart systems that include engineered interacting networks of physical and computational components [244]. It is also an umbrella term and concept that can represent many other words and phrases that describe similar or related intelligent systems, including the Internet of Things (IoT), machine-to-machine (M2M), industrial internet, digital city, etc. A CPS consists of the physical part – a device, a machine, or a building – and the digital or cyber part – the software system, communication network, and the data. The cyber part of CPS represents digitally the state of the physical part and impacts it by automated control or informing people of control actions.

Researchers working in the CPS field are trying to connect the digital systems (the Internet, data, software applications, etc.) and the physical realm (machines, infrastructure, building components, etc.) to enable innovative applications and services. The built environment is a critical component of the IoT-enabled Smart City paradigm. Buildings,

along with infrastructures and vehicular/transit systems, comprise the platform into which ubiquitous computing and IoT systems are embedded. Buildings represent highly structured spatial environments – organizational systems of connect spaces and components – that can provide a strong semantic overlay on the organization and interaction between IoT devices and their environments. Buildings provide intrinsic organizational information about the communities, businesses and operations of the people, equipment and systems they house. Buildings systems already incorporate proprietary networks of sophisticated sensors and devices in the form of energy systems, security systems, and emerging smart home devices, albeit with limited inter-system connectivity or exposure to the larger networks of IoT devices. These smart building sensor networks represent potential platforms for the deployment of more generalized IoT networks, and are sources of occupancy and space that can provide significantly enhanced value to new IoT systems.

7.1.2 The NIST CPS Framework

The National Institute of Standards and Technology (NIST) Engineering Laboratory is leading a program to advance Cyber-Physical Systems [243]. In this program, the NIST's CPS Public Working Group has developed a CPS Framework that presents a set of high-level concepts, their relationships, and a vocabulary for clear communication among stakeholders (e.g. designers, engineers, users) [244]. The goal of the CPS Framework is "to provide a common language for describing interoperable CPS architectures in various domains so that these CPS can interoperate within and across domains and form systems of systems" (SoS) [244]. Figure 1 shows the CPS framework developed by the NIST's CPS Public Working Group.

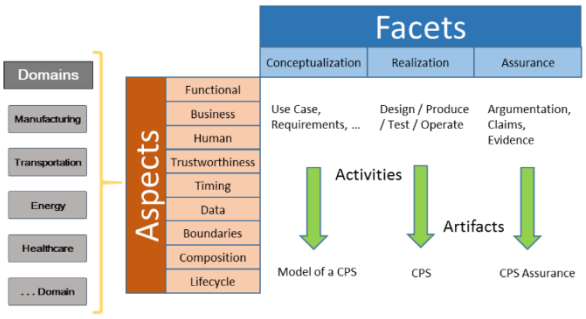


Figure 7. 1 CPS Framework – Domains, Facets, Aspects [244]

The NIST CPS framework can be used as guidance in designing, building, and verifying CPS and as a tool for analyzing complex CPS [244]. It consists of three major components: domains, aspects, and facets. The domains are the industries that the CPS can be specialized and applied to, such as manufacturing, transportation, and energy. The aspects are high-level groupings of cross-cutting concerns of CPS, involving functional, business, human, data, etc. Concerns are interests in a system relevant to one or more stakeholders. The facets are views on CPS encompassing identified responsibilities in the system engineering process. They contain well-defined activities and artifacts (outputs) for addressing concerns [244]. The three facets, conceptualization, realization, and assurance, deal with three major questions, respectively, which are 1) what things should be and what things are supposed to do, 2) how things should be made and operate, and 3) how to prove things work the way they should [244].

7.1.3 The NIST IoT - Enabled Smart City Framework

Effective smart city solutions are facing two major barriers. First, many current smart city ICT solutions are based on custom systems that are not interoperable, extensible, or cost-effective [245]. Second, many smart city standardization efforts, such as the ones carried out by ISO/IEC JTC1 [241], IEC [246], and IEEE [247], are currently underway but have not yet converged, creating uncertainty among stakeholders [245]. To reduce these barriers, NIST and its partners are convening an international public working group to develop a consensus framework of a common language/taxonomy and smart city architectural principles that can align existing smart city efforts that will support interoperable and portable smart applications [245]. The framework is named "IoT-Enabled Smart City Framework" (IES-City Framework). It adopts the concepts defined in the NIST CPS framework (e.g. domains, aspects, concerns, and facets) and maps them to exemplary smart city deployments.

The IES-City Framework is still under developing. This framework is requiring innovative IoT use cases to demonstrate how the concepts and tools enabled by IoT can be used in practice. The facility LCCA use case developed in this study is a proof of concept under the IES-City Framework.

7.2 A Vision for Future Smart City – An IoT Network of Smart Facilities

A conceptual BIM and IoT-enabled Smart City architecture in the perspective of a smart building network is proposed and shown in Figure 7.2.

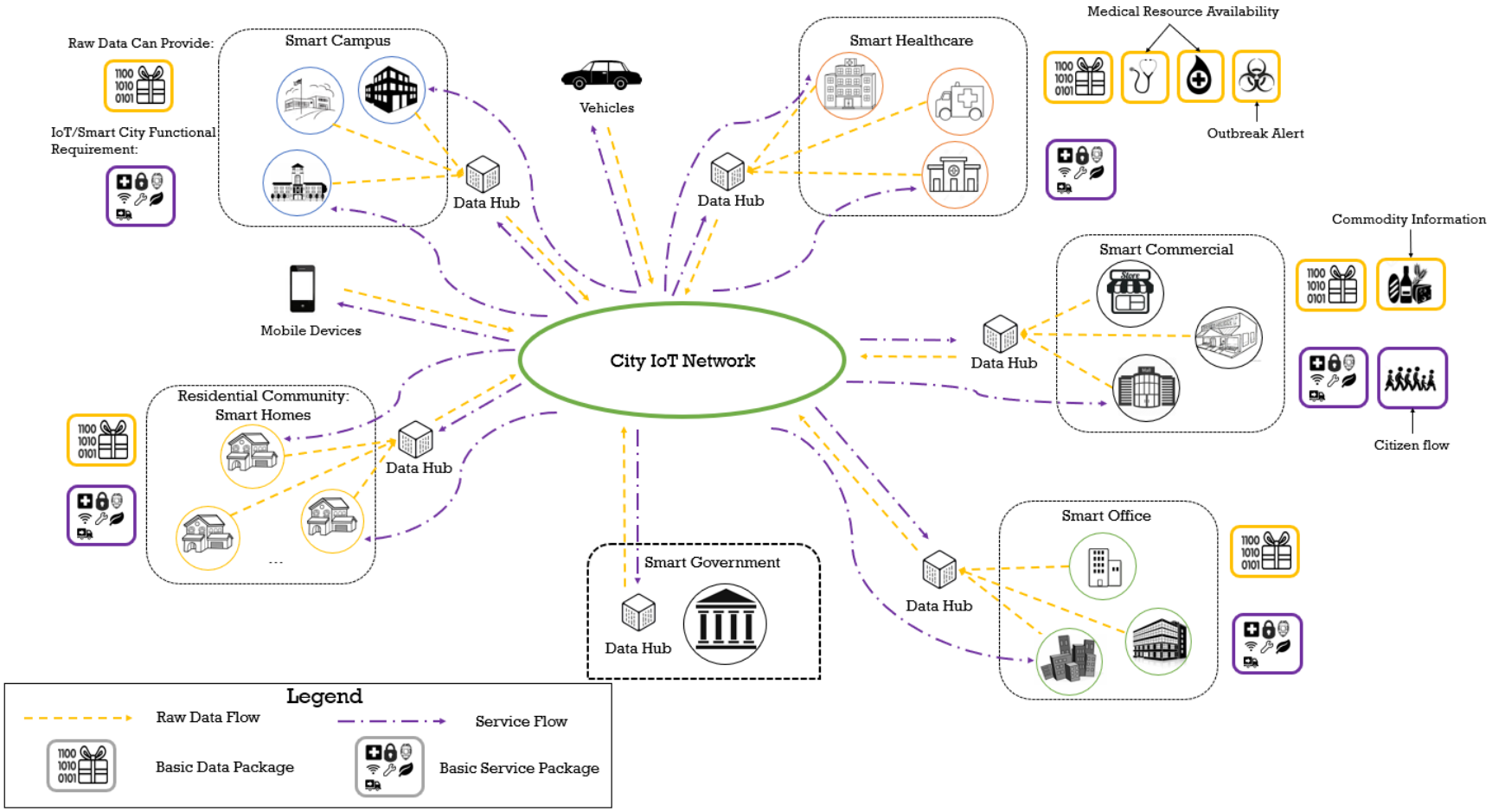


Figure 7. 2 An architecture of the envisioned future Smart City – an IoT network of smart facilities

In this envisioned smart city, multiple smart buildings form a community and many communities – residential community, campus, healthcare, commercial, office, government, etc. – form a smart city. In the future, each facility will be "smart" enough to provide a certain amount of data to the smart city's IoT network in real-time. The data flow generated in each building is collected by the data hub of each community, and then connected to the city-level IoT network. The data contents can vary based on the facility type but some of them are universal. The author names the data that will be provided by all smart buildings "the basic facility data package" (BFDP). The BFDP provides the fundamental data for the smart building IoT network and is the basis of innovative BIM and IoT-enabled smart city applications. The BFDP is evolving with time and may never be exhaustive. The LCCA data package presented in the Section 3.3.1 is a subset of the BFDP. The BFDP is a general dataset that can be used for many different use cases. A preliminary list of the BFDP's contents is shown in Table 7.1 [238].

Table 7. 1 A preliminary list of the basic facility data package's contents [238]

#	Basic Data	Description
1	Space occupancy rate	The ratio of rented or used space compared to the total amount of available space
2	People counting	Number of people in a space during a certain time
3	Electricity usage	Real-time electricity consumption
4	Water usage	Real-time water consumption
5	Lighting relevant information	e.g. illuminance, natural lighting information
6	Audio relevant information	e.g. noise level

7	Access relevant information	e.g. personnel attendance record, visitor access record
8	Logistics	e.g. inventory information, accessibility of the facility, RFID tagged goods in/out records
9	Environmental information	e.g. Real-time temperature, humidity, harmful substances content
10	Emergency information	e.g. smoke detection, fire alarm equipment information
11	Operation & maintenance	e.g. work order, maintenance record
12	Information and Communication Technology relevant information	e.g. Internet usage, WIFI coverage

Besides the basic data package, different data will be provided by certain types of facilities and the author names them “extra data”. For example, healthcare facilities (as shown in Figure 7.2) can provide information pertaining to medical resource availabilities, such as the doctors’ schedules and the blood bank inventory. They can also send an outbreak alert to the smart city network if an infectious disease case is identified. Another example of extra data is that the supermarket in the smart commercial community (as shown in Figure 7.2) can provide real-time commodity information to the smart city network so that citizens can locate the commodities they need. This is particularly crucial when natural disasters, such as the hurricane, tsunami, and sandstorm, are threatening the city and citizens are hoarding necessities.

The smart buildings in the proposed architecture not only provide data to the network but also require services from it. The service requirements may vary based on the facility types but there are some common services required by all. The author names them “the

basic service package". Some examples of the basic service package involve security, emergency assistance, data connection, operation and maintenance, etc. Besides the basic service package, different services may be requested by certain types of facilities and the author names them "extra service". For example, a shopping mall may request the real-time citizen flow information from the smart city network to predict the customer flow (Figure 7.2). Similar to the BFDP, the basic service package is also evolving with time and may never be exhaustive. A preliminary list of the basic service package's contents is shown in Table 7.2 [238].

Table 7. 2 A preliminary list of the basic facility service package [238]

#	Basic Services	Description
1	Security	Protection personnel and property from damage or harm, e.g. police patrol.
2	Emergency Assistance	e.g. first-aid
3	Data Connection	e.g. Internet connection
4	Resource	e.g. electricity, water, gas
5	Operation & Maintenance	Services pertaining to maintain the facility
6	Sustainability relevant service	e.g. garbage disposal and recycle resource
7	Logistics	e.g. the accessibility of a facility

7.3 Use Cases for the Smart Built Environment

The research presented in this dissertation is a use case under the proposed framework of BIM and IoT enabled smart built environment. This facility LCCA use case has proven that with sufficient data, the LCC of facilities within a community can be predicted by machine learning models. BIM and IoT-related technologies provide the data platform for this kind of innovation.

The envisioned future smart cities, as shown in Figure 7.2, will consist of smart communities. An innovation proven effective in a community level should also be feasible in a larger scope – the city level, given sufficient data availability and connectivity. The CPS is scalable – it may be a single device, or it can consist of one or multiple cyber-physical devices that form a system, or it can be a System of Systems (SoS) that consists of multiple systems [243]. A smart city is a CPS, so is a smart building, a building automation system, and a device in the system. Therefore, a valid data acquisition framework implementable in a smart community context should also be implementable in a smart city context. In an ideal future scenario, each smart building provides the BFDP to the smart city IoT network, and as a subset, the LCCA data package (Section 3.3.1) of each building will be utilized by the smart city's computation network. Thus, the machine learning-enabled LCCA can be conducted for every new and existing building in the city, based on the entire city's historical data.

This research uses an innovative use case enabled by BIM, IoT, and machine learning, to propose the author's vision for the future Smart City, which is presented in this chapter. The author envision that buildings will be able to serve as intelligent data hubs and actuators connected in the IoT network of smart cities. With the rapid development of ICT such as Data Mining, Machine Learning, and Artificial Intelligence, the data provided by

smart buildings, such as data related to human behavior, energy and resource consumption, information exchange, will be increasingly valuable for scientific research and technology applications. As the built environment evolving towards the envisioned future “smartness”, more innovative uses cases as the one presented in this research will be developed.

CHAPTER 8 CONCLUSIONS AND RECOMMENDATIONS

This thesis presents a research project that developed a machine learning-enabled facility life-cycle cost analysis (LCCA) framework using data provided by BIM and IoT. The main premise of this research is that the AECO industry demands facility LCCA but is facing the challenges of data shortage and complexity in predictive analysis. The framework proposed in this research uses data housed in BIM and IoT (ubiquitous building systems) to solve the data insufficiency issue, and implements machine learning methods to develop LCC forecasting models.

8.1 Research Contributions

This research contributes to the body of knowledge by systematically investigating the approach of obtaining a quick estimation of facilities' LCCs by implementing machine learning on historical data. First, a literature review and a questionnaire survey were conducted to determine the independent variables affecting the facility LCC. The potential data sources were summarized, and a data integration process introduced. Then, the framework for developing machine learning models for facility LCCA was proposed. A domain ontology for machine learning-enabled LCCA (LCCA-Onto) was developed to encapsulate knowledge about LCC components and their roles in relation to sibling ontologies that conceptualize the LCCA process. After that, a series of experiments were conducted on a university campus and the proposed machine learning-enabled LCCA framework demonstrated. Finally, the author's vision of the future smart built environment was discussed.

In the AECO industry, the data island (because of the unwillingness of sharing data, conflict of interest, and the lack of connectivity between systems) is hungering the development of facility-cost related decision making. This confirms the necessity of BIM and IoT-enabled data collection and analysis that can shed light on the inefficiencies of current facility LCCA practices.

The proposed LCCA framework 1) specifies the potentially influential attributes pertaining to the whole LCC of a facility, 2) utilizes BIM and IoT, which is embedded in heterogeneous building systems, as the data sources to provide robust, automated data stream for analysis, and 3) implements multiple machine learning techniques to forecast each critical LCC component and analyze their interrelationships.

This research innovatively implements machine learning to predict the whole life-cycle cost of facilities, which has never been done before. The data housed in BIM and IoT are utilized to solve the data insufficiency issue commonly faced by facility LCC analysts. Nowadays, the main reason of data insufficiency may no longer be the inexistence of data but rather the lack of connectivity [238]. This research proves the capability of BIM and IoT in providing facility data for more advanced analysis. Moreover, the proposed framework minimizes human involvement to the greatest extent possible. People make mistakes – the more people involved in the data processing and analysis process, the more risk the analysis is exposed to human errors. In addition, some stakeholders tend to be very protective of the money-related data, which makes collecting historical data extremely difficult [24]. This research bypassed some of the artificial barriers in cost analysis and uses the data from building systems directly. The more transparent approach provides reliable insights into facility LCC patterns.

From a practitioner's perspective, by exploring the new possibility for better prediction of a facility' LCC through leveraging historical data housed in heterogeneous building systems across a continuous network of buildings, this research has a greater impact than simply studying the LCC of an individual project in the design phase. The impact involves data-based LCC inputs in future facilities thus enabling cost benchmarking and informing project developments based on owned historical data. Using existing available data to benchmark facility costs can assist decision making, and new data can be incorporated as they become available. It is an iterative knowledge accumulation of facility costs that could not only identify performance trends and operation and maintenance expense "hot spots", but also identify the best practices of facility design, construction, and operation from a cost efficiency perspective.

8.2 Research Findings

According to the literature review, survey, and the experiments, this research has three major findings:

1) Current IoT networks (embedded in building systems) in buildings and BIM already contain the data that can be used for facility LCCA. 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 descriptive attributes used for machine learning can be found in BIM and building systems such as IWMS and SMS.
2) Integrating the data from IoT networks and BIM can streamline the machine learning-based LCCA process.

3) Machine learning methods are effective in facility LCC prediction. In the experiment that involves 123 buildings, the author developed single-target regression models – multilinear regression, kNN, random forest, SVM, and multilayer perceptron – and multi-target regression models – random forest and multilayer perceptron. Among these regression models, the SVM model gives the most precise prediction of facility LCC. Based on the results of 100 experiments, the SVM model have average MAE of 27.97%, 37.08%, and 39.40% for the predictions of initial cost, utility cost, and O&M cost, respectively.

8.3 Research Limitations

The experiments conducted in this study have several limitations. Firstly, the implementation of the proposed LCCA framework was limited to develop machine learning models for the 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 building LCC analysis and prediction can be developed as soon as the relevant data (discussed in Section 3.1) become available. Currently, the author cannot test the framework with more use case scenarios, such as LCC prediction during the design phase or construction phase, because of the lack of detailed design and construction documentation (during the detailed design phase, pre-construction phase, etc.). Moreover, the models developed in the experiments can only predict the lump sum 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 can be developed based on the proposed framework. In that case, the correlations between the

descriptive attributes (model input) and the cost breakdowns (model output, such as the concrete cost, the finish cost, and the building envelope cost) can be studied and prediction models can be established to estimate each cost elements. However, the author does not have enough detailed building cost data to implement the proposed framework.

Secondly, the author does not have a benchmarking tool (the baseline) to evaluate the improvements in prediction accuracy of the developed models. To the author's knowledge, the studied university does not have a prediction tool to use during the programming phase. The university's Budget Planning and Administration Department, and Facilities Management departments have not used the historical data for building LCC predictions before. Typically, cost estimators are hired to perform the initial cost prediction, but the author does not have these data to compare with the predictions produced by the developed machine learning models. Moreover, the utility and O&M cost estimation do not have a comparison base because the estimation of these costs, if any, is typically conducted after the design is available. The stakeholders in the university do not have a viable tool to conduct LCCA of utility and O&M costs during the programming phase.

Thirdly, the analysis in the experiment did not consider the influence of technology development on utility consumption and O&M costs. Most of the buildings studied were constructed in different years, with equipment of various ages, brands, and capacities. Except for the descriptive attribute *Building Age*, the developed machine learning models did not have any other indicators that describe the potential influence of technology development on the LCC. Moreover, the author studied the buildings' total costs for the past 20 years, while some of the buildings were less than 20 years old. These buildings

were treated as if they were built 20 years ago, with the data simulated. This simplification may have a negative influence on the accuracy of LCC prediction.

In addition, there are some opportunities for further applications of the proposed LCCA framework:

Large BIM model sample testing. In this research, an approach that extracts data from the BIM model to build the database of descriptive attributes is proposed and tested on several BIM models in the experiments. However, this approach is not fully implemented in the case study because most of the studied buildings do not have a well-developed BIM model. Future studies with enough BIM models may further test the approach and establish the database solely by extracting information from BIM.

Results visualization. BIM can serve as the platform for both acquiring building data and presenting LCC knowledge. This research does not involve the study of using BIM platforms (such as the Digital Building Lab Smart City [248]) to provide data visualizations of the analysis results. Hence, the LCCA results are presented in a "one-dimensional" form – only tables with numbers in them. Future studies may develop a BIM-based presentation platform to visualize the LCCA results in a multi-dimensional fashion that is comprehensible and intuitive for stakeholders.

The impacts on decision making. In this research, validation refers to the process by which the proposed framework and the machine learning models developed based on it are proved valid. How these models influence decision makings, such as during the capital planning or programming phase, is not tested. Future studies may focus on the impacts of LCCA machine learning models on human decision makings.

8.4 Recommendations for Future Research

Chapter 7 of this dissertation discusses the visions of future BIM and IoT-enabled smart built environment. This section proposes several recommended future research directions that can advance the AECO industry towards these visions.

Identifying IoT stakeholders' common data needs from facilities. Buildings – along with cities and transit systems – comprise the platform into which ubiquitous computing and IoT systems are embedded. However, currently, buildings are isolated and lack the data architecture foundation to connect with larger IoT networks. It is critical to identifying the building data requirements in the IoT paradigm and establishing a generalized facility data architecture in order a) to make the data generated in buildings, such as sustainability-related data and human-behavior-related data, available and standardized for further scientific research; b) to enable valid data collection and transmission, thereby to establish a foundation for building-related IoT innovative applications and services; c) to improve the efficiency and effectiveness of facility management activities.

Establishing the facility data infrastructure. With the rapid development of ICT such as Data Mining, Machine Learning, and Artificial Intelligence, the data provided by smart buildings, such as data related to human behavior, energy and resource consumption, information exchange, will be increasingly valuable for scientific research and technology applications. For example, given sufficient historical fire safety inspection data and fire detection device data, the data scientists can use machine learning techniques to recognize certain patterns related to fire hazard and identify the communities that have the biggest risk. Furthermore, if the fire detection devices of all the buildings in the same community

are connected through a cyber network, when a fire occurs in one building, occupants of the other buildings can be informed in real time. The widespread applications of this kind of technology require a well-established facility data infrastructure.

Identifying and realizing use cases of the smart built environment. Proposed in this thesis, the innovative approach that uses machine learning to conduct facility LCCA is a use case of the envisioned smart built environment (Chapter 7), which involves BIM and IoT. More innovative use cases in the smart built environment paradigm are currently unknown and more studies are needed to further identify and realize them.

APPENDIX A. QUESTIONNAIRES TO IDENTIFY THE INFLUENTIAL FACTORS OF FACILITY LCC

This appendix presents the questionnaires used for identifying the influential factors of facility LCC.

A.1 The questionnaire for estimators and project managers

Dear Participant,

Thank you very much for taking part in this survey. My name is Xinghua Gao, a Ph.D. candidate at Georgia Tech. I am trying to develop a building life-cycle cost prediction model using machine learning techniques. This questionnaire will help me identify the important factors that significantly affect the total construction cost of a building. Please imagine you are working in the programming phase of a building project. No design is available yet.

This survey should not take more than 20 minutes. Please provide following information:

Name: _____

Company/affiliation: _____

Title: _____

Email: _____

Number of years working in the cost estimation/project management/building design field: _____

According to your experience, please indicate the extent to which the available options for each factor cause variations in construction cost. You may choose from 0 to 5, or Y or N, where:

0 – the factor has **no influence** on construction cost variation at all

1 – the factor has a little influence but almost **negligible**

2 – the factor has **some influence** but **not significant**

3 – the factor has influence and the influence degree **varies depending on** the specific project

4 – the factor has **significant influence** on the construction cost

5 – this is one of the **determining factor**s of the construction cost

Y – I know the factor has an influence on the construction cost, but not sure about the influence degree.

N – I'm not sure about this.

Here are the factors:

Please input 0 to 5, or Y or N in this column	**Factor**	**Description**
	Building function/type	e.g. commercial building, medical building, residential building, educational building.

	LEED	LEED certified, silver, gold, or platinum
	Structure type	e.g. concrete, steel, masonry, timber structure.
	Type of floor structure	e.g. Cast-in-place (CIP) concrete, precast concrete.
	Building floor area (BFA)/built-up area	Gross floor area (GFA), gross internal area (GIA), usable floor area (UFA), etc.
	Number of floors	Including floor number aboveground and/or underground.
	Floor height	Average floor to ceiling height.
	Total height	Total building height.
	External wall area	External wall area is the difference between the external and internal gross areas.
	Internal perimeter length	The perimeter of building measured on the internal face of the enclosing structural walls.
	Footprint area	The gross area of the ground floor.
	Gross building volume (GRV)	The total volume of all interior spaces in a building over the gross floor area.
	The location of the core of the building	The location of the vertical circulation system including stairs, elevators, and the service ducts. It can be at the sides or in the middle.
	Fully enclosed covered area	The sum of all fully enclosed and covered building areas at all floor levels, such as

		basements, garages, floored roof spaces and attics.
	Unenclosed covered area	The sum of all unenclosed areas at all building floor levels, including roofed balconies, open verandahs, porches and porticos, etc.
	Number of rooms	e.g. apartment units, classrooms, households.
	Space distribution (percentage of usage)	e.g. 30% classroom, 50% laboratory, 20% office.
	Roof type	e.g. gable roof, flat roof, hip roof.
	Façade type	e.g. Exterior wall.
	Surface area of exterior wall	The value of external surface area subtracting the external window area.
	Window type	e.g. fixed windows, casement windows, sliding windows.
	Proportion of opening on external walls	Area of external doors and windows divided by external wall area ×100%.
	Finishing	Finishing type/grade (luxurious, medium, simple).
	Number of people	Designed/predicted number of occupants.
	Foundation type	e.g. Foundation system (pier, wall, slab), basement system (crawlspace, full, none, walkout).

	Foundation area	
	Topography & soil condition	e.g. plan irregularity, soil condition (hard, medium, soft).
	Parking area	
	Landscape area	
	Construction duration	
	The general contractor	
	Construction site area	
	Delivery method	e.g. CM at risk, design build, design bid build.
	Construction site access	e.g. free, semi-restricted, restricted.
	Mechanical, Electrical, Plumbing (MEP) Types	The type of mechanical systems, the vendors/subcontractors
	Number of elevators	

Are there any other factors you think could significantly influence the total construction cost but are not listed above? If there are, please specify:

_____

_____

_____

_____

A.2 The questionnaire for energy experts

Dear Participant,

Thank you very much for taking part in this survey. My name is Xinghua Gao, a Ph.D. candidate at Georgia Tech. I am trying to develop a building life-cycle cost prediction model using machine learning techniques. This questionnaire will help me identify the important factors that significantly affect life-cycle **energy cost** of a building. Please imagine you are working in the programming phase of a building project. No design is available yet.

This survey should not take more than 20 minutes. Please provide following information:

Name: _____

Company/affiliation: _____

Title: _____

Email: _____

Number of years working in the energy analysis field: _____

According to your experience, please indicate the extent to which each of the factors influence the energy cost, from 0 to 5, or you may choose Y or N, where:

0 – the factor has **no influence** on the total energy cost variation at all

1 – the factor has a little influence but almost **negligible**

2 – the factor has **some influence** but **not significant**

3 – the factor has influence and the influence degree **varies depending on** the specific building

4 – the factor has **significant influence** on the energy cost

5 – this is one of the **determining factor**s of the energy cost

Y – I know the factor has an influence on the energy cost, but not sure about the influence degree.

N – I'm not sure about this.

Here are the factors:

Please input 0 to 5, or Y or N in this column	**Attribute Name**	**Description**
	Building age	
	Building function/type	e.g. commercial building, medical building, residential building, educational building.
	LEED	LEED certified, silver, gold, or platinum
	Building floor area (BFA)/ built-up area	Gross floor area (GFA), gross internal area (GIA), usable floor area (UFA), etc.
	Number of floors	Including floor number aboveground and/or underground
	Floor height	Average floor to ceiling height

	Conditioned floor area (CFA)	The total floor area of enclosed conditioned space on all floors of a building.
	Gross building volume (GRV)	The total volume of all interior spaces in a building over the gross floor area.
	Space distribution (percentage of usage)	e.g. 30% classroom, 50% laboratory, 20% office.
	Room number	
	Number of computers and televisions	e.g. desktop PC, laptops, TVs
	Number of printers/ photocopiers	
	Number of kitchen electrical products	e.g. oven, microwave
	Occupants' average time spent in the building	None, medium, long
	Electric vehicle number	
	Occupancy percentage	The proportion of rooms occupied to the number of rooms available for the selected date or period.
	Reduce energy cost willingness	Yes/No

	Number of regular occupants	e.g. employees, students
	Total hours open per week	
	Heating percentage	The percentage of the total floor space within the facility that is served by mechanical heating equipment.
	Cooling percentage	The percentage of the total floor space within the facility that is served by mechanical cooling equipment.
	Programmable thermostat	Have programmable thermostat? Yes/No
	Apparent temperature	The temperature equivalent perceived by humans, caused by the combined effects of air temperature, relative humidity and wind speed.
	Photovoltaic (PV) system	Have PV system? Yes/No
	Percent lit when open	The percentage of lit area to total building area when open
	Percent lit off hours	The percentage of lit area to total building area when close

	Heating degree day (HDD)	The number of degrees that a day's average temperature is below the degree which buildings need to be heated.
	Cooling degree day (CDD)	The number of degrees that a day's average temperature is above the degree which buildings need to be cooled.
	Temperature	The temperature at the building location throughout the year
	Dew point	The temperature to which air must be cooled to become saturated with water vapor.
	Humidity	The amount of water vapor present in air.
	Daily average sky cover (Cloud cover)	The daily average fraction of the sky obscured by opaque clouds when observed from the building.
	Sprinkle head number	
	Total window area footage	
	Window types (e.g., operable window)	
	Lighting sensors (e.g., occupancy, daylight)	

Are there any other factors you think could significantly influence on the total energy cost but are not listed above? If there are, please specify:

_____

_____

_____

_____

A.3 The questionnaire for facility managers (operation and maintenance)

Thank you very much for taking part in this survey. My name is Xinghua Gao, a Ph.D. candidate at Georgia Tech. I am trying to develop a building life-cycle cost prediction model using machine learning techniques. This questionnaire will help me identify the important factors that significantly affect the life-cycle **operation and maintenance (O&M) cost** of a building. **Please do NOT consider utility cost (electricity, water, gas, etc.) in this survey**. Please imagine you are working in the programming phase of a building project. No design is available yet. You are asked to provide expert opinion on what factors will influence the overall life-cycle O&M cost of the new building.

This survey should not take more than 20 minutes. Please provide following information:

Name: _____

Company/affiliation: _____

Title: _____

Email: _____

Number of years working in the facilities management field: _____

According to your experience, please indicate the extent to which each of the factors influence the construction cost, from 0 to 5, or you may choose Y or N, where:

0 – the factor has **no influence** on the total O&M cost at all

1 – the factor has a little influence but almost **negligible**

2 – the factor has **some influence** but **not significant**

3 – the factor has influence and the influence degree **varies depending on** the specific project

4 – the factor has **significant influence** on the O&M cost

5 – this is one of the **determining factors** of the O&M cost

Y – I know the factor has an influence on the O&M cost, but not sure about the influence degree.

N – I'm not sure about this.

Here are the factors:

Please input 0 to 5, or Y or N in this column	**Factor**	**Description**
	Building age	
	Average number of staffs	The number of O&M staffs working on the building.

	Number of service types the building needs	e.g. a) many building services (skill trades) are needed, b) regular number needed, or c) only a few needed.
	The building's level of complexity	e.g. a) a complicated building like Klaus and Clough, b) a regular building like Mason, or c) a simple building like Caddell.
	Number of complicated equipment	Expensive, heavy, sophisticated mechanical, electrical, or plumbing equipment.
	Level of waste service needed	e.g. a) only need regular waste service, b) need solid waste division service, or c) no waste service needed.
	Level of building data availability	e.g. a) well-metered with many utility meters, lighting sensors, and occupancy sensors, b) have some meters, or c) do not have any meter at all.
	Number of shifts	Shifts of operation and maintenance team
	Building function/type	e.g. library, healthcare building, residential building, laboratory building.
	Building floor area (BFA)/built-up area	Gross floor area (GFA), gross internal area (GIA), usable floor area (UFA), etc.
	Number of floors	Including floor number aboveground and/or underground.
	LEED	LEED certified, silver, gold, or platinum
	Structural type	e.g. concrete, steel, masonry, timber structure.

	Percentage of usage	e.g. 30% classroom, 50% laboratory, 20% office.
	Number of rooms	e.g. apartment units, classrooms, households.
	Number of people	Designed/predicted number of occupants.
	Budget constraint	The budget constraint of operation and maintenance
	Building geometry in general	Footprint area, building volume, etc.
	Total height	Total building height
	Floor height	Average floor to ceiling height
	Number of elevators	
	Classroom area	
	Teachers' cabinets area	
	Hallways area	
	Sanitary area	
	Office area	
	Library area	
	Laboratory area	

Are there any other factors you think could significantly influence on the total O&M cost but are not listed above? If there are, please specify:

_____

_____

_____

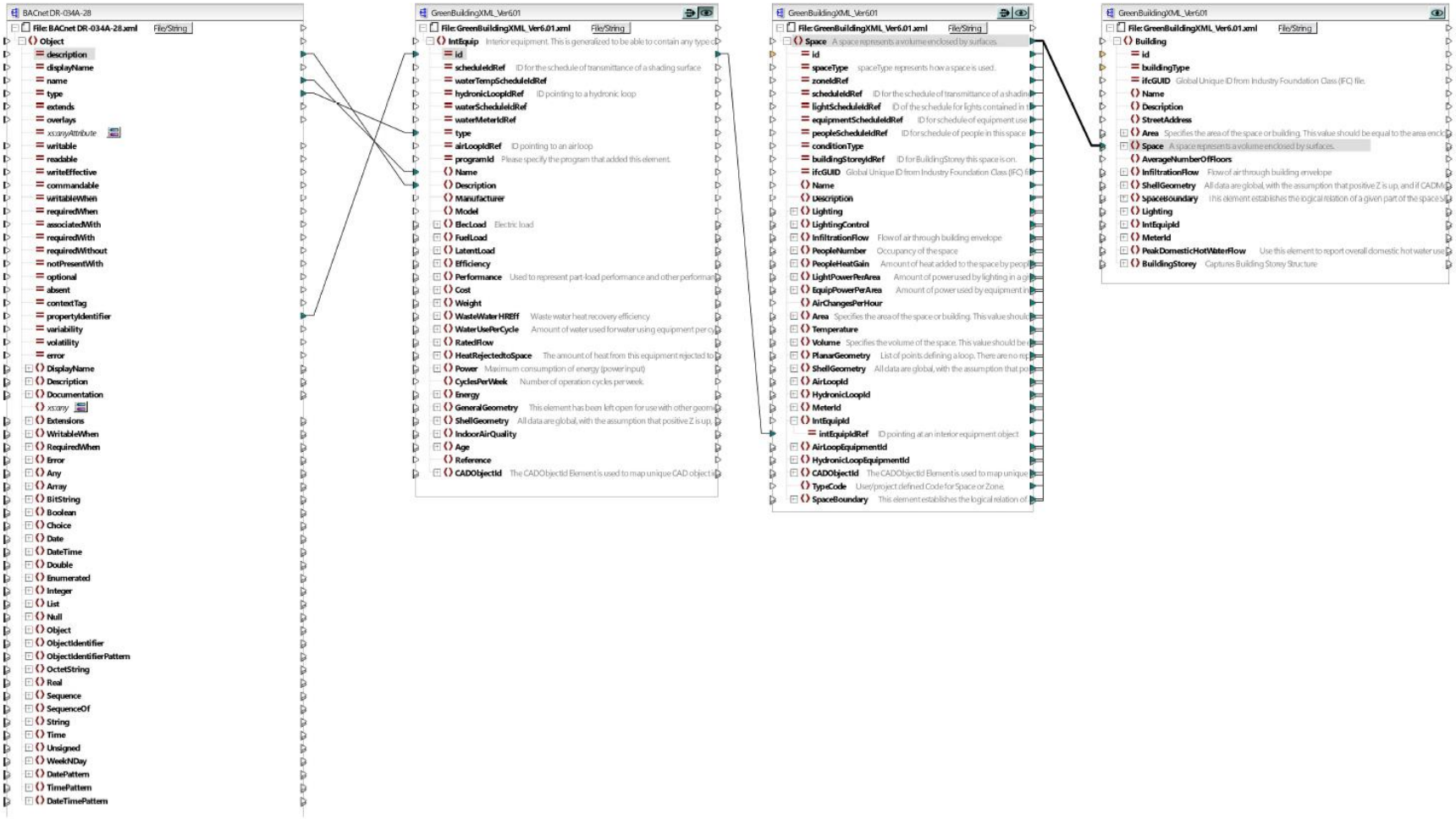
APPENDIX B. DATA MAPPING BETWEEN BUILDING DATA STANDARDS

This appendix presents the data mapping between data standards: BACnet XML and ifcXML (B1), BACnet XML and gbXML (B2), ifcXML and CityGML XML (B3).

B.1 Overlaps between BACnet XML and ifcXML


BACnet DR-034A-28
File BACnet DR-034A-28.xml File/String
CSML
defaultLocale
xs:anyAttribute
Definitions
Any
Array
BitString
Boolean
Choice
Date
DateTime
Double
Enumerated
Integer
List
Null
Object
description
displayName
name
type
extends
overlays
xs:anyAttribute
writable
readable
writeEffective
commandable
writableWhen
requiredWhen
associatedWith
requiredWith
requiredWithout
notPresentWith
optional
absent
contextTag
propertyIdentifier
variability
volatility
error
DisplayName
Description
Documentation
xs:any
Extensions
WritableWhen
RequiredWhen
Error
Any
Array
BitString
Boolean
Choice
Date
DateTime
Double
Enumerated
Integer
List
Null
Object
ObjectIdentifier
ObjectIdentifierPattern
OctetString
IFC4_ADD2
File IFC_Controller.xml; IFC_space.xml File/St
ifc:IfcController
href
ref
id
path
pos
GlobalId
Name
Description
ObjectType
Tag
PredefinedType
ifcOwnerHistory
ifcIsNestedBy
ifcIsDecomposedBy
ifcIsDeclaredBy
ifcIsTypedBy
ifcIsDefinedBy
ifcObjectPlacement TYPE
href
ref
id
path
pos
ifcRepresentation TYPE
ifcHasProjections
ifcHasOpenings
IFC4_ADD2
File IFC_space.xml; IFC_building.xml File/String
ifc:IfcSpace
href
ref
id
path
pos
GlobalId
Name
Description
ObjectType
LongName
CompositionType
PredefinedType
ElevationWithFlooring
ifcOwnerHistory
ifcIsNestedBy
ifcIsDecomposedBy
ifcIsDeclaredBy
ifcIsTypedBy
ifcIsDefinedBy
ifcObjectPlacement TYPE
href
ref
id
path
pos
ifcRepresentation TYPE
ifcContainsElements
ifcReferencesElements
IFC4_ADD2
File IFC_building.xml; IFC4_ADD2.xml Fil
ifc:IfcBuilding
href
ref
id
path
pos
GlobalId
Name
Description
ObjectType
LongName
CompositionType
ElevationOfRefHeight
ElevationOfTerrain
ifcOwnerHistory
ifcIsNestedBy
ifcIsDecomposedBy
ifcIsDeclaredBy
ifcIsTypedBy
ifcIsDefinedBy
ifcObjectPlacement TYPE
ifcRepresentation TYPE
ifcContainsElements
ifcReferencesElements
ifcBuildingAddress
IFC4_ADD2
File IFC4_ADD2.xml; IFC_space2.xml File/St
ifc:IfcSensor
href
ref
id
path
pos
GlobalId
Name
Description
ObjectType
Tag
PredefinedType
ifcOwnerHistory
ifcIsNestedBy
ifcIsDecomposedBy
ifcIsDeclaredBy
ifcIsTypedBy
ifcIsDefinedBy
ifcObjectPlacement TYPE
href
ref
id
path
IFC4_ADD2
File IFC_space2.xml; IFC_building2.xml File/Str
ifc:IfcSpace
href
ref
id
path
pos
GlobalId
Name
Description
ObjectType
LongName
CompositionType
PredefinedType
ElevationWithFlooring
ifcOwnerHistory
ifcIsNestedBy
ifcIsDecomposedBy
ifcIsDeclaredBy
ifcIsTypedBy
ifcIsDefinedBy
ifcObjectPlacement TYPE
href
ref
id
path
pos
ifcRepresentation TYPE
ifcContainsElements
ifcReferencesElements
IFC4_ADD2
File IFC_building2.xml; IFC4_ADD2.xml
ifc:IfcBuilding
href
ref
id
path
pos
GlobalId
Name
Description
ObjectType
LongName
CompositionType
ElevationOfRefHeight
ElevationOfTerrain
ifcOwnerHistory
ifcIsNestedBy
ifcIsDecomposedBy
ifcIsDeclaredBy
ifcIsTypedBy
ifcIsDefinedBy
ifcObjectPlacement TYPE
ifcRepresentation TYPE
ifcContainsElements
ifcReferencesElements
ifcBuildingAddress

B.2 Overlaps between BACnet XML and gbXML



B.3 Overlaps between ifcXML and CityGML XML

APPENDIX C. AN EXAMPLE OF THE ELECTRICITY CONSUMPTION RAW DATA

TimestampUTC2	TimestampUTC	Active Energy Delivered-Received
2012-10-01 04:00:00-05:00 EST	10/1/2012 4:00	769564.1875
2012-10-01 04:15:00-05:00 EST	10/1/2012 4:15	769571.1875
2012-10-01 04:30:00-05:00 EST	10/1/2012 4:30	769577.5625
2012-10-01 04:45:00-05:00 EST	10/1/2012 4:45	769585
2012-10-01 05:00:00-05:00 EST	10/1/2012 5:00	769591.3125
2012-10-01 05:15:00-05:00 EST	10/1/2012 5:15	769597.5625
2012-10-01 05:30:00-05:00 EST	10/1/2012 5:30	769604
2012-10-01 05:45:00-05:00 EST	10/1/2012 5:45	769610.1875
2012-10-01 06:00:00-05:00 EST	10/1/2012 6:00	769616.3125
2012-10-01 06:15:00-05:00 EST	10/1/2012 6:15	769622.4375
2012-10-01 06:30:00-05:00 EST	10/1/2012 6:30	769628.5
2012-10-01 06:45:00-05:00 EST	10/1/2012 6:45	769634.5
2012-10-01 07:00:00-05:00 EST	10/1/2012 7:00	769640.5
2012-10-01 07:15:00-05:00 EST	10/1/2012 7:15	769646.4375
2012-10-01 07:30:00-05:00 EST	10/1/2012 7:30	769652.375
2012-10-01 07:45:00-05:00 EST	10/1/2012 7:45	769658.375
2012-10-01 08:00:00-05:00 EST	10/1/2012 8:00	769664.25
2012-10-01 08:15:00-05:00 EST	10/1/2012 8:15	769670.125
2012-10-01 08:30:00-05:00 EST	10/1/2012 8:30	769677
2012-10-01 08:45:00-05:00 EST	10/1/2012 8:45	769682.5625

2012-10-01 09:00:00-05:00 EST	10/1/2012 9:00	769688.5
2012-10-01 09:15:00-05:00 EST	10/1/2012 9:15	769694.4375
2012-10-01 09:30:00-05:00 EST	10/1/2012 9:30	769700.375
2012-10-01 09:45:00-05:00 EST	10/1/2012 9:45	769706.375
2012-10-01 10:00:00-05:00 EST	10/1/2012 10:00	769712.3125
2012-10-01 10:15:00-05:00 EST	10/1/2012 10:15	769718.3125
2012-10-01 10:30:00-05:00 EST	10/1/2012 10:30	769724.25
2012-10-01 10:45:00-05:00 EST	10/1/2012 10:45	769730.1875
2012-10-01 11:00:00-05:00 EST	10/1/2012 11:00	769736.1875

APPENDIX D. A SMALL PORTION OF THE RAW DATA OF O&M WORK ORDER RECORDS

Columns 1 to 10

Work Order	Description	Created By	Date Created	Status	Region	Facility	Property	Project	Problem Code
42825-2010	CONSTRUCTION	TCORSO	15:45.0	90-COMP	GT-MAIN	NAA SOUTH PKNG	190	143-2010	
37767-2010	AREA 5 SMART WEEKLY	GLOCKERMAN	00:32.0	40	GT-MAIN	OKEEFE GYM	033A		
37772-2010	AREA 5 AIR COMPRESSOR WEEKLY	GLOCKERMAN	00:33.0	40	GT-MAIN	FOOD PROCESSING	159		
32931-2010	U&E-SEWER / STORM DRAINS WEEK-40	KCHAREPOO	52:50.0	40	GT-MAIN	GENERAL CAMPUS	GC		

72797-2010	364 REPLACE PARKING BRAKE CABLE	GBYRD	20:35.0	40	GT-MAIN	KING GARAGE WAR	67	08C
39146-2010	KING- EXHAUST FAN (SEMI-Y)	BHALABI	48:42.0	40	GT-MAIN	KING GARAGE WAR	67	
38295-2010	MS&E-WATER LEAK IN ROOM G272	GPRATER	22:32.0	40	GT-MAIN	MS&E	167	
45463-2010	PETTIT AIR COMP	OAUZLA	26:21.0	40	GT-MAIN	PETTIT	95	
52269-2010	MSE- CHECK AND CLEAN AHU DRAINS (M)	CSHEFFIELD	43:42.0	40	GT-MAIN	MS&E	167	

Columns 11 to 20

Type	Category	Organization	Requestor	Contact	Contact Phone	Contact Email	Budget	Desired Date	Customer Request	Reference	Shop
C	MAJOR	A-58-400	CONTRACT ADMINISTRATION				0				ACCOUNTING
R	ST	P-94-120	AREA 5				0				AREA 5 EAST
R	PM	P-94-120	AREA 5				0				AREA 5 EAST
R	PM	P-94-120	INFRA				0				U&E
S	SR	P-94-120	AREA 3				0				CENTRAL - MOTOR
R	PM	P-94-120	AREA 2				0				AREA 2 NORTH
S	SR	G-29-300	MOLECULAR DESIGN INSTITUT				0				AREA 2 NORTH

R	PM	B-12-401	MICROELECTRONIC RESEARCH	SCOTT FOWLER	0	AREA 3 CENTRAL
R	PM	P-94-120	FACILITIES		0	AREA 2 NORTH

Columns 21 to 30

Reference	Shop	Shop Person	Extra Description	Editor	Edit Date	Estimated Labor	Estimated Material	Estimated Equipment	Estimated Contract
	ACCOUNTING			QCARSWELL	05:06.0	0	0	0	2707.25
	AREA 5 EAST			GLOCKERMAN	15:16.0	0	0	0	0
	AREA 5 EAST			SROSA	27:47.0	0	0	0	0
	U&E	EST007		OREEVES	37:00.0	0	0	0	0
	CENTRAL - MOTOR	MTR008		GBYRD	14:31.0	0	0	0	0

			46:28.				
AREA 2 NORTH		GPRATER	0	0	0	0	0
			33:54.				
AREA 2 NORTH		GPRATER	0	0	0	0	0
AREA 3	ACM00		01:51.				
CENTRAL	1	CSHEFFIELD	0	0	0	0	0
			17:55.				
AREA 2 NORTH		ACLIFFORD	0	0	0	0	0

Columns 31 to 40

Estimated Total	Estimated Hours	Encumbered Labor	Encumbered Material	Encumbered Equipment	Encumbered Contract	Encumbered Total	Actual Labor	Actual Material	Actual Equipment
2707.25	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	67.92	0	0
0	1	0	0	0	0	0	39.16	0	0
0	3	0	0	0	0	0	19.58	0	0
0	1	0	0	0	0	0	2399.82	21.69	0
0	1.5	0	0	0	0	0	58.74	0	0
0	0	0	0	0	0	0	39.16	0	0
0	4.64	0	0	0	0	0	156.64	0	0
0	2	0	0	0	0	0	39.16	0	0

Columns 41 to 48

Actual Contract	Actual Total	Actual Hours	Billed Labor	Billed Material	Billed Equipment	Billed Contract	Billed Total
27642.38	27642.38	0	0	0	0	0	0
0	67.92	3	67.92	0	0	0	67.92
0	39.16	1	39.16	0	0	0	39.16
0	19.58	0.5	19.58	0	0	0	19.58
0	2421.51	1	2399.82	21.69	0	0	2421.51
0	58.74	1.5	58.74	0	0	0	58.74
0	39.16	1	39.16	0	0	0	39.16
0	156.64	4	156.64	0	0	0	156.64
0	39.16	1	39.16	0	0	0	39.16

APPENDIX E. THE MATLAB CODE FOR CLEANING THE UTILITY CONSUMPTION DATA

E.1 Utility consumption data cleaning and weekly consumption calculation

```
% This is a MATLAB program for cleaning the utility consumption data
% Developed by Xinghua Gao @ Georgia Tech
% Email: gaoxh@gatech.edu
% March 2019

% Function:
% Import all utility consumption (csv) files and process them
% Generate the weekly energy consumption matrix of the buildings
% The output of this program is the matrix "tar_merged" that records all the
% utility consumption and the "header" records the facility name, sensor ID
% and the start time.
% the results are wrote to a csv file named "results"

% IMPORTANT: there must be one or more building data in the folder to
% anlayze


clc;
clear;
```

```
%%%%%%%%%% Parameter Settings %%%%%%%%%%%%%
col_num = 300; % the column number of the target matrix
row_num = 210000; % the row number of the target matrix
start_path = fullfile('C:\test');% Define a starting folder.
    % IMPORTANT:theoutput folder (current folder of MATLAB)
    % must be the same folder.
path = "C:\test\1\";
max_week = 305;
showplot = 0; % =1 to show the plot of each building, =0 not to
plot = 0; % =1 to generate plot image files, =0 not to
%outlier1 = 40000; % set he outlier threshold
%%%%%%%%%% Settings End %%%%%%%%%%%%%%%%%%%%

% the target matrix
raw = zeros (row_num,col_num);
header = cell(4,col_num); % the header of target matrix

% Import the csv files in each folder
workspace;  % Make sure the workspace panel is showing.
format longg;
format compact;
```

```
% Ask user to confirm or change.
topLevelFolder = uigetdir(start_path);
if topLevelFolder == 0
    return;
end

% Get list of all subfolders.
allSubFolders = genpath(topLevelFolder);
% Parse into a cell array.
remain = allSubFolders;
listOfFolderNames = {};
while true
    [singleSubFolder, remain] = strtok(remain, ';');
    if isempty(singleSubFolder)
        break;
    end
    listOfFolderNames = [listOfFolderNames singleSubFolder];
end
numberOfFolders = length(listOfFolderNames)

% Process all csv files in those folders.

column = 1; % The columns in the comprehensive table
```

```
for k = 1 : numberOfFolders
    % Get this folder and print it out.
    thisFolder = listOfFolderNames{k};
    fprintf('Processing folder %s\n', thisFolder);

    filePattern = sprintf('%s/*.csv', thisFolder);
       baseFileNames = dir(filePattern);
    numberOfImageFiles = length(baseFileNames);

    % A list of all files in this folder.
    if numberOfImageFiles >= 1
          % Go through all those image files.
          for f = 1 : numberOfImageFiles
              fullFileName = fullfile(thisFolder, baseFileNames(f).name);
              fprintf('     Processing csv file %s\n', fullFileName);

              folder_name = extractAfter(thisFolder,path);
              folder_name_2 = folder_name + "\"; % for the format of file_name
              file_name = extractAfter(fullFileName,folder_name_2);

              % check if the csv file is energy consumption
              % 'Active Energy Delivered-Received'indicates that the
```

```
% file records energy consumption data

fileID = fopen(fullFileName,'r');
%col_head = cell(1,1);
col_head = textscan(fileID,'%*q %*q %q ',1,'Delimiter',',');
start_time = textscan(fileID,'%*q %q %*q ',1,'Delimiter',',');
fclose(fileID);

% if the csv file is about energy consumption then record
if isequal(col_head[241]{1,1},'Active Energy Delivered-Received')

   % write to the header table
   % the first row is the folder name (facility name)
   header{1,column} = folder_name;

   % the second row is the sensor ID
   header{2,column} = file_name;

   % the start time
   header{3,column} = start_time[241]{1,1};

   % write to the table
   temp = csvread(fullFileName,1,2); % skip 1 row, 2 columns
```

```
            % the size difference between temp and raw(:,column)
            [mm,nn] = size (temp);
            diff = row_num - mm;

            num_weeks = fix(mm/672);

            % the number of weeks
            header{4,column} = num_weeks;

            % Pad the temp matrix to make it the same size as the
            % column size of the target matrix
            raw(:,column) = padarray(temp,diff,0,'post');

            column = column +1;

        end
    end
    else
    fprintf('    Folder %s has no csv files in it.\n', thisFolder);
  end
end
```

```
% Transfer the 15min data to weekly data
% tar_total is the total consupmtion matrix (weekly record)
tar_total = zeros(fix(row_num/672),col_num);
for j = 1 : column - 1
   for i = 1 : header{4,j}-1
      temp_week = i*672;
      tar_total(i,j)= raw(temp_week,j);
   end
end

disp ('Total consumption matrix created');

% tar is the weekly consumption matrix
tar = zeros(size(tar_total));
for j = 1 : column - 1

%    % define outlier
%    outlier(j) = (max(tar_total(:,j)) - min(tar_total(:,j)))/200;
%    if outlier(j) <= 0 || outlier(j) >= outlier1;
%       outlier(j) = outlier1;
%    end

   for i = 2 : header{4,j}-1
```

```
        tar(i-1,j) = tar_total(i,j) - tar_total(i-1,j);


      % remove outliers
      if i >=3
         if  abs(tar(i-1,j)- tar(i-2,j)) >= abs(tar(i-2,j)*2)
             tar(i-1,j) =  tar(i-2,j);
          elseif  tar (i-1,j) <= 0 % if negative value shows, use the last week data
             tar (i-1,j) = tar(i-2,j);
          end
      end
    end
end


% remove outliers
% for j = 1 : column - 1
%     % set the outlier, which is the average of the weekly consumption * 3;
%     outlier(j) = mean(tar(:,j))*3;
%     %outlier(j) = (max(tar_total(:,j)) - min(tar_total(:,j)))/300;
%
%     if outlier(j) <= 0 || outlier(j) >= 25000;
%         outlier(j) = outlier1;
%     end
```

```
%
%     % remove outliers
%     for i = 2 : header{4,j}-1
%         if  tar(i,j)- tar(i-1,j)>= outlier(j)
%             tar (i,j)=  tar(i-1,j);
%         elseif  tar (i-1,j)<= 0 % if negative value shows, use the last week data
%             tar (i-1,j)= tar(i-2,j);
%         end
%     end
%
% end

disp ('Weekly consumption matrix created');

% merge the sensor data in the same building
tar_merged = tar;
header_merged = header;
x =1;
for j = 2:column -1
   if isequal(header{1,j},header{1,j-1}) && isequal(header{3,j},header{3,j-1})
       a(x) = j;
       x = x+1;
   end
```

```
end

[ma,na] = size (a);

for k = na:-1:1
    tar_merged(:,a(k)-1) = tar_merged(:,a(k)-1) + tar_merged(:,a(k));
    tar_merged(:,a(k)) = [];
    header_merged(:,a(k)) = [];
end

% trim the merged matrix
% max_week = max(cell2mat(header_merged(4,:)));
% tar_merged(max_week+1:end,:) = [];
tar_merged(max_week+1:end,:) = [];

[mt,nt] = size(tar_merged);
[mmt,nnt] = size(tar);
new_column = column - (nnt-nt) - 1;
tar_merged(:,new_column + 1:end)=[];

disp ('Weekly consumption matrix merged');

% write the results to a csv file
```

```
% fid = fopen('results.csv', 'w') ;
% for k=1:3
%     for j = 1: col_num
%         fprintf(fid, '"%s", ',header{k,j}) ;
%     end
%     fprintf(fid, '\n');
% end
%
% dlmwrite('results.csv',raw,'delimiter',',','-append');
%
% fclose(fid) ;


% plot
if plot ==1
   for i = 1: new_column
      if showplot == 0
         h = figure('visible','off');
      else
         h = figure('visible','on');
      end

      %NZ_tar_merged=(~tar_merged(:,i)==0);
      %plot(NZ_tar_merged);
```

```
        plot_tar_merged  = tar_merged(:,i);
        plot_tar_merged (plot_tar_merged==0) = nan;
        plot(plot_tar_merged);
        title([header_merged{1,i},' weekly electricity consumption']);
        grid on;
        saveas(h,sprintf('%s.png',header_merged{1,i}));
    end

end

disp ('Calculation ends')
```

E.2 CSV file generation

```
% This script is used for generate the csv file.
% Run the "one_utility_clean.m" first.

fid = fopen('results.csv', 'w') ;
for k=1:3
    for j = 1: new_column
        fprintf(fid, '"%s", ',header_merged{k,j}) ;
    end
    fprintf(fid, '\n');
end
```

```
dlmwrite('results.csv',tar_merged,'delimiter',',','-append');

fclose(fid) ;

disp ('Results.csv generated')
```

E.3 Monthly consumption calculation and CSV file generation

```
% This script is used for generate the monthly consumption csv file.
% Run the "one_utility_clean.m" first.

tar_month = zeros(ceil(mt/4),new_column); % assume 52 weeks per year
k = 1;
for j = 1 : new_column
   for i = 1:mt
      tar_month(k, j) = tar_month(k, j) + tar_merged(i,j);

      if rem(i,4)==0
         k = k+1;
      end
   end
   k = 1;
end
```

```
disp ('Monthly consumption calculated')


fid = fopen('results_monthly.csv', 'w') ;
for k=1:3
   for j = 1: new_column
      fprintf(fid, '"%s", ',header_merged{k,j}) ;
   end
   fprintf(fid, '\n');
end

dlmwrite('results_monthly.csv',tar_month,'delimiter',',','-append');


fclose(fid) ;


disp ('Results.csv generated')
```

E.4 Annual consumption calculation and CSV file generation

```
% This script is used for generate the annual consumption csv file.
% Run the "one_utility_clean.m" first.


tar_year = zeros(ceil(mt/52),new_column); % assume 52 weeks per year
k = 1;
```

```
for j = 1 : new_column
    for i = 1:mt
        tar_year(k, j) = tar_year(k, j) + tar_merged(i,j);

        if rem(i,52)==0
            k = k+1;
        end
    end
    k = 1;
end

disp ('Annually consumption calculated')

fid = fopen('results_annually.csv', 'w') ;
for k=1:3
    for j = 1: new_column
        fprintf(fid, '"%s", ',header_merged{k,j}) ;
    end
    fprintf(fid, '\n');
end

dlmwrite('results_annually.csv',tar_year,'delimiter',',','-append');
```

```
fclose(fid) ;


disp ('Results.csv generated')
```

APPENDIX F. THE OPENREFINE OPERATION HISTORY FOR CLEANING THE O&M WORK ORDERS

The OpenRefine operation.JSON:


```
[
  {
    "op": "core/column-split",
    "description": "Split column Work Order by separator",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Work Order",
    "guessCellType": true,
    "removeOriginalColumn": true,
    "mode": "separator",
    "separator": "-",
    "regex": false,
    "maxColumns": 2
  },
  {
    "op": "core/column-rename",
```

```
  "description": "Rename column Work Order 1 to Work Order Number",
  "oldColumnName": "Work Order 1",
  "newColumnName": "Work Order Number"
},
{
  "op": "core/column-rename",
  "description": "Rename column Work Order 2 to Year",
  "oldColumnName": "Work Order 2",
  "newColumnName": "Year"
},
{
  "op": "core/column-removal",
  "description": "Remove column Date Created",
  "columnName": "Date Created"
},
{
  "op": "core/column-removal",
  "description": "Remove column Status",
  "columnName": "Status"
},
{
  "op": "core/column-removal",
  "description": "Remove column Region",
```

```
  "columnName": "Region"
},
{
  "op": "core/column-removal",
  "description": "Remove column Project",
  "columnName": "Project"
},
{
  "op": "core/column-removal",
  "description": "Remove column Problem Code",
  "columnName": "Problem Code"
},
{
  "op": "core/column-removal",
  "description": "Remove column Contact",
  "columnName": "Contact"
},
{
  "op": "core/column-removal",
  "description": "Remove column Contact Phone",
  "columnName": "Contact Phone"
},
{
```

```
  "op": "core/column-removal",
  "description": "Remove column Contact Email",
  "columnName": "Contact Email"
},
{
  "op": "core/column-removal",
  "description": "Remove column Budget",
  "columnName": "Budget"
},
{
  "op": "core/column-removal",
  "description": "Remove column Desired Date",
  "columnName": "Desired Date"
},
{
  "op": "core/column-removal",
  "description": "Remove column Customer Request",
  "columnName": "Customer Request"
},
{
  "op": "core/column-removal",
  "description": "Remove column Reference",
  "columnName": "Reference"
```

```
},
{
  "op": "core/column-removal",
  "description": "Remove column Edit Date",
  "columnName": "Edit Date"
},
{
  "op": "core/text-transform",
  "description": "Text transform on cells in column Estimated Labor using expression value.toNumber()",
  "engineConfig": {
    "mode": "row-based",
    "facets": []
  },
  "columnName": "Estimated Labor",
  "expression": "value.toNumber()",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10
},
{
  "op": "core/text-transform",
```

"description": "Text transform on cells in column Estimated Material using expression value.toNumber()",
"engineConfig": {
"mode": "row-based",
"facets": []
},
"columnName": "Estimated Material",
"expression": "value.toNumber()",
"onError": "keep-original",
"repeat": false,
"repeatCount": 10
},
{
"op": "core/text-transform",
"description": "Text transform on cells in column Estimated Equipment using expression value.toNumber()",
"engineConfig": {
"mode": "row-based",
"facets": []
},
"columnName": "Estimated Equipment",
"expression": "value.toNumber()",
"onError": "keep-original",

```
    "repeat": false,
    "repeatCount": 10
   },
   {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Estimated Contract using expression value.toNumber()",
    "engineConfig": {
     "mode": "row-based",
     "facets": []
    },
    "columnName": "Estimated Contract",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
   },
   {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Estimated Total using expression value.toNumber()",
    "engineConfig": {
     "mode": "row-based",
```

```
      "facets": []
    },
    "columnName": "Estimated Total",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Estimated Hours using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Estimated Hours",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
```

```
    "op": "core/text-transform",
    "description": "Text transform on cells in column Encumbered Labor using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Encumbered Labor",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Encumbered Material using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Encumbered Material",
    "expression": "value.toNumber()",
```

```
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Encumbered Equipment using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Encumbered Equipment",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Encumbered Contract using expression value.toNumber()",
    "engineConfig": {
```

```
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Encumbered Contract",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Encumbered Total using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Encumbered Total",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
```

```
    {
      "op": "core/text-transform",
      "description": "Text transform on cells in column Actual Labor using expression value.toNumber()",
      "engineConfig": {
        "mode": "row-based",
        "facets": []
      },
      "columnName": "Actual Labor",
      "expression": "value.toNumber()",
      "onError": "keep-original",
      "repeat": false,
      "repeatCount": 10
    },
    {
      "op": "core/text-transform",
      "description": "Text transform on cells in column Actual Material using expression value.toNumber()",
      "engineConfig": {
        "mode": "row-based",
        "facets": []
      },
      "columnName": "Actual Material",
```

```
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Actual Equipment using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Actual Equipment",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Actual Contract using expression value.toNumber()",
```

```
      "engineConfig": {
        "mode": "row-based",
        "facets": []
      },
      "columnName": "Actual Contract",
      "expression": "value.toNumber()",
      "onError": "keep-original",
      "repeat": false,
      "repeatCount": 10
    },
    {
      "op": "core/text-transform",
      "description": "Text transform on cells in column Actual Total using expression value.toNumber()",
      "engineConfig": {
        "mode": "row-based",
        "facets": []
      },
      "columnName": "Actual Total",
      "expression": "value.toNumber()",
      "onError": "keep-original",
      "repeat": false,
      "repeatCount": 10
```

```
    },
    {
      "op": "core/text-transform",
      "description": "Text transform on cells in column Actual Hours using expression value.toNumber()",
      "engineConfig": {
        "mode": "row-based",
        "facets": []
      },
      "columnName": "Actual Hours",
      "expression": "value.toNumber()",
      "onError": "keep-original",
      "repeat": false,
      "repeatCount": 10
    },
    {
      "op": "core/text-transform",
      "description": "Text transform on cells in column Billed Labor using expression value.toNumber()",
      "engineConfig": {
        "mode": "row-based",
        "facets": []
      },
```

```
    "columnName": "Billed Labor",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Billed Material using expression
value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Billed Material",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
```

```
    "description": "Text transform on cells in column Billed Equipment using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Billed Equipment",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Billed Contract using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Billed Contract",
    "expression": "value.toNumber()",
    "onError": "keep-original",
```

```
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Billed Total using expression value.toNumber()",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Billed Total",
    "expression": "value.toNumber()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/column-addition",
    "description": "Create column Estimated Cost at index 36 based on column Billed Total using expression grel:\"Labor \" + cells[\"Estimated Labor\"].value + \";\" + \" Material \" + cells[\"Estimated Material\"].value + \";\" + \" Equipment \" + cells[\"Estimated Equipment\"].value + \";\" + \" Contract \" + cells[\"Estimated
```

Contract\"].value + \";\" + \" Total \" + cells[\"Estimated Total\"].value + \";\"+ \" Hours \" + cells[\"Estimated Hours\"].value + \";\"",

"engineConfig": {

"mode": "row-based",

"facets": []

},

"newColumnName": "Estimated Cost",

"columnInsertIndex": 36,

"baseColumnName": "Billed Total",

"expression": "grel:\"Labor \" + cells[\"Estimated Labor\"].value + \";\" + \" Material \" + cells[\"Estimated Material\"].value + \";\" + \" Equipment \" + cells[\"Estimated Equipment\"].value + \";\" + \" Contract \" + cells[\"Estimated Contract\"].value + \";\" + \" Total \" + cells[\"Estimated Total\"].value + \";\"+ \" Hours \" + cells[\"Estimated Hours\"].value + \";\"",

"onError": "set-to-blank"

},

{

"op": "core/text-transform",

"description": "Text transform on cells in column Actual Hours using expression grel:if (value==0, cells[\"Estimated Hours\"].value, value)",

"engineConfig": {

"mode": "row-based",

"facets": []

```
    },
    "columnName": "Actual Hours",
    "expression": "grel:if (value==0,  cells[\"Estimated Hours\"].value, value)",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in column Actual Total using expression grel:if(value==0,cells[\"Actual Labor\"].value + cells[\"Actual Material\"].value + cells[\"Actual Equipment\"].value + cells[\"Actual Contract\"].value, value)",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Actual Total",
    "expression": "grel:if(value==0,cells[\"Actual Labor\"].value + cells[\"Actual Material\"].value + cells[\"Actual Equipment\"].value + cells[\"Actual Contract\"].value, value)",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
```

```
},
{
  "op": "core/text-transform",
  "description": "Text transform on cells in column Actual Total using expression grel:if(value==0, cells[\"Estimated Total\"].value, value)",
  "engineConfig": {
    "mode": "row-based",
    "facets": []
  },
  "columnName": "Actual Total",
  "expression": "grel:if(value==0, cells[\"Estimated Total\"].value, value)",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10
}
{
  "op": "core/column-removal",
  "description": "Remove column Estimated Labor",
  "columnName": "Estimated Labor"
},
{
  "op": "core/column-removal",
  "description": "Remove column Estimated Material",
```

```
  "columnName": "Estimated Material"
},
{
  "op": "core/column-removal",
  "description": "Remove column Estimated Equipment",
  "columnName": "Estimated Equipment"
},
{
  "op": "core/column-removal",
  "description": "Remove column Estimated Contract",
  "columnName": "Estimated Contract"
},
{
  "op": "core/column-removal",
  "description": "Remove column Estimated Total",
  "columnName": "Estimated Total"
},
{
  "op": "core/column-removal",
  "description": "Remove column Estimated Hours",
  "columnName": "Estimated Hours"
},
{
```

```
    "op": "core/column-addition",
    "description": "Create column Encumbered Cost at index 15 based on column Encumbered Labor using expression grel:\"Labor \" + cells[\"Encumbered Labor\"].value + \";\" + \" Material \" + cells[\"Encumbered Material\"].value + \";\" + \" Equipment \" + cells[\"Encumbered Equipment\"].value + \";\" + \" Contract \" + cells[\"Encumbered Contract\"].value + \";\" + \" Total \" + cells[\"Encumbered Total\"].value + \";\"",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "newColumnName": "Encumbered Cost",
    "columnInsertIndex": 15,
    "baseColumnName": "Encumbered Labor",
    "expression": "grel:\"Labor \" + cells[\"Encumbered Labor\"].value + \";\" + \" Material \" + cells[\"Encumbered Material\"].value + \";\" + \" Equipment \" + cells[\"Encumbered Equipment\"].value + \";\" + \" Contract \" + cells[\"Encumbered Contract\"].value + \";\" + \" Total \" + cells[\"Encumbered Total\"].value + \";\"",
    "onError": "set-to-blank"
  },
  {
    "op": "core/column-move",
    "description": "Move column Encumbered Cost to position 16",
    "columnName": "Encumbered Cost",
```

```
  "index": 16
},
{
  "op": "core/column-move",
  "description": "Move column Encumbered Cost to position 31",
  "columnName": "Encumbered Cost",
  "index": 31
},
{
  "op": "core/column-removal",
  "description": "Remove column Encumbered Labor",
  "columnName": "Encumbered Labor"
},
{
  "op": "core/column-removal",
  "description": "Remove column Encumbered Material",
  "columnName": "Encumbered Material"
},
{
  "op": "core/column-removal",
  "description": "Remove column Encumbered Equipment",
  "columnName": "Encumbered Equipment"
},
```

```
{
  "op": "core/column-removal",
  "description": "Remove column Encumbered Contract",
  "columnName": "Encumbered Contract"
},
{
  "op": "core/column-removal",
  "description": "Remove column Encumbered Total",
  "columnName": "Encumbered Total"
},
{
  "op": "core/column-addition",
  "description": "Create column Actual Cost Detail at index 19 based on column Actual Total using expression grel:\"Labor \" + cells[\"Actual Labor\"].value + \";\" + \" Material \" + cells[\"Actual Material\"].value + \";\" + \" Equipment \" + cells[\"Actual Equipment\"].value + \";\" + \" Contract \" + cells[\"Actual Contract\"].value + \";\"",
  "engineConfig": {
    "mode": "row-based",
    "facets": []
  },
  "newColumnName": "Actual Cost Detail",
  "columnInsertIndex": 19,
  "baseColumnName": "Actual Total",
```

```
"expression": "grel:\"Labor \" + cells[\"Actual Labor\"].value + \";\" + \" Material \" + cells[\"Actual Material\"].value + \";\" + \" Equipment \" + cells[\"Actual Equipment\"].value + \";\" + \" Contract \" + cells[\"Actual Contract\"].value + \";\"",
"onError": "set-to-blank"
},
{
"op": "core/column-move",
"description": "Move column Actual Cost Detail to position 27",
"columnName": "Actual Cost Detail",
"index": 27
},
{
"op": "core/column-removal",
"description": "Remove column Actual Labor",
"columnName": "Actual Labor"
},
{
"op": "core/column-removal",
"description": "Remove column Actual Material",
"columnName": "Actual Material"
},
{
"op": "core/column-removal",
```

```
  "description": "Remove column Actual Equipment",
  "columnName": "Actual Equipment"
 },
 {
  "op": "core/column-removal",
  "description": "Remove column Actual Contract",
  "columnName": "Actual Contract"
 },
 {
  "op": "core/column-addition",
  "description": "Create column Billed Cost Detail at index 21 based on column Billed Total using expression grel:\"Labor \" + cells[\"Billed Labor\"].value + \";\" + \" Material \" + cells[\"Billed Material\"].value + \";\" + \" Equipment \" + cells[\"Billed Equipment\"].value + \";\" + \" Contract \" + cells[\"Billed Contract\"].value + \";\"",
  "engineConfig": {
   "mode": "row-based",
   "facets": []
  },
  "newColumnName": "Billed Cost Detail",
  "columnInsertIndex": 21,
  "baseColumnName": "Billed Total",
```

```
    "expression": "grel:\"Labor \" + cells[\"Billed Labor\"].value + \";\" + \" Material \" + cells[\"Billed Material\"].value + \";\" + \" Equipment \" + cells[\"Billed Equipment\"].value + \";\" + \" Contract \" + cells[\"Billed Contract\"].value + \";\"",
    "onError": "set-to-blank"
  },
  {
    "op": "core/column-move",
    "description": "Move column Billed Cost Detail to position 24",
    "columnName": "Billed Cost Detail",
    "index": 24
  },
  {
    "op": "core/column-removal",
    "description": "Remove column Billed Labor",
    "columnName": "Billed Labor"
  },
  {
    "op": "core/column-removal",
    "description": "Remove column Billed Material",
    "columnName": "Billed Material"
  },
  {
    "op": "core/column-removal",
```

```
    "description": "Remove column Billed Equipment",
    "columnName": "Billed Equipment"
  },
  {
    "op": "core/column-removal",
    "description": "Remove column Billed Contract",
    "columnName": "Billed Contract"
  },
  {
    "op": "core/column-rename",
    "description": "Rename column Estimated Cost to Estimated Cost Detail",
    "oldColumnName": "Estimated Cost",
    "newColumnName": "Estimated Cost Detail"
  },
  {
    "op": "core/column-rename",
    "description": "Rename column Encumbered Cost to Encumbered Cost Detail",
    "oldColumnName": "Encumbered Cost",
    "newColumnName": "Encumbered Cost Detail"
  },
]
```

APPENDIX G. THE MATLAB CODE FOR ANNUAL O&M COST CALCULATION

```
% Georgia Tech Operation & Maintenance work order history cleaning program
% Written by Xinghua Gao
% email: gaoxh@gatech.edu, gaoxinghua1988@gmail.com
% Date: October, 2018

clc;

%%%%%%%%% Input Parameters %%%%%%%%%
facility_num = 312;
start_year = 2006;
end_year = 2018;
table_name = temp3;
%%%%%%%% End Input %%%%%%%%%%%%%%%%%%

column_num = (end_year - start_year+1)*2 + 1;
data = table2cell(table_name);

% target table
tar = cell(facility_num,column_num);
tar (:,2:column_num) = {0};
```

```
[m,n] = size(data);


% create the map of facility name and number in this program
map = cell(facility_num,2);
for x = 1:facility_num
   map(x,1) = {x};
end


counter = 0;
for x = 1: m
   n = checkmap(data{x,2},facility_num,map);


   if n == 0
      counter = counter + 1;
      map{counter,2} = data{x,2};
      tar{counter,1} = data{x,2};


      for y = start_year:1:end_year
         if data{x,1} == y
            tar{counter,2*(y-2005)} = tar{counter,2*(y-2005)} + data {x,3};
            tar{counter,2*(y-2005)+1} = tar{counter,2*(y-2005)+1} + data {x,4};
         end
```

```
        end

    else
        for y = start_year:1:end_year
            if data{x,1} == y
                tar{n,2*(y-2005)} = tar{n,2*(y-2005)} + data {x,3};
                tar{n,2*(y-2005)+1} = tar{n,2*(y-2005)+1} + data {x,4};
            end
        end
    end

    % show progress
    r = rem (x,1000);
    if r == 0
        XX = [num2str(x), ' row calculated'];
        disp (XX);
    end
end

disp('Calculation ends');


function [y] = checkmap(x,facility_num,map)
```

```
    y = 0;
    for i = 1:facility_num
        if isequal(map{i,2},x)
            y = i;
        end
    end
end
```

APPENDIX H. THE R CODE FOR TIME SERIES BACKCASTING

H.1 The R code for utility consumption data backcasting (weekly)

```
# This is a R script for backcasting the utility consumption data (weekly)
# Developed by Xinghua Gao @ Georgia Tech
# Email: gaoxh@gatech.edu
# March 2019


############## Backcast functions #############


# Function to reverse time
#source: https://otexts.com/fpp2/backcasting.html
reverse_ts <- function(y)
{
  ts(rev(y), start=tsp(y)[1L], frequency=frequency(y))
}


# Function to reverse a forecast
#source: https://otexts.com/fpp2/backcasting.html
reverse_forecast <- function(object)
{
  h <- length(object[["mean"]])
  f <- frequency(object[["mean"]])
```

```
  object[["x"]] <- reverse_ts(object[["x"]])
  object[["mean"]] <- ts(rev(object[["mean"]]),
                         end=tsp(object[["x"]])[1L]-1/f, frequency=f)
  object[["lower"]] <- object[["lower"]][h:1L,]
  object[["upper"]] <- object[["upper"]][h:1L,]
  return(object)
}

############## Functions end #############

library(forecast)

############## Parameter setting ##########

# How many months to forecast/backcast
h <- 680

# Starting month, format: if from Oct 1, 2012, then write as "2012+9/12"
start_date = 2012+39/52

# Time seres data frequency, if monthly fre = 12, if weekly fre = 52
fre = 52
```

```
############ Parameter setting ends ########

# Import the monthly consumption data
utility <- read.csv("weekly.csv", header = TRUE)

ns <- ncol(utility)
fcast <- matrix(NA,nrow=h,ncol=ns)

 # i = 111
 ns = 2

for(i in 1:ns){
  zz <- as.numeric(utility[,i])

  # remove the data of last month (it is not correct)
  n <- length(zz)
  zz <- zz[1:(n-1)]

  # convert to time series format
  zz <- ts(zz, s= start_date, f = fre)

  # backcast
  zz %>%
```

```
    reverse_ts() %>%
    auto.arima() %>%
    forecast(h = h) %>%
    reverse_forecast() -> bc

  # record backcasted numbers
  fcast[,i] <- bc$mean

  # save the images to files
  #pdf(names(utility)[i])
  #png(paste0(names(utility)[i],".png"))

  #plot(bc, main = names(utility)[i])
  #autoplot(bc, main = names(utility)[i])
  #dev.off()

  print(paste0("round ", i))
}

# plot
autoplot(bc, main = names(utility)[i])

# write results to file
```

```
write(t(fcast),file="Backcasts.csv",sep=",",ncol=ncol(fcast))
```

H.2 The R code for utility consumption data backcasting (monthly)

```
# This is a R script for backcasting the utility consumption data (monthly)
# Developed by Xinghua Gao @ Georgia Tech
# Email: gaoxh@gatech.edu
# March 2019

############## Backcast functions #############

# Function to reverse time
#source: https://otexts.com/fpp2/backcasting.html
reverse_ts <- function(y)
{
  ts(rev(y), start=tsp(y)[1L], frequency=frequency(y))
}

# Function to reverse a forecast
#source: https://otexts.com/fpp2/backcasting.html
reverse_forecast <- function(object)
{
  h <- length(object[["mean"]])
  f <- frequency(object[["mean"]])
```

```
  object[["x"]] <- reverse_ts(object[["x"]])
  object[["mean"]] <- ts(rev(object[["mean"]]),
                         end=tsp(object[["x"]])[1L]-1/f, frequency=f)
  object[["lower"]] <- object[["lower"]][h:1L,]
  object[["upper"]] <- object[["upper"]][h:1L,]
  return(object)
}

############## Functions end #############

library(forecast)

############## Parameter setting ##########

# How many months to forecast/backcast
# h <- 199 # 2014 July
# h <- 188 # 2013 July
 h <- 178 # 2012 Oct
# h <- 4

# Input file name
input_file = "group2.csv"
```

```
# Starting month, format: if from Oct 1, 2012, then write as "2012+9/12"
# Where 12 is the frequency and 9 is the poropotion of time already passed.
start_date = 2012+9/12

############ Parameter setting ends ########

# Import the monthly consumption data
utility <- read.csv(input_file, header = TRUE)

ns <- ncol(utility)
fcast <- matrix(NA,nrow=h,ncol=ns)

# i = 8
# ns =2

# save the images to files
#pdf("myOut.pdf")

for(i in 1:ns){
  zz <- as.numeric(utility[,i])

  # remove the data of the first and last month (they may be incorrect)
  n <- length(zz)
```

```
  zz <- zz[1:(n-1)]
  zz <- zz[2:(n-1)]

  # convert to time series format
  zz <- ts(zz, s= start_date, f = 12)

  # backcast
  zz %>%
    reverse_ts() %>%
    auto.arima() %>%
    forecast(h = h) %>%
    reverse_forecast() -> bc

  # record backcasted numbers
  fcast[,i] <- bc$mean

  # plot
  #plot(bc, main = names(utility)[i])

  print(paste0("round ", i))
}

#dev.off()
```

```
autoplot(bc, main = names(utility)[i])

write(t(fcast),file="Backcasts.csv",sep=",",ncol=ncol(fcast))

#  ggtitle(paste("Backcasts from",bc[["method"]]))
```

H.3 The R code for utility consumption data forecasting (monthly)

In the experiments, sometimes forecasting is needed because some buildings have missing data from time to time. Hence, time series forecasting was used to simulate the data.

```
# This is a R script for forecasting the utility consumption data
# Developed by Xinghua Gao @ Georgia Tech
# Email: gaoxh@gatech.edu
# March 2019

library(forecast)

############## Parameter setting ##########

# How many months to forecast/backcast
h <- 65
```

```
# Input file name

input_file = "group2.csv"


start_date = 2012+9/12


############ Parameter setting ends ########



utility <- read.csv(input_file,header=TRUE)


ns <- ncol(utility)

fcast <- matrix(NA,nrow=h,ncol=ns)


for(i in 1:ns){

  zz <- as.numeric(utility[,i])


  # remove the data of the first and last month (they may be incorrect)

  n <- length(zz)

  zz <- zz[1:(n-1)]

  zz <- zz[2:(n-1)]


  # convert to time series format
```

```
  zz <- ts(zz, s= start_date, f = 12)


  # forecast
  fc <-forecast(zz,h=h)


  # record backcasted numbers
  fcast[,i] <- fc$mean


  # plot
  #plot(fc, main = names(utility)[i])


  print(paste0("round ", i))
}


autoplot(fc, main = names(utility)[i])


write(t(fcast),file="Forecasts.csv",sep=",",ncol=ncol(fcast))
```

H.4 The MATLAB code to prepare O&M cost data for backcasting

```
% This is a MATLAB program for preparing the O&M cost data for backcasting
% Developed by Xinghua Gao @ Georgia Tech
% Email: gaoxh@gatech.edu
% March 2019
```

```
% The annually monetary costs are saved in the variable "money"
% The annually labor hours are saved in the variable "hours"

% Use guide:

% Import the file "om_raw.csv" to get the cost table

%%%%%%%%%%%%%%%%%%%% Parameter setting %%%%%%%%%%%%%%%%%%%%
start_path = fullfile('C:\test\4\');% Define a starting folder.
    % IMPORTANT:theoutput folder (current folder of MATLAB)
    % must be the same folder.
path = "C:\test\4\";

start_year = 2006;

%%%%%%%%%%%%%%%%%%%% Parameter setting end %%%%%%%%%%%%%%%%%%%%

[m,n] = size (omraw); % the size of the raw cost data
end_year = start_year + m/2 -1;
```

```
money = zeros(m/2,n);
hours = zeros(m/2,n);

for j = 1:n
   for i = 1:m/2
      money(i,j) = omraw{(2*i-1),j};
      hours(i,j) = omraw{(2*i),j};
   end
end
```

H.5 The R code for O&M cost data backcasting

```
# This is a R script for backcasting the O&M costs
# Developed by Xinghua Gao @ Georgia Tech
# Email: gaoxh@gatech.edu
# March 2019

############## Backcast functions #############

# Function to reverse time
#source: https://otexts.com/fpp2/backcasting.html
reverse_ts <- function(y)
{
  ts(rev(y), start=tsp(y)[1L], frequency=frequency(y))
```

```
}

# Function to reverse a forecast
#source: https://otexts.com/fpp2/backcasting.html
reverse_forecast <- function(object)
{
  h <- length(object[["mean"]])
  f <- frequency(object[["mean"]])
  object[["x"]] <- reverse_ts(object[["x"]])
  object[["mean"]] <- ts(rev(object[["mean"]]),
                         end=tsp(object[["x"]])[1L]-1/f, frequency=f)
  object[["lower"]] <- object[["lower"]][h:1L,]
  object[["upper"]] <- object[["upper"]][h:1L,]
  return(object)
}

############## Functions end #############

library(forecast)

############## Parameter setting ##########

# How many years to forecast/backcast
```

```
h <- 8

# Input file name
input_file = "om_hours.csv"

# Starting month, format: if from Oct 1, 2012, then write as "2012+9/12"
# Where 12 is the frequency and 9 is the poropotion of time already passed.
start_date = 2012

############ Parameter setting ends ########

# Import the monthly consumption data
utility <- read.csv(input_file, header = TRUE)

ns <- ncol(utility)
fcast <- matrix(NA,nrow=h,ncol=ns)

# i = 8
# ns =2

# save the images to files
#pdf("myOut.pdf")
```

```
for(i in 1:ns){
  zz <- as.numeric(utility[,i])

  # remove the data of the first and last month (they may be incorrect)
  n <- length(zz)
  zz <- zz[1:(n-1)]
  zz <- zz[2:(n-1)]

  # convert to time series format
  zz <- ts(zz, s= start_date, f = 1)

  # backcast
  zz %>%
    reverse_ts() %>%
    auto.arima() %>%
    forecast(h = h) %>%
    reverse_forecast() -> bc

  # record backcasted numbers
  fcast[,i] <- bc$mean

  # plot
  #plot(bc, main = names(utility)[i])
```

```
  print(paste0("round ", i))
}

#dev.off()

autoplot(bc, main = names(utility)[i])

write(t(fcast),file="Backcasts.csv",sep=",",ncol=ncol(fcast))

#  ggtitle(paste("Backcasts from",bc[["method"]]))
```

APPENDIX I. THE LINEAR CORRELATION OF ATTRIBUTES

I.1 The linear correlation of attributes (total initial, utility, and O&M costs)

	initial	utility	om	gsf	floor	age	BLDG_SVC	CIRC	MECH	CLS_FAC
initial	1	0.684868	0.581039	0.568438	0.371307	-0.42186	-0.12596	0.113823	0.290264	0.146258
utility	0.684868	1	0.701122	0.378617	0.203526	-0.25845	-0.00911	0.066591	0.256681	0.180344
om	0.581039	0.701122	1	0.401407	0.213914	-0.09336	0.04279	0.092273	0.148799	0.259958
gsf	0.568438	0.378617	0.401407	1	0.660785	-0.25741	-0.15456	0.307756	0.119881	0.03611
floor	0.371307	0.203526	0.213914	0.660785	1	0.011293	-0.22561	0.216846	0.069083	0.07339
age	-0.42186	-0.25845	-0.09336	-0.25741	0.011293	1	0.148313	0.001716	-0.17005	0.05187
BLDG_SVC	-0.12596	-0.00911	0.04279	-0.15456	-0.22561	0.148313	1	-0.15885	-0.06879	0.112993
CIRC	0.113823	0.066591	0.092273	0.307756	0.216846	0.001716	-0.15885	1	-0.17041	0.158131
MECH	0.290264	0.256681	0.148799	0.119881	0.069083	-0.17005	-0.06879	-0.17041	1	-0.00608
CLS_FAC	0.146258	0.180344	0.259958	0.03611	0.07339	0.05187	0.112993	0.158131	-0.00608	1

LAB_FAC	0.169365	0.227244	0.202816	-0.04981	-0.0908	-0.12282	0.008543	-0.2074	0.076187	0.080344
OFF_FAC	-0.06888	0.064727	0.225134	-0.20373	-0.1695	0.155501	0.466483	-0.14753	0.056382	0.163798
STDY_FAC	-0.00724	-0.00674	0.001975	-0.03805	0.174829	0.044973	-0.04658	-0.08239	0.06768	-0.03932
SPEC_USE	-0.00796	0.001679	-0.00612	-0.02672	-0.13699	-0.14955	0.218155	-0.25061	-0.14189	-0.08146
GEN_USE	-0.04105	-0.00348	0.033601	-0.10351	-0.20006	0.002292	0.047134	-0.1581	-0.19024	-0.03251
SUPP_FAC	-0.10168	-0.1111	-0.13889	0.150942	-0.09731	-0.14072	-0.21225	0.380931	-0.28294	-0.10799
HLTH_FAC	-0.03798	-0.05734	0.012893	-0.04848	-0.11287	-0.08454	-0.01957	-0.08807	-0.01781	-0.04407
RES_FAC	-0.13395	-0.26521	-0.37184	0.049476	0.285506	0.14945	-0.34769	-0.03759	0.025331	-0.26871
other	0.44411	0.245702	0.111273	0.110487	0.092653	-0.09554	-0.0452	-0.03583	0.267505	-0.02927

	LAB_FAC	OFF_FAC	STDY_FAC	SPEC_USE	GEN_USE	SUPP_FAC	HLTH_FAC	RES_FAC	other
initial	0.169365	-0.06888	-0.00724	-0.00796	-0.04105	-0.10168	-0.03798	-0.13395	0.44411
utility	0.227244	0.064727	-0.00674	0.001679	-0.00348	-0.1111	-0.05734	-0.26521	0.245702
om	0.202816	0.225134	0.001975	-0.00612	0.033601	-0.13889	0.012893	-0.37184	0.111273

gsf	-0.04981	-0.20373	-0.03805	-0.02672	-0.10351	0.150942	-0.04848	0.049476	0.110487
floor	-0.0908	-0.1695	0.174829	-0.13699	-0.20006	-0.09731	-0.11287	0.285506	0.092653
age	-0.12282	0.155501	0.044973	-0.14955	0.002292	-0.14072	-0.08454	0.14945	-0.09554
BLDG_SVC	0.008543	0.466483	-0.04658	0.218155	0.047134	-0.21225	-0.01957	-0.34769	-0.0452
CIRC	-0.2074	-0.14753	-0.08239	-0.25061	-0.1581	0.380931	-0.08807	-0.03759	-0.03583
MECH	0.076187	0.056382	0.06768	-0.14189	-0.19024	-0.28294	-0.01781	0.025331	0.267505
CLS_FAC	0.080344	0.163798	-0.03932	-0.08146	-0.03251	-0.10799	-0.04407	-0.26871	-0.02927
LAB_FAC	1	0.070888	-0.15091	-0.12676	-0.24246	-0.14491	-0.06669	-0.40658	0.118302
OFF_FAC	0.070888	1	-0.10131	-0.11232	-0.09328	-0.18588	0.022323	-0.53826	-0.02844
STDY_FAC	-0.15091	-0.10131	1	-0.06275	-0.01546	-0.08506	-0.03001	-0.00414	-0.0434
SPEC_USE	-0.12676	-0.11232	-0.06275	1	0.018696	-0.05921	-0.02006	-0.12069	-0.0151
GEN_USE	-0.24246	-0.09328	-0.01546	0.018696	1	-0.13723	-0.04232	-0.14876	-0.04989

SUPP_FA C	-0.14491	-0.18588	-0.08506	-0.05921	-0.13723	1	-0.03324	-0.18509	-0.04604
HLTH_FA C	-0.06669	0.022323	-0.03001	-0.02006	-0.04232	-0.03324	1	-0.05395	-0.01389
RES_FAC	-0.40658	-0.53826	-0.00414	-0.12069	-0.14876	-0.18509	-0.05395	1	-0.08469
other	0.118302	-0.02844	-0.0434	-0.0151	-0.04989	-0.04604	-0.01389	-0.08469	1

I.2 The linear correlation of attributes (initial, utility, and O&M costs per square footage)

	initial	utility	om	gsf	floor	age	BLDG_SVC	CIRC	MECH	CLS_FAC
initial	1	0.271661	0.34364	-0.13415	-0.27608	-0.341	0.245097	-0.3204	0.084301	-0.04019
utility	0.271661	1	0.212271	-0.15073	-0.30479	-0.10745	0.28977	-0.31107	0.072601	-0.03649
om	0.34364	0.212271	1	-0.21926	-0.35417	0.094531	0.167198	-0.18463	0.01616	0.064605
gsf	-0.13415	-0.15073	-0.21926	1	0.660785	-0.25741	-0.15456	0.307756	0.119881	0.03611
floor	-0.27608	-0.30479	-0.35417	0.660785	1	0.011293	-0.22561	0.216846	0.069083	0.07339

age	-0.341	-0.10745	0.094531	-0.25741	0.011293	1	0.148313	0.001716	-0.17005	0.05187
BLDG_SVC	0.245097	0.28977	0.167198	-0.15456	-0.22561	0.148313	1	-0.15885	-0.06879	0.112993
CIRC	-0.3204	-0.31107	-0.18463	0.307756	0.216846	0.001716	-0.15885	1	-0.17041	0.158131
MECH	0.084301	0.072601	0.01616	0.119881	0.069083	-0.17005	-0.06879	-0.17041	1	-0.00608
CLS_FAC	-0.04019	-0.03649	0.064605	0.03611	0.07339	0.05187	0.112993	0.158131	-0.00608	1
LAB_FAC	0.248154	0.234957	0.11369	-0.04981	-0.0908	-0.12282	0.008543	-0.2074	0.076187	0.080344
OFF_FAC	-0.03616	0.234126	0.397904	-0.20373	-0.1695	0.155501	0.466483	-0.14753	0.056382	0.163798
STDY_FAC	-0.04463	-0.05776	-0.0179	-0.03805	0.174829	0.044973	-0.04658	-0.08239	0.06768	-0.03932
SPEC_USE	0.332761	-0.00593	0.033613	-0.02672	-0.13699	-0.14955	0.218155	-0.25061	-0.14189	-0.08146
GEN_USE	0.12215	0.160932	0.257984	-0.10351	-0.20006	0.002292	0.047134	-0.1581	-0.19024	-0.03251
SUPP_FAC	-0.08401	-0.10231	-0.16563	0.150942	-0.09731	-0.14072	-0.21225	0.380931	-0.28294	-0.10799
HLTH_FAC	0.013283	-0.04544	0.112292	-0.04848	-0.11287	-0.08454	-0.01957	-0.08807	-0.01781	-0.04407
RES_FAC	-0.24693	-0.26501	-0.40924	0.049476	0.285506	0.14945	-0.34769	-0.03759	0.025331	-0.26871
other	0.184681	0.043633	-0.03679	0.110487	0.092653	-0.09554	-0.0452	-0.03583	0.267505	-0.02927

	LAB_FAC	OFF_FAC	STDY_FAC	SPEC_USE	GEN_USE	SUPP_FAC	HLTH_FAC	RES_FAC	other
initial	0.248154	-0.03616	-0.04463	0.332761	0.12215	-0.08401	0.013283	-0.24693	0.184681
utility	0.234957	0.234126	-0.05776	-0.00593	0.160932	-0.10231	-0.04544	-0.26501	0.043633
om	0.11369	0.397904	-0.0179	0.033613	0.257984	-0.16563	0.112292	-0.40924	-0.03679
gsf	-0.04981	-0.20373	-0.03805	-0.02672	-0.10351	0.150942	-0.04848	0.049476	0.110487
floor	-0.0908	-0.1695	0.174829	-0.13699	-0.20006	-0.09731	-0.11287	0.285506	0.092653
age	-0.12282	0.155501	0.044973	-0.14955	0.002292	-0.14072	-0.08454	0.14945	-0.09554
BLDG_SVC	0.008543	0.466483	-0.04658	0.218155	0.047134	-0.21225	-0.01957	-0.34769	-0.0452
CIRC	-0.2074	-0.14753	-0.08239	-0.25061	-0.1581	0.380931	-0.08807	-0.03759	-0.03583
MECH	0.076187	0.056382	0.06768	-0.14189	-0.19024	-0.28294	-0.01781	0.025331	0.267505
CLS_FAC	0.080344	0.163798	-0.03932	-0.08146	-0.03251	-0.10799	-0.04407	-0.26871	-0.02927
LAB_FAC	1	0.070888	-0.15091	-0.12676	-0.24246	-0.14491	-0.06669	-0.40658	0.118302
OFF_FAC	0.070888	1	-0.10131	-0.11232	-0.09328	-0.18588	0.022323	-0.53826	-0.02844
STDY_FAC	-0.15091	-0.10131	1	-0.06275	-0.01546	-0.08506	-0.03001	-0.00414	-0.0434
SPEC_USE	-0.12676	-0.11232	-0.06275	1	0.018696	-0.05921	-0.02006	-0.12069	-0.0151

GEN_USE	-0.24246	-0.09328	-0.01546	0.018696	1	-0.13723	-0.04232	-0.14876	-0.04989
SUPP_FAC	-0.14491	-0.18588	-0.08506	-0.05921	-0.13723	1	-0.03324	-0.18509	-0.04604
HLTH_FAC	-0.06669	0.022323	-0.03001	-0.02006	-0.04232	-0.03324	1	-0.05395	-0.01389
RES_FAC	-0.40658	-0.53826	-0.00414	-0.12069	-0.14876	-0.18509	-0.05395	1	-0.08469
other	0.118302	-0.02844	-0.0434	-0.0151	-0.04989	-0.04604	-0.01389	-0.08469	1

APPENDIX J. THE R CODE FOR BASIC DATA ANALYSIS

```
# This is a R script for analyzing the facility life-cycle cost data

# Basic Data Analysis

# Developed by Xinghua Gao @ Georgia Tech

# Email: gaoxh@gatech.edu

# March 2019


# Load libraries

library(ggplot2)

library(keras)

library(mlbench)

library(dplyr)

library(magrittr)

library(neuralnet)

library(tensorflow)



################### Data importing and processing ###################


# Load raw data

data <- read.csv("train.csv", header = TRUE, check.names = TRUE)
```

```
# Change column name to solve the weird column header name issue
names(data)[1] <- "id"

# don't need the predictor "year" becasue already have the predictor "age"
data <- data[,-6]

# convert the predictors in Factor format to numeric
# data %<>% mutate_if(is.factor, as.numeric)

# Remove the instance with the largest utility cost: 189 Substation Control House
# It is a control house, its utility data are representing many other buildings
data <- data[-which.max(data$utility),]

# Remove the instance with the largest om cost: 73 McCamish Pavilion
# It is recently renovated in 2012 and the removation costs are recorded in the AiM system
data <- data[-which.max(data$om),]

# The utility consumption of building 138 is abnormal
data <- data[-which(data$id==138),]

# Remove the O'Keefe, Daniel C.building, which O&M cost is abnormal
# It just had a major renovation recently and the cost is recorded as maintenance cost
```

```
data <- data[-which(data$id==33),]

# Create the data frame for the cost per quare foot
data.persf <- data
data.persf$initial <- data.persf$initial*1000/data.persf$gsf
data.persf$utility <- data.persf$utility*1000/data.persf$gsf
data.persf$om <- data.persf$om*1000/data.persf$gsf

# Remove the outliers of utility cost, based on the cost per SF
data <- data[-which(data.persf$utility>=400),]
data.persf <- data.persf[-which(data.persf$utility>=400),]

# Remove the outliers of O&M cost, based on the cost per SF
data <- data[-which(data.persf$om>=1000),]
data.persf <- data.persf[-which(data.persf$om>=1000),]

# Preparing for checking the correlations of main paramters
data_2 <- data[, which(names(data) %in%
c("initial","utility","om","gsf","floor","age"))]
data.persf_2 <- data.persf[, which(names(data.persf) %in%
c("initial","utility","om","gsf","floor","age"))]

# the rate of initial to utility and om
```

```
rate_iu <- data$utility/data$initial

rate_iom <- data$om/data$initial



################## END Data importing and processing ##################



################## Plot ##################

### the correlations ###

# plot the correlations

pairs(data_2)

#cor(data_2) # the correlation table (less attributes)

cor <- cor(data[,-c(1:2,7:9,11,13)]) # the correlation table (all attributes)

# the correlations of per sf data

pairs(data.persf_2)

#cor(data.persf_2)

cor.persf <- cor(data.persf[,-c(1:2,7:9,11,13)])

### END the correlations ###

### the cost per SF histrograms ###
```

```
# plot the histograms
hist(data.persf$initial,breaks=100)
hist(data.persf$utility,breaks=100)
hist(data.persf$om,breaks=100)

# plot the histograms of the rate of initial to utility and om
hist(rate_iu, breaks=100)
hist(rate_iom, breaks=100)

### END the cost per SF histrograms ###

### Find the building with largest rate ###
# the id of the buidling with max rate of initial and utility
max_iu <- data$id[which.max(rate_iu)]

# the id of the buidling with max rate of initial and o&M
max_iom <- data$id[which.max(rate_iom)]

### END Find the building with largest rate ###

### the number of buildings by owner and cost per SF ###
# plot the count of buildings by owner and initial cost
ggplot(data.persf[data.persf$owner != "n/a",], aes(x = initial)) +
```

```
  facet_wrap(~owner) +
  geom_histogram(binwidth = 15) +
  ggtitle("The count of buildings by owner and initial cost") +
  xlab("Initial cost per SF") +
  ylab("The number of buildings")

# plot the count of buildings by owner and utility cost
ggplot(data.persf[data.persf$owner != "n/a",], aes(x = utility)) +
  facet_wrap(~owner) +
  geom_histogram(binwidth = 15) +
  ggtitle("The count of buildings by owner and utility cost") +
  xlab("Utility cost per SF") +
  ylab("The number of buildings")

# plot the count of buildings by owner and O&M cost
ggplot(data.persf[data.persf$owner != "n/a",], aes(x = om)) +
  facet_wrap(~owner) +
  geom_histogram(binwidth = 15) +
  ggtitle("The count of buildings by owner and O&M cost") +
  xlab("O&M cost per SF") +
  ylab("The number of buildings")

### END the number of buildings by owner and cost per SF ###
```

```
################## END Plot ##################

################## Write file ##################

write.table(cor, file = "Correlation.csv", sep = ",", col.names = NA,
            qmethod = "double")

write.table(cor.persf, file = "Correlation_persf.csv", sep = ",", col.names = NA,
            qmethod = "double")

################## END Write file ##################
```

APPENDIX K. THE R CODE FOR MODEL TRAINING AND VALIDATION

```
# This is a R script for analyzing the facility life-cycle cost data

# Machine learning models for single/multi target regression:

# Developed by Xinghua Gao @ Georgia Tech

# Email: gaoxh@gatech.edu

# March 2019



# Load libraries

library(ggplot2)

library(stringr)

library(hydroGOF)

library(Metrics)

library(class)

library(caret)

library(pls)
```

```
library(FNN)

library(rpart)

library(boot)

library(keras)

library(mlbench)

library(dplyr)

library(magrittr)

library(neuralnet)

library(tensorflow)

library(e1071)

library(gmodels)

library(psych)

library(randomForest)

library(MultivariateRandomForest)



################### Parameter setting ###################

# how many iterations? (100 by default)
```

```
loop = 3



# set a threshold for the buildng age;

# the buildings older than this age won't be used in

# the model development

age = 100



# set the model to train, 0 means not train, 1 means train

MLR = 0 # Linear regression

KNN = 0 # KNN

tree = 0 # random forest

SVM = 0 # SVM

MLP = 0 # multilayer perceptron

tree_multi = 0 # multi-output random forest

MLP_multi = 1 # multilayer perceptron (multi-target)



# number of descriptive attributes
```

```
num_des = 16


# the k value of the KNN model

knn_k = 3


# epochs of the multilayer perceptron model training

num_epo = 100


# batch size of the multilayer perceptron model training

num_batch = 90


# the validation split of the multilayer perceptron model training

# From 0.01 to 0.99, the percent of validation set.

val_split = 0.02


# descriptive attributes for initial cost prediction

attri_initial <- initial ~ gsf + age + floor +
```

```
  BLDG_SVC + CIRC +MECH + CLS_FAC + LAB_FAC +

  OFF_FAC + STDY_FAC + SPEC_USE + GEN_USE +

  SUPP_FAC  + HLTH_FAC + RES_FAC + other



# descriptive attributes for utility cost prediction

attri_utility <- utility ~ gsf + age + floor +

  BLDG_SVC + CIRC +MECH + CLS_FAC + LAB_FAC +

  OFF_FAC + STDY_FAC + SPEC_USE + GEN_USE +

  SUPP_FAC  + HLTH_FAC + RES_FAC + other



# descriptive attributes for O&M cost prediction

attri_om <- om ~ gsf + age + floor +

  BLDG_SVC + CIRC +MECH + CLS_FAC + LAB_FAC +

  OFF_FAC + STDY_FAC + SPEC_USE + GEN_USE +

  SUPP_FAC  + HLTH_FAC + RES_FAC + other



################## END Parameter setting ##################
```

```
################## Data importing and processing ##################


# Load raw data

data <- read.csv("train.csv", header = TRUE, check.names = TRUE)


# Change column name to solve the weird column header name issue

names(data)[1] <- "id"


# don't need the predictor "year" becasue already have the predictor "age"

data <- data[,-6]


# Remove the instance with the largest utility cost: 189 Substation Control House

# It is a control house, its utility data are representing many other buildings

data <- data[-which.max(data$utility),]
```

```
# Remove the instance with the largest om cost: 73 McCamish Pavilion

# It is recently renovated in 2012 and the removation costs are recorded in the AiM system

data <- data[-which.max(data$om),]



# Remove the O'Keefe, Daniel C.building, which O&M cost is abnormal

# It just had a major renovation recently and the cost is recorded as maintenance cost

data <- data[-which(data$id==33),]



# The utility consumption of building 138 is abnormal

data <- data[-which(data$id==138),]



# convert the predictors in Factor format to numeric

data %<>% mutate_if(is.factor, as.numeric)



# Create the data frame for the cost per quare foot

data.persf <- data
```

```
data.persf$initial <- data.persf$initial*1000/data.persf$gsf

data.persf$utility <- data.persf$utility*1000/data.persf$gsf

data.persf$om <- data.persf$om*1000/data.persf$gsf



# Remove the outliers of utility cost, based on the cost per SF

data <- data[-which(data.persf$utility>=400),]

data.persf <- data.persf[-which(data.persf$utility>=400),]



# Remove the outliers of O&M cost, based on the cost per SF

data <- data[-which(data.persf$om>=1000),]

data.persf <- data.persf[-which(data.persf$om>=1000),]



# Remove the instances with age older than the threshold

data <- data[!(data$age >age),]

data.persf <- data.persf[!(data.persf$age >age),]
```

```
################## END Data importing and processing ##################




################## Define the result data array ##################

# results <- data.frame("method" = c("MLR(single)", "KNN(single)",

#                                    "tree(single)", "SVM(single)", "MLP(single)",

#                                    "tree(multi)", "MLP(multi)"),

#                       "initial" = 1:7, "utility" = 1:7, "om" = 1:7)



# record each iteration

result_table <- array(0,dim=c(7,3,loop))



# record the mean of all iteration

result <- array(0,dim=c(7,3))



# The counter for valid loops

# Sometimes the loop may be skipped
```

```
counter = 0

################## END Define the result data frame ##################

################## Loop starts ##################

for (i in 1:loop){

################## Define the training set and test set ##################

# set the random seed if needed

# set.seed(123)

# data partition

ind <- sample(2, nrow(data), replace =T, prob = c(.8,.2))
```

```
training <- data.frame(data[ind==1,-c(1:5,7:9,11,13)])

test <- data.frame(data[ind==2,-c(1:5,7:9,11,13)])



trainingtarget <- data.frame(data[ind==1, c(3:5)])

testtarget <- data.frame(data[ind==2, c(3:5)])



# normalize



m <- colMeans(training)

s <- apply(training,2,sd)



# some times the m and s have zero elements, which make the training and test set have Nah

# in this case, skip to the next loop

if (any(m == 0)||any(s == 0)) {

  print(paste0("round ", i, ", the loop is skipped"))

  next
```

```
}


training <- data.frame(scale (training, center = m, scale =s))

test <- data.frame(scale (test, center = m, scale =s))


# normalize targets to test

m2 <- colMeans(trainingtarget) # here do use the mean and SD of the trainingtarget

s2 <- apply(trainingtarget,2,sd)

trainingtarget <- data.frame(scale (trainingtarget, center = m2, scale =s2))

testtarget <- data.frame(scale (testtarget, center = m2, scale =s2))


# individual targets for some R packages

training.initial <- trainingtarget[c(1)]

training.utility <- trainingtarget[c(2)]

training.om <- trainingtarget[c(3)]


test.initial <- testtarget[c(1)]
```

```
test.utility <- testtarget[c(2)]

test.om <- testtarget[c(3)]



# merge the training and test set for some R packages

training.merge <- data.frame(trainingtarget,training)

test.merge <- data.frame(testtarget,test)



################## END Define the training set and test set ##################



################## Model development and validation ##################



### Multilinear regression model (single target) ###



if (MLR == 1){
```

```
# MRL for initial cost

MLR_S_initial <- lm(attri_initial, data = training.merge)



# MRL for utility cost

MLR_S_utility <- lm(attri_utility, data = training.merge)



# MRL for O&M cost

MLR_S_om <- lm(attri_om, data = training.merge)



# Predictions using the developed linear models

pred_MLRSI <- predict (MLR_S_initial, test.merge) # initial

pred_MLRSU <- predict (MLR_S_initial, test.merge) # utility

pred_MLRSO <- predict (MLR_S_initial, test.merge) # om



# MLR validation
```

```
# Root Mean Squared Error (RMSE)

# RMSE = rmse(predY,test.merge$initial)



# Mean squared error (MSE)

#MSE = mse(predY,test.merge$initial)



# Relative absolute error (RAE)

#RAE = rae(test.merge$initial, predY)



# Mean absolute error (MAE)

MAE_MLRSI = mae(test.merge$initial,pred_MLRSI)

MAE_MLRSU = mae(test.merge$utility,pred_MLRSU)

MAE_MLRSO = mae(test.merge$om,pred_MLRSO)



# record the results

result_table[1,1,i] = MAE_MLRSI;

result_table[1,2,i] = MAE_MLRSU;
```

```
result_table[1,3,i] = MAE_MLRSO;



}



### END Multilinear regression model (single target) ###



### KNN regression model (single target) ###



if (KNN == 1){



# predictions based on the test set

# the experimens indicated that k = 3 yeilds best results

pred_KNNSI <- knn.reg(training, test, training.initial, k = knn_k)

pred_KNNSU <- knn.reg(training, test, training.utility, k = knn_k)

pred_KNNSO <- knn.reg(training, test, training.om, k = knn_k)



# Mean absolute error (MAE)
```

```
MAE_KNNSI = mae(testtarget$initial,pred_KNNSI$pred)

MAE_KNNSU = mae(testtarget$utility,pred_KNNSU$pred)

MAE_KNNSO = mae(testtarget$om,pred_KNNSO$pred)



# record the results

result_table[2,1,i] = MAE_KNNSI;

result_table[2,2,i] = MAE_KNNSU;

result_table[2,3,i] = MAE_KNNSO;



}



### END KNN regression model (single target) ###



### Regression Tree model (single target) ###



if (tree == 1){
```

```
# random forest for initial cost
forest_SI <- randomForest(attri_initial,
                          data = training.merge)

# random forest for utility cost
forest_SU <- randomForest(attri_utility,
                          data = training.merge)

# random forest for O&M cost
forest_SO <- randomForest(attri_om,
                          data = training.merge)

# predictions based on the test set
pred_treeSI <- predict(forest_SI, test)
pred_treeSU <- predict(forest_SU, test)
```

```
pred_treeSO <- predict(forest_SO, test)



# Mean absolute error (MAE)

MAE_treeSI <- mae(testtarget$initial,pred_treeSI)

MAE_treeSU <- mae(testtarget$utility,pred_treeSU)

MAE_treeSO <- mae(testtarget$om,pred_treeSO)



# record the results

result_table[3,1,i] = MAE_treeSI;

result_table[3,2,i] = MAE_treeSU;

result_table[3,3,i] = MAE_treeSO;



}



### END Regression Tree model (single target) ###
```

```
### SVM Regression model (single target) ###


if (SVM == 1){


# SVM regression for initial cost

SVM_S_initial <- svm(attri_initial,

                     data = training.merge)


# SVM regression for utility cost

SVM_S_utility <- svm(attri_utility,

                     data = training.merge)


# SVM regression for O&M cost

SVM_S_om <- svm(attri_om,

                data = training.merge)


# predictions based on the test set
```

```
pred_SVMSI <- predict(SVM_S_initial, test)

pred_SVMSU <- predict(SVM_S_utility, test)

pred_SVMSO <- predict(SVM_S_om, test)



# Mean absolute error (MAE)

MAE_SVMSI <- mae(testtarget$initial,pred_SVMSI)

MAE_SVMSU <- mae(testtarget$utility,pred_SVMSU)

MAE_SVMSO <- mae(testtarget$om,pred_SVMSO)



# record the results

result_table[4,1,i] = MAE_SVMSI;

result_table[4,2,i] = MAE_SVMSU;

result_table[4,3,i] = MAE_SVMSO;



}



### END SVM Regression model (single target) ###
```

```
# converge the data into matrix

training_m <- as.matrix(training)

trainingtarget_m <- as.matrix(trainingtarget)

training.initial_m <- as.matrix(training.initial)

training.utility_m <- as.matrix(training.utility)

training.om_m <- as.matrix(training.om)

test_m <- as.matrix(test)

testtarget_m <- as.matrix(testtarget)

test.initial_m <- as.matrix(test.initial)

test.utility_m <- as.matrix(test.utility)

test.om_m <- as.matrix(test.om)



### Multilayer perceptron model (single target) ###
```

```
if (MLP == 1){



# create MLP model (empty)

model_SI <- keras_model_sequential()

model_SI  %>%

  layer_dense(units = 10, activation = 'relu', input_shape = c(num_des)) %>%

  layer_dense(units = 8, activation = 'relu') %>%

  layer_dense(units = 5, activation = 'relu') %>%

  layer_dense(units = 1)



# compile

model_SI  %>% compile(loss = 'mse',

                optimizer = 'rmsprop',

                metrics = 'mae')



model_SU <- model_SI

model_SO <- model_SI
```

```
# fit the MLP model for initial cost prediction
MLP_SI <- model_SI  %>%
  fit(training_m,
      training.initial_m,
      epochs = num_epo,
      batch_size = num_batch,
      validation_split = val_split)


# fit the MLP model for utility cost prediction
MLP_SU <- model_SU %>%
  fit(training_m,
      training.utility_m,
      epochs = num_epo,
      batch_size = num_batch,
      validation_split = val_split)
```

```
# fit the MLP model for O&M cost prediction

MLP_SO <- model_SO %>%

  fit(training_m,

      training.om_m,

      epochs = num_epo,

      batch_size = num_batch,

      validation_split = val_split)



# predictions based on the test set

pred_MLP_SI <- model_SI %>% predict(test_m)

pred_MLP_SU <- model_SU %>% predict(test_m)

pred_MLP_SO <- model_SO %>% predict(test_m)



# Mean absolute error (MAE)

MAE_MLP_SI <- mae(testtarget$initial,pred_MLP_SI)

MAE_MLP_SU <- mae(testtarget$utility,pred_MLP_SU)

MAE_MLP_SO <- mae(testtarget$om,pred_MLP_SO)
```

```
# record the results

result_table[5,1,i] = MAE_MLP_SI;

result_table[5,2,i] = MAE_MLP_SU;

result_table[5,3,i] = MAE_MLP_SO;



}



### END Multilayer perceptron model (single target) ###




### Regression Tree model (multi target) ###



if (tree_multi == 1){



# Multivariate Random Forest model

# build_forest_predict(trainX, trainY, n_tree, m_feature, min_leaf, testX)
```

```
forest_M <- build_forest_predict(training_m, trainingtarget_m, 100, 10, 40, test_m)



# Evaluation

MAE_forest_MI <- mae(testtarget$initial,forest_M[,1])

MAE_forest_MU <- mae(testtarget$utility,forest_M[,2])

MAE_forest_MO <- mae(testtarget$om,forest_M[,3])



# record the results

result_table[6,1,i] = MAE_forest_MI;

result_table[6,2,i] = MAE_forest_MU;

result_table[6,3,i] = MAE_forest_MO;



}



### END Regression Tree model (multi target) ###
```

```
### Multilayer perceptron model (multi target) ###


if (MLP_multi == 1){


# create MLP model (empty)

model_M <- keras_model_sequential()

model_M  %>%

  layer_dense(units = 10, activation = 'relu', input_shape = c(num_des)) %>%

  layer_dense(units = 8, activation = 'relu') %>%

  layer_dense(units = 5, activation = 'relu') %>%

  layer_dense(units = 3)


# compile

model_M  %>% compile(loss = 'mse',

                     optimizer = 'rmsprop',

                     metrics = 'mae')
```

```
# fit the model
MLP_SI <- model_M  %>%
  fit(training_m,
      trainingtarget_m,
      epochs = num_epo,
      batch_size = num_batch,
      validation_split = val_split)

# predictions based on the test set
pred_MLP_M <- model_M %>% predict(test_m)

# Mean absolute error (MAE)
MAE_MLP_MI <- mae(testtarget$initial,pred_MLP_M[,1])
MAE_MLP_MU <- mae(testtarget$utility,pred_MLP_M[,2])
MAE_MLP_MO <- mae(testtarget$om,pred_MLP_M[,3])

# record the results
```

```
result_table[7,1,i] = MAE_MLP_MI;

result_table[7,2,i] = MAE_MLP_MU;

result_table[7,3,i] = MAE_MLP_MO;



}



### END Multilayer perceptron model (multi target) ###




################### Loop ends ###################

print(paste0("round ", i, " finished"))



counter = counter + 1



}



################### Results Output ###################
```

```
for (i in 1:7){
  for (j in 1:3){
    for (k in 1:counter){
        result[i,j] = result[i,j] + result_table[i,j,k]
    }
    result[i,j] = result[i,j]/counter
  }
}

print(result)
```

################## END Results Output ##################

REFERENCES

[1] A. Noshadravan, T.R. Miller, J.G. Gregory, A Lifecycle Cost Analysis of Residential Buildings Including Natural Hazard Risk, Journal of Construction Engineering and Management 143 (7) (2017) 10. 10.1061/(asce)co.1943-7862.0001286.

[2] N. Bakis, R. Amaratunga, M. Kagioglou, G. Aouad, An integrated environment for life cycle costing in construction, (2003).

[3] R.J. Cole, E. Sterner, Reconciling theory and practice of life-cycle costing, Building Research & Information 28 (5-6) (2000) 368-375.

[4] D.J. Ferry, R. Flanagan, Life cycle costing: A radical approach, Construction Industry Research and Information Association London, 1991. ISBN: 0860173224.

[5] ASTM, Standard practice for measuring life-cycle costs of buildings and building systems, ASTM, West Conshohocken, PA, 2017. ISBN.

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

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

[8] J.D. Kelleher, B.M. Namee, A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies, MIT Press, 2015. ISBN: 9780262029445.

[9] T.M. Mitchell, Machine Learning, McGraw-Hill, 1997. ISBN: 9780071154673.

[10] SAS, Machine Learning: What it is & why it matters, Available from: https://www.sas.com/it_it/insights/analytics/machine-learning.html, (2018) (Internet, cited March 29, 2019).

[11] H.F. Deng, D. Fannon, M.J. Eckelman, Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata, Energy and Buildings 163 (2018) 34-43. 10.1016/j.enbuild.2017.12.031.

[12] X. Gao, P. Pishdad-Bozorgi, D. Shelden, Y. Hu, Machine Learning Applications in Facility Life-cycle Cost Analysis: A Review, The 2019 ASCE International Conference on Computing in Civil Engineering, ASCE, Atlanta, GA, 2019.

[13] R. Jin, S. Han, C. Hyun, Y. Cha, Application of Case-Based Reasoning for Estimating Preliminary Duration of Building Projects, Journal of Construction Engineering and Management 142 (2) (2016). 10.1061/(asce)co.1943-7862.0001072.

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

[15] C.S. Li, S.J. Guo, 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, 2012, pp. 215-221.

[16] C.W. Koo, T. Hong, C.T. Hyun, S.H. Park, J.o. Seo, 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) (2010) 121-137.

[17] K. Bala, S. Ahmad Bustani, B. Shehu Waziri, 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) (2014) 519-530.

[18] S. Banihashemi, G. Ding, J. Wang, 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, 2017, pp. 371-376. 10.1016/j.egypro.2017.03.155.

[19] C.V. Gallagher, K. Leahy, P. O'Donovan, K. Bruton, D.T.J. O'Sullivan, Development and application of a machine learning supported methodology for measurement and verification (M&V) 2.0, Energy and Buildings 167 (2018) 8-22. 10.1016/j.enbuild.2018.02.023.

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

[21] H.W. Shi, W.Q. Li, The Integrated Methodology of Rough Set Theory and Artificial Neural-Network for Construction Project Cost Prediction, 2008. ISBN: 978-0-7695-3497-8.

[22] R.N. Milion, J.C. Paliari, L.H.B. Liboni, Improving consumption estimation of electrical materials in residential building construction, Automation in Construction 72 (2016) 93-101. 10.1016/j.autcon.2016.08.042.

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

[24] A. Weerasinghe, T. Ramachandra, J.O.B. Rotimi, A Simplified model for predicting running cost of office buildings in Sri Lanka, Proceedings of the 32nd Annual ARCOM Conference, 2016, pp. - 340.

[25] R. Minerva, A. Biru, D. Rotondi, Towards a definition of the Internet of Things (IoT), IEEE Internet Initiative 1 (2015) 1-86.

[26] C. Eastman, P. Teicholz, R. Sacks, K. Liston, BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors (2nd Edition), John Wiley & Sons, 2011. ISBN: 0470541377, 9780470541371.

[27] National Institute of Building Sciences, United States National Building Information Modeling Standard, Version 3, 2015.

[28] International Organization for Standardization, ISO 29481-1:2016. Building information models -- Information delivery manual -- Part 1: Methodology and format, 2016.

[29] R. Volk, J. Stengel, F. Schultmann, Building Information Modeling (BIM) for existing buildings — Literature review and future needs, Automation in Construction 38 (0) (2014) 109-127. http://dx.doi.org/10.1016/j.autcon.2013.10.023.

[30] E. Krygiel, B. Nies, Green BIM: successful sustainable design with building information modeling, John Wiley & Sons, 2008. ISBN.

[31] B. Hardin, D. McCool, BIM and construction management: proven tools, methods, and workflows, John Wiley & Sons, 2015. ISBN.

[32] R.S. Weygant, BIM content development: standards, strategies, and best practices, John Wiley & Sons, 2011. ISBN: 1118030478, 9781118030479.

[33] P. Teicholz, IFMA, BIM for Facility Managers, 2013. ISBN: 1118417623, 9781118417621.

[34] C. Samaras, A. Haddad, C.A. Grammich, K.W. Webb, Obtaining Life-cycle Cost-effective Facilities in the Department of Defense, Rand Corporation, 2013. ISBN: 0833080008.

[35] Y.P. Gupta, Life cycle cost models and associated uncertainties, Electronic Systems Effectiveness and Life Cycle Costing, Springer, 1983, pp. 535-549.

[36] Y. Asiedu, P. Gu, Product life cycle cost analysis: state of the art review, International journal of production research 36 (4) (2010) 883-908.

[37] National Research Council, Pay now or pay later: controlling cost of ownership from design throughout the service life of public buildings, Available from, (1991) (Internet, cited March 29, 2019).

[38] O. Alshamrani, Evaluation of school buildings using sustainability measures and life-cycle costing technique, Concordia University, 2012.

[39] A. Boussabaine, R. Kirkham, Whole life-cycle costing: risk and risk responses, John Wiley & Sons, 2008. ISBN: 0470759151.

[40] L.A. Zadeh, Fuzzy sets, Information and control 8 (3) (1965) 338-353.

[41] R.E. Bellman, L.A. Zadeh, Decision-making in a fuzzy environment, Management science 17 (4) (1970) B-141-B-164.

[42] C.-Y. Chiu, C.S. Park, Fuzzy cash flow analysis using present worth criterion, The Engineering Economist 39 (2) (1994) 113-138.

[43] C. Kahraman, D. Ruan, E. Tolga, Capital budgeting techniques using discounted fuzzy versus probabilistic cash flows, Information Sciences 142 (1-4) (2002) 57-76.

[44] M.G. Abdel-Kader, D. Dugdale, Evaluating investments in advanced manufacturing technology: A fuzzy set theory approach, The British Accounting Review 33 (4) (2001) 455-489.

[45] S.K. Fuller, S.R. Petersen, Life cycle costing manual for the Federal Energy Management Program, NIST Handbook 135 (1995).

[46] S. Fuller, Life-cycle Cost Analysis (LCCA) Available from: https://www.wbdg.org/resources/life-cycle-cost-analysis-lcca, (2016) (Internet, cited March 29, 2019).

[47] Federal Energy Management Program (FEMP), Methodology and Procedures for Life Cycle Cost Analyses (10 CFR Part 436, Subpart A), Available from: https://www.law.cornell.edu/cfr/text/10/part-436/subpart-A, (1990) (Internet, cited March 29, 2019).

[48] ASCE, Maximizing the Value of Investments Using Life Cycle Cost Analysis, Available from: https://www.asce.org/uploadedFiles/Issues_and_Advocacy/Our_Initiatives/Infrastructure/Content_Pieces/asce-eno-life-cycle-report.pdf, (2014) (Internet, cited March 29, 2019).

[49] A.S. Rushing, J.D. Kneifel, B.C. Lippiatt, Energy price indices and discount factors for life-cycle cost analysis, 2012, US Department of Commerce, National Institute of Standards and Technology, 2012. ISBN.

[50] A. Dell'Isola, S.J. Kirk, Life cycle costing for facilities, RSMeans, 2003. ISBN: 0876297025.

[51] D. Abate, M. Towers, R. Dotz, L. Romani, J. Miller, The Whitestone facility maintenance and repair cost reference 2014-2015, Available from, (2014) (Internet, cited March 29, 2019).

[52] L. Romani, D. Abate, J. Miller, R. Dotz, The Whitestone facility operation cost reference 2014-2015, Available from, (2014) (Internet, cited March 29, 2019).

[53] M. Clift, K. Bourke, Study on whole life costing, CRC London, 1999. ISBN: 1860812805.

[54] J. Bull, The way ahead for life cycle costing in the construction industry, Life Cycle Costing for Construction, Routledge, 2003, pp. 159-168.

[55] A.L. Samuel, Some studies in machine learning using the game of checkers, IBM Journal of research and development 3 (3) (1959) 210-229.

[56] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. ISBN: 9780387310732.

[57] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of machine learning, MIT press, 2012. ISBN: 0262304732.

[58] O.S. Alshamrani, Construction cost prediction model for conventional and sustainable college buildings in North America, Journal of Taibah University for Science 11 (2) (2017) 315-323. 10.1016/j.jtusci.2016.01.004.

[59] R. Jafarzadeh, S. Wilkinson, V. Gonzalez, J.M. Ingham, G.G. Amiri, Predicting Seismic Retrofit Construction Cost for Buildings with Framed Structures Using Multilinear Regression Analysis, Journal of Construction Engineering and Management 140 (3) (2014). 10.1061/(asce)co.1943-7862.0000750.

[60] D.J. Lowe, M.W. Emsley, A. Harding, Predicting construction cost using multiple regression techniques, Journal of Construction Engineering and Management 132 (7) (2006) 750-758.

[61] T.M. Zayed, D.W. Halpin, Productivity and cost regression models for pile construction, Journal of Construction Engineering and Management 131 (7) (2005) 779-789.

[62] H. Li, Q. Shen, P.E. Love, Cost modelling of office buildings in Hong Kong: an exploratory study, Facilities 23 (9/10) (2005) 438-452.

[63] S.M. Trost, G.D. Oberlender, Predicting accuracy of early cost estimates using factor analysis and multivariate regression, Journal of Construction Engineering and Management 129 (2) (2003) 198-204.

[64] S.Z. Dogan, D. Arditi, H.M. Gunaydin, Using decision trees for determining attribute weights in a case-based model of early cost prediction, Journal of Construction Engineering and Management-Asce 134 (2) (2008) 146-152. 10.1061/(asce)0733-9364(2008)134:2(146).

[65] S.Z. Dogan, D. Arditi, H.M. Gunaydin, Determining attribute weights in a CBR model for early cost prediction of structural systems, Journal of Construction Engineering and Management-Asce 132 (10) (2006) 1092-1098. 10.1061/(asce)0733-9364(2006)132:10(1092).

[66] O. Dursun, C. Stoy, Conceptual Estimation of Construction Costs Using the Multistep Ahead Approach, Journal of Construction Engineering and Management 142 (9) (2016). 10.1061/(asce)co.1943-7862.0001150.

[67] M.-Y. Cheng, H.-C. Tsai, E. Sudjono, Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry, Expert Systems with Applications 37 (6) (2010) 4224-4231.

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

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

[70] A.A. Aibinu, D. Dassanayake, T.-K. Chan, R. Thangaraj, Cost estimation for electric light and power elements during building design: A neural network approach, Engineering, Construction and Architectural Management 22 (2) (2015) 190-213.

[71] N. Sajadfar, Y.S. Ma, A hybrid cost estimation framework based on feature-oriented data mining approach, Advanced Engineering Informatics 29 (3) (2015) 633-647. 10.1016/j.aei.2015.06.001.

[72] C. Zhang, L.W. Cao, A. Romagnoli, On the feature engineering of building energy data mining, Sustainable Cities and Society 39 (2018) 508-518. 10.1016/j.scs.2018.02.016.

[73] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable & Sustainable Energy Reviews 81 (2018) 1192-1205. 10.1016/j.rser.2017.04.095.

[74] B.R. Park, E.J. Choi, J. Hong, J.H. Lee, J.W. Moon, Development of an energy cost prediction model for a VRF heating system, Applied Thermal Engineering 140 (2018) 476-486. 10.1016/j.applthermaleng.2018.05.068.

[75] E. Sala-Cardoso, M. Delgado-Prieto, K. Kampouropoulos, L. Romeral, Activity-aware HVAC power demand forecasting, Energy and Buildings 170 (2018) 15-24. 10.1016/j.enbuild.2018.03.087.

[76] J. Moon, J. Park, E. Hwang, S. Jun, Forecasting power consumption for higher educational institutions based on machine learning, Journal of Supercomputing 74 (8) (2018) 3778-3800. 10.1007/s11227-017-2022-x.

[77] E. Mocanu, P.H. Nguyen, W.L. Kling, M. Gibescu, Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning, Energy and Buildings 116 (2016) 646-655. 10.1016/j.enbuild.2016.01.030.

[78] R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor, 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 (2014) 168-178. 10.1016/j.apenergy.2014.02.057.

[79] J.S. Chou, N.T. Ngo, Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns, Applied Energy 177 (2016) 751-770. 10.1016/j.apenergy.2016.05.074.

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

[81] C. Ji, T. Hong, K. Jeong, S.B. Leigh, A model for evaluating the environmental benefits of elementary school facilities, Journal of Environmental Management 132 (2014) 220-229. 10.1016/j.jenvman.2013.11.022.

[82] S. Bouktif, A. Fiaz, A. Ouni, M.A. Serhani, Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches, Energies 11 (7) (2018). 10.3390/en11071636.

[83] C. Robinson, B. Dilkina, J. Hubbs, W.W. Zhang, S. Guhathakurta, M.A. Brown, R.M. Pendyala, Machine learning approaches for estimating commercial building energy consumption, Applied Energy 208 (2017) 889-904. 10.1016/j.apenergy.2017.09.060.

[84] D. Geysen, O. De Somer, C. Johansson, J. Brage, D. Vanhoudt, Operational thermal load forecasting in district heating networks using machine learning and expert advice, Energy and Buildings 162 (2018) 144-153. 10.1016/j.enbuild.2017.12.042.

[85] S. Idowu, S. Saguna, C. Ahlund, O. Schelen, Applied machine learning: Forecasting heat load in district heating system, Energy and Buildings 133 (2016) 478-488. 10.1016/j.enbuild.2016.09.068.

[86] C. Li, Z.X. Ding, D.B. Zhao, J.Q. Yi, G.Q. Zhang, Building Energy Consumption Prediction: An Extreme Deep Learning Approach, Energies 10 (10) (2017). 10.3390/en10101525.

[87] E.S. Neely, R. Neathammer, Life-cycle maintenance costs by facility use, Journal of Construction Engineering and Management-Asce 117 (2) (1991) 310-320. 10.1061/(asce)0733-9364(1991)117:2(310).

[88] C.P. Au-Yong, A.S. Ali, F. Ahmad, Prediction cost maintenance model of office building based on condition-based maintenance, Maintenance and Reliability - 16 (- 2) (2014) - 324.

[89] H. Krstić, S. Marenjak, Maintenance and operation costs model for university buildings, Technical Gazette - 24 (2017) - 200.

[90] K.J. Tu, Y.W. Huang, 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) (2013) 242-250.

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

[92] R. Liu, R.R.A. Issa, Survey: Common Knowledge in BIM for Facility Maintenance, Journal of Performance of Constructed Facilities (2015). 10.1061/.

[93] Y.C. Lin, Y.C. Su, Developing mobile- and BIM-based integrated visual facility maintenance management system, ScientificWorldJournal 2013 (2013) 124249. 10.1155/2013/124249.

[94] W. tShen, Q. Hao, Y. Xue, A loosely coupled system integration approach for decision support in facility management and maintenance, Automation in Construction 25 (2012) 41-48. https://doi.org/10.1016/j.autcon.2012.04.003.

[95] J. Irizarry, M. Gheisari, G. Williams, K. Roper, Ambient intelligence environments for accessing building information: A healthcare facility management scenario, Facilities 32 (3/4) (2014) 120-138. http://dx.doi.org/10.1108/F-05-2012-0034.

[96] S. Lee, Ö. Akin, Augmented reality-based computational fieldwork support for equipment operations and maintenance, Automation in Construction 20 (4) (2011) 338-352. DOI:10.1016/j.autcon.2010.11.004.

[97] R. Liu, R. Issa, 3D visualization of sub-surface pipelines in connection with the building utilities: Integrating GIS and BIM for facility management, Computing in Civil Engineering, 2012, pp. 341-348. https://doi.org/10.1061/9780784412343.0043.

[98] I. Motawa, A. Almarshad, A knowledge-based BIM system for building maintenance, Automation in Construction 29 (0) (2013) 173-182. http://dx.doi.org/10.1016/j.autcon.2012.09.008.

[99] C. Koch, M. Neges, M. König, M. Abramovici, Natural markers for augmented reality-based indoor navigation and facility maintenance, Automation in Construction 48 (2014) 18-30. https://doi.org/10.1016/j.autcon.2014.08.009.

[100] Y.C. Lin, Y.C. Su, Y.P. Chen, Developing mobile BIM/2D barcode-based automated facility management system, ScientificWorldJournal 2014 (2014) 374735. 10.1155/2014/374735.

[101] X. Yang, S. Ergan, Design and Evaluation of an Integrated Visualization Platform to Support Corrective Maintenance of HVAC Problem-Related Work Orders, Journal of Computing in Civil Engineering 30 (3) (2016). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000510.

[102] A. Golabchi, M. Akula, V. Kamat, Automated building information modeling for fault detection and diagnostics in commercial HVAC systems, Facilities 34 (3-4) (2016) 233-246. https://doi.org/10.1108/F-06-2014-0050.

[103] F. Shalabi, Y. Turkan, IFC BIM-Based Facility Management Approach to Optimize Data Collection for Corrective Maintenance, Journal of Performance of Constructed Facilities (2016) 04016081.

[104] A. Motamedi, A. Hammad, Y. Asen, Knowledge-assisted BIM-based visual analytics for failure root cause detection in facilities management, Automation in Construction 43 (0) (2014) 73-83. http://dx.doi.org/10.1016/j.autcon.2014.03.012.

[105] J. Zhou, P. Love, J. Matthews, B. Carey, C. Sing, Object-oriented model for life cycle management of electrical instrumentation control projects, Automation in Construction 49 (2015) 142-151.

[106] U. Isikdag, J. Underwood, G. Aouad, An investigation into the applicability of building information models in geospatial environment in support of site selection and fire response management processes, Advanced Engineering Informatics 22 (4) (2008) 504-519.

[107] F. Forns-Samso, T. Laine, B. Hensel, Building information modeling supporting facilities management, Proceedings ECPPM Ework and Ebusiness in Architecture, Engineering and Construction, 2012, pp. 51-57. ISBN:9780415621281.

[108] R. Liu, R.R.A. Issa, Automatically Updating Maintenance Information from a BIM Database, in: R.R. Issa, I. Flood (Eds.), Proceedings of the 2012 ASCE International Conference on Computing in Civil Engineering, American Society of Civil Engineers Tech. Council Comput. Inf. Technol. American Soc. Civil Eng. Tech. Council Comput. Inf. Technol. American Soc. Civil Eng., Place of Publication: Reston, VA, USA; Clearwater Beach, FL, USA. Country of

Publication: USA., 2012, pp. 373-380. https://doi.org/10.1061/9780784412343.0047.

[109] General Services Administration, BIM Guide For Facility Management Available from: https://www.gsa.gov/cdnstatic/largedocs/BIM_Guide_Series_Facility_Management.pdf, (2011) (Internet, cited March 29, 2019).

[110] W. Shen, Q. Hao, H. Mak, J. Neelamkavil, H. Xie, J. Dickinson, R. Thomas, A. Pardasani, H. Xue, Systems integration and collaboration in architecture, engineering, construction, and facilities management: A review, Advanced Engineering Informatics 24 (2) (2010) 196-207.

[111] N. Bakis, G. Aouad, M. Kagioglou, Towards distributed product data sharing environments - Progress so far and future challenges, Automation in Construction 16 (5) (2007) 586-595. https://doi.org/10.1016/j.autcon.2006.10.002.

[112] Z. Shen, L. Jiang, An augmented 3D iPad mobile application for communication, collaboration, and learning (CCL) of building MEP systems, Computing in Civil Engineering (2012), 2012, pp. 204-212.

[113] A. Motamedi, M.M. Soltani, A. Hammad, Localization of RFID-equipped assets during the operation phase of facilities, Advanced Engineering Informatics 27 (4) (2013) 566-579. http://dx.doi.org/10.1016/j.aei.2013.07.001.

[114] A. Papapostolou, H. Chaouchi, RFID-assisted indoor localization and the impact of interference on its performance, Journal of Network and Computer Applications 34 (3) (2011) 902-913. https://doi.org/10.1016/j.jnca.2010.04.009.

[115] M.H. Dawood, BIM based optimal life cycle cost of sustainable house framework, 2016 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016, 2016, pp. 279-283. ISBN.

[116] F. Jalaei, A. Jrade, M. Nassiri, INTEGRATING DECISION SUPPORT SYSTEM (DSS) AND BUILDING INFORMATION MODELING (BIM) TO OPTIMIZE THE SELECTION OF SUSTAINABLE BUILDING COMPONENTS, Journal of Information Technology in Construction 20 (2015) 399-420.

[117] S. Liu, X. Meng, C. Tam, Building information modeling based building design optimization for sustainability, Energy and Buildings 105 (2015) 139-153.

[118] M. Marzouk, M. Metawie, M. Hisham, I. Al-Sulahi, M. Kamal, K. Al-Gahtani, Modeling sustainable building materials in Saudi Arabia, Computing in Civil and Building Engineering (2014), 2014, pp. 1546-1553.

[119] M. Marzouk, S. Azab, M. Metawie, Framework for Sustainable Low-Income Housing Projects Using Building Information Modeling, Journal of Environmental Informatics 28 (1) (2016) 25-38. 10.3808/jei.201600332.

[120] M. MARZOUK, M. HISHAM, M. ELSHEIKH, K. AL-GAHTANI, Building information model for selecting environmental building materials, New Developments in Structural Engineering and Construction, Pages Su-11-268, Singapore, Singapore (2013).

[121] C. Fu, G. Aouad, A.M. Ponting, A. Lee, S. Wu, IFC implementation in lifecycle costing, China Architecture & Building Press, Beijing, 2003. ISBN: 7-112-02100-6.

[122] S. Dawood, R. Lord, N. Dawood, Development of a visual whole life-cycle energy assessment framework for built environment, Proceedings - Winter Simulation Conference, 2009, pp. 2653-2663. ISBN.

[123] P.H. Chen, C. Long, Y.Y. Chen, A BIM-based framework for selection of cost-effective green building design, Proceedings of the 13th East Asia-Pacific Conference on Structural Engineering and Construction, EASEC 2013, 2013. ISBN.

[124] O. Hosny, A. Elhakeem, Design buildings optimally: A lifecycle assessment approach, ISEC 2013 - 7th International Structural Engineering and Construction Conference: New Developments in Structural Engineering and Construction, 2013, pp. 1367-1372. ISBN.

[125] M. Nour, O. Hosny, A. Elhakeem, A BIM BASED APPROACH FOR CONFIGURING BUILDINGS' OUTER ENVELOPE ENERGY SAVING ELEMENTS, Journal of Information Technology in Construction 20 (2015) 173-192.

[126] Y.S. Shin, K. Cho, BIM Application to Select Appropriate Design Alternative with Consideration of LCA and LCCA, Mathematical Problems in Engineering (2015). 10.1155/2015/281640.

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

[128] M. Vukomanovic, M. Kararic, Model for cost prediction of prefabricated housing, Tehnicki Vjesnik-Technical Gazette 16 (3) (2009) 39-43.

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

[130] K. Amasyali, N. El-Gohary, Building lighting energy consumption prediction for supporting energy data analytics, in: O. Chong, K. Parrish, P. Tang, D. Grau, J. Chang (Eds.), Icsdec 2016 - Integrating Data Science, Construction and Sustainability, Vol. 145, 2016, pp. 511-517. 10.1016/j.proeng.2016.04.036.

[131] S. Mahmoud, M.F. Khamidi, A. Idrus, O.A.L. Ashola, Development of Maintenance Cost Prediction Model for Heritage Buildings, Jurnal Teknologi 74 (2) (2015) 51-57.

[132] N.A. Sallehta, M. Yakinl, K. Ismail, Y. Talib, Prelimanary Investigation on The Factors That Influencing The Maintenance Cost of Apartment, in: S.N.B. Kamaruzzaman, A.S.B. Ali, N.F.B. Azmi, S.J.L. Chua (Eds.), 4th International Building Control Conference 2016, Vol. 66, E D P Sciences, Cedex A, 2016. 10.1051/matecconf/20166600046.

[133] T.P. Lo, S.J. Guo, C.T. Chen, Application of the exponential grey model on the maintenance cost prediction for a large scale hospital, International Journal of Strategic Property Management 15 (4) (2011) 379-392. 10.3846/1648715x.2011.633775.

[134] K.J. Tu, Y.W. Huang, C.L. Lu, K.H. Chu, Predicting the operation and maintenance costs of apartment buildings at preliminary design stage: Comparing statistical regression and artificial neural network methods, CME 25 Conference Construction Management and Economics, 2007, pp. - 1483.

[135] C. Bahr, K. Lennerts, Quantitative validation of budgeting methods and suggestion of a new calculation method for the determination of maintenance costs, Journal of Facilities Management - 8 (- 1) (2010) - 63.

[136] G.A. Ehlers, R.D. Howerton, G.E. Speegle, Engery management and building automation system, in: U.S.P.a.T. Office (Ed.), 1996.

[137] J. Figueiredo, J. Martins, Energy production system management–renewable energy power supply integration with building automation system, Energy Conversion and Management 51 (6) (2010) 1120.

[138] M.J. Donnell, P.M. Herbst, T.M. Nitsch, R.E. Fransen, Building automation system, in: U.S.P.a.T. Office (Ed.), 2010.

[139] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy and Buildings 75 (2014) 109-118.

[140] ClimateTechWiki, Building Energy Management Systems (BEMS), Available from: https://www.climatetechwiki.org/technology/jiqweb-bems, (2019) (Internet, cited March 29, 2019).

[141] IEA (International Energy Agency), Technical Synthesis Report: A Summary of Annexes 16 & 17 Building Energy Management Systems, Available from: http://www.iea-ebc.org/Data/publications/EBC_Annex_16-17_tsr.pdf, (1997) (Internet, cited March 29, 2019).

[142] H. Doukas, K.D. Patlitzianas, K. Iatropoulos, J. Psarras, Intelligent building energy management system using rule sets, Building and Environment 42 (10) (2007) 3562-3569.

[143] D. Sapp, Computerized Maintenance Management Systems (CMMS), (2013).

[144] A. Lewis, D. Riley, A. Elmualim, Defining high performance buildings for operations and maintenance, International Journal of Facility Management 1 (2) (2010).

[145] Planon, Integrated Workplace Management Systems (IWMS), Available from: https://planonsoftware.com/us/whats-new/knowledge-center/glossary/iwms/, (2019) (Internet, cited March 29, 2019).

[146] G.S.A. GSA, BIM Guide For Facility Management 2011, pp. 1-82.

[147] X. Gao, P. Pishdad-Bozorgi, BIM-enabled Facilities Operation and Maintenance: A Review, (39) (2019) 227–247. doi.org/10.1016/j.aei.2019.01.005.

[148] P. Pishdad-Bozorgi, X. Gao, C. Eastman, A.P. Self, Planning and developing facility management-enabled building information model (FM-enabled BIM), Automation in Construction 87 (2018) 22-38. doi.org/10.1016/j.autcon.2017.12.004.

[149] X. Gao, P. Pishdad-Bozorgi, Past, Present, and Future of BIM-Enabled Facilities Operation and Maintenance, Construction Research Congress 2018, 2018, pp. 51-61. ISBN:9780784481264.

[150] buildingSMART, IFC Overview summary, Available from: http://www.buildingsmart-tech.org/specifications/ifc-overview, (2017) (Internet, cited March 29, 2019).

[151] buildingSMART, Alphabetical list of entities, Available from: http://www.buildingsmart-tech.org/ifc/IFC2x3/TC1/html/alphabeticalorder_entities.htm, (2017) (Internet, cited March 29, 2019).

[152] Green Building XML (gbXML) Schema Inc., gbXML, Available from: www.gbxml.org, (2018) (Internet, cited March 29, 2019).

[153] Open Geospatial Consortium, CityGML, Available from: https://www.citygml.org/, (2018) (Internet, cited March 29, 2019).

[154] bacnetwiki.com, BACnet Wiki, Available from: http://www.bacnetwiki.com/wiki/index.php?title=Main_Page, (2017) (Internet, cited March 29, 2019).

[155] Johnson Controls Inc., Metasys® Building Automation System, Available from: http://www.johnsoncontrols.com/buildings/building-management/building-automation-systems-bas, (2018) (Internet, cited March 29, 2019).

[156] Lawrence_Berkeley_National_Laboratory, Building Controls Virtual Test Bed, Available from: https://simulationresearch.lbl.gov/bcvtb, (2017) (Internet, cited March 29, 2019).

[157] W. Solihin, BIMRL-DBETL, Available from: https://github.com/BIMRL-Team/BIMRL-DBETL, (2018) (Internet, cited March 29, 2019).

[158] A. Ward, C. Benghi, M. Černý, S. Lockley, xBIM, Available from: https://github.com/xBimTeam, (2018) (Internet, cited March 29, 2019).

[159] Technische Universität München, 3D City Database, Available from: https://www.3dcitydb.org/3dcitydb/3dcitydbhomepage/, (2018) (Internet, cited March 29, 2019).

[160] Technische Universität München, 3D City Database for CityGML, Available from: http://www.3dcitydb.org/3dcitydb/fileadmin/TUM_Workshop/Documents/Tutorial.pdf, (2016) (Internet, cited March 29, 2019).

[161] BACnet XML Working Group, BACnet XML Applications, Available from: http://www.bacnet.org/WG/XML/, (2018) (Internet, cited March 29, 2019).

[162] buildingSMART, ifcXML, Available from: http://www.buildingsmart-tech.org/specifications/ifcxml-releases, (2018) (Internet, cited March 29, 2019).

[163] Open Geospatial Consortium, CityGML XML Schema, (2018).

[164] ALTOVA, MapForce, Available from: https://www.altova.com/mapforce, (2018) (Internet, cited March 29, 2019).

[165] D.C. Montgomery, E.A. Peck, G.G. Vining, Introduction to linear regression analysis, John Wiley & Sons, 2012. ISBN: 0470542810.

[166] E. Alpaydin, Introduction to machine learning, MIT press, 2014. ISBN: 0262325756.

[167] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and Computing 14 (3) (2004) 199-222.

[168] M. Gelfusa, A. Malizia, A. Murari, S. Parracino, M. Lungaroni, E. Peluso, J. Vega, L. DeLeo, C. Perrimezzi, P. Gaudio, First attempts at measuring widespread smoke with a mobile LiDAR system, (2015).

[169] O. Chapelle, V. Vapnik, Model selection for support vector machines, Advances in neural information processing systems, 2000, pp. 230-236. ISBN.

[170] A.J. Smola, Regression estimation with support vector learning machines, Master's thesis, Technische Universität München, 1996.

[171] Y.B. Dibike, S. Velickov, D. Solomatine, Support vector machines: Review and applications in civil engineering, Proc. of the joint workshop on Applications of AI in Civil Engineering, Cottbus-2000, Germany, Citeseer, 2000. ISBN.

[172] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician 46 (3) (1992) 175-185.

[173] J. Gleick, The information: A history, a theory, a flood, HarperCollins, UK, 2011. ISBN.

[174] J.D. Cryer, K.S. Chan, Time Series Analysis With Applications in R, 2009. ISBN: 0306-7734.

[175] T.B. Fomby, Deterministic trend/deterministic season model, Available from: http://faculty.smu.edu/tfomby/eco5375/data/notes/Det%20Time%20Trend%20Model.pdf, (2008) (Internet, cited March 29, 2019).

[176] O. Hellberg, Backcasting swedish industrial production, Workshop on Survey Sampling Theory and Methodology, Vilnius, Lithuania, 2010.

[177] S.S. Haykin, Neural networks and learning machines, Pearson Upper Saddle River, 2009. ISBN.

[178] F. Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, Available from, (1961) (Internet, cited March 29, 2019).

[179] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Available from, (1985) (Internet, cited March 29, 2019).

[180] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press Cambridge, 2016. ISBN.

[181] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of control, signals and systems 2 (4) (1989) 303-314.

[182] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, Springer, 2013. ISBN.

[183] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & Electrical Engineering 40 (1) (2014) 16-28.

[184] N. Sánchez-Maroño, A. Alonso-Betanzos, M. Tombilla-Sanromán, Filter methods for feature selection–a comparative study, International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2007, pp. 178-187. ISBN.

[185] K. Kira, L.A. Rendell, A practical approach to feature selection, Machine Learning Proceedings 1992, Elsevier, 1992, pp. 249-256.

[186] M.A. Hall, Correlation-based feature selection for machine learning, (1999).

[187] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 856-863. ISBN.

[188] Z. Zhao, H. Liu, Searching for Interacting Features, ijcai, Vol. 7, 2007, pp. 1156-1161. ISBN.

[189] R. Abraham, J.B. Simha, S.S. Iyengar, Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining, International Journal of Computational Intelligence Research 5 (2) (2009) 116-129.

[190] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern recognition letters 15 (11) (1994) 1119-1125.

[191] J. Reunanen, Overfitting in making comparisons between variable selection methods, Journal of Machine Learning Research 3 (Mar) (2003) 1371-1382.

[192] P. Somol, P. Pudil, J. Novovičová, P. Paclık, Adaptive floating search methods in feature selection, Pattern recognition letters 20 (11-13) (1999) 1157-1163.

[193] Y. Sun, C. Babbs, E. Delp, A comparison of feature selection methods for the detection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm, 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, 2006, pp. 6532-6535. ISBN:0780387414.

[194] D.E. Goldberg, J.H. Holland, Genetic algorithms and machine learning, Machine learning 3 (2) (1988) 95-99.

[195] L.J. Eshelman, The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination, Foundations of genetic algorithms, Vol. 1, Elsevier, 1991, pp. 265-283.

[196] O. Cordón, S. Damas, J. Santamaría, Feature-based image registration by means of the CHC evolutionary algorithm, Image and Vision Computing 24 (5) (2006) 525-533.

[197] C.-J. Tu, L.-Y. Chuang, J.-Y. Chang, C.-H. Yang, Feature selection using PSO-SVM, International Journal of Computer Science (2007).

[198] T.N. Lal, O. Chapelle, J. Weston, A. Elisseeff, Embedded methods, Feature extraction, Springer, 2006, pp. 137-165.

[199] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (Mar) (2003) 1157-1182.

[200] G. Vanwinckelen, H. Blockeel, On estimating model accuracy with repeated cross-validation, BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning, 2012, pp. 39-44. ISBN:9461970447.

[201] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Statistics surveys 4 (2010) 40-79.

[202] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Ijcai, Vol. 14, Montreal, Canada, 1995, pp. 1137-1145. ISBN.

[203] Bureau of Labor Statistics, Consumer Price Index, Available from: www.bls.gov/cpi, (2018) (Internet, cited March 29, 2019).

[204] Y. Hu, D. Castro-Lacouture, Clash Relevance Prediction Based on Machine Learning, Journal of Computing in Civil Engineering 33 (2) (2018) 04018060.

[205] M. Cristani, R. Cuel, A comprehensive guideline for building a domain ontology from scratch, proceeding of" International Conference on Knowledge Management (I-KNOW'04)", Graz, Austria, 2004. ISBN.

[206] N.F. Noy, D.L. McGuinness, Ontology development 101: A guide to creating your first ontology, Stanford knowledge systems laboratory technical report KSL-01-05 and …, 2001.

[207] R. Studer, V.R. Benjamins, D. Fensel, Knowledge engineering: principles and methods, Data and knowledge engineering 25 (1) (1998) 161-198.

[208] H. Burkhardt, B. Smith, Handbook of metaphysics and ontology, (1991).

[209] D. Fensel, Ontologies: A silver bullet for knowledge management and electronic-commerce, Berlin: Spring-Verlag 143 (2000).

[210] R.P. Charette, H.E. Marshall, UNIFORMAT II elemental classification for building specifications, cost estimating, and cost analysis, US Department of Commerce, Technology Administration, National Institute of Standards and Technology, 1999. ISBN.

[211] Construction Specifications Institute (CSI), MasterFormat 2016 Edition: Numbers and Titles, Available from: www.edmca.com/media/35207/masterformat-2016.pdf, (2016) (Internet, cited March 29, 2019).

[212] OWL Working Group, Web Ontology Language (OWL), Available from: www.w3.org/OWL, (2019) (Internet, cited March 29, 2019).

[213] Stanford Center for Biomedical Informatics Research, protégé, Available from: protege.stanford.edu, (2019) (Internet, cited March 29, 2019).

[214] gbXML Schema Inc., gbXML, Available from: www.gbxml.org, (2018) (Internet, cited March 29, 2019).

[215] openbimstandards.org, ifcOWL, Available from: openbimstandards.org/standards/ifcowl/, (2019) (Internet, cited March 29, 2019).

[216] OCCS Development Committee Secretariat, OmniClass: a strategy for classifying the built environment, Available from: www.omniclass.org/, (2019) (Internet, cited March 29, 2019).

[217] R.G. Kreider, An ontology of the uses of building information modeling, (2013).

[218] The Modbus Organization, Modbus Technical Resources, Available from: http://modbus.org/specs.php, (2018) (Internet, cited March 29, 2019).

[219] LonMark International., LonWorks Technology Achieves ISO/IEC Standardization Available from: https://www.lonmark.org/news_events/press/2008/1208_iso_standard, (2008) (Internet, cited March 29, 2019).

[220] AssetWorks, AiM Operations & Maintenance (O&M), Available from: https://www.assetworks.com/iwms/aim-oandm/, (2018) (Internet, cited March 29, 2019).

[221] The INSITE Consortium, INSITE Available from: www.insite.org, (2019) (Internet, cited March 29, 2019).

[222] Georgia Institute of Technology, Georgia Tech Capital Planning & Space Management System, Available from: https://tableau.gatech.edu/t/CPSM/views/BuildingInformationDashboard/Intro?%3Aembed=y&%3AshowShareOptions=true&%3Adisplay_count=no&%3AshowVizHome=no, (2019) (Internet, cited March 29, 2019).

[223] Tableau, Tableau, Available from: www.tableau.com, (2019) (Internet, cited March 29, 2019).

[224] Georgia Institution of Technology, The Facility Data of Georgia Tech, Available from: http://ionsvr2.fac.gatech.edu/ion/default.aspx?dgm=//IONSVR2/ION-Ent/config/diagrams/ud/network.dgm&node=&logServerName=QUERYSERVER.IONSVR2&logServerHandle=327952, (2018) (Internet, cited March 29, 2019).

[225] Ntrepid Corporation, Ion data grabber, Available from: ion.ntrepidcorp.com, (2018) (Internet, cited March 29, 2019).

[226] Aerospace System Design Laboratory, EnergyWatch for Georgia Institute of Technology, Available from: energywatch.gatech.edu, (2019) (Internet, cited March 29, 2019).

[227] openrefine.org, OpenRefine, Available from: http://openrefine.org/, (2018) (Internet, cited March 29, 2019).

[228] Microsoft, Microsoft R, Available from: mran.microsoft.com, (2019) (Internet, cited March 29, 2019).

[229] US Inflation Calculator, Historical Inflation Rates: 1914-2018, Available from: https://www.usinflationcalculator.com/inflation/historical-inflation-rates/, (2019) (Internet, cited March 29, 2019).

[230] Bureau of Labor Statistics, Average Energy Prices, Atlanta-Sandy Springs-Roswell – September 2018, Available from: www.bls.gov/regions/southeast/news-release/averageenergyprices_atlanta.htm, (2018) (Internet, cited March 29, 2019).

[231] Bureau of Labor Statistics, Current Employment Statistics - CES (National) Available from: www.bls.gov/ces/, (2019) (Internet, cited March 29, 2019).

[232] R-project.org, Documentation for package 'stats' version 3.6.0, Available from: stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html, (2019) (Internet, cited March 29, 2019).

[233] A. Beygelzimer, S. Kakadet, J. Langford, S. Arya, D. Mount, S. Li, FNN: Fast Nearest Neighbor Search Algorithms and Applications, Available from: CRAN.R-project.org/package=FNN, (2019) (Internet, cited March 29, 2019).

[234] L. Breiman, A. Cutler, A. Liaw, M. Wiener, randomForest: Breiman and Cutler's Random Forests for Classification and Regression, Available from: CRAN.R-project.org/package=randomForest, (2019) (Internet, cited March 29, 2019).

[235] Probability Theory Group, e1071: Misc Functions of the Department of Statistics, Available from: cran.r-project.org/web/packages/e1071/index.html, (2019) (Internet, cited March 29, 2019).

[236] J. Allaire, F. Chollet, Y. Tang, D. Falbel, W. Van Der Bijl, M. Studer, S. Keydana, keras: R Interface to 'Keras', Available from: CRAN.R-project.org/package=keras, (2019) (Internet, cited March 29, 2019).

[237] R. Rahman, MultivariateRandomForest: Models Multivariate Cases Using Random Forests, Available from: CRAN.R-project.org/package=MultivariateRandomForest, (2019) (Internet, cited March 29, 2019).

[238] X. Gao, S. Tang, P. Pishdad-Bozorgi, D. Shelden, Foundational Research in Integrated Building Internet of Things (IoT) Data Standards, Available from:

https://cdait.gatech.edu/sites/default/files/georgia_tech_cdait_research_report_on_integrated_building_-_iot_data_standards_september_2018_final.pdf, (2018) (Internet, cited March 29, 2019).

[239] UN DESA, World urbanization prospects: The 2014 revision, United Nations Department of Economics and Social Affairs, Population Division: New York, NY, USA (2015).

[240] A. Tascikaraoglu, Evaluation of spatio-temporal forecasting methods in various smart city applications, Renewable and Sustainable Energy Reviews 82 (2018) 424-435.

[241] ISO/IEC_JTC_1, Smart cities Preliminary Report 2014, Available from: https://www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/smart_cities_report-jtc1.pdf, (2014) (Internet, cited March 29, 2019).

[242] A. Buckman, M. Mayfield, S. BM Beck, What is a smart building?, Smart and Sustainable Built Environment 3 (2) (2014) 92-109.

[243] NIST, Cyber-Physical Systems, Available from: www.nist.gov/el/cyber-physical-systems, (2017) (Internet, cited March 29, 2019).

[244] The Cyber Physical Systems Public Working Group, Framework for Cyber-Physical Systems, Available from: https://pages.nist.gov/cpspwg/, (2016) (Internet, cited March 29, 2019).

[245] NIST, The Internet of Things-Enabled Smart City Framework, 2017.

[246] International Electrotechnical Commission (IEC), Electrotechnical aspects of Smart Cities, Available from: http://www.iec.ch/dyn/www/f?p=103:186:0::::FSP_ORG_ID:13073, (2017) (Internet, cited March 29, 2019).

[247] Institute of Electrical and Electronics Engineers (IEEE), IEEE Smart Cities, Available from: https://smartcities.ieee.org/, (2017) (Internet, cited March 29, 2019).

[248] Digital Building Laboratory, DBL SmartCity, Available from: dbl-smartcity.design.gatech.edu, (2019) (Internet, cited March 29, 2019).