ISSN 2044-124X
Volume 9 Number 4 2019

# Built Environment Project and Asset Management

## Data analytics and big data in construction project and asset management

Guest Editors: Ajibade A. Aibinu,
Fernando Koch and
S. Thomas Ng

# Editorial boards

Built Environment Project and
Asset Management
Vol. 9 No. 4, 2019
p. 473
© Emerald Publishing Limited
2044-124X

# Guest editorial

## Data analytics and big data in construction project and asset management

The increasing adoption of digital technology and the rapid proliferation of data have spurred the application of data analytics and big data to drive smart project and asset management. We are likely to see the rise of new approaches to information management and data usage within the architecture, engineering and construction (AEC) sector. Transforming data and information into knowledge and intelligence would change the way projects and assets are managed and will facilitate optimal solutions across the sector. This special journal issue provides a forum to explore, develop and disseminate emerging concepts of data analytics and big data and their potential applications and opportunities in the AEC.

In the opening paper of this special issue, Marzouk and Enaba explores how data in building information model (BIM) can be furnished with descriptive analytics within the BIM environment to analyse construction project performance. They developed a dynamic model that helps in detecting hidden patterns and different progress attributes from construction project raw data. Their study assumes that integrating BIM and data analytics in a construction project is beneficial. They then validated this assumption on a case study project. Marzouk and Enaba study is interesting because by leveraging data embedded in BIM to gain actionable insight using analytics, the AEC sector can increase the benefit and value derived from BIM efforts on projects thereby increasing the adoption of BIM and big data.

The second paper by Farghaly, Abanda, Vidalakis and Wood investigates "the transfer of information from the BIM models to either conventional or advanced asset management platforms using Linked Data". Employing a participatory action research approach with focus group and interviews as well as prototyping, they proposed a process for generating linked data in the asset management context and its integration with BIM data. In view of the very limited application of BIM data for asset management, the process proposed in their paper can improve the data exchange from BIMs to asset management systems during handover stage and consequently improve asset management outcomes during the operation and maintenance stage. Applying the process that they have developed to a real-world case study of BIM for facilities management would provide empirical evidence of the value and challenges of BIM-linked data integration for asset management.

Amit Mitra and Kamran Munir focus on how big data can influence asset management. Drawing from interviews of leaders of digital transformation projects in three organisations that are within the insurance industry, natural gas and oil, and manufacturing industries, they argued that as asset information becomes a project deliverable, and as data increases in volume, velocity and variety; and as it is aggregated and re-used, there is need to improve semantic processors to deal with the vast array of data in variable formats. Evidently, the increasing volume and variety of asset management data will make the implementation data analytics inevitable if organisation must derive value from their asset information. Although the research is still in its infancy, it is interesting because it reminds us that the value of asset information does not lie in the volume of the data itself but in the insight that organisations can garner from the data to achieve better asset management outcomes.

Still on the application of big data and analytics, Jafari and Akhavian use data analytics models to predict the driving forces of housing prices arguing that this can be extremely useful in the built environment and real estate decision-making processes. Based on a data set of 13,771 houses, they developed a hedonic pricing method of the key factors that affect the residential housing prices across the USA. They tested and validated their model using 22 houses not used in models. The results of their study show that the main driving

Built Environment Project and
Asset Management
Vol. 9 No. 4, 2019
pp. 474-475
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-09-2019-139

forces for housing transaction price are square footage of the unit, followed by location, and number of bathrooms and bedrooms. They also highlighted that the impact of neighbourhood characteristics such as distance to open spaces and business centres on the housing prices is not as strong as the impact of housing unit characteristics and location characteristics. Perhaps these factors could vary according to housing submarkets. This needs to be explored further to better understand the conditions under which specific factors are more important than the others. Over the years, there have been many analytic models of this nature developed in the academic domain. It will be useful to understand the extent to which these models are being adopted in practice and the barriers.

Turning to review papers, Madanayake and Egbu identify the gaps and potential future research avenues in the big data research in the construction industry. Based on a systematic review of publications produced over a period of 10 years from 2007 to 2017, they concluded that although there are plenty of research in the application of big data in construction, research on the implications of big data in the overall sustainability – social, economic and environmental dimensions – is lacking. They proposed up to 15 topic areas of research needed to address the use and implications of big data within these three dimensions. This paper should be useful to researchers when setting up agenda for research in this area.

In a second systematic review of the literature, Yap, Ho and Ting examine the application of different multi-criteria decision-making (MCDM) methods for the tasks of selecting the most suitable location for energy generation, logistics, public services and retail facilities. They identified the most commonly used MCDM methods as well as the most frequently used criteria. They made a very important observation that, in the near future, as built environment implement Internet of Things sensors to capture data in various formats, we would need new MCDM methods which can account for the changing nature of the various criteria in order to determine optimal site selection outcomes. Evidently, MCDM methods can benefit from big data and analytics for better decision making.

In response to the low rate of adoption of big data and analytics in practice, Ram, Afridi and Khan argue that gaining an understanding of the adoption process and the factors that drive big data adoption in construction will facilitate devising strategies and plans to increase the adoption as well as it will help digitalization of the industry. As a result, they developed a conceptual model of the factors which drive big data adoption in construction. They propose several factors that influence adoption from a theoretical perspective. Their work is useful towards the development of context specific factors influencing big data adoption. However, their model would need further empirical testing to confirm some of the factors put forward.

Despite its significance, the application of data analytics is an emerging and growing area in the AEC sector. With the increasing digitisation of built environment project and asset management, implementing big data and analytics is evitable if the AEC must derive greater benefits from digitisation efforts. The variety of papers contained in this special issue provide readers with starting point for pursuing further research in this area. Action research approach could help in developing specific strategies for effective implementation of big data and analytics in various contexts. The guest editors would like to thank all authors who submitted the papers for this special issue as well as to the reviewers for their valuable contributions. Many thanks to Professor Mohan Kumaraswamy, the Editor-in-Chief, for supporting this special issue. His encouragement made this special issue possible.

**Ajibade A. Aibinu and Fernando Koch**
*Melbourne School of Design, Faculty of Architecture Building and Planning,*
*University of Melbourne, Melbourne, Australia, and*
**S. Thomas Ng**
*Department of Civil Engineering, University of Hong Kong, Hong Kong*

# Analyzing project data in BIM with descriptive analytics to improve project performance

Mohamed Marzouk and Mohamed Enaba
*Structural Engineering Department, Faculty of Engineering, Cairo University, Cairo, Egypt*

Received 4 May 2018
Revised 1 November 2018
18 March 2019
27 March 2019
Accepted 1 April 2019

## Abstract

**Purpose** – The purpose of this paper is to expand the benefits of building information modeling (BIM) to include data analytics to analyze construction project performance. BIM is a great tool which improves communication and information flow between construction project parties. This research aims to integrate different types of data within the BIM environment, then, to perform descriptive data analytics. Data analytics helps in identifying hidden patterns and detecting relationships between different attributes in the database.
**Design/methodology/approach** – This research is considered to be an inductive research that starts with an observation of integrating BIM and descriptive data analytics. Thus, the project's correspondence, daily progress reports and inspection requests are integrated within the project 5D BIM model. Subsequently, data mining comprising association analysis, clustering and trend analysis is performed. The research hypothesis is that descriptive data analytics and BIM have a great leverage to analyze construction project performance. Finally, a case study for a construction project is carried out to test the research hypothesis.
**Findings** – The research finds that integrating BIM and descriptive data analytics helps in improving project communication performance, in terms of integrating project data in a structured format, efficiently retrieving useful information from project raw data and visualizing analytics results within the BIM environment.
**Originality/value** – The research develops a dynamic model that helps in detecting hidden patterns and different progress attributes from construction project raw data.

**Keywords** Big data, Building information modelling, Trend analysis, Data analytics, Association analysis, Data clustering

**Paper type** Research paper

## Introduction

An abundance of data can be generated with different formats and databases to support construction projects throughout their lifecycle. Sharing data and supporting collaboration between project parties throughout the project's lifecycle is essential to enhance productivity (Martínez-Rojas *et al.*, 2015). Data-driven analytics is gaining wide popularity due to its ability to identify trends and patterns for business intelligence (Naderpajouh *et al.*, 2015). Abbaszadegan and Grau (2015) stated that from 50 to 80 percent of construction site problems arise from the lack of data or a delay in the receipt of information. Therefore, effective management of project data is vital to achieving construction project objectives. For example, Brynjolfsson *et al.* (2011) concluded that data-driven decision making improves productivity and output by 5–6 percent. The construction industry should rely on fact-based decision making to ensure industry survival. Moreover, data analytics application is still novel in the construction management field with a very limited number of research (Ahmed *et al.*, 2018).

BIM significantly helps to make project data organized and unified. BIM is considered to be a hub holding project information. Correa (2015) argued that BIM processes have been expanding rapidly, and shall include artificial intelligence, machine learning and data mining, which assists in automating many tasks including project workflow. BIM changes the way of information exchange between parties. Currently, there is a gap regarding the use of BIM as a communication platform between project parties. Using BIM in communication between project parties helps in minimizing project failure, innovation, efficient communication between project parties and better decision making (Zahiroddiny, 2016).

Built Environment Project and Asset Management
Vol. 9 No. 4, 2019
pp. 476-488
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-04-2018-0069

Zahiroddiny (2016) stated that in a traditional communication system data needs to be summarized to ensure the smooth delivery of information. El-Diraby *et al.* (2017) stated that one of the challenges currently facing the BIM environment is that the communication and interactions take place outside the BIM model which results in wasting valuable knowledge and delaying projects. Using BIM to describe and analyze construction project status based on the raw data produces a drastic improvement to projects, as this helps with discovering knowledge, predicting outcomes, automating processes and decision making (Correa, 2015). Furthermore, integrating BIM and construction document helps to ensure documents consistency of the project. Moreover, integration eases the process of document retrieval and display (Goedert and Meadati, 2008).

Big data analytics is a booming technology that effectively handles construction projects data. Big data has three chief characteristics (or the 3Vs): volume (size of data), velocity (rate of data generation) and variety (numerical or categorical, audio format files or video files) (Hafiz *et al.*, 2015). The construction industry meets the 3Vs of big data. In terms of volume, the construction industry generates a huge quantity of data (e.g. design data, schedules, enterprise resource planning, financial data). Regarding variety, there are numerous different data formats generated during the course of construction projects, including images, sensors, drawings, documents, schedules, excel sheets, etc. With reference to velocity, the construction industry is highly dynamic, and during a project's lifecycle there is an enormous stream of data produced daily (Bilal *et al.*, 2016). Mainly there are three types of data analytics: descriptive analytics, predictive analytics and prescriptive analytics (Wang *et al.*, 2016). Descriptive analytics deals with unsupervised data, where there are no predefined patterns. Descriptive data analytics aims of studying current processes to identify problems and opportunities. The scope of this research is to apply descriptive data analytics for construction project data, i.e. correspondence, daily progress report (DPR) and inspection request (IR).

It is worth mentioning, that nowadays construction projects are getting more complex with a relatively long duration and are highly technological and capital-intensive projects; therefore, appropriate analytics can help managers to achieve successful projects by effectively report project status. Descriptive data analytics is an efficient way to report construction project status. Limited efforts have studied the potential of BIM to include data analytics. In order to apply data mining in the construction project, data has to be standardized, organized and unified (Ahmed *et al.*, 2018).

Retrieving useful information to enhance a project from raw data is a challenging task (Han and Golparvar-Fard, 2017). Therefore, the aim of this paper is to develop a platform within the BIM environment to include correspondence, DPR and IR. Not only the developed model allows the project parties to communicate within the BIM environment, but also data analytics is performed for the gathered data. The developed model consists of three tiers which are: data integration within the BIM model per each model element, data analytics by applying association analysis, clustering technique and trend analysis and visualizing analytics results within the BIM environment.

## Research background

Big data management is divided into big data engineering and big data analytics. Big data engineering concerns data storage and processing (Bilal *et al.*, 2016). In contrast, big data analytics is concerned with tools and techniques used to extract insights from data, and drive decision making (Bilal *et al.*, 2016). Today, industries across a variety of sectors are shifting toward big data to maximize value and take advantage of the multiple benefits it affords. In particular, big data offers a better platform for decision making and optimization (Zhang *et al.*, 2015). For instance, Wal-Mart contracted with Hewlett-Packard to construct a data warehouse to record every transaction at all of their 6,000 stores worldwide. By applying machine learning to these data, Wal-Mart can detect patterns that allow it to evaluate pricing strategies, target advertising campaigns and optimize the supply chain. On the other hand, companies and institution in the medical sector also collect a huge volume of data, much of it regarding patients. Through data mining, medical researchers can gain insights into the causes of diseases and therefore create diagnostic tools that are more effective. Intelligence agencies are also concerned with data management, as it supports threat assessment by readily connecting multiple sources of information (Bryant *et al.*, 2008).

Although construction projects create huge amounts of unstructured data, however, data mining is not yet widespread in construction projects (Alsubaey *et al.*, 2015). In addition, as project data expands, data processing becomes more difficult. The construction industry is renowned for applying tools and techniques developed based on the experiences of other industries. Herein the paper explores the benefits of descriptive data analytics, i.e. association analysis and clustering technique to the construction industry.

### *Association analysis*

Association rule mining was proposed by Agrawal *et al.* (1993), and there are multiple benefits from studying association rule mining. Association rule mining aims to discover relationship variables in a large database. One of the most popular applications of association is market basket analysis. Market basket analysis is a process which analyzes customer buying habits; this is done by finding associations between different items. Market basket analysis helps in gaining insights into different items correlation (Han *et al.*, 2011). Association rule mining is useful for categorical data only. Generally, association rule strength lies in measuring support and confidence between items. Support measures the probability of two or more items containing combinations using Equation (1). On the other hand, confidence measures the degree of certainty that $X$ also contains $Y$, using Equation (2) (Ma *et al.*, 1998). Therefore, the output of association rule mining is that when $X$ occurs, $Y$ occurs with a certain probability:

$$\text{Support} = \frac{(X \cup Y).\text{count}}{n}, \tag{1}$$

$$\text{Confidence} = \frac{(X \cup Y).\text{count}}{X.\text{count}}, \tag{2}$$

where $X$ and $Y$ are data objects and $n$ is the number of data sets in a database.

Association rule mining applications in the construction industry include studying the occupational injuries in the industrial buildings sector, disputes analysis, energy performance analysis, analyzing accidents in the construction sites and studying the causes of defects (Lin and Fan, 2018). For Lee *et al.* (2016) applied association analysis to study the concrete defects in the construction industry. Lin and Fan (2018) studied the inspection records for 990 projects, and then 499 types of defects were extracted. Finally, an association is developed to measure the support and confidence between defects types. In addition, Wang *et al.* (2007) applied association data mining in order to monitor cranes health conditions. Cheng and Li (2015) proposed a genetic algorithm model which aims to monitor project quality defects. The model has two main modules which are database module and data mining module. The database module consists of defects tables which contain elements IDs and the list of attributes (material, element, defects type, description, etc.). The data mining module identifies the patterns of defects.

### *Clustering technique*

Clustering can be defined as the process of grouping a set of data objects into multiple groups based on their similarity. Clustering is a simple data mining method which enables easy class grouping. Clustering applications in the construction industry include clustering the facilities deficiencies, clustering construction accidents occurred, clustering project by customer satisfaction and clustering construction documents (Bilal *et al.*, 2016). However, there is no research regarding clustering project elements (slab, beam, masonry, etc.) during the course of the project to identify problems and to group project elements.

The partitioning clustering method is adopted as it is the most fundamental and commonly used clustering method. Partitioning clustering divides the data into K groups, where each group has at least one object. K-means clustering is the most common clustering algorithm due to its simplicity. In K-means clustering, K partitions are created based on the number of attributes (*n*). K-means clustering is performed in five steps. The first step is to produce a K random of centroids which is determined based on the user's requirements. The second step is to allocate each data point to nearest centroid based on proximity measure. Euclidean distance is the most widely used method to measure proximity distance which is estimated using Equation (3) (Kotu and Deshpande, 2014). The third step is modifying centroids for each cluster and introducing new centroids. The purpose of modifying centroids is to minimize the sum of squared errors (SSE) which is estimated using Equation (4). The fourth step is to iteratively repeat Step No. 2, where data points proximity is measured and calculated based on new centroids. The fifth step is iterating the second and third step until no major changes in the centroids are noted (Kotu and Deshpande, 2014):

$$\text{Distance} = \sqrt{(X_1-C_1)^2+(X_2-C_2)^2+(X_3-C_3)^2+\cdots+(X_n-C_n)^2}, \tag{3}$$

where $C$ is the number of centroids; $X$ the number of attributes and $n$ the number of data points.

$$\text{SSE} = \sum_{i=1}^{k}\sum_{x_i \in c_i}||x_i-\mu_1||2, \tag{4}$$

where $c_i$ is the $i$th cluster; $j$ the data points in a given cluster; $\mu_i$ the centroid for $i$th cluster; $x_j$ a specific data object.

## Research methodology

The research methodology is composed of three main stages. The first stage is data integration. At this stage, templates for correspondence, DPR and IR are formed and integrated within the BIM model. The documents are allocated per each model element. Second, data analytics is conducted, where data mining is applied to analyze project performance. Finally, the data mining results are presented in the BIM model. The outcome of the model is a tool able to answer questions about project data and develop an understanding of it. The answers to the questions are visualized on the BIM model. The rest of the section discusses each research tier.

### *Data integration*

The BIM tool implemented in this study is REVIT which is an Autodesk product (Autodesk, 2014). One of the features of REVIT is that it contains the REVIT Platform application program interface (API), the different applications of which can be included in a BIM model using .NET compliant language (Autodesk, 2014). REVIT API is applied to develop documents template windows in the REVIT model. One of the features of REVIT is that every project element has a unique ID code. In addition, the user can create a code for each element. This helps when organizing and managing construction project documents and ensures all the information is stored by element(s) within the model. So, aside from schedule and cost (5D) of data in the BIM model other critical data has been integrated within the BIM model.

*Correspondence procedures*. The correspondence template allows a consistent format for all letters in the construction project. To send correspondence, first, the sender identifies the element(s) where she/he has a concern. Then, the sender opens the messaging template. The sender identifies from, to and CC from a drop-down box. Then, the sender chooses the subject from a list of options which are: delay in submitting a document, delay of progress, variation order, test report, value engineering or other (specify). The sender should choose the type of letter from the available options (e.g. error, warning or contractual notice) as well as the action required by field. The sender also has to mention references to the letter (i.e. contract clause, meeting or verbal instruction). Finally, the sender identifies the potential cost and time impact per chosen element and overall project.

As each construction project is unique, the template is designed to be dynamic, allowing the project parties to add more fields. In addition, there is an option in the field with a drop-down menu to add more options. Therefore, the letter subjects and type are iteratively updated to include more subjects and types according to the project criteria. There is a free text box in which the sender can describe the purpose of the letter in details. Table I shows how the correspondence stored data from the correspondences template is converted into tabular form. Table I first column represents the correspondence number and the subsequent columns represent correspondences' attributes.

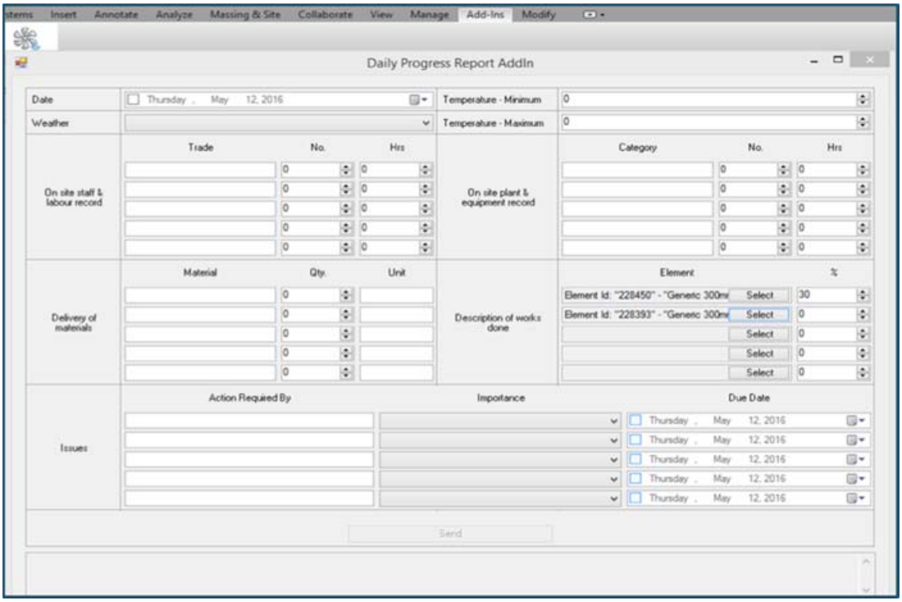
*Daily progress report*. DPRs are a crucial document in any construction project. The DPR template demonstrates different daily site data such as labor and staff, equipment, material delivered on-site and daily progress for each element (as shown in Figure 1). The percentage of completion of each element in the BIM model is stored and updated for each model element. This helps in monitoring project element(s) progress on a daily basis. Moreover, hidden progress habits and attributes driving the progress of each type of elements can be easily extracted. Thus, the DPR helps in monitoring project resources.

There are two main types of tables extracted from DPR template. First, there is the tracking of project elements' daily progress per each element. Second, there is the tracking of resources on-site, which include equipment, labor and materials on site (as shown in Table II). Table II first column represent calendar date and the subsequent columns represent the weather, daily number of labors, equipment and materials delivered on-site, respectively.

*Inspection request*. The IR template is created and integrated into the BIM model. In the IR window, the contractor chooses from the model element(s) that need to be inspected and then, the contractor fills the inspection date, time and describes the work to be inspected (see Figure 2). Finally, the contractor checks the purpose of inspection from a checklist, which includes surveying, earthwork, structure, architecture, mechanical, electrical, etc. However, if there is a new purpose of inspection not included in the template, the contractor chooses to add a new item (as shown in Figure 2), and then the new item will be integrated into the template checklist. Then, the consultant gets notified that there are works that need to be inspected. Thus, the responsible person evaluates the work and documents it in the IR template. There are three categorical evaluation statuses: approved (A), approved as noted (B) or rejected (C). After the inspection, the consultant evaluates the works and includes his/her comments if needed. The IR data are exported in tabular form. The tabular form includes calendar date, different work items that are inspected and the consultant evaluation of the works, i.e. approved (A), approved as noted (B) or rejected (C).

Table I. Correspondence table exported from BIM model

| No. | Element code | From | To | CC | Subject | Type | Action required by | Reference | Schedule impact | Cost impact |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | D02-01 | Contractor | Consultant | Employer | Test report | Error | Consultant | Null | Null | Null |
| 2 | D03-01 | Contractor | Consultant | Employer | Variation order | Error | Consultant | Clause no | Null | $100,000 |
| 3 | D03-02 | Consultant | Contractor | Employer | Delay of progress | Warning | Contractor | Null | 12 days | Null |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |



Daily Progress Report AddIn

| | | | |
|---|---|---|---|
| Date | ☐ Thursday , May 12, 2016 | Temperature - Minimum | 0 |
| Weather | | Temperature - Maximum | 0 |

On site staff & labour record

| Trade | No. | Hrs |
|---|---|---|
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 0 |

On site plant & equipment record

| Category | No. | Hrs |
|---|---|---|
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 0 |

Delivery of materials

| Material | Qty. | Unit |
|---|---|---|
| | 0 | |
| | 0 | |
| | 0 | |
| | 0 | |
| | 0 | |

Description of works done

| Element | | % |
|---|---|---|
| Element Id: "228450" - "Generic 300m | Select | 30 |
| Element Id: "228393" - "Generic 300m | Select | 0 |
| | Select | 0 |
| | Select | 0 |
| | Select | 0 |

Issues

| Action Required By | Importance | Due Date |
|---|---|---|
| | | ☐ Thursday , May 12, 2016 |
| | | ☐ Thursday , May 12, 2016 |
| | | ☐ Thursday , May 12, 2016 |
| | | ☐ Thursday , May 12, 2016 |
| | | ☐ Thursday , May 12, 2016 |

Send

**Figure 1.**
Daily progress report BIM template

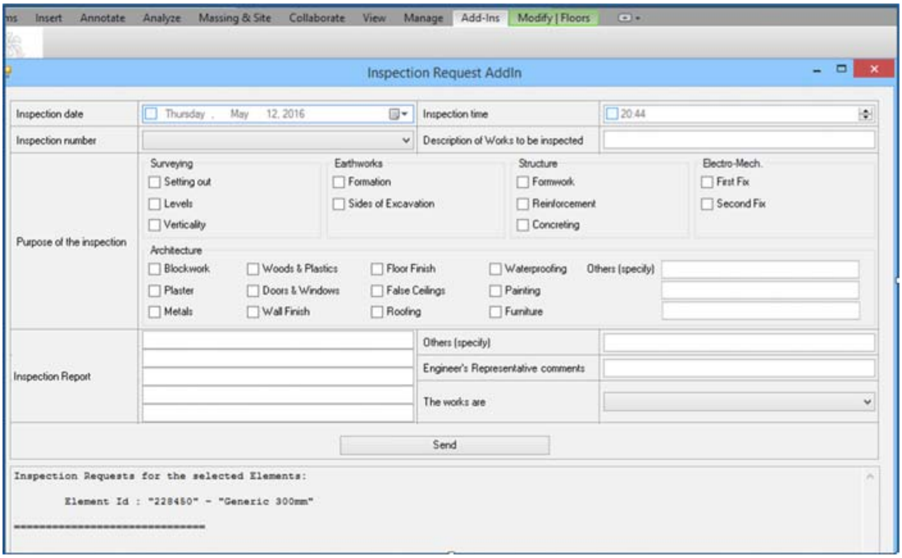
**Table II.**
Daily equipment, staff, labors and material on site

| | | Equipment | | | Staff and labor | | | Material on site | |
|---|---|---|---|---|---|---|---|---|---|
| Date | Weather | Loader | Total station | Fork left | Contractor engineers | Contractor labors | Consultant engineers | Cement (ton) | Reinforcement bars (ton) |
| January 1, 2017 | 10°C | 1 | 1 | 3 | 10 | 100 | 6 | Null | Null |
| January 2, 2017 | 8°C | 1 | 2 | 3 | 9 | 120 | 6 | Null | Null |
| January 3, 2017 | 8°C | 2 | 3 | 3 | 10 | 150 | 6 | 50 | 3,000 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

### *Data analytics*

After developing Tier 1, the data are ready to be analyzed. Three data mining techniques adopted in this research which are association, clustering and trend analysis. Association supports the assessment of the strength of co-occurrence between one element and another per item of correspondence. The output of association specifies that if a party chooses one element there is then a tendency to choose another specific element for the correspondence. The developed model calculates support and confidence between different project elements. Another example of implementing association in this research involves studying the different correspondence subjects stored in relation to each element. The main advantage is that the project parties can gain an understanding regarding how to reduce argumentative correspondence and identify potential failure causes per each element.

Clustering is implemented when studying the allocated correspondence subjects by element. There are numerous attributes that can be included in clustering, i.e., trade, allocated contractor, cost impact, schedule impact, IR grade, temperature, element allocated contractor, correspondence importance and availability of resources. Clustering elements



| | | | |
|---|---|---|---|
| Inspection date | □ Thursday , May 12, 2016 | Inspection time | □ 20:44 |
| Inspection number | | Description of Works to be inspected | |

Purpose of the inspection

Surveying: □ Setting out □ Levels □ Verticality

Earthworks: □ Formation □ Sides of Excavation

Structure: □ Formwork □ Reinforcement □ Concreting

Electro-Mech.: □ First Fix □ Second Fix

Architecture: □ Blockwork □ Plaster □ Metals □ Woods & Plastics □ Doors & Windows □ Wall Finish □ Floor Finish □ False Ceilings □ Roofing □ Waterproofing □ Painting □ Furniture Others (specify)

Inspection Report

Others (specify)

Engineer's Representative comments

The works are

```
Inspection Requests for the selected Elements:

        Element Id : "228450" - "Generic 300mm"

==============================
```

**Figure 2.**
Inspection request
BIM template

based on correspondence subject helps to identify potential problems and to develop how different elements groups can impact project performance. Therefore, clustering a correspondence subject provides an early warning regarding the diverse issues that may cause cost or schedule overrun. This can help to resolve issues efficiently. Additional, clustering is performed to group delayed activities and to analyze project schedule performance. Finally, clustering is performed to group elements according to IR grades, i.e. A (approved), B (approved as noted) or C (rejected). There are numerous attributes that can be considered here, including element type, element allocated contractor, period (month or day) and weather. After developing the IR template, different attributes can be easily recorded within the BIM model.

Different trends can be extracted from construction projects to explore different attributes affecting project progress. Trend analysis uses historical data to extrapolate what will happen in the future. One of the strengths of big data is that it allows decision makers to draw on practice rather than theories. Thus, comparing various attributes, including different direct and indirect equipment, staff, laborers and material on site, against progress over time can help to identify what drives progress. Every construction project is unique; therefore, trends analysis helps us to take creative action based on the project data analyzed. Furthermore, trend analysis helps afford insights into probable schedule performance.

Based on the above, the developed model provides a tool to answer questions about project data and guides the team by using questions that help to analyze the construction project performance, utilizing the collected data. Table III provides examples of dynamic queries proposed by the model to assist the project team to understand the data. The queries template is dynamic, and there are a lot of options to be completed as shown by blank spaces. Therefore, the project team gains an understanding of the different attributes that can cause quality problems, schedule overrun or cost overrun. The first column in Table III shows the data source. The second column shows the data mining algorithm used to solve the questions, and the third column the questions proposed by the model, in which the dots present different options that the user can choose from the last column.

**Table III.**
BIM model dynamic queries

| Data source | Algorithm | Criteria | Blanks alternatives |
|---|---|---|---|
| Correspondence | Association | Measure the strength co-occurrence between one element and another per each correspondence with support of … % and confidence of … % | Percentage from 0% to 100% |
| | | Measure different correspondence subject stored per each element with support of … % and confidence of … % | Percentage from 0% to 100% |
| | Clustering | Display the most common Subjects of correspondence | Not applicable |
| | | Display the most common Types of correspondence | Not applicable |
| | | Show Elements Which has correspondence subject of … | Delay in submitting document<br>Delay of progress<br>Variation order<br>Test report<br>Value engineering<br>Other (specify) |
| | | Show Elements Which has correspondence Type of … | Error<br>Warning<br>Contractual notice |
| | | Cluster Elements with cost impact less than … Less than … And Greater than … | Amounts ($) |
| | | Filter Elements with Schedule less than … Less than … And Greater than … | Duration (days) |
| Daily progress report | Trend analysis | Show the trend of element progress over … | Weather<br>Labor<br>Consultant engineers<br>Contractor engineers<br>Equipment |
| | | Calculate amount of work that can be executed based on Material On Site (MOS) | Steel bars<br>Cement<br>Blocks<br>Painting |
| Inspection request | Clustering | Visualize the Inspection Requests' grade per element | Not applicable |
| | | Visualize Inspection Requests' per … | Surveying work<br>Concrete works<br>Block works<br>Mechanical works<br>Electrical works |

### *Data visualization*

After applying data analytics, it is important to present results efficiently. Visualizing results in an efficient way is a cornerstone of this research, which aims to help project teams benefit from insights and hidden patterns in data. Therefore, the majority of the results (where applicable) are presented on the model. NAVISwork was adopted as a visualization tool in this research. NAVISwork is an Autodesk product, which provides advanced simulation, visualization and analysis of 3D models (Autodesk, 2014). The data analytical model developed using BIM allows the project team to apply descriptive analytics and obtain insights, even if they are not themselves data scientists.

## Case study

A case study has been performed to validate the implemented methodology. The case study aims to proof a concept which is studying the benefits of integrating BIM and data analytics in a construction project. The implemented case study is for a construction project located in Cairo, Egypt. The scope of work is to execute entire civil and infrastructure works on turnkey basis for a residential compound. The project consists of 237 units. The project total land area is 236,000 $\mathrm{m}^2$. The projects mainly consist of the residential units, public areas and infrastructure. The project data (i.e. correspondence, DPR and IR) are acquired and integrated into the project 5D model, although the case study focuses only on three sets of data, that is not considered alone as a big data. However, the case study shows that there are a lot of hidden patterns and insights in these data to track and analyze the construction projects performance.

### *Results and discussion*

At this point, a question arises concerning whether the construction industry meets the pillars of big data? In terms of volume, the case study data collected to integrate into the 5D BIM model (correspondence, DPRs and IRs), reached 150 GB. The size of the model is relatively large, despite not including process documentation, GIS, augmented reality, data conflict detection, energy efficiency data or real-time progress monitoring using CCTV, which are also significant areas of research in the construction industry. Were these data sources included, the model size could easily reach into Terabytes. In terms of velocity, there is a continuous stream of data updates daily, including but not limited to progress updates, invoices, quality management and document control management. Recent construction management technology studies automated progress monitoring and live streaming. This also increases the velocity of construction project data generation. Finally, in terms of variety, the gathered data includes design documents, excel sheets, text documents and images. Therefore, it can be concluded that construction projects fulfill the big data management pillars. Moreover, when the construction projects become more complex, the only way to handle the project data is by applying big data concepts and techniques. The rest of this section discusses the association, clustering and trend analysis case study results.

Association is developed to measure co-occurrence between one element and another per each correspondence. The elements are consolidated into categories which are earthwork, plain concrete, slab on grade, columns, slab, blockwork, mechanical works, electrical works, roads and infrastructure. The project team can sieve confidence from 0 up to 100 percent. It was found that the highest association is between earthwork and roads, as there is a 26.5 percent support and 100 percent confidence between roads and earthwork for correspondence. Thus, if the project team is able to solve earthwork problems, then the roads problems will be solved accordingly with a support of 26.5 percent and confidence of 100 percent.

Another association has been developed to measure (support/confidence) of correspondence stored for each model element. So, it has been found that there is a support of 60 percent and a confidence of 90 percent between correspondence subject variation order and delay in submitting documents. After digging the results deeper, it is found that the process of approving the variation order is a lengthy process. Moreover, the contractor always delays in providing the supporting documents of variation order. In addition, the developed model shows that there always a delay in the elements with an allocated subject of variation order and delay of submitting documents. Therefore, the project team shall investigate corrective action regarding managing the changes in the project.

The above analysis shows applying association in construction projects helps in detecting the correlation between different elements or correspondences subjects. Hence, studying these correlations help the project team regarding the criticality of different documents. Moreover, the hidden patterns can be extracted from the project data and analyzed, in order to avoid delay or cost overrun.

The developed model allows developing different clusters based on different attributes, i.e. elements allocated contractor, trade, correspondence subject, delay severity, elements total float, number of laborer's, equipment and material availability. Clustering delayed activities means delays are recorded within the BIM model, making it possible to identify the different factors affecting project schedule performance.

Clustering model is developed to study stored correspondence subject per each model element. The project team can choose any kind of subject as discussed in Table III. Moreover, the project team can filter using the correspondence type (i.e. error, warning or contractual notice). In addition, a clustering model is developed to group IR grades (i.e. A, B or C) for each trade. The clustering results are being visualized in the BIM model (as shown in Figure 3). The IRs with grades A, B and C are being highlighted in green, yellow and red, respectively. Storing the IR grade per element leverages the project team to analyze the IR grade per each trade (as discussed in Table III). The clustering results show that block works always obtain grade C, and this illustrates that there is a quality problem in this trade. Furthermore, the majority of IR got a grade B. Figure 3 shows that there is a quality problem, where the majority of works got grades B or C. The results show that corrective actions are needed to improve the quality of work. Thus, avoiding any rework cost and schedule overrun.

Segregation based on IR grade helps to further understanding regarding work quality per trade under various circumstances. Moreover, adding more attributes assists in assessing different project contractors' competency, strength and weakness. Clustering inspections require help discovering meaningful groups of trades. Therefore, this advance understanding of how different work elements are associated with different groups, helping to clarify the kind of work elements that require attention in order to avoid rework.

A number of patterns are discoverable here, since project data are constantly being recorded and structured. For instance, studying the period over which a project has progressed the most (based on recorded elements percentage of completion) can drive optimum ratios between attributes, and ratios can be developed between the number of laborers and contracted engineers or the optimum ratio between contracted engineers and consultant engineers. Moreover, labor productivity rates for labor and equipment can be calculated for a specific trade or period. Studying these trends as set out in DPRs helps the project team to make accurate decisions based on project trends. It is noteworthy, however, that trend analysis is performed on numerical attributes only, and that these have to first be normalized, as the ranges between different attributes can vary significantly.

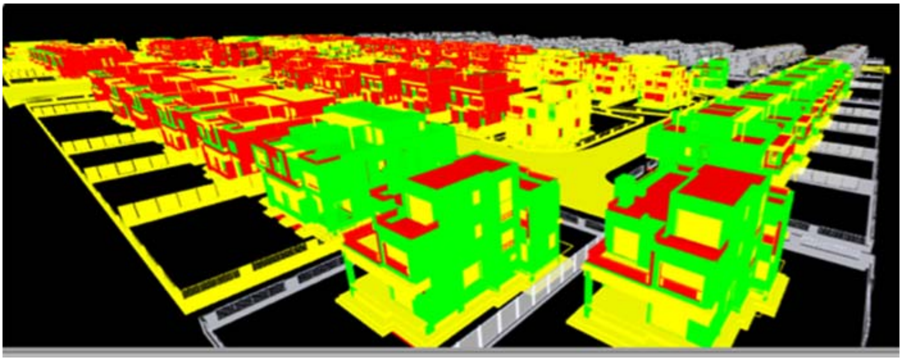


**Figure 3.**
IRs clustering visualization in the BIM model

## Conclusion

The research emphasizes the importance of integrating different project documents within the BIM models. In this research, unstructured data have been structured and integrated per each element in the BIM model which allows efficient communication, data consistency and data analytics. Then, data analytics adopted to study the impact of integrating BIM and data analytics then the results are validated through applying a case study of a construction project. Association, clustering and trend analysis are applied in this research. Through applying association, the correlation between construction elements is extracted. Moreover, the correlations between correspondence subjects are tackled. Applying, clustering allows segmenting different correspondences' subject based on time and cost impact, besides, grouping elements based on IR grades (i.e. A, B or C). The paper also argues that integrating BIM and data analytics has a lot of benefits such as providing the project team with a rich overview of the project status, analyzing the project performance, avoiding misconceptions in data sets and reducing the time of explaining information.

A limitation of this research is that there is a large amount of data that has to be added to the model on a daily basis. Therefore, the data entry has to be accurate in order to avoid misleading results. Another limitation of this research is not studying different platforms, i.e. Hadoop for handling the model massive amount of data. A critical success factor to achieve the claimed analytics results is that the project team shall collaborate and use BIM as a communication platform. Future research will consider applying text mining to all project correspondences, as it will help in gaining insights regarding potential problems from correspondences between project parties.

## References

Abbaszadegan, A. and Grau, D. (2015), "Assessing the influence of automated data analytics on cost and schedule performance", *Procedia Engineering*, Vol. 123, pp. 3-6.

Agrawal, R., Imieliński, T. and Swami, A. (1993), "Mining association rules between sets of items in large databases", *ACM Sigmod Record*, Vol. 22 No. 2, pp. 207-216.

Ahmed, V., Aziz, Z., Tezel, A. and Riaz, Z. (2018), "Challenges and drivers for data mining in the AEC sector", *Engineering, Construction and Architectural Management*, Vol. 25 No. 11, pp. 1436-1453.

Alsubaey, M., Asadi, A. and Makatsoris, H. (2015), "A Naive Bayes approach for EWS detection by text mining of unstructured data: a construction project case", *SAI Intelligent Systems Conference (IntelliSys), IEEE*, pp. 164-168.

Autodesk (2014), "Autodesk Navisworks products", available at: www.cadac.com/media/1207/autodesk-navisworks-2014-brochure.pdf (accessed April 5, 2017).

Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A. and Pasha, M. (2016), "Big data in the construction industry: a review of present status, opportunities, and future trends", *Advanced Engineering Informatics*, Vol. 30 No. 3, pp. 500-521.

Bryant, R., Katz, R.H. and Lazowska, E.D. (2008), "Big-data computing: creating revolutionary breakthroughs in commerce, science and society", Computing Communication Consortium (CCC), Washington, DC.

Brynjolfsson, E., Hitt, L.M. and Kim, H.H. (2011), "Strength in numbers: how does data-driven decision-making affect firm performance?", SSRN 1819486.

Cheng, Y. and Li, Q. (2015), "GA-based multi-level association rule mining approach for defect analysis in the construction industry", *Automation in Construction*, Vol. 51, pp. 78-91.

Correa, F.R. (2015), "Is BIM big enough to take advantage of big data analytics?", *Proceedings of the International Symposium on Automation and Robotics in Construction, Vilnius Gediminas Technical University, Department of Construction Economics and Property*, Vol. 32, p. 1.

El-Diraby, T., Krijnen, T. and Papagelis, M. (2017), "BIM-based collaborative design and socio-technical analytics of green buildings", *Automation in Construction*, Vol. 82, pp. 59-74.

Goedert, J.D. and Meadati, P. (2008), "Integrating construction process documentation into building information modeling", *Journal of Construction Engineering and Management*, Vol. 134, No. 7, pp. 509-516.

Hafiz, A., Lukumon, O., Muhammad, B., Olugbenga, A., Hakeem, O. and Saheed, A. (2015), "Bankruptcy prediction of construction businesses: towards a big data analytics approach", *2015 IEEE First International Conference on Big Data Computing Service and Applications (BigDataService), IEEE*, pp. 347-352.

Han, J., Kamber, M. and Pei, J. (2011), "Data mining: concepts and techniques", 3rd ed., Morgan Kaufmann Publishers Inc., San Francisco, CA.

Han, K.K. and Golparvar-Fard, M. (2017), "Potential of big visual data and building information modeling for construction performance analytics: an exploratory study", *Automation in Construction*, Vol. 73, pp. 184-198.

Kotu, V. and Deshpande, B. (2014), *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*, Morgan Kaufmann.

Lee, S., Han, S. and Hyun, C. (2016), "Analysis of causality between defect causes using association rule mining. World academy of science, engineering and technology", *International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering*, Vol. 10 No. 5, pp. 659-662.

Lin, C.L. and Fan, C.L. (2018), "Examining association between construction inspection grades and critical defects using data mining and fuzzy logic", *Journal of Civil Engineering and Management*, Vol. 24 No. 4, pp. 301-317.

Ma, B.L., Liu, B. and Hsu, Y. (1998), "Integrating classification and association rule mining", *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, New York, NY, pp. 24-25.

Martínez-Rojas, M., Marín, N. and Vila, M.A. (2015), "The role of information technologies to address data handling in construction project management", *Journal of Computing in Civil Engineering*, Vol. 30 No. 4, p. 04015064.

Naderpajouh, N., Choi, J. and Hastak, M. (2015), "Exploratory framework for application of analytics in the construction industry", *Journal of Management in Engineering*, Vol. 32 No. 2, p. 04015047.

Wang, G., Gunasekaran, A., Ngai, E.W. and Papadopoulos, T. (2016), "Big data analytics in logistics and supply chain management: certain investigations for research and applications", *International Journal of Production Economics.*, Vol. 176, pp. 98-110.

Wang, Z., Hu, X. and Chen, Z. (2007), "Mining association rules on data of crane health-condition monitoring", *International Conference on Transportation Engineering*, pp. 2054-2059.

Zahiroddiny, S. (2016), "Understanding the impact of building information modelling (BIM) on construction projects' communication patterns", doctoral thesis, Northumbria University, available at: http://nrl.northumbria.ac.uk/30221/ (accessed August 10, 2017).

Zhang, Y., Luo, H. and He, Y. (2015), "A system for tender price evaluation of construction project based on Big data", *Procedia Engineering*, Vol. 123, pp. 606-614.

## Further reading

Autodesk (2017), "Revit 2014 platform API developers guidelines", Autodesk, available at: www.autodesk.com/developer-network/platform-technologies/revit (accessed January 4, 2017).

Provost, F. and Fawcett, T. (2013), *Data Science for Business: What you Need to Know about Data Mining and Data-analytic Thinking*, O'Reilly and Associates, Sebastopol, CA.

## Corresponding author

Mohamed Enaba can be contacted at: moh.enaba@gmail.com

# BIM-linked data integration for asset management

Karim Farghaly, F.H. Abanda, Christos Vidalakis and Graham Wood
*Faculty of Technology Design and Environment,*
*Oxford Brookes University, Oxford, UK*

Received 22 November 2018
Revised 19 May 2019
Accepted 28 May 2019

## Abstract

**Purpose** – The purpose of this paper is to investigate the transfer of information from the building information modelling (BIM) models to either conventional or advanced asset management platforms using Linked Data. To achieve this aim, a process for generating Linked Data in the asset management context and its integration with BIM data is presented.

**Design/methodology/approach** – The research design employs a participatory action research (PAR) approach. The PAR approach utilized two qualitative data collection methods, namely; focus group and interviews to identify and evaluate the required standards for the mapping of different domains. Also prototyping which is an approach of Software Development Methodology is utilized to develop the ontologies and Linked Data.

**Findings** – The proposed process offers a comprehensive description of the required standards and classifications in construction domain, related vocabularies and object-oriented links to ensure the effective data integration between different domains. Also the proposed process demonstrates the different stages, tools, best practices and guidelines to develop Linked Data, armed with a comprehensive use case Linked Data generation about building assets that consume energy.

**Originality/value** – The Linked Data generation and publications in the domain of AECO is still in its infancy and it also needs methodological guidelines to support its evolution towards maturity in its processes and applications. This research concentrates on the Linked Data applications with BIM to link across domains where few studies have been conducted.

**Keywords** BIM, Building maintenance, Asset management, Information management, Information exchange, Building lifecycle

**Paper type** Research paper

## 1. Introduction

In the last decade, building information modelling (BIM) has been recognized as an evolving technological innovation which can facilitate the transformation of the construction industry (Li *et al.*, 2017) and create several opportunities to exchange data between different stakeholders (Zadeh *et al.*, 2017). BIM data function as "back-end data" in the Computer Aided Facility Management systems for activities such as space management and maintenance management (Kučera and Pitner, 2018). Therefore, private and public owners are interested in the application of BIM in the built environment in general and in the asset management domain in particular (Becerik-Gerber *et al.*, 2012; Teicholz, 2013). In order to facilitate asset management, the delivered building information BIM models in the handover stage have to contain accurate information related to the building assets for better operation and maintenance (O&M) (Zadeh *et al.*, 2017). To achieve this aim, it has been argued that effective syntactic and semantic interoperability between building information models and different assets database has to be achieved (Ibrahim *et al.*, 2016). Pärn *et al.* (2017) critiqued that semantic interoperability is the single most important interoperability challenge to overcome the integration of BIM data with other systems including AM platforms. Despite the different classifications and long lists of diverse and required information have been developed by academics and industry professionals to improve the BIM implementation in asset management, it has been argued that still there is a lack of interoperability between building information models and different data domains prevents the cross-domain use of data at an enterprise level in the O&M industry (Corry *et al.*, 2014). Hu *et al.* (2018) argued an ontology is required to cross-link building performance with other building information and Linked Data

Built Environment Project and Asset Management
Vol. 9 No. 4, 2019
pp. 489-502
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-11-2018-0136

offers a mechanism to facilitate meaningful sharing of cross-domain building information. Therefore, providing object-oriented cross-domain linking instead of identifying the required information can be more efficient and adequate solution to achieve semantic interoperability between BIM and asset management (Kim *et al.*, 2018).

Linked Data generation and publication is a comprehensive process which requires a pursuit and implementation sequential tasks to ensure effective linking of different data sets. To this end, several guidelines and best practices have been developed and being advocated. However, it has been argued that general guidelines alone cannot provide the required level of detail and adequate process maps including all the different tasks. This is mainly due to the general guidelines not taking into consideration specific characteristics and vocabularies of particular domains (Villazon-Terrazas *et al.*, 2012). As Linked Data application is still in its infancy in the AECO domain (Radulovic *et al.*, 2015), a specific process map for Linked Data generation and publication is required. The main aim of this research is to address the suggested gap by providing a structured process map for linking data between building information models and different information stored in silos related to building assets. This process can improve the data exchange from the building information BIM models to asset management systems during handover stage and consequently improve asset management during the O&M stage. Achieving the set aim would allow a series of different ontological sources to talk to each other and also enhance compatibility with present and future versions of platforms and databases adapting the selected ontological sources. The proposed process offers a comprehensive description of the required standards and classifications in construction and operation domains, related vocabularies and their link to each other. Also, the proposed process presents the different stages, tools, best practices and guidelines to develop Linked Data, armed with a complete example on the generation and publication of Linked Data concerning assets that consume energy in buildings. The selection of assets that consume energy was underpinned by the UK strong environmental and economical mandates to improve the existing buildings energy performance and to ensure new constructions ensuring sustainable performance during the whole building lifecycle.

## 2. Linked Data

Linked Data aims to define a process to publish and share machine readable inter-Linked Data on the web, based on a set of design principles. Semantic Web term has been coined before Linked Data term in 1994 and documented in a scientific American article in 2001 (Berners-Lee *et al.*, 2001). The Semantic Web is "Web of actionable information – information derived from data through a semantic theory for interpreting the symbols" (Shadbolt *et al.*, 2006). The Semantic Web is an extension to the current web where information (Data and documents) is well-defined to ensure better cooperation between computers and humans. The Semantic Web purpose is to achieve data universality and data linking with any other data. However, Berners-Lee (2006) noticed that some semantic data published on the web are not linked to other outside semantic data. Therefore Bernes-Lee outlaid four principles that need to be adopted to obtain truly Linked Data. In 2010, Bernes-Lee suggested a five-star deployment schema for Linked Data based on the four Linked Data principles. The Semantic Web of data, unlike the document web, requires standards to ensure a highly interconnected network where the huge amount of heterogeneous data has been given a well-defined meaning. Therefore, an ecosystem of standards, named Semantic Web Stack, to support LOD has been developed by the World Wide Web Consortium team.

### *2.1 Linked Data and Semantic Web in AECO domain*

Linked Data application has enjoyed great popularity in other domains including biology, medical records, accounting and social media (Schmachtenberg *et al.*, 2014). These success stories encourage the implementation of Semantic Web and Linked Data in AECO domain (Radulovic *et al.*, 2015). Several researchers have illustrated the different benefits can be acquired by the implementation of Linked Data in the built environment domain, also the barriers to overcome the integration of data through the building lifecycle phases. Abanda *et al.* (2013) categorized Semantic Web implementation studies in the built environment domain based on area of application. While Pauwels *et al.* (2017) categorized the works differently to three main categories based on aims and barriers to overcome.

Curry *et al.* (2013) proposed the use of Linked Data in order to manage and operate building assets holistically. They argued that Linked Data can provide a cross-domain integration between building silo systems in a homogenous format. Based on that, O'donnell *et al.* (2013) combined Linked Data and complex event processing technologies to enhance the efficiency of building energy management activities. Several silo domains such as human resources source, architecture source which are represented in 3D models, inventory source, legislation source and building energy performance (BEP) source can be cross-mapped and integrated through the use of Linked Data. They argued that this approach can reduce the time required by building managers to analyse and optimize BEP. Corry *et al.* (2014) demonstrated several examples of integration and publishing of building-related data following Semantic Web rules to improve the building performance. One of the examples was the cross-domain integration of the scheduling data with the building operation strategy. They linked the data from building management systems, the room booking system, human resources management systems and building information model based on the room entity. Lee *et al.* (2016) proposed a framework for sharing construction defect information through the applicability of BIM and Linked Data. For creating the ontology, they adapted OmniClass classification's taxonomy to build the classes and properties. Kim *et al.* (2018) proposed a method to link the Industry Foundation Classes (IFC) objects with the FM work information. They developed a semantic relation between the classes of IFC, COBie and historical maintenance work concepts. They argued that the proposed approach can enable facility managers to semantically link the BIM objects to the maintenance records in the Semantic Web during the O&M phase in order to provide a BIM environment without the specific BIM authoring application.

Despite the available studies, the Linked Data generation and publications in the domain of AECO is still in its infancy and it also needs methodological guidelines to support its evolution towards maturity in its processes and applications (Radulovic *et al.*, 2015). Pauwels *et al.* (2017) argued that highest number of used cases in the construction industry lies under linking across domains because of the limited number of involved technologies and the simpler approaches. This research concentrates on the Linked Data applications with BIM to link across domains where few studies have been conducted. The following section – Section 3 – represents the research methods adopted for the study and the proceeding section – Section 4 – presents the proposed process for generating Linked Data in the context of asset management and its integration with BIM data. Finally, the last section – Section 5 – provides some concluding remarks and suggests future lines of work.

## 3. Methodology

Research design is the map for the research process to achieve its aims and objectives. The aim of this research complies with the aim of the theory behind data integration. Data integration started in the 1980s with discussions about "data exchange" between different applications. It has been agreed that data integration between BIM and AM systems can provide unified view across data sources and enable the analysis of combined data sets to unlock new knowledge (Farghaly *et al.*, 2017). Based on the workflow of data integration suggested by Sherman (2015), the research design of this research is formulated (Figure 1).

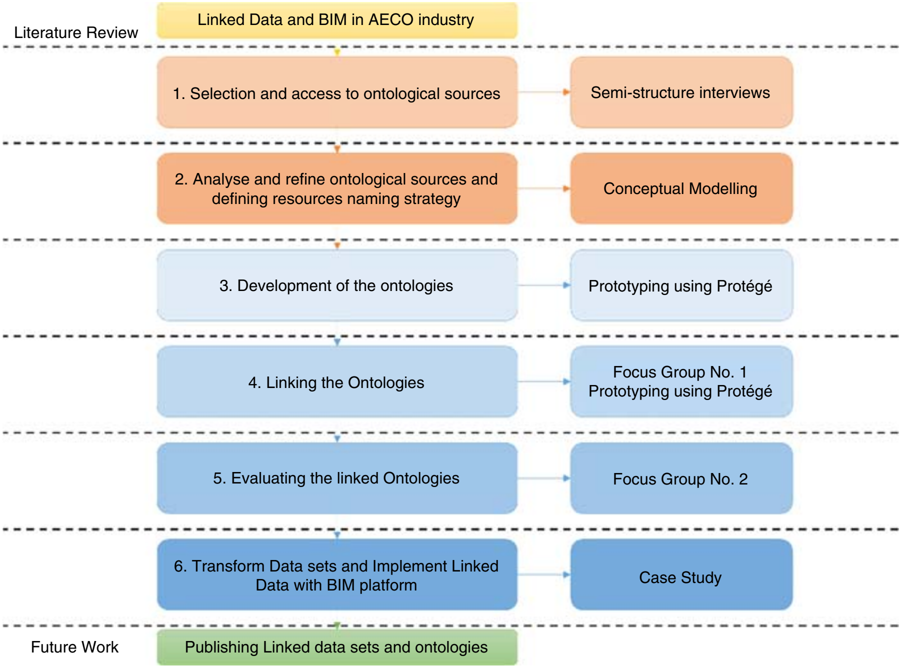

Literature Review
Linked Data and BIM in AECO industry
1. Selection and access to ontological sources
Semi-structure interviews
2. Analyse and refine ontological sources and defining resources naming strategy
Conceptual Modelling
3. Development of the ontologies
Prototyping using Protégé
4. Linking the Ontologies
Focus Group No. 1
Prototyping using Protégé
5. Evaluating the linked Ontologies
Focus Group No. 2
6. Transform Data sets and Implement Linked Data with BIM platform
Case Study
Future Work
Publishing Linked data sets and ontologies

**Figure 1.**
Research design of this research

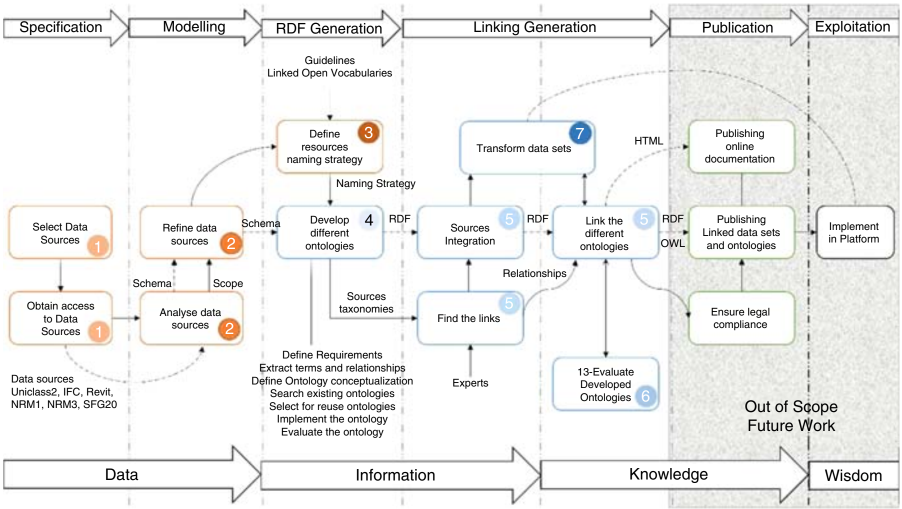
Each phase in the research design requires different research methods to enrich the ontologies with the required information. The first phase includes the engagement of experts to identify the required standards for the mapping of different domains. Four semi-structured interviews with facility managers were conducted in the first phase for that purpose. Once the ontological sources such as: new rules of measurement (NRM) 1 and 3, Uniclass2, SFG20, IFC and finally Revit as a BIM platform are identified, the different sources were accessed. Phase 2 requires analysing and refining the ontological sources to be suitable for the research scope. Conceptual modelling approach has been adapted to build the required conceptual frameworks. Conceptual framework has been constructed from combining different paradigms and concepts based on the literature review and conducted interviews with the research goal and objectives. The main sources of the developed data sets are previous researches, applicable theories, researcher's own experience knowledge and research's thought experience (Maxwell, 2008). Phase 3 has been achieved through prototyping which is an approach of Software Development Methodology. The prototype model is selected as the approach in this research as it is a top-down, iterative approach that continues until the user's requirements are accomplished. The fourth phase includes the involvement of experts from the AECO industry to define the link between classes in different ontologies. As cross-domain integration of assets' data and data from building information models are relatively new and no enough information around that topic, focus groups is conducted in this research as the qualitative data collection method. This interaction between participants in focus groups may produce spontaneous responses and more cognitive views. The focus group was conducted with eight experts in the construction and/or operation industry. The expertise for eligibility to participate in the focus group was determined based on different criteria, namely; five years' experience in BIM and/or asset management and mechanical or electrical engineer. The focus group started with high level of involvement of the interviewer, by giving an introduction to the different classifications for the building assets and brief introduction to Linked Data and ontologies. The first question is then enquired, leading to an unstructured discussion about the potential answers. During discussion, the interviewer level of involvement was low, then moved to high by concluding the discussion and then moving to the next question. Another focus group was conducted with the same participants in the first focus group in Phase 5. The second focus group has been lead to evaluate and validate the mapping developed between all the different ontologies for each individual asset that consumes energy during the development of ontologies stage. The interviewer showed the developed mapping between the different standards for each asset and consequently the participants interact to agree/disagree/refine the developed mapping. The focus group's discussion was whether the abstracted concepts were precise and accurate or not. Finally, a case study has been conducted for a new extension of an educational building in Phase 6 to evaluate the implementation of the proposed mapping between the different ontological sources.

## 4. Linked Data development

Figure 2 represents an overview of whole process of Linked Data development and its relevant tasks. The sequential relations between the tasks are represented with full lines, while the outputs from each task are represented with continuous lines. The process also is represented by other four main phases, namely; data, information, knowledge and finally wisdom. These phases have been adapted from the Big Data domain. Following, each task of the Linked Data development process is discussed in details.

### 4.1 Selection and access to ontological sources

The first two tasks in Linked Data generation process are the selection of the required ontological sources and obtaining access for the selected ontological sources, respectively. Different ontological sources were selected in the research to achieve the required goal such as: NRM1, NRM3, Uniclass2, SFG20, IFC and finally Revit as a BIM Platform. All these ontological sources achieved the specified requirements explicitly; to include data/classification or vocabularies about assets in buildings, to be available to use, standards



**Source:** Adapted from Radulovic *et al.* (2015)

**Figure 2.**
Process of generation and publication of Linked Data

and/or guidelines related to the best practice in UK and finally to be presented in a structured way to be easily adopted.

NRM is a suite of documents issued by the Royal Institution of Chartered Surveyors group. NRM1 has been published to provide a standard and guideline on the quantification of building works for cost estimate purpose based on the UK practice. While, NRM3 has been written to provide a standard and guideline for the quantification and description of the maintenance works for cost estimate purpose during building phases. Both NRM1 and NRM3 have been selected for the proposed Linked Data as they are understandable by all stakeholders involved in the project and associated elemental classifications can aid the communication between the project team and the employer (RICS, 2012). The data source of the NRM1 and NRM3 data set is available in the public domain from the rics.org.uk webpage (RICS, 2012) and is provided in PDF format as tables.

IFC provides the benchmark for sharing of information of any built environment asset through its lifecycle between all the stakeholders, notwithstanding of the used software application. The data source of IFC data set and vocabulary is available on BuildSMART website and is provided in ifcXML and Web Ontology Language (OWL) formats. SFG20 provides the benchmark for optimum maintenance, avoiding over or under maintaining of assets and the backbone to building engineering services maintenance industry. Although SFG20 is not specified in PAS 1192, it can be easily figured out as it is aligned with NRM3. The authors accessed the SFG20 through a free trial request where the standard is provided in a tree online taxonomy.

Uniclass2 is the new UK implementation of the international framework for construction information. Uniclass2 was been developed to form a structured classification which is endorsed by all construction and property bodies and professional institutions. The data source of Uniclass2 data set and vocabulary is available as structured tables on the NBS BIM toolkit website and is provided in PDF and xls formats. Finally, Revit is one of the most popular BIM platforms. It is important to take in consideration the classification of Revit to be able to link all the different standards with the Revit elements. Revit is accessed through the university computers.

### *4.2 Analyse and refine ontological sources*

Once the ontological sources are selected and access is obtained, the next step is to analyse the data in order to observe how much the data is structured and organized, understand the structure/schema of the data and the relationships between them and finally define the required data sets to form the classes and concepts for an ontology. Subsequently, the next step is to refine the data by correcting errors in the schema and creating mapping between columns and rows if the schema is SQL based.

As mentioned before in the Introduction section, this research concentrates on assets that consume energy. Therefore, all data sets related to assets that consume energy are used to develop the ontologies in the next stages. The developed taxonomy by Farghaly *et al.* (2018) was chosen for the refinement of the data sets. The taxonomy classified the assets consume energy to nine main categories, namely; water heating, ventilation, refrigeration, lighting, electronics, kitchen, computers, space cooling/heating and others. Each category contains the related assets (Farghaly *et al.*, 2018).

### *4.3 Define resources naming strategy*

The vocabularies used to represent data are a key to form Linked Data. Meanwhile, one of main principles of Linked Data states that URIs must be used for naming resources such as vocabularies and terms. In this section, the strategies to define the URIs for generating resources are discussed.

There are two main forms of URI, namely; slash URI and hash URI. In slash URIs, the resource is accessed as individual or group. While the hash URIs contain a fragment, which separates the normal URI and fragment identifier by a hash character (#). Most known available domains are Semantic Web, DBpedia and RDF-ized version of Wikipedia (Heath and Bizer, 2011).

Accordingly and armed with the tips provided by Heath and Bizer (2011), as the data set will contain a significant amount of data and it can grow in the future, slash URIs are adopted for data sets. However, smaller amount of data is entered in the development of ontologies, therefore, the hash URIs are used. The ontologies will have the path form /ontology/ < ontologyName > # < className > for classes and /ontology/ < ontologyName > # < propertyName > . The domain of semanticweb.org is selected to adapt the developed ontologies and the naming conventions of the classes are written as the selected construction standards and classifications.

### *4.4 Development of ontologies*

The ontology is developed through seven different steps (Noy and Mcguinness, 2001). The first step is to define the requirements that have to be fulfilled by the ontology. As the research concentrates only on assets that consume energy, the developed ontologies will only cover these assets. Since the data sets for that scope is small and the speed of processing is not an issue, the Turtle serialization was selected because it is easy to read by humans. Most of data sets are available in PDF format, therefore the ontologies have been created from scratch using Protégé 5.2 as ontology editor.

The second step is to consider reusing existing ontologies. Abanda *et al.* (2017) developed an ontology based on NRM1 concepts to facilitate the cost estimation process in AEC industry. Pauwels and Terkaj (2016) developed an ontology based on IFC EXPRESS schema to allow the conversion of IFC instance files into equivalent RDF graphs. These existing ontologies were utilized in the ontologies development by referencing them, instead of by importing the existing ontologies as a whole. It has been observed that developing ontologies with the required classes for the research scope is simpler and quicker than importing all the existing ontologies.

The third step is to enumerate important terms in the ontology. In this step, terms are extracted to form a list of concepts (classes, relationships and slots) from the data schema regardless any overlap between concepts they represent. The names of the selected terms have followed the resource naming strategy.

The fourth step is to define the classes and develop the class hierarchy. Several approaches can be used for developing class hierarchy, namely; top-down, bottom-up and combination. In this research, the top-down approach was adapted in the development of the class hierarchy in the different proposed ontologies. For example, as the schema of the NRM1 and NRM3 is available in a tabular format in PDF documents. The ontology concepts were extracted manually from the elemental work breakdown structure (WSP) illustrated in the NRM documents. The WSP includes eight different classes/concepts which were adopted as the top level concepts, namely; Group 1: substructure; Group 2: superstructure; Group 3: internal finishes; Group 4: fittings, furnishes and equipment; Group 5: services; Group 6: prefabricated buildings and building units; Group 7: work to existing buildings and Group 8: external works. Concepts were categorized into four hierarchical levels. The second level concepts were obtained from the immediate breakdown of first level concepts as in the NRMs. The third and fourth concepts were obtained from the first and fourth columns, respectively, from the tables under each second level concept. Columns four and five in the NRM1 and NRM3 tables represent the included and excluded elements, respectively.

The fifth and sixth steps are closely interconnected and usually are done together. The fifth step is to define properties of classes (slots). The definition of classes alone will not provide enough information to answer competency questions asked in Step 1. Slots describe the internal structure of concepts in ontology and they have to be attached to the most general class that can have these properties. While the sixth step is to define the facets of the slots. The values of slots are described in different facets such as; value type, allowed values, cardinality and other facet features. The value type facet can be described in different value types such as; string, number, Boolean and enumerated. Allowed values facets define the range of slot and cardinality facets define how many values the slot can have. In this research, the properties attached to the classes and subclasses described value type facet, and the string and number value types are used for defining most of the slots. Finally, the seventh step is to create the instance of classes in the hierarchy. This step consists requires three tasks, respectively, which are choosing the class, creating the instance of that class and finally filling in the slot's values. Figure 3 represents a screenshot for the developed ontologies using Protégé 5.2.0.

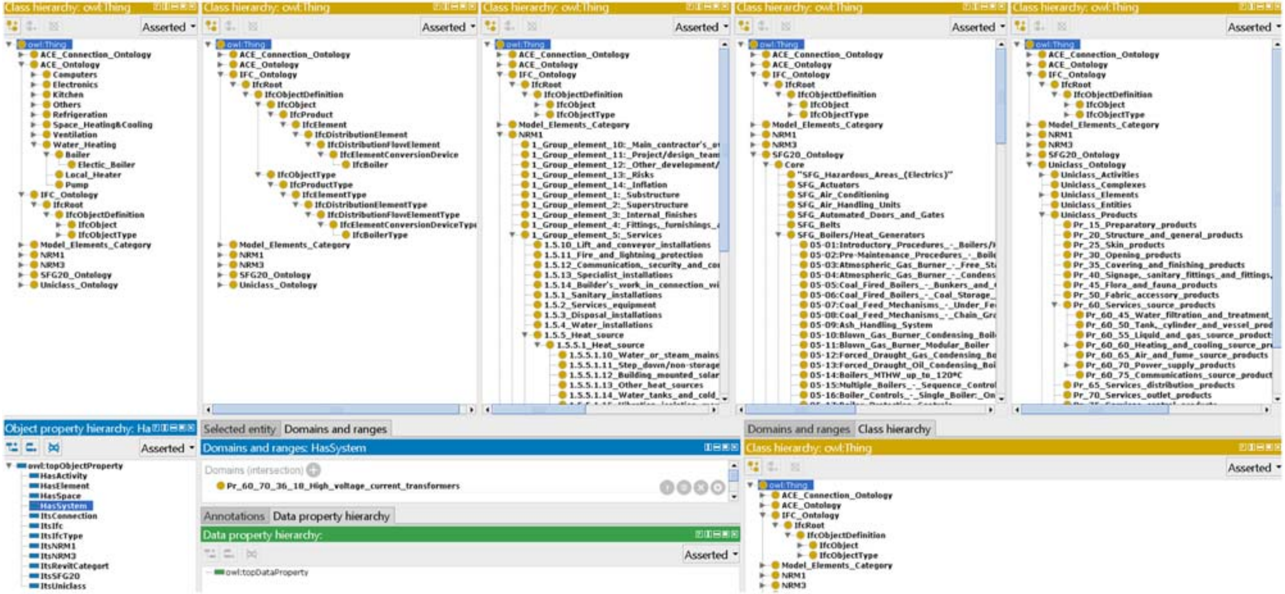
### *4.5 Link the ontologies*

Linked Data relies on setting RDF links between URI aliases in order to be able to track the different information providers refer to the same asset. This stage aims to make visible indicators that have not been previously harvested such as: the interconnections, incoming and outgoing links between vocabularies/classifications. In the end of the first focus group, the link between the classes of the different classifications has been documented. Figure 4 illustrates the link and mapping between the different ontological sources for a specific asset (Boiler). The hierarchical sequence for each standard is represented to reach the class to symbolize electrical boiler. The different relationships are colour-coded depicted in the bottom left of Figure 4.

### *4.6 Evaluate the ontologies*

Ontology evaluation is vital activity to assure that what is developed meets the application requirement (Gomez-Perez *et al.*, 2006). In this step, syntactic and semantic correctness of the developed ontologies and also the mapping between the different ontologies have to be verified. Also the developed ontology has to be evaluated against the purpose of development. Logical-based approach is adapted for syntactic evaluation. Several SPARQL queries are executed to observe the credibility of obtained results and HermiT reasoner was used to validate the consistency with the used ontologies. HermiT is reasoner for ontologies written using the OWL and it is preinstalled Protégé plug-in. Meanwhile, Manchester OWL syntax validator was used to evaluate the OWL syntax compliance for the developed ontologies.

After the syntactic evaluation, the ontology has been revised and the final version deemed to reflect practice was semantically correct. Feature-based approach is adapted for semantic evaluation of the developed ontologies. Feature-based approach evaluates the ontologies quality by engaging users and expertise. In the final revision, a new ontology is added which represents the connection of the asset to electricity. The subclasses of this ontology are socket, diffuser, isolator, control panel and battery. An object property was added to map the connection to electricity ontology and the main ACE ontology. Also, during the focus group, several challenges and gaps have been discussed between the author and the expertise which need to overcome to achieve appropriate cross linking between different sources. The first challenge highlighted was that the export option of IFC in Revit cannot fulfil the required link by expertise. By default, Revit exports building elements to an IFC file based on the categories (and subcategories) to which the elements belong based on the mapping defined by International Alliance for Interoperability data exchange standards. However, it is required to map with Revit families instead. The authors highlighted it the participants that: on the Autodesk knowledge network website, a solution is proposed to overcome that challenge. The solution requires adding two shared parameters named IFCExportAs and IFCExportType where the Revit user needs to fulfil with the class of IFC for each family. The second challenge is exporting the whole model which is timing consuming and



```
owl:Thing
  ACE_Connection_Ontology
  ACE_Ontology
    Computers
    Electronics
    Kitchen
    Others
    Refrigeration
    Space_Heating&Cooling
    Ventilation
    Water_Heating
      Boiler
        Electic_Boiler
      Local_Heater
      Pump
  IFC_Ontology
    IfcRoot
      IfcObjectDefinition
        IfcObject
        IfcObjectType
  Model_Elements_Category
  NRM1
  NRM3
  SFG20_Ontology
  Uniclass_Ontology
```

```
owl:Thing
  ACE_Connection_Ontology
  ACE_Ontology
  IFC_Ontology
    IfcRoot
      IfcObjectDefinition
        IfcObject
          IfcProduct
            IfcElement
              IfcDistributionElement
                IfcDistributionFlowElement
                  IfcElementConversionDevice
                    IfcBoiler
        IfcObjectType
          IfcProductType
            IfcElementType
              IfcDistributionElementType
                IfcDistributionFlowElementType
                  IfcElementConversionDeviceTyp
                    IfcBoilerType
  Model_Elements_Category
  NRM1
  NRM3
  SFG20_Ontology
  Uniclass_Ontology
```

```
owl:Thing
  ACE_Connection_Ontology
  ACE_Ontology
  IFC_Ontology
    IfcRoot
      IfcObjectDefinition
        IfcObject
        IfcObjectType
  Model_Elements_Category
  NRM1
    1_Group_element_10:_Main_contractor's_o
    1_Group_element_11:_Project/design_team
    1_Group_element_12:_Other_development/
    1_Group_element_13:_Risks
    1_Group_element_14:_Inflation
    1_Group_element_1:_Substructure
    1_Group_element_2:_Superstructure
    1_Group_element_3:_Internal_finishes
    1_Group_element_4:_Fittings,_furnishings_
    1_Group_element_5:_Services
      1.5.10_Lift_and_conveyor_installations
      1.5.11_Fire_and_lightning_protection
      1.5.12_Communication,_security_and_co
      1.5.13_Specialist_installations
      1.5.14_Builder's_work_in_connection_wi
      1.5.1_Sanitary_installations
      1.5.2_Services_equipment
      1.5.3_Disposal_installations
      1.5.4_Water_installations
      1.5.5_Heat_source
        1.5.5.1_Heat_source
          1.5.5.1.10_Water_or_steam_mains
          1.5.5.1.11_Step_down/non-storage
          1.5.5.1.12_Building_mounted_solar
          1.5.5.1.13_Other_heat_sources
          1.5.5.1.14_Water_tanks_and_cold_
```

```
owl:Thing
  ACE_Connection_Ontology
  ACE_Ontology
  IFC_Ontology
    IfcRoot
      IfcObjectDefinition
        IfcObject
        IfcObjectType
  Model_Elements_Category
  NRM1
  NRM3
  SFG20_Ontology
    Core
      "SFG_Hazardous_Areas_(Electrics)"
      SFG_Actuators
      SFG_Air_Conditioning
      SFG_Air_Handling_Units
      SFG_Automated_Doors_and_Gates
      SFG_Belts
      SFG_Boilers/Heat_Generators
        05-01:Introductory_Procedures_-_Boilers/
        05-02:Pre-Maintenance_Procedures_-_Boil
        05-03:Atmospheric_Gas_Burner_-_Free_St
        05-04:Atmospheric_Gas_Burner_-_Condens
        05-05:Coal_Fired_Boilers_-_Bunkers_and_
        05-06:Coal_Fired_Boilers_-_Coal_Storage_
        05-07:Coal_Feed_Mechanisms_-_Under_Fe
        05-08:Coal_Feed_Mechanisms_-_Chain_Gr
        05-09:Ash_Handling_System
        05-10:Blown_Gas_Burner_Condensing_Boil
        05-11:Blown_Gas_Burner_Modular_Boiler
        05-12:Forced_Draught_Gas_Condensing_Bo
        05-13:Forced_Draught_Oil_Condensing_Boi
        05-14:Boilers_MTHW_up_to_120°C
        05-15:Multiple_Boilers_-_Sequence_Control
        05-16:Boiler_Controls_-_Single_Boiler:_On
```

```
owl:Thing
  ACE_Connection_Ontology
  ACE_Ontology
  IFC_Ontology
    IfcRoot
      IfcObjectDefinition
        IfcObject
        IfcObjectType
  Model_Elements_Category
  NRM1
  NRM3
  SFG20_Ontology
  Uniclass_Ontology
    Uniclass_Activities
    Uniclass_Complexes
    Uniclass_Elements
    Uniclass_Entities
    Uniclass_Products
      Pr_15_Preparatory_products
      Pr_20_Structure_and_general_products
      Pr_25_Skin_products
      Pr_30_Opening_products
      Pr_35_Covering_and_finishing_products
      Pr_40_Signage,_sanitary_fittings_and_fittings,
      Pr_45_Flora_and_fauna_products
      Pr_50_Fabric_accessory_products
      Pr_60_Services_source_products
        Pr_60_45_Water_filtration_and_treatment_
        Pr_60_50_Tank,_cylinder_and_vessel_prod
        Pr_60_55_Liquid_and_gas_source_product
        Pr_60_60_Heating_and_cooling_source_pr
        Pr_60_65_Air_and_fume_source_products
        Pr_60_70_Power_supply_products
        Pr_60_75_Communications_source_product
      Pr_65_Services_distribution_products
      Pr_70_Services_outlet_products
```

```
owl:topObjectProperty
  HasActivity
  HasElement
  HasSpace
  HasSystem
  ItsConnection
  ItsIfc
  ItsIfcType
  ItsNRM1
  ItsNRM3
  ItsRevitCategort
  ItsSFG20
  ItsUniclass
```

```
Pr_60_70_36_10_High_voltage_current_transformers
```

```
owl:topDataProperty
```

```
owl:Thing
  ACE_Connection_Ontology
  ACE_Ontology
  IFC_Ontology
    IfcRoot
      IfcObjectDefinition
        IfcObject
        IfcObjectType
  Model_Elements_Category
  NRM1
  NRM3
```

**Figure 3.**
Screenshot from Protégé 5.2.0 for the developed ontologies

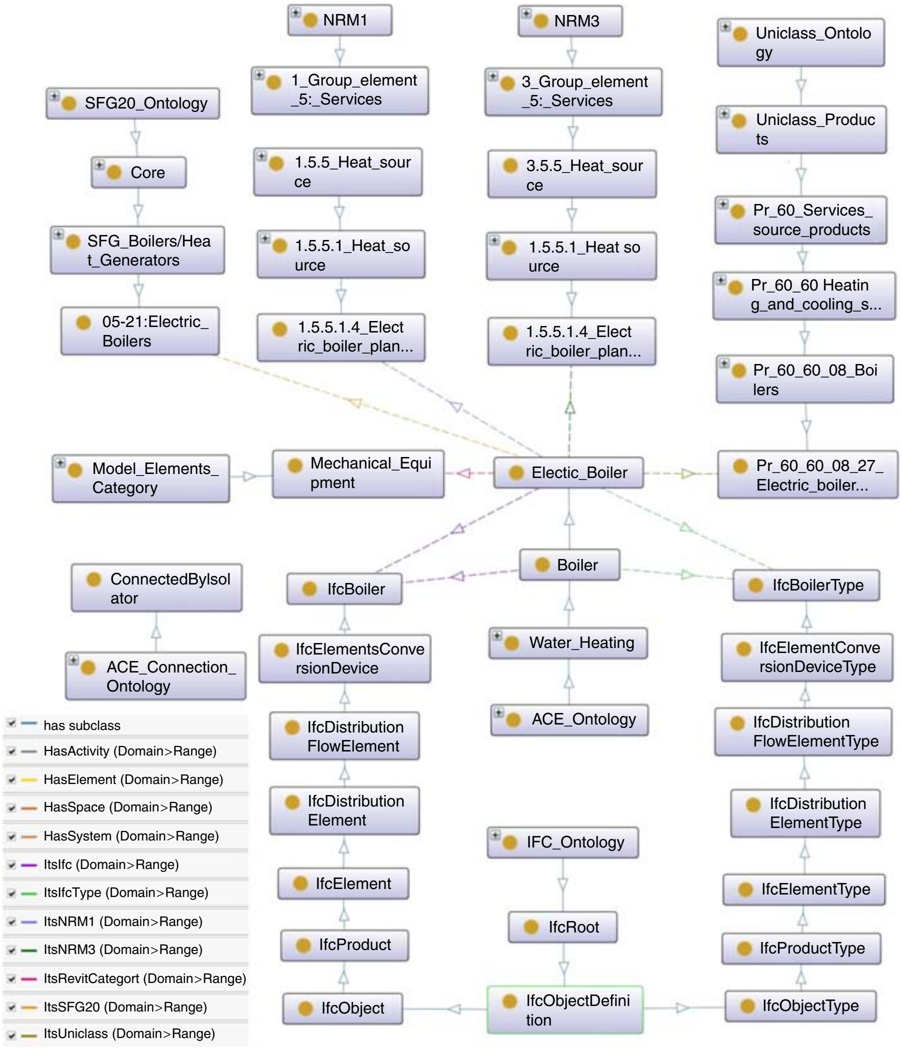


**Figure 4.**
Example of linked classifications for an electrical boiler

not required. One of the participants suggested to add two parameters for the Revit elements. The first parameter identifies, first, if the asset is maintainable or not, while the second parameter classifies the asset's importance from the operation point of view. The asset importance parameter was suggested to be a Camel Case string type. Using the two parameters, we can distinguish which assets required to be exported and consequently develop an MVD to achieve the purpose. The third highlighted point is linking the different tables of Uniclass2 to each other's.

### *4.7 Transform data sets*

The developed ontologies and the defined resource naming strategy are used in the transformation process of data into RDF format. Figure 5 illustrates the proposed framework for

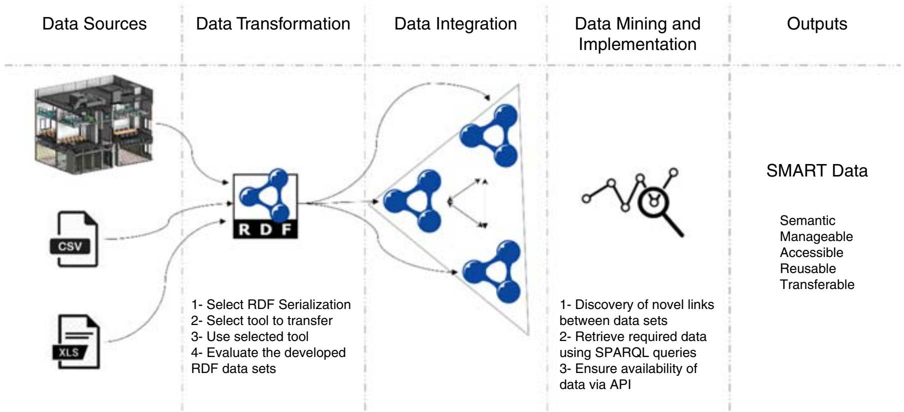

Data Sources
Data Transformation
Data Integration
Data Mining and Implementation
Outputs
CSV
XLS
R D F
SMART Data
Semantic
Manageable
Accessible
Reusable
Transferable
1- Select RDF Serialization
2- Select tool to transfer
3- Use selected tool
4- Evaluate the developed RDF data sets
1- Discovery of novel links between data sets
2- Retrieve required data using SPARQL queries
3- Ensure availability of data via API

**Figure 5.** Proposed framework for data transformation

data transformation in this research. Three different data sources were selected for our case study, namely; the BIM models, the operations and maintenance schedules and the procurement documents. The Turtle serialization was selected as the research RDF serialization. Turtle serialization is easily readable by humans and the research data sets are small which would not affect the speed of processing. Since the data are available in the CSV and XML format, OpenRefine with the RDF extension for transforming the data into RDF is selected. The OpenRefine tool is widely known in the community and it is easy to use. To achieve the transformation, a mapping between the data and the ontology has to be defined taking in consideration the defined resource naming strategy. This has been achieved in several tasks, first, initial transformations to the data are made in order to correct errors. Second, mapping the datasets (the columns and rows in the table) to the developed ontology taking in consideration the resource naming strategy. Consequently, the RDF syntax is chosen and the data sets are generated and evaluated semantically and syntactically.

After the conversion of data sets into RDF formats, they are now machine readable and more interoperable as they are represented using the defined standardized vocabularies. To cross-link the different data sets, the authors manually have mapped the different assets to their related information in the several data sets. With different data sets inter-linked to each other, the discovery of novel links between the data sets and the retrieve of required data are now possible using SPARQL queries. Also, it will provide a potential for the availability of required information in the BIM models using Application Programming Interface to import required data in the BIM platforms. Finally, the proposed framework can provide outputs which are characterized with maximally semantical, manageable, accessible, reusable and transferable data.

## 5. Conclusion and further works

The holistic management and maintenance of assets is a multi-domain problem encompassing data from different sources such as building information models, sensors, assets databases. To effectively conduct AM-related activities, identifying required information alone will not achieve the purpose. Cross-domain integration with clear definition of the required information would be more effective to systematically manage activities for assets' O&M.

The IFC format has been utilized as the format for providing information exchange between BIM and AM, however, it still presents many challenges. Although the IFC schema is a rich and vast data model that can contain the required data for different applications and needs in the AECO domain, the facilities managers do not normally use them, as IFC models either do not contain the required information or they contain superfluous information which makes it difficult to extract the required ones. BIM already has been moving in the direction of knowledge processing, with the development of ifcOWL, thus being able to leverage web Linked Data as a tool to extend interoperability to other knowledge domains, which were not previously considered.

This research presented a process map for Linked Data generation for building assets to improve asset management. By providing a detailed description of all the tasks and related tools and technologies in the generation and publication processes, the developed process map can help both owners and facilities mangers to manage the building assets information from different databases with semantically Linked Data in the Semantic Web. The different developed ontologies reused terms of widely deployed vocabularies and standards in AECO industry such as IFC, NRM1, NRM3, Uniclass2 and SFG20. Vocabularies linking using schema-level constructs of classes and properties provides a shared knowledge representative of a conceptual model. The different classes of the ontologies from the selected standards were object/asset based linked and mapped. The proposed process map aims to help researchers and practitioners interested in managing and operating building assets without authorization of BIM platforms and with exploiting Linked Data technologies. Although it is possible to create the same mapping system using traditional SQL-based technology, the usage of the Linked Data approach can provide a foundation for enabling modularity of new technologies and future extensions for the systems.

Once the mapping has been executed, a case study for an educational building has been conducted to evaluate the proposed mapping. The case study had several limitations due to the absence of required sensors and databases to provide information related to different aspects such as the occupant behaviour, rooms temperature and the lectures schedules. Future work including several use cases will be conducted to ensure the outputs of the proposed processes and corresponding tools and techniques. Authors believe that publishing assets information from the different data sets as RDF along with the developed ontologies can provide both the syntactic and semantic integration of BIM data and other assets ontological sources as long as promised by Semantic Web technologies.

## References

Abanda, F.H., Kamsu-Foguem, B. and Tah, J. (2017), “BIM – new rules of measurement ontology for construction cost estimation”, *Engineering Science and Technology, an International Journal*, Vol. 20 No. 2, pp. 443-459.

Abanda, F.H., Tah, J.H. and Keivani, R. (2013), “Trends in built environment Semantic Web applications: where are we today?”, *Expert Systems with Applications*, Vol. 40 No. 14, pp. 5563-5577.

Becerik-Gerber, B., Jazizadeh, F., Li, N. and Calis, G. (2012), “Application areas and data requirements for BIM-enabled facilities management”, *Journal of Construction Engineering and Management*, Vol. 138 No. 3, pp. 431-442.

Berners-Lee, T. (2006), “Linked data-design issues”, available at: www.w3.org/DesignIssues/LinkedData.html (accessed 22 June 2018).

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), “The Semantic Web”, *Scientific American*, Vol. 284, pp. 28-37.

Corry, E., O’donnell, J., Curry, E., Coakley, D., Pauwels, P. and Keane, M. (2014), “Using Semantic Web technologies to access soft AEC data”, *Advanced Engineering Informatics*, Vol. 28 No. 4, pp. 370-380.

Curry, E., O’donnell, J., Corry, E., Hasan, S., Keane, M. and O’Riain, S. (2013), “Linking building data in the cloud: integrating cross-domain building data using linked data”, *Advanced Engineering Informatics*, Vol. 27 No. 2, pp. 206-219.

Farghaly, K., Abanda, F.H., Vidalakis, C. and Wood, G. (2018), "Taxonomy for BIM and asset management semantic interoperability", *Journal of Management in Engineering*, Vol. 34 No. 4, p. 04018012, doi:10.1061/(ASCE)ME.1943-5479.0000610.

Farghaly, K., Abanda, H., Vidalakis, C. and Wood, G. (2017), "BIM big data system architecture for asset management: a conceptual framework", *Proceedings of Lean & Computing in Construction Congress (LC 3), Heraklion, October 31-November 2 2016.*

Gomez-Perez, A., Fernández-López, M. and Corcho, O. (2006), *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*, Springer Science & Business Media, London.

Heath, T. and Bizer, C. (2011), "Linked data: evolving the web into a global data space", *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 1 No. 1, pp. 1-136.

Hu, S., Corry, E., Horrigan, M., Hoare, C., Dos Reis, M. and O'Donnell, J. (2018), "Building performance evaluation using openmath and linked data", *Energy and Buildings*, Vol. 174, pp. 484-494.

Ibrahim, K.F., Abanda, F.H., Vidalakis, C. and Woods, G. (2016), "BIM for FM: input versus output data", *Proceedings of the 33rd CIB W78 Conference 2016, Brisbane, July 4-12 2017.*

Kim, K., Kim, H., Kim, W., Kim, C., Kim, J. and Yu, J. (2018), "Integration of ifc objects and facility management work information using Semantic Web", *Automation in Construction*, Vol. 87, pp. 173-187.

Kučera, A. and Pitner, T. (2018), "Semantic BMS: allowing usage of building automation data in facility benchmarking", *Advanced Engineering Informatics*, Vol. 35, pp. 69-84.

Lee, D.-Y., Chi, H.-L., Wang, J., Wang, X. and Park, C.-S. (2016), "A linked data system framework for sharing construction defect information using ontologies and BIM environments", *Automation in Construction*, Vol. 68, pp. 102-113.

Li, X., Wu, P., Shen, G.Q., Wang, X. and Teng, Y. (2017), "Mapping the knowledge domains of building information modeling (BIM): a bibliometric approach", *Automation in Construction*, Vol. 84, pp. 195-206.

Maxwell, J.A. (2008), "Designing a qualitative study", *The SAGE Handbook of Applied Social Research Methods*, Vol. 2, pp. 214-253.

Noy, N.F. and Mcguinness, D.L. (2001), "Ontology development 101: a guide to creating your first ontology", Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, Stanford, CA.

O'donnell, J., Corry, E., Hasan, S., Keane, M. and Curry, E. (2013), "Building performance optimization using cross-domain scenario modeling, linked data, and complex event processing", *Building and Environment*, Vol. 62, pp. 102-111.

Pärn, E., Edwards, D. and Sing, M. (2017), "The building information modelling trajectory in facilities management: a review", *Automation in Construction*, Vol. 75, pp. 45-55.

Pauwels, P. and Terkaj, W. (2016), "EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology", *Automation in Construction*, Vol. 63, pp. 100-133.

Pauwels, P., Zhang, S. and Lee, Y.-C. (2017), "Semantic Web technologies in AEC industry: a literature overview", *Automation in Construction*, Vol. 73, pp. 145-165.

Radulovic, F., Poveda-Villalón, M., Vila-Suero, D., Rodríguez-Doncel, V., García-Castro, R. and Gómez-Pérez, A. (2015), "Guidelines for linked data generation and publication: an example in building energy consumption", *Automation in Construction*, Vol. 57, pp. 178-187.

RICS (2012), *NRM 1: Order of Cost Estimating and Cost Planning for Capital Building Works RICS New Rules of Measurement*, Royal Institution of Chartered Surveyors, London.

Schmachtenberg, M., Bizer, C. and Paulheim, H. (2014), "Adoption of the linked data best practices in different topical domains", *13th International Semantic Web Conference Riva del Garda, October 19-23, 2014 Proceedings, Part I*, Springer, pp. 245-260.

Shadbolt, N., Berners-Lee, T. and Hall, W. (2006), “The Semantic Web revisited”, *IEEE Intelligent Systems*, Vol. 21 No. 3, pp. 96-101.

Sherman, R. (2015), “Chapter 12 – data integration processes”, in Sherman, R. (Ed.), *Business Intelligence Guidebook*, Morgan Kaufmann, Boston, MA, pp. 301-333.

Teicholz, P. (2013), *BIM for Facility Managers*, John Wiley & Sons, New Jersey.

Villazon-Terrazas, B., Vila-Suero, D., Garijo, D., Vilches-Blazquez, L.M., Poveda-Villalon, M., Mora, J., Corcho, O. and Gomez-Perez, A. (2012), “Publishing linked data-there is no one-size-fits-all formula, European Data Forum 2012, 6 June, Copenague, Dinamarca.

Zadeh, P.A., Wang, G., Cavka, H.B., Staub-French, S. and Pottinger, R. (2017), “Information quality assessment for facility management”, *Advanced Engineering Informatics*, Vol. 33, pp. 181-205.

**Corresponding author**
Karim Farghaly can be contacted at: karim.ibrahim-2016@brookes.ac.uk

# Influence of Big Data in managing cyber assets

Amit Mitra
*Bristol Business School, University of the West of England, Bristol, UK, and*
Kamran Munir
*Department of Computer Science and Creative Technologies, University of the West of England, Bristol, UK*

Received 17 July 2018
Revised 21 January 2019
16 May 2019
Accepted 27 May 2019

## Abstract

**Purpose** – Today, Big Data plays an imperative role in the creation, maintenance and loss of cyber assets of organisations. Research in connection to Big Data and cyber asset management is embryonic. Using evidence, the purpose of this paper is to argue that asset management in the context of Big Data is punctuated by a variety of vulnerabilities that can only be estimated when characteristics of such assets like being intangible are adequately accounted for.

**Design/methodology/approach** – Evidence for the study has been drawn from interviews of leaders of digital transformation projects in three organisations that are within the insurance industry, natural gas and oil, and manufacturing industries.

**Findings** – By examining the extant literature, the authors traced the type of influence that Big Data has over asset management within organisations. In a context defined by variability and volume of data, it is unlikely that the authors will be going back to restricting data flows. The focus now for asset managing organisations would be to improve semantic processors to deal with the vast array of data in variable formats.

**Research limitations/implications** – Data used as evidence for the study are based on interviews, as well as desk research. The use of real-time data along with the use of quantitative analysis could lead to insights that have hitherto eluded the research community.

**Originality/value** – There is a serious dearth of the research in the context of innovative leadership in dealing with a threatened asset management space. Interpreting creative initiatives to deal with a variety of risks to data assets has clear value for a variety of audiences.

**Keywords** Social media, Databases, Forecasting, Estimation

**Paper type** Research paper

## 1. Introduction

Assets are at the heart of the capacity of organisations that can be leveraged to acquire sustainable competitive advantages. Although traditional conceptualisations of assets purport to physical characterisations, yet resource richness of organisations is increasingly based on intangible digital assets held in clouds (Mitra and O'Regan, 2019). With ever increasing interactive use through stakeholder participation, the size of such intangible data assets is likely to snowball in volume. Instant nature of interactions and data exchanges that are all in real time tends also to ratchet up the velocity of such data usage. Almost unique in value contribution to such assets is the multiplicity of data formats that is premised on a diverse range of platforms through which variability of data generation is envisaged (McCreary and Kelly, 2013).

As organisations transform themselves digitally (Westerman *et al.*, 2014), we can discern a gradual shift in the way that they value assets. Although physical assets have been the cornerstone of traditional business, today data and consequent digital assets would be the institutional preference. It is clear that as the expectation of customers to be able to access assets 24/7 has become an accepted mode of practice, organisations are also swiftly moving to convert their physical assets into digital assets. Damage to physical assets like building and infrastructure can be costly and can be replaced through reconstruction; on the contrary, data losses or corruption of data can result in even larger losses and can be harder to replace (Mitra *et al.*, 2018). Losses because of data loss can

Built Environment Project and Asset Management
Vol. 9 No. 4, 2019
pp. 503-514
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-07-2018-0098

have debilitating effects on the organisation and can create abiding vulnerabilities for organisations in a market where information assurance tends to play a key role in bolstering trust. The latter would lure customers of buying into the products and services on offer and, thus, extend trade.

The nature of vulnerabilities of the two types of assets could not be any more distinct. Although damage from fire, natural calamities and destruction would be the norm of susceptibilities of physical assets like buildings, digital assets are prone to a myriad of malware attacks as well as are threatened by hackers who may make illegal use of these. At the end of the day, although a small number of people could be affected through the destruction of physical assets, millions of peoples' lives could get adversely affected when digital assets are compromised. Beale (2018) talking about the insurance provision of Lloyds of London for cloud-based digital assets mentioned that digital assets are stored in the cloud supervised by a few firms. If one of the cloud providers was to come down because of a cyber-attack for three days, it would affect 12.5m business users in the USA alone. The size of the population who would be affected tells us of the enormity of the risk that digital asset loss could bring upon us. So, the reality that stares us on our face is one which is complex – here, we are compelled to translate our physical assets into digital ones, on the one hand, and, on the other, compound the risk of loss due to the vulnerabilities that such digital assets inherently contain.

Beale (2018) went on to narrate another instance, which is likely to provide a glimpse of some of the challenges of the type of compromises that assets management might unknowingly bump into:

> An employee in a chain of opticians received an email saying she had been caught speeding on camera. She clicked on the link and it offered to show her a photograph of her being caught in the act of speeding. This was a cyber attack as the email was not genuine. By clicking on the link, she triggered a virus that infected all the files on her company's servers. Then she received the email that said, your files are all encrypted, and we need a fee from you, payable in bitcoin to unlock the encryption. The files contained sensitive patient records as well as the software to run the business. Without access to them the business couldn't operate. They had no choice but to pay up to these hackers, whoever they were. The company's insurer's paid for the ransom and provided the reimbursement for the entire costs for getting up and running. But of course, it didn't end there because the encryption key that the hackers then released only covered 90% of the files and the company needed an IT contractor to come in and rebuild and recover the remainder of them. The company eventually got up and running again but it was a traumatic experience for everyone involved.

The abovementioned account vividly demonstrates vulnerabilities of cyber assets and the way Big Data is the bed of an interconnected range of strengths and weaknesses. While data need to be freely available for companies to do business, at the same time, there are these constant threats from hackers who are omnipresent and growing in numbers with the passage of time. As mentioned before, Big Data enables organisations to maintain and grow digital assets while at the same time offering potential avenues to hackers who are waiting to pounce on unwitting asset managers.

The valuation of digital assets also tends to have a dynamic that in a way determines its net worth. This is where Big Data tends to play an imperative role in the creation and sustenance of such assets. The value of digital assets is usually borne out by among others a preponderance of customer information. Along with this, there are also the reviews and comments that customers write about their experiences of consuming products and services. There could also be organisations that engage with customers before designing their products (Mitra *et al.*, 2018). In any case, the size of user generated data is of such a magnitude that cloud-based infrastructure is the only appropriate platform to accommodate these sorts of unlimited volumes. The dynamic nature of these assets is also something that needs to be factored in if we are to effectively monitor their value.

Essentially, Big Data involves a couple of specific types of challenges. The first relates to the messiness as a result of heterogeneous sources of data along with the exponential growth of data within very short-time horizons. The second is about the semantic processes that will be employed to make sense, discern patterns when making sense of this vast volume of data. In tandem with these challenges, contemporary asset management is also not without its unique characteristics. First, increasingly, we need to deal with digital assets or data assets instead of physical infrastructure. Second, asset destruction or any kind of compromise is immensely harmful to firm reputation in any industry (Ransbotham *et al.*, 2016). Therefore, security and protection of Big Data assets usually resident in clouds are priority for organisations. Third, the greater the movement of data between different stakeholders, the higher the value addition (Hawlitschek *et al.*, 2018).

## 2. Review aim

Despite a growing interest in Big Data use by businesses, researchers know relatively little about estimating influence of specific characteristics of Big Data on cyber asset management. Industries such as insurance, natural gas and oil, and manufacturing can be characterised as generating Big Data that is central to efficient cyber asset management. Because of their unique products and services, cyber asset management within the mentioned industries today is tied up with vast cyber assets whose vulnerabilities are the target by interests that could profit from jeopardising the reliance of millions of peoples' livelihoods. Given recent historical and potentially increasing importance of Big Data for all types of organisations and industries and given the centrality of cyber asset management in organisational capacity, addressing issues of vulnerabilities of such Big Data assets is timely. The paper primarily examines the relationship between Big Data and asset management with specific reference to the insurance, oil and natural gas, and manufacturing sectors.

The paper is structured such that it can explore the relationship between Big Data and cyber asset management through anecdotal as well as extant literature-based evidence. Following the section on introduction, we have considered the nature of data from the perspective of challenges to format and processing. This if followed by data characteristics in the context of analytics that are now widely used. The next section is on analytics, and Big Data is primarily focused on extant literature implications. We briefly dwell on the methodology that we have followed to develop the paper next. Following the section on the extant literature, we have a section on discussion that essentially considers the principal relationship of Big Data and asset management. Finally, the section on conclusion is the next section in which we have considered the contribution that this paper makes, some of the key limitations and areas of further research that may be worth pursuing in future work.

## 3. Method

The literature review that is the central basis of this paper is entirely premised on desk research. The latter was undertaken to explore a couple of key dimensions of Big Data that have been alluded to in the previous sections. Unlike traditional data that conform to specific formats, Big Data is messy and exists in various forms. Inherently, this agglomeration of different types of data creates challenges for processing to come up with meaningful outputs. So, the characteristic messiness of Big Data was the first filter that we used to locate papers in the survey for this study. Within the extant literature, we selected papers that reported on the messiness of data spanning across a variety of formats. Second criterion we used to include papers in our survey is that of locating research that were focused on micro-dimensions of Big Data. So far, Big Data seems to be successfully used when it comes to macro-data. However, to locate person specific data, specific algorithms need to be written that are both costly as well as resource intensive to develop.

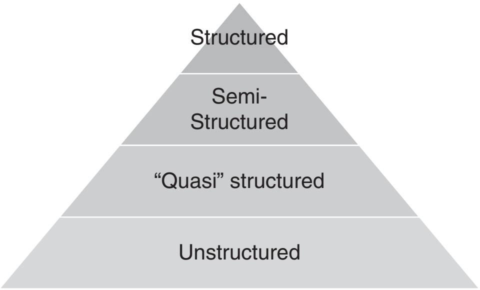
At the same time, there are three types of sources, namely, examination of organisations that have been influenced in their asset management by their use of some types of Big Data. The insurance industry, natural resources and manufacturing are the three sectors in which organisational Big Data instances have been drawn upon for this study. Second, we have also examined a wide range of the extant literature on Big Data and asset management. We did find that although there is the burgeoning literature in Big Data or for that matter literature with a specialist focus on asset management, the literature with a focus on the relationship between Big Data and its influence on asset management was either non-existent or was embryonic in nature. Third, we have drawn from anecdotal accounts of key managers in well-known organisations who are now managing their assets through extensive reliance on Big Data sources.

## 4. Data characteristics

The diagram in Figure 1 is aimed to illustrate the voluminous nature of Big Data and the ensuing types of data structures that are compatible with it. The diagram illustrates the increasingly unstructured nature of the growth of Big Data.

From a management perspective, the advent of Big Data has made it impossible to think of businesses in the same way as in the past, whereas traditional approaches have used analytics to understand and fine tune processes to keep management informed while alerting them to anomalies (business intelligence (BI) is driven by the idea of "exception reporting"). In contrast, Big Data has flipped such an analytical orientation on its head. The central view of Big Data is that the world and the data that it describes are in a state of change and flux, and those organisations that can recognise and react quickly have the upper hand in this space. The sought-after business and IT capabilities are discovery and agility, rather than stability (Davenport, 2014). Table I projects a few key differences between Big Data and traditional analytics.

The dimension of primary purpose of Big Data in Table I reveals how the entire industrial world is reorienting itself with regard to customer data. The traditional information



**Source:** Adapted from EMC Education Services (2015)

**Figure 1.** Big Data growth is increasingly unstructured

**Table I.** Big Data and traditional analytics

| Dimensions | Big Data | Traditional analytics |
|---|---|---|
| Type of data | Unstructured formats | Formatted in rows and columns |
| Volume of data | 100 terabytes to petabytes | Tens of terabytes or less |
| Flow of data | Constant flow of data | Static pool of data |
| Analysis methods | Machine learning | Hypothesis based |
| Primary purpose | Data-driven products | Internal decision support and services |

**Source:** Adapted from Davenport (2014) management approach has been to use analytics to cater a better internal decision making through the creation of reports and presentations that advise senior executives on internal decisions. In contrast, data scientists are today involved in data-driven products and services through the creation of customer facing and customer touching applications.

## 5. Analytics and Big Data

Generally, the business world could in theory question any differences between standard analytics and that of Big Data. However, there are three key dimensions that seem to make it obvious that Big Data would require different ways of capturing as well as novel techniques of analysis in comparison to traditional analytics. These three dimensions include volume, velocity and variability (McCreary and Kelly, 2013). The Big Data context of volume focuses on being able to process large amounts of data – here the guiding logic is always more data are better than smaller amounts of higher quality data. The primary considerations here relate to scalability, distribution, the ability to process the acquired data and the like. The speed at which data get generated is the dimension that relates to the velocity of Big Data. Obviously, here, we relate to the amount of time that is taken for action to be initiated after the receipt of that data. Issues of concern here include granularity of data streams, appreciating what is irrelevant and the amount of inactivity that may be tolerable in relation to data, decision making and action taking. In an interconnected world of numerous data sources through which data get generated – it is often unstructured, punctuated with errors, and inconsistent in nature. Relevant issues in this messiness context include amount of information loss in data clean up, semantic integration and versatility in representation (Lycett, 2013). As early as 2012, about 2.5 exabytes of data were generated each day and further this amount doubled every 40 months or so. Every second sees the passage of data across the internet compared to what was stored in the entire internet 20 years ago (McAfee and Brynjolfsson, 2012).

With the burgeoning growth of data assets in comparison to physical ones, we are also gradually moving into a world where intangibleness of assets is also becoming a key feature of the value of organisational capacity. All the feedback generated through customer comments on satisfaction levels along with counter comments of other users would today be considered as an important contributor to the asset value of a company. As a matter of fact, in manufacturing organisations like modern car manufacturers, the whole design process is becoming dependent on the engagement of prospective customers whose design expectations get embedded in the design phase of car model development (Mitra *et al.*, 2018). Data assets in the form of blog posts, tweets, and likes on Facebook all make up the gamut of data assets that are vital for the modern organisation that is trying to get near to customer expectations and in the process garner competitive advantages in an environment that expects swift reactions to online feedback. In a way, Big Data enriches the ties between customer and manufacturers as it strengthens the relationship between the two. In the contemporary context, every organisation wants to develop and strengthen customer relationship management (CRM) capacities so that they are able to sell all of their products without actually having to deal with unsold stock. If organisations can use the data generated by customer participation in social media, it might be feasible for them to then enhance their CRM capacities, leading to greater customer satisfaction and reduction of losses from excessive unsold stock.

Closely tied to the growth of intangible assets of an organisation is the need to be able to measure it. As Green (2008) by using an engineering concept has decomposed intangible assets by rendering them into their primitive formats of BI metadata to align with operational data. Green (2008) argues that by aligning operational data with BI implies that it might be feasible to create cross-tab views of data that can then be modelled into multidimensional views that are compatible with establishing accountability and valuation of tangible assets. Perhaps knowledge assets have a good deal of intangibility embedded in them. In a sense, Big Data could be the only mechanism to get to developing or assessing the value of intangible assets.

An example of intangible knowledge assets would be the accessibility of innovative approaches to develop solutions. During the first gulf war (circa 1990), there were a lot of Westland helicopters that were used by coalition forces in Iraq. Because the Iraq war was primarily fought in desert like environments, the amount of sand in the air was much higher than normal. In this operation, the Westland helicopter pilots of the Royal Airforce (RAF) had to deal with sand getting into the engine and thus the flights/sorties were disrupted and soon it was almost impossible to use these helicopters as they were not built to operate in these sandy conditions. The RAF personnel devised unique ways through which they could keep the helicopters flying despite the high percentage of sand in the air. So, after the Iraq war ended most of the helicopter pilots retired as it was normal for Royal Airforce pilots at the earmarked points in their forces career. However, with the passage of a decade, there was the second gulf war (circa 2003) when pilots were expected to operate Westland helicopters in amongst the desert sand. But this time all the intangible knowledge that was an operational asset was lost forever as there was no compilation of those experiences anywhere that could be used by pilots in the second gulf war.

The issue of garnering greater, deeper insights from Big Data to enhance competitive capacities of organisations is an abiding dimension of Big Data use that has also been reflected within the extant literature. However, Dutta and Bose (2015) are probably unique in their quest to conceptualise a framework through which analytics and project implementation using outcomes of data analysis could lead to effective implementation of big data projects in the realm of asset management. According to Bizer *et al.* (2011), there are essentially a couple of key issues with regard to successful implementation of Big Data use. Although the first relates to managing such exorbitant volume of data, the second has to be getting the right decoding mechanisms to locate patterns and make sense of the data. It is perhaps intriguing to appreciate that even though potential of exploitation of Big Data has been around for a while now yet hardly anything out of the ordinary seems to be undertaken by large conglomerates. For instance, it was reported by the Economic Intelligence Unit in 2015 that as high a proportion of firms as 56 per cent of manufacturing firms did not seem to have made much progress in using or applying Big Data to progress current business potential of the organisation. Lee *et al.* (2013) reported that there seems to be an upsurge in the uptake of technologies by manufacturing companies to use technologies such as advanced analytics and cyber physical approaches to augment productivity and make systems more effective. The remit of Big Data enables organisations to be ambitious in the way they can predict consumption patterns of consumers such that substantial advantages can be garnered by having exactly what customers might be looking for. Obviously, predictive analysis using Big Data is able to focus on a variety of sources that would not be considered relevant in normal data mining.

Dutta and Bose (2015) go on to devise a framework for implementing Big Data projects. But this framework seems particularly simplistic. As if it would be feasible to pre-define all stages to then take the implementation forward. In essence, Big Data given its three Vs, i.e. volume, velocity, variability, has a nature that seems unlikely to conform to linear progression. Therefore, it is very different from other types of technology projects that have hitherto been envisaged for organisations. Various stages have been mentioned by Dutta and Bose (2015) – these could have more clearly defined KPIs or measures by which you could assess if that part of the development conforms with expectations for the stage. Chang *et al.* (2014) have drawn attention to the fact that despite the advent of a multidimensional data frame through Big Data use, there is a need to recognise a paradigm shift in the way data are viewed and then used as evidence. Chang *et al.* (2014) have argued that theory continues to be relevant in the domain of data analytics. Something that has been taken for granted is the interdisciplinary nature of the data that need to be factored into the compilation and analysis of Big Data. Unlike the world of social media where demons are being created through the relentless generation of data, Chang *et al.* (2014) have stressed that there needs to be connections to some simple logic and understanding of the expectations of the audience of a business if the management were to succeed in garnering advantages through the use of vast amounts of data that Big Data provides. As McAfee and Brynjolfsson (2012) have pointed out in the end, it will be humans who will be responsible to design the processes, while discovering insights will be as a result of a combination of algorithm and system-based data analysis and intuition of people using the systems. So, following Chang *et al.* (2014), we could say that in the context of asset management the same type of mindset would work well whereby we use intuition and our knowledge of the assets and then marry this with the outcomes of Big Data analysis.

Kwon *et al.* (2014) have looked at benefit perceptions of adoption intentions of Big Data analytics by firms. Kwon *et al.*'s (2014) work is interesting from the differentiation between internal corporate and external source data use by firm's perspective and its connections to Big Data analytics that their empirical data seem to project. On the one hand, they contend that firm's intentions for Big Data analytics can be positively affected by its competence in maintaining the quality of corporate data. Furthermore, a firm's encouraging experience in using external data sources could promote the future acquisition of Big Data analytics. Remarkably, Kwon *et al.*'s (2014) work goes on to show that a firm's positive experience in using internal source data could hamper its adoption intention for Big Data analytics.

Whyte *et al.* (2016) looked at the role of big data and its role within asset management of complex projects. Using three organisations, namely Airbus, CERN and Crossrail, Whyte *et al.* (2016) pointed out that new challenges arise as asset information has become a project deliverable. as data increase in volume, velocity and variety and as it is aggregated and reused, with connections (and potential connections) across internally and externally held data sets. This dimension of asset management becoming a project deliverable is an important development in the context of Big Data. Such a change in expectations from a project perspective is probably indicative of the extensive role that metadata or dynamic data generation within the realm of Big Data is able to provide. As project management is becoming more customer directed, it is also clear that flexibility that Big Data provides is ultimately a sought-after attribute that attracts asset managers. As a matter of fact, Whyte *et al.*'s (2016) focus mainly stems from the flexibility that Big Data is able to provide in the context of complex projects. Change management can also be a complex process in a project setting that is again facilitated by Big Data. However, when integrity is central to a project, there can also be critical requirements of flexibility. So, Whyte *et al.* (2016) have unearthed the limits of flexibility within the context of complex projects when integrity is of paramount importance.

An important feature of contemporary project delivery is to do with managing change in asset information as data sets are aggregated and reused through life. Volume, velocity and variety of data bring new challenges of version control, linkages across project stages and with other data sets, and approaches to arranging and organising. Whyte *et al.* (2016) also pointed out that there might be increasing integration between data sets in project delivery, yet digital systems are not seamlessly connected but are heterogeneous with significant modifications in the use of data through the project life cycle.

A significant issue regarding asset management alluded to by Whyte *et al.* (2016) concerns users who do not follow prescribed processes involved in configuration management. At a time, when Big Data was not so expressly recognised as today, Mitra (2009) reported similar experiences when IT project management for the Commonwealth Games in Manchester revealed engineers' propensity to move away from prescribed solution approaches and to use personal preferences and choices to determine a unique path to create solutions for clients.

In Figure 2, we have shown how asset valuation may be envisaged to grow in a Big Data context. The greater the movement of assets among key stakeholders, the greater will be the propensity of valuation of the asset within specific industry contexts. The challenge of protection of assets from being compromised is also another important caveat when assets are dynamic and moving. So, although greater movement of assets is likely to generate higher valuation, at the same time, vulnerability of such assets being compromised would also gradually ratchet up. Asset characterisation would be reliant on the three Vs (McCreary and Kelly, 2013) as well as the type of semantic tools that are going to be applied to process the burgeoning data within specific assets.

## 6. Discussion

To answer the question of how Big Data influences asset management in organisations, we considered the use of a framework through which antecedents of cloud computing in multinational companies have been assessed. Borrowing from Mitra *et al.*'s (2018) study on resource-based valuation of cloud implementation among multinational companies, it is feasible to discern patterns within influences of Big Data on the way organisations manage assets. Mitra *et al.*'s (2018) study has special relevance here as cloud computing is usually the type of platforms on which Big Data is resident and second the whole issue of resource-based valuation is also clearly connected to assets of organisations. It is important to bear in mind that Big Data capacities enable asset management to be both focused on the macro- and the micro-dimensions of a business (Bizer *et al.*, 2011). Predictive analysis can substantially enhance an organisation's ability to acquire and generate assets. We agree with Boyd and Crawford (2012) that bigger data are not always better data. Without taking into account the sample of a data set, the size of the data set is meaningless. For instance, a researcher may seek to assess the topical frequency of tweets, at the same time, if Twitter was to remove all tweets that contain tricky words or content from the stream, the topical frequency would be erroneous.

### *6.1 Industry expectations*

It is clear from the assessment of evidence so far that as assets of organisations become digital in nature, their reliance on Big Data increases exponentially. However, this messiness of data is also a key feature of the burgeoning growth in data. Data generated by users are a significant contributor to Big Data. In turn, assets of organisations are also made up of the

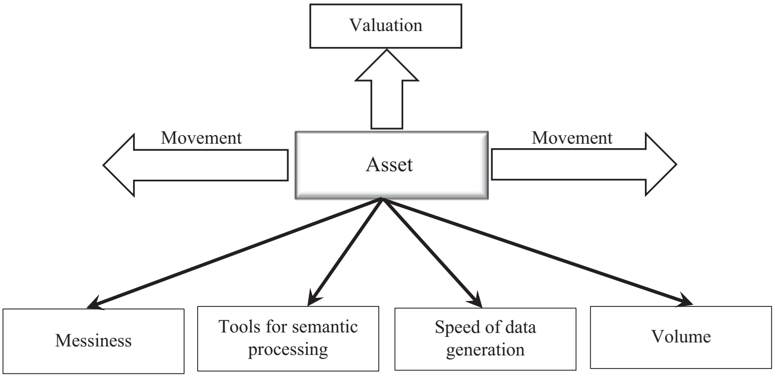


**Figure 2.** Dynamics of asset valuation

type of customer data that get generated through interaction among and with customers. As a matter of fact, the greater the volume of customer feedback, the greater will be the likely asset valuation. The type of semantic processors that will be used can also determine the eventual curation of meaning and enable organisations to direct offers that fit with customer expectations. Although user generated data are a boon to the valuation of organisations, at the same time "Without taking into account the sample of a data set, the size of the data set is meaningless" (Boyd and Crawford, 2012, p. 669). So, it is borne out by the analysis so far that ultimately protection of customer generated data would be an obvious way to protect assets in a fast moving data space. Using Green's (2008) arguments, it might be feasible to decompose intangible organisational assets that could then enable the creation of accountability of assets that is an industry expectation.

### *6.2 Process standardisation*

Cost reduction by creating standardised features that organisations use to deploy various functionalities is increasingly the goal of capacity development in organisations. Just as cloud computing enabled organisations to move away from excessive customisation (Mitra *et al.*, 2018) and in the process reduce costs, Big Data-based asset management can do both, i.e. it will be able to provide both a macro-visualisation of assets and, at the same time through the reliance on analytics it will be able to provide specific solutions for individual clients/customers for the organisation. As data generation becomes more oriented towards customer feedback so will it become possible for industries to fine tune offerings to individual audiences. So, here both process standardisation (Dutta and Bose, 2015) could likely be feasible along with the fact that there will not be any significant cost rises due to specific customisations.

### *6.3 Scalability*

Another type of development that affects large multinational organisations with assets that serve various countries around the world is the issue of scalability. With the reliance of cloud computing infrastructure organisations in the insurance industry, the oil and natural gas sector as well as manufacturing organisations would find it imperative to rely on Big Data analytical capacity. One of the multinational oil and natural gas companies that were part of the study conducted by Mitra *et al.* (2018) needed to scale their information management deployment to cater for 835 k employees who worked in more than seventy locations around the world. Here, also to reduce duplication of effort and to manage assets that are increasing in size almost every day, asset managers would be using Big Data approaches to make sense of global patterns that provide indicators for initiating action. Another key dimension here is speed of response. Probably, without semantic processors that are being used for the acquired Big Data on assets, it would be impossible to act and of course not being able to act quickly could lead to major calamities for organisations that have critical assets being deployed in such a way that they impact lives of users.

### *6.4 Investment optimisation*

Use of data banks across various locations around the world is the likely home of Big Data among all of the organisations that have been used for this study. All of these data banks are cloud repositories, and hence, they do not have capital expenses. In a way, operational expenses are what would be the driving logic for analytics that are applied on the assets to generate various insights that would then enable greater fit with customer or client expectations. Use of third-party resources to store Big Data and also use of third-party tools including NoSQL and Hadoop clusters for predictive analysis is something that is becoming common across most Big Data using organisations. So, here there are clear ways through which organisations can focus on creating reliable assets that they then use to query/mine using Big Data tools without much of a commitment of capital expenses up front.

### *6.5 Focus on core capacities*

Although information management has become an integral part of most of the industries included in this study as well as the fact that reliance on IT tools is also a clear part of all the organisations that we used, yet the organisations and their employees actually have certain competencies which are different than pure IT use (Chang *et al.*, 2014). With the use of semantic processors on the data assets, it is feasible that employees could get on with their actual specialism-related work while big data tools could work on developing insights using various pre-defined constructs. In a way, uniqueness of employee competencies could also lead to garnering specific advantages within specific industry contexts (Mitra and Neale, 2014). Of course, given the three Vs of Big Data (McAfee and Brynjolfsson, 2012), asset data are different from traditional data and would require almost continuous processing that would also enable the key players in the organisation to concentrate on core competencies while enriching the data with insights that are emerging through the Big Data analysis.

## 7. Conclusion

The paper has so far demonstrated that Big Data has clear and substantial influence in the way asset management is envisaged and administered. Although the three Vs are a principal mechanism of visualising Big Data related assets, yet both variability and the volume of data are quite substantial challenges that stakeholders have to grapple with. We are in a world where even when data are not born digital, and it becomes part of data assets through various semantic processors that are available within industry. But although there is considerable success that predictive analysis has brought to organisations, yet the messiness of data brought about by variability is something that is an ongoing challenge in Big Data-based asset management. Although customer expectations are being successfully gleaned through specific semantic processor, yet more confusing data get generated by the minute that then extends challenges for interpreting it. The other specific vulnerability of a data defined asset environment is its availability. Any time there is any attack that compromises Big Data on assets, organisations will suffer with consequences that will affect multiple markets. Beale's (2018) experiences in the insurance sector vividly projects some of these challenges both for individual organisations as well as for entire industries.

From a technology perspective, Big Data always seems to point to greater capacity to develop certainty about the asset management space. However, Boyd and Crawford (2012) have reminded us of the instinct of apophenia, i.e. seeing patterns where there are none. So, some of the confidence because of data mining capacities brought about by the access to large amounts of data might be misplaced. Boyd and Crawford (2012) has cited Leinweber's (2007) research in which it was shown how data mining techniques could show a strong but spurious correlation between changes in S&P 500 stock index and butter production in Bangladesh. Limited archiving capacities can also lead to uncertainties about historical data on assets. If Twitter and Facebook were considered as examples of Big Data sources, then they offer very poor archiving and search functions.

The paper by examining the extant literature traced the type of influence that Big Data has over asset management among organisations. It is clear that in a cloud-based world, we are likely to improve predictive analytics and there is no chance of reverting to static comparative static data sets any more. Also, in a world that is primarily defined by variability and volume of data, it is unlikely that we will be going back to restricting data flows, rather the focus now among asset managing organisations would be to improve semantic processors to deal with this vast volume of variable format data. The fact that we used only secondary data and anecdotal evidence has restricted the type of inferences we have been able to draw from the study. Second, we have considered industries that may be unique in the way they handle Big Data and so assuming that asset management will be in all likelihood be the same across the oil and natural gas, manufacturing and insurance industry might have been simplistic. At the same time, we have been referring to user generated or data on users of these industries so there might have been commonalities among assets. Use of more real time data aided by quantitative analysis could be a way forward for developing abiding insights into Big Data's influence on specific assets or organisations.

## References

Beale, I. (2018), "Enabling human progress in a digital world", *Bristol Distinguished Address Series Lecture Podcast*, 21 February, available at: www1.uwe.ac.uk/whatson/bristoldaseries/previoustalks/ingabealedbe.aspx

Bizer, C., Heath, T. and Berners-Lee, T. (2011), "Enabling scalable semantic reasoning for mobile services", in Sheth, A. (Ed.), *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, IGI Global, pp. 205-227.

Boyd, D. and Crawford, K. (2012), "Critical questions for big data", *Information, Communication & Society*, Vol. 15 No. 5, pp. 662-679.

Chang, R.M., Kauffman, R.J. and Kwon, Y.O. (2014), "Understanding the paradigm shift to computational social science in the presence of big data", *Decision Support Systems*, Vol. 63, pp. 67-80.

Davenport, T. (2014), *Big Data at Work: Dispelling Myths, Uncovering the Opportunities*, Harvard Business Review Press.

Dutta, D. and Bose, I. (2015), "Managing a big data project: the case of Ramco Cements Limited", *International Journal of Production Economics*, Vol. 165, pp. 293-306.

EMC Education Services (Ed.) (2015), *Data Science and Big Data Analytics. Discovering, Analyzing, Visualizing and Presenting Data*, John Wiley & Sons.

Green, A. (2008), "Intangible asset knowledge: the conjugality of business intelligence (BI) and business operational data", *VINE*, Vol. 38 No. 2, pp. 184-191.

Hawlitschek, F., Notheisen, B. and Teubner, T. (2018), "The limits of trust-free systems: a literature review on blockchain technology and trust in the sharing economy", *Electronic Commerce Research and Applications*, Vol. 29, pp. 50-63.

Kwon, O., Lee, N. and Shin, B. (2014), "Data quality management, data usage experience and acquisition intention of Big Data analytics", *International Journal of Information Management*, Vol. 34 No. 3, pp. 387-394.

Lee, J., Lapira, E., Bagheri, B. and Kao, H. (2013), "Recent advances and trends in predictive manufacturing systems in big data environment", *Manufacturing Letters*, Vol. 1 No. 1, pp. 38-41.

Leinweiber, D. (2007), "Stupid data miner tricks: overfitting the S&P 500", *The Journal of Investing*, Vol. 16 No. 1, pp. 15-22.

Lycett, M. (2013), " 'Datafication': making sense of (big) data in a complex world", *European Journal of Information Systems*, Vol. 22 No. 4, pp. 381-386.

McAfee, A. and Brynjolfsson, E. (2012), "Big Data: the management revolution", *Harvard Business Review*, Vol. 90 No. 10, pp. 61-68.

McCreary, D. and Kelly, A. (2013), *Making Sense of NoSQL*, Manning.

Mitra, A. (2009), "Evolution of an IS development effort: an ANT interpretation", *Journal of Systems and Information Technology*, Vol. 11 No. 2, pp. 150-167.

Mitra, A. and Neale, P. (2014), "Visions of a pole position: developing inimitable resource capacity through enterprise systems implementation in Nestlé", *Strategic Change*, Vol. 23 Nos 3/4, pp. 225-235.

Mitra, A. and O'Regan, N. (2019), "Creative leadership in the cyber asset market: an interview with Dame Inga Beale", *Journal of Management Inquiry*, doi: 10.1177/105649261982883.

Mitra, A., O'Regan, N. and Sarpong, D. (2018), "Cloud resource adaptation: a resource based perspective on value creation for corporate growth", *Technological Forecasting and Social Change*, Vol. 130, May, pp. 28-38.

Ransbotham, S., Fichman, R.G., Gopal, R. and Gupta, A. (2016), "Ubiquitous IT and digital vulnerabilities", *Information Systems Research*, Vol. 27 No. 4, pp. 834-847.

Westerman, G., Bonnet, D. and McAfee, A. (2014), *Leading Digital: Turning Technology into Business Transformation*, Harvard Business Review Press.

Whyte, J., Stasis, A. and Lindkvist, C. (2016), "Managing change in the delivery of complex projects: configuration management, asset information and 'Big Data' ", *International Journal of Project Management*, Vol. 34 No. 2, pp. 339-351.

## Further reading

Gandomi, A. and Haider, M. (2015), "Beyond the hype: big data concepts, methods, and analytics", *International Journal of Information Management*, Vol. 35 No. 2, pp. 137-144.

**Corresponding author**
Amit Mitra can be contacted at: amit.mitra@uwe.ac.uk

# Driving forces for the US residential housing price: a predictive analysis

Amirhosein Jafari
*Department of Construction Management, College of Engineering, Louisiana State University, Baton Rouge, Louisiana, USA, and*
Reza Akhavian
*School of Engineering, California State University East Bay, Hayward, California, USA*

Received 20 July 2018
Revised 22 December 2018
6 May 2019
Accepted 16 May 2019

## Abstract

**Purpose** – The purpose of this paper is to determine the key characteristics that determine housing prices in the USA. Data analytical models capable of predicting the driving forces of housing prices can be extremely useful in the built environment and real estate decision-making processes.
**Design/methodology/approach** – A data set of 13,771 houses is extracted from the 2013 American Housing Survey (AHS) data and used to develop a Hedonic Pricing Method (HPM). Besides, a data set of 22 houses in the city of San Francisco, CA is extracted from Redfin real estate brokerage database and used to test and validate the model. A correlation analysis is performed and a stepwise regression model is developed. Also, the best subsets regression model is selected to be used in HPM and a semi-log HPM is proposed to reduce the problem of heteroscedasticity.
**Findings** – Results show that the main driving force for housing transaction price in the USA is the square footage of the unit, followed by its location, and its number of bathrooms and bedrooms. The results also show that the impact of neighborhood characteristics (such as distance to open spaces and business centers) on the housing prices is not as strong as the impact of housing unit characteristics and location characteristics.
**Research limitations/implications** – An important limitation of this study is the lack of detailed housing attribute variables in the AHS data set. The accuracy of the prediction model could be increased by having a greater number of information regarding neighborhood and regional characteristics. Also, considering the macro business environment such as the inflation rate, the interest rates, the supply and demand for housing, and the unemployment rates, among others could increase the accuracy of the model. The authors hope that the presented study spurs additional research into this topic for further investigation.
**Practical implications** – The developed framework which is capable of predicting the driving forces of housing prices and predict the market values based on those factors could be useful in the built environment and real estate decision-making processes. Researchers can also build upon the developed framework to develop more sophisticated predictive models that benefit from a more diverse set of factors.
**Social implications** – Finally, predictive models of housing price can help develop user-friendly interfaces and mobile applications for home buyers to better evaluate their purchase choices.
**Originality/value** – Identification of the key driving forces that determine housing prices on real-world data from the 2013 AHS, and development of a prediction model for housing prices based on the studied data have made the presented research original and unique.

**Keywords** Stepwise regression, Data analytics, Residential property, Hedonic pricing method, Housing prices, Predictive model

**Paper type** Research paper

## Introduction

Among all the construction and built environment sectors, the residential sector has the most impact on people's lives. It deals with homes as a property that almost everyone in modern society pays for its ownership and maintenance expenses one way or the other. As such, housing price fluctuation has always been a major topic for construction, real estate and built environment research studies. The potential increase in house value affects the quality of life of residents as well as the national economy. Therefore, it is important to study the composition of housing prices from a micro-level perspective and determine the root causes of the price increase.

Built Environment Project and Asset Management
Vol. 9 No. 4, 2019
pp. 515-529
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-07-2018-0100

This paper uses the Hedonic Pricing Model (HPM) to determine the key factors that affect the residential housing prices across the USA. Toward this goal, a data set of 13,771 houses is extracted from the 2013 American Housing Survey (AHS) and used to determine which factors have a significant impact on the overall price of a housing unit in the USA. The developed HPM is also validated through a sample of 22 houses located in representative San Francisco area as a sample, including both new homes and resales (extracted from Redfin real estate brokerage database). What is unique about San Francisco area is that the housing price has increased significantly recently and that is primarily due to the high-tech boom. From 2016 3rd quarter to 2017 3rd quarter, the equity gain in the San Francisco metropolitan area was $73,217. This number is $39,096, $39,887 and $14,888 for San Diego, Los Angeles, and national average, respectively (CoreLogic, 2018).

For the purpose of this study, different types of data are collected including the market value of housing units, the floor area of the unit, service year of the unit, the number of bedrooms and bathrooms, the location of the unit, among others. The collected variables are categorized into three main categories, include housing unit characteristics, location characteristics and neighborhood characteristics. Data pre-processing methods including data transformation and outlier detection have also been employed. Finally, a framework is proposed that uses: a correlation analysis; a stepwise regression model development; and the best subsets regression model selection, to first, identify the key driving forces that determine housing prices; and second, develop a hedonic pricing model based on identified factors to be used in housing price prediction. The presented study contributes to the body of knowledge by: identifying of the key driving forces that determine housing prices on real-world data; and developing a prediction model for housing prices based on identified variables, using an HPM.

## Literature review

### *Key factors in forming the housing price*

The topic of house pricing method is very well researched. Since housing is a major investment in the world, there is a continuous trend of research studies that focus on understanding the factors that may affect the price of houses. Many research groups including Candas *et al.* (2015), Chen and Jim (2010) and He *et al.* (2010), among others, published the results of their studies on factors affecting housing price changes by utilizing the practical and recreational values related to the geographical aspect of houses. As an example, according to Chen and Jim (2010), buyers willing to pay 17 percent more in value for an attractive landscape.

### *Hedonic regression analysis and pricing models*

Hedonic Price Analysis was invented by Andrew Court in the 1930s when he used the Hedonic Model to analyze automobile value in pricing and quality characteristics. However, it is argued that it was Griliches and Rosen who expanded the Hedonic Model into a wide range of applications in many systems (Goodman, 1998). The classic literature includes the "Hedonic-Price-Indexes for Automobiles: An Econometric Analysis of Quality Change" published by Zvi Grilichesval (Griliches, 1971) and Shervin Rosen's "Hedonic-Prices and Implicit Markets: Product Differential in Pure Competition" (Rosen, 1974). Rosen (1974) defined Hedonic-Prices as "the implicit prices of attributes and are revealed to economic agents from observed prices of differentiated products and the specific amounts of characteristics associated with them." These implicit prices are estimated by the first step regression analysis (Rosen, 1974). Since Rosen introduced the Hedonic Model to the residential housing price analysis process after 40 years of practice and continuous improvement, the HPM has developed into one of the widely used models in the real estate market analysis.

In Rosen's (1974) Hedonic theory the present values of rents per unit of each hedonic characteristic can be seen as implicit market prices. The embedded option to redevelop to higher intensity per unit land value, however, is not included in this model. A study by Clapp *et al.* (2012) showed the change in the Hedonic equilibrium where this option is incorporated into the model and the resulting value driven by a Wiener process (Clapp *et al.*, 2012). There are a number of simplifying and limiting assumptions in this study including the presence of a single call option. A study by Wallace and Meese (1997) looks at some hypotheses that explore the repeat-sales and hedonic price indices when they are constructed on municipality-level data sets. Their data included individual home sales in Oakland and Fremont California over 18 years. The results of their study indicated that Hedonic techniques are better suited to problems with index numbers and they recommend researchers relying on the hedonic indices at the municipality level (Wallace and Meese, 1997). Another study evaluates residential real estate market at the county level using a HPM. They have criticized this method which can be affected by missing variable bias and suggested environmental characteristics such as air and water quality be considered in residential price models (Hanink *et al.*, 2012). In the presented study of US housing market, the environmental characteristics have been considered. Another study compared the performances of three most frequently used house price measurement methods. The three methods included (1) the simple average method with no quality adjustment, (2) the matching approach with the repeat sales modeling framework and (3) the HPM. Method (1) was identified as biased as a result of the recent trend of housing suburbanization which is common in some U.S. states; method (2) was also proved biased when adopted in the newly built house price measurement due to the unique pricing behavior of newly built housing units which again can be seen in US markets. Therefore, the HPM model showed the best performance among the three most frequently used house price measurement methods.

### *Major issue, debates and the limitations in housing price determinants*

A great deal of research has been focused on an in-depth study of certain variables, but a higher number of studies have been aiming at comparison-based approaches. Most of the researchers used regression analysis and drew data from open geospatial and economic data sources. Despite wide coverage, the current stages of research generally stay fragmented, more specifically, they are often conducted from a distinct perspective within the housing market. Each research serves a different purpose for a different population in a different environment of housing transaction. Since the housing market is so massive and dynamic, the existing empirical approach can only render as referential assistance rather than an absolute truth. The HPM has been widely applied in previous studies to estimate the value of housing. Sirmans *et al.* (2005) examined and analyzed 125 different Hedonic Pricing Models and listed top 20 characteristics appeared most often in Hedonic Pricing Model Studies, including: lot size, ln lot size, square feet, ln square feet, brick, age, number of stories, number of bathrooms, number of rooms, bedrooms, full baths, fireplace, air-conditioning, basement, garage spaces, deck, pool, distance, time on market and time trend (Sirmans *et al.*, 2005). Based on previous studies, we have compiled a list of characteristics data sets for analyzing which is explained in the remainder of this paper.

## Data

### *Data collection*

Data for this study were obtained from the 2013 AHS Metropolitan public use file. This microdata set contains individual responses to survey questions for which the basic unit is an individual housing unit (US Census Bureau (USCB), 2016). Such information richness at the micro-level is appropriate for the development of a prediction model of housing unit price based on their attributes.

In addition, to validate the prediction model, a sample of 22 houses located in the San Francisco city were analyzed. Redfin real estate brokerage database (www.redfin.com) is used to collect this data and Google map (www.maps.google.com) is employed to calculate the required distances. Most of Redfin information was obtained using the first multiple listing services.

### *Variable selection*

In addition to the market value of housing units in US Dollar (VALUE), three categories of features are considered including housing unit characteristics, location characteristics and neighborhood characteristics. For each category, the most important variables are selected based on the literature review and available data.

Housing unit characteristics considered in this analysis include the floor area of the unit in ft$^2$ (UNITSF), service year of the unit (SERVICE) which can be calculated by the subtraction of age of building (BUILT) from the study year, which considers the number of years in which the unit was built and used, the total number of bedrooms (BEDRMS) and the number of bathrooms (BATHRMS). Location characteristics considered include regional census division in which the unit is located (DIVISION) and central city or suburban status of the unit's location (METRO). Neighborhood characteristics include access of the unit to open spaces such as parks or ranches within half a block (EGREEN), access of the unit to businesses or institutions, such as stores, restaurants, schools or hospitals within half block (ECOM), access of the unit to convenience grocery store within 15 min (GROCERY), access of the unit to convenience drug store within 15 min (DRUGSTORE), and access of the unit to any railroads, airports or highways with at least four lanes within half block (ETRANS).

### *Data preparation*

The purpose of data preparation is to: remove the data points having missing information; and remove the outliers from the data set. An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Outliers detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data. A 5–95 percent method is used to detect and remove outliers, in which 2.5 percent of data from the minimum side and 2.5 percent data from the maximum side are detected and removed from the data set.

The published AHS microdata for household-level information contains 84,355 responses or data points, of which 13,771 data points were selected based on: data points having no missing information about the selected variables; and data points with no outliners.

### *Data summary*

Variable definitions and descriptive statistics are given in Tables I and II.

## Methodology

### *Hedonic pricing model*

As stated earlier, the goal of the study is to determine which characteristics are meaningful for house buyers in the USA and how we can predict the market value of a house using those factors. An HPM is used to determine which characteristics have a significant impact on the overall price of a house. The house transaction price changes according to a certain ratio when some of the characteristics change. The fundamental premise of the Hedonic Pricing Model is shown in the following equation:

$$P = f(X_1, X_2, X_3, \ldots, X_n), \tag{1}$$

where $P$ is the dependent variable reflecting the listing housing price, and $X_1, X_2, X_3, \ldots, X_n$

**Table I.** Descriptive statistics for continuous variables

| Variable | Description | AHS data |  |  |  | San Francisco area data |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  | Mean | SD | Minimum | Maximum | Mean | SD | Minimum | Maximum |
| VALUE | Market value of housing unit ($) | 229,038 | 172,342 | 10,000 | 1,000,000 | 2,502,681 | 1,955,143 | $698,000 | 9,995,000 |
| UNITSF | Square footage of unit (SF) | 1,999 | 878 | 400 | 6,000 | 2,466 | 1,154 | 1,077 | 6,505 |
| SERVICE | Number of years the unit has been in service | 44.7 | 26.2 | 0 | 94 | 86.0 | 37.0 | 11 | 123 |
| BEDRMS | Number of bedrooms in unit | 3.2 | 0.81 | 1 | 6 | 3.5 | 0.96 | 2 | 6 |
| BATHRMs | Number of bathrooms in unit | 1.6 | 0.76 | 0 | 4 | 2.9 | 1.16 | 1 | 6.5 |

| | | | Frequency[a] (%) | |
|---|---|---|---|---|
| | | | AHS | San Francisco |
| Variable | Description | Category | data | Area data |
| DIVISION | Census division | New England | 3.6 | 0 |
| | | Middle Atlantic | 16.7 | 0 |
| | | East North Central | 22.8 | 0 |
| | | West North Central | 6.5 | 0 |
| | | South Atlantic and East South Central | 23.9 | 0 |
| | | West South Central | 9.6 | 0 |
| | | Mountain and Pacific | 16.7 | 100 |
| METRO | Central city/suburban status | Central city of MSA | 19.3 | 100 |
| | | Inside MSA, but not in central city – urban | 40.7 | 0 |
| | | Inside MSA, but not in central city – rural | 16.2 | 0 |
| | | Outside MSA, urban | 7.1 | 0 |
| | | Outside MSA, rural | 16.6 | 0 |
| EGREEN | Open spaces within 1/2 block of unit | Yes | 42.9 | 45.4 |
| | | No | 57.1 | 54.6 |
| ECOM | Business/institutions within 1/2 block | Yes | 17.0 | 36.4 |
| | | No | 83.0 | 63.6 |
| GROCERY | Grocery store nearby (15 min) | Yes | 94.9 | 95.4 |
| | | No | 5.1 | 4.6 |
| DRUGSTORE | Drugstore nearby (15 min) | Yes | 84.8 | 100 |
| | | No | 15.2 | 0 |
| ETRANS | Railroad/airport/4-lane highway within 1/2 block | Yes | 8.9 | 50 |
| | | No | 91.1 | 50 |

**Notes:** MSA, Metropolitan statistical area. [a]Frequency: represents the ratio of number of that variables to the total number of data points

**Table II.**
Descriptive statistics for categorical variables

are independent variables reflecting the housing unit, location, and neighborhood characteristics that described in the "Data" Section.

The Hedonic Pricing Model can be classified into four simple parametric functional forms: linear specification, semi-log specification, log-log specification, box-cox transform (Xiao *et al.*, 2017). In the linear specification, both the dependent and explanatory variables enter the regression with linear form. The formula is shown in the following equation:

$$y = \alpha + \beta X + e \; P = \beta_0 + \sum_{k=1}^{k} \beta_{kx_k} + \sum, \tag{2}$$

where $p$ is the listing housing price; $\sum$ the vector of the random error term; $\beta_k$ the marginal change of the unit price of the $K$th characteristic $x_k$ of the good (Xiao *et al.*, 2017).

In the semi-log specification, the dependent variable is log form, and explanatory variables are linear. The equation is shown in the following equation:

$$\ln P = L_n \beta_0 + \sum_{k=1}^{k} \beta_k x_k + \Sigma, \tag{3}$$

where $p$ is the listing housing price; $\sum$ the vector of the random error term; $\beta_k$ the rate at which the price increases at a certain level given the characteristics $x_k$ (Xiao *et al.*, 2017).

Since the house prices are skewed, a semi-log model is the most commonly used in the specification of Hedonic Housing Price because the log transformation reduces the problem of heteroscedasticity. Therefore, in the developed hedonic pricing model, the $P$ (dependent variable in Equation (1)) will be the natural log of house prices.

### *Developed framework*

For the purpose of this study, a four-step framework is developed to analyze the collected data, as it is presented in Figure 1.

In the first step, the correlation analysis is performed to find correlated factors with the housing unit price. The correlation analysis results in an understanding of how closely the independent variables are correlated to the dependent variable. Coefficients approaching 1 or −1 indicate highly correlated variables and a perfect linear relationship. The $p$-value smaller than 0.05 means that there is convincing evidence of the correlation in 95 percent confidence level. In this analysis, a correlation coefficient is used to identify which variables have a high correlation to dependent variables; those variables were typically identified as significant in the following regression analysis.

After identifying the strong correlation factors, then stepwise regression model provides a regression model that contains the most significant variables for the AHS data set. Stepwise regression model determines the best combination of predictor variables which have the smallest $t$-test $p$-value of smaller than 0.05. The stepwise regression model enters or removes a predictor based on the results of $t$-tests. Each of the predictors is a candidate to be entered the stepwise regression model; at each step, the predictor with smallest $t$-test $p$-value is added to the stepwise model. The predictor with the largest $p$-value is removed from the stepwise model because the variable is the least significant variable. The stepwise regression automatically stops running when all the variables in the model have a $p$-value less than the $\alpha$ value and all the variables not in the model have a $p$-value above the $\alpha$ value. In this analysis, the $\alpha$ value is set up to be 0.05 which is a common setting in this type of analysis. The $p$-values fall

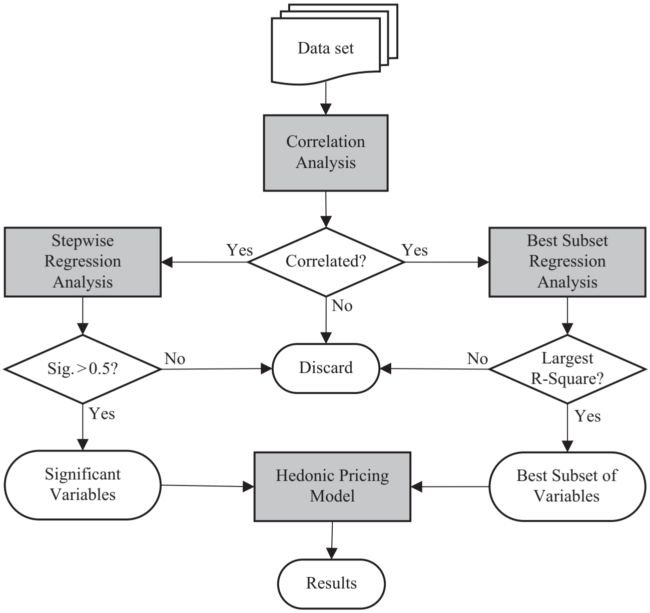


**Figure 1.** Methodology framework

below 0.05 considers to be statistically significant in this analysis. The final model of stepwise regression analysis is not guaranteed to be the optimal model; it is a model contains the most significant variables for each data set.

The best subsets regression model is used to perform further analysis. The subsets regression analysis is a screening tool to reduce the number of possible regression models since it identifies all the potential regression models derived from all the possible combinations of the predictors. The higher the $R^2$, the better the model fits the data. In addition, the Mallow Cp estimates the size of the bias into the predicted responses. When the Cp value is near the p (the number of the parameters in this model), that means the bias is small. When Cp is greater than $p$, the bias is substantial, and when Cp is below $p$, there is a sampling error.

The last step for this analysis is to use variables identifies in correlation analysis, stepwise regression analysis and best subsets regression analysis to perform the Hedonic Pricing Model with Semi-Log specification.

## Results and discussions

### *Correlation analysis*

After running the correlation analysis, the correlation relationships between the listing housing price and other variables are tested. The correlated variables list in Table III. As it is shown, almost all of the selected variables are correlated to housing prices in AHS data set. For the AHS data set, the housing unit characteristics including "UNITSF," "BEDRMS" and "BATHRMS" variables are positively related to the housing price, while "SERVICE" variable is negatively related to the housing price, at 0.01 significant levels since their $p$-value are smaller than 0.01. Regarding location characteristics, "DIVISION" variable is positively related to the housing price at 0.01 significant level, while the division of Mountain and Pacific (where San Francisco area located) is most positively related to housing price and East North Central is most negatively related to housing price. In addition, regarding the "METRO" variables, inside metropolitan statistical area (but not in central city) is most positively related

| Variables | Correlation | $p$-value | Strength of the relationship |
|---|---|---|---|
| UNITSF | 0.449 | < 0.0001 | Positive strong |
| SERVICE | −0.173 | < 0.0001 | Negative strong |
| BEDRMS | 0.336 | < 0.0001 | Positive strong |
| BATHRMS | 0.323 | < 0.0001 | Positive strong |
| DIVISION_New England | 0.079 | < 0.0001 | Positive strong |
| DIVISION_Middle Atlantic | 0.173 | < 0.0001 | Positive strong |
| DIVISION_East North Central | −0.165 | < 0.0001 | Negative strong |
| DIVISION_West North Central | −0.098 | < 0.0001 | Negative strong |
| DIVISION_South Atlantic and East South Central | −0.090 | < 0.0001 | Negative strong |
| DIVISION_West South Central | −0.114 | < 0.0001 | Negative strong |
| DIVISION_Mountain and Pacific | 0.232 | < 0.0001 | Positive strong |
| METRO_Central city of MSA | −0.041 | < 0.0001 | Negative strong |
| METRO_ Inside MSA, but not in central city – urban | 0.170 | < 0.0001 | Positive strong |
| METRO_ Inside MSA, but not in central city – rural | 0.010 | 0.234 | None |
| METRO_ Outside MSA, urban | −0.117 | < 0.0001 | Negative strong |
| METRO_ Outside MSA, rural | −0.131 | < 0.0001 | Negative strong |
| EGREEN | 0.018 | 0.034 | Positive weak |
| ECOM | −0.071 | < 0.0001 | Negative strong |
| GROCERY | 0.047 | < 0.0001 | Positive strong |
| DRUGSTORE | 0.068 | < 0.0001 | Positive strong |
| ETRANS | −0.064 | < 0.0001 | Negative strong |

**Table III.** Analysis results of variable

to housing price and outside metropolitan statistical area (rural) is most negatively related to housing price. Finally, regarding the neighborhood characteristics, "EGREEN" is positively related to housing price at 0.05 significant level, "GROCERY" and "DRUGSTORE" are positively related to housing price at 0.01 significant levels, and "ECOM" and "ETRANS" are negatively related to housing price at 0.01 significant levels.

Results of correlation analysis showed the main factors that correlate with housing transaction prices. Based on the results of this correlation comparison, we would expect to see those strongly correlated factors in the proposed regression models.

### *Stepwise regression analysis*

The next step for this analysis is to perform a stepwise regression analysis. The stepwise regression model should identify the most significant variables in the model. The natural log of "VALUE" is the dependent variable, and the other listed characteristics are considered the independent variables.

For the first step, the predictor "UNITSF" is entered into the stepwise model. The estimated constant intercept is 11.3, and the *p*-value for testing "UNITSF" is 0.000. The $R^2$ is the percentage of variation in the response that is explained by the model. The higher the $R^2$ value, the better the model fits the data. However, the $R^2$ value increases when additional predictors added to the model; the three-predictor model will always have higher $R^2$ value than the two- predictor model. So, $R^2$ is more useful when comparing the same size models. The adjusted $R^2$ incorporates the number of predictors in the model, using adjusted $R^2$ value when compared models that have different numbers of predictors. The adjusted $R^2$ value is 20.02 percent. Mallows' Cp-statistic is 4,626.3, and it measures the bias and the variation in the predicted response. When the Cp value is near or smaller to the number of parameters, the model is unbiased.

In the second step, the predictor "Division_Mountain & Pacific" is entered the Stepwise Model. The estimated constant intercept is 11.22, and the *p*-value for testing "Division_Mountain & Pacific" is 0.000. The adjusted $R^2$ value is 25.78 percent. Mallows' Cp-statistic is 3,452 which is not small enough yet.

After 16 steps, the estimated constant intercept is 10.87, and the estimated *S* is 4,578. The predicted $R^2$ value is 40.72 percent, and the adjusted $R^2$ value is 40.65 percent. Mallows' Cp-statistic is 16 which is smaller to the number of parameters, and consequently, the model is unbiased. The five variables that are not entered the model include "DIVISION_East North Central," "DIVISION_West North Central," "METRO_Central City of MSA," "METRO_Inside MSA but not in central city – rural" and "GROCERY." It can be implied that the distance to grocery stores is not an important factor when people want to purchase a house. Table IV shows the stepwise regression analysis results for ASH data set.

### *Best subsets regression analysis*

The best subsets regression analysis identifies all the potential regression models derived from the possible combinations of the predictors. The higher Adjusted $R^2$ value, the better the model fits the data. For different numbers of variables in the prediction model (range from 1 to 19), the best subset of variables is selected. Figure 2 illustrates the relation of the number of variables in the model and the highest adjusted $R^2$ that can be resulted in the regression. As it is shown, the highest $R^2$ is 40.73 percent from a subset of 19 (out of 22) variables. Although the adjusted $R^2$ value of 40.73 percent might not be very high, it is a reasonable value for the high number of data points while knowing that the signs and significance of the estimated coefficients support overall model validity (Jafari *et al.*, 2017). Also as it is shown in Figure 2, increasing the number of variables in the prediction model to more than 15 has no significant impact on the $R^2$ of the regression model.

**Table IV.**
Stepwise regression analysis

| Step | Parameter | "Sig Prob" | Seq SS | $R^2$ | Cp |
|---|---|---|---|---|---|
| 1 | UNITSF | 0.0000 | 1,600.446 | 0.2072 | 4,626.3 |
| 2 | DIVISION_Mountain and Pacific | 0.0000 | 391.4284 | 0.2579 | 3,452.4 |
| 3 | DIVISION_Middle Atlantic | 0.0000 | 369.364 | 0.3057 | 2,344.8 |
| 4 | BATHRMS | 0.0000 | 220.483 | 0.3343 | 1,684.4 |
| 5 | DIVISION_New England | 0.0000 | 186.1789 | 0.3584 | 1,127.1 |
| 6 | METRO_ Inside MSA, but not in central city – urban | 0.0000 | 116.4435 | 0.3735 | 779.31 |
| 7 | SERVICE | 0.0000 | 78.81215 | 0.3837 | 544.55 |
| 8 | BEDRMS | 0.0000 | 68.60272 | 0.3926 | 340.46 |
| 9 | METRO_ Outside MSA, urban | 0.0000 | 25.60256 | 0.3959 | 265.54 |
| 10 | METRO_ Outside MSA, rural | 0.0000 | 21.04553 | 0.3986 | 204.32 |
| 11 | DIVISION_South Atlantic and East South Central | 0.0000 | 21.90335 | 0.4014 | 140.52 |
| 12 | DRUGSTORE | 0.0000 | 12.78911 | 0.4031 | 104.1 |
| 13 | ECOM | 0.0000 | 13.71161 | 0.4049 | 64.904 |
| 14 | DIVISION_West South Central | 0.0000 | 6.498887 | 0.4057 | 47.381 |
| 15 | ETRANS | 0.0000 | 5.529242 | 0.4064 | 32.77 |
| 16 | EGREEN | 0.0000 | 6.220614 | 0.4072 | 16.082 |

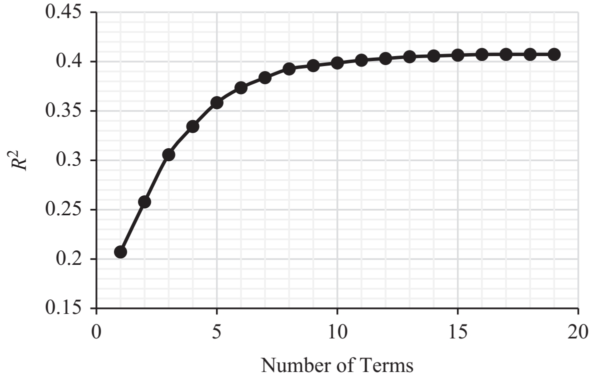


**Figure 2.**
Best subsets regression analysis for different numbers of variables

The subsets of variables for each best regression models are presented in Table V. As it is shown, the most important variables are repeated in each model. The first seven factors impacting on the housing price are in order of "UNITSF," "DIVISION_Mountain & Pacific," "DIVISION_Middle Atlantic," "BATHRMS," "DIVISION_New England," "METRO_ Inside MSA, but not in central city – urban," "SERVICE," and "BEDRMS." In addition, the results show that the neighborhood characteristics ("EGREEN," "ECOM," "GROCERY," "DRUGSTORE" and "ETRANS") have the least significant impact on the housing price in comparison to housing characteristics and location characteristics.

Since the model with 15 variables has an $R^2$ of near to maximum, this model is selected to perform the Hedonic Pricing Model with Semi-Log specification (highlighted in Table V).

*Hedonic pricing model with semi-log specification*

The results of HPM with Semi-Log specification for AHS data is shown in Table VI. "LnVALUE" is the natural log of "VALUE," and it is the dependent variable in the Hedonic Pricing Model. In total, 15 independent variables are selected from the previous step, as they are highlighted in Table V. As it is shown, all the estimated coefficients of the independent variables are significant at the 1 percent level.

**Table V.** Variables in best subsets regression analysis

| No. of var. | $R^2$ | UNITSF | SERVICE | BEDRMS | BATHRMS | DIVISION_New England | DIVISION_Middle Atlantic | DIVISION_East North Central | DIVISION_West North Central | DIVISION_South Atlantic and East South Central | DIVISION_West South Central |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.2072 | X | | | | | | | | | |
| 2 | 0.2579 | X | | | | | | | | | |
| 3 | 0.3057 | X | | | | | X | | | | |
| 4 | 0.3343 | X | | | X | | X | | | | |
| 5 | 0.3584 | X | | | X | X | X | | | | |
| 6 | 0.3735 | X | | | X | X | X | | | | |
| 7 | 0.3837 | X | X | | X | X | X | | | | |
| 8 | 0.3926 | X | X | X | X | X | X | | | | |
| 9 | 0.3959 | X | X | X | X | X | X | | | | |
| 10 | 0.3986 | X | X | X | X | X | X | | | | |
| 11 | 0.4014 | X | X | X | X | X | X | | | X | |
| 12 | 0.4031 | X | X | X | X | X | X | | | X | |
| 13 | 0.4049 | X | X | X | X | X | X | | | X | |
| 14 | 0.4057 | X | X | X | X | X | X | | | X | X |
| 15 | 0.4065 | X | X | X | X | X | X | | | X | |
| 16 | 0.4072 | X | X | X | X | X | X | | | X | X |
| 17 | 0.4073 | X | X | X | X | X | X | | | X | X |
| 18 | 0.4073 | X | X | X | X | | X | X | X | X | X |
| 19 | 0.4073 | X | X | X | X | | X | X | X | X | X |

| No. of var. | DIVISION_Mountain and Pacific | METRO_Central city of MSA | METRO_ Inside MSA, but not in central city – urban | METRO_ Inside MSA, but not in central city – rural | METRO_ Outside MSA, urban | METRO_ Outside MSA, rural | EGREEN | ECOM | GROCERY | DRUGSTORE | ETRANS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | |
| 2 | X | | | | | | | | | | |
| 3 | X | | | | | | | | | | |
| 4 | X | | | | | | | | | | |

(*continued*)

Table V.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | X | | | | | | | | | |
| 6 | X | | X | | | | | | | |
| 7 | X | | X | | | | | | | |
| 8 | X | | X | | | | | | | |
| 9 | X | | X | X | | | | | | |
| 10 | X | | X | X | X | | | | | |
| 11 | X | | X | X | X | | | | | |
| 12 | X | | X | X | X | | | | X | |
| 13 | X | | X | X | X | | X | | X | |
| 14 | X | | X | X | X | | X | | X | |
| 15 | X | | X | X | X | X | X | | X | X |
| 16 | X | | X | X | X | X | X | | X | X |
| 17 | X | | X | X | X | X | X | X | X | X |
| 18 | X | | X | X | X | X | X | X | X | X |
| 19 | X | X | X | X | X | X | X | X | X | X |

| Term | Estimate | SE | $p$-value |
|---|---|---|---|
| Intercept | 10.842262 | 0.029559 | < 0.0001 |
| UNITSF | 0.0002565 | 6.766e-6 | < 0.0001 |
| SERVICE | −0.002584 | 0.000211 | < 0.0001 |
| BEDRMS | 0.0994013 | 0.007199 | < 0.0001 |
| BATHRMS | 0.1283736 | 0.007558 | < 0.0001 |
| DIVISION_New England | 0.6766997 | 0.027084 | < 0.0001 |
| DIVISION_Middle Atlantic | 0.503868 | 0.014722 | < 0.0001 |
| DIVISION_South Atlantic and East South Central | 0.1071566 | 0.013027 | < 0.0001 |
| DIVISION_Mountain and Pacific | 0.5558795 | 0.014579 | < 0.0001 |
| METRO_Inside MSA, but not in central city – urban | 0.1191956 | 0.011535 | < 0.0001 |
| METRO_Outside MSA, urban | −0.213675 | 0.02033 | < 0.0001 |
| METRO_Outside MSA, rural | −0.127884 | 0.015094 | < 0.0001 |
| EGREEN | 0.0466613 | 0.010274 | < 0.0001 |
| ECOM | −0.077771 | 0.013667 | < 0.0001 |
| DRUGSTORE | 0.0926327 | 0.014084 | < 0.0001 |
| ETRANS | −0.078908 | 0.017718 | < 0.0001 |

**Table VI.** Hedonic pricing model with semi-log specification

Regarding the housing characteristics, the results show that if the floor area of the house is increased by 100 square feet, its value increase by 2.6 percent. In addition, when the service year of the building is one year less, its value increase by 0.3 percent. Furthermore, having one more bedroom or one more bathroom would increase the house price by 10.4 or 13.7 percent, respectively. Regarding the neighborhood characteristics, neighborhoods within half a block to open spaces (e.g. parks) and 15 min access to drug stores have higher housing values by 4.8 and 9.7 percent, respectively. Also, the neighborhoods within half a block to business areas and half a block to highways or railroads have lower housing values by 7.5 and 7.6 percent, respectively.

*Prediction model implementation*

The developed hedonic pricing model from AHS 2013 data is used to predict the housing price of 22 selected houses in the San Francisco area. It should be mentioned that the market values in AHS data are based on 2013 US Dollar, while the San Francisco Area data are based on 2018 US Dollar. In addition, the average price per square foot for the houses in AHS 2013 data is $119.2 per SF, while for San Francisco Area data it is $951.00. San Francisco is one of the most expensive cities in the entire USA and has experienced a booming increase in the housing market value since 2013. Therefore, in order to predict the recent housing market values in the San Francisco Area using the developed HPM based on the 2013 AHS data, a quantitative adjustments approach is required. The paired data analysis (PDA) is one of the most common methods for obtaining market-derived grid adjustments used in the sales comparison approach (Lipscomb and Gray, 1995). PDA involves comparing two set of properties to one another to examine one specific difference. In this paper, the average value of the variable that represents the house value (LnValue) is used to calculate the adjustment rate between the AHS and San Francisco Area data.

The average of LnValue for the actual San Francisco data is 14.56, while the average of the predicted LnValue is 12.57. The difference between these two values (1.99) is the adjustment rate for increasing the price of the houses in the last five years. The adjusted predicted value can be calculated by adding this adjustment rate to the predicted LnValue. Figure 3 shows a comparison between actual housing prices in the San Francisco Area and the predicted and adjusted predicted values from the 2013 AHS data.

The mean squared deviation between the adjusted predicted values and the actual values of LnVALUE is 0.092, illustrating that the prediction model could work accurately to

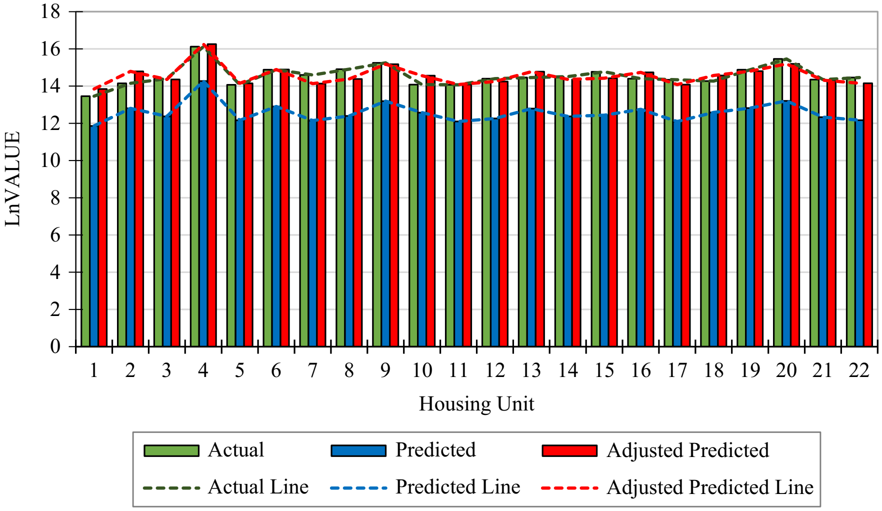


**Figure 3.**
The result of HPC on San Francisco data

estimate the housing prices, using a benchmark adjustment. Therefore the developed prediction model resulted from 2013 AHS data could be used to predict the market value of the housing prices based on available benchmarks.

## Conclusion

The developed framework which is capable of predicting the driving forces of housing prices and predict the market values based on those factors could be useful in the built environment and real estate decision-making processes. Researchers can also build upon the developed framework to develop more sophisticated predictive models that benefit from a more diverse set of factors. Finally, predictive models of housing price can help develop user-friendly interfaces and mobile applications for home buyers to better evaluate their purchase choices.

Evaluating the AHS data for housing price modeling is a novel method that is investigated in this research. However, an important limitation of this study is the lack of detailed housing attribute variables in the AHS data set. The accuracy of the prediction model could be increased by having a greater number of information regarding neighborhood and regional characteristics. Also, considering the macro business environment such as the inflation rate, the interest rates, the supply and demand for housing, and the unemployment rates, among others could increase the accuracy of the model. The authors hope that the presented study spurs additional research into this topic for further investigation. Another limitation of the presented work is that the analysis have been performed in a regional market. Housing markets in the USA are often localized and there could also be sub-markets under a larger one. Future studies can use the results of this research and concentrate more on localized markets for more focused analysis.

## References

Candas, E., Kalkan, S.B. and Yomralioglu, T. (2015), "Determining the factors affecting housing prices", FIG Working Week 2015, Sofia, May 17-21.

Chen, W.Y. and Jim, C.Y. (2010), "Amenities and disamenities: a hedonic analysis of the heterogeneous urban landscape in Shenzhen (China)", *Geographical Journal*, Vol. 176 No. 3, pp. 227-240.

Clapp, J.M., Jou, J.B. and Lee, T. (2012), "Hedonic models with redevelopment options under uncertainty", *Real Estate Economics*, Vol. 40 No. 2, pp. 197-216.

CoreLogic (2018), "Special report: evaluating the housing market since the great recession", available at: www.corelogic.com/downloadable-docs/corelogic-peak-totrough-final-030118.pdf (accessed December 2018).

Goodman, A.C. (1998), "Andrew court and the invention of hedonic price analysis", *Journal of Urban Economics*, Vol. 44 No. 2, pp. 291-298.

Griliches, Z. (1971), "3. Hedonic price indexes for automobiles: an econometric analysis of quality change", in Griliches, Z. (Ed.), *Price Indexes and Quality Change*, Harvard University Press, Cambridge, MA and London, pp. 173-196.

Hanink, D.M., Cromley, R.G. and Ebenstein, A.Y. (2012), "Spatial variation in the determinants of house prices and apartment rents in China", *The Journal of Real Estate Finance and Economics*, Vol. 45 No. 2, pp. 347-363.

He, C., Wang, Z., Guo, H., Sheng, H., Zhou, R. and Yang, Y. (2010), "Driving forces analysis for residential housing price in Beijing", *Procedia Environmental Sciences*.

Jafari, A., Valentin, V. and Berrens, R.P. (2017), "Estimating the economic value of energy improvements in US residential housing", *Journal of Construction Engineering and Management*, Vol. 143 No. 8, pp. 04017048-1-04017048-9.

Lipscomb, J. and Gray, B. (1995), "A connection between paired data analysis and regression analysis for estimating sales adjustments", *Journal of Real Estate Research*, Vol. 10 No. 2, pp. 175-183.

Rosen, S. (1974), "Hedonic prices and implicit markets: product differentiation in pure competition", *Journal of Political Economy*, Vol. 82 No. 1, pp. 34-55.

Sirmans, G.S., Macpherson, D.A. and Zietz, E.N. (2005), "The composition of hedonic pricing models", *Journal of Real Estate Literature*, Vol. 13 No. 1, pp. 3-43.

US Census Bureau (USCB) (2016), "American housing survey (AHS)", available at: www.census.gov/programs-surveys/ahs.html (accessed December 2018).

Wallace, N.E. and Meese, R.A. (1997), "The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches", *The Journal of Real Estate Finance and Economics*, Vol. 14 Nos 1-2, pp. 51-73.

Xiao, Y., Chen, X., Li, Q., Yu, X., Chen, J. and Guo, J. (2017), "Exploring determinants of housing prices in Beijing: an enhanced hedonic regression with open access POI data", *ISPRS International Journal of Geo-Information*, Vol. 6 No. 11, pp. 358-366.

## Further reading

Wu, J., Deng, Y. and Liu, H. (2014), "House price index construction in the nascent housing market: the case of China", *The Journal of Real Estate Finance and Economics*, Vol. 48 No. 3, pp. 522-545.

Zhang, Y. and Dong, R. (2018), "Impacts of street-visible greenery on housing prices: evidence from a hedonic price model and a massive street view image dataset in Beijing", *International Journal of Geo-Information*, Vol. 7 No. 3, pp. 104-116.

Zheng, S., Sun, W. and Wang, R. (2014), "Land supply and capitalization of public goods in housing prices: evidence from Beijing: land supply and capitalization of public goods", *Journal of Regional Science*, Vol. 54 No. 4, pp. 550-568.

## Corresponding author

Reza Akhavian can be contacted at: reza.akhavian@csueastbay.edu

Received 14 May 2018
Revised 17 November 2018
18 April 2019
Accepted 17 May 2019

# Critical analysis for big data studies in construction: significant gaps in knowledge

Upeksha Hansini Madanayake and Charles Egbu
*School of the Built Environment and Architecture,*
*London South Bank University, London, UK*

## Abstract

**Purpose** – The purpose of this paper is to identify the gaps and potential future research avenues in the big data research specifically in the construction industry.

**Design/methodology/approach** – The paper adopts systematic literature review (SLR) approach to observe and understand trends and extant patterns/themes in the big data analytics (BDA) research area particularly in construction-specific literature.

**Findings** – A significant rise in construction big data research is identified with an increasing trend in number of yearly articles. The main themes discussed were big data as a concept, big data analytical methods/techniques, big data opportunities – challenges and big data application. The paper emphasises "the implication of big data in to overall sustainability" as a gap that needs to be addressed. These implications are categorised as social, economic and environmental aspects.

**Research limitations/implications** – The SLR is carried out for construction technology and management research for the time period of 2007–2017 in Scopus and emerald databases only.

**Practical implications** – The paper enables practitioners to explore the key themes discussed around big data research as well as the practical applicability of big data techniques. The advances in existing big data research inform practitioners the current social, economic and environmental implications of big data which would ultimately help them to incorporate into their strategies to pursue competitive advantage. Identification of knowledge gaps helps keep the academic research move forward for a continuously evolving body of knowledge. The suggested new research avenues will inform future researchers for potential trending and untouched areas for research.

**Social implications** – Identification of knowledge gaps helps keep the academic research move forward for continuous improvement while learning. The continuously evolving body of knowledge is an asset to the society in terms of revealing the truth about emerging technologies.

**Originality/value** – There is currently no comprehensive review that addresses social, economic and environmental implications of big data in construction literature. Through this paper, these gaps are identified and filled in an understandable way. This paper establishes these gaps as key issues to consider for the continuous future improvement of big data research in the context of the construction industry.

**Keywords** Construction industry, Built environment, Systematic review, Data analysis, Big data analytics, Knowledge gaps

**Paper type** Literature review

## Introduction

In essence, "big data" is an artefact generated as a collective intelligence of human individuals shared mainly through the technological environment (Eadie *et al.*, 2013). This environment devices a common platform, where virtually anything and everything can be documented, measured and captured digitally where the digital capture is then transformed into data (Sivarajah *et al.*, 2017). Mayer-Schonberger and Cukier (2012) referred to this process as "datafication". As the world has now inundated with data, big data analytics (BDA) is increasingly becoming a trending practice and has received substantial attraction from both academics and practitioners regardless of the sector. This massive growth in data generation brings significant opportunities for data scientists to capture meaningful insights and knowledge. Arguably, the accessibility of data and then its management can improve the status quo of many sectors by strengthening existing statistical analytical techniques (Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016). Notwithstanding the

Built Environment Project and
Asset Management
Vol. 9 No. 4, 2019
pp. 530-547
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-04-2018-0074

current situation, it is apparent that this trend is going to be improving in the future. In compliance with the concept of "datafication" and the technological advancements, it is conspicuous that the future will majorly rely on the data which is being generated and shared through machines, as machines communicate with each other over data networks- doing so will result less human involvement (van Dijck, 2014).

Having said the state-of-art in big data generally, the construction industry is not an exception to the pervasive digital revolution. Undoubtedly, the industry is dealing with significant amount of data arising from diverse disciplines (i.e. Building Information Modelling (BIM) data) throughout the life cycle of a facility and if they were better harnessed (by discovering the latent patterns, trends and correlations buried inside) it could help deriving useful social, economic and environmental insights that would support data-driven decision-making for competitive advantage. These interests have made construction organisations for the adoption of BDA, with the intention of generating valuable strategic business insights for enhanced decision making and thereby achieve organisational competitive advantage as there is a massive value creation potential from analysis of big data in construction sector (Cook, 2015). While big data appear to be one of the innovative trends for construction industry, it is still developing at a slow-moving pace as it is just starting to see the transformative effects of big data (Cook, 2015). To that end, it is important that the academia informs some of the influential implications of big data in construction for the better understanding of its implementers and/or adopters. This paper therefore investigates the previous big data research in construction literature to find out the gaps that have not being addressed with regards to social, economic and environmental implications as the effort will then could help both industry practitioners as well as academics to carry forward for the continuous improvement of the industry. There is currently no comprehensive survey of the literature, targeting the application of big data in the context of the construction industry tied-up for social, economic and environmental implications. This paper fills the void and presents a wide-ranging interdisciplinary study of fields such as machine learning, data warehousing, data mining, etc., and their applications.

## Literature review – big data studies in construction

The built environment and project management research over the past 50 years has focused on emerging research and development topics. "Big Data" is one such concern where research interests include the limitations of utilising big data; challenges associated with its visualisation; the role of improved forecasting in BDA – such as machine learning; access and ownership of data; and the opportunities for experimenting big data techniques for smart city and infrastructure components.

The construction industry generates significant amount of data that can be quickly voluminous from diverse disciplines throughout the life cycle of a facility. BIM is a perfect example for such large data gathering. BIM captures multi-dimensional CAD information systematically to support multidisciplinary collaboration and integration among stakeholders (Eadie *et al.*, 2013). With the emergence of embedded devices and sensors, constructions have even started to produce massive lump of data during the operations and maintenance stage. These are also converted into "big; BIM data". This situation has lead construction industry to enter the big data era (Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016).

Optimisation and automation of work processes with more collaboration technologies (digitalisation) were the key concerns during mid-2007. During the last immediate few years, research has begun to discuss more of a lifecycle perspective on costs and benefits of big data based decisions made during design and construction (Levitt and M.ASCE, 2007). Nowadays, considerations have been more focused on many different perspectives of big data as a way of boosting productivity and gain competitive advantages over business rivals (Oyedele, 2016; McGuire *et al.*, 2012; Gandomi and Haider, 2015; Marr, 2017). However, regardless of the approach or the ontological perspective used, many of the articles claim that Big data is a powerful tool that can be considered as a source when “properly managed, processed and analysed, have a powerful potential to generate new knowledge thus proposing innovative and actionable insights for businesses” (Jukić *et al.*, 2015, p. 4). In addition to these extant insights on the current field, it is important to identify the significant gaps and potential future research directions – which are addressed in this study.

## Methodology

The methodology is two-fold. The first part is to identify the key themes discussed in the previous big data research, while the second part identifies the significant gaps in existing body of knowledge. First, a systematic literature review (SLR) is carried out to observe and understand past trends and extant patterns/themes in the BDA research area particularly in construction, that lead to identify potential gaps in research. It is important to mention that this review excludes social science research that investigates human behaviour. To achieve the said goal, a comprehensive “keywords” search is been conducted in specialised databases, journals and few conference proceedings. Prospective articles are also used to capture more relevant papers through bibliography. This significantly helped enhance rigour of database as well as to increase the number of papers investigated.

The identified studies are analysed according to their contribution, summarising the key themes discussed, extant knowledge dimensions, thereby identifying limitations, implications and potential further research avenues to fill the knowledge gaps and support the academic community in exploring research themes/patterns.

Thus, in the process of tracing big data studies, a profiling method is employed to analyse peer-reviewed articles published between 2007–2017, extracted from the Scopus and emerald databases. The analysis presented in this paper has identified relevant BD research studies that have contributed both conceptually and empirically to the expansion and accrual of intellectual wealth to the BDA in technology and organisational resource management discipline specifically in construction.

The reason behind the selection of Scopus lies at its comprehensiveness for academic journal articles consisting of nearly 22,000 titles from over 5,000 publishers, of which 20,000 are peer-reviewed journals in the scientific, technical, medical and social sciences (including arts and humanities) (Elsevier, 2017). Besides, it is frequently updating, unbiased and a reliable source that covers the research spectrum with more than 50 per cent of key research publishers (Cheng *et al.*, 2012). Emerald data base is selected for the consistency of data to be derived from AEC sectors.

### *Research scope*

Although there is a growing demand, big data research for construction is still in the nascent stage (Qadir *et al.*, 2016). Therefore, comparatively a paucity of research studies is expected. It is appearing that BD and BDA as a research discipline are still evolving and not yet established. Thus, a comprehensible understanding of the phenomenon, its definitions, classifications and applications are yet to be fully established.

Although there is a notable body of research that provides understanding on different aspects of BD and BDA area, there seems to be a lack of comprehensive and methodical approaches to understand the phenomenon of BD. Further, there is a lack of studies that empirically reports about the practical applications of big data techniques-more precisely the types of BDA methods used in particular construction case studies (projects or organisations). This study not only aims to grasp these gaps, but also to capture the conclusions offered by each review article to analyse, synthesise and present a state-of-the-art structured analysis of the normative literature on big data and BDA to support future research avenues.

*Data collection process*

The SLR presented in this study aims to evaluate the existing research published on BD and BDA in relation to the construction sector. An established profiling approach is used to investigate and analyse different themes discussed. Following search strings were used to identify the relevant articles in the sections of "title, abstract, keywords" of articles belonging to Scopus and Emerald databases. The search is further filtered to subject areas related to construction sector journals only. The main keywords used in the search were "big data" AND "construction". Relevant literature was majorly found in the journal for automation in construction and Journal of Information Technology in Construction. Since those journals are relevant data sources, even though the papers do not use the phrase "big data", it was able to find relevant articles through the "suggestions" facility. Searches were also made with "construction engineering projects" as search keywords, but with limited additional search results. The latter search resulted with more technical data. Few exclusion keywords used to avoid the search being biased to pure engineering/technical papers.

Using the aforementioned search strings, the search was limited to published "journal articles" filtered for the two subject areas in Scopus: engineering and energy, for the time period of 2007–2017. For emerald, the subject areas filtered to Information and communications technology, Information and knowledge management, Management science/operations research and Information systems. In the search results, there were several papers that are not available or do not specifically deal with construction industry. Such papers were excluded by the excluding keywords. Some papers were found within the areas of ICT and software development and data/information management, but are not directly related to the construction industry. Papers only dealt with construction industry/projects were selected among them. The search initially offered with 113 and 93 journal articles to review from Scopus and Emerald respectively. The selected material was subsequently examined in more detail to identify the relevance of articles. Only papers that give a clear contribution or are of clear relevance to big data in the construction industry were selected. However, after cleansing out of the total number of 221 articles, 115 articles were discarded and finally 106 articles were remained and taken forward for further interrogation. The selected 106 articles were then profiled according to following sub-categories:

(1) type of publication (e.g. research or technical paper, literature review, viewpoint);

(2) type of research method(s) employed (e.g. case study, mixed method, analytical);

(3) yearly publications from 2007 until 2017; and

(4) main theme of the article (definition/challenges-opportunities/techniques, etc.).

For the fourth sub-category above, this paper focuses on the themes discussed in selected papers generally and thereby extracts the gaps that have not being addressed. However, the authors suggest a further investigation incorporating few different dimensions such as time, cost quality. An example for time dimension can be; application of big data techniques in stages of a typical construction project. The themes were further sub-categorised as definition/concept, challenges-opportunities and application of big data techniques into construction. The selected papers were carefully observed to identify the key themes discussed and counted into one sub-category among the three (as mentioned above). To all the above sub and sub-sub categories, frequency of citations was calculated. This profiling helps developing and understanding the state-of-the-art of big data research as well as the significant gaps both in theory and practice. If an article speculates more than one topic, it was classified into the category perceived as predominant.

Identification of knowledge gaps is the second part of the analysis, which was again carried out systematically. The future directions, suggestions, limitations, uncovered areas were carefully read and scrutinised in each selected paper and a list of gaps was developed. The list (with short running titles) initially resulted with 26 key gap areas. These were further brought down into seven using simple manual factor analysis. The final list of gaps includes; sustainability, mass-scale data use, contribution to supply chain, skills/knowledge dimensions, qualitative/quantitative data difference, other innovative technology synergies with big data, interoperability issues and whole life cycle perspective; which are detailed described in the subsequent paragraphs.

## Literature review analysis

### *Data analysis*

Majority of the papers reviewed were research papers (40 per cent), technical papers (22 per cent) and case studies (20 per cent) predominantly based on case/study research methods (35 per cent). A significant rise in construction big data research is also identified with an increasing trend in number of yearly articles. Figure 1 shows how research interests in big data has risen almost exponentially from 2007 to 2017 implying that BDA is an extremely emerging topic which reports to be in its pinnacle at this time.

According to Figure 2, the main theme discussed in the literature was big data opportunities and challenges/drivers and barriers for adoption (42 per cent) that vary from project domain to firm level as well as to industry and national level. The opportunities as identified by SLR in this study inter alia: value creation, generation of business intelligence for informed business decisions, visualisation of patterns trends-correlations, process optimisation, enhancing the flexibility of supply chain and resource allocation, assets/resources management, productivity growth, competitive advantage, etc.

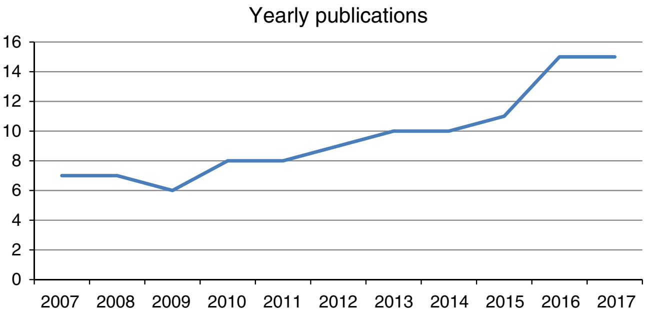


**Figure 1.** Number of articles published on construction-specific big data research in each year (2007–2017)

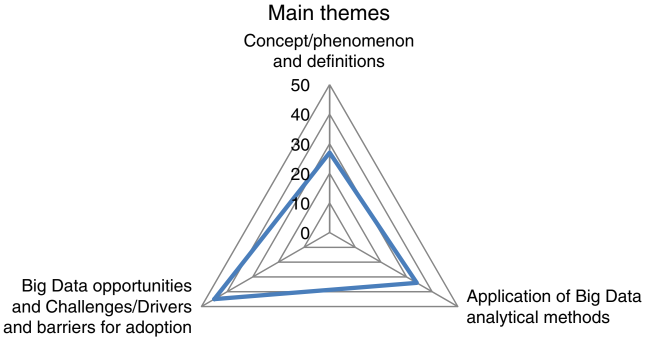


**Figure 2.** Main themes discussed in reviewed articles

The challenges include, data privacy data security, infrastructure and maintenance costs, lack of skills and training, hardware and software complexities for data integration and data synchronising.

Application of big data analytical methods (31 per cent) was secondly on the discourse while big data as a concept/phenomenon (21 per cent) is the least discussed theme.

### *Data synthesis*

*Big data as a concept.* The data concept of being "big" is difficult to define as what appears to be big today may appear to be smaller in the near future (MIT, 2013). Further, no evidence proves that massive data sets are always complex or small data sets are always simple (Sivarajah *et al.*, 2017). Therefore, the complexity of the data is not confined to data size, since data sets will increase in the future. Supporting this argument, McGuire *et al.* (2012, p. 559) refer big data to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse. A similar definition was provided by Amir Gandomi and Haider (2015) with an explanation of three 'v's (volume, variety and velocity). Bilal, Oyedele, Akinade *et al.* (2016); Bilal, Oyedele, Qadir *et al.* (2016) also bring out the 3'V's concept as clear evident in construction data. Because, construction data is large, heterogeneous and dynamic (Aouad *et al.*, 1999), they become voluminous due to large volumes of data gathered at a fast moving speed. These data as suggested by Bilal, Oyedele, Akinade *et al.* (2016); Bilal, Oyedele, Qadir *et al.* (2016) include: design data, schedules, enterprise resource planning (ERP) systems, financial data, etc. The latter authors have studied the diversification of these data in such a deep level that they can be categorised into single format (e.g. DWG, DXF, DGN, RVT, ifcXML, XML, ifcOWL, OWL, DOC/XLS/PPT, RM/MPG and JPEG). This dynamic nature of construction data has allowed systems to stream through sensors, RFIDs and BMS (building management sensors). This rich intellectual performance of sensors has been successfully tested by Akhavian and Behzadan (2015) with the use of smartphone sensors and machine learning classifiers to improve the quality and reliability of project decision-making and control. Thus, Bilal, Oyedele, Akinade *et al.* (2016); Bilal, Oyedele, Qadir *et al.* (2016) purport that utilisation of these voluminous data in an optimum manner would indisputably be the next frontier of innovation in the construction industry. The authors further distinguish the concept of big data engineering (BDE) and BDA as two concepts that needs careful consideration.

*Big data analytical methods/techniques.* Amir Gandomi and Haider (2015) describe several popular BDA techniques specifically can be used in construction sector. These techniques include text analytics (data mining) like information extraction (IE), entity recognition (ER) and relation extraction (RE), text summarisation, question answering (QA), sentiment analysis – data mining (Fan *et al.*, 2015), content-based analytics, structure-based analytics (Chen *et al.*, 2016 for cloud-based system framework for structured BIM data); audio analytics (ref) and visual-image/video analytics (Han and Golparvar-Fard, 2017); social media analytics (Tang *et al.*, 2017), GIS analytics (Buffat *et al.*, 2017) and predictive analytics (Li, 2017; Fan *et al.*, 2017).

Bilal, Oyedele, Akinade *et al.* (2016); Bilal, Oyedele, Qadir *et al.* (2016) suggest possible BDE and BDA methods that can be successfully applied into construction practices, seeing BDE as an infrastructure provider to support BDA. The authors further introduce big data processing techniques (Map Reduce (MR), Directed Acyclic Graphs (DAG)) and storage techniques (distributed file systems, NoSQL databases) for BDE. Map Reduce process has been widely used in construction as a mean of laser scanning/point scanning for high definition surveying (HDS) where important information can be extracted from LiDAR point clouds such as the location, orientation and size of objects and possible damage. Aljumaily *et al.* (2016) propose a big data approach-mapreduce process to automatically identify and extract buildings from a digital surface model created from aerial laser scanning data. Han and Golparvar-Fard (2017) use a similar point cloud system to analyse the building performance. On the other hand, for BDA techniques such as statistics: data mining, Machine learning techniques, regression, classification and clustering are reported. The construction industry has employed some of these statistical methods in a variety of application areas, such as identifying causes of construction delays (Chau *et al.*, 2003) learning from post-project reviews (PPRs) (Carrillo *et al.*, 2011), decision support for construction litigation (Jordan and Mitchell, 2015; Mahfouz, 2009), detecting structural damages of buildings (Jiang and Mahadevan, 2008), identifying actions of workers and heavy machinery (Gong *et al.*, 2011; Huang and Beck, 2013).

Chau *et al.* (2003) in their study of identifying critical factors for construction delays has employed data mining techniques to capture ML algorithms to produce knowledge discovery dataset (KDD). Another study conducted by Buchheit *et al.* (2000) also showcased a KDD process for a project related to construction of infrastructure. KDD process is reinvented by Soibelman and Kim (2002) to illustrate it is applicability to construction industry in identifying construction disputes such as; delays, cost overrun and quality failures. While Carrillo *et al.* (2011) employ data mining techniques to use past projects as learning material for risk free future projects, Liao and Perng (2008) attempted to employ association rule mining technique to investigate the prevention extent of workplace as well as occupational health safety threats. A similar study has been conducted by Cheng *et al.* (2012) using data mining to investigate the occupational injuries in construction sites. Interestingly, the system was competent enough to reveal the most impactful causes as; falls and collapses.

Data warehousing (DW) is another technique employed by many construction-related studies. Chau *et al.* (2003) and Kimball and Ross (2011) used DW to evaluate construction productivity data by an SQL multi-layer analysis. SQL has been reported for its wide usage specifically in construction for is facilitation for querying partial BIM models query languages such as express query language (EQL) and Building Information Modelling Query Language (BIMQL) (Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016). Such studies include Kimball and Ross (2011) and Koonce and Judd (2001).

Machine learning is a type of predictive statistics that is widely applied in construction data predictions. Arditi and Pulket (2005) a closer version of artificial intelligence (AI) allowing a programme to learn from data about specific task automatically and predict the possible future outcomes. Machine learning rule-based learning is an industry-wide application where many researchers found beneficial with a variety of applications such as artificial neural networks methods, case-based reasoning techniques and hybrid methodologies (Ahn and Kim, 2009; Arditi and Pulket, 2005, 2010; Chau, 2005, 2006, 2007; Chen and Hsu, 2007; Cheng *et al.*, 2009; Choi *et al.*, 2014; Du *et al.*, 2010; Pulket and Arditi, 2009; Sanyal *et al.*, 2014). A similar study carried out by Sacks *et al.* (2018) introduced a time and cost saving automated method of checking building designs for code compliance using machine learning technique. This method is considered to be highly beneficial at the phase of pre-processing and preparing BIM models for checking. Predictive analytics has been widely used in construction activity predictions through simulations. Li (2017) in his investigation of stadium construction used a constrained parametric index analysis model of the progress analysis to monitor, predict and control the resource (i.e manpower, scheduling) progress.

Regression models have also been in the use for many years in construction research. The use of regression model often comes with a machine learning technique. Many of the studies have used regression models to predict construction tender prices, cost estimates and material price fluctuations (Cheng *et al.*, 2009; Fallis, 2013; Lau *et al.*, 2010; Narbaev and De Marco, 2014). Using a similar technique Shrestha *et al.* (2017) used dynamic items basket (DIB) method, which is based on regression modelling for large amount highway bid data to establish a framework for improved calculation process of Highway Construction Cost Index (HCCI).

*Big data opportunities and challenges.* There is a considerable body of research on the opportunities and challenges offered by BDA (Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016; Wamba *et al.*, 2016; Wang *et al.*, 2016; Devlin, 2016).

The actual challenge of tackling with big data as suggested by Mishra *et al.* (2017, p. 28) "was to deal with diversified data types (variety), timely response requirements (velocity) and uncertainties in the data (veracity)". Mishra and Sharma (2015) advocate handling both semi-structured and unstructured data is challenging especially when they are not received in a timely manner. The reason as latter authors mention is mainly be due to the lack of insufficient sources needed to gather, store and analyse big data but within a particular time frame. The authors further mention the reliability of data is also a big issue where additionally cleansing methods required to be applied in order to mitigate the uncertainty which may consume much more time and resources. Manyika *et al.* (2011) state identifying the exact area of applications as one of the biggest challenges in big data. Besides, Data security, privacy and protection, quality of construction industry data sets, cost implications for big data in construction industry, internet connectivity for big data applications, exploiting big data to its full potential are also discussed in the existing literature (Tene and Polonetsky, 2013; Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016).

The greatest opportunities as mentioned by many articles are, Resource and waste optimisation (Bilal *et al.*, 2015; Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016; Oyedele, 2016; Oyedele *et al.*, 2013; Lu *et al.*, 2015, 2016), Generative designs and clash detection and resolution (Nima, 2014; Wang and Leite, 2013), performance prediction (Abaza *et al.*, 2004; Kobayashi *et al.*, 2010), visual analytics (Goodwin and Dykes, 2012; Löfström and Palm, 2008), social networking services/analytics (Demirkesen and Ozorhon, 2017; Wolf *et al.*, 2009), personalised services (Liu *et al.*, 2012; Singh *et al.*, 2010), facility management (Isikdag *et al.*, 2013; Liu *et al.*, 2012; Rueppel and Stuebbe, 2008; Taneja *et al.*, 2012), Energy management and analytics (Hong *et al.*, 2012; Linda *et al.*, 2012; Sanyal and New, 2013), big data integration with BIM (Volk *et al.*, 2014), big data integration with IOT and cloud computing (Elghamrawy and Boukamp, 2010), big data for augmented reality (AR) (Williams *et al.*, 2015), use of big data for smart buildings and smart city/urban infrastructure (Khan and Hornbæk, 2011; Liu *et al.*, 2014). Further, Mishra *et al.* (2017) and Janssen *et al.* (2017) purport that exploitation of BDA can lead to gain competitive advantage at any level.

*Big data application.* There are number of studies address the applicability of big data techniques into construction specifically for performance and process optimisation (Eriksson *et al.*, 2017; Becerik-Gerber *et al.*, 2012; Bilal *et al.*, 2015; Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016; Demchenko *et al.*, 2014; Koseleva and Ropaite, 2017; Lu *et al.*, 2015; Zhang *et al.*, 2015; Qadir *et al.*, 2016; Hao *et al.*, 2015; Alaka *et al.*, 2015; Rathorea *et al.*, 2016; An, 2014).

Motawa (2017) deployed BDA into BIM system to capture buildings operation knowledge, particularly for building maintenance and refurbishment. The proposed big data technique in this study was cloud-based spoken dialogue system and case-based reasoning BIM system. Thus, the study tried to answer problems specific to building maintenance.

In a similar study conducted to investigate the leading Hong Kong construction firms' efficiency by Chiang *et al.* (2013) used variation in weights for estimation and enhances the adequacy for individual contractor's efficiency scores. A study conducted by Mansouri and Akhavian (2018) discussed the importance of early engagement of stakeholders as well as early incorporation of big data principles into construction projects, to use big data in its full potential. The authors explain how it improves the forecasting/planning process and manage both explicit and tacit knowledge provisions of the project.

## Discussion

### *Significant knowledge gaps- social, economic and environmental implications of big data in construction*

Although there are plenty of research in the application of big data in construction. There is a lack in focus on the implications of BDA on to the social, economic and environmental dimensions of sustainability. Subsequent paragraphs critically review some of the literature and discuss these implications.

### *Social implications*

Implication of big data on the society as cited by many of the relevant studies are tied-up to people's quality of living and their behaviours/attitude for the post-occupancy of buildings. Few see big data as a convergence of people, place and technology as it helps to improve understanding on common data environments (Cook, 2015) as well as to control and/or maximise the social interactions through social media (Tang *et al.*, 2017). In opposed to this positive impact, the negative impact on human relationships caused by humans' constant connection with digital data which ends up as an addiction is an under-researched area in construction. As a finding of Tang *et al.*' s (2017) study, it has been presented the mostly tweeted words by construction workers, but the study does not address how this finding benefit to improve the health and well-being of them. Thus, it is advisable to undertake further research on how social media BDA can be used to prevent threats such as safety, injury or mental illness caused by work-related stress. On a different dimension, Big data techniques have proven to monitor and analyse indoor quality of space such as air quality, noise, light, etc. Such empowers occupiers on useful information on health and well-being attributes of spaces and it is sign of improved sustainability to the society. In light of same direction, Zhu and Ge (2014) investigate the social impact of green buildings focusing on user satisfaction of green buildings using a big data post-occupancy evaluation. However, many of these studies related to green building performance evaluation lacks occupiers' subjective as well as objective opinion evaluation. Further, big data with its asset management have massive potential for positive social implications as it allows for meaningful business decisions based upon the life cycle of building not limited to the capital cost alone (Cook, 2015). However, the studies lack the robustness of lifecycle studies as to how these data can be reused for future benefit. The ways to leverage any drop of data, is to use them over and over again, before it becomes stale and that indeed saves massive amount of time as well. Akhavian and Behzadan (2015) profess through mobile sensor data quality and reliability of project decision-making could be improved, Despite the advantages big data offers, some of the social implications like disruption to social interactions, quality assurance, need for mutually agreed standards/guidelines, intellectual property, privacy and security issues are still not entirely addressed. Finally, need for skills, knowledge and training is another important social implication emerged from the review. Currently there is a huge demand for data management capabilities among professionals, help embed use of new technology across the built environment sector and generate new avenues for value creation from the vast amounts of data available in construction industry as well as all other sectors that is linked into it.

### *Economic implications*

Cook (2015) in his paper based upon the 2015 RICS/SPR Cutting Edge conference purports that there is major change impacting from the rise of digital technology to creating value from big data. One of the main areas of impact as he suggested would be changes to job roles and business structures, requiring continuous learning and greater flexibility and adaptability for economic stability of the construction industry as its greater potential to many add value and generate more income. Again, Big data with its potential for future prediction models thrives longer term view on total life-cycle costs rather than short term financial implications, (Cook, 2015). However, the studies that address the economic value big data do not exactly evaluate and state as to how big data could be exploited to thrive competitiveness. The effectiveness of big data cannot be measured just by accumulating large volumes of data; it is more of the use cases or industrial problems that dictate the usefulness of these technologies (Bilal, Oyedele, Akinade *et al.*, 2016; Bilal, Oyedele, Qadir *et al.*, 2016). Thus, an evidently proven case study analysis would help practitioners to see how big data could maximise competitive edge to stay on top of the market. Progress prediction model produced by Li (2017) and Shrestha *et al.* (2017) is good examples of how big data improve estimation and bidding process. Again, how this index can be used to improve bidding success is an undiscussed area. Further, Cook (2015) emphasises, as a result of big data hype, businesses are starting to be keener on grouping and sharing economies with short term needs such as entrepreneurships. On what complex and dynamic ways big data facilitate these sharing economies is still unexplored. BDA supported by mobile technology with its geospatial capability adds a valuable dimension (Akhavian and Behzadan, 2015) that is already of great interest for real estate and building developers (Buffat *et al.*, 2017; Shrestha *et al.*, 2017). By understanding data and big analytical techniques, these businesses can better understand the current and future client demand and better target their new potential clients – which is a massive implication for sectoral economy as a whole. Nevertheless, client behaviour analysis is very common in sectors like retail but not much in construction. The cost implications of big data, considered amongst the low-profit-margin businesses (construction industry) is another under-researched area. Chau *et al.* (2003) mention that adoption of big data incorporates costly endeavours such as data centre purchases and software licensing, such costly add-ons to projects are more likely to be opposed difficult to be defended on low-profit margins. Hence it would be more worthwhile conducting more "quantified" research that presents the business case on the extent of financial return for big data investment which would help strategic decision makers for their investment decisions. More studies on cost-benefit analysis of using big data technologies in construction projects are required to this end. Since BDA is often involved with large data sets, it would be beneficial if these percentages of revenue making could be examined in larger scale (predominantly considering the entire construction industry). The current body of knowledge predominantly limits to micro levels studies like project (Han and Golparvar-Fard, 2017; Zhang *et al.*, 2015; An, 2014) level.

It was noticeable that many of the studies used "Building Information Modelling-BIM" (Han and Golparvar-Fard, 2017; Motawa, 2017) and "Internet of things-IOT" (Akhavian and Behzadan, 2015) interchangeability to explain the application of big data as there is a closer relationship between the technologies in terms data sharing. However, considering the dynamic and competitive environment of today's world, it is imperative to embrace innovative technologies and their synergies (Demirkesen and Ozorhon, 2017). Thus, yet, no research has been undertaken exploring the synergies between them.

### *Environmental implications*

Numerous studies have conducted addressing the issues with energy demand and presented new approaches for building stock modelling (Buffat *et al.* 2017; Fan *et al.*, 2015, 2017; Moreno *et al.*, 2016; Mathew *et al.*, 2015; Yu *et al.*, 2016; Sanyal *et al.*, 2014) using geographical/spatial data sets such as building footprints and digital elevation models and building automation systems. The model helps controlling the building heat demand for various climate conditions and improves energy efficiency and conservation – which shows a positive environmental impact of BDA. This improves understanding of the impact of climate change in future years and allows designers ascertain heating and cooling loads in different parts of the world which ultimately saves massive amount of energy. Further, an accurate estimation of future climate conditions supported by an energy simulation has the advantage of decreasing the extensive supply and demand for energy without neglecting the extremes and variations of the possible climate changes (Nik, 2016; Nik *et al.*, 2016). However, in most of the energy studies reviewed, selecting a suitable measure is an important step- typically a multi-criteria decision-making procedure (Nik, 2016), affected by several factors related to economy, availability, etc., which have not been considered in most of the work. It is also apparent that more research required to be produced as solutions to disasters which is a major effect of climate change. Disaster resilience and coordination in humanitarian operations and supply chains, is a shortage in existing body of literature. Another study conducted by Chen and Lu (2017) looked in to the use of big data sets to minimise demolition waste in Hong Kong which had numerous environmental implications including mitigating adverse impacts – i.e. land deterioration, resource depletion and various forms of pollution such as noise, dust, air and discharge of toxic waste. It is worthwhile noting that, many of the suggested waste reduction tools are yet to be validated with use cases. Although there are many research studies claiming how BDA encourage efficient process efficiency and optimisation (Motawa, 2017; Li, 2017) there is a lack of connection between how it actually makes agile and adaptable to dynamic business environments. There is a gap identified in the areas of how BDA contributes supply chain design by focusing on main characteristics of supply chain including agility, adaptability, alignment and integration.

### *Summary of findings – gaps in existing body of knowledge for future research*

#### *Social*

- How social media bid data analytics can be used to prevent threats such as safety, injury or mental illness caused by work-related stress.
- The negative impact on human relationships caused by humans' constant connection with digital data which ends up as an addiction.
- How to leverage any drop of data in life cycle studies by reusing data.
- Green building post-occupancy evaluation should consider occupiers' both subjective and objective opinions.
- Issues related to Quality assurance, intellectual property, privacy and security.
- Need for mutually agreed standards/guidelines.
- Current and future needs for skills, knowledge and training for big data.

#### *Economic*

- Use cases for the exploitation of big data in its full potential for competitive advantage (at every level i.e. project, organisation, sector).
- How bidding progress prediction models can be used to improve bidding success rate.
- In what complex and dynamic ways big data facilitate sharing economies within organisations.
- Client behaviour analysis to predict market conditions.
- Business case on the extent of financial return for big data investment which would help strategic decision makers for their investment decisions.

- Cost-benefit analysis of using big data technologies in construction projects
- Synergies between BDA, BIM and IOT.

*Environmental*

- Predictive analytics for Disaster risk reduction and climate change.
- Multi-criteria decision-making procedure for energy studies.
- Validation of waste reduction tools with use-cases.
- Supply chain agility and adaptability.

## Conclusion

The main purpose of this paper is to identify the gaps and potential future research avenues in the big data research in the context of construction industry. The authors observe a greater demand as well as prospects for increased use of big data methods and applications within construction and highlight that as the need for this research. Gaps in knowledge in current research efforts are identified through an SLR. After in-depth discussion on gaps-in-knowledge, insights in different research areas are discussed in line with the social, environmental and economic implications to provide a comprehensive big picture toward which big data- related areas need further attention for the competitive advantage of the construction industry. The implications reflect both positive and negative impacts while the issues around big data, i.e. data privacy is still an on-going research topic that needs further investigation. Some of the gaps identified inter alia: impact of big data on health and wellbeing of construction workers, need for mutually agreed standards/guidelines, contribution to supply chain agility, impact of big data in disaster risk reduction, skill/knowledge dimensions. It also interesting that many authors have purported that there is a potential to combine different types of innovative technologies with big data techniques to maximise the effectiveness potential, although they have not precisely outlined the "how" part. It is crucial that the real power of big data is properly discoursed (i.e. how the predictability helps generating valuable insights and thereby informed decisions) through academic channels. To that end, identification of contemporary gaps aids continuous improvement.

It is worthwhile mentioning that the SLR presented in this paper limits to peer-reviewed journal articles only. In fact, there has been increasing number of indexed conferences in big data, smart systems and digital information and communication technology-related areas. However, this has been identified as a limitation of this paper and therefore the conclusions may have impacts on the generalisability as well as the representation of the sample of papers reviewed. The paper acknowledges this as further Research Avenue for future investigation.

## References

Abaza, K.A., Ashur, S.A. and Al-Khatib, I.A. (2004), "Integrated pavement management system with a markovian prediction model", *Journal of Transportation Engineering*, Vol. 130 No. 1, pp. 24-33.

Ahn, H. and Kim, K. (2009), "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach", *Applied Soft Computing*, Vol. 9 No. 2, pp. 599-607.

Akhavian, R. and Behzadan, A.H. (2015), "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers", *Advanced Engineering Informatics*, Vol. 29 No. 4, pp. 867-877.

Alaka, H., Oyedele, L., Bilal, M., Akinade, O., Owolabi, H. and Ajayi, S. (2015), "Bankruptcy prediction of construction businesses: towards a big data analytics approach", *2015 IEEE First International Conference on Big Data Computing Service and Applications*, *San Francisco, CA, 30 March-3 April*, pp. 347-352.

Aljumaily, H., Laefer, D.F. and Cuadra, D. (2016), "Big-data approach for three-dimensional building extraction from aerial laser scanning", *Journal of Computing in Civil Engineering*, Vol. 30 No. 3, pp. 1-10.

An, Q. (2014), "The effective classification process analysis of the big data in construction project", *Applied Mechanics and Materials*, Vol. 44 No. 50, pp. 1749-1751.

Aouad, G., Kagioglou, M., Cooper, R., Hinks, J. and Sexton, M. (1999), "Technology management of IT in construction: a driver or an enabler?", *Logistics Information Management*, Vol. 12 No. 2, pp. 130-137.

Arditi, D. and Pulket, T. (2005), "Predicting the outcome of construction litigation using boosted decision trees", *Journal of Computing in Civil Engineering*, Vol. 19 No. 4, pp. 387-393.

Arditi, D. and Pulket, T. (2010), "Predicting the outcome of construction litigation using an integrated artificial intelligence model", *Journal of Computing in Civil Engineering*, Vol. 24 No. 1, pp. 73-80.

Becerik-Gerber, B., Jazizadeh, F., Li, N. and Calis, G. (2012), "Application areas and data requirements for BIM-enabled facilities management", *Journal of Construction Engineering and Management*, Vol. 138 No. 3, pp. 431-442.

Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Akinade, O.O., Ajayi, S.O., Alaka, H.A. and Owolabi, H.A. (2015), "Analysis of critical features and evaluation of BIM software: towards a plug-in for construction waste minimization using big data", *International Journal of Sustainable Building Technology and Urban Development*, Vol. 6 No. 4, pp. 211-228.

Bilal, M., Oyedele, L.O., Akinade, O.O., Ajayi, S.O., Alaka, H.A., Owolabi, H.A., Qadir, J., Pasha, M. and Bello, S.A. (2016), "Big data architecture for construction waste analytics (CWA): a conceptual framework", *Journal of Building Engineering*, Vol. 6 No. 2, pp. 44-156.

Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A. and Pasha, M. (2016), "Big data in the construction industry: a review of present status, opportunities, and future trends", *Advanced Engineering Informatics*, Vol. 30 No. 3, pp. 500-521.

Buchheit, R., Garrett, J., Lee, S. and Brahme, R. (2000), "A knowledge discovery framework for civil infrastructure: a case study of the intelligent workplace", *Engineering with Computers*, Vol. 16 No. 4, pp. 264-274.

Buffat, R., Froemelt, A., Heeren, N., Raubal, M. and Hellweg, S. (2017), "Big data GIS analysis for novel approaches in building stock modelling", *Applied Energy*, Vol. 208 No. 2017, pp. 277-290.

Carrillo, P., Harding, J. and Choudhary, A. (2011), "Knowledge discovery from post-project reviews", *Construction Management and Economics*, Vol. 29 No. 7, pp. 713-723.

Chau, K.W. (2005), "Predicting construction litigation outcome using particle swarm optimization", in Ali, M. and Esposito, F. (Eds), *Innovations in Applied Artificial Intelligence*, Lecture Notes in Computer Science, IEA/AIE, Springer, Heidelberg, Berlin, Vol. 3533.

Chau, K.W. (2006), "Prediction of construction litigation outcome – a case-based reasoning approach", in Ali, M. and Esposito, F. (Eds), *Innovations in Applied Artificial Intelligence*, Lecture Notes in Computer Science, IEA/AIE, Springer, Heidelberg, Berlin and Annecy, pp. 548-553.

Chau, K.W. (2007), "Application of a PSO-based neural network in analysis of outcomes of construction claims", *Automation in Construction*, Vol. 16 No. 5, pp. 642-646.

Chau, K.W., Cao, Y., Anson, M. and Zhang, J. (2003), "Application of data warehouse and decision support system in construction management", *Automation in Construction*, Vol. 12 No. 2, pp. 213-224.

Chen, H.M., Chang, K.C. and Lin, T.H. (2016), "A cloud-based system framework for performing online viewing, storage, and analysis on big data of massive BIMs", *Automation in Construction*, Vol. 71 No. 2016, pp. 34-48.

Chen, J.H. and Hsu, S.C. (2007), "Hybrid ANN-CBR model for disputed change orders in construction projects", *Automation in Construction*, Vol. 17 No. 1, pp. 56-64.

Chen, X. and Lu, W. (2017), "Identifying factors influencing demolition waste generation in Hong Kong", *Journal of Cleaner Production*, Vol. 141 No. 2017, pp. 799-811.

Cheng, C.W., Leu, S.S., Cheng, Y.M., Wu, T.C. and Lin, C.C. (2012), "Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry", *Accident Analysis & Prevention*, Vol. 48 No. 2012, pp. 214-222.

Cheng, M.Y., Tsai, H.C. and Chiu, Y.H. (2009), "Fuzzy case-based reasoning for coping with construction disputes", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 4106-4113.

Chiang, At.H., Li, J., Choi, T.N.Y. and Man, K.F. (2013), "Evaluating construction contractors' efficiency in Hong Kong using data envelopment analysis assurance region model", *Journal of Facilities Management*, Vol. 11 No. 1, pp. 52-68.

Choi, S., Kim, D.Y., Han, S.H. and Kwak, Y.H. (2014), "Conceptual cost-prediction model for public road planning via rough set theory and case-based reasoning", *Journal of Construction Engineering and Management*, Vol. 140 No. 1, pp. 401-426.

Cook, D. (2015), "RICS futures: turning disruption from technology to opportunity", *Journal of Property Investment and Finance*, Vol. 33 No. 5, pp. 456-464.

Demchenko, Y., De Laat, C. and Membrey, P. (2014), "Defining architecture components of the big data ecosystem", *2014 International Conference on Collaboration Technologies and Systems, Minneapolis, MN, 19-23 May*, pp. 104-112.

Demirkesen, S. and Ozorhon, B. (2017), "Impact of integration management on construction project management performance", *International Journal of Project Management*, Vol. 35 No. 8, pp. 1639-1654.

Devlin, B. (2016), "Cultivating success in big data analytics", *Cutter, IT Journal*, Vol. 29 No. 6, pp. 8-16.

Du, Y., Wen, W., Cao, F. and Ji, M. (2010), "A case-based reasoning approach for land use change prediction", *Expert Systems with Applications*, Vol. 37 No. 8, pp. 5745-5750.

Eadie, R., Browne, M., Odeyinka, H., McKeown, C. and McNiff, S. (2013), "BIM implementation throughout the UK construction project lifecycle: an analysis", *Automation in Construction*, Vol. 36, pp. 145-151.

Elghamrawy, T. and Boukamp, F. (2010), "Managing construction information using RFID-based semantic contexts", *Automation in Construction*, Vol. 19 No. 8, pp. 1056-1066.

Elsevier (2017), *Scopus Content Overview*, Elsevier, Amsterdam, available at: www.elsevier.com/solutions/scopus/content (accessed 15 January 2018).

Eriksson, C., Cheng, I., Pitman, K., Dixon, T., Van De Wetering, J. and Sexton, M. (2017), "Smart cities, big data and the built environment: what's required?", available at: www.rics.org/Global/RICS-Smart-Cities-Big-Data-REPORT-2017.pdf (accessed 12 February 2018).

Fallis, A. (2013), "Applying regression analysis to predict and classify construction cycle time", *Journal of Chemical Information and Modeling*, Vol. 53 No. 9, pp. 1689-1699.

Fan, C., Xiao, F. and Zhao, Y. (2017), "A short-term building cooling load prediction method using deep learning algorithms", *Applied Energy*, Vol. 195, pp. 222-233.

Fan, C., Xiao, F., Madsen, H. and Wang, D. (2015), "Temporal knowledge discovery in big BAS data for building energy management", *Energy and Buildings*, Vol. 109, pp. 75-89.

Gandomi, A. and Haider, M. (2015), "Beyond the hype: big data concepts, methods, and analytics", *International Journal of Information Management*, Vol. 35 No. 2, pp. 137-144.

Gong, J., Caldas, C.H. and Gordon, C. (2011), "Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and Bayesian network models", *Advanced Engineering Informatics*, Vol. 25 No. 4, pp. 771-782.

Goodwin, S. and Dykes, J. (2012), "Visualising variations in household energy consumption", *IEEE Conference on Visual Analytics Science and Technology 2012 – Proceedings*, pp. 217-218.

Han, K.K. and Golparvar-Fard, M. (2017), "Potential of big visual data and building information modeling for construction performance analytics: an exploratory study", *Automation in Construction*, Vol. 73, January, pp. 184-198.

Hao, J., Zhu, J. and Zhong, R. (2015), "The rise of big data on urban studies and planning practices in China: review and open research issues", *Journal of Urban Management*, Vol. 4 No. 2, pp. 92-124.

Hong, I., Byun, J. and Park, S. (2012), "Cloud computing-based building energy management system with ZigBee sensor network", *Proceedings – 6th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2012*, pp. 547-551.

Huang, Y. and Beck, J.L. (2013), "Novel Sparse Bayesian learning for structural health monitoring using incomplete modal data", ASCE International Workshop on Computing in Civil Engineering, Los Angeles, CA.

Isikdag, U., Zlatanova, S. and Underwood, J. (2013), "A BIM-oriented model for supporting indoor navigation requirements", *Computers, Environment and Urban Systems*, Vol. 41, September, pp. 112-123.

Janssen, M., van der Voort, H. and Wahyudi, A. (2017), "Factors influencing big data decision-making quality", *Journal of Business Research*, Vol. 70, January, pp. 338-345.

Jiang, X. and Mahadevan, S. (2008), "Bayesian probabilistic inference for nonparametric damage detection of structures", *Journal of Engineering Mechanics*, Vol. 134 No. 10, pp. 820-831.

Jordan, M.I. and Mitchell, T.M. (2015), "Machine learning: trends, perspectives, and prospects", *Science*, Vol. 349 No. 6245, pp. 255-260.

Jukić, N., Sharma, A., Nestorov, S. and Jukić, B. (2015), "Augmenting data warehouses with big data", *Information Systems Management*, Vol. 32 No. 3, pp. 200-209.

Khan, A. and Hornbæk, K. (2011), "Big data from the built environment", *Proceedings of the 2nd International Workshop on Research in the Large – LARGE 2011*, pp. 29-34.

Kimball, R. and Ross, M. (2011), *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling*, Nachdr, Wiley, New York, NY.

Kobayashi, K., Do, M. and Han, D. (2010), "Estimation of Markovian transition probabilities for pavement deterioration forecasting", *KSCE Journal of Civil Engineering*, Vol. 14 No. 3, pp. 343-351.

Koonce, D.A. and Judd, R.P. (2001), "A visual modelling language for EXPRESS schema", *International Journal of Computer Integrated Manufacturing*, Vol. 14 No. 5, pp. 457-472.

Koseleva, N. and Ropaite, G. (2017), "Big data in building energy efficiency: understanding of big data and main challenges", *Procedia Engineering*, Vol. 172 No. 2017, pp. 544-549.

Lau, S.C., Lu, M. and Ariaratnam, S.T. (2010), "Applying radial basis function neural networks to estimate next-cycle production rates in tunnelling construction", *Tunnelling and Underground Space Technology*, Vol. 25 No. 4, pp. 357-365.

Levitt, R.E., M.ASCE (2007), "CEM Research for the next 50 years: maximizing economic, environmental, and societal value of the built environment", *Journal of Construction Engineering and Management*, Vol. 133 No. 9, pp. 619-628.

Li, H. (2017), "Estimation of stadium construction schedule based on big data analysis", *International Journal of Computers and Applications*, Vol. 41 No. 4.

Liao, C.W. and Perng, Y.H. (2008), "Data mining for occupational injuries in the Taiwan construction industry", *Safety Science*, Vol. 46 No. 7, pp. 1091-1102.

Linda, O., Wijayasekara, D., Manic, M. and Rieger, C. (2012), "Computational intelligence-based anomaly detection for building energy management systems", *Resilient Control Systems (ISRCS), 2012 5th International Symposium*, pp. 77-82.

Liu, J., Yao, R., Wang, J. and Li, B. (2012), "Occupants' behavioural adaptation in workplaces with non-central heating and cooling systems", *Applied Thermal Engineering*, Vol. 35 No. 1, pp. 40-54.

Liu, X., Iftikhar, N. and Xie, X. (2014), "Survey of real-time processing systems for big data", *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS 2014*, pp. 356-361.

Löfström, E. and Palm, J. (2008), "Visualising household energy use in the interest of developing sustainable energy systems", *Housing Studies*, Vol. 23 No. 6, pp. 935-940.

Lu, W., Chen, X., Ho, D.C.W. and Wang, H. (2016), "Analysis of the construction waste management performance in Hong Kong: the public and private sectors compared using big data", *Journal of Cleaner Production*, Vol. 112, pp. 521-531.

Lu, W., Chen, X., Peng, Y. and Shen, L. (2015), "Benchmarking construction waste management performance using big data", *Resources Conservation and Recycling*, Vol. 105, Part A, pp. 49-58.

McGuire, T., Chui, M. and Manyika, J. (2012), "Why big data is the new competitive advantage", *Ivey Business Journal*, Vol. 76, July/August, pp. 1-4.

Mahfouz, T.S. (2009), *Construction Legal Support for Differing Site Conditions (DSC) Through Statistical Modeling and Machine Learning (ML)*, Graduate Theses and Dissertation No. 10698, Lowa State University.

Mansouri, S. and Akhavian, R. (2018), "The status quo and future potentials of data analytics in aec/fm: a quantitative analysis of academic research and industry outlook", *Proceedings of Construction Research Congress, American Society of Civil Engineers*, *New Orleans, LA*, *2-4 April*, pp. 90-100.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byres, A.H. (2011), *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, San Francisco, CA, June, p. 156.

Marr, B. (2017), *Big Data Analytics- What it is and Why it Matters*, SAS, Cary, NC, available at: www.sas.com/en_gb/insights/analytics/big-data-analytics.html

Mathew, P.A., Dunn, L.N., Sohn, M.D., Mercado, A., Custudio, C. and Walter, T. (2015), "Big-data for building energy performance: lessons from assembling a very large national database of building energy use", *Applied Energy*, Vol. 140, 15 February, pp. 85-93.

Mayer-Schonberger, V. and Cukier, K. (2012), "Big data: a revolution that will transform how we live, work, and think", *CEUR Workshop Proceedings*, Houghton Mifflin Harcourt, Boston, MA.

Mishra, D., Luo, Z., Jiang, S., Papadopoulos, T. and Dubey, R. (2017), "A bibliographic study on big data: concepts, trends and challenges", *Business Process Management Journal*, Vol. 23 No. 3, pp. 555-573.

Mishra, R. and Sharma, R. (2015), "Big data: opportunities and challenges", *International Journal of Computer Science and Mobile Computing*, Vol. 4 No. 6, pp. 27-35.

MIT (2013), "The big data conundrum: how to define it?", *MIT Technology Management Review*, available at: www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/ (accessed 3 March 2018).

Moreno, M.V., Dufour, L., Skarmeta, A.F., Jara, A.J., Genoud, D., Ladevie, B. and Bezian, J.-J. (2016), "Big data: the key to energy efficiency in smart buildings", *Soft Computing*, Vol. 20 No. 5, pp. 749-1762.

Motawa, I. (2017), "Spoken dialogue BIM systems – an application of big data in construction", *Facilities*, Vol. 35 No. 14, pp. 787-800.

Narbaev, T. and De Marco, A. (2014), "An Earned Schedule-based regression model to improve cost estimate at completion", *International Journal of Project Management*, Vol. 32 No. 6, pp. 1007-1018.

Nik, V.M. (2016), "Making energy simulation easier for future climate – synthesizing typical and extreme weather data sets out of regional climate models (RCMs)", *Applied Energy*, Vol. 177 No. 9, pp. 204-226.

Nik, V.M., Mata, E., Sasic Kalagasidis, A. and Scartezzini, J.-L. (2016), "Effective and robust energy retrofitting measures for future climatic conditions – reduced heating demand of Swedish households", *Energy and Buildings*, Vol. 121 No. 2016, pp. 176-187.

Nima, K.O. (2014), "Comparison of machine learning techniques for developing performance prediction models", *Computing in Civil and Building Engineering*, pp. 1222-1229.

Oyedele, L.O. (2016), "Big data and sustainability: the next step for circular economy", *Presentation Material, 14th International Annual Conference, World Association for Sustainable Development (WASD), Canary Wharf, London, 20-22 July*.

Oyedele, L.O., Regan, M., von Meding, J., Ahmed, A., Ebohon, O.J. and Elnokaly, A. (2013), "Reducing waste to landfill in the UK: identifying impediments and critical solutions", *World Journal of Science, Technology and Sustainable Development*, Vol. 10 No. 2, pp. 131-142.

Pulket, T. and Arditi, D. (2009), "Construction litigation prediction system using ant colony optimization", *Construction Management and Economics*, Vol. 27 No. 3, pp. 241-251.

Qadir, J., Ahad, N., Mushtaq, E. and Bilal, M. (2016), "SDNs, clouds and big data : mutual opportunities", *Automation in Construction*, Vol. 12 No. 3, pp. 1-6.

Rathorea, M.M., Ahmad, A.A., Paul, A.A. and Rho, S. (2016), "Urban planning and building smart cities based on the internet of things using big data analytics", *Computer Networks*, Vol. 101 No. 4, pp. 63-80.

Rueppel, U. and Stuebbe, K.M. (2008), "BIM-based indoor-emergency-navigation-system for complex buildings", *Tsinghua Science and Technology*, Vol. 13 No. 1, pp. 362-367.

Sacks, R., Bloch, T. and Katz, M. (2018), "Application of machine learning for automated code compliance", *17th International Conference on Computing in Civil and Building Engineering Proceedings*, *Tampere*, *5-7 June*.

Sanyal, J. and New, J. (2013), "Simulation and big data challenges in tuning building energy models", 2013 Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES), pp. 1-6.

Sanyal, J., New, J., Edwards, R.E. and Parker, L. (2014), "Calibrating building energy models using supercomputer trained machine learning agents", *Concurrency Computation Practice and Experience*, Vol. 26 No. 13, pp. 2122-2133.

Shrestha, K.J., Jeong, H.D. and Gransberg, D.D. (2017), "Multidimensional Highway Construction Cost Indexes using dynamic item basket", *Journal of Construction Engineering and Management*, Vol. 143 No. 8, pp. 1-11.

Singh, H., Muetze, A. and Eames, P.C. (2010), "Factors influencing the uptake of heat pump technology by the UK domestic sector", *Renewable Energy*, Vol. 35 No. 4, pp. 873-878.

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017), "Critical analysis of big data challenges and analytical methods", *Journal of Business Research*, Vol. 70 No. 1, pp. 263-286.

Soibelman, L. and Kim, H. (2002), "Data preparation process for construction knowledge generation through knowledge discovery in databases", *Journal of Computing in Civil Engineering*, Vol. 16 No. 1, pp. 39-48.

Taneja, S., Akcamete, A., Akinci, B., Garrett, J.H., Soibelman, L. and East, E.W. (2012), "Analysis of three indoor localization technologies for supporting operations and maintenance field tasks", *Journal of Computing in Civil Engineering*, Vol. 26 No. 6, pp. 708-719.

Tang, L., Zhang, Y., Dai, F., Yoon, Y., Song, Y. and Sharma, R.S. (2017), "Social media data analytics for the US construction industry: preliminary study on Twitter", *Journal of Management in Engineering*, Vol. 33 No. 6, pp. 1-15.

Tene, O. and Polonetsky, J. (2013), "Big data for all: privacy and user control in the age of analytics", *Northwestern Journal of Technology and Intellectual Property*, Vol. 11 No. 5, pp. 240-273.

Van-Dijck, J. (2014), "Datafication, dataism and dataveillance: big data between scientific paradigm and ideology", *Surveillance and Society*, Vol. 12 No. 2, pp. 197-208.

Volk, R., Stengel, J. and Schultmann, F. (2014), "Building Information Modeling (BIM) for existing buildings- literature review and future needs", *Automation in Construction*, Vol. 38, pp. 109-127.

Wamba, S.F., Gunasekaran, A., Akter, S., Ren, S.J., Dubey, R. and Childe, S.J. (2016), "Big data analytics and firm performance: effects of dynamic capabilities", *Journal of Business Research*, Vol. 70 No. 8, pp. 356-365.

Wang, H., Xu, Z., Fujita, H. and Liu, S. (2016), "Towards felicitous decision making: an overview on challenges and trends of big data", *Information Sciences*, Vol. 367 No. 7, pp. 747-765.

Wang, L. and Leite, F. (2013), "Knowledge discovery of spatial conflict resolution philosophies in BIMenabled MEP design coordination using data mining techniques: a proof-of-concept", *Computing in Civil Engineering*, pp. 419-426.

Williams, G., Gheisari, M., Chen, P.-J. and Irizarry, J. (2015), "BIM2MAR: an efficient BIM translation to mobile augmented reality applications", *Journal of Management in Engineering*, Vol. 31 No. 1, pp. 401-432.

Wolf, T., Schröter, A., Damian, D. and Nguyen, T. (2009), "Predicting build failures using social network analysis on developer communication", *Proceedings – International Conference on Software Engineering*, pp. 1-11.

Yu, Z., Haghighat, F. and Fung, B.C.M. (2016), "Advances and challenges in building engineering and data mining applications for energy-efficient communities", *Sustainable Cities and Society*, Vol. 25 No. 8, pp. 33-38.

Zhang, Y., Luo, H. and He, Y. (2015), "A system for tender price evaluation of construction project based on big data", *Procedia Engineering*, Vol. 123, pp. 606-614.

Zhu, B.-F. and Ge, J. (2014), "Discussion of the evaluation method and value of green building's POE in the era of large data", *Journal of Harbin Institute of Technology (New Series)*, Vol. 21 No. 5, pp. 10-14.

## Further reading

Hedges, K. (2017), "Closing the gap between big data and the design and construction industries with Building Information Modeling (BIM) schema", *Journal of Architectural Engineering Technology*, Vol. 1 No. 1, pp. 11-18.

## Corresponding author

Upeksha Hansini Madanayake can be contacted at: madanayu@lsbu.ac.uk

# A systematic review of the applications of multi-criteria decision-making methods in site selection problems

Jeremy Yee Li Yap, Chiung Chiung Ho and Choo-Yee Ting
*Faculty of Computing and Informatics,*
*Multimedia University, Cyberjaya, Malaysia*

Received 19 June 2018
Revised 25 November 2018
8 May 2019
13 June 2019
Accepted 14 June 2019

## Abstract

**Purpose** – The purpose of this paper is to perform a systematic review on the application of different multi-criteria decision-making (MCDM) methods in solving the site selection problem across multiple problem domains. The domains are energy generation, logistics, public services and retail facilities. This study aims to answer the following research questions: Which evaluating criteria were used for each site selection problem domain? Which MCDM methods were frequently applied in a particular site selection problem domain?

**Design/methodology/approach** – The goals of the systematic review were to identify the evaluating criteria as well as the MCDM method used for each problem domain. A total of 81 recent papers (2014–2018) including 32 papers published in conference proceedings and 49 journal articles from various databases including IEEE Xplore, PubMed, Springer, Taylor and Francis as well as ScienceDirect were evaluated.

**Findings** – This study has shown that site selection for energy generation facilities is the most active site selection problem domain, and that the analytic hierarchy process (AHP) method is the most commonly used MCDM method for site selection. For energy generation, the criteria which were most used were geographical elements, land use, cost and environmental impact. For logistics, frequently used criteria were geographical elements and distance, while for public services population density, supply and demand, geographical layout and cost were the criteria most used. Criteria useful for retail facilities were the size (space) of the store, demographics of the site, the site characteristics and rental of the site (cost).

**Research limitations/implications** – This study is limited to reviewing papers which were published in the years 2014–2018 only, and only covers the domains of energy generation, logistics, public services and retail facilities.

**Practical implications** – MCDM is a viable tool to be used for solving the site selection problem across the domains of energy generation, logistics, public services and retail facilities. The usage of MCDM continues to be relevant as a complement to machine learning, even as data originating from embedded IoT devices in built environments becomes increasingly Big Data like.

**Originality/value** – Previous systematic review studies for MDCM and built environments have either focused on studying the MCDM techniques itself, or have focused on the application of MCDM for site selection in a single problem domain. In this study, a critical review of MCDM techniques used for site selection as well as the critical criteria used during the MCDM process of site selection was performed on four different built environment domains.

**Keywords** Logistics, TOPSIS, Analytical hierarchy process, Retail, Public service, ELECTRE, Site selection, Energy generation, Multi-criteria decision-making, PROMTHEE

**Paper type** Literature review

## 1. Introduction

Multi-criteria decision-making (MCDM) has been used for selecting the most preferred alternative from a pool of alternatives, especially when the evaluation measures (criteria) are numerous and are often conflicting with each other. As such, MCDM is a good choice as a tool for solving site selection problems (Hsieh *et al.*, 2004). An emerging challenge for site selection is the increasing availability of Big Data and Internet-of-Things (IoT) data which has led to

The authors thank the sponsors Telekom Malaysia Research & Development for funding the project (TM RnD MMUE/160021) which has enabled this paper to be written.

Built Environment Project and Asset Management
Vol. 9 No. 4, 2019
pp. 548-563
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-05-2018-0078

increasingly complex MCDM models which incurs additional computational resources to solve (Rikalovic *et al.*, 2014). Ting *et al.* (2018) have studied the usage of machine learning (ML) and Big Data for performing site selection and have concluded that feature selection is of utmost importance for ensuring efficient results. Feature selection is the act of only choosing the most significant factors that contribute to the accuracy of a predictive ML model in order to minimize processing time. In that study, site-related criteria were sampled to construct a prediction model for selecting the most preferred site for a payment point, as processing the full set of criteria required too much time. Although the volume of Big Data is advantageous for analytics, it needs to be reduced to its most important components in order to be useful for site selection. In this scenario, MCDM can be used as an alternative to ML feature selection (Peng *et al.*, 2012) as the human experts can be relied to choose the most important factors.

Site selection is the process of determining the most favorable location for a building or facility, which balances the needs of the building or facility against the advantages of multiple candidate locations This process is impacted by many conflicting factors or criteria (Hoover, 1948) and originated from the public sector's problem on deciding the site of their strategic assets (Marianov and Serra, 2004). In the beginning, the earliest application of private sector site selection was focused on retail facilities(Ghosh and Craig, 1983; Kohsaka, 1989), however recent applications of site selection have focused on the domain of energy generation and logistics.

MCDM methods used for solving site selection problems include the following: analytic hierarchy process (AHP), fuzzy AHP, analytic network process (ANP), Elimination Et Choix Traduisant la REalité (ELECTRE), Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) and Preference Ranking Organization Method for Enrichment Evaluations (PROMTHEE). The AHP proposed by Saaty (1977) is a pairwise comparison-based MCDM methodology. AHP was further improved by fuzzy AHP, which incorporated the fuzzy set theory that "allows solving a lot of problems related to dealing with imprecise and uncertain data" (Balmat *et al.*, 2011) as a solution for uncertainty. ANP is another MCDM method proposed by Saaty (1996) as a complement to AHP. ELECTRE was originally proposed by Roy (1968) as a MCDM to find a set of preferred alternatives based on two indices called the concordance index and the discordance index. TOPSIS was proposed by Hwang and Yoon (1981) and later extended by Hwang *et al.* (1993) and Lai *et al.* (1994). TOPSIS utilizes distance from an ideal solution as a means of ranking alternatives. PROMTHEE was developed by Brans *et al.* (1986) and by Brans and Vincke (1985) as a means of optimizing the selection of alternatives in a MCDM setting. Each MCDM technique will be further elaborated and evaluated in the context of site selection in a subsequent section.

### *1.1 Review aims*

Prior systematic review studies on MCDM methods and site selection have focused on either a systematic review of MDCM methods or the application of one or more MCDM methods in solving a site selection problem in a single domain. Literature review of MCDM techniques and their generic applications has been performed by Toloie-Eshlaghy and Homayonfar (2011) as well as Velasquez and Hester (2013) and Zavadskas *et al.* (2014). Application of MCDM methods on specific site selection problem domains are widely reported and will be discussed in Section 3. To date, there is no systematic review of the application of MCDM across multiple site selection problem domains, nor are there attempts to group criteria for MCDM applications by common site selection problem domains. Thus, this systematic review study aims to answer the following research questions:

*RQ1.* Which evaluating criteria were used for each site selection problem domain?

*RQ2.* Which MCDM methods were frequently applied in a particular site selection problem domain?

In this study, the application of site selection in the emerging domains of energy generation (Kumar *et al.*, 2017; Şengül *et al.*, 2015) and logistics (Aguezzoul, 2014; Büyüközkan and Göçer, 2017), as well as more traditional domains such as public services (Aydinoglu *et al.*, 2015; Zhao and Li, 2016) and retail facilities (Chang, 2014; Chang and Yang, 2015) have been critically reviewed.

In Section 2, the research methodology used in the review is explained. Section 3 shows the analysis and reporting of the papers reviewed followed by a discussion in Section 4. Lastly, the conclusion and future directions are discussed in Section 5.

## 2. Review methodology

This systematic review was performed from two different perspectives. The first perspective is site selection based on the problem domains. The 81 papers reviewed were from the problem domains of energy generation, logistic, public service and retail facilities from the years 2014 to 2018. A five-year period was chosen as we wanted to review the current state of the art for site selection. The same papers were also reviewed from the second perspective that of the MCDM methods used to aid the decision makers in performing site selection.

A systematic review was used as the research methodology of this study. The definition of a systematic review was given by Denyer *et al.* (2008). It states that a systematic review is "a specific methodology that locates existing studies, selects and evaluates contributions, analyses and synthesizes data, and reports the evidence in such a way that allows reasonably clear conclusions to be reached about what is and is not known" (Denyer *et al.*, 2008). A systematic review approach was chosen over a narrative review approach as the former approach reduces shortcuts and biases, as well as being more evidence-based as compared to the latter approach (Pae, 2015; Rother, 2007). The combination of specific research questions, comprehensive sources with explicit search approach, criterion-based selection and critical evaluation of the reviewed material all contributed toward the choice of conducting this study using a systematic review approach as opposed to a narrative approach.

Step 1: formulating the question: the focus of the review was established by asking questions which will determine which paper is to be reviewed. The CIMO logic that was proposed by Denyer *et al.* (2008) was used to achieve this purpose. CIMO stands for context, intervention, mechanisms and outcome. The CIMO logic is synthesized as: in this class of problematic contexts, use this intervention type to invoke these generative mechanism(s), to deliver these outcome(s) (Denyer *et al.*, 2008). CIMO as a question-structure is well established as a suitable question-structure for synthesis, as opposed to the 3WH question-structure of a narrative review (Booth, 2018).

The question formulated to identify the four main elements of the CIMO logic is as follows:

> The decision-making process to selecting a particular site for a project by any organization is one that involves multiple criteria and multiple alternatives (C), where the alternatives should be evaluated using a single or a combination of multi-criteria decision-making methods (I) based on the opinions of the decision makers to identify the value of each alternative (M) to obtain the best alternative in the selection process (O).

Step 2: locating the papers: the papers to be reviewed were located, selected and appraised. The papers were obtained from scientific websites and databases including Elsevier, Springer, ScienceDirect and IEEE Xplore. The search for papers was narrowed down using keywords search and by selecting papers which were published in the years 2014–2018. Search strings, simple operators and boolean logic were used during the search in order to limit the search to related papers. The search code were formulated as such: ("site select*" OR select* OR site) AND (MCDM OR MCDA OR "multi* criteria" OR MADM OR *AHP OR ANP OR *TOPSIS OR ELECTRE OR PROMTHEE), and were conducted on January 7, 2018.

Step 3: selecting the papers to be reviewed: Step 2 resulted in a total of 152 papers. In this step, a filtering process was performed to focus on the relevant problem domains. The abstract of all the papers were read and papers that were reporting on problem domains other than energy generation, logistic, public service and retail were removed. This process resulted in 81 papers selected for the systematic analysis process.

Step 4: analysis: in this step, the 81 papers from Step 3 were systematically reviewed. The MCDM methods used for site selection, as well as the criteria considered in each MCDM methods were extracted from the papers.

Step 5: reporting: the findings of the systematic review are presented as a formatted report upon the conclusion of the systematic review process.

Step 1 until Step 3 have been discussed in this section. The following section will focus on Step 4 and Step 5 of the methodology by presenting the findings and results of the systematic review.

## 3. Findings

The literature reviewed was categorized based on two different perspectives: problem domains and MDCM methods used. Table I shows a sample of the paper's categorization based on these two aspects. The complete list of reviewed papers can be found at http://bit.ly/SiteSelectionReviewedPapers.

### *3.1 Problem domain categorization*

*3.1.1 Energy generation facilities*. Energy generation is defined as the process of obtaining or harnessing energy from natural resources such as biomass, waves, wind, etc., and so forth. In Table I, it can be seen that 32 out of the 81 reviewed papers were studies on the problem of selecting a site to set up an energy generation facility. As the majority of reviewed papers lie in this problem domain, we can conclude that the study of site selection for renewable energy is a growing research area.

The papers reviewed can be categorized into a few categories, each with its own set of criteria. The categories are wind, solar, tidal, hydro, offshore and biofuel. For wind-power generation site selection, criteria were wind speed, rainfall, slope, altitude, land use, land cover, location, environmental parameters, agrological capacity, distance, environmental, economic, social, highways, railways, built-up, forest zone, scenic zone, access to the grid, land costs, reduce noise blade, geographic standards and infrastructure. For solar-energy generation site selection, the criteria studied were cost, biological environment, physical environment, economic development, agrological capacity, slope, area, field orientation, distance, potential solar radiation, temperature, climatic, location, geomorphological, solar radiation, land availability, water availability, cost of land, population benefitted, transmission losses, number of rainy days, proximity to power grid and transportation. For tidal-power generation site selection, the criteria which have been investigated were magnitude of current speed, tidal current direction, mean turbulent kinetic energy, energy flux, wave breaking, shipping traffic, energy resource, site characteristics and environmental suitability. For hydro-power generation site selection, criteria studied included terrain, climate, location and risk. For offshore power generation site selection, studied criteria included wind speed, wave power density, depth range, minimum distance to shore, exclusion areas, energy resources and profitability, conservation areas, view protection, human activities and power grid access. The final category, biofuel, used the

| | Energy generation | Logistics | Public service | Retail facilities |
|---|---|---|---|---|
| AHP/fuzzy AHP | Wiguna *et al.* (2016), Garni and Awasthi (2017), Al-Shabeeb *et al.* (2016), Guptha *et al.* (2015), Multazam *et al.* (2016), Rezaian and Jozi (2016), Sánchez-Lozano *et al.* (2016), Ubando *et al.* (2015), Vasileiou *et al.* (2017) and Zoghi *et al.* (2017) | Abedi-Varaki and Davtalab (2016), Andarani and Budiawan (2015), Bahrani *et al.* (2016), Çetinkaya *et al.* (2016), Chabuk *et al.* (2017), Chauhan and Singh (2016), Dai (2016), Djokanović *et al.* (2016), Pramanik (2016), Rahmat *et al.* (2017), Trivedi and Singh (2017) and Vučijak *et al.* (2016) | Chaudhary *et al.* (2016), Li *et al.* (2017), Liu *et al.* (2017), Matteo *et al.* (2016) and Triantono and Susetyarto (2017) | |
| ANP | Gigović *et al.* (2017) | Morteza *et al.* (2016) | | |
| TOPSIS/ fuzzy TOPSIS | Sánchez-Lozano *et al.* (2016) and Villacreses *et al.* (2017) | Aghajani Mir *et al.* (2016), Arıkan *et al.* (2017), Çetinkaya *et al.* (2016), Chauhan and Singh (2016), Mangalan *et al.* (2016) and Morteza *et al.* (2016) | Liu *et al.* (2017) | Chang (2014) |
| ELECTREE | Wu *et al.* (2016) | | | |
| PROMTHEE | Wiguna *et al.* (2016) | Arıkan *et al.* (2017) | | |
| Others | Aredes *et al.* (2017), Abaei *et al.* (2017), Birjandi *et al.* (2015), Cradden *et al.* (2016), El-Azab and Amin (2015), Ghosh *et al.* (2016), Jangid *et al.* (2016), Kim *et al.* (2016), Lee *et al.* (2017), Martinkus *et al.* (2017), Noorollahi *et al.* (2016), Shaheen and Khan (2016), Shimray *et al.* (2017), Thongpun *et al.* (2017), Wang, *et al.* (2016), Wu *et al.* (2017) and Xu *et al.* (2016) | Chen *et al.* (2017), Fraile *et al.* (2016), Krylovas *et al.* (2016), Kumar and Bansal (2016), Liu *et al.* (2016), Liu *et al.* (2017), Sultana and Rasel (2016), Temur (2016), Wang *et al.* (2017), Wechtaisong *et al.* (2014), Wibowo *et al.* (2014), Wu and Xie (2016), Yongfei *et al.* (2017) | Aydinoglu *et al.* (2015), Chen *et al.* (2017), Min *et al.* (2015), Song *et al.* (2015), Wang *et al.* (2016), Yao and Cheng (2017), Zhang *et al.* (2017) and Zhao and Li (2016) | Chang and Yang (2015) and Chen and Tsai (2016) |

**Table I.**
Problem domain/methodologies categorization of the reviewed papers

following criteria for site selection: natural capital, built capital, human capital, available natural resources, social aspect, installed plants, and fuel demand per region.

Five papers employed hybrid MCDM methods to solve the energy generation site selection problem. For example, a hybrid method combining interpretive structural modeling (ISM), fuzzy analytic network process and VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) was used to select the best site for a photovoltaic solar plant (Lee *et al.*, 2017). Eight papers incorporated the geographical information system (GIS) into their decision support systems for site selection (Algarni and Awasthi, 2017; Al-Shabeeb *et al.*, 2016; Gigović *et al.*, 2017; Guptha *et al.*, 2015; Noorollahi *et al.*, 2016; Sánchez-Lozano *et al.*, 2016; Vasileiou *et al.*, 2017; Villacreses *et al.*, 2017).

The usage of GIS is highly recommended for energy generation site selection as GIS allows for the gathering, managing and analyzing of geographical data efficiently and effectively. In the domain of energy generation site selection, the geographical elements, land use, cost and environmental impact are the important criteria to be considered.

*3.1.2 Logistic facilities*. Logistic facilities are facilities that house the detailed organization and implantation of complex operations. These operations involve the flow of things from one point to another and were reported in 32 of the reviewed papers.

11 papers reviewed reported on the problem of selecting a site for a waste management facility. The usage of GIS is very important in this particular problem domain. For example, GIS and AHP were used for a landfill site selection in Behbahan, Iran (Rahmat *et al.*, 2017). The research concluded that groundwater and surface waters were the most important criteria when selecting the site for a landfill facility, whereas the least important criterion is the slope of the land. Other types of logistic site selection studied focused on the placement of warehouses (Chen *et al.*, 2017; Mangalan *et al.*, 2016), railway data centers (Liu *et al.*, 2016) and transport facilities in urban environments (Fraile *et al.*, 2016).

The number of past papers that incorporated the usage of GIS when researching on the problem of site selection for logistic facilities is high as well, with six of the reviewed papers in this category. GIS is highly effective for selection of sites for logistic facilities as this problem domain focuses on complex operations which are heavily impacted by the geographical elements and distance of the logistic site.

*3.1.3 Public service facilities*. Public service facilities refer to facilities that are used to provide for the public needs. The 14 papers reviewed showed that the site selection problem for the public services domain exists, and that the number and types of services provided as a public service has evolved over time. Public services were grouped into four groups of services – core public needs (payment points, culture and education, town planning), emergency needs (emergency shelters and emergency command centers), transportation needs (ports, trains, parking, ride sharing) and sensory needs (meteorology and sensor sites). Each of these groups of services has their own set of criteria. For core public services, criteria of importance include consumer passion, fairness and benefits. Emergency needs emphasizes on criteria such as distance, land cover, population density, costs, reachability, communications, impact and dangerousness. The criteria for transportation needs are the potential users, potential travel demand, potential travel purposes, distances, demographic variables, accessibility, attractiveness of the site, potential for development, competitiveness, economy, society, environment and technology. The final group of public services, sensory needs, will have criteria such as accuracy, cost, terrain conditions and coverage density.

Selecting the site of a public service facility often required the use of various combination of MCDM techniques. Seven of the papers reviewed employed various combination of MCDM to solve this problem. For example, a study looked into the site selection for public bikes rental stations in Taipei using spatial-temporal analysis and retail location theory (Wang *et al.*, 2016). Another paper focused on determining the site for a multi-function terminal in Merak Port, Indonesia using AHP and Delphi (Triantono and Susetyarto, 2017). Other papers focused on selecting the site for electric vehicles charging stations (Zhao and Li, 2016), parking places (Aydinoglu *et al.*, 2015) and car sharing stations (Li *et al.*, 2017).

For public service facilities site selection, criteria such as population density, supply and demand, geographical layout and cost are important. Cost is an important criterion as public services are mostly funded by public funds. Criteria affecting the selection of the site for public service facilities will change as new buildings or new developments emerge. Therefore, future studies should focus on newer technologies or methods which are not affected by physical development changes within and near the site itself.

*3.1.4 Retail facilities*. Retail facilities are sites which provide services for consumer goods and are focused on the sale of products. Shopping complexes, small grocery centers and restaurants are part of this category. This domain has only three papers out of the pool of 81 papers reviewed which indicates the need for more research. Two of the three papers focused on retail chain location (Chang, 2014; Chang and Yang, 2015) and the last paper focused on the site selection for a restaurant chain (Chen and Tsai, 2016).

Three different methodologies were used in the three different papers that were reviewed regarding retail facilities. The simplest method to be used was the TOPSIS method. The TOPSIS method deals with multiple criteria that are central in the decision-making process. The criteria that were proposed to be included in the model were crowds, store cluster, site features, store spaces and rent costs (Chang, 2014). In another paper, a data mining framework was designed to determine the most significant location factors that affect store performance (Chen and Tsai, 2016). The results were store size, availability of parking area, store visibility and population growth rate of the vicinity area. An optimization method was used to determine an ideal site selection for a retail chain operation in China (Chang and Yang, 2015) which indicated that the criteria of importance were stream of people, group of stores, location characteristics, store space and rent proportional to sales.

As a conclusion, the main criteria used in determining the site selection for a retail store would be the size (space) of the store, demographics of the site, the site characteristics and rental cost of the site.

### *3.2 MCDM methods categorization*

*3.2.1 AHP/fuzzy AHP/ANP.* The AHP proposed by Saaty (1977) is a pairwise comparison-based MCDM methodology. The decision problem is structured as a hierarchy which consists of several levels. The goals of the decision problem are represented as the first level. The second level represents the main decision criteria, followed by their sub-criteria in the subsequent levels. The final level represents the different alternatives to the decision problem. The elements of each level are compared in a pairwise fashion which results in a pairwise comparison matrix. Different methods are used to calculate the weights of the elements; the principal eigenvector technique (Saaty, 1977), the weighted least square method, the logarithmic least square method, geometric mean method (Crawford and Williams, 1985) and goal programming method (Bryson, 1995; Lin, 2006). In order to identify the best alternative to the decision problem, the weights from each level obtained using either of the methods are aggregated.

AHP is one of the most popular MCDM methods due to its many advantages. One of the advantages is that this method is easy to use. The method uses pairwise comparisons, allowing decision makers to weigh the criteria and compare alternatives easily. The method is also scalable, allowing for the adjustment of size to accommodate the decision problem. This is due to the hierarchical structure of the method. However, AHP faces the disadvantage of uncertainty and inconsistency in judgment and ranking criteria. It also “does not allow [individuals] to grade one instrument in isolation, but in comparison with the rest, without identifying weakness and strengths” (Konidari and Mavrakis, 2007). Another flaw of the general form of AHP is that it is possible for rank reversal to occur. During rank reversal, the order of the initial alternatives identified by the AHP model is reversed as more alternatives are considered. Since AHP uses pairwise comparisons to rank the alternatives, any addition of alternatives to the decision problem could cause a rank reversal. When AHP is used for performing site selection, the addition of more site alternatives may cause a rank reversal.

Fuzzy AHP incorporates the fuzzy set theory that “allows solving a lot of problems related to dealing with imprecise and uncertain data” (Balmat *et al.*, 2011). This allows it to handle imprecise input and overcome the disadvantage of AHP not being able to handle uncertainty and inconsistency. While the conventional AHP uses a numeric scale (1–9) to compare the elements in the hierarchy, fuzzy AHP uses linguistic variables (e.g. “slightly more important”), and their corresponding fuzzy numbers to compare the elements. However, a fuzzy AHP model is not easy to develop. It requires many simulations before being able to be deployed in the real world. There are many variants of fuzzy AHP (Chang, 1996; Csutora and Buckley, 2001; van Laarhoven and Pedrycz, 1983; Mikhailov, 2000).

ANP is another MCDM method proposed by Saaty (1996). The purpose of the development of ANP was to address the interdependency and feedback problems between the criteria in the hierarchy of the decision problem. ANP can be considered as one form of AHP with the added concern for network structure. Despite that, its major disadvantage, apart from those of AHP, is that "it ignores the different effects among clusters" (Wang, 2012).

From a total of 81 papers reviewed, 28 papers have applied AHP or fuzzy AHP, and two papers have used ANP.

*3.2.2 ELECTRE.* ELECTRE was originally proposed by Roy (1968) as an MCDM technique to find preferred alternatives (the kernel set) based on two indices called the concordance index and the discordance index. These two indices can be used to compare the merits of two alternatives. This relation will determine which alternative is better than the other in terms of ranking. The original model proposed by Roy, ELECTRE I, cannot be used for ranking the alternatives (Roy, 1968). Instead, it was used to obtain the kernel set. Other variants of ELECTRE have been used to improve the initial model. For instance, ELECTRE II (Roy and Bertier, 1971) was proposed to address the problem of inefficiency of ELECTRE I in ranking. ELECTRE III (Roy, 1978) extends the crisp outranking relations to fuzzy outranking relations, and ELECTRE IV (Roy and Hugonnard, 1982) was an attempt to simplify ELECTRE III.

The major advantage of ELECTRE is that it considers uncertainty and vagueness, a quality lacking in many other MCDM methods. However, the processes and results can be difficult to be understood by non-experts. Apart from that, the alternatives that have the lowest performance according to certain criteria will not be displayed. The nature of ELECTRE being an outranking method does not directly identify the strength and weaknesses of the alternatives, nor verify the results and impacts of the results (Konidari and Mavrakis, 2007).

ELECTRE III was used as the MCDM method in two of the reviewed papers, one of which was to study the decision framework for offshore wind power station site selection (Wu *et al.*, 2016). The proposed methodology in that work was to combat three problems faced by canonical MCDM methods, i.e. the compensation problems in processing criteria information, the loss of decision information and the interaction problem in a fuzzy environment.

*3.2.3 TOPSIS.* TOPSIS was proposed by Hwang and Yoon (1981) and later extended by Hwang *et al.* (1993) and Lai *et al.* (1994). In the original TOPSIS method, the best alternative is the one which has "the shortest distance from the positive ideal solution (PIS) and the farthest from the negative ideal solution (NIL)" (Lai *et al.*, 1994).

TOPSIS is a relatively easy method to use. It is scalable as the steps of the method remain the same regardless of any number of criteria or alternatives (Iç, 2012). However, the distance measure used in its process, the Euclidean distance, does not take into account the correlation between the elements of the model (criteria and alternatives). It is hard to weight the elements and keep the consistency of the decision makers' judgments, especially when new elements are added into the model.

Among the 81 papers reviewed, 11 papers have applied TOPSIS in the research. Among these, the TOPSIS method was used as a methodology for a chain store location selection (Chang, 2014). The criteria used in this particular paper were crowds, store cluster, site features, store spaces and rent costs. The store chosen by the resultant model proved to be the best when the store's performance reached the annual target within a period of two years.

*3.2.4 PROMTHEE.* PROMTHEE was developed by Brans *et al.* (1986) and Brans and Vincke (1985). PROMTHEE works in a way that does not offer the right alternative but rather helps decision makers obtain the most optimal alternative that best matches their objectives and understanding of the problem. It provides a comprehensive and rational framework for structuring a decision problem, identifying and quantifying its conflicts and synergies, clusters of actions, and highlights the main alternatives and the structured reasoning behind the alternative.

This MCDM method is easy to use and it is not necessary to assume that the criteria of the decision problem are proportionate. However, this method does not provide a clear way to assign the weights and values to the criteria and alternatives. PROMTHEE was the chosen MCDM method for site selection in two of the papers reviewed.

A paper researching on the solid waste disposal methodology selection using MCDM methods (Arıkan *et al.*, 2017) explores the PROMTHEE method in comparison to TOPSIS and fuzzy TOPSIS. The PROMTHEE method was able to rank the solid waste disposal methodologies almost similarly to the other two methodologies. The paper however does not critically analyze the performance of the methods being compared.

*3.2.5 Other/hybrid methodologies*. A combination of multiple MCDM methods or novel methodologies in the research was used in 43 of the reviewed papers for the purpose of solving a site selection problem. Most of the papers reviewed incorporated the usage of GIS in their hybrid methodologies to solve the problem of site selection. The system was incorporated with AHP (Djokanović *et al.*, 2016; Rahmat *et al.*, 2017) and fuzzy AHP (Algarni and Awasthi, 2017; Çetinkaya *et al.*, 2016) to name a few. The usage of hybrid methodologies is highly suggested as these combinations serve to overcome the disadvantages of each of the method.

For example, the combination of TOPSIS and VIKOR was researched to develop an optimized municipal solid waste management model (Aghajani Mir *et al.*, 2016). Since the original TOPSIS was unable to "convert negative points to positive points in comparison, extra measurements in the residual parts of the algorithm were used" (Aghajani Mir *et al.*, 2016). The combination of the two methods by sensitivity analysis leads to an improved method of ranking the alternatives.

Another hybrid methodology was developed to select a sustainable location of healthcare waste disposal facility (Chauhan and Singh, 2016). This methodology incorporated ISM, fuzzy AHP and fuzzy TOPSIS. ISM was used to eliminate dependent variables, which reduced the time and complexity of the decision-making process. Fuzzy AHP was then subsequently used to calculate the weights of each criterion. Finally, the ranking of the alternatives was then derived from fuzzy TOPSIS to complete the selection methodology.

There are other novel methodologies developed as well such as the site selection based on Voronoi Diagram for electric vehicle charging stations (Song *et al.*, 2015) and the cloud-based design optimization for the decision-making process involving sites selection under high uncertainty (Temur, 2016).

## 4. Discussion

After the systematic review has been conducted, it was found that the reasons for selecting a particular MCDM technique for solving a site selection problem has been described briefly in almost all the reviewed literature. The criteria that were of importance for each MCDM technique were also identified clearly. The research gap that is identified is the relevance of MDCM techniques for solving site selection problems given the increased application of ML techniques using Big Data.

Of the four problem domains, energy generation and logistic facilities were most studied, followed by public service and retail facilities. In terms of research trends, the problem of selecting energy generation facilities shows the biggest growth, while the research interest in deciding on the location of retail facilities has shown minimal growth.

AHP continues to be used frequently as a MCDM method for solving site selection problems, although it is rarely used as a single method. The usage of multiple combinations of MCDM methods seems to be the approach that is most promising, given that many reviewed papers reported on the usage of combined MCDM methods.

The usage of GIS as an enabler for a MCDM approach toward site selection can be observed to be particularly useful for the problem domains of energy generation and logistic facility site selection. This indicates that for these two areas, geographical constraints continue to play a major role in these problem domains.

MCDM methods are able to tap on Big Data to draw on increasingly higher number of criteria, due to advances in sensor technology such as IoT sensors. IoT sensors are internet-connected devices which can acquire information from a built environment. This is reported by Huang and Wey (2019), who described how Big Data acquired in Taipei was used to develop land reuse strategies through the usage of ANP models. The increasing number of criteria due to Big Data collected via IoT devices will result in complex MCDM models which needs greater computing resources and development of more time efficient MCDM models.

## 5. Conclusion and future directions

In this paper, a systematic review of the application of MCDM methods for site selection has been performed. In total, 81 papers from journals and conferences were systematically reviewed. The systematic review was conducted to identify important criteria in each problem domain, as well as to investigate the usage of MCDM for the purpose of site selection. For site selection problem domains, energy generation is the most studied problem domain followed by logistics, public services and retail facilities. AHP and TOPSIS were among the most commonly used MCDM methods for site selection, although hybrid methods are becoming increasingly popular. In the near future, built environments will continuously capture data in various formats via the usage of embedded IoT sensors. Therefore, new MCDM methods which can accommodate dynamic criteria for continuous optimal site selection need to be developed in the future.

## References

Abaei, M.M., Arzaghi, E., Abbassi, R., Garaniya, V. and Penesis, I. (2017), "Developing a novel risk-based methodology for multi-criteria decision making in marine renewable energy applications", *Renewable Energy*, Vol. 102, Part B, pp. 341-348.

Abedi-Varaki, M. and Davtalab, M. (2016), "Site selection for installing plasma incinerator reactor using the GIS in Rudsar county, Iran", *Environmental Monitoring and Assessment*, Vol. 188 No. 6, p. 353.

Aghajani Mir, M., Taherei Ghazvinei, P., Sulaiman, N.M.N., Basri, N.E.A., Saheri, S., Mahmood, N.Z., Jahan, A., Begum, R.A., Aghamohammadi Yongfei, N. *et al.* (2016), "Application of TOPSIS and VIKOR improved versions in a multi criteria decision analysis to develop an optimized municipal solid waste management model", *Journal of Environmental Management*, Vol. 166, pp. 109-115.

Aguezzoul, A. (2014), "Third-party logistics selection problem: a literature review on criteria and methods", *Omega*, Vol. 49, pp. 69-78.

Algarni, H. and Awasthi, A. (2017), "A fuzzy AHP and GIS-based approach to prioritize utility-scale solar PV sites in Saudi Arabia", available at: https://doi.org/10.1109/SMC.2017.8122783 (accessed November 25, 2018).

Al-Shabeeb, A.R., Al-Adamat, R. and Mashagbah, A. (2016), "AHP with GIS for a preliminary site selection of wind turbines in the North West of Jordan", *International Journal of Geosciences*, Vol. 07 No. 10, pp. 1208-1221.

Andarani, P. and Budiawan, W. (2015), "Multicriteria decision analysis for optimizing site selection of electronic and electricity equipment waste dismantling and sorting facility (Case study: In Indonesia, using AHP)", *2015 International Conference on Science in Information Technology (ICSITech)*, pp. 264-269.

Aredes, M.A., Oliveira, D.S., de, Aredes, M., Guijun, L., Jinjun, L. and Bo, W. (2017), "A Brazilian PMU-WAMS pilot project: a methodology for PMU site selection", *IECON 2017 – 43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 5-10.

Arıkan, E., Şimşit-Kalender, Z.T. and Vayvay, Ö. (2017), "Solid waste disposal methodology selection using multi-criteria decision making methods and an application in Turkey", *Journal of Cleaner Production*, Vol. 142, pp. 403-412.

Aydinoglu, A.C., Senbil, M., Saglam, D. and Demir, S. (2015), "Planning of parking places on transportation infrastructure by geographic information techniques", *3rd International Istanbul Smart Grid Congress and Fair (ICSG)*, pp. 1-5.

Bahrani, S., Ebadi, T., Ehsani, H., Yousefi, H. and Maknoon, R. (2016), "Modeling landfill site selection by multi-criteria decision making and fuzzy functions in GIS, case study: Shabestar, Iran", *Environmental Earth Sciences*, Vol. 75 No. 4, p. 337.

Balmat, J.F., Lafont, F., Maifret, R. and Pessel, N. (2011), "A decision-making system to maritime risk assessment", *Ocean Engineering*, Vol. 38 No. 1, pp. 171-176.

Birjandi, A.H., d'Auteuil, S., Ridd, C. and Bibeau, E.L. (2015), "An innovative low cost hydrokinetic site selection technique for cold climate regions", *OCEANS 2015 – Genova*, pp. 1-4.

Booth, A. (2018), "Alternative question structures for different types of systematic review", 26 January, available at: www.networks.nhs.uk/nhs-networks/nwas-library-and-information-service/documents/alternative-question-structures-for-different-types-of-systematic-review (accessed November 23, 2018).

Brans, J.P. and Vincke, P. (1985), "Note – a preference ranking organisation method: (the PROMETHEE method for multiple criteria decision-making)", *Management Science*, Vol. 31 No. 6, pp. 647-656.

Brans, J.P., Vincke, P. and Mareschal, B. (1986), "How to select and how to rank projects: the PROMETHEE method", *European Journal of Operational Research*, Vol. 24 No. 2, pp. 228-238.

Bryson, N. (1995), "A goal programming method for generating priority vectors", *The Journal of the Operational Research Society*, Vol. 46 No. 5, pp. 641-648.

Büyüközkan, G. and Göçer, F. (2017), "Application of a new combined intuitionistic fuzzy MCDM approach based on axiomatic design methodology for the supplier selection problem", *Applied Soft Computing*, Vol. 52, pp. 1222-1238.

Çetinkaya, C., Özceylan, E., Erbaş, M. and Kabak, M. (2016), "GIS-based fuzzy MCDA approach for siting refugee camp: a case study for southeastern Turkey", *International Journal of Disaster Risk Reduction*, Vol. 18, pp. 218-231.

Chabuk, A., Al-Ansari, N., Hussain, H., Knutsson, S., Pusch, R. and Laue, J. (2017), "Landfills site selection in Babylon, Iraq", *Earth Sciences and Geotechnical Engineering*, Vol. 7 No. 4, pp. 1-15.

Chang, D.-Y. (1996), "Applications of the extent analysis method on fuzzy AHP", *European Journal of Operational Research*, Vol. 95 No. 3, pp. 649-655.

Chang, H.J. (2014), "A TOPSIS model for chain store location selection", *Review of Intgrative Business and Economics Research*, Vol. 4 No. 1, pp. 410-416.

Chang, H.J., Hsieh, C.M. and Yang, F.M. (2015), "Acquiring an optimal retail chain location in China", *Proceedings – 2015 2nd International Conference on Information Science and Control Engineering, ICISCE 2015*, pp. 96-99.

Chaudhary, P., Chhetri, S.K., Joshi, K.M., Shrestha, B.M. and Kayastha, P. (2016), "Application of an analytic hierarchy process (AHP) in the GIS interface for suitable fire site selection: a case study from Kathmandu Metropolitan City, Nepal", *Socio-Economic Planning Sciences*, Vol. 53, pp. 60-71.

Chauhan, A. and Singh, A. (2016), "A hybrid multi-criteria decision making method approach for selecting a sustainable location of healthcare waste disposal facility", *Journal of Cleaner Production*, Vol. 139, pp. 1001-1010.

Chen, C., Liu, J., Li, Q., Wang, Y., Xiong, H. and Wu, S. (2017), "Warehouse site selection for online retailers in inter-connected warehouse networks", *Proceedings – IEEE International Conference on Data Mining, ICDM, November*, pp. 805-810.

Chen, L.-F. and Tsai, C.-T. (2016), "Data mining framework based on rough set theory to improve location selection decisions: a case study of a restaurant chain", *Tourism Management*, Vol. 53, pp. 197-206.

Chen, M., Liu, J., Li, Z., Ma, W. and Sun, Z. (2017), "Research on site selection of rescue sites at sea based on NSGA II", *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 460-465.

Cradden, L., Kalogeri, C., Barrios, I.M., Galanis, G., Ingram, D. and Kallos, G. (2016), "Multi-criteria site selection for offshore renewable energy platforms", *Renewable Energy*, Vol. 87, pp. 791-806.

Crawford, G. and Williams, C. (1985), "A note on the analysis of subjective judgment matrices", *Journal of Mathematical Psychology*, Vol. 29 No. 4, pp. 387-405.

Csutora, R. and Buckley, J.J. (2001), "Fuzzy hierarchical analysis: the Lambda-Max method", *Fuzzy Sets and Systems*, Vol. 120 No. 2, pp. 181-195.

Dai, X. (2016), "Dam site selection using an integrated method of AHP and GIS for decision making support in Bortala, Northwest China", Lund University GEM Thesis Series, Lund University, available at: http://lup.lub.lu.se/student-papers/record/8886448 (accessed November 20, 2018).

Denyer, D., Tranfield, D. and Van Aken, J.E. (2008), "Developing design propositions through research synthesis", *Organization Studies*, Vol. 29 No. 3, pp. 393-413.

Djokanović, S., Abolmasov, B. and Jevremović, D. (2016), "GIS application for landfill site selection: a case study in Pančevo, Serbia", *Bulletin of Engineering Geology and the Environment*, Vol. 75 No. 3, pp. 1273-1299.

El-Azab, R. and Amin, A. (2015), "Optimal solar plant site selection", *Presented at the SoutheastCon 2015*, pp. 1-6.

Fraile, A., Larrodé, E., Alberto Magreñán and Sicilia, J.A. (2016), "Decision model for siting transport and logistic facilities in urban environments: a methodological approach", *Journal of Computational and Applied Mathematics*, Vol. 291, pp. 478-487.

Garni, H.Z.A. and Awasthi, A. (2017), "A fuzzy AHP and GIS-based approach to prioritize utility-scale solar PV sites in Saudi Arabia", *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1244-1249.

Ghosh, A. and Craig, C.S. (1983), "Formulating retail location strategy in a changing environment", *Journal of Marketing*, Vol. 47 No. 3, pp. 56-68.

Ghosh, S., Chakraborty, T., Saha, S., Majumder, M. and Pal, M. (2016), "Development of the location suitability index for wave energy production by ANN and MCDM techniques", *Renewable and Sustainable Energy Reviews*, Vol. 59, pp. 1017-1028.

Gigović, L., Pamučar, D., Božanić, D. and Ljubojević, S. (2017), "Application of the GIS-DANP-MABAC multi-criteria model for selecting the location of wind farms: a case study of Vojvodina, Serbia", *Renewable Energy*, Vol. 103, pp. 501-521.

Guptha, R., Puppala, H. and Kanuganti, S. (2015), "Integrating fuzzy AHP and GIS to prioritize sites for the solar plant installation", *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 465-470.

Hoover, E.M. (1948), *Location of Economic Activity*, Mcgraw-Hill, New York, NY.

Hsieh, T.-Y., Lu, S.-T. and Tzeng, G.-H. (2004), "Fuzzy MCDM approach for planning and design tenders selection in public office buildings", *International Journal of Project Management*, Vol. 22 No. 7, pp. 573-584.

Huang, J.-Y. and Wey, W.-M. (2019), "Application of big data and analytic network process for the adaptive reuse strategies of school land", *Social Indicators Research*, Vol. 142 No. 3, pp. 1075-1102.

Hwang, C.-L. and Yoon, K. (1981), *Multiple Attribute Decision Making: Methods and Applications*, Springer-Verlag, New York, NY.

Hwang, C.-L., Lai, Y.-J. and Liu, T.-Y. (1993), "A new approach for multiple objective decision making", *Computers & Operations Research*, Vol. 20 No. 8, pp. 889-899.

Iç, Y.T. (2012), "An experimental design approach using TOPSIS method for the selection of computer-integrated manufacturing technologies", *Robotics and Computer-Integrated Manufacturing*, Vol. 28 No. 2, pp. 253-254, available at: https://doi.org/10.1016/j.rcim.2011.09.005

Jangid, J., Bera, A.K., Joseph, M., Singh, V., Singh, T.P., Pradhan, B.K. and Das, S. (2016), "Potential zones identification for harvesting wind energy resources in desert region of India – a multi criteria evaluation approach using remote sensing and GIS", *Renewable and Sustainable Energy Reviews*, Vol. 65, pp. 1-10.

Kim, T., Park, J.-I. and Maeng, J. (2016), "Offshore wind farm site selection study around Jeju Island, South Korea", *Renewable Energy*, Vol. 94, pp. 619-628.

Kohsaka, H. (1989), "A spatial search-location model of retail centers", *Geographical Analysis*, Vol. 21 No. 4, pp. 338-349.

Konidari, P. and Mavrakis, D. (2007), "A multi-criteria evaluation method for climate change mitigation policy instruments", *Energy Policy*, Vol. 35 No. 12, pp. 6235-6257.

Krylovas, A., Zavadskas, E.K. and Kosareva, N. (2016), "Multiple criteria decision-making KEMIRA-M method for solution of location alternatives", *Economic Research-Ekonomska Istraživanja*, Vol. 29 No. 1, pp. 50-65.

Kumar, A., Sah, B., Singh, A.R., Deng, Y., He, X., Kumar, P. and Bansal, R.C. (2017), "A review of multi criteria decision making (MCDM) towards sustainable renewable energy development", *Renewable and Sustainable Energy Reviews*, Vol. 69, pp. 596-609.

Kumar, S. and Bansal, V.K. (2016), "A GIS-based methodology for safe site selection of a building in a hilly region", *Frontiers of Architectural Research*, Vol. 5 No. 1, pp. 39-51.

Lai, Y.J., Liu, T.Y. and Hwang, C.L. (1994), "TOPSIS for MODM", *European Journal of Operational Research*, Vol. 76 No. 3, pp. 486-500.

Lee, A.H.I., Kang, H.Y. and Liou, Y.J. (2017), "A hybrid multiple-criteria decision-making approach for photovoltaic solar plant location selection", *Sustainability (Switzerland)*, Vol. 9 No. 2, p. 19, available at: https://doi.org/10.3390/su9020184

Li, W., Li, Y., Fan, J. and Deng, H. (2017), "Siting of carsharing stations based on spatial multi-criteria evaluation: a case study of Shanghai EVCARD", *Sustainability*, Vol. 9 No. 1, pp. 1-16.

Lin, C.C. (2006), "An enhanced goal programming method for generating priority vectors", *Journal of the Operational Research Society*, Vol. 57 No. 12, pp. 1491-1496.

Liu, J., Li, P., Shi, T. and Ma, X. (2016), "Optimal site selection of China railway data centers by the PSO algorithm", *12th World Congress on Intelligent Control and Automation (WCICA)*, pp. 251-257.

Liu, Y., Li, L., Xie, Z., Zhu, G. and Xu, G. (2017), "Urban emergency shelter site selection", *2017 International Conference on Behavioral, Economic, Socio-Cultural Computing (BESC)*, pp. 1-6.

Mangalan, A.V., Kuriakose, S., Mohamed, H. and Ray, A. (2016), "Optimal location of warehouse using weighted MOORA approach", *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, pp. 662-665.

Marianov, V. and Serra, D. (2004), "Location models in the public sector", economics working paper, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, available at: https://econpapers.repec.org/paper/upfupfgen/755.htm (accessed November 22, 2018).

Martinkus, N., Rijkhoff, S.A.M., Hoard, S.A., Shi, W., Smith, P., Gaffney, M. and Wolcott, M. (2017), "Biorefinery site selection using a stepwise biogeophysical and social analysis approach", *Biomass and Bioenergy*, Vol. 97, pp. 139-148.

Matteo, U.D., Pezzimenti, P.M. and Garcia, D.A. (2016), "Methodological proposal for optimal location of emergency operation centers through multi-criteria approach", *Sustainability*, Vol. 8 No. 1, pp. 1-12.

Mikhailov, L. (2000), "A fuzzy programming method for deriving priorities in the analytic hierarchy process", *Journal of the Operational Research Society*, Vol. 51 No. 3, pp. 341-349.

Min, L., Zaigui, Y., Tianlu, Q., Xianzhe, Z. and Qing, D. (2015), "The optimization of Nanjing's public cultural facility location based on genetic algorithm", *23rd International Conference on Geoinformatics*, pp. 1-6.

Morteza, Z., Reza, F.M., Seddiq, M.M., Sharareh, P. and Jamal, G. (2016), "Selection of the optimal tourism site using the ANP and fuzzy TOPSIS in the framework of integrated coastal zone management: a case of Qeshm Island", *Ocean & Coastal Management*, Vol. 130, pp. 179-187.

Multazam, T., Putri, R.I., Pujiantara, M., Priyadi, A. and Hery, P.M. (2016), "Wind farm site selection base on fuzzy analytic hierarchy process method; case study area Nganjuk", 2016 international seminar on intelligent technology and its applications (ISITIA)", *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 545-550.

Noorollahi, Y., Yousefi, H. and Mohammadi, M. (2016), "Multi-criteria decision support system for wind farm site selection using GIS", *Sustainable Energy Technologies and Assessments*, Vol. 13, pp. 38-50.

Pae, C.-U. (2015), "Why systematic review rather than narrative review?", *Psychiatry Investigation*, Vol. 12 No. 3, pp. 417-419.

Peng, Y., Kou, G., Ergu, D., Wu, W. and Shi, Y. (2012), "An integrated feature selection and classification scheme", *Studies in Informatics and Control*, Vol. 21 No. 3, p. 241, available at: https://doi.org/10.24846/v21i3y201202

Pramanik, M.K. (2016), "Site suitability analysis for agricultural land use of Darjeeling district using AHP and GIS techniques", *Modeling Earth Systems and Environment*, Vol. 2 No. 2, p. 56.

Rahmat, Z.G., Niri, M.V., Alavi, N., Goudarzi, G., Babaei, A.A., Baboli, Z. and Hosseinzadeh, M. (2017), "Landfill site selection using GIS and AHP: a case study: Behbahan, Iran", *KSCE Journal of Civil Engineering*, Vol. 21 No. 1, pp. 111-118.

Rezaian, S. and Jozi, S.A. (2016), "Application of multi criteria decision-making technique in site selection of wind farm- a case study of Northwestern Iran", *Journal of the Indian Society of Remote Sensing*, Vol. 44 No. 5, pp. 803-809.

Rikalovic, A., Cosic, I. and Lazarevic, D. (2014), "GIS based multi-criteria analysis for industrial site selection", *Procedia Engineering*, Vol. 69, pp. 1054-1063.

Rother, E.T. (2007), "Systematic literature review X narrative review", *Acta Paulista de Enfermagem*, Vol. 20 No. 2, pp. v-vi.

Roy, B. (1968), "Classement et choix en présence de points de vue multiples", *Revue Fran{$\backslash$k {i}}aise d'informatique et de Recherche Opérationnelle*, Vol. 2 No. 8, pp. 57-75.

Roy, B. (1978), "ELECTRE III: Un algorithme de classement fondé sur une représentation floue des préférences en présence de critères multiples", *Cahiers Du CERO*, Vol. 20 No. 1, pp. 3-24.

Roy, B. and Bertier, P. (1971), "La methode ELECTRE II: une methode de classement en presence de critteres multiples", SEMA (Metra International), Direction Scientifique, Note de Travail No. 142, Paris, p. 25.

Roy, B. and Hugonnard, J.C. (1982), "Ranking of suburban line extension projects on the Paris metro system by a multicriteria method", *Transportation Research Part A: General*, Vol. 16 No. 4, pp. 301-312.

Saaty, T.L. (1977), "A scaling method for priorities in hierarchical structures", *Journal of Mathematical Psychology*, Vol. 15 No. 3, pp. 234-281.

Saaty, T.L. (1996), *Decision Making with Dependence and Feedback: The Analytic Network Process*, RWS Publication.

Sánchez-Lozano, J.M., García-Cascales, M.S. and Lamata, M.T. (2016), "GIS-based onshore wind farm site selection using fuzzy multi-criteria decision making methods. Evaluating the case of Southeastern Spain", *Applied Energy*, Vol. 171, pp. 86-102.

Şengül, Ü., Eren, M., Eslamian Shiraz, S., Gezder, V. and Şengül, A.B. (2015), "Fuzzy TOPSIS method for ranking renewable energy supply systems in Turkey", *Renewable Energy*, Vol. 75, pp. 617-625.

Shaheen, M. and Khan, M.Z. (2016), "A method of data mining for selection of site for wind turbines", *Renewable and Sustainable Energy Reviews*, Vol. 55, pp. 1225-1233.

Shimray, B.A., Singh, K.M., Khelchandra, T. and Mehta, R.K. (2017), "Optimal ranking of hydro power plant sites based on MLP-BP and fuzzy inference approach", *8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 189-193.

Song, L., Liu, C. and Li, B. (2015), "Optimal selection of location for community hospitals a case of Huilongguan region in Beijing", *2015 IEEE International Conference on Information and Automation*, pp. 2803-2806.

Sultana, N. and Rasel, R.I. (2016), "Evaluation of geographic locations for river bridge construction: a multi-criteria decision analysis with evidential reasoning approach", *3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1-4.

Temur, G.T. (2016), "A novel multi attribute decision making approach for location decision under high uncertainty", *Applied Soft Computing*, Vol. 40, pp. 674-682.

Thongpun, A., Nasomwart, S., Peesiri, P. and Nananukul, N. (2017), "Decision support model for solar plant site selection", *2017 IEEE International Conference on Smart Grid and Smart Cities (ICSGSC)*, pp. 50-54.

Ting, C.-Y., Ho, C.C., Yee, H.J. and Matsah, W.R. (2018), "Geospatial analytics in retail site selection and sales prediction", *Big Data*, Vol. 6 No. 1, pp. 42-52.

Toloie-Eshlaghy, A. and Homayonfar, M. (2011), "MCDM methodologies and applications: a literature review from 1999 to 2009", No. 21, p. 53.

Triantono, H.B. and Susetyarto, M.B. (2017), "Technical use of analytical hierarchy process and Delphi method in determining terminal location multi function of Merak Port", *2017 International Conference on Information Management and Technology (ICIMTech)*, pp. 350-355.

Trivedi, A. and Singh, A. (2017), "A hybrid multi-objective decision model for emergency shelter location-relocation projects using fuzzy analytic hierarchy process and goal programming approach", *International Journal of Project Management*, Vol. 35 No. 5, pp. 827-840.

Ubando, A.T., Promentilla, M.A.B., Culaba, A.B. and Tan, R.R. (2015), "Application of spatial analytic hierarchy process in the selection of algal cultivation site for biofuel production: a case study in the Philippines", *2015 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 1-6.

van Laarhoven, P.J.M. and Pedrycz, W. (1983), "A fuzzy extension of Saaty's priority theory", *Fuzzy Sets and Systems*, Vol. 11 Nos 1–3, pp. 199-227.

Vasileiou, M., Loukogeorgaki, E. and Vagiona, D.G. (2017), "GIS-based multi-criteria decision analysis for site selection of hybrid offshore wind and wave energy systems in Greece", *Renewable and Sustainable Energy Review*, Vol. 73, p. 745, available at: https://doi.org/10.1016/j.rser.2017.01.161

Velasquez, M. and Hester, P. (2013), "An analysis of multi-criteria decision making methods", *International Journal of Operations Research*, Vol. 10, pp. 56-66.

Villacreses, G., Gaona, G., Martínez-Gómez, J. and Jijón, D.J. (2017), "Wind farms suitability location using geographical information system (GIS), based on multi-criteria decision making (MCDM) methods: the case of continental Ecuador", *Renewable Energy*, Vol. 109, pp. 275-286.

Vučijak, B., Kurtagić, S.M. and Silajdžić, I. (2016), "Multicriteria decision making in selecting best solid waste management scenario: a municipal case study from Bosnia and Herzegovina", *Journal of Cleaner Production*, Vol. 130, pp. 166-174.

Wang, J., Tsai, C.H. and Lin, P.C. (2016), "Applying spatial-temporal analysis and retail location theory to pubic bikes site selection in Taipei", *Transportation Research Part A: Policy and Practice*, Vol. 94, pp. 45-61.

Wang, K., Peng, Y., Hu, C., Xu, J. and Guan, Q. (2017), "Deployment optimization method for hydrological sensor network to maximize spatial coverage", *6th International Conference on Agro-Geoinformatics*, pp. 1-6.

Wang, T.C. (2012), "The interactive trade decision-making research: an application case of novel hybrid MCDM model", *Economic Modelling*, Vol. 29 No. 3, pp. 926-935.

Wechtaisong, C., Sutthitep, T. and Prommak, C. (2014), "Multi-objective planning and optimization for base station placement in WiMAX network", *11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 1-4.

Wibowo, S., Deng, H. and Zhang, X. (2014), "Fuzzy multicriteria decision support for solid waste disposal method and site selection", *9th IEEE Conference on Industrial Electronics and Applications*, pp. 1774-1779.

Wiguna, K.A., Sarno, R. and Ariyani, N.F. (2016), "Optimization solar farm site selection using multi-criteria decision making fuzzy AHP and PROMETHEE: case study in Bali", *2016 International Conference on Information Communication Technology and Systems (ICTS)*, pp. 237-243.

Wu, L. and Xie, H. (2016), "Research on the location of the railway logistics center based on existing railway freight station", *2016 International Conference on Logistics, Informatics and Service Sciences (LISS)*, pp. 1-5.

Wu, Y., Chen, K., Zeng, B., Yang, M., Li, L. and Zhang, H. (2017), "A cloud decision framework in pure 2-tuple linguistic setting and its application for low-speed wind farm site selection", *Journal of Cleaner Production*, Vol. 142, pp. 2154-2165.

Wu, Y., Zhang, J., Yuan, J., Geng, S. and Zhang, H. (2016), "Study of decision framework of offshore wind power station site selection based on ELECTRE-III under intuitionistic fuzzy environment: a case of China", *Energy Conversion and Management*, Vol. 113, pp. 66-81.

Xu, H., Deng, G., Li, Y., Wang, X., Wu, H., Zhou, Q. and Jiang, B. (2016), "Numerical simulation for tidal current turbine siting metrics of Zhoushan Archipelago", *OCEANS 2016 – Shanghai*, pp. 1-7.

Yao, S. and Cheng, S. (2017), "Visualized data analysis for site selection for remedial education institutions – a case study of educational open data", *10th International Conference on Ubi-Media Computing and Workshops (Ubi-Media)*, pp. 1-5.

Yongfei, M., Canghai, W., Huabiao, W., Yanjiao, H., Chunlai, L., Shujie, Z., Jiaxin, Z., Yu, Z., Zhengxi, L., Yujie, D., Guobin, F., Libin, Y. and Xianmin, W. (2017), "Research on site selection and protection configuration of distributed power supply in microgrid system", *2017 International Conference on Smart Grid and Electrical Automation (ICSGEA)*, pp. 131-133.

Zavadskas, E.K., Turskis, Z. and Kildienė, S. (2014), "State of art surveys of overviews on MCDM/MADM methods", *Technological and Economic Development of Economy*, Vol. 20 No. 1, pp. 165-179.

Zhang, X., Hui, G., Gao, Q., Ren, X., Zhou, B., Yang, D. and Bi, Y. (2017), "A multi-object optimization model of electricity fee payment site selection based on multiple payment methods", *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1540-1546.

Zhao, H. and Li, N. (2016), "Optimal siting of charging stations for electric vehicles based on fuzzy Delphi and hybrid multi-criteria decision making approaches from an extended sustainability perspective", *Energies*, Vol. 9 No. 4, p. 19, available at: https://doi.org/10.3390/en9040270

Zoghi, M., Houshang Ehsani, A., Sadat, M., Javad Amiri, M. and Karimi, S. (2017), "Optimization solar site selection by fuzzy logic model and weighted linear combination method in arid and semi-arid region: a case study Isfahan-IRAN", *Renewable and Sustainable Energy Reviews*, Vol. 68, pp. 986-996.

**Corresponding author**
Jeremy Yee Li Yap can be contacted at: jeremyyap.kylmac@gmail.com

Received 9 May 2018
Revised 7 November 2018
23 March 2019
Accepted 26 March 2019

# Adoption of Big Data analytics in construction: development of a conceptual model

Jiwat Ram
*University of South Australia – Mawson Lakes Campus, Mawson Lakes, Australia, and*
Numan Khan Afridi and Khawar Ahmed Khan
*Shandong University, Jinan, China*

## Abstract

**Purpose** – Big Data (BD) is being increasingly used in a variety of industries including construction. Yet, little research exists that has examined the factors which drive BD adoption in construction. The purpose of this paper is to address this gap in knowledge.

**Design/methodology/approach** – Data collected from literature (55 articles) were analyzed using content analysis techniques. Taking a two-pronged approach, first study presents a systematic perspective of literature on BD in construction. Then underpinned by technology–organization–environment theory and supplemented by literature, a conceptual model of five antecedent factors of BD adoption for use in construction is proposed.

**Findings** – The results show that BD adoption in construction is driven by a number of factors: first, technological: augmented BD–BIM integration and BD relative advantage; second, organizational: improved design and execution efficiencies, and improved project management capabilities; and third, environmental: augmented availability of BD-related technology for construction. Hypothetical relationships involving these factors are then developed and presented through a new model of BD adoption in construction.

**Research limitations/implications** – The study proposes a number of adoption factors and then builds a new conceptual model advancing theories on technologies adoption in construction.

**Practical implications** – Findings will help managers (e.g. chief information officers, IT/IS managers, business and senior managers) to understand the factors that drive adoption of BD in construction and plan their own BD adoption. Results will help policy makers in developing policy guidelines to create sustainable environment for the adoption of BD for enhanced economic, social and environmental benefits.

**Originality/value** – This paper develops a new model of BD adoption in construction and proposes some new factors of adoption process.

**Keywords** Sustainability, Technological innovation, BIM, Project management, Asset management, Data analysis

**Paper type** Research paper

## 1. Introduction

Big Data (BD) is changing the operational dynamics of businesses by facilitating innovations in products and services; and improvements in productivity, decision making and organizational capabilities. Construction industry is no exception to these changes. In fact, given that the construction industry is plagued by severe inefficiency problems which translate into low productivity, costing global economy an estimated $1.6 trillion a year (Barbosa *et al.*, 2017); adopting a technology like BD seems to be unavoidable.

More so, the above-mentioned problems exclude loss of human lives due to safety and hazardous nature of work. Design optimization, lack of digitalization, project management, workers safety, green-house gas contributions and economization of construction work are some of the other ongoing challenges faced by the construction industry (Li, Xu and Zhang, 2017; Li, Wu, Shen, Wang and Teng, 2017). Resultantly, recent efforts have been directed at achieving efficiencies and reducing impact of construction industry's environmental footprint (Barbosa *et al.*, 2017). Using latest technologies such as BD, data analytics and

Built Environment Project and Asset Management
Vol. 9 No. 4, 2019
pp. 564-579
© Emerald Publishing Limited
2044-124X
DOI 10.1108/BEPAM-05-2018-0077

building information modeling (BIM) seems to be a logical way forward to ease some of the pressures faced by the industry (Barbosa *et al.*, 2017).

Ahmed *et al.* (2017) concur and argue that the analytics of BD collected at various stages in the construction cycle will facilitate gaining new insights, thus improving predictions and decision making (e.g. improved design decision making). The common approach used in different construction segments is that alternative designs are generated and imposed during the execution phase. Such practices lead to delays, material wastage and are considered impractical. Inappropriate decisions at the design stage result in an approximately 33 percent of construction waste (Bilal *et al.*, 2015). The use of BD can overcome or minimize such waste and enhance resource efficiencies.

BD of past projects can be mined using appropriate analytics tools to perform text analysis, link analysis and dimensional analysis; which will help achieve BIM efficiencies. Data mining of BD can help in identifying the triggers to safety problems and preventing occupational injuries to workers (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016).

With the increased use of BIM technologies for construction design and execution processes, integrating BD–BIM will yield many benefits such as, improved decision making; enhanced modeling and design efficiencies; identifying the causes of construction failures; detecting damages to the building structures; monitoring the actions of heavy machinery and workers (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016; Motawa, 2017). Using BD for gaining insights on stakeholders' engagement and project planning can result in productive and efficient management of projects (Ekambaram *et al.*, 2018). Marr (2016) highlights one example where a new BD-driven BIM system helped in achieving the savings of $11m in costs and shortened project completion time by 12 weeks.

The above-discussed advantages and savings seem to be just the tip of the iceberg of the value that the adoption and use of BD can create in improving construction performance and stimulating economic activities.

Despite the potential benefits, little work has been done to examine the DB adoption in construction. The existing work (e.g. Bilal *et al.*, 2015; Raguseo, 2018) offer generic insights. While studies (e.g. Amasyali and El-Gohary, 2018; Bibri, 2018a, b; Chen and Lu, 2017; Ekambaram *et al.*, 2018; Koseleva and Ropaite, 2017; Mawed and Al-Hajj, 2017) have examined the application of BD in many industrial contexts; the in-depth theoretically driven investigations on the BD adoption in construction are clearly lacking.

We argue that such a situation is counterproductive, which not only hampers the transfer of knowledge to industry but also dilutes the efforts to build academic knowledge in a cohesive manner.

Adding to the predicament, the current estimates on the use of BD for construction show very light to nominal use (Alavi and Gandomi, 2017). As such, there are calls (e.g. Alavi and Gandomi, 2017; Barbosa *et al.*, 2017) for enhanced digitalization of construction industry to improve productivity and efficiencies.

We contend that gaining an understanding of the BD adoption process in construction and the factors that drive it is timely and needed for two reasons (Alavi and Gandomi, 2017). First, such an understanding will facilitate devising strategies and plans to accentuate the adoption and use of BD for achieving much needed construction productivity and efficiency improvements. Second, it will help digitalization of industry. Moreover, it will also help address the calls for identifying factors that drive BD adoption (Raguseo, 2018).

Addressing the above-discussed gap in knowledge and fulfilling calls (e.g. Raguseo, 2018), this study thus investigates the following question:

*RQ1*. What are the factors that drive adoption of BD for use in construction industry?

The study draws upon technology–organization–environment (TOE) framework to propose new factors and model their relationships to BD adoption in construction, thus uniquely contributing to advancement of academic knowledge in an area with little research.

For the purpose of clarity, BD is defined here as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation" (Gartner, 2018). Whereas, BIM is defined "as a set of interrelating policies, processes and technologies that generate a systematic approach to managing the critical information for building design and project data in digital format throughout the lifecycle of a building" (Wong and Zhou, 2015).

## 2. Methodology

The study used secondary data involving articles published on BD and construction. The data collection involved search of two key databases, i.e. ScienceDirect and Emerald. The keywords strings such as BD adoption, factors, role and impact on construction industry were used to maximize the search output. Construction industry-related variants such as built environment and buildings were used to enlarge the search yield (Table I). Given the research question of the study, rationale was to use keywords that help find articles on the BD adoption-related issues in construction industry context, as can be seen in Table I.

The fact that the study was able to collect a size-able number of relevant articles shows that the keywords used were appropriate for the study. It also transpired in the search that the research papers examining BD adoption in construction remain limited to none, thus confirming the need for this study.

The data search yielded a sample size of 55 relevant articles. These 55 articles were then reviewed and categorized in broad themes and sub-themes. The review led to a realization that almost all the 55 papers either discussed BD or BIM in construction context, resulting in broad themes, i.e. BD, BIM or BD opportunities and challenges. This initial categorization was further analyzed to consolidate and identify the relevant sub-themes. We reviewed the abstracts, study's objective(s), and findings to develop a more granular understanding. It led to further categorization of articles in sub-themes such as, BD applications in multiple segments (e.g. construction, facility management (FM)), Governance and management, or BIM application in project management.

We explain sub-theme categorization through an example. For instance, the article by Bradley *et al.* (2016) discussed BIM within the infrastructure domain and associated modeling standards, among other issues. Hence the article has been categorized under sub-theme BIM applications in infrastructure (see Table II). The article also discussed the importance of BIM integration with technologies such as BD, which supports this study's conceptualization of the proposed factor "Augmented BD–BIM integration" (Figure 1).

**Table I.**
List of keyword phrases

| Keywords | Keywords |
|---|---|
| "Big Data adoption" + "Construction" | "Big Data adoption factors" + "Construction" |
| "Big Data adoption" + "Built environment" | "Big Data adoption" + "Buildings" |
| "Role of Big Data" + "Construction industry" | "Big Data" + "Building information modeling" |
| "Big Data" + "Industrialized construction" | "Big Data applications" + "Construction industry" |
| "Big Data impacts" + "Construction" | "Big Data adoption" + "Opportunities" + "Construction" |
| "Big Data" + "Project management" | "Competitive advantage of Big Data" + "Construction" |
| "Role of predictive analytics" + "Construction industry" | "Relative advantage" + "Big Data" + "construction" |

| Themes | Sub-themes | References |
|---|---|---|
| BIM | 1. BIM application in infrastructure | Bradley *et al.* (2016) |
| | 2. Smart cities and buildings | Yamamura *et al.* (2017) |
| | 3. BIM application in waste management | Akinade *et al.* (2017, 2018), Lu, Webster, Chen, Zhang, and Chen (2017) |
| | 4. BIM applications in sustainability | Lu, Wu, Chang and Li (2017), Wong and Zhou (2015) |
| | 5. BIM applications in construction, project management | Li, Xu and Zhang (2017), Li, Wu, Shen, Wang and Teng (2017), Park *et al.* (2018), Rowlinson (2017), Smith (2016) |
| | 6. BIM adoption | Ahmed and Kassem (2018) |
| | 7. Governance and management | Alreshidi *et al.* (2017) |
| | 8. BIM–BD integration | Aziz *et al.* (2017), Buffat *et al.* (2017), Motawa (2017), Zhong *et al.* (2017) |
| | 9. Building performance | Chien *et al.* (2017), Gerrish *et al.* (2017) |
| | 10. BIM application in facility management | Edirisinghe *et al.* (2017) |
| BD | 1. BD Applications in multiple fields (civil engineering, construction, energy, facilities, real estate; sustainability, urban planning, utilities) | Ahmed *et al.* (2017), Akhavian and Behzadan (2015, 2016), Ang and Seng (2016), Amasyali and El-Gohary (2018), Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha (2016), Cook (2015), Du *et al.* (2014), Koseleva and Ropaite (2017), Mawed and Al-Hajj (2017), Shen *et al.* (2017), Walker (2016) |
| | 2. Smart cities and smart buildings applications | Bibri (2018a, b), Bibri and Krogstie (2017a, b), Hashem *et al.* (2016), Kim (2018), Mehmood *et al.* (2017), Ng *et al.* (2017), Pan *et al.* (2016), Plageras *et al.* (2018) |
| | 3. Project management | Ekambaram *et al.* (2018), Zhang *et al.* (2015) |
| | 4. Construction waste analytics | Bilal, Oyedele, Akinade, Ajayi, Alaka, Owolabi, Qadir, Pasha and Bello (2016), Bilal *et al.* (2017), Chen and Lu (2017), Lu *et al.* (2015), Lu *et al.* (2016) |
| BD opportunities and challenges | 1. Management efficiencies | Deutsch and Leed (2015), Hao *et al.* (2015), Wang and Zhai (2016) |
| | 2. Adoption | Kwon *et al.* (2014), Raguseo (2018) |

**Table II.** Themes and sub-themes as identified from literature

The categorization of the articles in themes and sub-themes (Table II) helped not only in a systematic structuring and organizing of the literature, but also in underlining and presenting a critical perspective of some of the key issues, triggers and factors that facilitate the adoption of BD in construction.

Once the categorization was finalized, the relevant articles were analyzed using content analysis technique to understand the factors and associated issues that drive adoption of BD in construction.

The entire process of data collection, themes/sub-themes classification, using content analysis and model development is consistent with prior studies (e.g. Edirisinghe *et al.*, 2017; Li, Xu and Zhang, 2017; Li, Wu, Shen, Wang and Teng, 2017) in BD construction context.

Next, we present the review of literature (Section 3) and then based on the issues identified in the review and underpinned by TOE (Section 4), we develop hypotheses and a new model of BD adoption in construction (Section 5).

## 3. Literature review

### *3.1 Advantages and challenges of BD adoption in construction*

The construction industry is known to be besieged by inefficiency, poor performance of labor and resource utilization problems (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade,

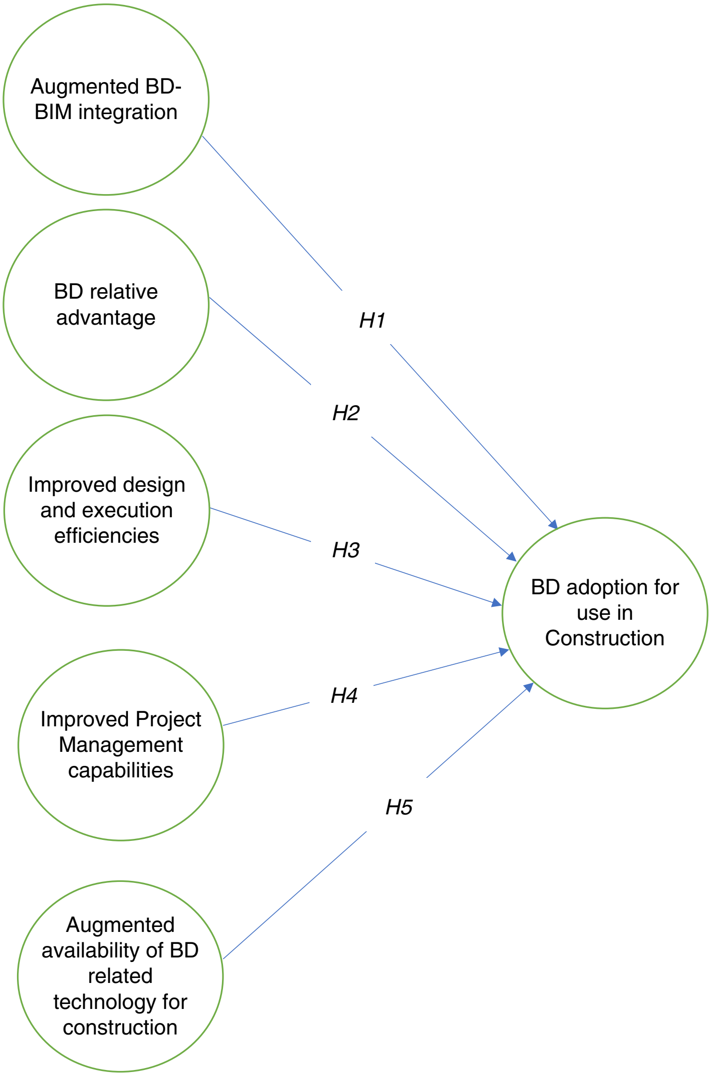

Augmented BD-BIM integration
BD relative advantage
Improved design and execution efficiencies
Improved Project Management capabilities
Augmented availability of BD related technology for construction
H1
H2
H3
H4
H5
BD adoption for use in Construction

**Figure 1.**
Model of BD adoption in construction

Owolabi, Alaka and Pasha, 2016). Such a situation has resulted in concentrated efforts to adopt technologies, e.g. BIM, BD and data analytics to improve productivity and enhance efficiencies in planning, design and overall delivery of construction projects (Raguseo, 2018).

BD adoption is considered pivotal to achieving construction efficiencies as it could facilitate data mining and finding new insights such as gaining an understanding of the factors that cause work delays (Kim *et al.*, 2008). Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha (2016) concur suggesting that knowledge discovery in databases is an efficient approach to analyze large construction data sets to identify causes of construction delays, cost overrun and quality controls. BD enables analyzing (e.g. using Multivariate statistical techniques) large data sets to predict total cycle time of construction operation and predict the accuracy of estimation at the early stages of construction projects

(Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016). Mawed and Al-Hajj (2017) found that BD helped improve performance in FM and enabled change in business and the operation models facilitating informed, smarter and quick decisions. BD adoption provides capabilities that can enable an organization to achieve competitive advantage (Kwon *et al.*, 2014).

Despite the benefits, construction industry lags behind in BD adoption (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016). Kwon *et al.* (2014) conducted a survey from a wide variety of industries including construction and found that the adoption of BD is influenced by an organization's capabilities in collecting and maintaining quality data. Raguseo (2018) examined risks and benefits and concluded that the adoption of BD improves productivity and organizational capabilities in skills development and collection of large amounts of data assets. The author argued that privacy and security of data remain the top two risks and recommended conducting further studies to understand the factors and associated issues to BD adoption.

*3.1.1 BD adoption across multiple fields*. Safety remains one of the priorities in construction sector. Akhavian and Behzadan (2016) used automated sensor data capture technique to understand construction workers' behaviors through activity recognition technology. They used smartphones to capture body movements of workers and simulated the data to understand the workers' productivity and occupational safety on site. The experiment could prove to be a useful starting point to use BD not only for improving the occupational health and workplace safety of workers on site, but also for enhancing overall productivity and efficiencies of other professionals involved in the back-office construction-related work. Big sensor data can also be used for urban planning and management activities including "air pollution monitoring, assistive living, disaster management systems and intelligent transportation" (Ang and Seng, 2016).

BD can help in understanding energy consumption behaviors and achieving efficiencies (Amasyali and El-Gohary, 2018). As such efforts to develop energy consumption prediction models predominantly for commercial and/or educational buildings with a particular interest in overall cooling, heating and lighting energy consumption patterns have been the focus of investigations (Amasyali and El-Gohary, 2018). However, Koseleva and Ropaite (2017) point to the challenges in using energy-related BD for achieving consumption efficiencies and suggest that limited applications are available to process such BD, particularly when several energy-related dimensions are involved.

Building occupancy has direct bearing on the energy consumption efficiencies. Shen *et al.* (2017) examined 50 projects/systems, and reviewed and compared them in "terms of occupancy sensing type, occupancy resolution, accuracy, ground truth data collection method, demonstration scale, data fusion and control strategies." Authors suggest that implicit occupancy sensing can provide capabilities to achieve efficiencies in building energy management "through optimal delivery of building services (including lighting, heating, ventilating and air conditioning) with lower costs compared to traditional explicit sensing approaches."

Mawed and Al-Hajj (2017) examined the data collected from multiple sources and technologies to improve the efficiency of the FM services. Their preliminary findings show that asset owners and service providers in FM are still in early stages of assessing the applications of BD and the benefits and challenges of using BD. Ahmed *et al.* (2017) concur and argue that while BD has a significant role to play in FM, yet firms operating within FM lack an understanding of the value of BD, thus awareness need to be improved. One of the challenges they pointed out was the imbalance between the data capture and data analysis and hence the need for more work to analyze the data. They also identified a number of factors such as data privacy, data security, data heterogeneity and lack of available cases of BD use which could influence its adoption in FM.

Extending the discourse, Aziz *et al.* (2017) proposed that FM can be improved by using data from design and construction stages to plan maintenance schedules. They suggested that the integration of disparate data such as road networks, mobiles and sensors could help decision makers in designing efficient built environment solutions.

BD also plays a core role in infrastructure development and planning, as Pan *et al.* (2016) investigated and confirmed the importance of urban BD for achieving efficiencies in city intelligence.

The growing availability of BD technologies is one of the factors driving adoption of BD in construction (Raguseo, 2018). However, Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha (2016) and Walker (2016) argue that while application of BD and knowledge created by BD has many potential benefits, yet its adoption is still relatively slow. With developments in BIM, Internet of Things (IoT), and cloud computing; the chances are BD uptake will amplify in coming years.

### *3.2 Factors driving BD adoption in construction*

*3.2.1 BD–BIM integration and technological capabilities*. BIM is an emerging solution which facilitates integration and management of information for the whole building lifecycle (Wong and Zhou, 2015). BD captured throughout the lifecycle can be used for improved BIM output, thus enhancing the effectiveness of BIM. As such, increased interests in using integrated BD–BIM solutions and the growth in availability of integration technologies are driving BD adoption in construction (Aziz *et al.*, 2017).

BIM is used for Green buildings design and development, as it enables addressing sustainability issues during designing processes (Wong and Zhou, 2015). Complemented by BD, BIM is used for performance analyses and simulations. These include energy performance analysis, $CO_2$ emission analysis, simulation of lighting and also some integrated optimization of building performance. The integration of BD–BIM helps designers at early design phase as it enables gaining an integrated and visualized view of building performance (Lu, Wu, Chang and Li, 2017).

BD–BIM integrated solutions are also used in FM for predicting operational efficiencies, facilitating efficient design and minimizing waste (Bilal *et al.*, 2015). Edirisinghe *et al.* (2017) proposed a BD integrated BIM-enabled FM framework covering elements including planning, value realization, leadership, data capture and integration techniques, and legal and policy context. They argued that framework will offer proactive decision making and response capabilities.

*3.2.2 BD's role in improved management*. BD adoption is expected to enable changes in stakeholder engagement resulting in productive and efficient management of projects. Rowlinson (2017) highlighted the role of BIM in integrated project delivery (IPD). The author argues that more efforts are needed to integrate BIM into IPD through a process of change management which can be achieved by involving relevant institutional stakeholders including policy makers.

Collection of real time information about stakeholders and different phases of construction and operations enables optimizing the building portfolio and developing sustainable and smart cities (Mawed and Al-Hajj, 2017). Walker (2016) agrees as the author acknowledges that impact of BD on overall construction management and in particular project management is a game changer. The author stressed that the changes in knowledge management and learning due to BD and associated technologies will spur innovations and efficiencies.

Cokins (cited in Mawed and Al-Hajj, 2017) suggested that BD can bring performance improvements in construction in at least six areas, i.e. strategic planning and execution; cost visibility and driver behavior; customer intelligence; forecasting, planning and predictive analytics; enterprise risk management and process improvement. These components align with quality standards, such as six sigma, to reduce or eliminate waste, and streamline processes in order to reduce the cycle times, eventually leading to productivity and efficiency improvements.

Smith (2016) highlighting the challenges argued that "designers not providing full access to the models" and technological incompatibilities are hindering efforts in utilization of BIM and BD in construction.

## 4. Development of conceptual model of BD adoption

The model developed by this study is informed by TOE framework. TOE has been used widely by prior studies to model adoption of technologies including, e.g. Enterprise 2.0 (Jia *et al.*, 2017). Given its multi-context coverage and inclusiveness, TOE provided a robust theoretical underpinning to conceptualize various factors that drive the adoption of BD in construction, hence adopted by this study.

### *4.1 Technology–organization–environment framework*

TOE posits that a technological innovation process is an ensemble of three inter-connected contexts: technological, organizational and environmental (Tornatzky and Fleischer, 1990).

The "Technological" context is meant to consider technologies within and outside an organizational ecosystem which influences organizations to adopt and use the available latest technologies, and deploy change. An organization considering adopting new technologies will look at the benefits of the technology and how its adoption will enhance efficiencies and add value to its operations (Baker, 2012). As such a host of factors including, but not limited to, relative advantage, higher technological competence, perceived benefits/usefulness, cost efficiencies have been identified by earlier research (Jia *et al.*, 2017).

Organizational context is another element of TOE which involves assessing inwardly the organizational strengths, weaknesses and characteristics. It includes the structure; processes; means of communication; human and physical capabilities; top management involvement and support; size and slack resources among others (Baker, 2012).

Environmental context provides a lens to look into those factors that are related to an organization's business ecosystem and the opportunities and challenges present in the corresponding business environment (Baker, 2012). Researchers have identified a host of factors that either influence or influenced by the environmental context. Some of these factors are competitive pressures, regulations and policies, industrial protocols, marketing opportunities, and value chain dynamics (Jia *et al.*, 2017).

## 5. The conceptual model

Informed by the literature review (Section 3) and underpinned by TOE, this study proposes five factors that influence the adoption of BD in construction (Table III). These factors and their relationships to adoption are presented in the conceptual model (Figure 1).

### *5.1 Augmented BD–BIM integration*

The developments in BD are accentuating efforts to integrate BD and BIM technologies (Aziz *et al.*, 2017). The integration of BD–BIM facilitates improved cost, time estimates and optimized scheduling (Aziz *et al.*, 2017). Akinade *et al.* (2017) argue that augmented BD–BIM and BIM–IoT integrations are providing value added capabilities to construction organizations.

Yamamura *et al.* (2017) investigated a solution for optimizing energy performance through an integrated geographic information systems (GIS) and BIM. The BD collected from GIS was fed into BIM for modeling, leading to development of an optimal design for energy management and renewable energy utilization. Through their case study they tested

**Table III.**
Factors influencing adoption of BD in construction

| Factors | TOE classification | Factor drawn from the discussion in Literature Review Section(s) | Some reference |
|---|---|---|---|
| 1. Augmented BD–BIM integration | Technological | 3.2.1, 3.2.2 | Aziz *et al.* (2017), Akinade *et al.* (2017) |
| 2. BD relative advantage | Technological | 3.1, 3.1.1 | Barima (2017), Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha (2016), Raguseo (2018) |
| 3. Improved design and execution efficiencies | Organizational | 3.1, 3.1.1, 3.2.2 | Motawa (2017), Amasyali and El-Gohary (2018) |
| 4. Improved project management capabilities | Organizational | 3.1.1, 3.2.2 | Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha (2016), Ekambaram *et al.* (2018) |
| 5. Augmented availability of BD-related technology for construction | Environmental | 3.1.1, 3.2.2 | Bibri (2018a, b), Plageras *et al.* (2018) |

the integrated GIS–BIM system for urban planning of Tokyo city and found that the integrated GIS–BIM system provides an appropriate solution for effective energy planning and renewable energy management.

The integration of social media-based BD with BIM is helping in facility planning and management (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016). Similarly, RFID-based BD and BIM integration is used for FM (Meadati *et al.*, 2010). An integrated BD–BIM system that enables capturing building operational knowledge is another development highlighting the growing synergies among two technologies and benefits they offer (Motawa, 2017).

Following the above arguments, we assert that the increased technological integration between BD and BIM are enabling the use of BD in construction to enhance modeling efficiencies and achieve benefits such as improved planning, designing, implementation and control capabilities. Hence the following hypothesis is proposed:

*H1*. Augmented BD–BIM integration is positively associated with the adoption of BD for use in construction.

### *5.2 BD relative advantage*

The increased use of BD to gain insights for decision making, new product development and achieving operational and strategic capabilities is becoming a source of competitive advantage for the adopting organizations (Raguseo, 2018).

Kwon *et al.* (2014) established that benefits proposition of using BD significantly influence its adoption among organizations including construction organizations. They argued that effective management of data helps develop managerial and operational capabilities which become a source of competitive advantage.

The enormous volume and variety of data generated at various stages in construction projects' lifecycle make BD an important asset for organizations. These data in several forms such as drawings, text, numbers (estimates), videos and photos can be used for improved decision making and efficiencies related to planning, execution and maintenance of construction assets (Barima, 2017). BD adoption can facilitate advanced simulation to improve whole lifecycle performance of built environment products (Wang and Zhai, 2016). Such uses can help organizations become productive and reduce operational risks, gaining competitive advantage in a saturated construction marketplace (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016).

Akinade *et al.* (2017) highlight the value created by BD adoption in construction, such as, in cost savings, speed and efficiencies and informed decision making, just to mention a few. Construction organizations adopting BD gain relative advantage in "end product improvement, design improvements, improved procurement, physical construction processes improvement, construction maintenance processes improvement, enhanced new materials development and boosted technical skills development" (Barima, 2017).

BD enabled data analytics facilitates scenario planning and forecasting (Deutsch and Leed, 2015). Hao *et al.* (2015) echoed the views and argued that BD enables better planning and stakeholder engagement.

The above discussion clearly highlights that those construction organizations who adopt BD are expected to gain multiple advantages relative to their competitors in the industry. Thus, we propose:

*H2*. The relative advantage of using BD is positively associated with the adoption of BD for use in construction.

### *5.3 Improved design and execution efficiencies*

The adoption of BD enables augmented design and execution capabilities (Motawa, 2017). The variety of data collected from the use of constructed infrastructure and facilities help understand the requirements, problems/challenges and advantages in a better way which enhances organizational capabilities in design and execution of construction work (Akinade *et al.*, 2017).

BD in the form of internal data (e.g. temperature, humidity, occupancy, energy consumption) and external data (e.g. weather, traffic, business activities, economy) can enable gaining insights to improve construction work productivity and post construction environment through reduction in air pollution, improved disaster management, and intelligent transportation (Ang and Seng, 2016).

BD can help reduce the negative impact of construction activities on environment by enabling development of efficient consumption models, improving consumption predictive capabilities and construction design processes that maximize natural use of resources and minimize generation of un-recyclable waste (Amasyali and El-Gohary, 2018; Shen *et al.*, 2017).

Construction waste and its treatment also remains one of the thorny issues, resulting in a growing number of studies that have investigated the importance of BD in enabling improved management of construction waste (Bilal *et al.*, 2017; Chen and Lu, 2017). Lu *et al.* (2016) used BD to analyze the construction waste management (CWM) performance and recommended that "the value of environment protection leadership" should be promoted for improved CWM performance.

One of the ways to minimize the negative impact of environmental footprint of construction activities is to use BD for smart city planning and development. Bibri (2018a, b) argued that IoT-driven BD can help in design and development of smart cities. The author argued that state-of-the-art sensor-based BD can be used for planning and design of built environment, waste management, water management and transportation, just to mention a few. This coincided with earlier work by Pan *et al.* (2016) in urban planning context, and Ng *et al.* (2017) who proposed a master data management (MDM) solution that uses IoT BD for smart cities planning.

In light of above arguments, the study hypothesizes:

*H3*. Improved design and execution efficiencies facilitated by BD are positively associated with the adoption of BD for use in construction.

### 5.4 Improved project management capabilities

Construction is predominantly project-based activity. A significant amount of data is collected throughout various phases of construction project lifecycle, which become an important source for enhancing project management capabilities (Ekambaram *et al.*, 2018).

Rowlinson (2017) proposed an IPD approach where project-based data collected at various stages in lifecycle can be integrated with BIM for enhanced capabilities. The use of BD induces agility and effectiveness of project management (Zhang *et al.*, 2015). Information generated from planning, design, execution and post construction stages is stored digitally. This information includes cost and schedule estimates, design-related information, construction performance data, variance and issues, risks, quality and procurement data. BD enables understanding these data for enhanced decision making, effective risk management, quality improvements, improved planning, safe and effective construction, and overall improved management (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016).

BD, therefore, can help build project management maturity within organization, as organizations will be able to analyze the past mistakes and become more consistent in their project management-based activities. Such a situation will lead to using project management standards more effectively, and hence we propose:

*H4.* Improved project management capabilities facilitated by BD are positively associated with the adoption of BD for use in construction.

### 5.5 Augmented availability of BD-related technology for construction

The increased availability of BD-related technology is one of the driving factors of BD adoption for use in construction (Bibri, 2018a). Off-the-shelf or open source master data management applications are available to analyze BD (Ang and Seng, 2016).

The growth in technologies that serve as BD capture sources is helping adoption too. Availability of technologies, such as, IoT, large scale wireless sensor systems. GIS, RFID and POS (point of sales) are key sources of data collection resulting in increased adoption of BD in construction. Further, open standards for interoperability in the Construction and Urban Planning fields, such as IFC (buildings and infrastructure) and CityGML are also contributing toward favorable view of adopting BD in construction. Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha (2016) provided a comprehensive review of BD applications and the associated technologies.

The development in data storage technology is also leading to the adoption of BD. Distributed data storage file technology such as Hadoop and Tachyon are facilitating the BD developments (Raguseo, 2018). Advancements in Relational database technology such as the development of "Not only SQL" systems, which provide improved traditional data management in numerous ways are also furthering adoption of BD (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016).

Availability of visual analytics software to view the results in graphic/visual formats is also contributing toward BD adoption (Raguseo, 2018). Technologies such as Social–BIM, BIMCloud (to store user interaction with building models' data through IFC, Apache Cassandra, hosted on Amazon EC2) are proposed to be used in construction industry further helps in BD adoption (Bilal, Oyedele, Qadir, Munir, Ajayi, Akinade, Owolabi, Alaka and Pasha, 2016).

Plageras *et al.* (2018) propose use of IoT-based BD applications for smart buildings. MDM is another BD-based application used in infrastructure and smart cities planning and development (Ng *et al.*, 2017). The above discussion highlights the vast opportunities involving availability of BD-related technology for construction (Barima, 2017). Hence, we propose:

*H5.* Augmented availability of BD-related technology for construction is positively associated with the adoption of BD for use in construction.

## 6. Conclusions

BD and associated technologies are being increasingly adopted for achieving efficiencies and improved productivity. Yet, little work has been done to examine the factors that drive BD adoption for use in construction. Taking a two-fold approach, the study first presents a structured review of literature. Then, underpinned by TOE theory, the study proposes five new factors and builds a corresponding model explaining the relationship of these factors to BD adoption for use in construction.

We argue that BD adoption is influenced by increased technological synergies between BD and BIM. These synergies facilitate BD–BIM integration allowing feeding BD into BIM and leveraging this connectivity for design and development efficiencies.

Organizations that adopt BD are expected to enjoy relative advantage compared to the competitors due to improved capabilities in gaining new insights from BD. Technological superiority achieved by adopting BD provides added value and help construction organizations remain comparatively sustainable in highly intense construction market.

The findings also suggest that achieving technological sophistication is not the only reason to adopt BD, but capacity development in project management and resource management are also some of the factors weighing in the favorable adoption decision.

With the increased use of BD across various business sectors, the technological competence is also growing. These fast-paced technological developments and availability of technologies facilitating capture, storage and processing of BD is building confidence among the potential adopter in construction organizations to adopt BD.

### *6.1 Implications for theory and practice*

The study makes several theoretical and managerial contributions. First, the study proposes a set of five antecedent factors explaining the adoption of BD in construction encapsulating technology, organization and environment context.

Second, underpinned by TOE theory, the study develops hypotheses and a corresponding conceptual model. The newly developed model contributes toward extending application of TOE (particularly for the adoption stage of innovation process) to new form of BD technological innovation.

It is pertinent to note that the growing trends in the use of BD warrant upgrading innovation theories and making them more inclusive with factors related to technologies such as BD. The current theories on technological innovation process may not be fittingly applicable to BD given novel technology. Therefore, we believe that the proposed model contributes significantly toward extending current knowledge on innovation process and development of new academic insights and thoughts.

Third, by taking a theory-driven approach to examine BD adoption in construction, the study contributes to an area where theoretically informed existing research is scarce.

Fourth, through a systematic literature review, the study has categorized a large body of literature on BD in construction in various themes/sub-themes which will serve as a platform for future research work and development of knowledge in a cohesive manner.

Finally, fulfilling the calls, the study builds knowledge that will help in efforts toward digitalization of construction industry.

Managerially, findings will help chief information officers, IT/IS managers, business development managers and senior executives in construction organizations to understand the factors that drive the adoption of BD and evaluate their organizational environment for adoption considerations. The knowledge developed in the study can also be used for business case development for adopting BD, making informed investment decisions and strategies for realizing expected returns once BD is adopted.

Findings will help policy makers in devising policy guidelines and regulations for uptake of BD in construction. Given that the construction work produces long-term environmental impacts and BD adoption can help mitigate some of the negative impacts, policy makers can form policies to help organizations adopt BD and minimize the negative impact on environment.

### *6.2 Limitations and future directions*

The study opens up new avenues of scholarly investigations. First, further studies can identify more antecedent factors of BD adoption in construction, which will extend the findings presented here. Factors such as availability of BIM (software tools), GIS and sensors (as RFIDs) as enabling tools driving BD adoption, and BD being as an enabler of Lean and Green Building design, construction activities could be tested.

Second, future work can consider articles focused on manufacturing as industrialized construction, pre-fabrication and modularization also plays a big role in diving BD adoption.

Third, more work can look into the differences and similarities in factors driving adoption across different types of construction environments, e.g. traditional vs sustainable or green building construction.

Fourth further studies can actually collect the data for the factors identified in this study and examine the causal relationships presented in the model. Finally, more work is needed to identify the item measures for the constructs shown in the model of the study.

The study has some limitation too. The model proposed in the study is based on secondary literature-based data, and primary data can be collected to examine the hypothetical relationships. The generalization of findings therefore needs to be done with lot of caution. Second, like other qualitative content analyses-based studies, findings are based on subjective understanding and a longitudinal or cross-sectional study may be done to extend this work. Finally, BD research is still in its infancy, so the work presented here is of fundamental knowledge building type which necessitates further studies to advance knowledge in this area.

## References

Ahmed, A.L. and Kassem, M. (2018), "A unified BIM adoption taxonomy: conceptual development, empirical validation and application", *Automation in Construction*, Vol. 96, pp. 103-127.

Ahmed, V., Tezel, A., Aziz, Z. and Sibley, M. (2017), "The future of big data in facilities management: opportunities and challenges", *Facilities*, Vol. 35 Nos 13/14, pp. 725-745.

Akhavian, R. and Behzadan, A.H. (2015), "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers", *Advanced Engineering Informatics*, Vol. 29 No. 4, pp. 867-877.

Akhavian, R. and Behzadan, A.H. (2016), "Smartphone-based construction workers' activity recognition and classification", *Automation in Construction*, Vol. 71, pp. 198-209.

Akinade, O.O., Oyedele, L.O., Ajayi, S.O., Bilal, M., Alaka, H.A., Owolabi, H.A. and Arawomo, O.O. (2018), "Designing out construction waste using BIM technology: stakeholders' expectations for industry deployment", *Journal of Cleaner Production*, Vol. 180, pp. 375-385.

Akinade, O.O., Oyedele, L.O., Omoteso, K., Ajayi, S.O., Bilal, M., Owolabi, H.A., Alaka, H.A., Ayris, L. and Looney, J.H. (2017), "BIM-based deconstruction tool: towards essential functionalities", *International Journal of Sustainable Built Environment*, Vol. 6 No. 1, pp. 260-271.

Alavi, A.H. and Gandomi, A.H. (2017), "Big data in civil engineering", *Automation in Construction*, Vol. 79, pp. 1-2.

Alreshidi, E., Mourshed, M. and Rezgui, Y. (2017), "Factors for effective BIM governance", *Journal of Building Engineering*, Vol. 10, pp. 89-101.

Amasyali, K. and El-Gohary, N.M. (2018), "A review of data-driven building energy consumption prediction studies", *Renewable and Sustainable Energy Reviews*, Vol. 81, pp. 1192-1205.

Ang, L.M. and Seng, K.P. (2016), "Big sensor data applications in urban environments", *Big Data Research*, Vol. 4, pp. 1-12.

Aziz, Z., Riaz, Z. and Arslan, M. (2017), "Leveraging BIM and big data to deliver well maintained highways", *Facilities*, Vol. 35 Nos 13/14, pp. 818-832.

Baker, J. (2012), "The technology–organization–environment framework", in Dwivedi, Y., Wade, M. and Schneberger, S. (Eds), *Information Systems Theory. Integrated Series in Information Systems*, Vol. 28, Springer, New York, NY.

Barbosa, F., Woetzel, J., Mischke, J., Ribeirinho, M.J., Sridhar, M., Parsons, M., Bertram, N. and Brown, S. (2017), *Reinventing Construction through a Productivity Revolution*, McKinsey Global Institute, Minneapolis, available at: https://goo.gl/1Nqqf8 (accessed February 6, 2017).

Barima, O. (2017), " 'BIG Data' and construction value delivery", *Construction Projects: Improvement Strategies, Quality Management and Potential Challenges*, Nova Science Publishers, New York, NY, pp. 113-135.

Bibri, S.E. (2018a), "The IoT for smart sustainable cities of the future: an analytical framework for sensor-based big data applications for environmental sustainability", *Sustainable Cities and Society*, Vol. 38, pp. 230-253.

Bibri, S.E. (2018b), "A foundational framework for smart sustainable city development: theoretical, disciplinary, and discursive dimensions and their synergies", *Sustainable Cities and Society*, Vol. 38, pp. 758-794.

Bibri, S.E. and Krogstie, J. (2017a), "Smart sustainable cities of the future: an extensive interdisciplinary literature review", *Sustainable Cities and Society*, Vol. 31, pp. 183-212.

Bibri, S.E. and Krogstie, J. (2017b), "ICT of the new wave of computing for sustainable urban forms: their big data and context-aware augmented typologies and design concepts", *Sustainable Cities and Society*, Vol. 32, pp. 449-474.

Bilal, M., Oyedele, L.O., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A. and Alaka, H.A. (2017), "The application of web of data technologies in building materials information modelling for construction waste analytics", *Sustainable Materials and Technologies*, Vol. 11, pp. 28-37.

Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Akinade, O.O., Ajayi, S.O., Alaka, H.A. and Owolabi, H.A. (2015), "Analysis of critical features and evaluation of BIM software: towards a plug-in for construction waste minimization using big data", *International Journal of Sustainable Building Technology and Urban Development*, Vol. 6 No. 4, pp. 211-228.

Bilal, M., Oyedele, L.O., Akinade, O.O., Ajayi, S.O., Alaka, H.A., Owolabi, H.A., Qadir, J., Pasha, M. and Bello, S.A. (2016), "Big data architecture for construction waste analytics (CWA): a conceptual framework", *Journal of Building Engineering*, Vol. 6, pp. 144-156.

Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A. and Pasha, M. (2016), "Big Data in the construction industry: a review of present status, opportunities, and future trends", *Advanced Engineering Informatics*, Vol. 30 No. 3, pp. 500-521.

Bradley, A., Li, H., Lark, R. and Dunn, S. (2016), "BIM for infrastructure: an overall review and constructor perspective", *Automation in Construction*, Vol. 71, pp. 139-152.

Buffat, R., Froemelt, A., Heeren, N., Raubal, M. and Hellweg, S. (2017), "Big data GIS analysis for novel approaches in building stock modelling", *Applied Energy*, Vol. 208, pp. 277-290.

Chen, X. and Lu, W. (2017), "Identifying factors influencing demolition waste generation in Hong Kong", *Journal of Cleaner Production*, Vol. 141, pp. 799-811.

Chien, S.C., Chuang, T.C., Yu, H.S., Han, Y., Soong, B.H. and Tseng, K.J. (2017), "Implementation of cloud BIM-based platform towards high-performance building services", *Procedia Environmental Sciences*, Vol. 38, pp. 436-444.

Cook, D. (2015), "RICS futures: turning disruption from technology to opportunity", *Journal of Property Investment & Finance*, Vol. 33 No. 5, pp. 456-464.

Deutsch, R. and Leed, A.P. (2015), "Leveraging data across the building lifecycle", *Procedia Engineering*, Vol. 118, pp. 260-267.

Du, D., Li, A. and Zhang, L. (2014), "Survey on the applications of big data in Chinese real estate enterprise", *Procedia Computer Science*, Vol. 30, pp. 24-33.

Edirisinghe, R., London, K.A., Kalutara, P. and Aranda-Mena, G. (2017), "Building information modelling for facility management: are we there yet?", *Engineering, Construction and Architectural Management*, Vol. 24 No. 6, pp. 1119-1154.

Ekambaram, A., Sørensen, A.Ø., Bull-Berg, H. and Olsson, N.O. (2018), "The role of big data and knowledge management in improving projects and project-based organizations", *Procedia Computer Science*, Vol. 138, pp. 851-858.

Gartner (2018), "Gartner IT glossary", available at: www.gartner.com/it-glossary/big-data/ (accessed April 18, 2019).

Gerrish, T., Ruikar, K., Cook, M., Johnson, M., Phillip, M. and Lowry, C. (2017), "BIM application to building energy performance visualisation and management: challenges and potential", *Energy and Buildings*, Vol. 144, pp. 218-228.

Hao, J., Zhu, J. and Zhong, R. (2015), "The rise of big data on urban studies and planning practices in China: review and open research issues", *Journal of Urban Management*, Vol. 4 No. 2, pp. 92-124.

Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E. and Chiroma, H. (2016), "The role of big data in smart city", *International Journal of Information Management*, Vol. 36 No. 5, pp. 748-758.

Jia, Q., Guo, Y. and Barnes, S.J. (2017), "Enterprise 2.0 post-adoption: extending the information system continuance model based on the technology-organization-environment framework", *Computers in Human Behavior*, Vol. 67, pp. 95-105.

Kim, H., Soibelman, L. and Grobler, F. (2008), "Factor selection for delay analysis using knowledge discovery in databases", *Automation in Construction*, Vol. 17 No. 5, pp. 550-560.

Kim, P.W. (2018), "Operating an environmentally sustainable city using fine dust level big data measured at individual elementary schools", *Sustainable Cities and Society*, Vol. 37, pp. 1-6.

Koseleva, N. and Ropaite, G. (2017), "Big data in building energy efficiency: understanding of big data and main challenges", *Procedia Engineering*, Vol. 172, pp. 544-549.

Kwon, O., Lee, N. and Shin, B. (2014), "Data quality management, data usage experience and acquisition intention of big data analytics", *International Journal of Information Management*, Vol. 34 No. 3, pp. 387-394.

Li, X., Xu, J. and Zhang, Q. (2017), "Research on construction schedule management based on BIM technology", *Procedia Engineering*, Vol. 174, pp. 657-667.

Li, X., Wu, P., Shen, G.Q., Wang, X. and Teng, Y. (2017), "Mapping the knowledge domains of building information modeling (BIM): a bibliometric approach", *Automation in Construction*, Vol. 84, pp. 195-206.

Lu, W., Chen, X., Ho, D.C. and Wang, H. (2016), "Analysis of the construction waste management performance in Hong Kong: the public and private sectors compared using big data", *Journal of Cleaner Production*, Vol. 112, pp. 521-531.

Lu, W., Chen, X., Peng, Y. and Shen, L. (2015), "Benchmarking construction waste management performance using big data", *Resources, Conservation and Recycling*, Vol. 105, pp. 49-58.

Lu, W., Webster, C., Chen, K., Zhang, X. and Chen, X. (2017), "Computational building information modelling for construction waste management: moving from rhetoric to reality", *Renewable and Sustainable Energy Reviews*, Vol. 68, pp. 587-595.

Lu, Y., Wu, Z., Chang, R. and Li, Y. (2017), "Building information modeling (BIM) for green buildings: a critical review and future directions", *Automation in Construction*, Vol. 83, pp. 134-148.

Marr, B. (2016), "How Big Data and analytics are transforming the construction industry", *Forbes*, available at: www.forbes.com/sites/bernardmarr/2016/04/19/how-big-data-and-analytics-are-transforming-the-construction-industry/#4d3effdd33fc (accessed April 18, 2019).

Mawed, M. and Al-Hajj, A. (2017), "Using big data to improve the performance management: a case study from the UAE FM industry", *Facilities*, Vol. 35 Nos 13/14, pp. 746-765.

Meadati, P., Irizarry, J. and Akhnoukh, A.K. (2010), "BIM and RFID integration: a pilot study", *Advancing and Integrating Construction Education, Research and Practice, Second International Conference on Construction in Developing Countries*, *Cairo*, *August 3-5*, pp. 570-578.

Mehmood, R., Meriton, R., Graham, G., Hennelly, P. and Kumar, M. (2017), "Exploring the influence of big data on city transport operations: a Markovian approach", *International Journal of Operations & Production Management*, Vol. 37 No. 1, pp. 75-104.

Motawa, I. (2017), "Spoken dialogue BIM systems – an application of big data in construction", *Facilities*, Vol. 35 Nos 13/14, pp. 787-800.

Ng, S.T., Xu, F.J., Yang, Y. and Lu, M. (2017), "A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning", *Procedia Engineering*, Vol. 196, pp. 939-947.

Pan, Y., Tian, Y., Liu, X., Gu, D. and Hua, G. (2016), "Urban big data and the development of city intelligence", *Engineering*, Vol. 2 No. 2, pp. 171-178.

Park, Y.N., Lee, Y.S., Kim, J.J. and Lee, T.S. (2018), "The structure and knowledge flow of building information modeling based on patent citation network analysis", *Automation in Construction*, Vol. 87, pp. 215-224.

Plageras, A.P., Psannis, K.E., Stergiou, C., Wang, H. and Gupta, B.B. (2018), "Efficient IoT-based sensor Big Data collection – processing and analysis in smart buildings", *Future Generation Computer Systems*, Vol. 82, pp. 349-357.

Raguseo, E. (2018), "Big data technologies: an empirical investigation on their adoption, benefits and risks for companies", *International Journal of Information Management*, Vol. 38 No. 1, pp. 187-195.

Rowlinson, S. (2017), "Building information modelling, integrated project delivery and all that", *Construction Innovation*, Vol. 17 No. 1, pp. 45-49.

Shen, W., Newsham, G. and Gunay, B. (2017), "Leveraging existing occupancy-related data for optimal control of commercial office buildings: a review", *Advanced Engineering Informatics*, Vol. 33, pp. 230-242.

Smith, P. (2016), "Project cost management with 5D BIM", *Procedia-Social and Behavioral Sciences*, Vol. 226, pp. 193-200.

Tornatzky, L.G. and Fleischer, M. (1990), *The Processes of Technological Innovation*, Lexington Books, Lexington, MA.

Walker, D.H. (2016), "Reflecting on 10 years of focus on innovation, organisational learning and knowledge management literature in a construction project management context", *Construction Innovation*, Vol. 16 No. 2, pp. 114-126.

Wang, H. and Zhai, Z.J. (2016), "Advances in building simulation and computational techniques: a review between 1987 and 2014", *Energy and Buildings*, Vol. 128, pp. 319-335.

Wong, J.K.W. and Zhou, J. (2015), "Enhancing environmental sustainability over building life cycles through green BIM: a review", *Automation in Construction*, Vol. 57, pp. 156-165.

Yamamura, S., Fan, L. and Suzuki, Y. (2017), "Assessment of urban energy performance through integration of BIM and GIS for smart city planning", *Procedia Engineering*, Vol. 180, pp. 1462-1472.

Zhang, Y., Luo, H. and He, Y. (2015), "A system for tender price evaluation of construction project based on big data", *Procedia Engineering*, Vol. 123, pp. 606-614.

Zhong, R.Y., Peng, Y., Xue, F., Fang, J., Zou, W., Luo, H., Ng, S.T., Lu, W., Shen, G.Q. and Huang, G.Q. (2017), "Prefabricated construction enabled by the Internet-of-Things", *Automation in Construction*, Vol. 76, pp. 59-70.

**Corresponding author**
Jiwat Ram can be contacted at: jiwat.ram@gmail.com