# Big Data Analytics system for costing power transmission projects

1. Juan Manuel Davila Delgado, PhD., manuel.daviladelgado@uwe.ac.uk
*Associate Professor, University of the West of England, UK*
2. Lukumon Oyedele*, PhD., l.oyedele@uwe.ac.uk
*Professor, University of the West of England, UK*
3. Muhammad Bilal, PhD., muhammad.bilal@uwe.ac.uk
*Associate Professor, University of the West of England, UK*
4. Anuoluwapo Ajayi, PhD., anuoluwapo.ajayi@uwe.ac.uk
*Associate Professor, University of the West of England, UK*
5. Lukman Akanbi, PhD., lukman.akanbi@uwe.ac.uk
*Associate Professor, University of the West of England, UK*
6. Olugbenga Akinade, PhD., olugbenga.akinade@uwe.ac.uk
*Associate Professor, University of the West of England, UK*

*Corresponding author.

Inaccurate cost estimates have significant impacts on the final cost of power transmission projects and erode profits. Methods for cost estimation have been investigated thoroughly, but they are not used widely in practice. The purpose of this study is to leverage a Big Data architecture, to manage the large and diverse data required for predictive analytics. This paper presents a Predictive Analytics and Modelling System (PAMS) that facilitates the use of different data-driven cost prediction methods. A 2.75 million-point dataset of power transmission projects has been used as a case study. The proposed Big Data architecture is fit for purpose. It can handle the diverse datasets used in the construction sector. The three most prevalent cost estimation models were implemented (linear regression, support vector regression, and artificial neural networks). All models performed better than the estimated human-level performance. The primary contribution of this study to the Body of Knowledge is an empirical indication that data-driven methods analysed in this study are on average 13.5% better than manual methods for cost estimation of power transmission projects. Additionally, the paper presents a Big Data architecture that can manage and process large varied datasets and seamless scalability.

**Keywords**: Predictive Analytics; Data-Driven; Big Data; Cost Estimation.

## Introduction

Reliable cost estimation is a critical factor for the successful delivery of construction projects. Poor cost estimates have been identified as significant factors contributing to cost overruns and profit erosion (Ahiaga-Dagbui and Smith, 2014; Sridarran et al., 2017). Accurate and systematic cost estimates are of utmost importance for the construction sector where low-profit margins are typical. The average profit margin of the top 100 UK construction companies was 1.5% in 2016 and 2.5% in 2017 (TCI, 2018). Cost estimation for construction projects has been a topic of intense study for many years. Estimating costs in construction is a challenging task due to many factors including the limited information at early phases of the project, the many different clients, the dissimilarity among projects, the diverse contexts and sites, and the large labour force required (Wilson, 2005). Cost estimation of power transmission projects has specific challenges such as large and complex construction sites and difficult accessibility (e.g. highways and river crossings), varying contexts (urban, rural, and protected natural areas), and unknown soil conditions. However, compared with traditional construction projects, power transmission projects also represent advantages for implementing data-driven cost estimation methods including a smaller number of potential clients, a considerable similarity among projects, and limited types of materials and plant equipment used; which facilitates data collection and management.

Many data-driven cost estimation methods have been developed and tested (e.g. Hwang, 2009; Lowe et al., 2006; Sonmez, 2008); and most of the research efforts have been placed on developing more accurate methods and to compare their performance among them, i.e. to find the best possible method (e.g. G.-H. Kim et al., 2004). However, there is no consensus on the best method to address cost estimation yet. More importantly, most construction companies have not adopted advanced cost estimation predictive models; and, in practice, cost estimation still relies heavily on human expertise rather than on systematic data-driven methods (Carr, 1989; Meredith et al., 2014).

A major indication that current cost estimation practices in construction are still unreliable is the number of projects experiencing cost overruns. The construction industry is well known for cost overruns. For example, a study found that large infrastructure projects across the world are on average 80% overbudget (Agarwal et al., 2016). Many factors contribute to cost overruns, but poor cost estimation practices are a major influencing factor (e.g. Adam et al., 2017; Larsen et al., 2016). Aljohani et al. (2017) point out that estimation methods used in practice are still affected by the estimator's bias and varying degrees of experience. Aljohani et al. (2017) also note that in practice data-driven methods are not fully used due to unavailable and unreliable data sources. The many different clients, project types, sites, materials and subcontractors complicate data collection. If data is available, it is usually not accessible through a single interface, and it is stored in different locations and formats, which limits its use.

Despite the fact that data unavailability and poor data management practices are major factors limiting the use of data-driven cost estimation methods (Aljohani et al., 2017), methods reported in literature usually do not present data management frameworks and approaches required to enable data-driven methods. This is becoming more relevant as the amount of data collected by construction companies is increasing substantially and traditional methods for data management cannot cope with the increasing amounts of generated data. For example, none of the traditional methods provide a way to integrate different types of data (Davila Delgado et al., 2015; Gerrish et al., 2015), support dynamic visualisations (Davila Delgado et al., 2018; Mousa et al., 2016), or provide real-time links with Big Data repositories (Bilal et al., 2016). Equally important is the lack of comparisons between the performances of predictive data-driven methods and methods usually used in practice. In most studies in literature, only an indication of how accurate the proposed methods are, is presented. But no comparison is presented with the actual methods used in practice. There is no clear indication of how much

better the predictive data-driven methods are compared with the ones used in practice, and if the increase in performance will justify the investment required to implement predictive data-driven methods.

This study seeks to address: the lack of comparisons between data-driven cost estimation methods and manual methods, and the lack of demonstrations of data management approaches to cope with large amounts of diverse data. The objectives of this study are:

(1) To get a quantitative indication of the difference in the performance between the most common predictive data-driven cost estimation methods and traditional manual methods used in practice.

(2) To demonstrate the implementation of a Big Data architecture that can manage large amounts of data from diverse sources and in different formats.

This paper presents the Predictive Analytics and Modelling System (PAMS), which integrates uses historical financial and project data to predict costs. PAMS enables the extraction of valuable insights by integrating large and varied datasets and performing predictive analytics

## Context and related works

### *Big Data in Architecture, Engineering and Construction (AEC)*

Big Data is the term coined to define sets of data that are too large, complex, and heterogeneous so that traditional software applications cannot process them. The main defining attributes of Big Data are (i) *volume*, i.e. the amount of data; (ii) *variety*, different file formats and structures; and (iii) *velocity*, i.e. the speed at which the data is queried and processed (Erl et al., 2016). Irrespective of the term "Big Data", the size of the datasets is only one challenging aspect of many, and it is, in most cases, not the most significant one (Boyd and Crawford, 2011). Additional attributes have been identified such as value, vision, validation; but for the AEC industry *veracity* (i.e. the consistency, completeness, and reliability of data) and *variety* (i.e.

different file formats, e.g. 2D drawings, 3D models, pictures, animations, spreadsheets) are the key Big Data attributes that constitute a significant challenge for Big Data adoption (Bilal et al., 2016).

Compared with other more modern and less established industries, the amount of data being generated in the AEC industry is smaller by some orders of magnitude. Nevertheless, it also must deal with increasing amounts of data that is generated during the entire life cycle of built assets. The increasing amounts data are driven by the push towards the adoption of the Smart Cities (Batty et al., 2012; Zanella et al., 2014) and Smart Infrastructure (Al-Hader and Rodzi, 2009; Hoult et al., 2009) paradigms. Both of which rely on the use of sensing and monitoring systems and on 3D digital representations of built assets that include performance and condition data (Khan and Hornbæk, 2011). Built assets have been instrumented with various types of sensors and embedded devices, which generate large and dynamic sets of data. For example, sensors are used in buildings to monitor temperature variations (Chen et al., 2014), indoor air quality (Kumar et al., 2016), and occupancy (Akkaya et al., 2015). They are also used to monitor power consumption (Suryadevara et al., 2015), structural condition (Davila Delgado et al., 2017; 2016), and surrounding environmental conditions (Martín-Garín et al., 2018). However, current data management frameworks used in the AEC industry cannot handle the ever increasing and diverse data sets. Big Data management frameworks and programming models must be adopted to handle and process the data effectively. Otherwise, relevant insights and value could not be extracted from the generated data.

Big Data analytics is the broad term that refers to the various methods used to extract insights from data. Big Data analytics draws techniques from various existing fields such as statistics, data mining, business analytics, and applies them to large and diverse datasets. The types of analyses carried out in Big Data analytics can be classified into the following categories (Figure 1 presents the categories and lists examples of methods used in each category): (i) Descriptive

analytics, which uses statistical methods to describe and quantify basic features of the dataset. (ii) Predictive analytics, which uses diverse techniques to analyse current data and make predictions about future events. Lastly (iii) prescriptive analytics, which uses the insights gained by descriptive and predictive analytics to devise actions that lead to a defined goal, e.g. cost reduction. All these types of analyses are beneficial to support the AEC industry in general and have the potential to unleash a new period of productivity improvement, which is a crucial interest for the sector as a whole (Abdel-Wahab and Vogl, 2011; Liberda et al., 2003).

Research efforts reported in literature have been focused mainly on identifying challenges and potential architectures. For example, the Big Data challenges for storing and visualising massive BIM models have been studied (Chen et al., 2016; Gao et al., 2017). Challenges for handling geospatial data (Yang et al., 2017), Earth observation data (Xia et al., 2018), challenges regarding building energy efficiency (Koseleva and Ropaite, 2017), and for managing big visual data and BIM (Han and Golparvar-Fard, 2017) have been reported as well. Nonetheless, there is a significant interest in leveraging Big Data technologies to improve a wide variety of tasks in the AEC industry. These tasks range from support at the conceptual design stage, construction and planning, to tasks related to operations and facility management. However, work has focused on operations and facility management due to easier access to data. For example, Big Data analytics have been used to predict air passenger demands in airports (Kim and Shin, 2016), to model commuting patterns (Wan et al., 2018), and to infer transport mode using data from mobile devices (Semanjski et al., 2017). Also, Jeong et al. (2017; 2019) presented a cloud-based Big Data management and analytics framework that handles the massive and diverse datasets used for bridge monitoring. Wang et al. (2018) presented a Big Data approach to identify potential quality issues in construction components.

However, Big Data has not been widely employed to support tasks in the preconstruction stage such as bidding, tendering, costing, scheduling. This stage represents a massive opportunity

for cost reduction as flawed decision-making and planning leads to mounting cost and time delays during construction (Chan et al., 2004; Olawale and Sun, 2010). Poor cost estimation often defines whether a project is profitable or not. This is the focus of the study presented in this paper as is illustrated in Figure 1.

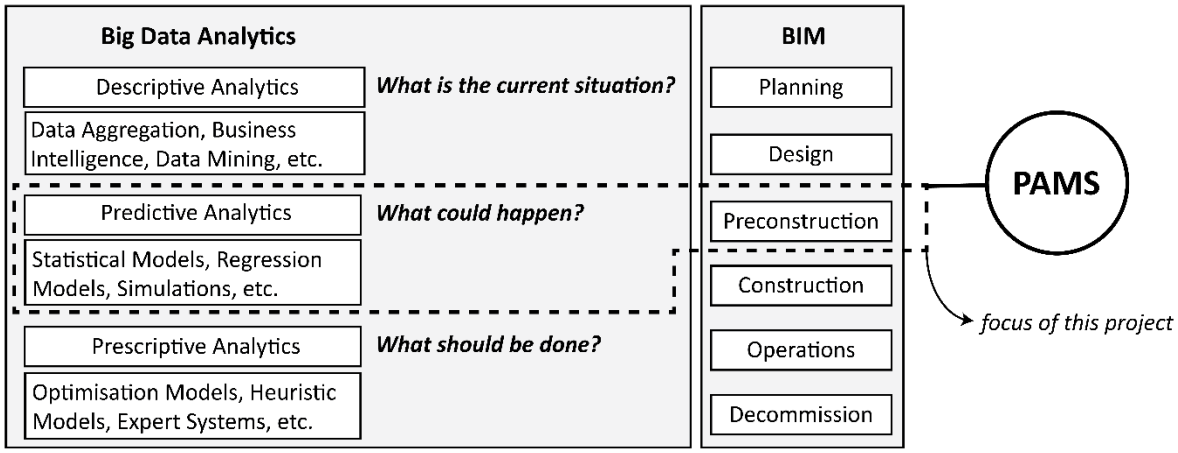


**Figure 1**. Categories of Big Data Analytics and their use on the construction projects' life-cycle. Preconstruction is the focus of study in this project.

### *Predictive analytics*

Numerous efforts have been carried out to apply intelligent systems to address the increasingly complex problems of the AEC area (Irani and Kamal, 2014). A significant application is predictive analytics, which is a process that uses data and statistical algorithms to predict future outcomes. It is mostly used in the financial, healthcare and marketing sectors. The most widely-used predictive technique is linear regression, which is a simple approach to model the relationship between two variables. The relationship is modelled using linear predictor functions whose unknown model parameters are estimated from available data. Other more sophisticated techniques exist as well such as Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANNs). Predictive analytics has not been used widely in the AEC industry, where simulation approaches are preferred. For example, creating buildings in virtual environments and simulating its potential energy use is another approach to predict energy consumption (Elbeltagi et al., 2017).

However, the idea to generate solutions to given problems based on data has been broadly investigated. For example, ANNs, a resurgent and prominent method for predictive analytics, has been used to aid in the design of water harvesting structures (Chandwani et al., 2016), predicting and controlling cooling loads in buildings (Venkatesan and Ramachandraiah, 2018), predicting risks for building maintenance (de Silva et al., 2013), and predicting the escalation of highway construction cost over time (Wilmot and Mei, 2005). Generative and Genetic Algorithms (GAs) have been used to predict structural designs solutions given limited spatial data (Davila Delgado et al. 2013; Hofmeyer et al., 2013; 2015). GAs have been used to analyse the static security of electric power systems (Canto dos Santos et al., 2015) and to generate optimal construction schedules (Faghihi et al., 2014). GAs have been combined with ANNs to optimise environmentally friendly buildings (Sun et al., 2015). SVM has been used to predict failures of construction companies (Horta and Camanho, 2013). Linear interpolation methods have been used to predict the annual electricity consumption of elevators (Tukia et al., 2016) and cooling loads (Geekiyanage and Ramachandra, 2018). Intelligent decision support systems have been developed to facilitate the management of construction processes (Hajdasz, 2014) and rule-based support systems have been used to check regulatory compliance (Beach et al., 2015). Simulation approaches have been combined with ANN models for managing Heating, Ventilation and Air Conditioning (HVAC) systems (Faizollahzadeh Ardabili et al., 2016).

***Cost Estimation***

Cost estimation for construction has been thoroughly investigated. Until now, there is no consensus regarding the supremacy of one single method, even though their performance has been compared and evaluated (e.g. Shane et al., 2009; Trost & Oberlender, 2003). This subsection presents a quick snapshot of the construction cost estimation landscape.

(1) *Analytical and numerical approaches*. These approaches define cost estimation as a regression problem and use different techniques of varying complexity to solve it (e.g. Hwang,

2009; Lowe et al., 2006; Sonmez, 2008). For example, ANNs have been used to predict cost of road tunnel construction based on geological attributes (Petroutsatou et al., 2012), to predict cost of retrofitting due to earthquakes (Jafarzadeh et al., 2014), and to predict costs of irrigation improvement projects (ElMousalami et al., 2018). Firouzi et al. (2016) present a method to predict total cost with dependent cost items, and Dursun & Stoy (2016) present a method that stacks predictive models to create a multistep estimation method.

Approaches that use existing BIM models or design documents to come up cost estimates exist as well. For example, methods to automate the manual process of generating a cost estimate of structural elements (Jadid and Idrees, 2007) and of steel frames (Barg et al., 2018) have been reported in the literature. An approach to cost estimation based on the level of detail of the BIM model (Cheung et al., 2012); and a method that uses Industry Foundation Classes (IFC) data to estimate costs (Ma et al., 2013) have been reported as well. Lastly, Asmar et al. (2011) present an approach to systematise the cost estimation of highway projects at planning stages. The disadvantage of these methods is that they require an existing BIM model, which in most cases is not available at early tendering stages.

(2) *Knowledge-based approaches*. These approaches use codified expert knowledge to support cost estimation. For example, Choi et al. (2014) present a method to estimate the cost of roads using case-based reasoning. Yildiz et al. (2014) present a knowledge-based tool that maps risks to support cost estimation. Ahn et al. (2014) present a case-based reasoning approach to identify attributes that impact cost estimation and in (Ahn et al., 2017) presents a method to determine the effect of covariance in case-based reasoning approaches for cost estimation. The main disadvantage of these methods is that their performance is restricted by the quality of the encoded expert knowledge and cases; and their limited transferability.

(3) *Improvement of cost estimation factors and processes*. These approaches seek to improve the factors affecting the accuracy of cost estimation methods. For example, Yu et al. (2006)

present an approach to define cost indexes in real time; and Cheng et al. (2013) present a model that uses economic and financial data (e.g. Consumer Price Indexes, oil prices, stock market indexes) to predict variation in costs using a modified version of SVM. Hwang (2011) presents a method that uses time series indexes to consider the trends of costs in the market during construction.

More research efforts are required to address cost overruns in the construction industry. While various factors are at play, accurate cost estimates are essential to reduce cost overruns. More importantly, in the existing cost estimation literature for construction, few studies address the use of Big Data architectures and approaches to deal with large amounts of diverse data for cost estimation. Also, there is an insufficient number of comparisons between existing methods used in practice and data-driven methods. The motivation of this study is to address these gaps and contribute to the reduction of cost overruns in the construction industry.

## Predictive Analytics and Modelling System (PAMS)

The main component of the PAMS is the Big Data Analytics Environment shown in Figure 2. It is a server application deployed using the Oracle Big Data Lite, a collection of software that supports Big Data applications, in a virtual machine running Red Hat Enterprise Linux. The Oracle Big Data Lite enables Big Data warehousing, Big Data analytics and machine learning running in the cloud or on premises. The Big Data Analytics Environment can fit and load large volumes of structured and semi-structured data, support various analytic models, and process complex models on large datasets very quickly, all of which are the main features of Big Data architectures (Sagiroglu and Sinanc, 2013). It has three layers: the Unstructured Data Storage layer, the Structured Data Storage layer, and the Data Services Layer (Figure 2).

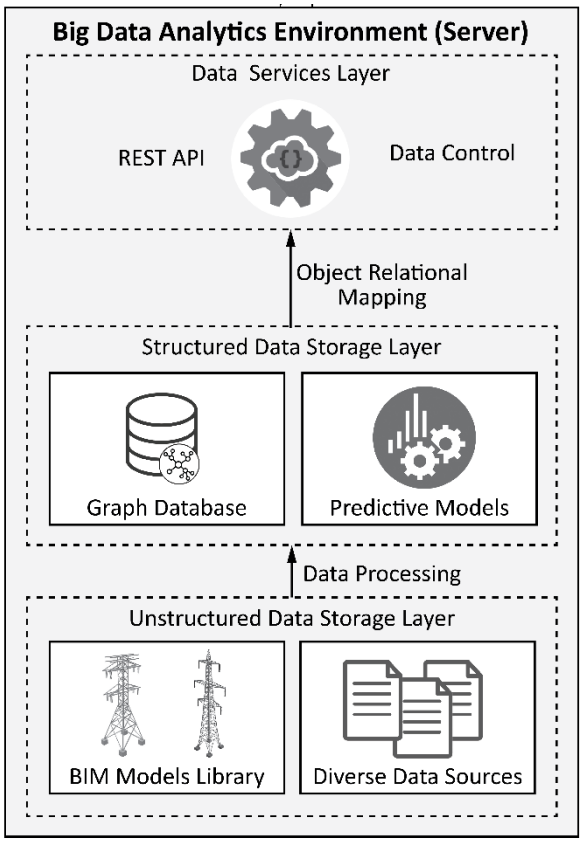


**Figure 2**. Diagram of the Big Data Analytics environment.

The Big Data Analytics Environment has the three principal attributes of Big Data applications (Erl et al., 2016):

**Variety**. The Unstructured Data Storage Layer contains a library of BIM models and a collection of historical data in diverse file formats and structures (i.e. project financial records, material quantities and cost estimates, work plans, labour and plant financial data). The different data in this layer has been cleaned, formatted, and stored in a graph database in the Structured Data Storage Layer. The graph database uses the Resource Description Framework (RDF) to store the data in graph triplets (Klyne and Carroll, 2006). RDF enables to merge and model data from different sources without a defined schema, which facilitates the quick loading of data from varied sources and with different structures. A number of domain ontologies were developed for the graph database. These ontologies standardised the data extracted from the diverse data sources.

**Volume**. The Cloudera Distribution including Apache Hadoop (CDH), which is a collection of software based on the Hadoop Distributed File System (HDS), was used for Big Data management because it provides reliable and high-performance access to large datasets in distributed file systems and it enables parallel processing. The graph database was implemented using Apache HBase, which is an open-source, non-relational, distributed database that runs in combination with HDS providing a robust way of storing and querying large quantities of sparse data.

**Velocity**. A vital requirement of the PAMS is the ability to analyse the stored historical data quickly. For Big Data processing, the Big Data programming model so-called Spark was used for processing the developed predictive models. Spark is at the core of the Berkeley Big Data Analytics Stack (BDAS) framework, which is regarded as the as a next-generation framework for large-scale data processing (Ryza et al., 2015). BDAS is an open source data analytics framework that provides speedy response times for complex computations on large datasets. Compared with other traditional analytics frameworks, BDAS achieves very fast processing times by enabling large-scale in-memory data processing. The analytical pipelines for processing the predictive models were implemented using the R language and environment through SparkR, an R package that provides frontend use to Apache Spark; and MLLib, a Spark machine learning library with an R interface. R was selected because is widely used in academia, is open source, and supports multicore task distribution. The popularity of R for Big Data analytics is growing, and it is becoming a de facto standard, which facilitates further developments. The results of the predictive models are mapped to objects in the Data Services Layer, which exposes them to the client application using a REST (Representational State Transfer) API (Application Programming Interface). This enables the client application to query data via the web robustly and quickly. The data requests and responses are elicited using the JSON (JavaScript Object Notation) format.

The Big Data Analytics Environment also deals with veracity, another attribute of Big Data. Methods to address missing and unreliable data that are common in the construction sector have been implemented and are presented in this paper. While these aspects are not the most commonly addressed in Big Data studies, they are –nevertheless– essential to the construction industry, which lags in the adoption of Big Data technologies.

## Big Data Analytics

Data from overhead power transmission projects constructed in the United Kingdom in the last ten years has been compiled and used as a basis to develop the predictive models in the Big Data Analytics Environment. It includes financial data, i.e. the total actual cost $\boldsymbol{C}$, the estimated cost $\boldsymbol{C}$', profit $\boldsymbol{P}$, profit margin $\boldsymbol{M}$, distance $\boldsymbol{d}$, and region $\boldsymbol{R}$ of each project. This section focuses on describing the analysis carried out to predict the total cost of projects using the financial data.

### *Data Pre-processing*

The data was collected from various heterogeneous sources, i.e. Microsoft Excel files, CSV (comma separated value) files, and relational databases. Incorrect, incomplete, and inconsistent records were identified and removed. E.g. typos and values that were many degrees of magnitude larger than the mean of that variable were removed. Missing and removed profit ($\boldsymbol{P}$) values were resolved by employing the so-called mean imputation technique, i.e. substituting the missing values with the mean of that variable for all other cases (Roderick and Rubin, 2002; Scheffer, 2002). This method was selected because it was assumed that the missing data was not a structural characteristic of the dataset and there were no indications of a strong correlation between profit and the other available variables, e.g. cost, distance, region, etc., for example, see Figure 4. Note that mean imputation does not change the sample mean for the variable but may decrease correlations among variables that have values replaced. To avoid that and because the cost ($\boldsymbol{C}$) and distance ($\boldsymbol{d}$) values are strongly correlated, the missing

and removed distance ($\boldsymbol{d}$) values were resolved using: $\hat{d} = \varrho_n \cdot \mathrm{RAND(a, b)}$, where ($\hat{d}$) is the calculated value for distance; $\mathrm{RAND(a, b)}$ is a function that calculates a pseudo-random number between the constants (a) and (b); $\varrho_n = \hat{C} \cdot d_n / C_n$, where ($d_n$) and ($C_n$) are corresponding distance and cost values of a project, and ($\hat{C}$) is the cost value that does not have a corresponding distance value. This approach maintains the correlation between cost ($\boldsymbol{C}$) and distance ($\boldsymbol{d}$) variables while providing variation to the generated data. The same approach was used for missing cost values. Once the data was cleansed, it was condensed and loaded into the graph database.

***Characteristics of the dataset***

The dataset used for this study is a compilation of financial data from overhead power transmission projects constructed in the United Kingdom in the last ten years, resulting in over 2.75 million data points. This dataset is not particularly large when compared with other datasets used for Big Data analytics in other fields, such as marketing and customer analytics that use data from hundreds of millions of users. Nevertheless, this dataset is very large when compared with the typical datasets used for cost estimation in construction, which usually range from a few dozen to around hundred projects (e.g. ElMousalami et al., 2018; Petroutsatou et al., 2012). Moreover, the Big Data architecture proposed in this paper can handle significantly larger datasets and enables seamless scalability to constantly add more data as it becomes available. Figure 3 shows a sample of the frequency distribution of the cost of the projects. The projects' cost ranges from £10k to £50m. The data has a positively skewed distribution and 94% of all the projects have costs of less than £4.8m. Figure 4 and 5 present the relative distribution of the region and cost up to £100K and more than £100K, respectively. Based on the characteristics of the dataset, it was decided only to use a sample of projects that have a cost ranging from £50k to £4.8 million to develop the predictive models.

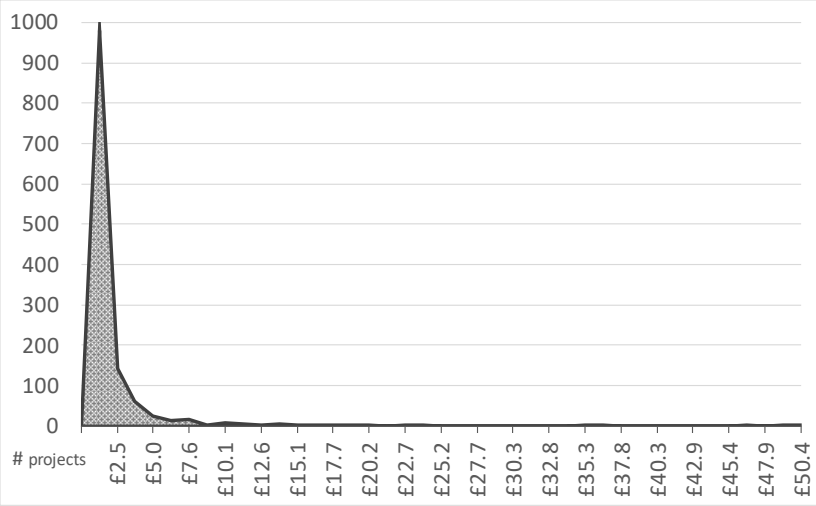


**Figure 3**. Frequency distribution of a sample of the cost of the projects. 94% of the projects have a cost of less than £4.8m.

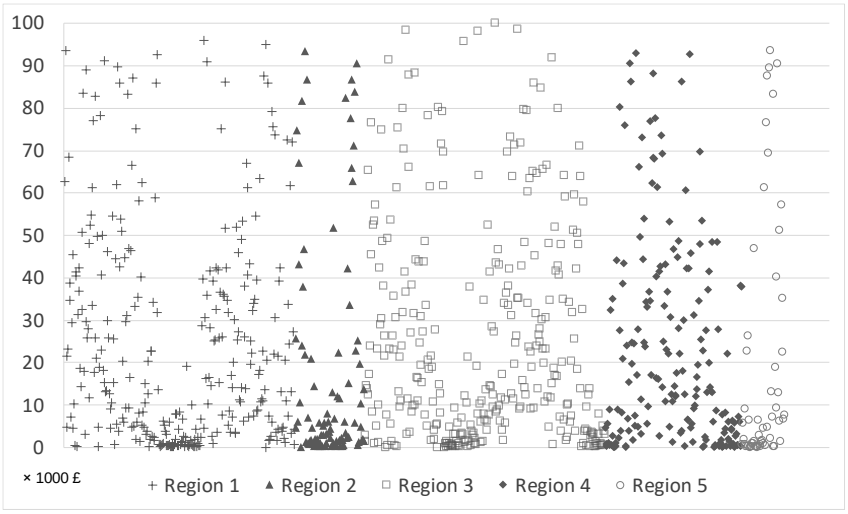


**Figure 4**. The relative distribution between cost and region for projects with costs smaller than £1 million.

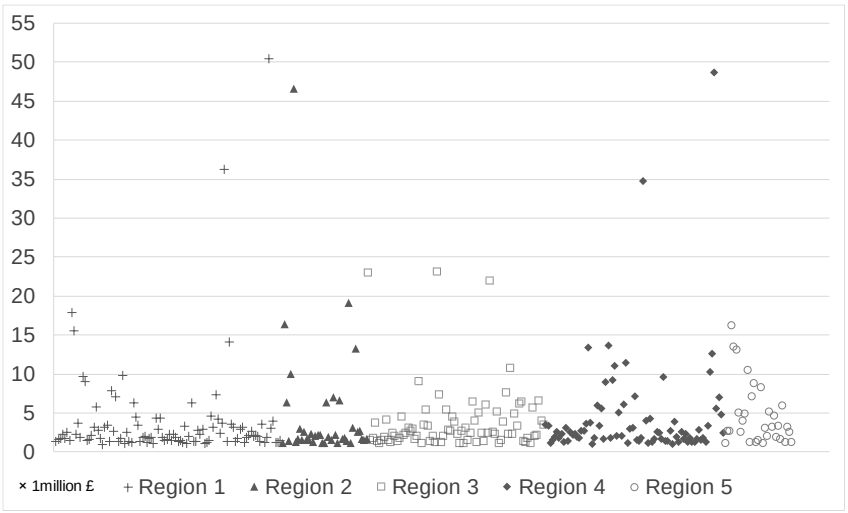


**Figure 5**. The relative distribution between cost and region for projects with costs larger than £1 million.

***Predictive analytics***

Three predictive models –linear regression (LR), support vector regression (SVR), and artificial neural network (ANN)– have been implemented to predict the total cost of the project given its distance. These models were selected because they are the most prevalent models used for cost estimation (Deng and Yeh, 2011). Five per cent of the projects have been used to test the predictive models, and the rest were used for developing the models. The traditional 80/20 per cent split between training and test data was not used due to the large size of the dataset, and to ensure a balanced variance between the parameter estimates and the performance statistic. The projects for testing were selected in a way that they were representative of the total cost distribution.

An LR model ($\hat{y} = \mathrm{w} + \mathrm{b}x$) has been implemented, in which $\hat{y}$ is the predicted *y* value, given $x$, $w = \sum y \sum x^2 - \sum x \sum xy / n\sum x^2 - (\sum x)^2$, and $b = n\sum xy - \sum x \sum y / n\sum x^2 - (\sum x)^2$. Where $x$ and $y$ are the investigated variables, in this case, the distance $\boldsymbol{d}$ and the total cost $\boldsymbol{C}$ of the project respectively. The SVR function used is: $\hat{\boldsymbol{y}} = \left(\boldsymbol{W}, (\Phi \boldsymbol{X})\right) + \mathrm{b}$. Variables $\zeta_i$ and $\xi_i^*$ are introduced to measure the deviation of samples, thus the SVR optimisation problem is expressed as $\min \frac{1}{2}\|\boldsymbol{W}\|^2 + C\sum(\zeta_i + \xi_i^*)$ subject to: $\{f\left(x^i\right) - y_i \leq \varepsilon + \zeta_i;\ y_i - f(x^i) \leq \varepsilon + \xi_i^*;\ \zeta_i, \xi_i^* \geq 0\}$. In which, *C* is the parameter that regulates the trade-off between the margin and the prediction error denoted by the variables $\zeta_i$ and $\xi_i^*$. The final regression function is $f(x) = \sum(a_i - a_i^*)K(x, y) + b$, where $a_i, a_i^*$ are the Lagrange multipliers and $K(x, y)$ is the kernel function. In this case, a linear kernel function was used ($K(x, y) = x^T y + \mathrm{b}$), $\mathrm{C} = 1.0$ , and $\varepsilon = 0.1$. For the ANN model, the following regression loss function was selected $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ and the so-called rectifier linear unit (ReLu) function $g(x) = \max(0, x)$ was used as the activation function. The vectorised forward propagation implementation is as follows: $\boldsymbol{Z}^l = \boldsymbol{W}^l \boldsymbol{X} + b^l$, $\boldsymbol{A}^l = g(Z^l)$, where *X* is the vector of input parameters, *W* is the vector of weights, *A* is the vector of activation functions, and *l* is the number of layers of the model. The

vectorised backward propagation is defined as follows: $\{\delta \boldsymbol{Z}^l = \delta \boldsymbol{A}^{l-1} * g^l(\boldsymbol{Z}^l);\ \delta \boldsymbol{W}^l = (1/2)\ \delta \boldsymbol{Z}^l \cdot \boldsymbol{A}^{(l-1)T};\ \delta b^l = (1/2 \sum \delta \boldsymbol{Z}^l;\ \delta \boldsymbol{A}^{l-1} = \boldsymbol{W}^{(l)T} \cdot \delta \boldsymbol{Z}^l\}$. Note that $\delta \boldsymbol{Z}^l$ is computed using an element wise product. Note that time-dependent factors that affect project cost such as inflation are not accounted in the predictive models thus the resulting cost predictions should be adjusted for inflation at the time that the prediction is carried out.

***Comparison of predictive models***

An indication of the coefficient of determination of the model is given by calculating the coefficient of determination $r^2$, which is computed as follows: $r^2 = 1 - \sum(y_i - \hat{y}_i)^2 / \sum(y_i - \bar{y})^2$. Where $\sum(y_i - \hat{y}_i)^2$ is the sum of squares of the difference between the actual values $y_i$ and the predicted values $\hat{y}_i$; and $\sum(y_i - \bar{y})^2$is the sum of squares between the difference of the actual values $y_i$ and their mean $\bar{y}$. The mean absolute error (MAE) is calculated using: $\frac{1}{n}\sum|y_i - \hat{y}_i|$, where *n* is the number of errors and $|y_i - \hat{y}_i|$ are the absolute errors. The root mean square error (RMSE) is computed as follows: $\sqrt[2]{1 - r^2} SD_y$, where $SD_y$ is the standard deviation of *Y*.

Figure 6 presents a comparison of the results of the three predictive models. Given the quasi-linear correlation between cost and distance, LR can be used as a baseline to measure the performance of SVR and ANN. This very useful because for more complex problems, these results can be used to help to select appropriate predictive models. The coefficient of determination for the three models are 73%, 80.5%, and 74.6% respectively. The effectiveness of LR and ANN are comparable, while SVR performed ~7% better. This indicates that SVR is a better choice when for this type of regression problems. The higher performance of SVR can be explained because –conversely to ANN– it requires fewer training examples to perform reasonably well. SVR is not affected and generalises quite well by small changes in data. On the other hand, ANN performs better for more complex regression problems and requires considerably larger amounts of data.

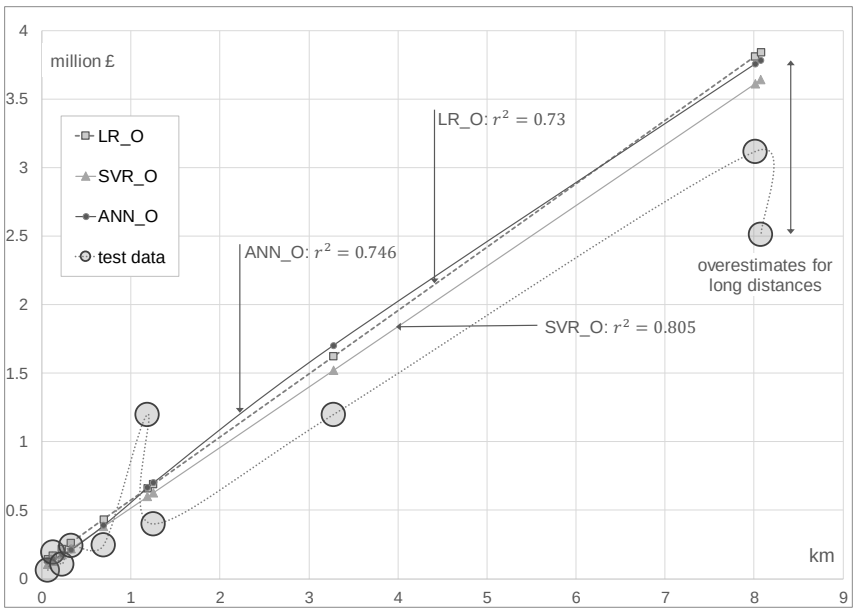


**Figure 6**. Comparison of the predictive models: the linear regression (LR), support vector regression (SVR), and Artificial Neural Network (ANN).

Additionally, the three models overestimate the cost as the distance increases. All the models perform better with projects costing less than £1.5million. A possible reason for this behaviour is that 78% of all the projects have costs of less than £2.5 million (Figure 3). This indicates that there is not enough data to develop accurate predictions because there are few data that delineates the model's behaviour for projects costing more than £2.5 million.

Table 1 presents the coefficient of determination $r^2$, mean absolute error (MAE), and root mean square error (RMSE) for the three predictive models. Table 1 presents results using only the distance as the predictor (rows 1-3) and using the region in which the line is located as an additional variable to predict the cost (rows 4-6). SVR achieved the best $r^2$ and the lowest MAE and RMSE for both sets of results, while LR and ANN achieved similar results. Adding the region, as an additional variable, to predict the cost did not improve the coefficient of determination of any of the models significantly. The MAE and RMSE only improved marginally. Therefore, in this case, the region is not a relevant variable to predict cost as the initial analysis suggested (Figure 4 and 5). Note that this study minimised the number of

variables included in the analysis as it has been found that using more variables do not necessarily increase accuracy (Gardner et al., 2016).

**Table 1**. Coefficient of determination ($r^2$), mean absolute error (MAE), and root mean square error (RMSE) for the three models: linear regression (LR), singular vector regression (SVR), and Artificial Neural Network (ANN).

| | | $r^2$ | MAE | RMSE |
|---|---|---|---|---|
| 1 | LR_O | 0.730 | 0.370 | 0.286 |
| 2 | SVR_O | 0.805 | **0.311** | 0.206 |
| 3 | ANN_O | 0.746 | 0.361 | 0.269 |
| 4 | LR_2V | 0.732 | 0.363 | 0.283 |
| 5 | SVR_2V | **0.805** | 0.312 | **0.206** |
| 6 | ANN_2V | 0.694 | 0.399 | 0.324 |

## Discussion

This paper presented an approach to developing a Big Data system to support cost estimation for power transmission projects. The proposed Big Data architecture proved to be fit for purpose and facilitated the integration of large and diverse data sources. The Big Data Analytics Environment enables the use of the most prevalent cost estimation models used in construction. The presented approach has the three principal attributes of Big Data (Erl et al., 2016): (1) Variety; the approach uses various types of structured and semi-structured data, i.e., financial data and project data in diverse file formats. It employs a standard model for merging data with different underlying schemas. (2) Volume, the approach uses a large dataset (over 2.75 million data points) and employs a Big Data platform (Cloudera Distribution) that uses a distributed file system and non-relational databases for high-performance access to large datasets and parallel processing. (3) Velocity, the approach uses a Big Data framework and programming model (BDAS and Spark) that facilitates in-memory processing to process complex models very quickly

There are many challenges for estimating costs of power transmission projects in a traditional manner such as the extensive construction sites. Estimators usually have to survey many

kilometres long routes in very far and inaccessible places. The construction sites are complex and located in different contexts (urban, rural, and protected natural areas). The power transmission lines usually cross through highways or rivers, which complicates construction and increases risks. Limited information and unknown soil conditions hinders reliable cost prediction. Many resources must be employed to achieve accurate cost prediction and thorough risk identification to support planning tasks effectively. In practice, this is prohibitively expensive, and usually, only rough estimates are carried out. However, power transmission projects have characteristics that facilitate the implementation of data-driven cost estimation methods such as a smaller number of potential clients, a considerable similarity among projects, and limited types of materials and plant equipment used. The main factor limiting the adoption of data-driven methods is the considerable efforts required to integrate diverse data. Power transmission projects have substantially less diverse data than traditional construction projects, which facilitates the implementation of data-driven methods.

An indication of the human level performance for estimating costs of power transmission projects was obtained by calculating $r^2$ using estimations –calculated by planners– and the actual cost at the end of the projects. The calculated human level performance is $r^2 = 63\%$. All three predictive models presented in this study perform better than the human level. SVR represents the highest increase in performance (17%). It is widely acknowledged that data-driven methods could outperform traditional manual methods. However, there was no empirical evidence of how much the difference in performance would be. This study presents an empirical comparison between traditional manual methods and data-driven methods. It provides a quantitative indication of the difference in performance. The results of this study indicate that the average difference in performance between predictive models is ~7.5%, while the average difference in performance between the data-driven methods and the manual method is approximately 13.5%, almost twice as much. This insight represents a relevant implication

for practice, as for example, stakeholders can start adopting simpler predictive models that are easier to implement and obtain a significant increase in performance.

This paper also presented a Big Data architecture to manage the vast and varied datasets required to process the predictive models that traditional methods cannot handle. Note that the main difference between Big Data Analytics and traditional data analytics is the manner in which the data is managed and processed. Algorithms used in Big Data Analytics are, in essence, the same algorithms used for traditional data analytics. The key difference is that the Big Data algorithms run on distributed file systems and non-relational databases using new computing frameworks. These new frameworks enable large-scale in-memory data processing. Traditional data analytics frameworks are very slow in handling queries because they have to sift through large amounts of data stored on disk and are not suitable for complex computations such as advanced predictive models. These limitations prevent extracting value or new insights from data.

## Conclusions

A Predictive Analytics and Modelling System (PAMS) has been presented that generates cost estimates using data from previously constructed designs. The presented Big Data Analytics Environment manages large and diverse datasets for predictive analytics. The proposed Big Data architecture is fit for purpose. The distributed file systems and Big Data frameworks and programming models employed can handle the large and diverse datasets used in the construction sector.

A 2.75 million-point dataset of power transmission projects has been used as a case study. The three most prevalent cost estimation models were implemented (linear regression, support vector regression, and artificial neural networks). The $R^2$s were 73%, 80% and 74% respectively, in line with other results reported in literature. All the implemented models performed better than the estimated human level performance (63%). Data-driven methods for

cost estimation have been studied extensively. However, due to the complexities of construction projects, there is no consensus regarding the best method. Incremental improvements to performance are being achieved constantly; however, adoption in practice remains very low. The intention of this paper is to provide an indication of the potential benefits of adopting predictive data-driven cost estimation methods and present an approach that can be useful in practice to support planning of power transmission projects.

The main contribution to the body of knowledge of this paper is an indication that data-driven cost estimation methods are on average 13.5% better than traditional manual methods for power transmission projects. This study has significant implications for practice because it enables to make data-driven decisions during preconstruction. In a highly competitive sector such as construction, making correct decisions could be the difference between the successful delivery of a project and the survivability of the company. For example, the presented approach will support stakeholders to make accurate cost estimates, to define accurate profit margins and to decide whether it is worthwhile to bid for a project or not. Future steps to improve the presented approach should focus on the following aspects: (i) to investigate whether the missing data in the dataset is a structural characteristic that reflects an underlying attribute of the dataset or are simply input and recording errors; (ii) to compile more granular data, for example to obtain the breakdown of the cost according to labour, materials and plant; and (iii) to develop an automatic adjustment of the predicted costs that takes into account the costs increases due to inflation at the time that the prediction is carried out.

## Data Availability Statement

Data generated or analysed during the study are available from the corresponding author by request.

## Acknowledgements

The authors would like to gratefully acknowledge the EPSRC and Innovate UK for funding this research under grant 102061, application number: 44746 – 322224.

## References

Abdel-Wahab, M., Vogl, B., 2011. Trends of productivity growth in the construction industry across Europe, US and Japan. Constr. Manag. Econ. 29, 635–644. https://doi.org/10.1080/01446193.2011.573568

Adam, A., Josephson, P.-E.B., Lindahl, G., 2017. Aggregation of factors causing cost overruns and time delays in large public construction projects. Eng. Constr. Archit. Manag. 24, 393–406. https://doi.org/10.1108/ECAM-09-2015-0135

Agarwal, R., Chandrasekaran, S., Sridhar, M., 2016. Imagining construction's digital future, Capital Projects and Infrastructure.

Ahiaga-Dagbui, D., Smith, S.D., 2014. Rethinking construction cost overruns: cognition, learning and estimation. J. Financ. Manag. Prop. Constr. 19, 38–54. https://doi.org/10.1108/JFMPC-06-2013-0027

Ahn, J., Ji, S.-H., Park, M., Lee, H.-S., Kim, S., Suh, S.-W., 2014. The attribute impact concept: Applications in case-based reasoning and parametric cost estimation. Autom. Constr. 43, 195–203. https://doi.org/10.1016/J.AUTCON.2014.03.011

Ahn, J., Park, M., Lee, H.-S., Ahn, S.J., Ji, S.-H., Song, K., Son, B.-S., 2017. Covariance effect analysis of similarity measurement methods for early construction cost estimation using case-based reasoning. Autom. Constr. 81, 254–266. https://doi.org/10.1016/J.AUTCON.2017.04.009

Akkaya, K., Guvenc, I., Aygun, R., Pala, N., Kadri, A., 2015. IoT-based occupancy monitoring techniques for energy-efficient smart buildings, in: 2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, pp. 58–63. https://doi.org/10.1109/WCNCW.2015.7122529

Al-Hader, M., Rodzi, A., 2009. The Smart City Infrastructure Development and Monitoring. Theor. Empir. Res. Urban Manag. 2, 87–94. https://doi.org/10.2307/24872423

Aljohani, A., Ahiaga-Dagbui, D., Moore, D., 2017. Construction Projects Cost Overrun: What Does the LiteratureTell Us? Int. J. Innov. Manag. Technol. 8, 137–143.

Asmar, M. El, Hanna, A.S., Whited, G.C., 2011. New Approach to Developing Conceptual Cost Estimates for Highway Projects. J. Constr. Eng. Manag. 137, 942–949. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000355

Barg, S., Flager, F., Fischer, M., 2018. An analytical method to estimate the total installed cost of structural steel building frames during early design. J. Build. Eng. 15, 41–50. https://doi.org/10.1016/J.JOBE.2017.10.010

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., Portugali, Y., 2012. Smart cities of the future. Eur. Phys. J. Spec. Top. 214, 481–518. https://doi.org/10.1140/epjst/e2012-01703-3

Beach, T.H., Rezgui, Y., Kasim, T., 2015. A rule-based semantic approach for automated regulatory compliance in the construction sector. Expert Syst. Appl. 42, 5219–5231. https://doi.org/10.1016/J.ESWA.2015.02.029

Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A., Pasha, M., 2016. Big Data in the construction industry: A review of present status, opportunities, and future trends. Adv. Eng. Informatics 30, 500–521. https://doi.org/10.1016/j.aei.2016.07.001

Boyd, D., Crawford, K., 2011. Six Provocations for Big Data, in: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. https://doi.org/10.2139/ssrn.1926431

Canto dos Santos, J.V., Farias Costa, I., Nogueira Thiago, 2015. New genetic algorithms for contingencies selection in the static security analysis of electric power systems. Expert Syst. Appl. 42, 2849–2856. https://doi.org/10.1016/J.ESWA.2014.11.041

Carr, R.I., 1989. Cost-Estimating Principles. J. Constr. Eng. Manag. 115, 545–551. https://doi.org/10.1061/(ASCE)0733-

9364(1989)115:4(545)

Chan, A.P.C., Scott, D., Chan, A.P.L., 2004. Factors Affecting the Success of a Construction Project. J. Constr. Eng. Manag. 130, 153–155. https://doi.org/10.1061/(ASCE)0733-9364(2004)130:1(153)

Chandwani, V., Gupta, N.K., Nagar, R., Agrawal, V., Jethoo, A.S., 2016. Artificial neural networks aided conceptual stage design of water harvesting structures. Perspect. Sci. 8, 151–155. https://doi.org/10.1016/j.pisc.2016.03.015

Chen, H.-M., Chang, K.-C., Lin, T.-H., 2016. A cloud-based system framework for performing online viewing, storage, and analysis on big data of massive BIMs. Autom. Constr. 71, 34–48. https://doi.org/10.1016/j.autcon.2016.03.002

Chen, J., Bulbul, T., Taylor, J.E.J., Olgun, G., 2014. A Case Study of Embedding Real-time Infrastructure Sensor Data to BIM. Constr. Res. Congr. 2014 269–278. https://doi.org/10.1061/9780784413517.028

Cheng, M.-Y., Hoang, N.-D., Wu, Y.-W., 2013. Hybrid intelligence approach based on LS-SVM and Differential Evolution for construction cost index estimation: A Taiwan case study. Autom. Constr. 35, 306–313. https://doi.org/10.1016/J.AUTCON.2013.05.018

Cheung, F.K.T., Rihan, J., Tah, J., Duce, D., Kurul, E., 2012. Early stage multi-level cost estimation for schematic BIM models. Autom. Constr. 27, 67–77. https://doi.org/10.1016/J.AUTCON.2012.05.008

Choi, S., Kim, D.Y., Han, S.H., Kwak, Y.H., 2014. Conceptual Cost-Prediction Model for Public Road Planning via Rough Set Theory and Case-Based Reasoning. J. Constr. Eng. Manag. 140, 04013026. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000743

Davila Delgado, J.M., Brilakis, I., Middleton, C.R., 2015. Open data model standards for structural performance monitoring of infrastructure assets, in: Beetz, J. (Ed.), CIB W78 Conference 2015. TU Eindhoven, Eindhoven, The Netherlands, pp. 1–10.

Davila Delgado, J.M., Butler, L.J., Brilakis, I., Elshafie, M.Z.E.B., Middleton, C.R., 2018. Structural performance monitoring using a dynamic data-driven BIM environment. J. Comput. Civ. Eng. 32.

Davila Delgado, J.M., Butler, L.J., Gibbons, N., Brilakis, I., Elshafie, M.Z.E.B., Middleton, C., 2017. Management of structural monitoring data of bridges using BIM. Proc. Inst. Civ. Eng. - Bridg. Eng. 170, 204–218. https://doi.org/10.1680/jbren.16.00013

Davila Delgado, J.M., de Battista, N., Brilakis, I., Middleton, C., 2016. Design and Data Modelling of Fibre Optic Systems to Monitor Reinforced Concrete Structural Elements, in: Construction Research Congress 2016. ASCE, San Juan, Puerto Rico, pp. 2291–2301. https://doi.org/10.1061/9780784479827.228

Davila Delgado, J.M., Hofmeyer, H., 2013. Automated generation of structural solutions based on spatial designs. Autom. Constr. 35, 528–541. https://doi.org/10.1016/j.autcon.2013.06.008

de Silva, N., Ranasinghe, M., de Silva, C.R., 2013. Use of ANNs in complex risk analysis applications. Built Environ. Proj. Asset Manag. 3, 123–140. https://doi.org/10.1108/BEPAM-07-2012-0043

Deng, S., Yeh, T.-H., 2011. Using least squares support vector machines for the airframe structures manufacturing cost estimation. Int. J. Prod. Econ. 131, 701–708. https://doi.org/10.1016/J.IJPE.2011.02.019

Dursun, O., Stoy, C., 2016. Conceptual Estimation of Construction Costs Using the Multistep Ahead Approach. J. Constr. Eng. Manag. 142, 04016038. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001150

Elbeltagi, E., Wefki, H., Abdrabou, S., Dawood, M., Ramzy, A., 2017. Visualized strategy for predicting buildings energy consumption during early design stage using parametric analysis. J. Build. Eng. 13, 127–136. https://doi.org/10.1016/J.JOBE.2017.07.012

ElMousalami, H.H., Elyamany, A.H., Ibrahim, A.H., 2018. Predicting Conceptual Cost for Field Canal Improvement Projects. J. Constr. Eng. Manag. 144, 04018102. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001561

Erl, T., Khattak, W., Buhler, P., 2016. Big data fundamentals : concepts, drivers & techniques, 1st Edition. ed. Prentice Hall.

Faghihi, V., Reinschmidt, K.F., Kang, J.H., 2014. Construction scheduling using Genetic Algorithm based on Building Information Model. Expert Syst. Appl. 41, 7565–7578. https://doi.org/10.1016/J.ESWA.2014.05.047

Faizollahzadeh Ardabili, S., Mahmoudi, A., Mesri Gundoshmian, T., 2016. Modeling and simulation controlling system of HVAC using fuzzy and predictive (radial basis function, RBF) controllers. J. Build. Eng. 6, 301–308. https://doi.org/10.1016/J.JOBE.2016.04.010

Firouzi, A., Yang, W., Li, C.-Q., 2016. Prediction of Total Cost of Construction Project with Dependent Cost Items. J. Constr. Eng. Manag. 142, 04016072. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001194

Gao, S., Li, L., Li, W., Janowicz, K., Zhang, Y., 2017. Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. Comput. Environ. Urban Syst. 61, 172–186. https://doi.org/10.1016/J.COMPENVURBSYS.2014.02.004

Gardner, B.J., Gransberg, D.D., Jeong, H.D., 2016. Reducing Data-Collection Efforts for Conceptual Cost Estimating at a Highway Agency. J. Constr. Eng. Manag. 142, 04016057. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001174

Geekiyanage, D., Ramachandra, T., 2018. A model for estimating cooling energy demand at early design stage of condominiums. J. Build. Eng. 17, 43–51. https://doi.org/10.1016/J.JOBE.2018.01.011

Gerrish, T., Ruikar, K., Cook, M.J., Johnson, M., Phillip, M., 2015. Attributing in-use building performance data to an as-built building information model for lifecycle building performance management, in: Beetz, J. (Ed.), Proceedings of the 32nd CIB W78 Conference. CIB, pp. 1–11.

Hajdasz, M., 2014. Flexible management of repetitive construction processes by an intelligent support system. Expert Syst. Appl. 41, 962–973. https://doi.org/10.1016/J.ESWA.2013.06.063

Han, K.K., Golparvar-Fard, M., 2017. Potential of big visual data and building information modeling for construction performance analytics: An exploratory study. Autom. Constr. 73, 184–198. https://doi.org/10.1016/J.AUTCON.2016.11.004

Hofmeyer, H., Davila Delgado, J.M., 2015. Coevolutionary and genetic algorithm based building spatial and structural design. Artif. Intell. Eng. Des. Anal. Manuf. 29, 351–370. https://doi.org/10.1017/S0890060415000384

Hofmeyer, H., Davila Delgado, J.M.M., 2013. Automated design studies: Topology versus One-Step Evolutionary Structural Optimisation. Adv. Eng. Informatics 27, 427–443. https://doi.org/10.1016/j.aei.2013.03.003

Horta, I.M., Camanho, A.S., 2013. Company failure prediction in the construction industry. Expert Syst. Appl. 40, 6253–6257. https://doi.org/10.1016/J.ESWA.2013.05.045

Hoult, N., Bennett, P.J., Stoianov, I., Fidler, P., Maksimović, Č., Middleton, C., Graham, N., Soga, K., 2009. Wireless sensor networks: creating 'smart infrastructure.' Proc. Inst. Civ. Eng. - Civ. Eng. 162, 136–143. https://doi.org/10.1680/cien.2009.162.3.136

Hwang, S., 2011. Time Series Models for Forecasting Construction Costs Using Time Series Indexes. J. Constr. Eng. Manag. 137, 656–662. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000350

Hwang, S., 2009. Dynamic Regression Models for Prediction of Construction Costs. J. Constr. Eng. Manag. 135, 360–367. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000006

Irani, Z., Kamal, M.M., 2014. Intelligent Systems Research in the Construction Industry. Expert Syst. Appl. 41, 934–950. https://doi.org/10.1016/J.ESWA.2013.06.061

Jadid, M.N., Idrees, M.M., 2007. Cost estimation of structural skeleton using an interactive automation algorithm: A conceptual approach. Autom. Constr. 16, 797–805. https://doi.org/10.1016/J.AUTCON.2007.02.007

Jafarzadeh, R., Ingham, J.M., Wilkinson, S., González, V., Aghakouchak, A.A., 2014. Application of Artificial Neural Network Methodology for Predicting Seismic Retrofit Construction Costs. J. Constr. Eng. Manag. 140, 04013044. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000725

Jeong, S., Hou, R., Lynch, J.P., Sohn, H., Law, K.H., 2019. A scalable cloud-based cyberinfrastructure platform for bridge monitoring. Struct. Infrastruct. Eng. 15, 82–102. https://doi.org/10.1080/15732479.2018.1500617

Jeong, S., Hou, R., Lynch, J.P., Sohn, H., Law, K.H., 2017. A Big Data Management and Analytics Framework for Bridge Monitoring, in: Structural Health Monitoring 2017. DEStech Publications, Inc., Lancaster, PA. https://doi.org/10.12783/shm2017/13862

Khan, A., Hornbæk, K., 2011. Big data from the built environment, in: Proceedings of the 2nd International Workshop on Research in the Large - LARGE '11. ACM Press, New York, New York, USA, p. 29.

https://doi.org/10.1145/2025528.2025537

Kim, G.-H., An, S.-H., Kang, K.-I., 2004. Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Build. Environ. 39, 1235–1242. https://doi.org/10.1016/j.buildenv.2004.02.013

Kim, S., Shin, D.H., 2016. Forecasting short-term air passenger demand using big data from search engine queries. Autom. Constr. 70, 98–108. https://doi.org/10.1016/J.AUTCON.2016.06.009

Klyne, G., Carroll, J., 2006. Resource Description Framework (RDF): Concepts and Abstract Syntax.

Koseleva, N., Ropaite, G., 2017. Big Data in Building Energy Efficiency: Understanding of Big Data and Main Challenges. Procedia Eng. 172, 544–549. https://doi.org/10.1016/J.PROENG.2017.02.064

Kumar, P., Skouloudis, A.N., Bell, M., Viana, M., Carotta, M.C., Biskos, G., Morawska, L., 2016. Real-time sensors for indoor air monitoring and challenges ahead in deploying them to urban buildings. Sci. Total Environ. 560–561, 150–159. https://doi.org/10.1016/J.SCITOTENV.2016.04.032

Larsen, J.K., Shen, G.Q., Lindhard, S.M., Brunoe, T.D., 2016. Factors Affecting Schedule Delay, Cost Overrun, and Quality Level in Public Construction Projects. J. Manag. Eng. 32, 04015032. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000391

Liberda, M., Ruwanpura, J., Jergeas, G., 2003. Construction Productivity Improvement: A Study of Human, Management and External Issues, in: Construction Research Congress. American Society of Civil Engineers, Reston, VA, pp. 1–8. https://doi.org/10.1061/40671(2003)5

Lowe, D.J., Emsley, M.W., Harding, A., 2006. Predicting Construction Cost Using Multiple Regression Techniques. J. Constr. Eng. Manag. 132, 750–758. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750)

Ma, Z., Wei, Z., Zhang, X., 2013. Semi-automatic and specification-compliant cost estimation for tendering of building projects based on IFC data of design model. Autom. Constr. 30, 126–135. https://doi.org/10.1016/J.AUTCON.2012.11.020

Martín-Garín, A., Millán-García, J.A., Baïri, A., Millán-Medel, J., Sala-Lizarraga, J.M., 2018. Environmental monitoring system based on an Open Source Platform and the Internet of Things for a building energy retrofit. Autom. Constr. 87, 201–214. https://doi.org/10.1016/J.AUTCON.2017.12.017

Meredith, J.R., Shafer, S.M., Mantel, S.J., Sutton, M.M., 2014. Project management in practice. Wiley.

Mousa, M., Luo, X., McCabe, B., 2016. Utilizing BIM and Carbon Estimating Methods for Meaningful Data Representation. Procedia Eng. 145, 1242–1249. https://doi.org/10.1016/j.proeng.2016.04.160

Olawale, Y.A., Sun, M., 2010. Cost and time control of construction projects: inhibiting factors and mitigating measures in practice. Constr. Manag. Econ. 28, 509–526. https://doi.org/10.1080/01446191003674519

Petroutsatou, K., Georgopoulos, E., Lambropoulos, S., Pantouvakis, J.P., 2012. Early Cost Estimating of Road Tunnel Construction Using Neural Networks. J. Constr. Eng. Manag. 138, 679–687. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000479

Roderick, J.A.L., Rubin, D.B., 2002. Statistical analysis with missing data, 2nd ed. Wiley.

Ryza, S., Laserson, U., Owen, S., Wills, J., 2015. Advanced analytics with Spark: Patterns for Learning from Data at Scale. O'Reilly.

Sagiroglu, S., Sinanc, D., 2013. Big data: A review, in: 2013 International Conference on Collaboration Technologies and Systems (CTS). IEEE, pp. 42–47. https://doi.org/10.1109/CTS.2013.6567202

Scheffer, J., 2002. Dealing with Missing Data. Res. Lett. Inf. Math. Sci 3, 153–160.

Semanjski, I., Gautama, S., Ahas, R., Witlox, F., 2017. Spatial context mining approach for transport mode recognition from mobile sensed big data. Comput. Environ. Urban Syst. 66, 38–52. https://doi.org/10.1016/J.COMPENVURBSYS.2017.07.004

Shane, J.S., Molenaar, K.R., Anderson, S., Schexnayder, C., 2009. Construction Project Cost Escalation Factors. J. Manag. Eng. 25, 221–229. https://doi.org/10.1061/(ASCE)0742-597X(2009)25:4(221)

Sonmez, R., 2008. Parametric Range Estimating of Building Costs Using Regression Models and Bootstrap. J. Constr. Eng. Manag. 134, 1011–1016. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:12(1011)

Sridarran, P., Keraminiyage, K., Herszon, L., 2017. Improving the cost estimates of complex projects in the project-based industries. Built Environ. Proj. Asset Manag. 7, 173–184. https://doi.org/10.1108/BEPAM-10-2016-0050

Sun, C., Han, Y., Feng, H., 2015. Multi-objective building form optimization method based on GANN-BIM model. Next Gener. Build. 2. https://doi.org/10.7480/NGB.2.1.1517

Suryadevara, N.K., Mukhopadhyay, S.C., Kelly, S.D.T., Gill, S.P.S., 2015. WSN-Based Smart Sensors and Actuator for Power Management in Intelligent Buildings. IEEE/ASME Trans. Mechatronics 20, 564–571. https://doi.org/10.1109/TMECH.2014.2301716

TCI, 2018. Top 100 Construction Companies 2017.

Trost, S.M., Oberlender, G.D., 2003. Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression. J. Constr. Eng. Manag. 129, 198–204. https://doi.org/10.1061/(ASCE)0733-9364(2003)129:2(198)

Tukia, T., Uimonen, S., Siikonen, M.-L., Hakala, H., Donghi, C., Lehtonen, M., 2016. Explicit method to predict annual elevator energy consumption in recurring passenger traffic conditions. J. Build. Eng. 8, 179–188. https://doi.org/10.1016/J.JOBE.2016.08.004

Venkatesan, K., Ramachandraiah, U., 2018. Climate responsive cooling control using artificial neural networks. J. Build. Eng. 19, 191–204. https://doi.org/10.1016/J.JOBE.2018.05.008

Wan, L., Gao, S., Wu, C., Jin, Y., Mao, M., Yang, L., 2018. Big data and urban system model - Substitutes or complements? A case study of modelling commuting patterns in Beijing. Comput. Environ. Urban Syst. 68, 64–77. https://doi.org/10.1016/J.COMPENVURBSYS.2017.10.004

Wang, D., Fan, J., Fu, H., Zhang, B., 2018. Research on Optimization of Big Data Construction Engineering Quality Management Based on RNN-LSTM. Complexity 2018, 1–16. https://doi.org/10.1155/2018/9691868

Wilmot, C.G., Mei, B., 2005. Neural Network Modeling of Highway Construction Costs. J. Constr. Eng. Manag. 131, 765–771. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:7(765)

Wilson, A., 2005. Experiments in probabilistic cost modelling, in: Skitmore, M., Marston, V. (Eds.), Cost Modelling. Routledge, pp. 169–180.

Xia, J., Yang, C., Li, Q., 2018. Using spatiotemporal patterns to optimize Earth Observation Big Data access: Novel approaches of indexing, service modeling and cloud computing. Comput. Environ. Urban Syst. https://doi.org/10.1016/J.COMPENVURBSYS.2018.06.010

Yang, C., Yu, M., Hu, F., Jiang, Y., Li, Y., 2017. Utilizing Cloud Computing to address big geospatial data challenges. Comput. Environ. Urban Syst. 61, 120–128. https://doi.org/10.1016/J.COMPENVURBSYS.2016.10.010

Yildiz, A.E., Dikmen, I., Birgonul, M.T., Ercoskun, K., Alten, S., 2014. A knowledge-based risk mapping tool for cost estimation of international construction projects. Autom. Constr. 43, 144–155. https://doi.org/10.1016/J.AUTCON.2014.03.010

Yu, W., Lai, C., Lee, W., 2006. A WICE approach to real-time construction cost estimation. Autom. Constr. 15, 12–19. https://doi.org/10.1016/J.AUTCON.2005.01.005

Zanella, A., Bui, N., Castellani, A., Vangelista, L., Zorzi, M., 2014. Internet of Things for Smart Cities. IEEE Internet Things J. 1, 22–32. https://doi.org/10.1109/JIOT.2014.2306328