

Requirements for the application of data sciences in construction

JA Wium¹ and CJ Jurgens²

^{1,2}Department of Civil Engineering
Stellenbosch University

^{1,2}Email: janw@sun.ac.za; cj@sun.ac.za

Abstract:

This paper investigates the use of data sciences in construction project management. A literature review identified limited use and application of data sciences in construction project management compared to other industries. Therefore, the paper addresses this gap by demonstrating the benefits of using data mining as a project management tool. It identifies the requirement for structured data capturing, facilitating the widespread application of data sciences in construction with demonstrated benefits. Following a description of the data mining methodology used, two cases are presented of the practical application of data mining on construction projects. In one case, predictions are made for the number of employment opportunities created in a project. In the other case, the total project duration is estimated/predicted. It is shown that improved results (in comparison with previously published cases) can be obtained from large databases with a 62.3% accuracy in one case study and 88.8% accuracy in the other. It also shows that the accuracy depends on the chosen algorithm. Similar to previous studies, the two cases identified the effort required to prepare suitable databases. This is a result of the current unstructured and uncoordinated capturing of data in the construction industry. It is recommended that companies develop standardised processes to capture data. It will be even more beneficial if such data can be stored in an industry-wide database to improve management processes and predictions. A current limitation to its application is the lack of knowledge in the industry about data sciences techniques.

Keywords: Data sciences, project efficiency, rework, project parameters

1 Introduction

Construction projects are prone to time and cost overruns and often suffer from delays due to various reasons (Mukuka, Aigbavboa and Thwala, 2015; Senouci, Ismail and Eldin, 2016; Niazi and Painting, 2017). Other demands are also sometimes placed on construction projects, such as creating employment opportunities (McCutcheon, 1995).

Data mining is effectively used in the communication, retail, insurance and medical sectors (among others) to reduce costs and improve business efficiency (Mcafee and Brynjolfsson, 2012; Pospieszny, 2017). A study by Bach, Bertoncel, Meško, Vugec and Ivančič (2020) amongst 3 European country clusters shows that the construction industry has some of the lowest usage of big data sources (6 - 15%) from a variety of other industries surveyed, with information and

communication (22 - 41%) and electricity, gas, air conditioning and water supply (14 - 36%) being the highest usage.

This paper aims to demonstrate the benefits of data sciences in managing construction projects and highlights the requirement for companies and the industry to standardise data capturing. The contribution of this paper is a recognition that structured data capturing and analysis is required to improve the management of construction projects. This can only be done if high volumes of data, which is mostly unstructured, are captured systematically.

The objectives are addressed by the paper layout as follows:

The paper first provides examples from the literature of data mining in construction project management and thereby identifies the need for further investigation. It then presents the methodology followed in two case studies, summarising the process of data mining. The paper concludes with a presentation of the outcome of two case studies completed as MEng research projects, where data mining is used effectively as project management tool to address employment creation and time management on projects.

2 Literature review: limited applications of data mining in construction

Fan, Abourizk and Kim (2008) determined the residual value of heavy construction equipment in the USA by using a national database of construction equipment auction results (LastBid) together with data mining techniques. They thereby developed a data-based decision support system which enables the better management of construction assets (equipment).

Lee, Hsueh and Tseng (2008) used a data mining process to obtain new knowledge from historic actions. They used a data mining algorithm to identify causes of poor quality in the building work of a large construction company to reduce ongoing maintenance costs. They successfully identified the three most important factors that caused leakage and concrete cracking in their buildings. The investigation therefore showed the benefits of using data mining to solve problems using existing project information.

Chaovalitwongse et al. (2012), used data mining and machine learning to create an application that would select the bid closest to the actual price of the project using project bid policies. By using bid data from the Texas and California Departments of Transport and a dataset of approximately 4000 projects, it was shown that data mining can be used for the evaluation of tender-selection policies. They also showed that data mining benefits are not limited to predictions or to knowledge generation only. A critical insight from the project highlighted the importance of normalising data or using data ratios to prevent overweighting of some features.

Williams and Gong (2014) used text mining together with numerical data to predict construction project cost overruns. The data was obtained from the California Department of Transport. It included numerical bid information but also a short, worded description of each project and its major cost items. Their investigation showed that more accurate predictions can be achieved if both numerical and text data is used in the database, as it allows for a more complete model to be constructed through the data mining. It was found that poor accuracy in predicting cost underruns

on projects is possibly due to the complexity of the construction environment or due to insufficient data that was used in the study.

Alsubaey, Asasi and Makatsoris (2015) used 46 files of project site meeting minutes to develop an early warning system on project outcomes by mining the unstructured text. They used the Naïve Bayes classifier to identify early signs of project risks materialising. Unfortunately, 86.95% of the outcomes were classified as “no-signal,” which was attributed to the classifier not being suited for the analyses of full sentences. Their case study does however show the potential of unstructured text as a data source. Text mining thus allows for the identification of valuable information that project personnel might oversee.

2.1 Synthesis:

The five examples above each demonstrated the potential of using data mining algorithms to support decision making and predictions in the construction industry. It is shown that data mining allows for the prediction of future outcomes, but also develops new knowledge from historic data.

A common thread from all studies was the recognition of the value of data, but also the need to develop practical and usable data sets. Limited data results in poor accuracy of prediction. Each study required considerable effort from researchers to prepare clean data sets from the available material. This crucial step might not happen in industry (no time), which highlights the requirement for correct data capturing the first time.

While the lessons learned from the five examples could be useful for construction project management, it was not the specific aim of these cases. Of the example studies presented above, there are limited cases of data mining being used to directly improve construction project management. The gaps identified in the existing literature which this article therefore seeks to address are the value of large data bases, the application of data sciences in construction project management and the nature of available data.

3 Methodology of the study

The process of data mining considers the use of large datasets from which computer algorithms recognise patterns and relationships between variables which may not be easily recognisable by the human mind. This process provides support for decision making based on the data (Syvajarvi and Stenvall, 2010). The methodology described below was used in two case studies to investigate the process, the possible application and/or benefits, and finally the requirements for data sciences in construction project management.

Throughout the different phases of a construction project, large amounts of data are generated and stored in a variety of ways and formats. This could include drawings, bills of quantities, contract documents, meeting minutes, instructions, requests for information, Building Information Models and a multitude of other sources.

For each of the two data mining studies below a clear definition of the goal was defined. In the first case it was to determine the number of employment opportunities created, and in the second case it was to determine final project duration during the execution of the project.

For each study pre-processing of the available data was required to produce a clean dataset. The process started with a thorough examination of the goals of the application based on the available data. This was then followed by three steps:

- Extracting of data that can model the system accurately;
- Cleaning of data from omissions, errors, and inconsistencies;
- Removal of excess data and/or transformation of data where needed.

The choice of the data mining model and data mining technique was made by a specialist with knowledge of statistics and machine learning. While data mining applications do not typically use all the available methods, the choice of application was dependent on the type of data and the goal. Aspects considered for the choice of the application included (as identified by Aggarwal, 2015):

- Similarity and distance functions;
- Association pattern mining;
- Cluster analysis;
- Outlier analysis;
- Classification and regression;
- Text mining.

The accuracy of the results either were validated or evaluated, depending on the nature of the process used, being either supervised or unsupervised. Clustering and outlier analyses were validated, whilst classification processes were evaluated. Criteria were defined with items such as coefficient of determination, mean absolute error, and root mean square error (see also Lee, S., Kim, C., Son, H., and Kim. C., 2011; Han, Kamber and Pei, 2012; Srivastava, 2014). For each study the data was divided into a training data set (80% and 67% of the data respectively for the two studies) and a testing data set (remainder of the data).

Several software packages are available that assist with the application of data mining. This allows a user to apply the techniques with relative ease. In the first case the Scikit-Learn Python module was used with the application of six data mining algorithms. In the second case Orange and Rapid Minor were chosen with similar algorithms to create prediction models. The outcome is described below.

4 Two case studies

Under the next two headings case studies are presented where the above methodology was applied to different aspects of construction project management. It is shown that there are real-life benefits for applying data mining methodologies to construction projects. More importantly however it also identifies the requirements that are needed for the widespread adoption of data sciences in construction.

Both of these case studies formed part of two MEng research projects (referenced below) for the qualification.

4.1 Predicting the employment opportunities in South African construction projects

South Africa had an official unemployment rate of 32.5% in the 4th quarter of 2020 (Stats SA, 2021). The South African government continues to address the problem by providing/regulating temporary employment opportunities in the construction industry.

To assist projects in successfully meeting the requirement of unskilled employment opportunities, Botha (2018) used data mining techniques to estimate at the initiation and/or planning phases of a project, the total temporary employment opportunities that would be created.

Botha (2018) used project details from 755 different road construction and rehabilitation projects of the Construction and Maintenance Directorate of a provincial government body.

Whilst the data base consisted of much more information, for the specific study he selected the following features:

- District (one of the districts in the Western Cape Province in South Africa)
- Project objective (such as ‘resurfacing’ which could be one of thirteen objectives)
- Tender cost estimate
- Number of created employment opportunities

Botha (2018) created a dataset after pre-processing the available data by normalising, cleaning and transforming data from 475 road construction or maintenance projects. Considerable effort was required to clean the available data. Cleaning of the data improved the R^2 value from 0.377 for the uncleaned data to 0.427 for cleaned data. The opportunities created versus the tender cost estimate is shown in Figure 1 for the cleaned dataset.

Using the above dataset, he evaluated six popular data mining algorithms as found in the Scikit-learn module of Python. The evaluation consisted of randomly splitting the dataset into training (80%) and testing (20%) parts. The process was repeated ten times on different random splits of the data.

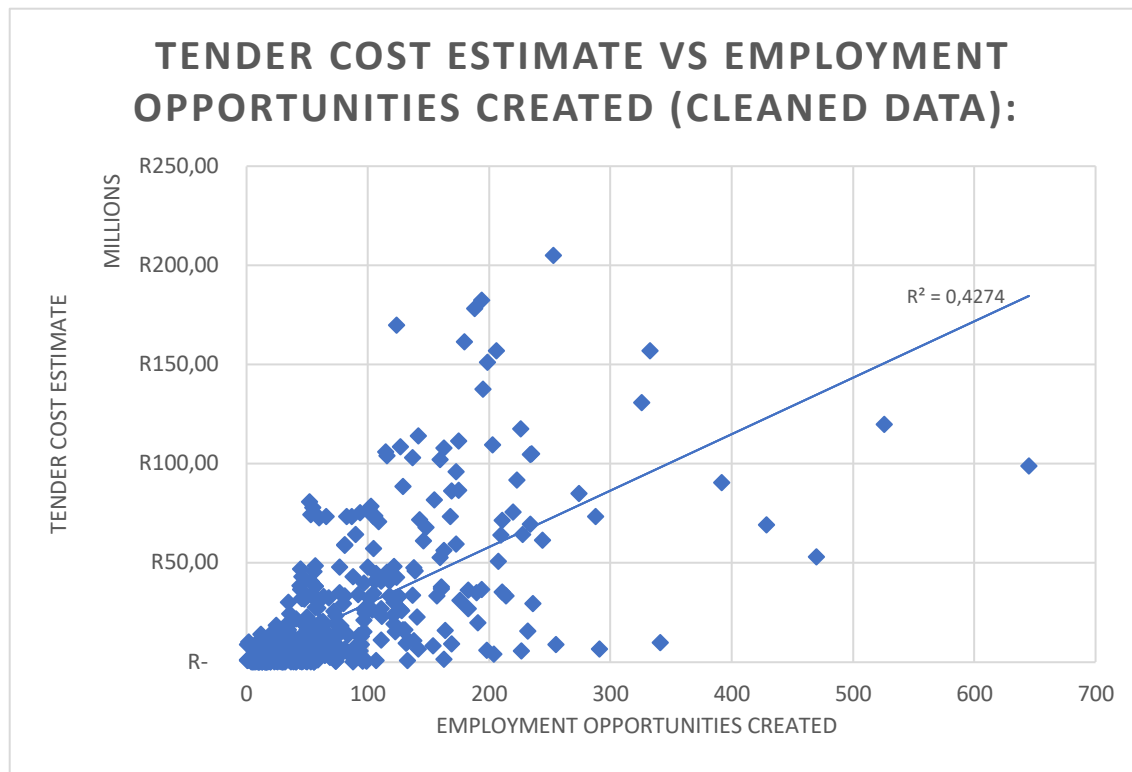


Figure 1. Tender cost estimate versus Employment opportunities (after cleaning dataset)(Botha, 2018)

Results from the four best performing algorithms are shown in Table 1. Botha (2018) found that the Naïve Bayes (Gaussian) classifier was the most accurate for his case study, predicting the correct class with 62.3% accuracy. His results compare favourably with the results of Williams and Gong (2014), who only managed to correctly predict the cost overrun of projects 44% of the time (probably due to their limited dataset).

Table 1. Prediction accuracies of different algorithms (Botha, 2018)

Algorithm	Total Accuracy
Naïves Bayes (Gaussian) Classifier	62.3%
K Nearest Neighbour Classifier	60.7%
Support Vector machine Classifier	58.3%
Neural Network Classifier	55.9%

The outcome of the study was that reasonable “ball park” estimates of employment opportunities can be made from sufficiently populated historic data. It also showed that cleaning of the data requires considerable effort, thus highlighting the requirement of having a well-structured data collection method to improve prediction accuracy but also reduce cleaning time. Lastly, the study found that some specialist knowledge is needed for the application of the various data mining algorithms.

4.2 Using project site meeting minutes to predict project duration

A study was performed to investigate data science's ability to timely predict the total project duration of construction projects.

The data for this study was obtained from construction project site meeting minutes (Van Niekerk, 2020). For this purpose, meeting minutes from 27 construction projects were obtained from a South African contractor. The projects were carried out over the last 10 years. Considerable effort was required for creating a suitable database for this study given the relatively unstructured nature of site meeting minutes. The data shown in Table 2 was manually extracted from the documents and entered into +500 rows of data in an Excel database.

Table 2. Time related factors recorded in the data warehouse (Van Niekerk, 2020)

Time related factors	Non-time related factors
% Progress	Meeting number
% Progress planned	Date
Expected completion date	Number of contractor employees at the meeting
Progress notes	Number of client representatives at the meeting
Original/contractual completion date	Health, safety, environment and community
Days ahead or behind programme	Security and IR
Class of progress	Quality
Weather delays	Contractual/commercial
Design delays	Number of FEQ's
Planning	Number of SI's
Project complete late	Early warning from contractor
	Early warning from client
	Number of compensation events
	Project man-hours
	Rework mentioned
	Repeat items on the minutes
	Meeting number

The data mining process was executed by applying the methodology as presented under Section 3 above, in conjunction with two freely available data mining applications. The objective was to determine the prediction accuracy of the two applications.

Training and testing datasets were set up by splitting the original dataset into 2/3 and 1/3 groups respectively. One hundred repetitions were performed during the model's training, each time splitting the dataset randomly between training and testing groups.

A variety of data mining algorithms were evaluated from each application. Table 3 shows the best results obtained for both applications, which used the same algorithms and yet gave different results for the two applications. It can be seen that the accuracy of the results is better than that of the investigation by Botha (2018), probably due to the larger dataset.

Table 3. Accuracy of results for Datasets A using Orange and Rapidminer (Van Niekerk, 2020)

	Orange	RapidMiner
Model type	Accuracy for Dataset A	Accuracy of Dataset A
Random Forest	88.8%	81.7%
Neural Network	85.3%	79.2%
Tree	78.6%	72.5%
Naïves Bayes	72.6%	72.5%

The outcome of the study demonstrated the benefit of using large amounts of data to make predictions of project outcomes (project programme). However, it was again realised that the process can only be practically implemented if data is captured systematically and in a well-defined structure. It was learned that without a defined structure for capturing data on construction projects, the widespread application of data sciences will most likely not happen.

5 Conclusion

This paper investigates the practical application of data mining, a sub-set of data sciences, as a management tool for the construction industry.

The cases identified in literature applied data sciences on industry data (equipment) and on project data. It is shown that data mining allows for the prediction of future outcomes, thereby developing new knowledge from historic data. In cases where the technique was applied to project data, it was seen that limited data resulted in relatively poor accuracy of prediction.

The first case demonstrates the value of historical data to find trends in and to predict job creation in public projects. It shows that a large database is needed, but when available it can be of much value to make suitable predictions.

The second case shows how project site meeting records can be used to predict project duration. The practical implication of the study is that with early warning information, project teams can take corrective actions at any time during the project.

Together with the cases from literature, the two studies emphasised that data sciences is a powerful technique that can be used by project teams for predictions or for improving overall project efficiencies. It is shown in this paper that the accuracy can be considerably better when having a database with large volumes of data. As in literature, it also identifies the effort in creating suitable databases from unstructured data, and thus emphasises the need for construction organisations to develop systems for capturing large volumes of structured data.

The real contribution of this paper is a recognition that the benefits of data mining in the construction industry is restricted by the current unstructured capturing of data.

At present the construction industry is shown to make little use of data sciences when compared to other industries, most probably due to the effort needed to create usable databases. Well-structured data is essential for the successful application of data mining techniques.

It is recommended that organisations standardise the format and data of items to be captured in their projects, starting with project site meetings minutes. A lack of expertise in data science techniques can however be a current restriction to the widespread application. This may require further investigation.

6 Acknowledgements

The authors wish to express their appreciation for the organisations and individuals who shared their information and knowledge for the study of the two case studies described in this paper.

7 References

- Aggarwal, C. C. (2015) 'Data Mining: The Textbook', *Springer International Publishing*. doi: 10.1007/978-3-319-14142-8.
- Bach, Bertoncel, Meško, Vugec and Ivančić. 'Big Data Usage in European Countries: Cluster Analysis Approach', *Data* 2020, 5, 25; doi:10.3390
- Botha, L.J. (2018) 'Data Mining Construction Project Information to aid Project Management', *MEng thesis, Stellenbosch: Stellenbosch University*, 2018.
- Fan, H., Abourizk, S. and Kim, H. (2008) 'Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model', 22(3), pp. 181–191.
- Han, J., Kamber, M. and Pei, J. (2012) 'Data Mining: Concepts and Techniques', 3rd edn, *San Francisco, CA, Morgan Kaufmann Publishers*. doi: 10.1016/B978-0-12-381479-1.00001-0.
- Lee, J., Hsueh, S. and Tseng, H. (2008) 'Utilising data mining to discover knowledge in construction enterprise performance records', *Journal of Civil Engineering and Management*, 14(2), pp. 79–84. doi: 10.3846/1392-3730.2008.14.2.
- Lee, S., Kim, C., Son, H., and Kim, C. 'Data mining-based predictive model to determine project financial success using project definition parameters', *28th International Symposium on Automation and Robotics in Construction, ISARC 2011*, pp. 473–478. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84863753135&partnerID=40&md5=0f5397513060366252a2792f36cecc00>.
- Mcafee, A. and Brynjolfsson, E. (2012) 'Big Data: The Management Revolution', *Harvard*

Business Review, October. Available at: <https://hbr.org/2012/10/big-data-the-management-revolution>.

- McCutcheon, R. (1995) 'Employment creation in public works: Labour-intensive construction in sub-saharan Africa: The implications for South Africa', *Habitat International*, 19(3), pp. 331–335. doi: [https://doi.org/10.1016/0197-3975\(95\)00001-V](https://doi.org/10.1016/0197-3975(95)00001-V)
- Mukuka, M., Aigbavboa, C. and Thwala, W. (2015) 'Effects of Construction Projects Schedule Overruns: A Case of the Gauteng Province, South Africa', *Procedia Manufacturing*. Elsevier B.V., 3(Ahfe), pp. 1690–1695. doi: 10.1016/j.promfg.2015.07.989.
- Niazi, G. A. and Painting, N. (2017) 'Significant Factors Causing Cost Overruns in the Construction Industry in Afghanistan', *Procedia Engineering*. The Author(s), 182, pp. 510–517. doi: 10.1016/j.proeng.2017.03.145.
- Pospieszny, P. (2017) 'Application of Data Mining Techniques in Project Management – an Overview', *CEA Annals*, (43), pp. 199–220.
- Senouci, A., Ismail, A. and Eldin, N. (2016) 'Time Delay and Cost Overrun in Qatari Public Construction Projects', *Procedia Engineering*. The Author(s), 164(June), pp. 368–375. doi: 10.1016/j.proeng.2016.11.632.
- Srivastava, S. (2014) 'Weka-01: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining', *International Journal of Computer Applications*, 88(10). doi: 10.5120/15389-3809.
- Stats SA (2021) 'Quarterly Labour Force Survey, Quarter 4: 2020', Department: Statistics South Africa. Pretoria, February 2021.
- Syvajarvi, A. and Stenvall, J. (2010) 'Data Mining in Public and Private Sectors: Organisational and Government Applications', , 1st edn. Edited by K. Klinger et al. Hershey: Information Science Reference. Available at: https://books.google.com/books?id=sI2fcLf_rV8C.
- Van Niekerk, J. (2020) 'A Study of the data mining of meeting minutes of construction projects', Stellenbosch: Stellenbosch University, 2018.
- Williams, T. P. and Gong, J. (2014) 'Predicting construction cost overruns using text mining, numerical data and ensemble classifiers', *Automation in Construction*. Elsevier B.V., 43, pp. 23–29. doi: 10.1016/j.autcon.2014.02.014.