# TOWARDS AUTOMATED COST ANALYSIS, BENCHMARKING AND ESTIMATING IN CONSTRUCTION: A MACHINE LEARNING APPROACH

Daqing Chen
*School of Engineering, London South Bank University*
*103 Borough Road, London SE1 0AA, UK*

Laureta Hajderanj
*School of Engineering, London South Bank University*
*103 Borough Road, London SE1 0AA, UK**

James Fiske
*Royal Institution of Chartered Surveyors*
*12 Great George Street, London, SW1P 3AD, UK*

**ABSTRACT**

In this paper, a novel machine learning based approach is proposed for automated cost analysis from priced bill of quantities prepared by tenders in the construction industry. The proposed approach features: 1) An effective integration of structured project-specific information with surveyor's domain knowledge in order to model the complex interrelationships between the specifications and descriptions of an item and its trade category; 2) An effective transformation to map the original data into a 2-dimensional space to tackle issues of high dimensionality in modelling, and 3) Simply classifiers with good classification capability. Relevant comparative experimental results have demonstrated the effectiveness of the proposed approach.

**KEYWORDS**

Construction Cost Benchmarking, Cost Analysis Construction Data Analysis, Bill of Quantities, Dimensionality Reduction, Supervised *t*-SNE.

## 1. INTRODUCTION

The ability to extract information and analyze it into a consistent format is essential in the construction industry in order to be able to learn from and to benefit future projects. In particular, historical project cost data is essential for early stage estimating and to be able to benchmark projects against others to identify areas to investigate, improve to demonstrate that value for money is being achieved.

The challenge is that construction projects are all procured differently with inconsistent reporting data, so data analysis is becoming more time consuming, often error prone and inconsistent as it heavily depends on an individual doing the analysis. Due to the time-consuming nature of the activity, the process of analyzing projects is often left to more junior roles, if done at all. Therefore, the amount of data that the industry has to learn from is becoming less in terms of samples and of quality so anything that solves this challenge is of great importance to the construction industry.

In this paper, a novel machine learning based approach is proposed to address the problem. The focus is placed on developing a knowledge base to represent the interrelationships among specifications and descriptions of items and their analysis categories for analysis of bill of quantities (BQ) into trades. Due to the measurement rules adopted in the projects available, SMM7 (Standard Method of Measurement) is used as a basis to create the knowledge base, and domain knowledge of experienced surveyors has also been integrated

into the knowledge representation effectively. As a result, a set of instances is generated. A supervised *t*-SNE algorithm (*t*-distributed stochastic neighbor embedding) is implemented to visualize the instances and map them into a 2-dimensional space. Consequently, a classifier can be trained to classify (assign) an identified item in a BQ into an appropriate trade category in this space, and the classifier itself effectively provides a form of knowledge representation. For comparison purposes, various classifiers have been created including multilayer perceptron (MP), *k*-nearest neighbors (*k*NN), decision tree, and Naïve Bayesian models, and all the models have demonstrated a high classification accuracy consistently.

Note that there are many different potential input data formats and structures from different procurement routes and measurement systems (e.g., NRM2, HMM, RMM, Uniclass, Uniformat, Masterformat etc.) and also many different analysis reporting formats (e.g. trade based, elemental based using the SFCA and international reporting standards such as ICMS etc.) (Gelder, J.E., 2015). This paper focuses on just one option in order to test the hypothesis and examines the analysis of SMM7 BQ into trades. In addition, how to identify items in different document formats, such as PDF, Excel, XML etc., is beyond the scope of this paper.

The remainder of the paper is organized as follows. Section 2 provides a brief critical review on the literature about machine learning based approaches to trade classification in priced BQ. Section 3 gives in detail the proposed methodology for effective trade classification. It also discusses how training samples can be created based on SMM and the domain knowledge of experienced surveyors. The training samples are visualized and mapped using *t*-SNE, and classification in the low-dimensional space is carried out with a number of classifiers for comparison purpose. This is followed by a discussion on the experimental findings in Section 4. And finally, in Section 5 a conclusion is given along with suggested future work.

## 2. A BRIEF CIRTICAL REVIEW ON RELATED WORK

The essence of a machine learning based approach to automate cost analysis lies in understanding and modelling the interrelationships among item specifications and the corresponding category of the item. On the one hand, item specifications and descriptions can be well-structured in a certain syntax, such as in SMM which specifies an item with respect to its category by using a set of codes structured hierarchically at several levels. On the other hand, uncertainties exist due to the fact that a same item could be described by different surveyors inconsistently with various wordings, particularly at a very low level of the description. Therefore, a good modelling of the relationships between an item's descriptions and its category should capture both structured and uncertain determining factors. In addition, a sufficient support of experienced surveyors is crucial in the modelling process.

Various machine algorithms have been considered in the literature for the modelling. In particular, case-based reasoning (CBR) was applied in several case studies (Kim, G.H., *et al.*, 2004, Ji, S.-H., *et al*., 2011, Zima, K., 2015, Lésniak, A., and Ziam, K, 2018). CBR is based on a simple philosophy that a construction project or work of similar type is more likely to have a similar cost. The approach is easy to implement; however, it heavily depends on the similarity measure used and how many similar cases to be considered. Usually Euclidean distance is employed as similarity measure. This measure may not work well in a high dimensional space which is often the case for cost analysis since most of the variables involved are of a categorical type.

Other algorithms recommended include regression analysis, MP, and support vector machine (Arage, S.S., Dharwadkar, N.V., 2017, Petruseva, S., *et al*., 2017). A comparison was made among CBR, MP, and regression algorithms by Kim, G.H., *et al.*, (2004). Although regression analysis is a typical technique, it may not be appropriate for cost analysis of BQ, since many regression variables are categorical type with each having many distinct values resulting in a very complex model. One of the main benefits of using MP model is that it has a powerful mapping capability as a universal approximator.

Several non-machine learning approaches have been proposed, such as ontology-based approaches (Lee, S.-L., *et al.*, 2013, Ma, Z., *et al.*, 2013) and OLAP cubes enabled cost analysis (Martınez-Rojas, M., *et al.*, 2015, Martınez-Rojas, M., *et al.*, 2016).

As a summary, although a range of machine learning models have been applied in different scenarios for cost analysis and estimation, there is a clear lack of adopting a modelling strategy that integrates structured information on item specifications and its classification category with domain knowledge of experienced participating surveyors to tackle uncertainties in classifying items. As discussed in a survey by Elfaki, A.O., *et*

*al*. (2014), there are two types of factors to be considered and incorporated in cost estimation: design-specific and estimator-specific, the former is well-defined and established, and the latter lacks standardization. In addition, the algorithms chosen should best fit the heterogeneous data sets associated.

# 3. METHODOLOGY

In the paper, the emphasis is placed on trade-based cost analysis from an SMM7 BQ. An integrated modelling strategy has been adopted which incorporates SMM-based structured information on item descriptions and their trade categories with the domain knowledge of experienced surveyors to tackle uncertainties in BQ. A supervised *t*-SNE algorithm has been chosen to transform the high dimensional BQ data into a 2-dimensional space, and further different classifiers can be built in this space.

## 3.1 Data Capture and Understanding

Capturing the relevant data for the modelling has been conducted by extraction of regulations from SMM7 and interviews with expert surveyors in the industry. The data essentially contains a set of rules that define and describe an item and its corresponding trade group. In general, an item is specified by three levels of specifications hierarchically using certain codes or textual descriptions as shown in Table 1. The resultant data set has 1245 rules (instances), and the number of unique values of the three levels 1, 2, and 3 are 7, 305, and 290, respectively. The total number of trades (classes) is 42. As all the variables in the data set are of categorical type, one-hot encoding transformation has been employed to transform the data, resulting in a target data set with 603 variables in total including the variable trade.

**Table 1.** Example of data description

| Trade | Item Code and Description at each Hierarchical level (Number of Unique Values) | | |
|---|---|---|---|
| | Level 1 (7) | Level 2 (305) | Level 3 (290) |
| Excavating | N/A | D11, D12, D20, D21, D40, D41 | excavation, excavating, excavate; working space; earthwork support; reduce levels; disposal; surface treatments. |
| Brickwork | N/A | F10, F11 | brickwork; common bricks; facing bricks; engineering bricks; walls; piers; projections; arches; closing cavity/cavities; facework |
| Miscellaneous | adaptation; alteration; builder's work; demolition | N/A | N/A |
| Paths | external | D20, E05, E20, E30, E40, Q10, Q20-26 | excavation, excavating, excavate; working space; earthwork support; reduce levels; disposal; surface treatments; filling; filling to make-up levels; in-situ concrete, beds; reinforcement; formwork; paving. |

The number of samples for each trade group is shown in Figure 1. Note that although many of the trade groups have only a few samples, treating them as outliers may not be appropriate as they are all valid from real BQ, and more importantly, in terms of cost, a small-sized trade group may attribute significantly.

## 3.2 Modelling

The target data set was mapped into a 2-dimensional space using the chosen supervised *t*-SNE. The transformation was considered necessary due to the high number of dimensions in the original space that could make classification difficult, and algorithms like *k*NN may not work appropriately because of the use of Euclidean distance for similarity measure. In addition, it provides an effective visualization of the data set. Figure 2 shows the mapped data instances with their trade labels using the original *t*-SNE and the supervised one, respectively. Apparently, instances of the same trade group have been made much more compacted by the

supervised *t*-SNE algorithm, and this could potentially facilitate the *k*NN. The settings of the supervised *t*-SNE were as follows: Perplexity 30, Number of iterations: 150, Learning rate: 100, and Momentum: 0.5.

The mapped data was used for modelling. Six models were built for the classification for comparison purposes: A multilayer perceptron model of single hidden layer with 2 input nodes, 20 hidden nodes, and 42 outputs nodes, respectively, and the inputs to the network was normalized using *z*-score standardization; A binary decision tree with 6 depths; A naïve Bayesian model, and *k*NNs with various number of nearest neighbors of 1, 5, and 10, respectively. The target data set was split randomly into three subsets of training, validation, and test with each having 33.3% of the samples. The MP model was trained with 1000 iterations and a set of randomly-selected initial connection weights and biases from an even distribution.

The experimental results are given in Table 2.

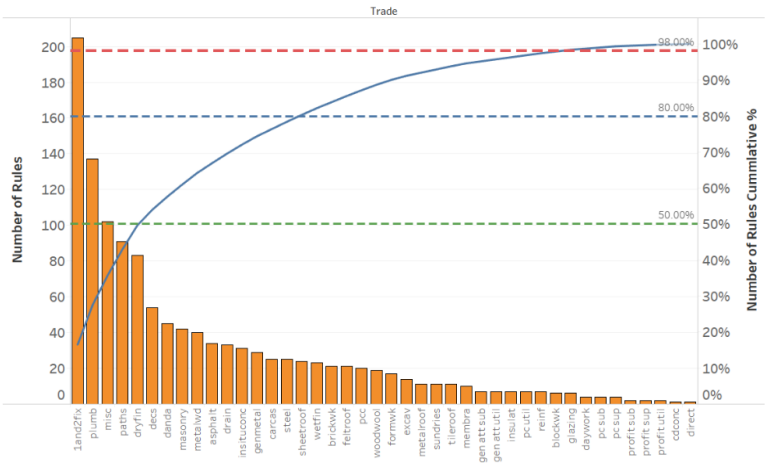


**Fig. 1.** The absolute and percentage cumulative number of instances (rules) for each trade group.

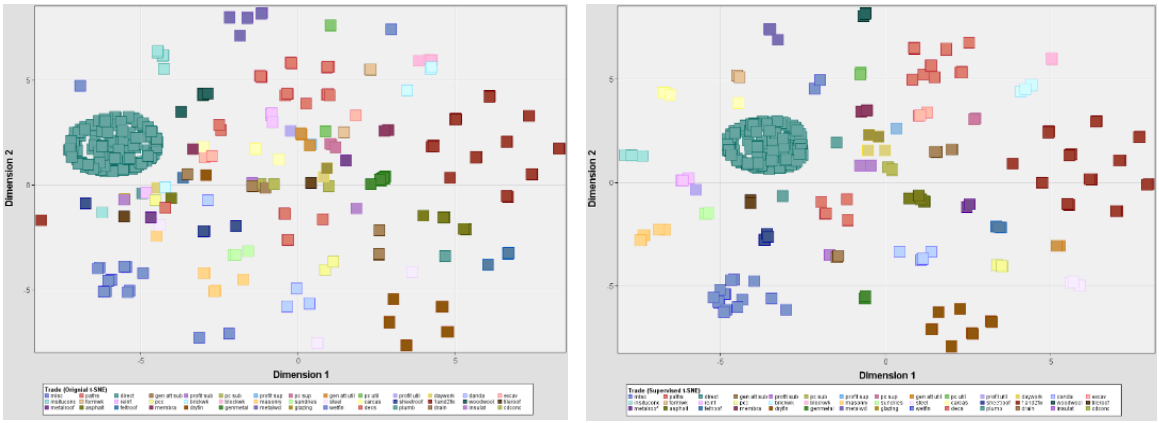


**Fig. 2.** Mapping the data from the original 602-dimesional space to a 2-dimensional space using original *t*-SNE (left) and supervised *t*-SNE (right). 42 trade groups considered.

**Table 2.** Classification accuracy by % (Total number of trade groups: 42)

| Dimensionality Reduction algorithm | Model | Multilayer Perception | Binary Decision Tree | Naïve Bayesian | *k*-Nearest Neighbors | | |
|---|---|---|---|---|---|---|---|
| | | | | | *k=1* | *k=5* | *k=10* |
| None | Training | 98.02 | 56.64 | 70.12 | 99.01 | 55.80 | 53.82 |
| | Validation | 48.52 | 45.30 | 56.69 | 43.07 | 46.34 | 43.87 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Test | 45.62 | 44.50 | 54.36 | 41.97 | 41.28 | 43.17 |
| Original *t*-SNE | Training | 96.30 | 86.42 | 90.12 | 100 | 92.71 | 84.67 |
| | Validation | 91.83 | 58.17 | 86.14 | 97.78 | 89.11 | 82.18 |
| | Test | 91.06 | 54.13 | 84.40 | 97.48 | 89.91 | 82.11 |
| Supervised *t*-SNE | Training | 97.28 | 97.28 | 95.80 | 100.00 | 96.23 | 92.96 |
| | Validation | 97.03 | 98.51 | 95.55 | 99.97 | 96.03 | 94.31 |
| | Test | 96.33 | 97.25 | 95.64 | 100.00 | 93.80 | 92.43 |

# 4. ANALYSIS AND DISCUSSION

From the simulation results a phenomenon has been observed: it is evident that all the models trained with the data mapped by the supervised *t*-SNE have achieved a consistently high classification accuracy with a good generalization capability. Contrastively, all the models trained by the original *t*-SNE have demonstrated a low generalization capability. In addition, all the models created in the original space have performed poorly with the lowest classification accuracy across all the model groups. These findings suggest that the proposed transformation using the supervised *t*-SNE was effective and promising.

Transforming data into a 2-dimensional space has also made it possible that a simple model can be good enough for a classification problem. Note that in the experiments, the simplest model, i.e., the *1*-nearest-neighbor classifier has shown the best performance.

On the other hand, due to the dimensionality reduction, the number of parameters whose values need to be optimized through training can be reduced significantly, especially for models like multilayer perceptron. Consider, for example, a multilayer perceptron with a single hidden layer. The network has $n$ input nodes, $m$ hidden nodes, and $l$ output nodes, respectively. Hence, the total number of the parameters that needs to be optimized is $m(n + l + 2)$. However, if the proposed transformation is used, the total number will be $m(4 + l)$ only. This will be beneficial for a modelling problem that lacks enough data.

From a modelling perspective, analyzing data into trades from BQ has posed a challenge about how to deal with a classification problem that involves a very high number of dimensions (over several hundred to several thousand) and a high number of classes coupled with imbalanced number of samples from each class.

# 5. CONCLUDING REMARKS

In this paper a novel approach has been proposed for cost analysis from BQ. The approach integrates structured information with surveyor's domain knowledge about the relationships between an item's descriptions and its trade membership. A supervised dimensionality reduction algorithm has been employed to transform the original data into a 2-dimensional space for classification. This has made a simple model work well with a high classification accuracy. The relevant experiments have demonstrated the effectiveness of the proposed approach.

Further research includes:

- To use real samples of priced BQ from tenders to conduct sensitivity analysis to identify which items are cost-sensitive and attribute considerably to the cost of BQ. The knowledge of this can be integrated into the objective function of the classification model.
- To explore how to interpret the mapped data created by *t*-SNE. From a business perspective, a model that can be easily interpreted would be useful. Currently, *t*-SNE has a poor interpretability.
- To consider how to accommodate a group of words appearing in BQ instead of only a single word. This could make the analysis more practical.
- To address class uncertainty problems since there are some rules that have identical descriptions but are associated with different trade groups.
- To then expand this research into different input formats and analysis structures.

# REFERENCES


Arage, S.S., Dharwadkar, N.V., 2017, Cost Estimation of Civil Construction Projects using Machine Learning Paradigm. *International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, Palladam, India, pp. 594 – 599.

Elfaki, A.O., et al., 2014, Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey. *Advances in Civil Engineering*, Vol. 2014, pp. 8 – 19.

Gelder, J.E., 2015, The design and development of a classification system for BIM. *Building Information Modelling (BIM) in Design, Construction and Operations*, Vol. 149, pp. 477 - 491.

Hajderanj, L., et al., 2019, A New Supervised *t*-SNE with Dissimilarity Measure for Effective Data Visualization and Classification. *2019 8th International Conference on Software and Information Engineering (ICSIE 2019)*, Cairo, Egypt. In print.

Ji, S.-H., et al., 2011, Cost estimation model for building projects using case-based reasoning. *Canadian Journal of Civil Engineering*, Vol. 38, No. 5, pp.570 – 581.

Kim, G.H., et al., 2004, Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, Vol. 39, pp. 1235 – 1242.

Lèsniak, A., and Ziam, K, 2018, Cost Calculation of Construction Projects Including Sustainability Factors Using the Case Based Reasoning (CBR) Method. Sustainability, Vol. 10, No. 5, p.1608.

Lee, S.-L., et al., 2013, BIM and ontology-based approach for building cost estimation. *Automation in Construction*, Vol. 41, pp. 96–105.Martınez-Rojas, M., et al., 2015, Cost Analysis in Construction Projects using Fuzzy OLAP Cubes. *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Istanbul, Turkey, pp. 96 – 105.

Ma, Z., et al., 2013, ONTOLOGY-BASED COMPUTERIZED REPRESENTATION OF SPECIFICATIONS FOR CONSTRUCTION COST ESTIMATION. the 30thCIB W78 International Conference, Beijing, China, pp.9 –12.

Martínez-Rojasa, M., et al., 2016, An intelligent system for the acquisition and management of information from bill of quantities in building projects. *Expert Systems with Applications*, Vol. 63, pp. 284 – 294.

Petruseva, S., et al., 2017, CONSTRUCTION COSTS FORECASTING: COMPARISON OF THE ACCURACY OF LINEAR REGRESSION AND SUPPORT VECTOR MACHINE MODELS. *Technical Gazette*, Vol. 24, No. 5, pp. 1431 – 1438.

Zima, K., 2015, The Case-Based Reasoning Model Of Cost Estimation At The Preliminary Stage Of A Construction Project. *Procedia Engineering*, Vol. 122, pp. 57 – 64.