A Method to Enable Automatic Extraction of Cost and Quantity Data from Hierarchical Construction Information Documents to Enable Rapid Digital Comparison and Analysis

Daniel Adanza Dopazo *, Lamine Mahdjoubi, Bill Gething

Posted Date: 17 August 2023

doi: 10.20944/preprints202308.1237.v1

Keywords: data mining; data extraction; data science; cost infrastructure projects

Article

A Method to Enable Automatic Extraction of Cost and Quantity Data from Hierarchical Construction Information Documents to Enable Rapid Digital Comparison and Analysis

Daniel Adanza Dopazo, Lamine Mahdjoubi and Bill Gething

Centre for Architecture and Built environment research, Coldharbour Ln, Stoke Gifford, Bristol BS16 1QY, UK

Abstract: Context: Despite the effort put into developing standards for structuring construction cost, and the strong interest into the field. Most construction companies still perform the process of data gathering and processing manually. That provokes inconsistencies, different criteria when classifying, misclassifications, and the process becomes very time-consuming particularly on big projects. Additionally, the lack of standardization makes it very difficult the cost estimation and comparison tasks. Objective: To create a method to extract and organise construction cost and quantity data into a consistent format and structure, to enable rapid and reliable digital comparison of the content. Method: The approach consists of a two-step method: Firstly, the system implements data mining to review the input document and determine how it is structured based on the position, format, sequence, and content of descriptive and quantitative data. Secondly, the extracted data is processed and classified with a combination of data science and experts' knowledge to fit a common format. Results: A big variety of information coming from real historical projects has been successfully extracted and processed into a common format with 97.5% of accuracy, using a subset of 5770 assets located on 18 different files, building a solid base for analysis and comparison. Conclusion: A robust and accurate method was developed for extracting hierarchical project cost data to a common machine-readable format to enable rapid and reliable comparison and benchmarking.

Keywords: data mining; data extraction; cost infrastructure projects; data science

---

Highlights:

- A fully automated method for the processes of data extraction and data wrangling which permits immediate access to data.
- The results present a strong accuracy of 97.5% when classifying the input structure.
- The solution achieves a higher level of efficiency due to the task automatization.

1. Introduction

Much effort has been put into developing standards for structuring construction cost and quantity information to streamline the process of estimating project costs, obtaining quotations and to enable comparison and analysis (Yan et al., 2020). Historically, information was broken down by building trade, the dominant UK standard being the Standard Method of Measurement (SMM), first introduced about 100 years ago.

Since the 1960s, there has been a shift to structuring information by building elements rather than by trade, first with the introduction of the Standard Form of Cost Analysis (SFCA) and then the New Rules of Measurement (NRM) suite of element-based standards introduced in 2009. However, some 12 years after it was officially superseded, organisations still continue to produce trade-based information, typically using SMM7 which is the most recent iteration of the trade-based standard (Symonds et al., 2015) . The industry clearly sees merits in both approaches depending on the purpose for which the data is being generated or used.

 doi:10.20944/preprints202308.1237.v1

Although construction information is now routinely transferred digitally between organisations, it is typically as PDFs or spreadsheets and is presented in a way that still emulates historic paper-based practice. Whilst a human, familiar with the standards, can interpret the visual clues and conventions used in these documents, such as pagination, alphanumeric codes, font changes, underlining, capitalisation and the position of text, that indicate the hierarchy of the information and thus its meaning, it is not digitally organised in a way that enables similar understanding by a computer. Further complexity is added by organisations developing their own bespoke variations of agreed standards,

This means that comparison and benchmarking between projects, even those ostensibly using the same measurement standard, relies on detailed expert review that is extremely time-consuming and prone to the inconsistencies that are inevitable where personal judgements are involved (Fisher et al., 1995).

Predictive data mining techniques can be used to automate the process of classifying construction cost data to a consistent common standard to enable robust comparison and analysis. This offers the potential to save time (and associated costs), allowing rapid assimilation of new data, and to circumvent the inconsistencies of manual data classification. The approach also offers a continuous learning capability making the system increasingly more accurate and robust as new data is registered in the dataset.

This paper presents an end-to-end methodology to automate and streamline the processes of extracting information from different input files with a variety of structures and formats and then to classify these data into a consistent format ready for comparison and analysis , to process the extracted information leaving room for comparison and predictions to finally implement some data analysis to bring a prove of the benefits of the suggested method in a real case scenario.

1.1. Related work

Analysis of cost data is of vital importance in the construction sector as the basis for establishing value for money on current projects and for accurate budgeting for future projects. The wider the sample of data, the more robust any comparison or benchmark is likely to be. However, the format and structure of available data is typically inconsistent even if ostensibly based on the same industry standard, either limiting the sample or involving time consuming, and costly manual re-classification of data by experts that is also potentially inconsistent as it involves personal judgement.

Due to the increased interest in the field, many approaches have been raised, the closely related studies have been classified and analyzed depending on the topic, the type of publication, their impact, and the date of publication. The most important ones have been gathered in Table 1 presenting the aim of the study, their approach, and their respective references:

Table 1. Summary presenting the aims and approaches of the related literature.

Aim	Approach	Reference
To identify similar construction projects for risk management.	The combination of NLP (Natural Language Processing) and Machine learning with a case base reasoning approach.	(Zou et al., 2017)
To enhance the attributes classification in construction projects.	The combination of data analysis and machine learning to identify the main factors that drive these classifications and provide reliable predictions.	(Desai, n.d.)
The optimization of risks applied to construction projects.	A two-step method is suggested based on the generation of the optimization attributes and the implementation of the algorithm C4.5.	(Zhong, n.d.)
Automatic text categorization of the project's assets.	A system that harnesses the benefits of NLP and machine learning for making an automatic text categorization.	(Sebastiani, 2002)

 doi:10.20944/preprints202308.1237.v1

To analyze the variability and the types of data structures used in construction projects.	A method that combines data extraction, data mining and analysis to assess the variability of structures among different projects.	(Soibelman et al., 2008)
To identify the non-flood areas in Poyang County, China.	To carry out different processes of data extraction and analysis that materialized in the identification of the flood risk areas.	(Moreno et al., n.d.)
To review and assess the current state of data mining in construction projects.	A systematic review of the historical application of data mining through the years to construction projects.	(Yan et al., 2020)
To decrease the transportation costs of prefabricated construction pieces	The approach extracts and processes geospatial data to feed the support vector machine for regression.	(Ahn et al., 2020)
To automatize the process of data extraction to support cost estimation	A method composed of three processes: The extraction of design information, to match the specified material from items in the database, to retrieve the price information of those materials	(Akanbi & Zhang, 2021)
To make a dictionary based on the WBS standard to support costs estimation	To carry out different surveys based on experts' opinions to develop the dictionary	(Ilmi et al., 2020)
To assess the main factors of the duration of construction projects	A data analysis is performed to assess the main factors that influence in determining the length of the construction projects.	(Stoy et al., n.d.)

(Zou et al., 2017) aims for retrieving similar cases with a novel approach using two different NLP techniques and a support vector machine, demonstrating in a practical scenario the implementation of these technologies to construction projects. The study is only suitable however for projects that share a specific type of document structure.

(Desai, n.d.) presents a method using decision trees and the inner correlation of the variables for enhancing the classification project which is similar to the presented full-fledged method excluding the data mining process.

A slightly different approach with similar ingredients can be found in (Zhong, n.d.) where the main aim consists of risk assessment with the usage of data science and the decision tree algorithm C4.5.

The increased importance of machine learning applied to construction projects can be demonstrated with (Sebastiani, 2002) which is a systematic review that analyses many studies performing an automatic text categorization over the previous years.

(Soibelman et al., 2008) consists of a study focusing on data mining aiming to analyses a big variety of data structures among a wide range of document types. The study incorporates not only file search but also text analysis and it avoids common mistakes through the usage of data mining. Despite the great results of the aforementioned study, the only critic it would be that the scope of the project is too wide to delve too deeply into the data extraction process.

(Hong et al., 2017) The approach intends to create an initial flood susceptibility map to identify the non-flood areas while analyses the importance of the flood-related variables. Their great results show an Area Under the Curve (AUC) of 0.98.

(Yan et al., 2020) is a systematic literature review which demonstrates the increased interest in applying data mining to the construction sector, especially after the year 2016 mostly due to China. Offering a general view that allows seeing the main trends of the market.

with a slightly different approach, (Ahn et al., 2020) improves the transportation costs for prefabricated construction parts extracting with the usage of geospatial data and support vector regression model. The results show the machine learning algorithm predicts with 87% of accuracy the number of trailers and the duration reducing 14% the costs

 doi:10.20944/preprints202308.1237.v1

A good example of the benefits of automatizing data extraction can be found in (Akanbi & Zhang, 2021) where a two-step method is presented for supporting cost estimation. First, the algorithm extracts design information from construction specifications and second, uses the extracted information to match the specified material from items in the database. The results show that they obtained 99.2% and 96.7% accuracy when extracting two types of information. However, the big percentages might be a bit misleading since a good data extraction system should always be near 100%.

After performing several surveys based on experts' opinions the study (Ilmi et al., 2020) presents a dictionary aimed at guidance for future cost project estimation based on the cost standard of work breakdown structure. Obtaining a handy way for guiding and earning productivity in the estimation process, as a constructive criticism it could be said that the project only applies to the scope of seaport project construction, and it presents some difficulties for implementing the same solution on a different scenario.

The study (Stoy et al., n.d.) presents a data analysis to assess the main factors that influence in determining the length of the construction projects. The main findings indicate that the project size and the followed standards have the biggest impact on the length. The main advantage of the study consists of filling the gap by studying the construction project length applied to German-speaking markets. However, it could be argued that by focusing on a small sample the knowledge cannot always be applied outside their scope.

1.2. The novelty of the method.

Based on the findings show by the state of the art, it can be inferred that the novelty of the presented method relies on the following points:

1. An end-to-end method: Many approaches managed to successfully solve a part of the data mining process, but very few ones encompass the process of data extraction, data wrangling and data preprocessing to make assets from different projects directly comparable.
2. Strong validation: The suggested method has been assessed with a big number of assets coming from real historical projects presenting reliable and robust results.
3. Different approach: The suggested approach relies on the usage of already existing technologies coming from the fields of data mining and machine learning assembled in an alternative way to target a different purpose. Making with this combination a unique method encompassing whole process.

2. The method

The paper presents a full-fledged method aiming for extracting relevant information in terms of project costs. In the given scenario, the information is located into a big number of files. The role of the method includes to process this information, making suitable for analysis and comparison by converting data into a common data structure. To achieve this goal, an approach is presented whose processes have been gathered in Figure 1:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2023 doi:10.20944/preprints202308.1237.v1

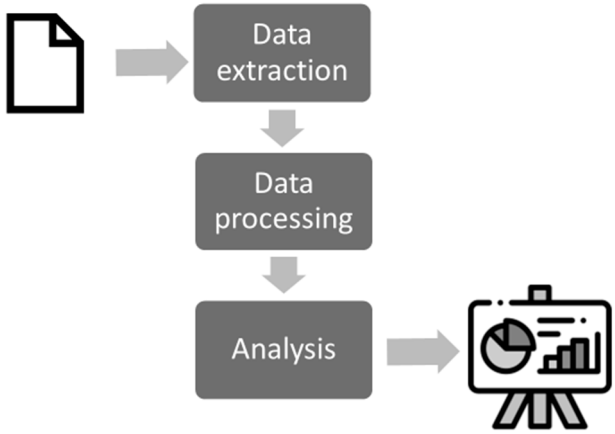

Figure 1. Diagram of the main processes of the presented method.

As presented in the picture above, the presented method can be summed up in a three-step process that happens on a sequential basis. Firstly, a data extraction process is carried out facing challenges such as: data variability, typos, and the appearance of different file structures. Secondly, a data process is being executed with the main function of transforming the already existing information into a common data format to make the projects comparable.

Finally, a process for data analysis will be carried out leaving some room to perform assessments and to better understand the inner relationship within the already gathered information. Finally, some inferences would be provided based on the received results.

2.1. Materials

For the development of the suggested method the following technologies have been used:

- Anaconda navigation version 2.2, for creating an environment.
- The IDE (Integrated development environment) Jupyter notebook version 6.4.5.
- Python language version 3.7.
- Different open-source libraries have been used, from them we can highlight: “pandas” for generating the data structures or “scikit learn” for providing the machine learning capabilities.

2.2. Input data

The presented method has been applied to a real case scenario extracting information from Bills of Quantities for two different projects measured using the SMM7 trade-based standard, which is a trade-based standard for structuring costs based on tender documentation(Murray, 1997). However, in both projects, the costs information has been structured quite differently.

- Costs for the first project were presented as a single PDF file with an elemental breakdown of the work for each of the 4 buildings included in the project; a total of 2217 items, grouped into 88 elemental Bills.
- Costs for the second project were presented as 17 separate Excel, trade based, work packages with a total of 1553 items.

Regardless of the type of document, both projects share a similar input structure when registering the costs. Each project asset is described by seven attributes containing information about cost categories and descriptions. The first attribute that describes each asset includes the most generic information. The information that contains the following attributes becomes progressively more specific and classified, being the seventh attribute the most specific of all. Additionally, each asset has specified other information such as the quantity needed for that project, the unit of measure, the rate, and the total cost.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2023 doi:10.20944/preprints202308.1237.v1

Despite the similarities between both projects, it is important to remark that whereas both projects have their assets classified following the SMM7 (Standard Method of Measurement) cost classification, one of the projects also contains additional information for classifying its costs into a trade-based standard.

For a better clarification, an example of some raw input data has been included in Figure 2:

				Robert Smillie Priced	
	DEMOLITION/ALTERATION/RENOVATION				
	C90: ALTERATIONS - SPOT ITEMS				
01	Various locations on site				
	Existing perimeter fencing and disposal off site; to be removed in sections as the new fence is erected				
	Complete; Provisional	113 m	£22.58	£2,551.54	
	Remove existing timber fencing internal to the site and dispose off site				
	Complete; Provisional	154 m	£22.58	£3,477.32	£6,028.86

Figure 2. An example of a registered asset in excel format.

Figure 2 presents a sample asset. The first line of the document indicates that this specific asset belongs to Category C for the standard SMM7 named: "C DEMOLITION/ALTERATION /RENOVATION" often used in many construction projects. The second row specifies the subcategory where the asset is located. In this case, it would be the category C90 for including the alteration works. Lately, the asset contains three more layers of descriptions to finally specify that the quantity would be 113, the unit of measure is in meters, the cost for each meter would be £22.58 also named as rate, and the final cost of the item taking into account the rate and the quantity would be £2551.54.

2.3. Understanding all processes of the method:

As stated before, the suggested method can be divided into three sequential steps: First, a data extraction process is carried out, followed by a data processing operation to convert it to a common data structure. Finally, some data analysis is carried out combined with some data analysis:

STEP 1: DATA EXTRACTION:

The input data for this step consists of two projects that come from different sources. On the one hand, a PDF file containing 2217 different assets for one of the projects. On the other hand, the 17 files containing information regarding 1553 assets for the second project.

The main role of this step includes the complete reading of all the input files and to recognize iteratively the relevant information regarding each asset. Additionally, the algorithm identifies the different types of information and is able to classify them and place them accordingly.

The main challenge for this step would be the data variability. Due to the manual inclusion of the data the assets present some variabilities in terms of structure and content. Unfortunately, this variability is amplified handling assets from different projects. The algorithm copes with this variability performing and flexible process since it extracts successfully the information applied to this subset.

As a main result. The information spread into the different input files is extracted and classified accordingly.

STEP 2: DATA PROCESSING:

The input data in this stage would be the information that has been extracted from the input files during the previous step. The main functionality now would be to process the information to make it process the information achieving two main goals. First, it makes the data comparable among different projects for the machine learning algorithms and second, it makes the data ready to be analyzed by the machine learning algorithms.

The main challenge for this step would be to cope with the typos and the errors that happen with the manual inclusion of data. The inclusion of manual notes for the accountant surveyor during

 doi:10.20944/preprints202308.1237.v1

the file or the break of the main structure of the data are two examples of that. Fortunately, the algorithm is able to identify these differences and is able to process the data accordingly.

As a main result, a common data structure is being created containing the following attributes for each registered asset.

- **Id:** It consists of an integer number that gets increased sequentially, it numerically identifies the number of assets that has been registered in the dataset.
- **Bill attribute:** It is a string type attribute that identifies the number of the bill where the asset has been located and a short description of it. For example: "Bill 123 Mechanical and plumbing"
- **Bill description:** Another string type attribute which contains redundant information including only a short description of the bill. It will be lately used for categorization purposes.
- **Category:** It is a categorical attribute containing a string that uniquely identifies the higher level of category for the SMM7 standard that the asset belongs to.
- **Subcategory:** Another categorical attribute that identifies the second layer of category for the standard SMM7 including a more specific categorization. For example, for the category: "D groundwork" we can find the subcategory: "D20: excavating and filling".
- **Description 1, 2 and 3:** As an additional information, each row contains three different descriptions where the first description contains the most generic information and the last one being the most specific. The information that the descriptions contain can vary a lot. For cite some examples, they can contain different unit of measures, for example: "maximum depth not exceeding 1.50m" or they can specify the type of work that has been carried out such as "Site preparation".
- **Quantity:** An integer number that specifies the number of items needed.
- **Unit:** An integer number which describes the unit of measure such as meter, item or square meter. For example, if the quantity of an item says 100 and the unit of measure indicates square meter. The dataset indicates that 100 square meters of that specific asset were needed on a specific project.
- **Rate:** A Boolean number including the price that is charged for each unit of measure. For example, it can say that for each square meter of a constructed wall, the client will be charged 157.57 GDP.
- **Total cost:** It is the number obtained because of multiplying the rate and the quantity. Following the previous examples. If the rate for each square meter of a wall would be 157.57 and the quantity would be 100. The total cost would be 15,757 GBP.
- **Letter:** The BoQ used as input files contain a letter that uniquely identify each asset located in the same categories and subcategories.
- **Page number:** As a helpful information, the processed data structure includes the pages number where the original item was registered in the input file. In this way, the accountant surveyor can doublecheck the correctness of the attributes in a faster way.
- **Trade based category name:** One of the projects also contains a trade-based classification of all their assets. Hence this string attribute works as a classification attribute identifying the categories that it belongs to.
- **Trade based category number:** Additionally, it specifies the amount of the total cost that would be located on that specific trade-based category. In the case where the asset only belongs to one category, this number would be the same than the total cost attribute.
- **Second Trade based category name:** Since SMM7 it is not a trade-based standard, there are a few cases were the same asset in SMM7 belongs to two categories with a trade-based approach. Hence, this attribute would be blank in most of the cases, and it would specify the second category that the asset belongs to in case of conflict.
- **Second Trade based category number:** In those cases where the asset belongs to more than one trade-based category, this number would indicate the cost that would be located on the second category. For example, for a fictitious asset classified the SMM7 class "Masonry" with a total cost of 10,000 GDP. On trade-based standard it could locate 4,000 GDP for "Substructure" and 6,000 GDP for "external walls".

For a better clarification, the first five assets for the project containing the excel package files is showed on Table 2 showing their information already extracted and processed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2023 doi:10.20944/preprints202308.1237.v1

Table 2. A sample of the first five registered assets with their information already extracted.

Id	Bill description	Category	Subcategory	Description level 1	Description level 2
0	Groundworks & substruct.	C demolition /…	C90 alterations…	Various loc. on site	Existing perimeter fencing and disp…
1	Groundworks & substruct.	C demolition /…	C90 alterations…	Various loc. on site	Remove existing timber fencing int…
2	Groundworks & substruct.	D groundwork	D20 excavating…	Site preparation	Site preparation
3	Groundworks & substruct.	D groundwork	D20 excavating…	excavating	To reduce levels
4	Groundworks & substruct.	D groundwork	D20 excavating…	excavating	Basements and the like

row	Description 3	quantity	unit	rate	Total cost	letter	Page num.
0	Complete; provisional	113	m	2258	255154	a	1
1	Complete; provisional	154	m	2258	347732	b	1
2	Brushes, scrub, undergrowth, hedges, trees and …	3328	m2	237	765036	a	1
3	Maximum depth not exceeding 2.00m	1140	m3	339	38646	b	1
4	Maximum depth not exceeding 1.00m	242	m3	339	82038	c	1

Row	Trade-based category name	Trade-based category code	Trade-based cat. name 2	Trade-based cat. code 2
0	Site works	255154	-	0
1	Site works	347732	-	0
2	Substructure	76036	-	0
3	Substructure	38646	-	0
4	Substructure	82038	-	0

STEP 3: ANALYSIS:

The input of this step consists of the already extracted and processed assets into a common data structure. Through this step a process of analysis and predictions is being carried out to have a deeper understanding of the data and to be able to perform future predictions allowing for automatization in the future projects.

The main challenge of this step consists in identifying the inner correlation between the different attributes in the dataset and to identify the main patters that allow the machine learning algorithm to make more accurate predictions.

As a main results, some knowledge that can be extrapolated is being extracted out of the initial dataset. Additionally, a mapping assessing a possible conversion between the SMM7 standard and a trade-based are also generated for a specified subset.

3. Results

As stated in the method section, the suggested method starts with a data extraction process. During this process the solution reads the input files and extracts their information sequentially. Secondly, the solution processed the data and is able to classify the information to construct an output structure to make the projects comparable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2023 doi:10.20944/preprints202308.1237.v1

The method is capable of extracting the information and processing it with 100% of accuracy. Secondly, the method is able to classify the extracted information and construct a common data structure successfully in 3679 assets out of the total 3770, which makes the solution 97.58% accurate on this step.

Finally, to demonstrate the benefits of standardization, an analysis process is being carried out. First of all, by analyzing the extracted data, it is important to take into account that the costs of all the registered assets are according to the SMM7 standard cost classification, by performing some analysis it is possible to appreciate that the categories where the accountant surveyors locate the costs are irregularly distributed. Where the most popular categories would be the combination of "S: PIPED SUPPLY" and "T: MECHANICAL HEATING /COOLING/REFRIGERATION SYSTEMS" with 22.02% of the occurrences followed by the categories "P BUILDING FABRIC SUNDRIES" encompassing 10.05% of the cases and the combination of "V: ELECTRICAL SUPPLY/POWER/LIGHTING" and "W: COMMUNICATIONS/SECURITY/CONTROL SYSTEMS" with 9.48% of the cases.

On the unpopular side we have categories like "C DEMOLITION/ALTERATION/RENOVATION" and "G STRUCTURAL/CARCASSING METAL/TIMBER" both appearing only in 1% of the assets.

For clarification, Figure 3 is shown presenting the top 20 most popular categories and their respective percentage of occurrences in the total dataset where it is possible to appreciate a big gap between the most popular category and the rest.

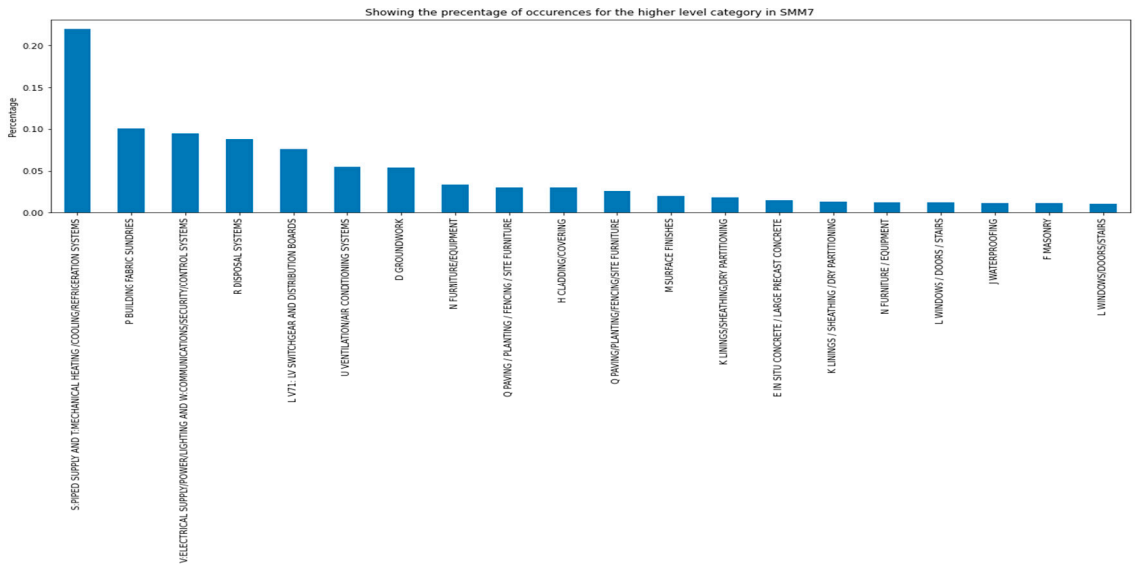

Figure 4. Top 20 most popular categories for the SMM7 standard.

Additionally, the SMM7 standard classifies the assets into different subcategory costs. In this more specific classification, it is possible to appreciate more uniform distribution of the occurrences spread among a much wider range of subcategories. For clarification, the top 20 most common subcategories have been gathered on Figure 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 August 2023 doi:10.20944/preprints202308.1237.v1

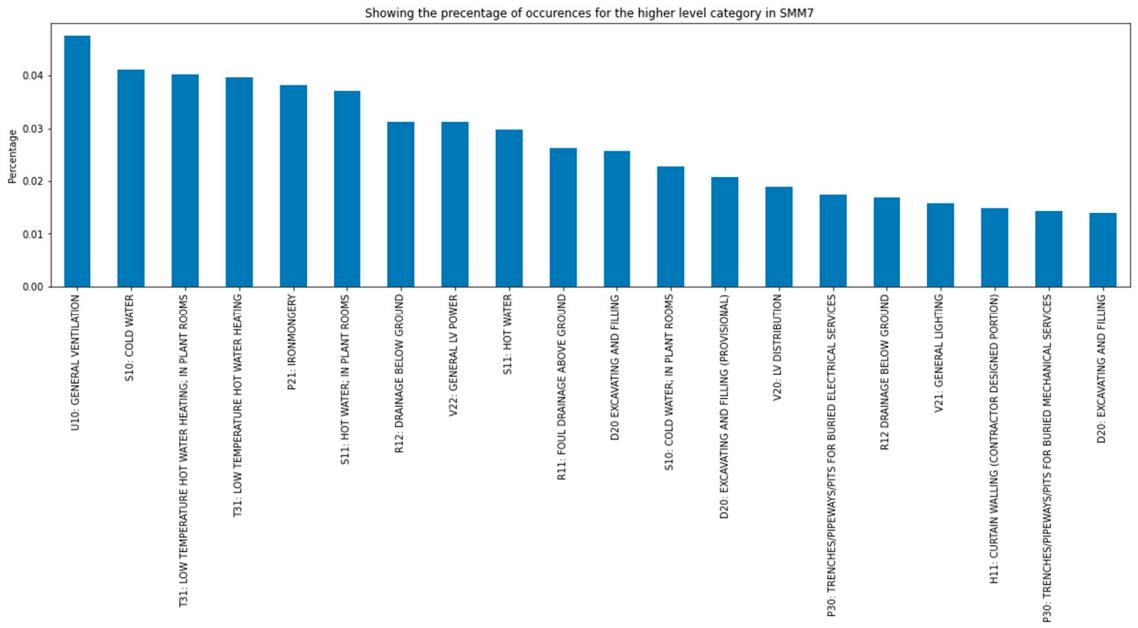

Figure 5. Top 20 most popular subcategories for the SMM7 standard.

As showed in the previous figure, the most common category would be "U10 GENERAL VENTILATION" encompassing 4.75% of the cases, followed closely by "S10 COLD WATER" with 4.11% and "T30 LOW TEMPERATURE HOT WATER HEATING in plant rooms" with 4.01%.

On the unpopular side, the most uncommon category would be "J30 LIQUID APPLIED TANKING / DAMP", "N25 SPECIAL PURPOSE FIXTURES /FURNISHINGS / EQUIPMENT" and "L30 STAIRS / WALKWAYS / BALUSTRADES" all of them 0.05% of the occurrences.

Second of all, to demonstrate one of the main capabilities of the suggested method, a mapping between the SMM7 and a trade-based approach has been provided, allowing for comparison of projects whose costs structure is radically different. The results have been gathered on Figure 6 showing the mapping 1553 assets coming from the first of the projects.

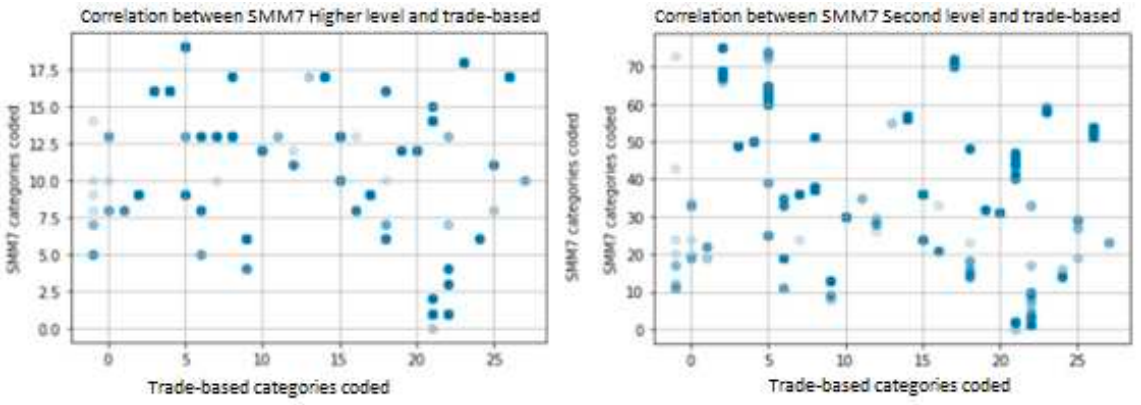

Figure 6. Top 20 most popular subcategories for the SMM7 standard.

The darkest dots in Figure 6 indicate that there are several instances belonging to those categories, whereas the lighter dots indicate that there are very few instances placed there. Firstly, it is possible to observe the irregular distribution of the categories among both standards. Secondly, it is possible to infer that there is indeed a correlation between both approaches which suggests that it would be feasible to implement a classifier algorithm to convert between both standards by extracting relevant features based on the text descriptions.

4. Conclusions

The main contribution of the presented paper relies on the creation of a full-fledged method, encompassing not only data gathering by harnessing data mining techniques and being able to extract information even in an scenario where the information it is spread among a big range of files including but also including an accurate data classification and the capability of converting the extracted data into a common and comparable data format.

The well validated results showing 97.5% of accuracy are reliable enough to prove the strength of the method in a real case scenario. The remaining 2.5% can be attributed to typos, and multiple irregularities located on the input files.

Finally, it is important to highlight the fully automated capability of the method. Despite the fact that it is able to classify the data emulating the expert´s knowledge. It is able to do so without any human intervention which makes it a fully automatized method able to work as a black box for the end user.

References

1. Ahn, S. J., Han, S. U., & Al-Hussein, M. (2020). Improvement of transportation cost estimation for prefabricated construction using geo-fence-based large-scale GPS data feature extraction and support vector regression. *Advanced Engineering Informatics*, *43*. https://doi.org/10.1016/J.AEI.2019.101012
2. Akanbi, T., & Zhang, J. (2021). Design information extraction from construction specifications to support cost estimation. *Automation in Construction*, *131*. https://doi.org/10.1016/J.AUTCON.2021.103835
3. Desai, V. S. (n.d.). Improved Decision Tree Methodology for the Attributes of Unknown or Uncertain Characteristics-Construction Project Prospective. *The International Journal of Applied Management and Technology*, *6*, 201.
4. Fisher, D., Miertschin, S., & PollockJr., D. R. (1995). Benchmarking in Construction Industry. *Journal of Management in Engineering*, *11*(1), 50–57. https://doi.org/10.1061/(ASCE)0742-597X(1995)11:1(50)
5. Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A.-X., & Chen, W. (2017). *Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China*. https://doi.org/10.1016/j.scitotenv.2017.12.256
6. Ilmi, A. A., Supriadi, L. S. R., Latief, Y., & Muslim, F. (2020). Development of dictionary and checklist based on Work Breakdown Structure (WBS) at seaport project construction for cost estimation planning. *IOP Conference Series: Materials Science and Engineering*, *930*(1). https://doi.org/10.1088/1757-899X/930/1/012007
7. Moreno, V., Génova, G., Parra, E., & Fraga, A. (n.d.). *Application of machine learning techniques to the flexible assessment and improvement of requirements quality*.
8. Murray, G. P. (1997). Rules and Techniques for Measurement of Services. *Measurement of Building Services*, 9–18. https://doi.org/10.1007/978-1-349-14282-8_2
9. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283
10. Soibelman, L., Wu, J., Caldas, C., Brilakis, I., & Lin, K. Y. (2008). Management and analysis of unstructured construction data types. *Advanced Engineering Informatics*, *22*(1), 15–27. https://doi.org/10.1016/j.aei.2007.08.011
11. Stoy, C., Dreier, F., & Schalcher, H.-R. (n.d.). *Construction duration of residential building projects in Germany*. https://doi.org/10.1108/09699980710716972
12. Symonds, B., Barnes, P., & Robinson, H. (2015). New Approaches and Rules of Measurement for Cost Estimating and Planning. *Design Economics for the Built Environment: Impact of Sustainability on Project Evaluation*, 31–46. https://doi.org/10.1002/9781118944790.CH3
13. Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). *Data mining in the construction industry: Present status, opportunities, and future trends*. https://doi.org/10.1016/j.autcon.2020.103331
14. Zhong, Y. (n.d.). *Research on Construction Engineering Project Management Optimization Based on C4.5 Improved Algorithm*. https://doi.org/10.1088/1757-899X/688/5/055036
15. Zou, Y., Kiviniemi, A., & Jones, S. W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Automation in Construction*, *80*, 66–76. https://doi.org/10.1016/J.AUTCON.2017.04.003