buildings

MDPI

*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; lamine.mahdjoubi@uwe.ac.uk (L.M.); bill.gething@uwe.ac.uk (B.G.)
* Correspondence: guede-91@hotmail.com

**Abstract:** Context: Despite the effort put into developing standards for structuring construction costs and the strong interest in the field, most construction companies still perform the process of data gathering and processing manually. This provokes inconsistencies, different criteria when classifying, misclassifications, and the process becomes very time-consuming, particularly in large projects. Additionally, the lack of standardization makes cost estimation and comparison tasks very difficult. Objective: The aim of this work was to create a method to extract and organize construction cost and quantity data into a consistent format and structure to enable rapid and reliable digital comparison of the content. Methods: The approach consisted of a two-step method: firstly, the system implemented data mining to review the input document and determine how it was structured based on the position, format, sequence, and content of descriptive and quantitative data. Secondly, the extracted data were processed and classified with a combination of data science and experts' knowledge to fit a common format. Results: A large variety of information coming from real historical projects was successfully extracted and processed into a common format with 97.5% accuracy using a subset of 5770 assets located on 18 different files, building a solid base for analysis and comparison. Conclusions: 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



**Citation:** Adanza Dopazo, D.; Mahdjoubi, L.; Gething, B. A Method to Enable Automatic Extraction of Cost and Quantity Data from Hierarchical Construction Information Documents to Enable Rapid Digital Comparison and Analysis. *Buildings* **2023**, *13*, 2286. https://doi.org/10.3390/buildings13092286

Academic Editor: Kwangbok Jeong

Received: 15 August 2023
Revised: 29 August 2023
Accepted: 6 September 2023
Published: 8 September 2023



## 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 enabling comparison and analysis [1]. Historically, information was broken down by building trade, with the dominant UK standard being the Standard Method of Measurement (SMM), which was 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, organizations still continue to produce trade-based information, typically using SMM7, which is the most recent iteration of the trade-based standard [2]. The industry clearly sees merits in both approaches, depending on the purpose for which the data are being generated or used.

Although construction information is now routinely transferred digitally between organizations, it is typically as PDFs or spreadsheets and is presented in a way that still emulates historic paper-based practice. Whilst a human who is familiar with the standards can interpret the visual clues and conventions used in these documents, such as pagination, alphanumeric codes, font changes, underlining, capitalization, and the position of text, that indicate the hierarchy of the information and, thus, its meaning, it is not digitally organized

in a way that enables similar understanding by a computer. Further complexity is added by organizations 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 [3].

Predictive data mining techniques can be used to automate the process of classifying construction cost data into 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 are registered in the dataset.

This paper presents an end-to-end methodology to automate and streamline the process of extracting information from different input files with a variety of structures and formats and then to classify these data into a consistent format that is ready for comparison and analysis; to process the extracted information, leaving room for comparison and predictions; and to finally implement some data analysis to prove 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 are typically inconsistent, even if they are ostensibly based on the same industry standard, either limiting the sample or involving time-consuming and costly manual reclassification of data by experts, which 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 were classified and analyzed depending on the topic, the type of publication, their impact, and the date of publication. The most important ones are gathered in Table 1, presenting the aims of the studies, their approaches, 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. | A combination of NLP (Natural Language Processing) and machine learning with a case-based reasoning approach. | [4] |
| To enhance the classification of attributes in construction projects. | A combination of data analysis and machine learning to identify the main factors that drive these classifications and provide reliable predictions. | [5] |
| 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. | [6] |
| Automatic text categorization of a project's assets. | A system that harnesses the benefits of NLP and machine learning for making an automatic text categorization. | [7] |
| 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. | [8] |
| 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. | [9] |
| To review and assess the current state of data mining in construction projects. | A systematic review of the historical application of data mining in construction projects through the years. | [1] |
| To decrease the transportation costs of prefabricated construction pieces. | The approach extracted and processed geospatial data to feed a support vector machine for regression. | [10] |
| To automatize the process of data extraction to support cost estimation. | A method composed of three processes: the extraction of design information, matching the specified material from items in the database, retrieving the price information of those materials. | [11] |

**Table 1.** *Cont.*

| Aim | Approach | Reference |
| --- | --- | --- |
| To form a dictionary based on the WBS standard [12] to support cost estimation. | To carry out different surveys based on experts' opinions to develop the dictionary. | [13] |
| To assess the main factors of the duration of construction projects. | A data analysis was performed to assess the main factors that had an influence in determining the length of construction projects. | [14] |

Ref. [4] aimed to retrieve similar cases with a novel approach using two different NLP techniques and a support vector machine, demonstrating the implementation of these technologies to construction projects in a practical scenario. The study is only suitable, however, for projects that share a specific type of document structure.

Ref. [5] presents a method using decision trees and the inner correlation of the variables for enhancing a classification project, which is similar to the present full-fledged method, excluding the data mining process.

A slightly different approach with similar characteristics can be found in [6], where the main aim consisted of risk assessment with the usage of data science and the decision tree algorithm C4.5.

The increased importance of machine learning being applied to construction projects can be demonstrated with [7], which is a systematic review that analyzes many studies, performing an automatic text categorization over the previous years.

Ref. [8] consists of a study focusing on data mining aiming to analyze a large variety of data structures among a wide range of document types. The study incorporates not only file searches 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 criticism would be that the scope of the project is too wide to delve too deeply into the data extraction process.

Ref. [15] suggest an approach intended to create an initial flood susceptibility map to identify non-flood areas while analyzing the importance of flood-related variables. Their great results show an Area Under the Curve (AUC) of 0.98.

Ref. [1] 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 the main trends of the market to be seen.

With a slightly different approach, [10] improve the transportation costs for prefabricated construction parts by extracting the usage of geospatial data and using the support vector regression model. The results show that the machine learning algorithm predicted the number of trailers and the duration with 87% accuracy, reducing 14% of the costs.

A good example of the benefits of automatizing data extraction can be found in [11], where a two-step method is presented for supporting cost estimation. First, the algorithm extracted design information from construction specifications, and second, it used 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 large 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, a study by [13] presents a dictionary aimed at guidance for future project cost estimation based on the cost standard of the 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 in different scenarios.

A study by [14] presents data analysis to assess the main factors that influence determining the length of construction projects. The main findings indicate that the project size and the standards followed have the greatest impact on the length. The main advantage of the study consists in filling the gap by studying construction project length in 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 Current Method*

Based on state-of-the-art findings, it can be inferred that the novelty of the present 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 encompass the processes of data extraction, data wrangling, and data preprocessing to make assets from different projects directly comparable.
2. Strong validation: the suggested method was assessed with a large number of assets coming from real historical projects, presenting reliable and robust results.
3. Different approach: the suggested approach relied on the usage of already existing technologies from the fields of data mining and machine learning, assembled in an alternative way to target a different purpose, making a unique method encompassing the whole process with this combination.

## 2. The Method

This paper presents a full-fledged method aiming to extract relevant information in terms of project costs. In the given scenario, the information was located in a large number of files. The role of the method included processing this information, making it suitable for analysis and comparison by converting the data into a common data structure. To achieve this goal, an approach whose processes are gathered in Figure 1 is presented.

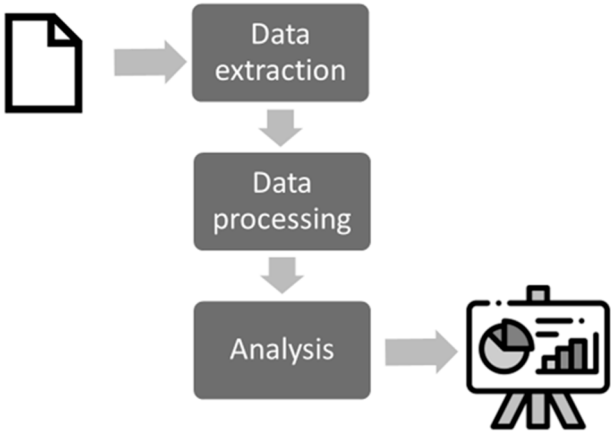


**Figure 1.** Diagram of the main processes of the present method.

As presented in the picture above, the present method can be summed up in a three-step process that happened on a sequential basis. Firstly, a data extraction process was carried out, facing challenges such as data variability, typos, and the appearance of different file structures. Secondly, a data process was 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 was carried out, leaving some room to perform assessments and to better understand the inner relationships within the already gathered information. Finally, some inferences were provided based on the received results.

### *2.1. Materials*

For the development of the suggested method, the following technologies were 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, from which we can highlight “pandas” for generating the data structures or “scikit learn” for providing the machine learning capabilities.

### *2.2. Input Data*

The present method was applied to a real case scenario of extracting information from bills of quantities for two different projects, as measured using the SMM7 trade-based standard, which is a trade-based standard for structuring costs based on tender documentation [16]. However, in both projects, the cost information was 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, with a total of 2217 items grouped into 88 elemental bills.
- Costs for the second project were presented as 17 separate trade-based Excel work packages, with a total of 1553 items.

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

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

For a better clarification, an example of some raw input data is included in Figure 2.

| | | | | | |
|---|---|---|---|---|---|
| | **DEMOLITION/ALTERATION/RENOVATION** | | | Robert Smillie Priced | |
| | **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" that is often used in many construction projects. The second row specifies the subcategory where the asset is located. In this case, it is in the category C90 for including alteration work. Lastly, the asset contains three more layers of descriptions to finally specify that the quantity is 113; the unit of measure is in meters; the cost for each meter is £22.58, which is also named as rate; and the final cost of the item, taking into account the rate and the quantity, is £2551.54.

### *2.3. Understanding All Processes of the Method*

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

Step 1: Data extraction:

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

The main role of this step included the complete reading of all the input files and iteratively recognizing the relevant information regarding each asset. Additionally, the algorithm identified the different types of information and was able to classify them and place them accordingly.

The main challenge for this step was data variability. Due to the manual inclusion of the data, the assets presented some variability in terms of structure and content. Unfortunately, this variability was amplified in handling assets from different projects. The algorithm coped with this variability performing and flexible process since it successfully extracted the information applied to this subset.

As a main result, the information spread throughout the different input files was extracted and classified accordingly.

Step 2: Data processing:

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

The main challenge for this step was to cope with the typos and the errors that happened with the manual inclusion of data. The inclusion of manual notes for the accountant surveyor in the files or the break of the main structure of the data were two examples of that. Fortunately, the algorithm was able to identify these differences and was able to process the data accordingly.

As a main result, a common data structure was created, containing the following attributes for each registered asset:

- ID: It consisted of an integer number that increased sequentially, and it numerically identified the number of assets that were registered in the dataset.
- Bill attribute: It was a string-type attribute that identified the number of the bill, where the asset was located, and a short description of it, for example, "Bill 123 Mechanical and plumbing".
- Bill description: Another string-type attribute which contained redundant information, including only a short description of the bill. It was later used for categorization purposes.
- Category: It was a categorical attribute containing a string that uniquely identified the higher level of the category for the SMM7 standard [17] that the asset belonged to.
- Subcategory: Another categorical attribute that identified the second layer of the category for the SMM7 standard, including a more specific categorization. For example, for the category "D groundwork", we found the subcategory "D20: excavating and filling".
- Description 1, 2 and 3: As additional information, each row contained three different descriptions, where the first description contained the most generic information and the last one was the most specific. The information that the descriptions contained varied a lot. To cite some examples, they could contain different units of measure, for example "maximum depth not exceeding 1.50 m", or they could specify the type of work that was carried out, such as "Site preparation".
- Quantity: An integer number that specified the number of items needed.
- Unit: An integer number which described the unit of measure, such as meters, item, or square meters. For example, if the quantity of an item was 100 and the unit of measure indicated square meters, the dataset indicated that 100 square meters of that specific asset was needed on a specific project.
- Rate: A Boolean number including the price that was charged for each unit of measure. For example, it might have stated that for each square meter of a constructed wall, the client would be charged 157.57 GDP.
- Total cost: It was the number obtained by multiplying the rate and the quantity. Following the previous examples, if the rate for each square meter of a wall was 157.57 and the quantity was 100, the total cost would be 15,757 GBP.

- Letter: The BoQs used as input files contained a letter that uniquely identified each asset located in the same categories and subcategories.
- Page number: As a helpful piece of information, the processed data structure included the page number where the original item was registered in the input file. In this way, the accountant surveyor could doublecheck the correctness of the attributes in a faster way.
- Trade-based category name: One of the projects also contained a trade-based classification of all their assets. Hence, this string attribute worked as a classification attribute, identifying the categories that it belonged to.
- Trade-based category number: Additionally, it specified the amount of the total cost that was located in that specific trade-based category. In cases where the asset only belonged to one category, this number was the same as the total cost attribute.
- Second trade-based category name: Since SMM7 is not a trade-based standard, there were a few cases where the same asset in SMM7 belonged to two categories with a trade-based approach. Hence, this attribute was blank in most of the cases, and it would specify the second category that the asset belonged to in case of conflict.
- Second trade-based category number: In those cases where the asset belonged to more than one trade-based category, this number indicated the cost that was located in the second category. For example, for a fictitious asset classified in the SMM7 class "Masonry" with a total cost of 10,000 GDP, on the trade-based standard, it would be located in 4000 GDP for "Substructure" and 6000 GDP for "external walls".

For better clarification, the first five assets for the project containing the Excel package files are shown in Tables 2–4, which shows their information already extracted and processed.

**Table 2.** The first attributes 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 and substruct. | C demolition/... | C90 alterations... | Various loc. on site | Existing perimeter fencing and disp... |
| 1 | Groundworks and substruct. | C demolition/... | C90 alterations... | Various loc. on site | Remove existing timber fencing int... |
| 2 | Groundworks and substruct. | D groundwork | D20 excavating... | Site preparation | Site preparation |
| 3 | Groundworks and substruct. | D groundwork | D20 excavating... | excavating | To reduce levels |
| 4 | Groundworks and substruct. | D groundwork | D20 excavating... | excavating | Basements and the like |

**Table 3.** The main attributes of a sample of the first five registered assets with their information already extracted.

| Row | Description 3 | Quantity | Unit | Rate | Total Cost | Letter | Page Num. |
|---|---|---|---|---|---|---|---|
| 0 | Complete; provisional | 113 | m | 2258 | 255,154 | a | 1 |
| 1 | Complete; provisional | 154 | m | 2258 | 347,732 | b | 1 |
| 2 | Brushes, scrub, undergrowth, hedges, trees and ... | 3328 | $m^2$ | 237 | 765,036 | a | 1 |
| 3 | Maximum depth not exceeding 2.00 m | 1140 | $m^3$ | 339 | 38,646 | b | 1 |
| 4 | Maximum depth not exceeding 1.00 m | 242 | $m^3$ | 339 | 82,038 | c | 1 |

**Table 4.** The last attributes of a sample of the first five registered assets with their information already extracted.

| Row | Trade-Based Category Name | Trade-Based Category Code | Trade-Based Cat. Name 2 | Trade-Based Cat. Code 2 |
|---|---|---|---|---|
| 0 | Site works | 255,154 | - | 0 |
| 1 | Site works | 347,732 | - | 0 |
| 2 | Substructure | 76,036 | - | 0 |
| 3 | Substructure | 38,646 | - | 0 |
| 4 | Substructure | 82,038 | - | 0 |

Step 3: Analisis:

The input of this step consisted of combining the already extracted and processed assets into a common data structure. Through this step, a process of analysis and predictions was carried out to gain a deeper understanding of the data and to be able to perform future predictions, allowing for automatization in future projects.

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

As a main result, some knowledge that was able to be extrapolated was extracted out of the initial dataset. Additionally, a mapping assessing a possible conversion between the SMM7 standard and a trade-based standard were also generated for a specified subset.

### *2.4. The Limitations of the Study*

Although the benefits and the main strengths of the study have been clearly stated through the paper, it is also important to take into account the limitations of the suggested method to present a more complete and unbiased solution to the reader:

- Representative data: The success of the method might depend on the diversity and representativeness of the historical project data used. If the subset of projects does not cover a wide range of project types, sizes, and complexities, the method's accuracy and applicability to real-world scenarios could be limited.
- Expert knowledge dependency: The method relies on a combination of data science techniques and experts' knowledge for data classification and standardization. Although some of the classification costs do not leave room for discussion, in some specific cases, this could lead to bias or inconsistencies if the experts' knowledge is not fully comprehensive or if different experts have differing interpretations.

## 3. Results

As stated in the methods section, the suggested method started with a data extraction process. During this process, the system read the input files and extracted their information sequentially. Secondly, the system processed the data and was able to classify the information to construct an output structure to make the projects comparable.

This method was capable of extracting the information and processing it with 100% accuracy. Secondly, the method was able to classify the extracted information and construct a common data structure successfully for 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 was carried out. First of all, in analyzing the extracted data, it is important to take into account that the costs of all the registered assets were according to the SMM7-standard cost classification; by performing some analysis, it was possible to appreciate that the categories where the accountant surveyors located the costs were irregularly distributed. The most popular categories were 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 had categories like "C Demolition/Alterations/Renovation" and "G Structural/Carcassing metal/Timber", both appearing only in 1% of the assets.

For clarification, Figure 3 presents the top 20 most popular categories and their respective percentages of occurrences in the total dataset, where it was possible to appreciate a large gap between the most popular category and the rest.

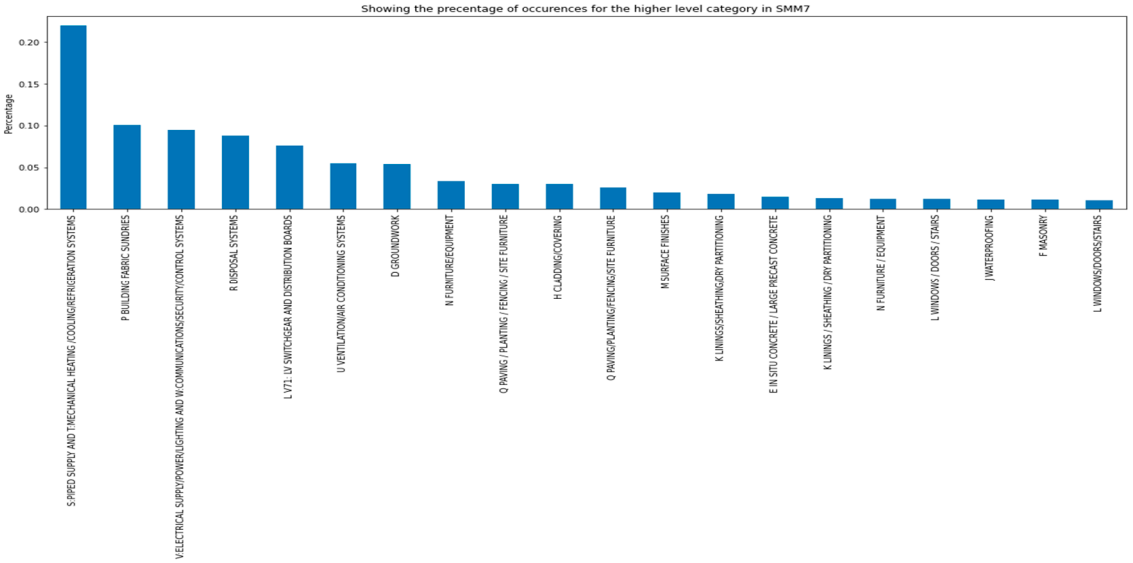


**Figure 3.** Top 20 most popular categories for the SMM7 standard.

Additionally, the SMM7 standard classifies assets into different subcategory costs. In this more specific classification, it was possible to appreciate a more uniform distribution of the occurrences spread among a much wider range of subcategories. For clarification, the top 20 most common subcategories are gathered in Figure 4.

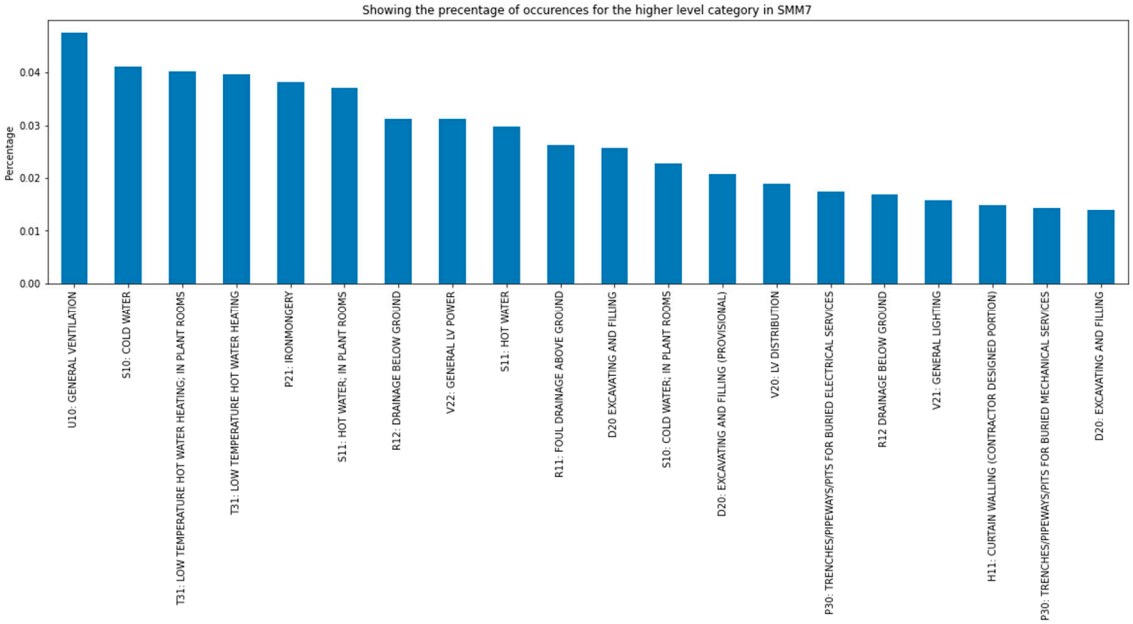


**Figure 4.** Top 20 most popular subcategories for the SMM7 standard.

As shown in the previous figure, the most common category was "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 categories were "J30 Liquid Applied Tanking/Damp", "N25 Spetial Purpose Fixtures/Furnishing/Equipment", and "L30 Stairs/Walkways/Balustrades", all of which made up 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 is provided, allowing for

comparison of projects whose cost structure is radically different. The results are gathered in Figure 5, showing the mapping of 1553 assets from the first of the projects.

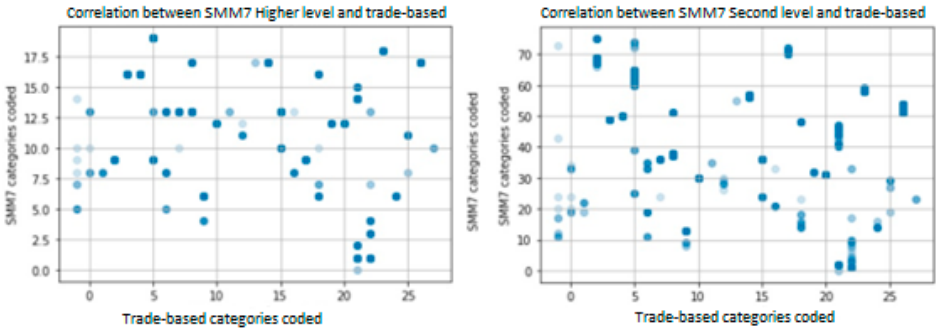


**Figure 5.** Correlation between SMM7 standard and trade-base approaches.

The darkest dots in Figure 5 indicate that there were several instances belonging to those categories, whereas the lighter dots indicate that there were 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 was indeed a correlation between both approaches, which suggests that it would be feasible to implement a classifier algorithm to convert information between both standards by extracting relevant features based on the text descriptions.

## 4. Conclusions

The main contribution of the present 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 a scenario where the information was spread among a large range of files, but also including 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% 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 was able to classify the data emulating the expert's knowledge, it was able to do so without any human intervention, which makes it a fully automatized method that is able to work as a black box for the end user.

### *Future Work*

Although the current study presents a full-fledged methodology that was complexly integrated and tested with real historical data, the study also opens the door to future research directions. Within them, it is worth highlighting the following ones:

- Firstly, research should investigate the challenges and barriers that construction companies might face when adopting and implementing the proposed method. This could include factors such as initial setup, integration with existing workflows, and overcoming resistance to change.
- Secondly, although the method was tested with a large set of assets, its validation could also be extended to more historical projects of different kinds and to implementing different data structures. This would bring the method more flexibility, credibility, and applicability.
- Thirdly, a comparative study between the proposed automated method and traditional manual methods of cost data extraction and organization could be conducted. This would help demonstrate the efficiency gains and accuracy improvements offered by the new approach.

- Finally, the present method could be explored and aligned with different existing industry standards for cost estimation and classification, enhancing its compatibility and encouraging its adoption by other companies.

**Author Contributions:** Writing—original draft, D.A.D.; Writing—review & editing, D.A.D.; Supervision, L.M. and B.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the UK Department for Transport and received funding under the Innovate UK project "Transport Infrastructure Efficiency Strategy Living Lab (45382)".

**Data Availability Statement:** Data is unavailable for confidentiality reasons. Access is restricted to protect sensitive information and the needs of the company that owns the restricted information. Inquiries for collaboration can be made to the authors for potential data access considerations.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Yan, H.; Yang, N.; Peng, Y.; Ren, Y. Data mining in the construction industry: Present status, opportunities, and future trends. *Automation in Construction.* **2020**, *119*, 103331. [CrossRef]
2. Symonds, B.; Barnes, P.; Robinson, H. New Approaches and Rules of Measurement for Cost Estimating and Planning. In *Design Economics for the Built Environment: Impact of Sustainability on Project Evaluation*; John Wiley & Sons: Hoboken, NJ, USA, 2015; pp. 31–46. [CrossRef]
3. Fisher, D.; Miertschin, S.; Pollock, D.R., Jr. Benchmarking in Construction Industry. *J. Manag. Eng.* **1995**, *11*, 50–57. [CrossRef]
4. Zou, Y.; Kiviniemi, A.; Jones, S.W. Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Autom. Constr.* **2017**, *80*, 66–76. [CrossRef]
5. Desai, V.S. Improved Decision Tree Methodology for the Attributes of Unknown or Uncertain Characteristics-Construction Project Prospective. *Int. J. Appl. Manag. Technol.* **2008**, *6*, 201.
6. Zhong, Y. Research on Construction Engineering Project Management Optimization Based on C4.5 Improved Algorithm. *IOP Conf. Serv. Mater. Sci. Eng.* **2019**, *688*, 055036. [CrossRef]
7. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv. (CSUR)* **2002**, *34*, 1–47. [CrossRef]
8. Soibelman, L.; Wu, J.; Caldas, C.; Brilakis, I.; Lin, K.Y. Management and analysis of unstructured construction data types. *Adv. Eng. Inform.* **2008**, *22*, 15–27. [CrossRef]
9. Moreno, V.; Génova, G.; Parra, E.; Fraga, A. Application of machine learning techniques to the flexible assessment and improvement of requirements quality. *Softw. Qual. J.* **2020**, *28*, 1645–1674. [CrossRef]
10. Ahn, S.J.; Han, S.U.; Al-Hussein, M. Improvement of transportation cost estimation for prefabricated construction using geofence-based large-scale GPS data feature extraction and support vector regression. *Adv. Eng. Inform.* **2020**, *43*, 101012. [CrossRef]
11. Akanbi, T.; Zhang, J. Design information extraction from construction specifications to support cost estimation. *Autom. Constr.* **2021**, *131*, 103835. [CrossRef]
12. Norman, E.S.; Brotherton, S.A.; Fried, R.T. *Work Breakdown Structures: The Foundation for Project Management Excellence*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
13. Ilmi, A.A.; Supriadi LS, R.; Latief, Y.; Muslim, F. Development of dictionary and checklist based on Work Breakdown Structure (WBS) at seaport project construction for cost estimation planning. *IOP Conf. Serv. Mater. Sci. Eng.* **2020**, *930*, 012007. [CrossRef]
14. Stoy, C.; Dreier, F.; Schalcher, H.-R. Construction duration of residential building projects in Germany. *Eng. Constr. Archit. Manag.* **2007**, *14*, 52–64. [CrossRef]
15. Hong, H.; Tsangaratos, P.; Ilia, I.; Liu, J.; Zhu, A.-X.; Chen, W. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. *Sci. Total Environ.* **2017**, *625*, 575–588. [CrossRef] [PubMed]
16. Murray, G.P. Rules and Techniques forMeasurement of Services. *Meas. Build. Serv.* **1997**, 9–18. [CrossRef]
17. Keily, P.; McNamara, P.H. *SMM7 Explained and Illustrated*; RICS Books: Coventry, UK, 2003.