Experiment proposal for data quality assessment in construction management

Raphael Hippe[1], Marco Binninger and Olli Seppanen[1]
[1]Department of Civil Engineering, Aalto University, Otakaari 5A, 02150 Helsinki, Finland
E-mail(s): raphael.hippe@aalto.fi, marco.binninger@googlemail.com, olli.seppanen@aalto.fi

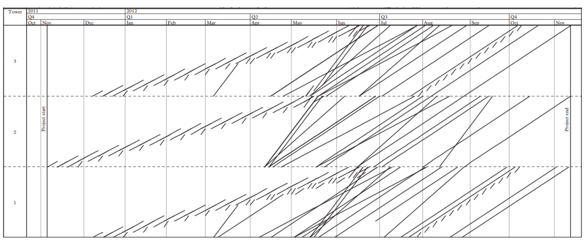
Abstract: In recent years the usage of digital tools in construction management has increased, and with it the need for automation and applications of *Artificial Intelligence* (AI). The basis for AI to work properly is clean and meaningful data. However, several publications have appeared in recent years documenting the lack of good data in the construction industry. Most of the previous studies discussing the use of AI in construction management do not take into account the assessment or definition of data quality. Therefore, this paper presents an experiment to measure data quality of three construction management methods: *Critical Path Method* (CPM), *Location Based Management System* (LBMS) and *Takt Planning and Takt Control* (TPTC). Since location-based techniques (LB), such as LBMS and TPTC, require information on a higher level of detail for input than activity-based techniques—especially in terms of connecting processes, times, and locations—the expectation is that they also provide better data as output. This lead to the following hypothesis: construction projects using TPTC or LBMS provide a higher data quality for AI than projects using CPM. The proposed experiment consists of five phases: Project Definition, Project Planning, Tool Selection, Data Entry & Export, and Evaluation. The experiment was carried out to find LB methods do indeed provide a higher score across all defined evaluation metrics. However, the authors identified several limitations for the presented results, with the strongest one being a small sample size of evaluated software solutions, which leads to a strong bias towards the individual implementations. Therefore, the next stage of the authors' research will be a more thorough execution of the experiment to verify the results.

Keywords: Data Quality, Artificial Intelligence, Construction Management, CPM, LBMS, TPTC

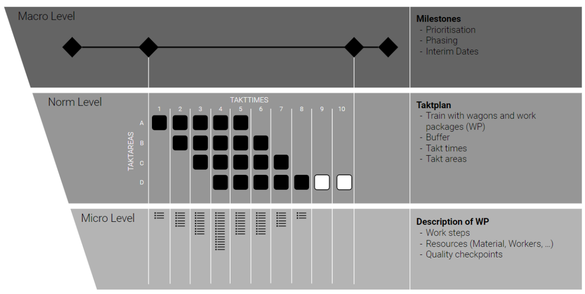
1 Introduction

There has been a growing interest in the usage of digital tools for construction management. Great effort has been devoted to the study of using *Artificial Intelligence* (AI) to automate processes within the construction industry. As reported by Abioye, Oyedele, Akanbi, *et al.* [1] the amount of papers published on *Machine Learning* (ML), Computer Vision and especially Optimisation has increased drastically over the last two decades. They reported that limitations in construction for those subfields

of AI, amongst others, are issues like incomplete data, scalability, and dealing with high-dimensional data. Further, Sacks, Girolami, and Brilakis [2] studied the use of *Building Information Modelling* (BIM) and AI in construction management and identified several challenges, which need to be overcome before AI can be widely used in the industry. AI solutions often are discipline-specific and lack multi-discipline collaboration. Many object relations and properties are not well defined in BIM and human intelligence is required in order to make sense of them. They discovered that data structures and representations often are optimised for the respective application to run and not for the use of AI [2]. To solve the issue of lacking data in construction management, Delgado and Oyedele [3] developed autoencoders to augment their data sets, which increased their model score by up to 11.5%.
Although several studies have indicated that data quality in construction is a limiting factor in terms of AI development and usage, little attention has been given to assessing and thus systematically increasing data quality in construction management [1]–[3]. In order to analyse this, it is important to know how construction projects are planned and managed. Olivieri, Seppänen, Alves, *et al.* [4] conducted a survey about the usage of the *Critical Path Method* (CPM), the *Last Planner System* (LPS), as well as *location-based techniques* (LB) in construction management across four countries and 532 projects. Their results show that CPM is the most used system for planning and controlling construction projects with a representation of 71%. However, the interest of LB systems has been growing and they were used in 40% of all considered projects. In recent years popular LB methods have been *Location-Based Management System* (LBMS) and, *Takt Planning and Takt Control* (TPTC) [5]–[7]. Figure 1 highlights, in (a) and (b) for LBMS and TPTC respectively, the structure created when adding the location information to time and process. Both LB methods have clear rules on how to integrate the locations of a construction site and how to connect it to all other relevant data points [5], [8]. While it is possible to add the location information when using an activity based method, like CPM, there is no clear procedure on how to select locations or on how to connect them to tasks or resources and thus there is no clear structure provided by the method itself.



(a) LBMS flowlines by Olivieri, Seppänen, and Denis Granja [5]



(b) Three levels of TPTC by Binninger [6] and Dlouhy, Binninger, Oprach, *et al.* [7]

Figure 1: Structure of LB methods LBMS and TPTC

Publications on the usage of AI in construction indicate that data quality in construction is a limiting factor. However, to the authors best knowledge, the impact of construction management methods on data quality has been scarcely investigated from the point of view of AI applications. Therefore, this work presents a method to assess data quality of construction management methods for AI applications. Because of the stricter rules on connecting data points by LB methods, this work further proposes the following hypothesis: construction projects using TPTC or LBMS provide a higher data quality for AI than projects using CPM.

The remainder of the paper is organized as follows: section 2 describes the method used to assess the data quality of construction management methods. Tentative results are presented in section 3 and section 4 concludes with a summary.

2 Method

This work presents an experiment to determine the data quality produced by construction management methods, which is illustrated in Figure 2. The experiment consists of five phases: *Project Definition*, *Project Planning*, *Tool Selection*, *Data Entry & Export*, and *Evaluation*. During the *Project Definition*

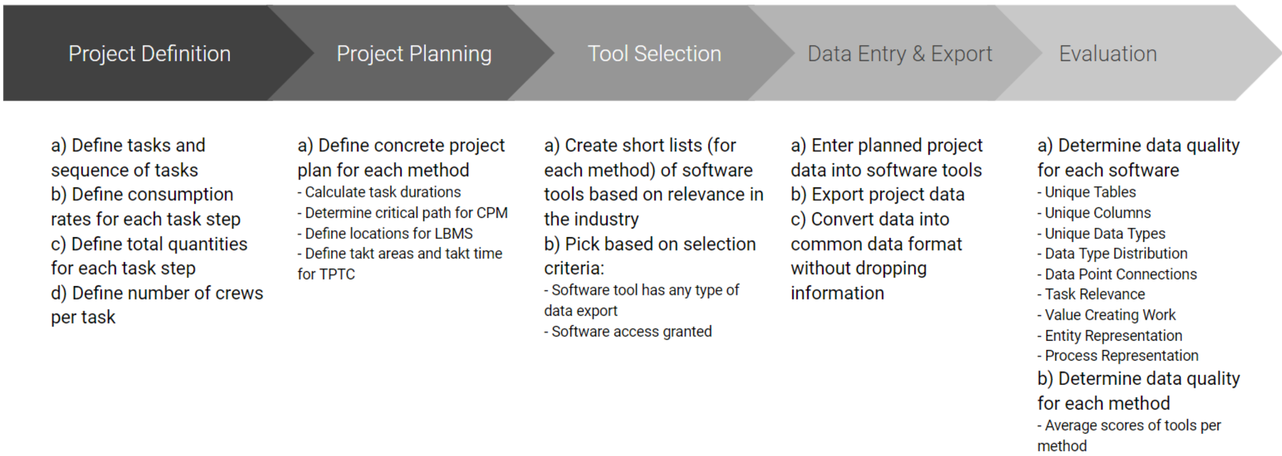


Figure 2: Experiment setup

phase, one project is defined as a foundation for the experiment. Therefore tasks and their sequence have to be set. The sequence of tasks describes the fixed order the tasks need to be executed in. To calculate the duration of each task, tasks should consist of at least one or more task steps, with each task step having a consumption rate, some quantity of materials, and a number of crews working on this task. Next, in the *Project Planning* phase, the base project has to get prepared for each construction management method. This means that the exact duration for each task is calculated, using the consumption rates, quantities, and number of crews. For CPM, the critical path needs to be identified by using the sequence of tasks. Additionally, the locations for LBMS, as well as the takt areas for TPTC, have to be defined. The third phase—*Tool Selection*—has the purpose of identifying software tools, which are used in practice to run construction projects using either CPM, LBMS or TPTC. Therefore, short lists of software solutions for each construction management method have to be created using search engines. Finally, the desired amount of software tools are selected from the short lists by applying the following selection criteria: the software tool has any type of data export, and it is possible to receive access to the software solution. Both criteria are necessary conditions, because otherwise there would be no data to evaluate. *Data Entry & Export* is the fourth phase and describes the process of using the previously selected software tools to plan the project according to the preparations done in phase one and two. Next, the project data has to be exported from the tools using their defined export functions. The authors suggest to convert all data sets into a shared data format, such as CSV, to simplify the evaluation phase. Finally, the *Evaluation* phase uses dimensions, elements as well as indicators, presented by Cai and Zhu [9], to determine the data quality. Cai and Zhu [9] defined five dimensions of data quality: *Availability*, *Usability*, *Reliability*, *Relevance*, and *Presentation Quality*. However, they argue from a computer science point of view for the goal of optimizing software solutions in the big data era, whereas this work aims to assess data quality created by construction management methods. Thus, several of their dimensions are being dismissed in the evaluation process, because they are only affected by the implementation of a software solution. The authors found the following elements to be relevant for this work: *Accuracy*, *Integrity*, and *Completeness* from the *Reliability* dimension, *Fitness* from *Relevance* dimension, and *Readability* and *Structure* from the *Presentation Quality* dimension. Cai and Zhu [9] describe the *Accuracy* element as a measurement of how well the data compares to actual reference values. Additionally, in the authors' opinion, this is reflected by the level of detail used by the data set. The *Integrity* of data is displayed by how precisely tasks, as well as definitions and relations of entities, on the construction site are represented. This is further supported by the *Completeness* element, which indicates, whether the data set describes the actual components in their totality. Indicators such as how much has the data already been summarized, or are appropriate data types being used, are important factors for AI applications. They stem from the *Fitness*, *Readability*, and *Structure* elements. Table 1 provides an overview of the Data Quality Elements and their respective Evaluation Questions and Metrics. Since the answers to those questions are very complex, subjective, and not very tangible,

Table 1: Evaluation Questions and Evaluation Metrics

Data Quality Elements	Evaluation Question	Evaluation Metrics
Accuracy, Completeness, Fitness, Structure	How precisely are the tasks described?	Unique Tables, Unique Columns
Accuracy, Completeness	How accurate are tasks linked to the actual work?	Task Relevance, Value Creating Work
Accuracy, Completeness, Integrity	How accurate are tasks linked to entities on site?	Entity Representation
Accuracy, Completeness, Integrity	How much of the actual process on site is represented in the data?	Process Representation
Structure, Fitness, Readability	How is the representation of data types?	Data Type Distribution
Structure, Fitness, Readability, Integrity	How well are data points connected to each other?	Connections

the authors derived eight metrics from evaluation questions. The metrics *Unique Tables* and *Unique Columns* simply represent the number of exported tables and columns. Which is supported by the basic idea of more columns means more information, which allows for more AI use cases. The *Task Relevance* metric describes the amount of features, which contribute to describing the broader tasks in a meaningful way. All value creating activities that are described by the data are counted in the *Value Creating Work* metric. All columns, which represent any entities on the construction site, such as locations, workers, or materials, are added up in the *Entity Representation* metric. More columns for those metrics, imply a higher level of detail and less abstraction, which provides more opportunities for feature engineering and thus richer AI solutions. The *Data Type Distribution* metrics provides an overview of all data types used in the data set. Finally, the *Connections* metric indicates how many references between data points can be found in a data set. All metrics are defined so that higher results mean higher data quality.

3 Results

The experiment described in section 2 has been conducted for the three construction management methods CPM, LBMS and TPTC. During the *Project Definition* phase a project for a six storey building consisting of seven tasks has been defined, which is displayed in Table 2. The tasks have been enriched by steps, consumption rates, quantities, and number of crews. Note that the consumption rates and quantities for this project are not taken from a real project.

Table 2: Project Definition: Tasks, Consumption Rates, Quantities, and Crews for a 6 storey building. *(*Reinforce and pour slab on metal deck)*

Subcontractor	Task	Steps	Consumption Rate [hrs]	Quantity [LF/SF]	Crews
Concrete	R/P SOMD*	Concrete	0.0219	60000	10
Concrete	R/P SOMD*	Reinforcement	0.0219	6000	10
Drywall	LAYOUT TOP TRACK	Layout Top Track	0.0073	60000	4
Drywall	LAYOUT TOP TRACK	Install Top Track	0.0073	6000	4
MEP	OVERHEAD MEP	OH MEP INSTALL	0.032	12000	2
Drywall	STUDS	2nd Pass Int Walls	0.0138	60000	3
MEP	In-Wall MEP Rough-in	Rough Inwall	0.016	60000	6
Drywall	Drywall Install	Hang Drywall	0.032	60000	10
Finishes	Finishes	Finish Materials	0.128	60000	15

Next, the project has been prepared for the three construction management methods. For CPM the initial task sequence has been selected as the critical path, since each task depends on the previous one. Locations and takt areas have been identified for LBMS and TPTC respectively. Building storeys have been selected as locations for LBMS. Since TPTC typically uses smaller locations, the authors decided on using 18 takt areas with a takt time of 4 days.

During the *Tool Selection* phase three short lists were created. While the short list of CPM contained 15 software solutions, both LB methods found less results, with LBMS only providing three, and TPTC only nine options. A few software solutions had to be discarded, because they did not provide a data export option. Others could not be considered for this experiment due to the lack of accessibility. However, one tool per method was found, matching all criteria.

The described project data has been entered and exported from all three software solutions. The LBMS tool required a data set conversion from XML to CSV. The other tools exported to CSV directly. The evaluation metrics, defined in Table 1 have been applied to all exported data sets. Therefore, tables and columns meeting the respective evaluation metric criteria, have been counted. The *Connection* metric takes all columns into account, which create a relation between data points. All columns, which contribute to describing the tasks, defined during the first two phases of the experiment, are summed up in the *Task Relevance* metric. This includes all processes on the construction site, which are relevant to the respective task, e.g. design or logistic. Next, the *Value Creating Work* metric adds up all columns of a data set, which describe the actual tasks and task steps. The *Entity Representation* metric counts all columns representing entities like workers, materials, or tools on the construction site. All features expressing ongoing processes on the construction site, such as the state of the tasks are added up in the *Process Representation Metric*. Finally, the *Data Type Distribution* metric accounts for all different data types used in a data set.

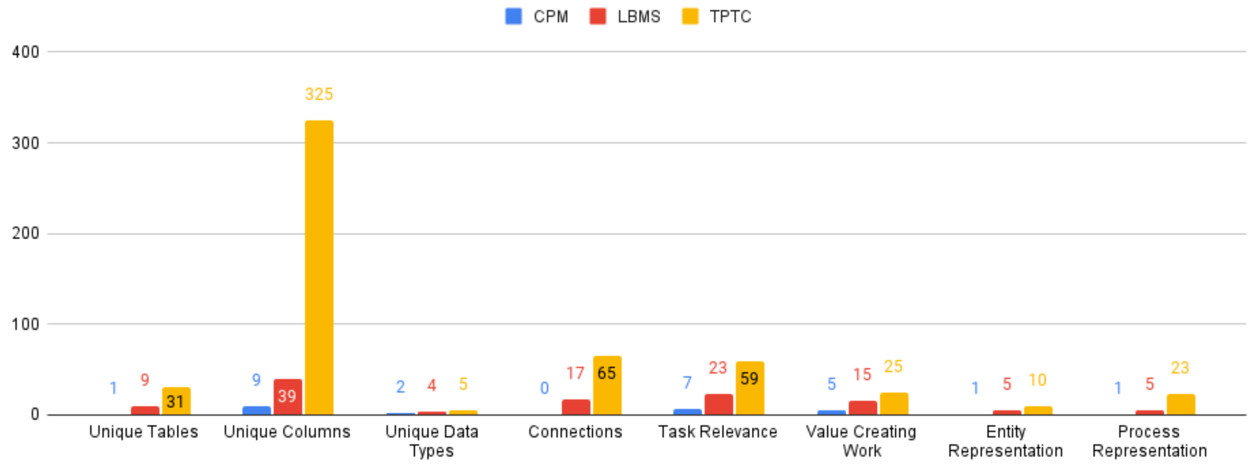


Figure 3: Results of Evaluation Metrics for one CPM, LBMS, and TPTC software solution respectively

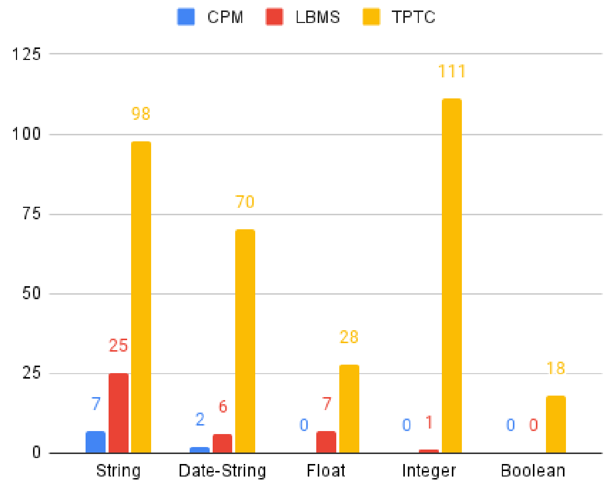


Figure 4: Results of Data Type Distribution evaluation metric for one CPM, LBMS and TPTC software solution respectively

Figure 3 presents the results of eight evaluation metrics for one CPM, LBMS, and TPTC software tool respectively. From this figure it can be seen that both LBMS and TPTC software scored higher than the CPM tool in every metric. Especially the results of *Unique Tables*, *Unique Columns*, and *Connections* indicate the structural differences of activity-based and location-based methods. The higher level of detail provided by LBMS and TPTC software, which was assumed based on Figure 1, can be seen through the better results in the *Task Relevance*, *Value Creating Work*, *Entity Representation* and *Process Representation* metrics. LBMS and TPTC tools also show a higher *Data Type Distribution* score than the CPM software solution, which can be found in Figure 4. Striking is the complete lack of numerical data in the data set of the CPM tool, which is preferential to categorical data.

While the results are a good first indication to support the hypothesis, the evaluation is only based on a single project and a single software tool per construction method. Thus the results are extremely biased towards the individual implementation of the respective software applications. Nonetheless, the results of this experiment support the proposed hypothesis—construction projects using TPTC or LBMS provide a higher data quality for AI than projects using CPM—to be true.

4 Conclusion

Based on the results, it can be concluded that the research into assessing data quality in construction management has been very successful and data quality of construction management applications can be measured. The proposed experiment works as a method for assessing data quality and the first performance of the experiment clearly shows differences between the data quality of projects using LBMS or TPTC over CPM. However, generalized conclusions cannot be drawn based on one software solution per construction method. Additional projects as well as expanding the experiment to include project controlling, would further strengthen the experiment. On the basis of the promising findings presented in this paper, work on the remaining issues is continuing and will be presented in future papers.

References

[1] S. O. Abioye, L. O. Oyedele, L. Akanbi, *et al.*, “Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges”, *Journal of Building Engineering*, vol. 44, p. 103 299, 2021.

[2] R. Sacks, M. Girolami, and I. Brilakis, “Building information modelling, artificial intelligence and construction tech”, *Developments in the Built Environment*, vol. 4, p. 100 011, 2020.

[3] J. M. D. Delgado and L. Oyedele, “Deep learning with small datasets: Using autoencoders to address limited datasets in construction management”, *Applied Soft Computing*, vol. 112, p. 107 836, 2021.

[4] H. Olivieri, O. Seppänen, T. d. C. Alves, *et al.*, “Survey comparing critical path method, last planner system, and location-based techniques”, *Journal of construction engineering and management*, vol. 145, no. 12, p. 04 019 077, 2019.

[5] H. Olivieri, O. Seppänen, and A. Denis Granja, “Improving workflow and resource usage in construction schedules through location-based management system (lbms)”, *Construction management and economics*, vol. 36, no. 2, pp. 109–124, 2018.

[6] M. Binninger, *Untersuchungen zum arbeitsfluss in getakteten bauproduktionssystemen aus perspektive von generalunternehmern*, 2022.

[7] J. Dlouhy, M. Binninger, S. Oprach, and S. Haghsheno, “Three-level method of takt planning and takt control–a new approach for designing production systems in construction”, *Proceedings (IGLC 24). Boston, USA*, 2016.

[8] J. Dlouhy, S. Oprach, M. Binninger, T. Richter, and S. Haghsheno, “Using takt planning and takt control in production projects–comparision of construction and equipment phases”, in *Proceedings of the 26th Annual Conference of the International Group for Lean Construction, Chennai, India*, 2018, pp. 16–22.

[9] L. Cai and Y. Zhu, “The challenges of data quality and data quality assessment in the big data era”, *Data science journal*, vol. 14, 2015.