Automation in Construction 124 (2021) 103564

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

A BIM-data mining integrated digital twin framework for advanced project management

Yue Pan, Limao Zhang *

School of Civil and Environmental Engineering, Nanyang Technological University, Singapore, 50 Nanyang Avenue, 639798, Singapore

ARTICLE INFO

Keywords:
Digital twin
Building information modeling (BIM)
Process mining
Time series analysis

ABSTRACT

With the focus of smart construction project management, this paper presents a closed-loop digital twin framework under the integration of Building Information Modeling (BIM), Internet of Things (IoT), and data mining (DM) techniques. To be specific, IoT connects the physical and cyber world to capture real-time data for modeling and analyzing, and data mining methods incorporated in the virtual model aim to discover hidden knowledge in collected data. The proposed digital twin has been verified in a practical BIM-based project. Based on large inspection data from IoT devices, the 4D visualization and task-centered or worker-centered process model are built as the virtual model to simulate both the task execution and worker cooperation. Then, the high-fidelity virtual model is investigated by process mining and time series analysis. Results show that possible bottlenecks in the current process can be foreseen using the fuzzy miner, while the number of finished tasks in the next phase can be predicted by the multivariate autoregressive integrated moving average (ARIMAX) model. Consequently, tactic decision-making can realize to not only prevent possible failure in advance, but also arrange work and staffing reasonably to make the process adapt to changeable conditions. In short, the significance of this paper is to build a data-driven digital twin framework integrating with BIM, IoT, and data mining for advanced project management, which can facilitate data communication and exploration to better understand, predict, and optimize the physical construction operations. In future works, more complex cases with multiple data streams will be used to test the developed framework, and more detailed interpretations with the actual observations of construction activities will be given.

1. Introduction

Rather than a simple virtual model or software, Buildings Information Modeling (BIM) can be regarded as a process of creating models with semantically rich information in a common data environment (CDE) to accelerate the digitalization in the Architecture, Engineering, Construction, and Operation (AECO) industry [37]. In particular, the cloud BIM has been developed to facilitate a higher level of cooperation among project participants with varying knowledge and backgrounds. It provides digital collaboration spaces (i.e., project servers, cloud-based systems) for information gathering, managing, and sharing in a unified and digital manner throughout the whole life cycle of a civil engineering project. That is to say, BIM can be seen as a data repository to store massive data gathered from data-rich objects, inputs, documents, sensors, building management tools, and others during project execution [39,43]. As the adoption of BIM grows, the amount of BIM data will increase exponentially, resulting in some characteristics of "big data" [34]. It is easy for BIM data files to reach a large size in dozens or hundreds of gigabytes [12]. For instance, the BIM project for an airport terminal with 548,300 m^2 can reach approximately 50 GB, which is saved within a scalable NoSQL database in a cloud environment [25]. This kind of heavily accumulated data captures details of the parametric model and executing process to offer affluent evidence for decision making, which is worthy of deep exploration to seek hidden knowledge and further enhance the value of BIM [36].

Great attention should be paid to the BIM data layer. Notably, the Industry Foundation Classes (IFC) developed by the International Alliance of Interoperability (IAI, also called BuilindSMART) is an open and neutral data schema to save digital building descriptions, mainly serving as a global standard for BIM data exchange [35]. However, most data mining (DM) techniques lack the capability in handling IFC directly, and a single IFC file containing basic object information is insufficient for resourceful decision making. Therefore, data preparation is an essential task to represent IFC schema and integrate external data related to BIM

* Corresponding author.
E-mail addresses: pany0010@e.ntu.edu.sg (Y. Pan), limao.zhang@ntu.edu.sg (L. Zhang).

https://doi.org/10.1016/j.autcon.2021.103564
Received 11 July 2020; Received in revised form 2 January 2021; Accepted 18 January 2021
Available online 31 January 2021

objects on a semantic level [28]. A more appropriate data format, which could be easily understood by computers, is expected to be generated to describe physical and conceptual entities along with relevant activities and logic. To be more specific, the IFC4 schema using *IfcProcess*, *IfcControl*, *IfcActor*, and others can map data about activities or processes of a construction project into the IFC standard. Then, important information associated with cases and events can be retrieved from IFC files and saved in Comma Separated Values (CSV) files. The output of the ideal data structure in CSV is called the event log, which is made up of sequential cases and events with typical attributes, like timestamp, activity, actor, and others [54]. It should be noted that a lot of valuable knowledge regarding project evolution will be embedded in event logs. A special focus can be on various DM techniques to exploit the growing availability of BIM event logs in a meaningful way, aiming to reveal valuable insights into the real executed processes.

Process mining, as a relatively young research discipline specialized in analyzing affluent evidence from event logs, can be performed to evaluate the process performance in an objective manner [58]. On the one hand, it is an automated process to discover and create process models conforming to event logs without human intervention, which offers visibility to make the actual process clearer and more comprehensible. On the other hand, it learns and analyzes the dynamic behaviors from the end-to-end process models, which provides evidence to diagnose the process and improve the workflows. Some studies have reported promising results from the successful implementation of process mining towards smart manufacturing management, such as to evaluate and predict the production flow [14], to identify the anomalous behavior and cyber-attacks [33], to analyze operators' behavior and knowledge [56], to determine the cause of underperforming blocks and provide guidance for improvement [38], and others. Similarly, we can extend the application scope of process mining into construction project management, which can eliminate the great dependency on expert judgment. In addition, process mining can be conducted on a monthly basis rather than an entire project period, and thus it can give month-by-month feedback and guidance in construction monitoring and optimization. Since the sequence of information from logs attaches a period to each value, time series analysis can also be taken into account to identify underlying trends and patterns evolving over time, which helps in fitting models to predict future events. There are many mature approaches for time series analysis involving complicated mathematical calculations, such as autoregressive integrated moving average (ARIMA), support vector machine (SVM), artificial neural network (ANN), and others [10]. Although these methods are extensively applied in the field of financial and business, they have not been well-integrated in the BIM data with a series of particular time intervals for knowledge discovey.

Since BIM plays a central role in information delivery and management across the whole project lifecycle, another point should be noted that BIM can support the synchronization of data gathered from IoT devices through communication standards (i.e., O-MI/O-DF, IFC models) [9]. The value of BIM from IoT integration perspectives lies in its powerful ability to create an information loop among different systems, participants, and phases [8]. Some studies have highlighted the effective use of BIM-IoT integration in improving construction and operational efficiencies, which could provide real-time feedback, draw up predictive planning, and others [7,52]. However, most of these researches are still at the conceptual stage [48]. It is known that the emerging BIM-IoT integration paves a new way of collecting, storing, and exchanging data, and thus we intend to merge it into digital twins for digital revolutions in a practical application. More specifically, the deployment of IoT largely relies on a variety of devices to detect the actual operations in real time. The BIM platform has the potential synergy with IoT to synchronize and store various data. After a large amount of data is interpreted and processed into the proper structure, it can be utilized to build the digital twin by mirroring the physical system in an isolated virtual environment. For a better understanding of the true project state., there is a two-way dynamic mapping in the digital twins to create a cyber-physical closed-loop [41]. On the one hand, realistic inspection data gathered in the physical part will be sent to the virtual part for processing and analyzing. On the other hand, the virtual model will simulate real-world behavior in a digital way and output actionable insights by a variety of DM algorithms, which has the practical value in dynamically adapting the physical system to the changing environment.

The major objective of this research is to develop a digital twin-based framework through integrating BIM, IoT, and DM, which acts as a practical method to control and optimize the complex construction process under a high degree of automation and intelligence. It aims to bridge the gap between data science and construction process analysis. For one thing, the 4D BIM visualization can be synchronized with the IoT data. For another, we resort to advanced DM approaches to deeply explore BIM construction event logs from IoT data in terms of process modeling, bottleneck diagnoses, and progress prediction in an objective manner. It leaves three main research questions to be answered: (1) How to design a rational architecture of digital twin with the help of BIM, IoT, and DM to support intelligent process control and project management; (2) How to automatically construct the high-fidelity virtual model as a digital replica of the physical object, which can simulate the as-happened construction process; and (3) How to fully mine the large amount of BIM event logs delivered from the physical to the virtual side in both the current and future perspective, aiming to detect possible risk and predict the construction progress. The proposed BIM-IoT-enabled digital twin in a data-driven loop is proven effective in pairing the virtual and physical model for advanced construction management, which has been verified by a case study about a real-world construction project. From the level of knowledge, experiments based on process mining and time series analysis have been done in the virtual model, contributing to converting data into the strategic value of information for process understanding and trend prediction in the complex construction project. From the level of application, the usage of DM techniques, in turn, gives continual feedback about developing and adjusting project planning and staffing to maximize efficiency, reliability, and sustainability on sites, which can efficiently reduce the dependency of decision making in project management on expert knowledge and subjective judgment.

The rest of the paper is structured as follows. A literature review on digital twin and its potential use in the construction industry is provided in Section 2. The developed digital twin architecture is described in Section 3. Critical DM methods for analyzing and validating the virtual model of the digital twin are introduced in Section 4. A case study in a practical construction project is performed to show the effectiveness of the proposed architecture in Section 5. The time series data and algorithms are discussed in Section 6. Conclusions and future works are outlined in Section 7.

2. Literature review on digital twin

The term "digital twin" initially proposed in 2003 is not a new concept, but it gains increasing popularity in the current industrial revolution 4.0 (digitalization). More specifically, the re-emergence of interest in digital twins is largely inspired by the study from the National Aeronautics and Space Administration (NASA) to continuously simulate, forecast, and evaluate the spacecraft state, aiming to mitigate the degradation and failure in the vehicle [15]. Afterward, digital twins have been increasingly recognized by more and more researchers, and the Gartner research firm in 2018 even predicted the idea as one of the top ten most promising technology trends over the next ten years [51]. In general, digital twins refer to a mirror and digital depiction of the actual production process, which can imitate all aspects of physical processes under the integration of physical products, virtual products, and relevant connection data.

To date, digital twins play a crucial role in pursuing the deep cyber-physical integration of intelligent manufacturing towards a greater level of flexibility, adaptability, and predictability in production management. The digital twin system has been widely applied in product design and production, which can assist in understanding customer demands quickly, identifying or even predicting weaknesses in models early, controlling production processes to respond to the changing environment timely, and making valuable suggestions to optimize plant operation and maintenance before failure occurrence [32,44,49,53]. Moreover, some leading companies, such as General Electric (GE), Siemens, British Petroleum (BP), and Airbus, have implemented digital twins in the practical production and relevant patents for production technical innovation [50]. Due to the success of digital twin in manufacturing, some efforts have been devoted to building the cyber-physical model for supporting digital development in the construction industry. It has been proved that a system architecture of digital twin potentially has a wide application prospect in representing, predicting, and managing the current and future conditions of the infrastructure itself, built environment, or city assets. For instance, Yuan et al. [57] monitored the temporary structure by the bi-directional coordination between physical and virtual systems, where the virtual components were built by the real-time data from sensors in the physical part to make early warning and immediate instruction for structural failure prevention. Srewil and Scherer [46] utilized data from Radio-frequency identification (RFID) to map the actual process into the virtual model, which could provide a comprehensive solution for real-time construction process monitoring. Linares et al. [26] adopted the advanced equipment of an Augmented/Virtual Reality (AR/VR) coupled with sensors to capture images or videos on the physical site, which was helpful in safety monitoring, risk warning, and remote instruction. Lu et al. [27] designed a digital twin at both the building and city levels following data integration, synchronization, and analysis, in order to realize anomaly detection, ambient environment monitoring, maintenance optimization and prioritization, and energy planning. To sum up, the superiority of digital twin lies in its value-added services in automatic data collection, conceptual development, dynamic analysis, problem diagnosis and optimization for smart design, operation, control, and maintenance. In other words, real-time data derived from the physical products are the basis to align the real world into the virtual parts. Through automatically detecting issues and evaluating performance ahead of time, optimized solutions can be formulated in a data-driven manner and put into operation in time to bring benefits of improved reliability and efficiency. Thus, there are reasons to believe that the concept of digital twins will become increasingly important in the rise and progression of the construction industry revolution.

From these above-mentioned pieces of literature, it can be found that the effectiveness of the virtual part largely depends on the great volumes of collected data and the corresponding data analysis. Commonly, IoT devices are responsible for data acquisition, which can capture various information about the real-time and recordable status of the current operations for further utilization [48]. Since BIM has evolved into an open platform for information sharing and management, it is able to synchronize with multiple data sources from IoT. That is to say, the integration of BIM and IoT can store and update a variety of information, including object properties, site and facility conditions, physical measurements, time series data about the progress, and others, which offers rich data sources for DM-supported knowledge learning and decision making. Hence, it can be considered to establish a well-defined framework of a digital twin based upon BIM, IoT, and DM, which can be presented as a "physical-data-virtual' paradigm for higher interoperability, automation, and intelligence in delivering smarter construction services [1]. In existing research, the developed digital twins mainly provide a crucial and analytical edge to BIM-IoT integration. For instance, Lu and Brilakis [29] automated the geometric modeling in the digital twin part for existing reinforced concrete bridges from 3D cloud points, which could reach a relatively high spatial accuracy. Stojanovic et al. [47] reconstructed and visualized the captured state of the built environment using the basic data from 3D point clouds and related IFC, which could be helpful in enhancing collaboration, decision making, and forecasting among facility management stakeholders. Shim et al. [45] adopted the 3D scanning technology to duplicate an existing bridge structure as the object-based digital twin model, from which data about damage and repair history could be analyzed to orient long-term strategies for bridge assessment and maintenance. However, they mostly emphasize the 3D geometry and model evaluation in digital twins, while less attention has been paid to knowledge discovery from the DM layer.

It should be noted that BIM-IoT integration can provide a constantly updated and rich data influx about both the functional and performance features of a facility [13], but BIM itself lacks data manipulation capabilities to evaluate and predict the real-time status of assets, processes, systems, or even services. To address issues in information integration and data analysis, Cheng et al. [7] connected various kinds of information from the as-built BIM models and IoT sensor networks, which were used to train machine learning algorithms (SVM and ANN) to make predictive maintenance planning for building facilities. Ma et al. [30] adopted BIM and GIS in an integrated manner to provide related geometric, attributive, and spatial data, and then Reliability Centered Maintenance (RCM) algorithms were performed on these prepared data for decision-making on equipment maintenance of business parks. In other words, DM techniques can offer a wealth of digital insights into the collected data for making more informed and proactive decisions in condition assessment, prediction, and improvement, which no longer rely on subjective judgment with bias and uncertainty. Since the BIM-IoT-enabled digital twin will contain a lot of data with hidden knowledge, appropriate DM methods need to be performed to realize the full value of data for two major purposes. For one thing, DM can promote bidirectional interaction in the physical and cyberspace. For another, DM helps to continuously guide and adjust the construction process towards the project goals using actual data rather than observation or intuition. Despite the importance of DM approaches, the integration of BIM, IoT, and DM for digital twin is still in infancy. For this concern, we intend to develop a data-driven framework of a digital twin, which can be strategically leveraged and integrated with the BIM, IoT, and DM to yield significant value in intelligently improving construction efficiency, collaboration, and reliability.

3. Architecture of the proposed digital twin

Based upon the great amounts of IoT data from the BIM-enabled construction project, a data-driven digital twin framework is put forward to build a closed-loop between the physical and digital world. Fig. 1 presents the conceptual architecture of the digital twin, which can take effect throughout the project life cycle for smart construction monitoring and management. Noticeably, it is an integration of BIM with real-time data collected by IoT devices and knowledge extraction from data analytics, which is comparatively a new development. The workflow of the proposed digital twin incorporating BIM, IoT, and DM can be briefly presented below.

To begin with, the unmanned aerial vehicle (UAV) equipped with the 3D Light Detection and Ranging (LiDAR) can deliver IoT services from great heights over the construction site [23]. It takes 3D point clouds to sense and act upon the actual (as-built) environment for real-time operational monitoring. Subsequently, this inspection data is sent to the BIM cloud system for storing. Cloud storage offers a large resource pool to address the problem of information overload [12]. It can be seen from Fig. 1 that the BIM cloud performs as a bridge of the physical-cyber system to continuously collects the comprehensive set of information from the physical entity and send data to the virtual part. To make full use of the point cloud, it is compared with the as-planned IFC by a tool named "Real-Time and Automated Monitoring and Control (RAAMAC)" in the BIM server (https://bimserver.org/). The developed tool is responsible for identifying and communicating discrepancies between actual and planned performance, resulting in as-built IFC for the purpose of automated construction progress monitoring [11,16]. However, IFC saving the digital building description is in a plain text file, which is

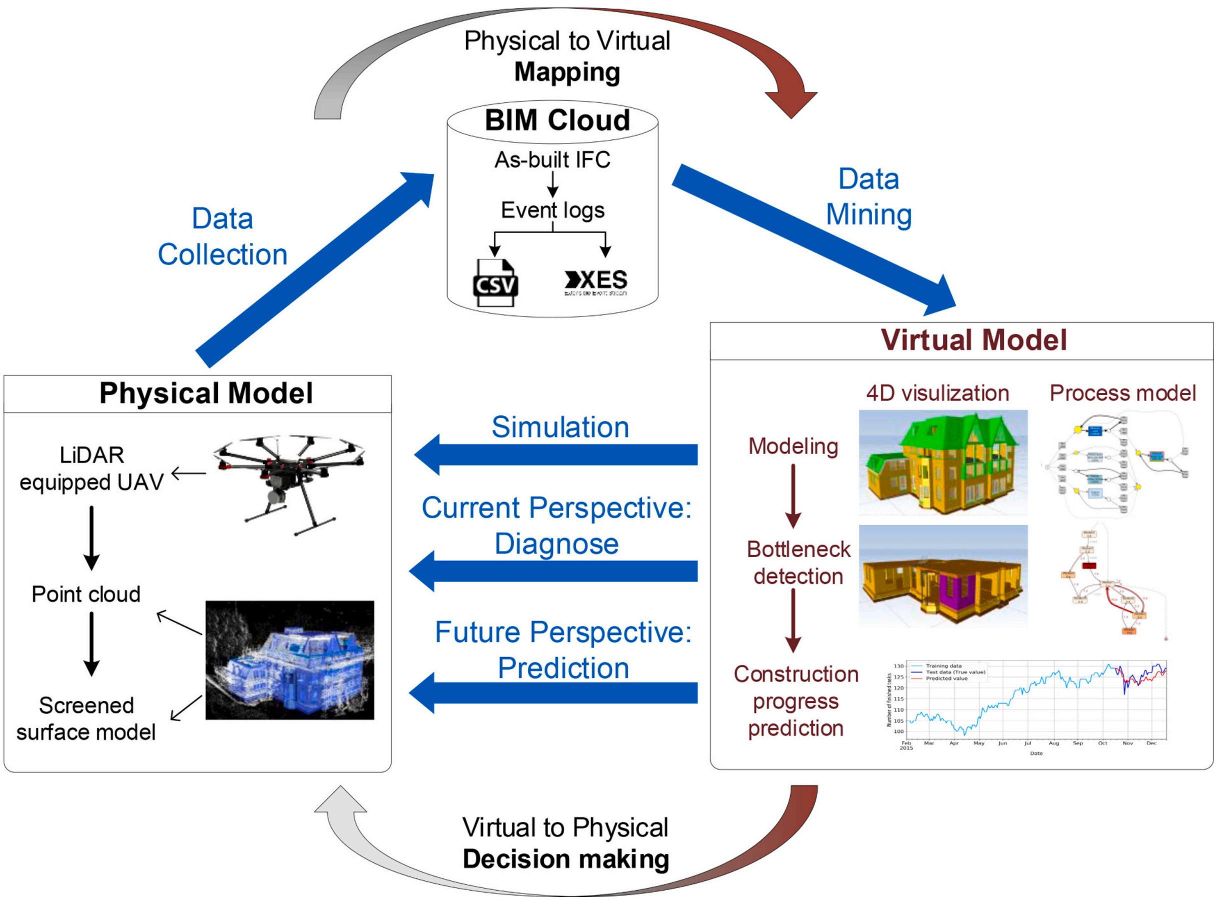

Fig. 1. Architecture of the proposed digital twin for a BIM-enabled construction project.

unreadable by DM algorithms. As a solution, another existing tool named "IFC Logger" [22] is employed to automatically parse useful data from IFC, such as construction tasks, workers, time, and others, which can output event logs in a comprehensible form for computers. To further ensure the data quality, data cleaning methods are conducted to remove noise. Lastly, the latest and prepared data gathered via IoT devices offer opportunities to pair physical entity into the high-fidelity virtual models along with vivid simulation, such as the 4D model and refined process model. Various DM techniques are applied in the virtue of digital twin integrating large data to realize process modeling, bottleneck diagnoses, and progress prediction automatically, which can return positive and timely feedback to managers. For instance, the 4D model in the combination of the 3D model and construction schedule owns a strong capacity in information visualization. As for the process model, it provides a concise and graphical representation of the complicated process, which demonstrates the practical implication for comprehending and managing the workflows and collaboration in the construction phase.

As elaborated in Fig. 1, the knowledge discovery and reasoning in the virtual part are mainly conducted from two views. On the one hand, process mining is adopted to provide a current perspective of the construction project implementation. A better understanding of workflow and collaboration can be realized from the discovered process model. Moreover, possible bottlenecks arising in the actual process can be detected easily, and thus response measures can be taken to avoid these unnecessary delays before occurring. On the other hand, time series analysis is performed to intelligently measure and predict the successive construction progress from the future perspective. Managers can keep abreast of the workers' current performance and the related trend. Since these predictions from the updated information provide directions for controlling and improving the construction work, they should be fully utilized to draw up reasonable plans and adjustments at an early stage. In other words, the prominent advantage of data analysis in the digital model is that it helps in exploring observed data timely and automate strategic decisions for process optimization, and thus managers no longer depend too much on expert experience and domain knowledge. The feedback can be delivered back to the physical side in time to dynamically regulate the construction scheduling and worker arrangement. In short, the developed digital twin architecture under the inclusion of BIM, IoT, and DM techniques realizes the remote and efficient interaction between physical and virtual objects, allowing for smart construction process management and assessment.

4. Model analysis and validation

The methodology of this research is based on the concept of design science [19,31], whose objective is to develop an innovative and purposeful artifact named the digital twin to provide technology-based solutions for specific problems in the construction industry. The research process is basically structured in a build-and-evaluate loop with three major steps, namely the model construction, model implementation, model evaluation. The loop is performed on the designed digital twin several times to test and assess its utility and quality, allowing for continuous improvement on the artifact to satisfy the requirements of the problem. To begin with, a new artifact is conceptually designed and formally represented targeting a heretofore unsolved issue. Then, with the stepping into the digital era, various digital technologies (i.e., BIM, IoT, DM) are applied as the intellectual and computational tools in implementing the artifact in a more effective and efficient manner, which are outstanding in extending the boundaries of human cognitive and problem-solving. Finally, some commonly-used metrics are calculated to mathematically evaluate the designed artifact, which, in turn, give informed feedback and solution to the model construction phase. To sum up, the research contributes to developing and presenting a novel and useful digital twin under the combined knowledge of BIM, IoT, and DM to address unsolved problems for smart construction project management, which can eventually guide the model implementation in practical engineering. To reach the desired conclusions, the proposed research framework will be conducted in a case study, which can intuitively show its feasibility and explain its implication. Notably, a great challenge from the research method termed design science tends to remain in exploring and validating the established model from the data layer, which may impose difficulties in affording actional insights to drive the project management towards the increased automation and intelligence. In this regard, special focus should be put on the knowledge discovery in the high-fidelity virtual model of the digital twin. From the current perspective, process mining tools can automatically identify and diagnose processes from event logs to comprehend the actual construction execution. From the future perspective, time series analysis coded by Python can estimate the work progress using a series of previously observed data. The adopted DM methods in the virtual model along with the validation metrics are theoretically introduced below.

4.1. Current perspective: Process discovery and diagnosis

4.1.1. Process mining

By extracting and exploiting the process-related information embedded in construction event logs, process mining can provide a holistic and deep insight into the current process to examine sequences of activities taken by actors. As a result of the digital transformation, actual process deviations and bottlenecks can be easily identified from the discovered end-to-end process, which will further support the potential improvements towards greater operational efficiency. Particularly, automated discovery of the process model by proper process mining algorithms is the most basic task, which only takes event logs with no prior information as input and then returns results in a visually structured and comprehensive process graph. This discovered process model can, therefore, be deemed as a special virtual twin of the physical part to capture and explore data from logs. It is noteworthy that the functionality of the process model is far more than simulating the complex construction process, which is also applicable in evaluating and forecasting construction performance. Herein, we adopt two desired modeling languages named Petri net and business process modeling notation (BPMN) as an abstract representation of the virtual model. Both graphical notations can describe the order dependencies between executed activities by various flow logics, including sequences, choices, concurrencies, and loops, which can be mutually translated under certain mapping rules [42]. The obvious difference is the place on behalf of an event between two activities in the Petri net, while BPMN has no need for these events to connect activities. To be specific, a Petri net can be defined as a triplet (P, T, F), where P and T are a finite set of places and transitions, respectively, and $F \subseteq (P \times F) \cup (T \times F)$ stands for a set of directed arcs (flow relation).

Inductive miner [24] is one of the leading algorithms in automatically constructing sound virtual process models, which can well handle less-frequent behavior to describe correct causal dependencies from event logs. It recursively splits the original event logs into sublogs according to four operators, namely the exclusive-choice operator ($\times$), sequence operator ($\rightarrow$), parallel operator ($\wedge$), and redo-loop operator ($\circlearrowleft$). The output is the process tree replaying all behaviors seen in the log to preserve the model soundness, which is an abstract representation of a sound block-structured workflow net with a leaf node referring to a single event and a non-leaf node denoting an operator [20]. For instance, the inductive miner can produce a process model expressed as $Q = \rightarrow(a, \times(\wedge(b, c), e), d)$ to replay process in an event log $L = [<a, b, c, d>^3, <a, c, b, d>^2, <a, e, d>]$ recording 6 cases and 23 events [54]. Then, the process tree can be easily converted into an equivalent Petri net and BPMN. Due to the quality, flexibility, and scalability of the process model from the inductive miner, its important application is the conformance checking to identify undesirable deviations between the virtual process model and the corresponding observations in the actual situation. Notably, the captured discrepancies help in not only judging the great alignment of activity sequences, but also suggesting proper adjustments of the virtual model to make it closer to reality.

A concern is that the inductive miner is likely to return intricate and misleading "spaghetti-like" models. As an alternative solution, another algorithm termed the fuzzy miner [17] can refine the complicated process to a desired level of granularity using a "map" metaphor. To distinguish the frequent activities and paths from infrequent ones, the fuzzy miner calculates a significant weight for each element and a correlation weight for each edge [2]. Two metrics named significant and correlation are utilized to describe the importance of behavior in a certain process, where the significance is measured by the frequency of events to facilitate model simplicity, and correlation is calculated by the distance of events to determine behavior aggregation and abstraction. Accordingly, the fuzzy model can retain all highly significant behavior, aggregate insignificant behavior with high correlation, and delete lowly correlated behavior, resulting in a good approximation of complex behavior at a relatively fast speed. In regard to the process analysis, the advantage of the fuzzy miner lies in its diagnose ability, which can intuitively project bottlenecks into the current process map under the consideration of frequency and duration attached in each event. It has been proved useful for bottleneck detection in practice [18,21,40]. What's more, animation movies based on the fuzzy miner provide a powerful tool in visualizing the bottlenecks, which assist to better explain and resolve possible delays for flow time reduction in the actual process.

4.1.2. Process model evaluation

Since the discovered process model acts as a virtual replica of the real process to support construction improvement, it is of necessity to validate the effectiveness of the model derived from process mining algorithms. Herein, three concrete indicators named fitness, precision, and generalization in Eqs. (1)–(3) are deployed to comprehensively evaluate the model quality [5,6]. To be specific, fitness quantifies the ability of the process model to replay activity sequences in event logs. As seen in Eq. (1), alignment costs are given for skipping and inserting activities, which are not corresponding to logs. Precision in Eq. (2) associated with underfitting calculates the fraction of behavior allowed in the process model, which is not observed in the event log. Clearly, poor precision roots in too many different behaviors between the physical and virtual parts. Rather, generalization defined in Eq. (3) is related to overfitting, which estimates how generic the process model is to describe unknown behavior not limited in the event logs. Greater generalization ability is confirmed when more parts of the discovered process model can be frequently visited. All three indicators are in the range of [0,1], and the value closer to 1 implies better quality. There is an obvious trade-off in underfitting and overfitting, which means that a model with more exact behavior probably owns a greater fitness or precision but a lower generalization.

$$f = 1 - \frac{fcost(L, M)}{move_L(L) + |L| \times move_M(M)} \tag{1}$$

where L stands for the event log, M refers to the process model, $fcost(L, M)$ represents the total alignment cost for L and M, $move_L(L)$ and $move_M(M)$ are the cost in L and M, respectively. Its denominator measures the minimal cost when no match is in L and M occurs.

$$p = \frac{1}{|E|} \sum_{e \in E} \frac{|en_L(e)|}{|en_M(e)|} \tag{2}$$

where e represents an event with a sequence of activities in L, $|E|$ is the number of events in L, $|en_L(e)|$ is the actual number of activities in L, and $|en_M(e)|$ is the number of activities enabled in M.

$$g = 1 - \frac{\sum_n \left(\sqrt{|executions|}\right)^{-1}}{|n|} \quad (3)$$

where |execution| is the number of executions of certain parts of the process tree, and $|n|$ is the number of nodes in the process tree.

4.2. Future perspective: Process prediction and analysis

4.2.1. Time series prediction

The sequence data sent from the physical model to the virtual one is ordered in clearly defined time components, which is the time series data to carefully track the evolution of construction work. It is common to carry out proper algorithms for pattern discovery in the time series data, which are likely to persist in the future. That is to say, the virtual model can be explored from a future perspective through examining characteristics of changes and predicting the coming amount of finished tasks, which can potentially guide the construction schedule and optimize the workflow in turn. Particularly, the Autoregressive Integrated Moving Average (ARIMA) model [3] is one of the most popular statistical methods to tackle time series data. Eq. (4) defines the ARIMA model to specify the current observation in terms of the linear relationship with past values, which can be decomposed into three components: autoregressive part (AR), integrated part (I), and moving average part (MA) with three non-negative parameters p, d, and q, respectively. To be specific, AR (p) describes a regression involving dependencies between the current observation and the observations over a prior period, which means the variable of interest is regressed on its own lagged values. I (d) identifies the times in differencing the observations to ensure a stationary time series with constant mean and variance over time. MA (q) provides a regression error in a linear combination of error terms, which takes into consideration dependencies between an observation and a residual term from the moving average to the lagged observations.

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)(1-L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t \quad (4)$$

$$L^k X_t = X_{t-k} \quad (5)$$

where t is the index, L is the lag operator provided in Eq. (5), X_t represents the time series data, ε_t refers to the residual, ϕ_i and θ_i stand for the numerical coefficient for the value associated with the ith lag in the AR and MA mode, respectively. Besides, p and q are the order of the AR and MA model, respectively, and i denotes the degree of difference.

It should be noted that ARIMA is primarily proven useful in analyzing univariate stochastic time series. Indeed, values for every period are possibly influenced by not only past periods but also one or more outside factors associated with each time period. Therefore, it is convinced that the model forecasting performance can be raised in view of some extra explanatory variables in categorical or numerical form. In this regard, the multivariate ARIMA termed the ARIMAX model is developed to integrate covariates into the ARIMA model using Eqs. (5)–(7), which is the variation of Eq. (4) [4]. Specifically speaking, the ARIMAX model is fully capable of handling the time series of interest and its orders along with additional inputs called the exogenous variables (augments).

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)(1-L)^d (X_t - m_t) = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t \quad (6)$$

$$m_t = c + \sum\nolimits_{i=0}^{b} \eta_i y_{t,i} \quad (7)$$

where L is the lag operator from Eq. (5), $y_{t,i}$ is a set of exogenous variables affecting the time series, η_i is the weight of exogenous variables fitted based on the model selection, and b is the size of the set of exogenous variables.

4.2.2. Model selection and evaluation

In the pursuit of promising model performance, how to figure out proper order parameters for the ARIMAX model becomes the main priority. The most intuitive method is to read the correlogram plot of the autocorrelation function (ACF) and partial autocorrelation (PACF) by Eqs. (8), (9), respectively. Specifically, ACF calculates the autocorrelation between an observation X_t and the lagged observation X_{t-k}, while PACF is the correlation in X_t and X_{t-k} conditioned on observations between these two observations. However, when the data is in high complexity, it could be a little confusing to determine parameters directly by viewing the decay from plots. Thus, a more effective method called the grid research can be utilized to iteratively run the developed ARIMAX model on multiple combinations of p, d, and q, and then make the comparison of model performance based on the criteria of goodness-of-fit, namely the log-likelihood, Akaike information criteria (AIC), Bayesian information criterion (BIC). Especially for AIC and BIC in Eqs. (10), (11), they have penalized likelihood with similar expressions, and the major difference is that BIC penalizes the model complexity more heavily. Regarding model selection, we prefer the fitting model with higher log-likelihood and lower AIC and BIC.

$$\varphi_k = corr(X_t, X_{t-k}) \quad (8)$$

$$\varphi_{kk} = corr(X_t, X_{t-k} | X_{t-1}, \ldots, X_{t-k+1}) \quad (9)$$

where $k = 0, 1, 2, \ldots$ represents the lag.

$$AIC = -2logL + 2(p + q + k + 1) \quad (10)$$

$$BIC = -2logL + (p + q + k + 1)log(n) \quad (11)$$

where p and q are the parameters of AR and MA model, respectively, L denotes the likelihood function, k represents the number of parameters in the model, and n stands for the number of data points.

Besides, the fitting model determined from the training set needs to perform a forecast on the test set to return continuous values. For comprehensively assessing the quality of predictions, two basic evaluation metrics named Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are adopted in comparison of the paired true and predicted value on the test set. MAE presented in Eq. (12) is an arithmetic average of the absolute errors in a set of predictions. Although MAE is easy to understand where individual differences have equal weight, it fails to alert very large errors. To deal with the issue, RMSE in Eq. (13) is expressed in a quadratic scoring rule to measure the average magnitude of errors. RMSE can make large errors more noted through assigning a higher weight to them. The minimum value of both MAE and RMSE is 0, and the smaller value indicates better prediction performance of the fitting model.

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n} |y_i - \hat{y}_i| \quad (12)$$

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (13)$$

where n is the number of data points, y_i is the predicted value, and $\hat{y}_i$ is the true value.

5. Case study

The conceptual architecture of the digital twin proposed in Section 3 is performed in a real-world construction project to verify its effectiveness. Special emphasis is put on the major aspects of data analytics in the virtual replica. Following the methods of model analysis and validation in Section 4, several experiments are conducted for process discovery, diagnosis, and prediction, which largely depend on real-time data to better control and optimize the construction process instead of expert knowledge and judgment.

5.1. Data acquisition and description

This case study is implemented in a dataset about an actual BIM-enabled construction work of a three-story house in the Netherlands, which has already been prepared by Schaijk from Eindhoven University of Technology [55]. That is to say, data acquisition based on the IoT-based process has been finished by the previous study. Our work is to perform the developed digital twin framework in this existing dataset about a project carried out as a joint effort of 11 workers from Feb 2015 to Dec 2015. To make the process of data acquisition clearer, a brief introduction about it is given below. A UAV carrying the LiDAR scanner is taken as the IoT device. That is because the laser scanning is less susceptible to the effects of the outdoor environment, which gains dominance over the traditional photo scanning. The UAV flies above the construction site covering most parts of the building surface and surrounding space during the project, in order to efficiently capture scanned-surface models and the current operation status represented by high-quality point clouds in real-time. It is important to emphasize that the BIM cloud storage system is essentially used to store and manage these IoT data in great volumes. The tool "RAAMAC" in the BIM server helps to parse the information in point clouds and convert them into the desired IFC, while the tool "IFC Logger" further translates the IFC file into the event log as a collection of cases. That is to say, point clouds are automatically uploaded, saved, and maintained in the BIM cloud to create a real-time database, which can be accessed by different users and shared between the physical and virtual sides. In the meantime, real-time information regarding cases and events can be extracted from the IFC and organized in the event log. As is known to all, the event log is the properly formatted time series data with multiple attributes concerning events, ordered cases, and their associated properties to trace detailed flows of construction. All the crucial preliminary work in data acquisition has been done. Based on these prepared data, we build a data-driven digital twin and mainly focus on one of the most important layers in the system called data analytics.

It is noteworthy that event logs are the output to track the as-happened construction process in machine-interpretable formats, including CSV and eXtensible Event Stream (XES). Process mining is especially used to discover knowledge from such data, which provides a new way of monitoring and improving the process. To be more specific, one event log describes a process made up of several cases, while one case occurs based on a sequence of ordered events (tasks). In this case study, the extracted CSV file contains 26,970 lines and 5 columns, where each line corresponds to a specific construction event and each column stands for an attribute. Table 1 shows an example of the event log data, where "IfcClass" is regarded as the case identifier. Events with the same name in the attribute "IfcClass" belong to the same case and have the same properties. For instance, "IfcSlab" can donate occurrences of slabs. In total, the case owns 13 unique types of "IfcClass", among which "IfcSlab", "IfcWall", and "IfcColumn" are the three key cases comprising the largest number of tasks (>3000). "TaskName" stands for a well-defined event in the construction process. In terms of "Worker", it refers to a certain worker to execute an event. There are 11 different workers participating and collaborating in this project, and workers 7, 1, and 3 are the top three most hard-working ones to carry out the most tasks. The last two attributes named "TaskStart" and "TaskFinish" are the timestamp to state the sequences of events related to a case. In short, this prepared event log in the size of 26,970×5 is the data basis for constructing a digital twin, which needs to be deeply explored using advanced DM techniques. Relying on the high level of bidirectional coordination between the physical and virtual structures, it is expected to bring potential benefits in the timely service of knowledge discovery and reasoning for process optimization purposes.

Table 1
Example of continuous records from construction event logs in the CSV format.

IfcClass	TaskName	Worker	TaskStart	TaskFinish
IfcSlab	Casting channel plate	Worker11	4/3/2015	5/3/2015
IfcSlab	Casting channel plate	Worker2	5/3/2015	6/3/2015
IfcWall	Framing lift walls	Worker1	6/3/2015	7/3/2015
IfcBeam	Steel beams	Worker1	6/3/2015	8/3/2015
IfcBeam	Steel beams	Worker1	6/3/2015	8/3/2015

5.2. Modeling of construction process

The prepared IFC and event log associated with day-to-day operations in the construction phase are accessible in the cloud database, which can be employed to recreate and simulate the progress in a virtual environment. In the context of cyber-physical synchronicity, we intend to build digital entities as a reflection of the actual activity sequences under ideal accuracy and update them through dynamic reconfiguration. The virtual model plays a crucial role in better simulating and understanding the construction logistics, which can then communicate closely with the physical system based upon their comprehensive data analysis. Herein, we perform two ways of building the virtual counterparts incorporating temporal information, namely the 4D model and process model, which are introduced below.

For one thing, the data-rich 4D model can synchronize with IoT data, which links the traditional 3D geometrical model with timelines to produce a digital description of the current project status. The clear visual context is established by importing IFC files generated based on point clouds. Moreover, animations with great visibility and transparency can also be performed to effectively imitate the execution of physical activities over the notion of space and time, particularly targeting at a continuous process monitoring and simulation for further investigation. In consequence, some schedule problems can be disclosed at an early stage to reduce unwanted conflicts and failures of the project before it occurs. Fig. 2 takes the constructed as-built models at the end of Feb, May, Aug, and Dec as examples to reveal how the construction work proceeds as time passes. Especially for Fig. 2(d), it can be observed that the virtual model and its corresponding point clouds demonstrate a pretty good match, which simply validates the correctness of the virtual visual expression.

For another, process mining relying on the inductive miner is performed to realize the automation of process discovery. As a view on reality, we map the as-happened construction work into a process model on a monthly basis using the tool of ProM (http://www.promtools.org). Figs. 3 and 4 show what the process looks like in May from views of the task and worker, separately. The process models are expressed as BPMN and Petri nets with causal relationships of sequence, concurrency, loop, choice, and others. To overcome the complexity in construction, the discovered model is abstracted from noise (i.e. infrequent/exceptional events), and thus only representative behavior covering 99% of records in event logs is taken into account. As a result of model simplicity, the task-centered model in Fig. 3 preserves 7 core tasks (out of 11 in total), which are executed by 2296 times (out of 2325 total records). Similarly, 7 productive workers remain in the worker-centered model in Fig. 4, who are responsible for 98.41% of tasks. To be more specific, Fig. 3 starts with an XOR split to create four clusters of tasks, which are "prefabricated stairs and land" (Cluster 1), "masonary work" and "external facade work" (Cluster 2), "placing window frames" (Cluster 3), and "deposit" (Cluster 4). Tasks in the four clusters can be executed parallelly. Fig. 4 provides a clear insight into collaboration among workers. In the beginning, Worker 1 involves in the process execution together with Worker 10, or Worker 3 and 4, or Worker 3 and 6. Then, either Worker 7 or 8 takes over the work and finishes it. Moreover, the virtual part in the process model format can be animated to dynamically display sequences of construction work and track the progress over time.

In terms of evaluating the discovered virtual model, the Petri nets in Figs. 3(b) and 4(b) directly integrate with the conformance checking, where the first number in the bracket is the number of records aligned correctly with event logs and the second number represents undesirable

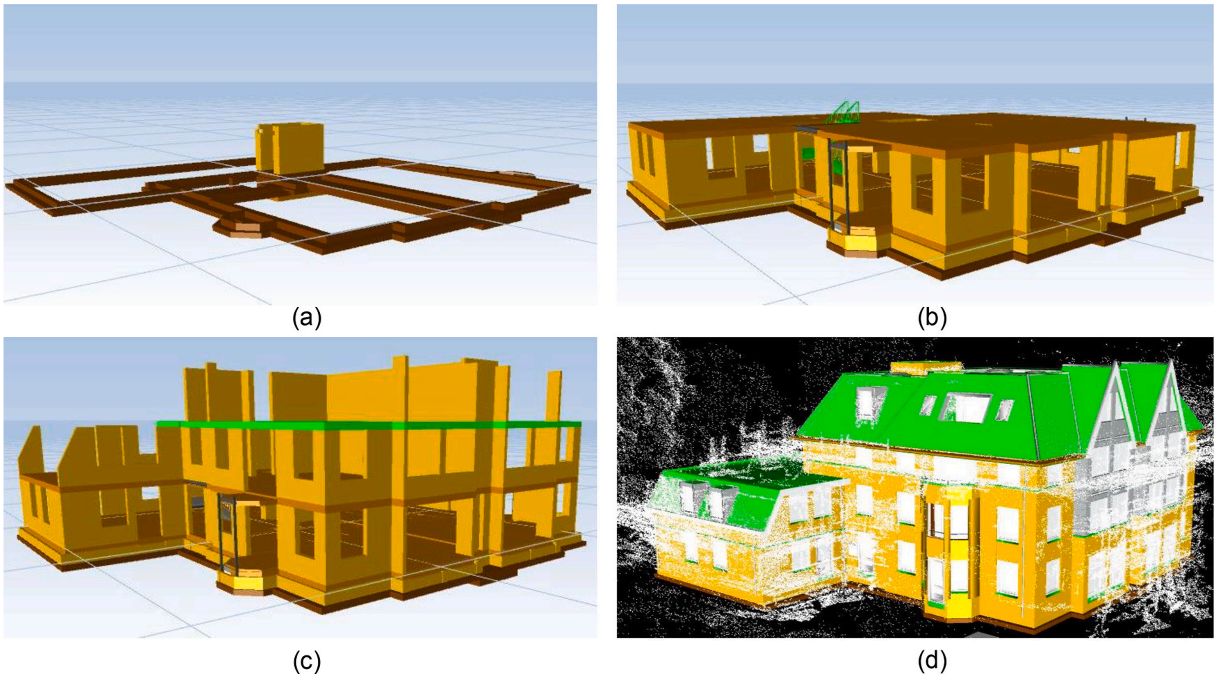

Fig. 2. 4D snapshots for the virtual model at the end of (a) Feb; (b) May; (c) Aug; and (d) Dec. (Note: Point clouds are also provided in (d).)

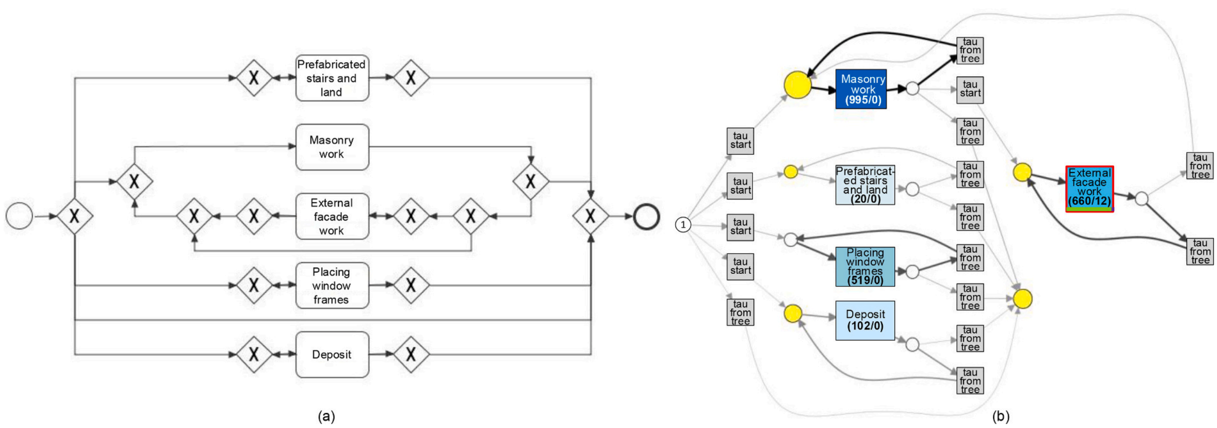

Fig. 3. Task-centered process model represented by (a) BPMN; and (b) Petri nets.

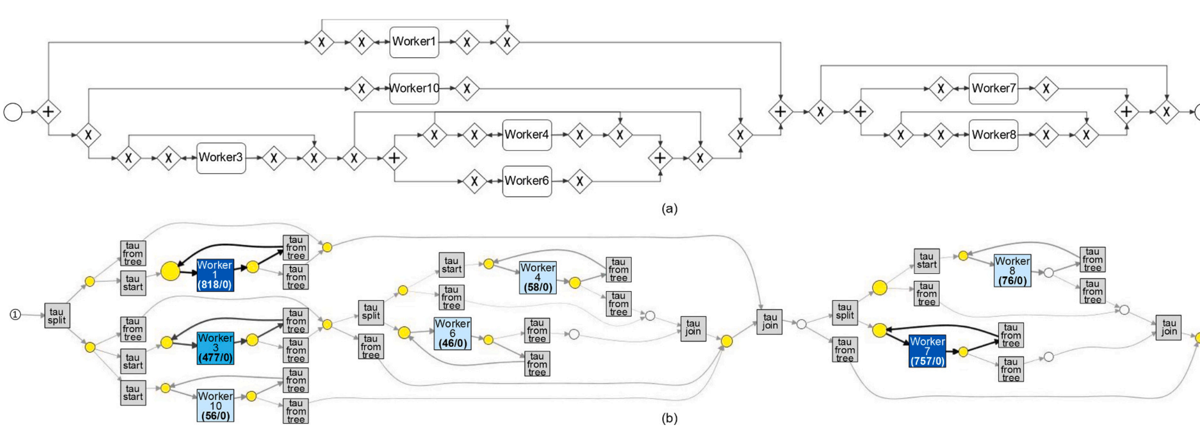

Fig. 4. Worker-centered process model represented by (a) BPMN; and (b) Petri nets.

deviations between the modeled and observed behavior. Only the part of "external facade work" has the deviation, which is highlighted by the red border frame in Fig. 3(b). More precisely, 1.81% of this certain task (12 out of 660) cannot correspond to the event log correctly. It can be seen from Figs. 3(b) and 4(b) that there is a relatively high degree of agreement to well match the discovered and actual process. To further measure the quality of discovered models in reflecting the actual behavior from log data, evaluation metrics in Eqs. (1)–(3) are calculated. As listed in Table 2, precision is approximately 0.3 lower than the reply fitness, which means that there is a trade-off between underfitting and overfitting. Fitness and generalization are guaranteed with a value closer to 1, indicating that both the task-centered and worker-centered process models are generalized enough to replay the most executed sequences of events observed in the logs. Precision larger than 0.7 is also acceptable to characterize the process credibility.

5.3. Diagnosis of construction process

With an understanding of the frequent activities and paths during construction, the discovered process model based on a fuzzy miner can diagnose and foresee the most frequently occurring bottlenecks, which are not visible via observation. Feedback from the diagnosis is expected to strengthen operations and collaboration, bringing an inherent benefit in construction efficiency enhancement. By Disco Fluxicon software (https://fluxicon.com/disco/), the fuzzy model can be generated and simplified into the desired level to be easily comprehended, as shown in Fig. 5. The average duration spent in the process is projected into the model by the coloring of boxes and the thickness of the arrows. The diagnostic results from process mining are presented below, which can then be explained reasonably to inform tactical decisions for process improvement.

(1) Regarding the task-centered model in Fig. 5 (a), the most significant bottleneck highlighted by the software was the construction path between "Deposit" and "Adhesive work sand-lime brick elements", which took up 4 days. It is worth noting that the long path "Edge processing – Reinforcement – Deposit – Adhesive work sand-lime brick element" was prone to be slower than others. It can be inferred that the delay in a certain task could pass to negatively influence another, resulting in a chain reaction. Except for the process diagnosis, another key point is to raise some conjectures for reaching a richer understanding of the identified bottleneck. For example, one possible explanation herein is that there was a lag after the task "Deposit" concerned with pouring concrete during the real construction. After workers finished the activities for curing the concrete on objects, they just waited with nothing to do, leading to a great waste of manpower and time. For this concern, managers can arrange these workers to do other tasks once they complete the "Deposit". As for a single task, the chart presents that the task named "Masonry work" spent the longest time. It is rational since the real case shows that the task number of "Masonry work" was the largest accounting for 42.80% of workloads in May. In contrast, the task named "External facade work" constituted less than 1% of the total work in the actual process, but it took the second-longest days (10 days) to complete. That is to say, this task should be underlined as a root cause of delays in May.

(2) For the worker-centered model in Fig. 5(b), there were two big red arrows in the path of "Worker 3-Worker 1" and "Worker 1-Worker 3". As an interpretation from the chart, a possible bottleneck between Worker 1 and Worker 3 was recognized by the software, which needed to be investigated at first. One of the potential reasons causing the particular bottleneck may be a lack of proper cooperation and communication between the two workers. Managers can therefore target Worker 1 and 3 to adjust their inappropriate workflows and promote greater cooperation. Then, we go back to the actual construction process to check whether the bottleneck shown in the chart has occurred. It could be found that there was actually a real observed record of conflicts between Worker 1 and 3, which was consistent with the process diagnosis from the chart to validate the practicability of the process mining results. To explain the bottleneck in the terms of factuality, that is because both of the workers were carpenters with the same duties. If the work arrangement was unreasonable or communication between them was poor, they tended to take tasks with significant overlapping to slow down the progress. This fact also suggests that it is necessary to optimize the workflow in workers with the same occupations for minimizing duplication in efforts. Besides, although Worker 8 kept active in the highest number of days (19 days), he only completed 3.26% amount of work within the month in the real case. In other words, Worker 8 was more likely to generate delays than other workers participating in May due to his poor efficiency. More instructions should be given to Worker 8, aiming to facilitate him to carry out construction more skillfully and quickly.

(3) Self-loops in "Pedestal sand-lime brick", "Prefabricated stairs and land", and "Worker1" from Fig. 5 stood for unnecessary reworks, which should not be ignored. These recognized reworks from process maps were supposedly problematic to cause additional time and costs, which also deserved careful checks and serious consideration. In comparison to the real case, the more amount of reworks actually appeared in the two tasks "Pedestal sand-lime brick" and "Prefabricated stairs and land", since these finished works were more likely to fail in meeting the acceptable quality criterion. Besides, Worker 1 was an unskilled carpenter without much work experience, who was unable to perform construction tasks in a reliable and efficient manner. The physical truth has proven that the undesired reworks could negatively impact project period and cost, and thus, managers should strive to decrease reworks in the pursuit of a more linear and branching process.

Table 2
Evaluation of the discovered process model.

Model	Reply fitness	Precision	Generalization
Task-centered	0.997	0.698	0.999
Worker-centered	1	0.727	0.981

Apart from the process model, the 4D model provides another intuitive way to visually highlight unwanted bottlenecks. When the possible delays are detected, color schemes can be given to the specific components of the 4D model causing the bottlenecks as a visual representation. For example, Fig. 6 assigns magenta to the important cause of delay named "External facade work", and thus this noteworthy part can be easily distinguished from others. It offers an opportunity in triggering warnings on the possible delays before they emerge in physical conditions. Based on the early warning, managers can provide guidance and adjustment to construction workers ahead of time. In return, workers can take more notice of the inefficient parts, who can then implement corresponding actions to effectively reduce or even eliminate negative effects from potential bottlenecks if possible.

5.4. Prediction of construction process

Since the event logs cover 11 months of the construction process, it can be organized into a new dataset with 230 lines and 3 features for time series analysis. As outlined in Table 3, each line of the dataset describes daily work using three attributes, including the date, number of finished tasks, and active workers. Remarkably, the number of finished tasks is worthy of being forecasted to describe its variation tendency in a quantitative manner. That is to say, predictions based on

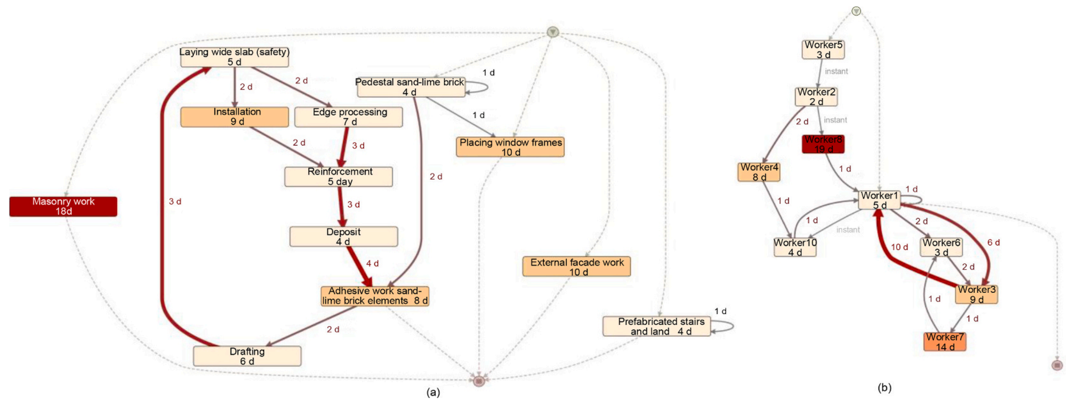

Fig. 5. Fuzzy process model about May for bottleneck detection: (a) Task-centered model; and (b) Worker-centered model.

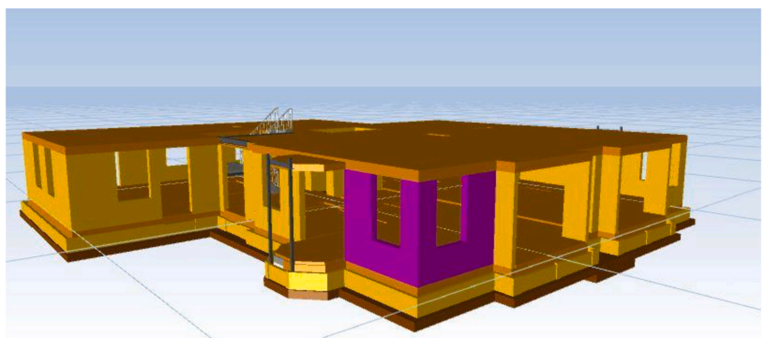

Fig. 6. 4D model visualization of the certain bottleneck in task "External facade work".

Table 3
Summary of time series data.

Characteristic	Date	Number of finished tasks	Number of workers
Range	2/2/2015–18/12/2015	[98, 131]	[8,11]
Mean (Std)	–	117.261 (9.572)	9.543 (0.631)
Median	–	120.500	10

time series data are possible to provide an overview of the construction progress in advance, which instruct real-time decision making in optimizing the work arrangement to ensure satisfactory performance. For the prediction from a future perspective, the ARIMAX model is integrated into the data-driven virtual system, in order to fit the temporal evolution of the construction phase by learning historical data of task numbers along with the outside factor termed worker number.

From the beginning, the Ljung-Box test is performed in the time series data to test its randomness on a series of lags. It returns a p-value smaller than 0.05 to reject the null hypothesis that the original data is white noise. In other words, the time series data embedding patterns deserve in-depth exploration. Then, the meaningful dataset is partitioned into a training set and a test set under an 80%–20% split, where the test set is the most recent end of data (16/10/2015–18/12/2015) accounting for typically 20% of the total sample. It can be seen in Fig. 7 (a) that the original data of task number is non-stationary in nature, which is also checked statistically by the augmented Dickey-Fuller test to accept the null hypothesis that the time series sample has a unit root (p-value >0.05). Since stationary processes with constant mean and variance over time can make reliable predictions with ease, the time series scale is necessarily transformed into the stationarity with a p-value below 0.05 using the first-order difference ($d=1$), as displayed in Fig. 7 (b). Thirdly, two important orders q and p in ARIMAX can be roughly identified from ACF and PACF plots, which visualize the correlation of present with lags and the correlation of residuals with the next lags, respectively. It is observed that the second points in Fig. 8(a) and (b) fall on the lower edge of the blue area, indicating the levels at which the autocorrelation is significant. Meanwhile, a too complex model with many lags is not required due to its risk of overfitting. Therefore, the value of p and q can be primarily set as 2. To further verify the

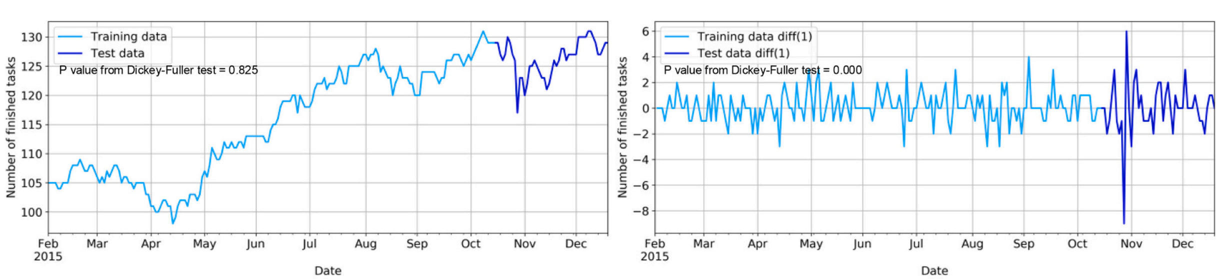

Fig. 7. Plots and the augmented Dickey-Fuller test for: (a) Original time series data; and (b) Stationary data after the first-order difference.

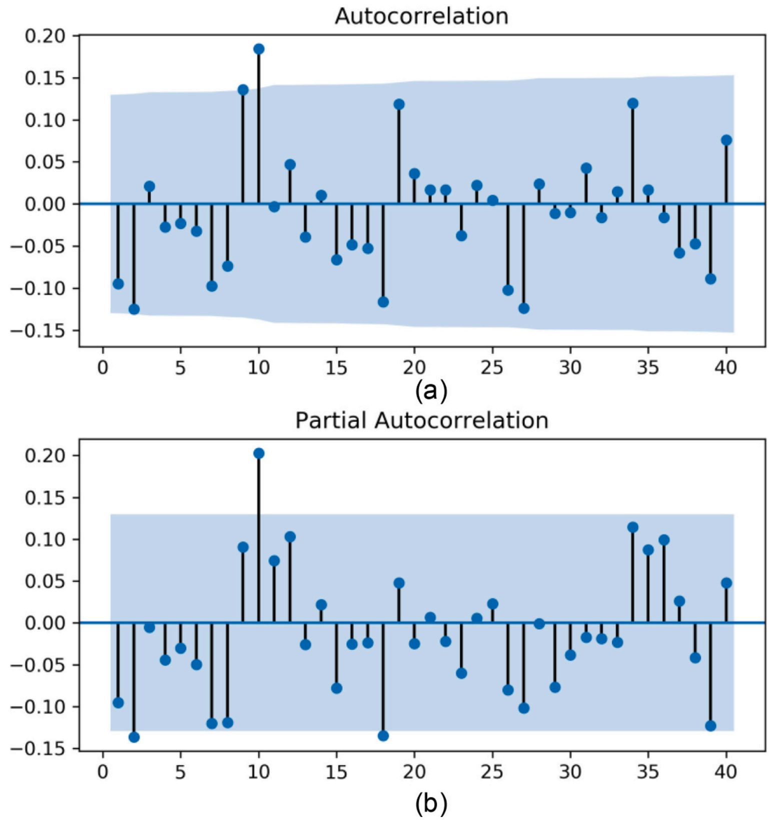

Fig. 8. ACF and PACF plots for stationary data after the first-order difference.

Table 4
Goodness of fit for six candidate ARIMAX models.

Model	Log-likelihood	AIC	BIC
ARIMAX (1,1,1)	−284.941	579.881	595.929
ARIMAX (1,1,2)	−282.400	576.799	596.056
ARIMAX (2,1,1)	−285.019	582.038	601.295
ARIMAX (3,1,3)	−282.586	579.173	601.639
ARIMAX (2,1,2)	**−273.855**	**565.711**	**594.596**
ARIMAX (4,1,4)	−281.653	585.306	620.611

determined orders, six ARIMAX models under different combinations of p and q have been built in Table 4. The examination of the goodness of fit turns out that ARIMAX (2,1,2) with the maximal log-likelihood and the minimal AIC and BIC is the best-fitted one for producing dependable forecasts of future points in the time series.

For developing a predictive model, the training set is used to estimate coefficients of the ARIMAX (2, 1, 2) model associated with the lagged worker number as the covariate. Table 5 summarizes the optimal coefficients as the weights of each term derived from the maximum likelihood estimation. Notably, a p-value less than 0.05 indicates the statistical significance of all coefficients. Based on the fitted ARIMAX (2, 1, 2) model, we can predict the number of tasks on a certain day relying on the full history up to the day. In Fig. 9, the predicted value (red line) denoting the number of tasks thought to be executed in the following days is plotted against the true value (blue line), which appears to be in the correct trend and scale. That is to say, the developed model in a satisfactory fit is able to make promising forecasts aligned with the truth well, contributing to evaluating the next construction workload numerically. Also, the red line with the mean value of 124.894 is averagely below the blue line with a mean of 126.348, implying that our predictions are relatively conservative. To better understand the accuracy of prediction, Fig. 10(a) visualizes the residual error, which oscillates near zero to demonstrate the great quality of the forecasts. Clearly, Fig. 10(b) and (c) reveal that residual errors in both the training set and the test set have approximately normal distributions, which are centered on 0.085 and − 1.454, respectively. Although there exists a bias in the prediction, the value of the residual seems acceptable. The negative sign in the average residual error of the test set also proves that the prediction of construction efficiency is slightly lower than the actual value.

Table 5
Coefficient estimation of ARIMAX (2, 1, 2) model.

Item	Coefficient	Std error	p-value	97.5% confidence interval
Constant	−5.641	0.111	0.000	[−5.858, −5.423]
Workers	0.013	0.002	0.000	[0.009, 0.017]
AR. ϕ_1. Tasks	1.802	0.000	0.000	[1.802, 1.802]
AR. ϕ_2. Tasks	−0.802	0.000	0.000	[−0.802, −0.802]
MR. θ_1. Tasks	−0.999	0.078	0.000	[−1.152, −0.845]
MR. θ_2. Tasks	0.141	0.075	0.000	[−0.007, 0.288]

In sum, the developed ARIMAX model allows for learning time series data in the virtual model, which possesses a strong predictive ability in estimating the trend of construction progress in the next few months. It can give back pieces of numerical evidence to managers for schedule design, task allocation, and workflow optimization. For one thing, it seems that the number of finished tasks is on the rising trend as the construction process runs. Hence, managers can reasonably arrange more workers and tasks after Jun. For another, if the manager hopes to fulfill the project ahead of schedule, he'd better focus on the work during Feb – Jun at slow construction speed through optimization of the relevant construction process and worker arrangement. Moreover, since the number of finished tasks estimated by the developed ARIMAX model tends to be slightly smaller than observations, the project duration in the proposed scheduling could be a little longer than the reality. When workers proceed to work as planned, the rate of progress in the physical part is likely to exceed managers' expectations through speedy actions.

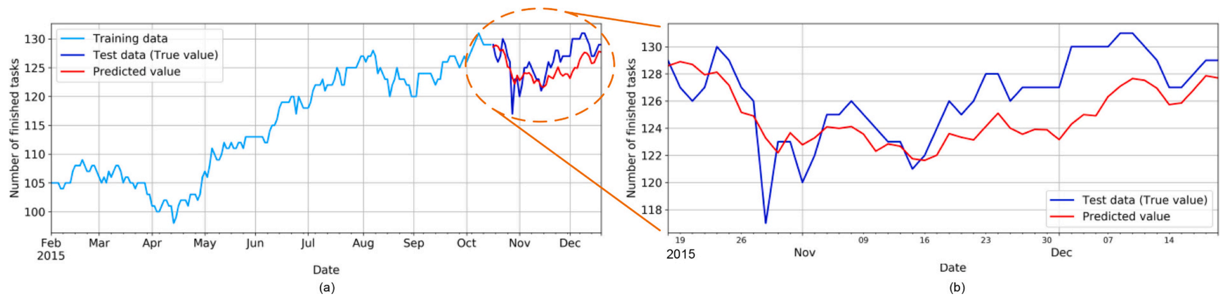

Fig. 9. Plots of the forecast line and corresponding true value in: (a) Whole dataset; and (b) Test set.

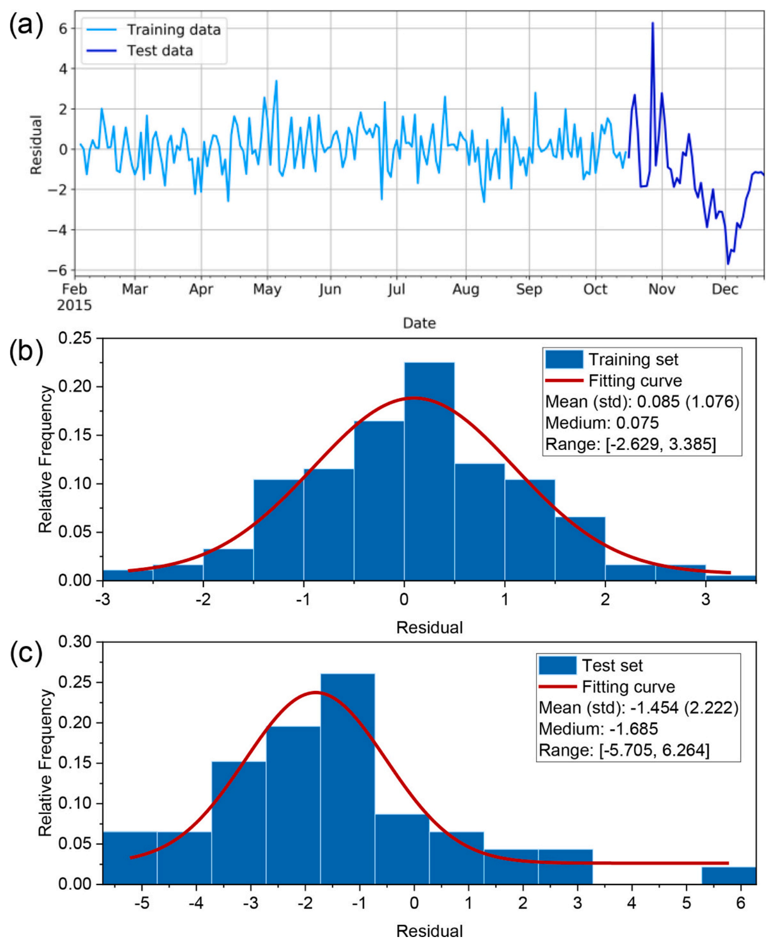

Fig. 10. Residual errors in: (a) Whole dataset; (b) Training set; and (c) Test set.

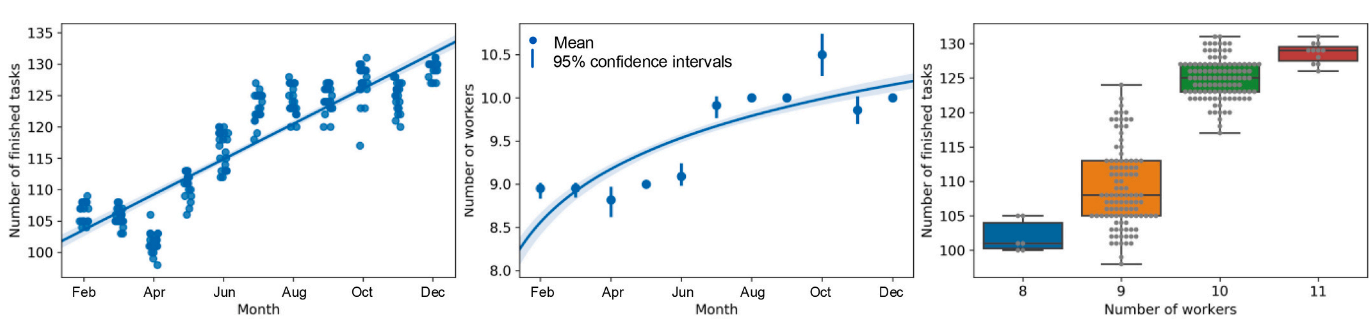

Fig. 11. (a) and (b) Variation of task number and worker month by month; and (c) Relationship between the number of tasks and workers.

6. Discussions

Remarkably, the time series data contains lots of hidden knowledge about tasks and workers, which can shed light on the nature of project evolution. Besides, the superiority of ARIMAX in forecasting the construction progress can be further validated based on the comparison against four common time series algorithms. The discussions are summarized as follows.

(1) Characteristics of finished tasks and involved workers can be observed directly from time series data, which can serve as direct evidence for managers in project management. Linear regression and a variation of linear regression in the form of $y \sim \log(x)$ are fitted well along with a 95% confidence interval in Fig. 11(a) and (b), respectively, which manifest a growing tendency in the number of both tasks and workers over the month. That is to say, as a building rises through its floors, more trades can perform work. More workers involved especially after Jun is entirely expected to increase the task number. It has been proved in Fig. 11 (c) that there is a positive correlation between the number of finished tasks and workers. In particular, 10 or more workers can averagely execute more than 9 tasks each day than workers fewer than 9. Apart from more workers, it can be assumed that the more skilled techniques and closer collaboration can be another method to accelerate the project process. As the construction proceeds, workers will gradually be more and more familiar with the tasks and their co-workers. Accordingly, managers can consider assigning more than 10 skilled workers every weekday in the intermedia-late course of the project.

(2) The developed ARIMAX model is compared with other popular time series algorithms to exhibit its outstanding predictive ability. Specifically, SARIMA and SARIMAX stand for the seasonal ARIMA and ARIMAX model incorporating the seasonal order argument. It is found in Fig. 12 that predictions from the ARIMAX model (green line) and SARIMAX model (red dash line) show the consistent trend as the true value (blue line), verifying the necessity of exogenous variables in achieving precise forecasting of the task number. Meanwhile, the green line gets much closer to the blue line, indicating that the ARIMAX model is prone to ensure the prediction quality. Although the two lines from AR and ARIMA model taking no account of outside factors can also be near the blue line, both of them have an obvious downward trend, which is just the opposite of the reality. According to evaluation metrics RMSE and MAE in Eqs. (12), (13), the performance of five candidate models are measured quantifiably in Table 6, resulting in the rank as ARIMAX > ARIMA > AR > SARIMA > SARIMAX. It suggests that our model choice in ARIMAX (2, 1, 2) associated with the number of workers turns out to be the best one under the smallest RMSE (2.635) and MAE (2.204). Noteworthily, SARIMA and SARIMAX considering seasonality are the two most inaccurate models, whose RMSE and MAE are at least 62.24% and 33.89% lower than the most appropriate ARIMAX. That is to say, construction performance does not experience obvious seasonal variation. Besides, a more complex time series model does not always mean better.

Table 6
Evaluation of predictions from different time series algorithms.

Model	RMSE	MAE
AR (1,0)	3.681	2.930
ARIMA (1,1,1)	3.678	2.901
ARIMAX (2, 1, 2) with number of workers	**2.635**	**2.204**
SARIMA (1,1,0) (2, 0, 1, 5)	4.275	3.334
SARIMAX (2,1,0) (2, 0, 1, 5) with number of workers	5.545	6.076

7. Conclusions and future works

Towards a smart construction service, a detailed framework of the digital twin containing a physical model, a virtual model, and connection data is developed under the integration of BIM, IoT, and DM. The significance of the established loop lies in facilitating the automation, digitalization, and intelligence of the construction project management, which could be less susceptible to human cognitive errors. More specifically, IoT devices are deployed to collect real-time data about the actual status of the construction operation with little manual interaction. The rich data source from IoT serves as the foundation of the cyber-physical synchronicity, which needs to be mapped into the IFC scheme for model interoperability and then saved as event logs for data analysis and smart reasoning. The contribution of this research can be outlined as (a) the state of knowledge by integrating advanced DM techniques into the virtual part, including process mining and time series analysis, aiming to gain deep insights into massive IoT data gathered from the physical side and stored in the cloud BIM; and (b) the state of practice by realizing process simulation, bottleneck diagnose, and progress prediction objectively in the virtual space, which can ensure a comprehensive view of the entire process, support fast and cost-effective troubleshooting, and inform data-driven decisions to improve the workflows and staffing in the physical world at an early stage.

A case study in an actual BIM-based construction project is implemented to make continuous process analysis, prediction, and optimization. Specifically, a data-driven digital twin creates a constant loop between the physical and virtual parts, which largely relies on the point clouds taken by IoT devices during the real-time operational monitoring. The BIM cloud storage system acts as the data repository to continuously synchronize with IoT data and the BIM server interprets IoT data into proper formats. Noticeably, this updated data can be passed to the cyber world, which is helpful in automatically building the virtual model paired with physical features and conducting knowledge discovery for tactical decision making. That is to say, both the IoT and DM algorithms contribute to making these digital replicas far more useful. Herein, the virtual model is built in two formats with identical fidelity, namely 4D visualization and process models, both of which emphasize the nature of task execution and worker collaboration through process simulation. Especially for the process model, it is established in the view of tasks and

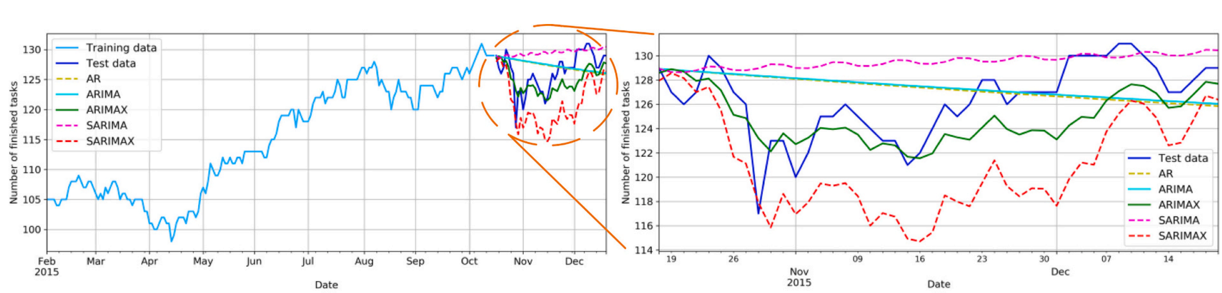

Fig. 12. Comparisons of predictions from different time series algorithms visualized in: (a) Whole dataset; and (b) Test set.

workers, which can well reply to the event log with the value of fitness and generalization around 1 and precision larger than 0.7. Moreover, the process mining algorithms (i.e., inductive and fuzzy miner) and ARIMAX (2, 1, 2) model associated with the lagged worker number in the minimal RMSE (2.635) and MAE (2.204) are two DM techniques encompassed in the virtual model. They deeply explore large log data from current and future perspectives to facilitate effective interaction between the real and virtual worlds. On the one hand, bottlenecks causing delays can be easily detected in the current process. On the other hand, the number of future construction tasks to be accomplished can be predicted over time. These predictions about possible problems and future workload can generate early risk warnings and realize performance assessment for the optimization purpose. Accordingly, suggestions can be output in a dynamic manner to guide the physical process, which can even respond to changes in the real construction site. Managers can, therefore, formulate more rational construction scheduling with well-arranged workflows, workloads, and workers, aiming to promptly improve operational efficiency and strengthen cooperation in the physical construction process.

There are some lilitations open for future studies. We mainly rely on a single monitoring source that is the point clouds from a UAV in this case for simplicity, but it could be hard to reveal the complex nature of a large-scale project in reality. Besides, by interpreting the collected single source of information, it is inadequate to gain an overall sense of the construction project and put forward improved schemes. Since possible explanations are subjected to limited data and a low degree of confidence, there should be room for errors in the current research. Therefore, it is desirable to collect and merge multiple sources of monitoring data using information fusion [59] in future studies, which can be explored and interpreted from different perspectives to support more convincing conclusions. For example, detailed information about the occupations of workers can be provided to better explain the construction logic from the stem of data. In the end, more meaningful interpretations of what people were actually doing and what the meaning of the patterns is can be generated.

Declaration of Competing Interest

None.

Acknowledgment

The Ministry of Education Tier 1 Grant, Singapore (No. 04MNP002126C120, No. 04MNP000279C120) and the Start-Up Grant at Nanyang Technological University, Singapore (No. 04INS000423C120) are acknowledged for their financial support of this research.

References

[1] C. Boje, A. Guerriero, S. Kubicki, Y. Rezgui, Towards a semantic construction digital twin: directions for future research, Autom. Constr. 114 (2020) 103179, https://doi.org/10.1016/j.autcon.2020.103179.
[2] R.J.C. Bose, E.H. Verbeek, W.M. van der Aalst, Discovering Hierarchical Process Models Using ProM, International Conference on Advanced Information Systems Engineering, Springer, 2011, pp. 33–48, https://doi.org/10.1007/978-3-642-29749-6_3.
[3] G.E. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015, 978–1–118-67502-1.
[4] D.A. Broniatowski, M. Dredze, M.J. Paul, A. Dugas, Using social media to perform local influenza surveillance in an inner-city hospital: a retrospective observational study, JMIR Public Health Surveill. 1 (1) (2015), https://doi.org/10.2196/publichealth.4472.
[5] J.C. Buijs, B.F. Van Dongen, W.M. van Der Aalst, On the role of fitness, precision, generalization and simplicity in process discovery, in: OTM Confederated International Conferences "On the Move to Meaningful Internet Systems", Springer, 2012, pp. 305–322, https://doi.org/10.1007/978-3-642-33606-5_19.
[6] J.C. Buijs, B.F. van Dongen, W.M. van der Aalst, Quality dimensions in process discovery: the importance of fitness, precision, generalization and simplicity, Int. J. Coop. Inform. Syst. 23 (01) (2014) 1440001, https://doi.org/10.1142/S0218843014400012.
[7] J.C. Cheng, W. Chen, K. Chen, Q. Wang, Data-driven predictive maintenance planning framework for MEP components based on BIM and IoT using machine learning algorithms, Autom. Constr. 112 (2020) 103087, https://doi.org/10.1016/j.autcon.2020.103087.
[8] B. Dave, A. Buda, A. Nurminen, K. Främling, A framework for integrating BIM and IoT through open standards, Autom. Constr. 95 (2018) 35–45, https://doi.org/10.1016/j.autcon.2018.07.022.
[9] B. Dave, S. Kubler, K. Främling, L. Koskela, Opportunities for enhanced lean construction management using internet of things standards, Autom. Constr. 61 (2016) 86–97, https://doi.org/10.1016/j.autcon.2015.10.009.
[10] C. Deb, F. Zhang, J. Yang, S.E. Lee, K.W. Shah, A review on time series forecasting techniques for building energy consumption, Renew. Sust. Energ. Rev. 74 (2017) 902–924, https://doi.org/10.1016/j.rser.2017.02.085.
[11] A. Dimitrov, M. Golparvar-Fard, Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections, Adv. Eng. Inform. 28 (1) (2014) 37–49, https://doi.org/10.1016/j.rser.2017.02.085.
[12] L. Ding, X. Xu, Application of cloud storage on BIM life-cycle management, Int. J. Adv. Robot. Syst. 11 (8) (2014) 129, https://doi.org/10.5772/58443.
[13] L. Ding, Y. Zhou, B. Akinci, Building information modeling (BIM) application framework: the process of expanding from 3D to computable nD, Autom. Constr. 46 (2014) 82–93, https://doi.org/10.1016/j.autcon.2014.04.009.
[14] C.M. Flath, N. Stein, Towards a data science toolbox for industrial analytics applications, Comput. Ind. 94 (2018) 16–25, https://doi.org/10.1016/j.compind.2017.09.003.
[15] E. Glaessgen, D. Stargel, The digital twin paradigm for future NASA and US Air Force vehicles, in: 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, 2012, p. 1818, https://doi.org/10.2514/6.2012-1818.
[16] M. Golparvar-Fard, F. Peña-Mora, S. Savarese, D4AR–a 4-dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication, J. Inform. Technol. Construct. 14 (13) (2009) 129–153. https://www.itcon.org/2009/13.
[17] C.W. Günther, W.M. Van Der Aalst, Fuzzy Mining–Adaptive Process Simplification Based on Multi-Perspective Metrics, International Conference on Business Process Management, Springer, 2007, pp. 328–343, https://doi.org/10.1007/978-3-540-75183-0_24.
[18] T. Gurgen Erdogan, A. Tarhan, A goal-driven evaluation method based on process mining for healthcare processes, Appl. Sci. 8 (6) (2018) 894, https://doi.org/10.3390/app8060894.
[19] A.R. Hevner, S.T. March, J. Park, S. Ram, Design science in information systems research, MIS Q. (2004) 75–105, https://doi.org/10.2307/25148625.
[20] I. Hwang, Y.J. Jang, Process mining to discover shoppers' pathways at a fashion retail store using a WiFi-base indoor positioning system, IEEE Trans. Autom. Sci. Eng. 14 (4) (2017) 1786–1792, https://doi.org/10.1109/TASE.2017.2692961.
[21] M. Jans, J.M. Van Der Werf, N. Lybaert, K. Vanhoof, A business process mining application for internal transaction fraud mitigation, Expert Syst. Appl. 38 (10, 2011) 13351–13359, https://doi.org/10.1016/j.eswa.2011.04.159.
[22] S. Kouhestani, M. Nik-Bakht, IFC-based process mining for design authoring, Autom. Constr. 112 (2020) 103069, https://doi.org/10.1016/j.autcon.2019.103069.
[23] T. Lagkas, V. Argyriou, S. Bibi, P. Sarigiannidis, UAV IoT framework views and challenges: towards protecting drones as "things", Sensors 18 (11, 2018) 4015, https://doi.org/10.3390/s18114015.
[24] S.J. Leemans, D. Fahland, W.M. van der Aalst, Discovering Block-Structured Process Models from Event Logs-a Constructive Approach, International Conference on Applications and Theory of Petri Nets and Concurrency, Springer, 2013, pp. 311–329, https://doi.org/10.1007/978-3-642-38697-8_17.
[25] J.R. Lin, Z.Z. Hu, J.P. Zhang, F.Q. Yu, A natural-language-based approach to intelligent data retrieval and representation for cloud BIM, Comp. Aided Civil Infrastruct. Eng. 31 (1) (2016) 18–33, https://doi.org/10.1111/mice.12151.
[26] D.A. Linares, C. Anumba, N. Roofigari-Esfahan, Overview of supporting technologies for cyber-physical systems implementation in the AEC industry, Comput. Civil Eng. (2019), https://doi.org/10.1061/9780784482438.063.
[27] Q. Lu, A.K. Parlikad, P. Woodall, G. Don Ranasinghe, X. Xie, Z. Liang, E. Konstantinou, J. Heaton, J. Schooling, Developing a digital twin at building and city levels: case study of West Cambridge campus, J. Manag. Eng. 36 (3) (2020), 05020004, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000763.
[28] Q. Lu, X. Xie, A.K. Parlikad, J.M. Schooling, Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance, Autom. Constr. 118 (2020) 103277, https://doi.org/10.1016/j.autcon.2020.103277.
[29] R. Lu, I. Brilakis, Digital twinning of existing reinforced concrete bridges from labelled point clusters, Autom. Constr. 105 (2019) 102837, https://doi.org/10.1016/j.autcon.2019.102837.
[30] Z. Ma, Y. Ren, X. Xiang, Z. Turk, Data-driven decision-making for equipment maintenance, Autom. Constr. 112 (2020) 103103, https://doi.org/10.1016/j.autcon.2020.103103.
[31] S.T. March, V.C. Storey, Design science in the information systems discipline: an introduction to the special issue on design science research, MIS Q. (2008) 725–730, https://doi.org/10.2307/25148869.
[32] Q. Min, Y. Lu, Z. Liu, C. Su, B. Wang, Machine learning based digital twin framework for production optimization in petrochemical industry, Int. J. Inf. Manag. 49 (2019) 502–519, https://doi.org/10.1016/j.ijinfomgt.2019.05.020.
[33] D. Myers, S. Suriadi, K. Radke, E. Foo, Anomaly detection for industrial control systems using process mining, Comp. Secur. 78 (2018) 103–125, https://doi.org/10.1016/j.cose.2018.06.002.

[34] Y. Pan, L. Zhang, BIM log mining: exploring design productivity characteristics, Autom. Constr. 109 (2020) 102997, https://doi.org/10.1016/j.autcon.2019.102997.
[35] Y. Pan, L. Zhang, BIM log mining: learning and predicting design commands, Autom. Constr. 112 (2020) 103107, https://doi.org/10.1016/j.autcon.2020.103107.
[36] Y. Pan, L. Zhang, Z. Li, Mining event logs for knowledge discovery based on adaptive efficient fuzzy Kohonen clustering network, Knowl.-Based Syst. 209 (2020) 106482, https://doi.org/10.1016/j.knosys.2020.106482.
[37] Y. Pan, L. Zhang, M.J. Skibniewski, Clustering of designers based on building information modeling event logs, Comp. Aided Civil Infrastruct. Eng. 35 (7) (2020) 701–718, https://doi.org/10.1111/mice.12551.
[38] J. Park, D. Lee, J. Zhu, An integrated approach for ship block manufacturing process performance evaluation: case from a Korean shipbuilding company, Int. J. Prod. Econ. 156 (2014) 214–222, https://doi.org/10.1016/j.ijpe.2014.06.012.
[39] Y. Peng, J.-R. Lin, J.-P. Zhang, Z.-Z. Hu, A hybrid data mining approach on BIM-based building operation and maintenance, Build. Environ. 126 (2017) 483–495.
[40] W. Premchaiswadi, P. Porouhan, Process modeling and bottleneck mining in online peer-review systems, SpringerPlus 4 (1) (2015) 1–18, https://doi.org/10.1186/s40064-015-1183-4.
[41] Q. Qi, F. Tao, Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison, IEEE Access 6 (2018) 3585–3593, https://doi.org/10.1109/ACCESS.2018.2793265.
[42] I. Raedts, M. Petkovic, Y.S. Usenko, J.M.E. van der Werf, J.F. Groote, L.J. Somers, Transformation of BPMN models for behaviour analysis, MSVVEIS 2007 (2007) 126–137. https://www.scitepress.org/papers/2007/24288/24288.pdf.
[43] R. Sacks, C. Eastman, G. Lee, P. Teicholz, BIM Handbook: A Guide to Building Information Modeling for Owners, Designers, Engineers, Contractors, and Facility Managers, John Wiley & Sons, 2018 (ISBN: 1119287537).
[44] B. Schleich, N. Anwer, L. Mathieu, S. Wartzack, Shaping the digital twin for design and production engineering, CIRP Ann. 66 (1) (2017) 141–144, https://doi.org/10.1016/j.cirp.2017.04.040.
[45] C.-S. Shim, N.-S. Dang, S. Lon, C.-H. Jeon, Development of a bridge maintenance system for prestressed concrete bridges using 3D digital twin model, Struct. Infrastruct. Eng. 15 (10, 2019) 1319–1332, https://doi.org/10.1080/15732479.2019.1620789.
[46] Y. Srewil, R.J. Scherer, Effective Construction Process Monitoring and Control through a Collaborative Cyber-Physical Approach, Working Conference on Virtual Enterprises, Springer, 2013, pp. 172–179, https://doi.org/10.1007/978-3-642-40543-3_19.
[47] V. Stojanovic, M. Trapp, R. Richter, B. Hagedorn, J. Döllner, Towards the generation of digital twins for facility management based on 3D point clouds, in: Proceeding of the 34th Annual ARCOM Conference, 2018, pp. 270–279.
[48] S. Tang, D.R. Shelden, C.M. Eastman, P. Pishdad-Bozorgi, X. Gao, A review of building information modeling (BIM) and the internet of things (IoT) devices integration: present status and future trends, Autom. Constr. 101 (2019) 127–139, https://doi.org/10.1016/j.autcon.2019.01.020.
[49] F. Tao, F. Sui, A. Liu, Q. Qi, M. Zhang, B. Song, Z. Guo, S.C.-Y. Lu, A. Nee, Digital twin-driven product design framework, Int. J. Prod. Res. 57 (12, 2019) 3935–3953, https://doi.org/10.1080/00207543.2018.1443229.
[50] F. Tao, H. Zhang, A. Liu, A.Y. Nee, Digital twin in industry: state-of-the-art, IEEE Trans. Ind. Inform. 15 (4) (2018) 2405–2415, https://doi.org/10.1109/TII.2018.2873186.
[51] F. Tao, M.J.I.A. Zhang, Digital twin shop-floor: a new shop-floor paradigm towards smart manufacturing, IEEE Access 5 (2017) 20418–20427, https://doi.org/10.1109/ACCESS.2017.2756069.
[52] J. Teizer, M. Wolf, O. Golovina, M. Perschewski, M. Propach, M. Neges, M. König, Internet of Things (IoT) for integrating environmental and localization data in Building Information Modeling (BIM), ISARC, in: Proceedings of the International Symposium on Automation and Robotics in Construction Vol. 34, IAARC Publications, 2017, https://doi.org/10.22260/ISARC2017/0084.
[53] J. Vachálek, L. Bartalský, O. Rovný, D. Šišmišová, M. Morháč, M. Lokšík, The digital twin of an industrial production line within the industry 4.0 concept, in: 2017 21st International Conference on Process Control (PC), IEEE, 2017, pp. 258–262, https://doi.org/10.1109/PC.2017.7976223.
[54] W. Van der Aalst, Data Science in Action. Process Mining, Springer, Heidelberg, 2016, https://doi.org/10.1007/978-3-662-49851-4_1.
[55] S. van Schaijk, Building Information Model (BIM) Based Process Mining Enabling Knowledge Reassurance and Fact-Based Problem Discovery within the Architecture, Engineering, Construction and Facility Management Industry, Master Construction Management and Engineering, Eindhoven University of Technology, 2016. https://research.tue.nl/en/studentTheses/.
[56] B.N. Yahya, M. Song, H. Bae, S.-o. Sul, J.-Z. Wu, Domain-driven actionable process model discovery, Comput. Ind. Eng. 99 (2016) 382–400, https://doi.org/10.1016/j.cie.2016.05.010.
[57] X. Yuan, C.J. Anumba, M.K. Parfitt, Cyber-physical systems for temporary structure monitoring, Autom. Constr. 66 (2016) 1–14, https://doi.org/10.1016/j.autcon.2016.02.005.
[58] Yue Pan, Limao Zhang, Roles of artificial intelligence in construction engineering and management: A critical review and future trends, Autom. Constr. 122 (2021) 103517, https://doi.org/10.1016/j.autcon.2020.103517.
[59] Yue Pan, Limao Zhang, Xianguo Wu, Miroslaw J. Skibniewski, Multi-classifier information fusion in risk analysis, Inf. Fusion 60 (2020) 121–136, https://doi.org/10.1016/j.inffus.2020.02.003.