www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

CONSTRUCTION SCHEDULE RISK ANALYSIS – A HYBRID MACHINE LEARNING APPROACH

SUBMITTED: February 2021
REVISED: November 2021
PUBLISHED: January 2022
EDITOR: Robert Amor
DOI: 10.36680/j.itcon.2022.004

John Patrick Fitzsimmons, Principal Planner
Planning and Project Controls, Laing O'Rourke
jpf36@cam.ac.uk

Ruodan Lu, Senior Member (corresponding author)
Darwin College, University of Cambridge
rl508@cam.ac.uk

Ying Hong, Research Associate
Department of Engineering, University of Cambridge
yh448@cam.ac.uk

Ioannis Brilakis, Laing O'Rourke Reader
Department of Engineering, University of Cambridge
ib340@cam.ac.uk

SUMMARY: *The UK commissions about £100 billion in infrastructure construction works every year. More than 50% of them finish later than planned, causing damage to the interests of stakeholders. The estimation of time-risk on construction projects is currently done subjectively, largely by experience despite there are many existing techniques available to analyse risk on the construction schedules. Unlike conventional methods that tend to depend on the accurate estimation of risk boundaries for each task, this research aims to proposes a hybrid method to assist planners in undertaking risk analysis using baseline schedules with improved accuracy. The proposed method is endowed with machine intelligence and is trained using a database of 293,263 tasks from a diverse sample of 302 completed infrastructure construction projects in the UK. It combines a Gaussian Mixture Modelling-based Empirical Bayesian Network and a Support Vector Machine followed by performing a Monte Carlo risk simulation. The former is used to investigate the uncertainty, correlated risk factors, and predict task duration deviations while the latter is used to return a time-risk simulated prediction. This study randomly selected 10 projects as case studies followed by comparing their results of the proposed hybrid method with Monte Carlo Simulation. Results indicated 54.4% more accurate prediction on project delays.*

KEYWORDS: *Construction Scheduling, Machine Learning, Risk Analysis*

REFERENCE: *John Patrick Fitzsimmons, Ruodan Lu, Ying Hong, Ioannis Brilakis (2022). Construction schedule risk analysis – a hybrid machine learning approach. Journal of Information Technology in Construction (ITcon), Vol. 27, pg. 70-93, DOI: 10.36680/j.itcon.2022.004*

1. INTRODUCTION

Around £100bn is spent in the UK each year on infrastructure investments (Infrastructure and Projects Authority 2016), making the delivery of infrastructure 70% of the total spending on the National Health Service. From the construction of schools and hospitals to the delivery of road and rail infrastructure, research has shown that time and again projects tend to exceed their initial time estimates (Drury et al. 2018). In the last 10 years more than 50% of construction projects undertaken in the UK have finished late (Glenigan 2015). Worldwide, projects are being delivered later than intended, with a regularity which suggests the problem is not being improved by the combined efforts of research to date (Salling and Leleur 2015; Vick and Brilakis 2018). The money wasted on poorly delivered infrastructure would equate to as much as £35bn per year in unrealised economic benefit (HM Treasury 2014). If this trend were extrapolated worldwide, the cost to the global economy would be $620bn each year, with around $1.1tn of potential lost (Oxford Economics 2017).

There are many factors leading to the late delivery of projects, from tendency towards over optimism (Flyvbjerg, 2008), through to inexperience/incompetence in client/contractor teams (Baker et al., 2008), the inability of individuals to manage the complexity associated with large and multi-facetted desired outcomes (Koppenjan et al., 2011) and the practice of strategic misrepresentation (Flyvbjerg, 2008) and many others. To balance these competing risks, a project must be planned robustly. One of the key factors determining the outcome of a project, in relation to its original schedule and budget, is the quality and integrity of the up-front planning (Elzomor et al., 2018). Developing the project scope definition is one of the major tasks in the up-front planning process. Poor scope definition adversely affects projects in cost, schedule, and operational characteristics (Gibson & Gebken, 2003). The Project Definition Rating Index (PDRI) developed by the Construction Industry Institute (CII) is a weighted score sheet used for determining the level of scope definition in the building sector during up-front planning (Cho & Gibson, 2001). Too often, the delivery of construction projects gets caught up resolving highly complex technical problems and focuses very little on the decisions made during the estimating phase of the project. The estimation of time-risk on construction projects is generally done subjectively, largely by experience, and the penetration of academic concepts into the realm of common practice is negligible (Ortiz-González et al., 2014). Additionally, the planning of a project must include for some level of time contingency, which is clearly defined and managed to minimise the risk of delay (Mubarak, 2015). Although some software packages such as Monte Carlo™ or Deltek Acumen Risk™ (Deltek, 2019), have begun to allow for limited predefined delay correlations in modelling risk; the analyses are still inarticulate as they deal with a level of uncertainty which is difficult to calculate and understand without detailed knowledge of historical records or a strong evidentiary basis. To tackle the above-mentioned challenge, this research aims to perform risk analysis for construction scheduling more objectively and more accurately using a proposed machine learning method. The hypothesis of this work is that the proposed method is more accurate and outperforms the traditional Monte Carlo Simulation (MCS) for the time-risk prediction. This hypothesis is elaborated in the following sections. The novelty of this research lies in the fact that it combines the strengths of machine learning and Monte Carlo to generate a robust predictive model by training a large number of real-life completed infrastructure construction projects with thousands of tasks, taking into consideration of multiple risk factors. We present a literature review in Section 2 followed by the elaboration on the proposed method in Section 3. Data and experiments are presented in Section 4. Finally, we conclude this research in Section 5.

2. RESEARCH BACKGROUND

This section discusses research efforts that have been devoted to improving the risk analysis for construction scheduling. We group existing research works into traditional methods and data-driven methods, which are discussed in the following texts.

2.1 Traditional Methods

Many established risk analysis techniques exist to assist planners in modelling uncertainty on construction projects and give insight into the likelihood of meeting a particular completion date. The most common way of defining the completion date and critical path is the Critical Path Method (CPM) (Kelley, 1961), with estimates ranging from 22% to 70% industry usage (CIOB, 2009). CPM uses a series of tasks with defined dependency links to create a directed acyclic graph (DAG) network, which determines the earliest start and latest finish for each of the tasks in the network. It is a deterministic method and does not account for any uncertainty in task or project duration

(Kelley, 1961). To add uncertainty to the CPM technique, several established risk analysis practices have emerged. These techniques include MCS (Zio, 2013), Programme Evaluation and Review Technique (PERT) (Gładysz et al., 2015) and the Critical Chain Method (CCPM) (Su et al, 2016). MCS is a technique used to understand the impact of risk and uncertainty in financial, project management, cost, and other forecasting models. A MC simulator helps one visualize most or all of the potential outcomes to have a better idea regarding the risk of a decision (Goodfellow et al., 2016). In the context of computational schedule risk analysis, MCS uses best, worst and most likely duration estimates for each activity to run thousands of 'dummy runs' through the project, assigning a random but bounded value from a custom distribution for the duration of each task (Zio, 2013). The results of these 'dummy runs' are averaged to give a completion date estimate and percentage-based likelihood of the project succeeding in meeting its completion date (Steyn, 2018). Risk estimates are, however, calculated either subjectively based on the experience, arbitrarily or using work studies which can be based on a small sample, thereby relying on outputs recorded when working practices in respect of health, safety, environment and quality differed. These traditional risk analysis methods suffer from subjectivity due to the insufficient evidence base and inconsistent personal professional experience. It has been contended for a long time that these models do not sufficiently account for correlation or covariance among different risk factors and groups of task (Jun-yan, 2012), which can be a significant compounding factor over many tasks (Wang & Demsetz, 2000). To add covariance and correlation to PERT and MCS, Ökmen & Öztaş (2008) constructed a system using Correlated Schedule Risk Analysis (CSRAM). It uses correlation coefficients of a range of project risks including weather, soil conditions, labour productivity, and material/resource availability factors to calculate the possible distribution of estimated duration deviation based on subjective human inputs. This is an effective approach for modelling risk on various activities. Its subjectivity, however, leads to potentially flawed and biased outcomes. Probability-Impact (P-I) and its variants are predominant models used in risk assessment, which is a matrix to represent the likelihood of an event occurring and consequences (Bartlett, 2004). Previous studies (i.e. Dey et al. (1994) and Zhi (1995)) combined Analytic Hierarchy Process with P-I model to assess risk objectively. Some researchers (i.e. Ward & Chapman (2003)) argued that P-I matrix oversimplifies the impact and probability. Therefore, more recent research (including Han et al. (2008) and Dikmen & Birgonul (2006)) further developed decision making support systems with multiple criteria to assess project risks quantitatively. In general, the robustness of traditional risk analysis methods (including PERT and P-I models) depends on the experience of managers and planners when approximating the impact and likelihood of an uncertainty.

2.2 Data-driven Methods

A promising solution to the problem of subjectivity is learning from historic data. Bayesian Network (BN), a probabilistic DAG, is frequently used to enhance schedule risk analysis (Jun-yan, 2012; Khodakarami & Abdi, 2014) BN analyses rely on a series of "conditional independence" tests to determine that each node satisfies Markov conditions such that it is conditionally independent of all its non-descendants given its parents (Neapolitan, 2004). With large networks, this can be computationally prohibitive.

Where large datasets are available, there is a possibility of making subjective BN more objective in nature. BN can be learned from data, although some argue that this evidence-based approach to network analysis makes the science more frequentist than Bayesian (Bayarri & Berger, 2004). The combination of Bayesian and frequentist should yield improvement in predictive accuracy for modelling risk. This approach is known as Empirical Bayes.

Factors that influences project time can vary from engineering design to project management (Mulholland & Christian, 1999). A frequently reported factor to project time is "uncertain weather conditions" (Moselhi et al., 2011). Other factors include macroeconomic condition (Azar & Menassa, 2012), task-specific criticality (Fortin et al., 2010), schedule quality (Kähkönen, 2015), inaccurate estimation of material and labour productivity (Kaming et al., 1997), and the shortage of resources (e.g. material and labour) (Shen, 1997). For example, Luu et al. (2009) identified a total of 16 project-related and owner-related factors that affect project delay. These factors can be included in an Empirical Bayesian Network, leading themselves to a data-driven approach. The use of reliable evidence is noted as being key to improving accuracy of prediction models. Kirytopoulos et al. (2008) showed that the MCS and PERT techniques are significantly improved when using information about previous occurrences when planning future works.

MCS is a probabilistic risk analysis method that models the probability distribution of any risk-prone parameters and simulates the results based on the value randomly selected from the defined distribution (Koulinas et al., 2020).

As compared to MCS, PERT emphasized scheduler's experiences in approximating the longest and shortest possible time of task completion time. Whereas MCS focuses on modelling the probability distribution of each activity Tokdemir et al. (Tokdemir et al., 2019); and thus, MCS is more objective in risk assessment.

Researchers have made significant attempts to predict project costs and durations using machine learning techniques, such as Artificial Neural Network (ANN), Support Vector Machine (SVM), and Bayesian Network. For example, (Peško et al., 2017) developed a framework for predicting road costs and durations using ANN has been presented with limited success on six different model architectures. Attal (2010) used the overall duration and cost of highways projects to train a series of ANN to determine the key project features to be used in a duration prediction model while Hola & Schabowicz (2010) took a more specific approach, using ANN to predict earthworks durations. However, it remains challenging to interpret the results of ANN since the backpropagation algorithms made it difficult to interpret the correlation between input data and output results.

Another adequate supervised classification method – Support Vector Machine (SVM) is more interpretable, since SVM is solving a constrained quadratic programming problem (Joachims, 1999). SVM distinguishes different classes by learning a separating hyperplane (Vapnik, 2000). An important component in SVM is kernel, which enlarges the feature space in a higher dimensional space and accommodates a non-linear boundary between classes (James et al., 2013). The function of kernel is to take data as input and transform it into the required form. The most frequently used SVM kernels include linear kernel, polynomial kernel, and radial basis function kernel are (James et al., 2013). The advantage of SVM is that it suits small sample size and the optimal solution can be found based on the existent information of sampling rather than infinite sampling which, in return, improves computational efficiency (Guo et al., 2006). SVM seems to be less frequently used in construction schedules, but has been used to predict construction contract default (Huang & Tserng, 2018) and contractor's qualification (Lam et al., 2010). Compared with ANN, SVM has more interpretable results and supports small sample training with less computational costs.

Table 1 summarizes existing risk analysis approaches for construction scheduling. It is notable that none of the studies present an attempt to use historical schedule data to train their project outcome prediction models. Some use information about project outcomes for label data, but do not utilise the detailed information contained within the schedule file due to a lack of open or ease of accessible big data.

TABLE 1. Summary of existing risk analysis methods for construction scheduling

Method	Includes for duration uncertainty?	Accounts for co-variance between task and risk groups?	Is objective – uses empirical evidence to model risk?
CPM	×	×	×
MCS	√	×	×
PERT	√	×	×
CCPM	√	×	×
BN	√	√	×
CSRAM	√	√	×
ANN & SVM	×	×	√

3. PROPOSED METHOD

The objective of this research is to generate a time-risk prediction model using significant quantities of historic schedule data. We propose a novel hybrid data-driven method that combines the strengths of SVM and MCS to train a robust predictive model to achieve this goal. FIG. 1 illustrates the proposed hybrid method. Specifically, the input of this work is the construction schedule data. Then, the proposed method consists of two major processes: 1) Date preparation and 2) Training and Simulation. The final output is a time-risk prediction made by the predictive model and simulation. In the construction practice, clients define the project scope and contractors initiate the project planning to deliver the project within scope and time. The scope of this work is analysing risks of project delay caused by contractor works including the late delivery of critical construction milestones; whereas project scope changes are not considered in this study, since such changes may lead by many external and uncontrollable factors (i.e., macro-economics and organisation's strategic plan). The following subsections elaborate on each process in details.

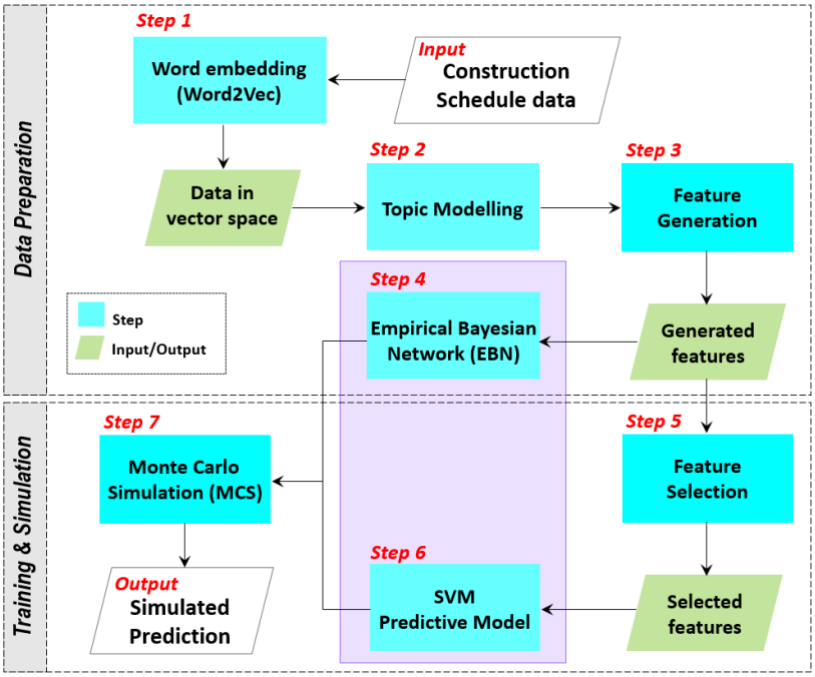


FIG. 1: Workflow of the proposed machine learning method

3.1 Data Collection

The schedule data was collected within two construction and engineering firms by approaching the directors with whom the authors had an existing relationship. One of the organisations is a tier one contractor who directly delivers large infrastructure projects on behalf of public and private sector clients. The other is a schedule management consultancy, who provides services to clients and contractors in many engineering disciplines. The files were provided in the software native Primavera P6 .XER format. In total 560 project files were collected. Of which 302 were valid. Some were invalid due to corrupt native files or a lack of "actual versus planned" information, i.e., the ground truth data. Valid projects were constructed between 2007 and 2019 in the UK, worth of £8.9bn. With a relatively broad range of project included, FIG. 2 illustrates the break down by project discipline. In total, the dataset contains 293,263 tasks, which are classified as "task dependent" in the Primavera P6, meaning that they directly represent a required project deliverable and have a non-zero duration. These task files were converted from the native format into .txt format. All project-related information was structured using the JSON Schema.

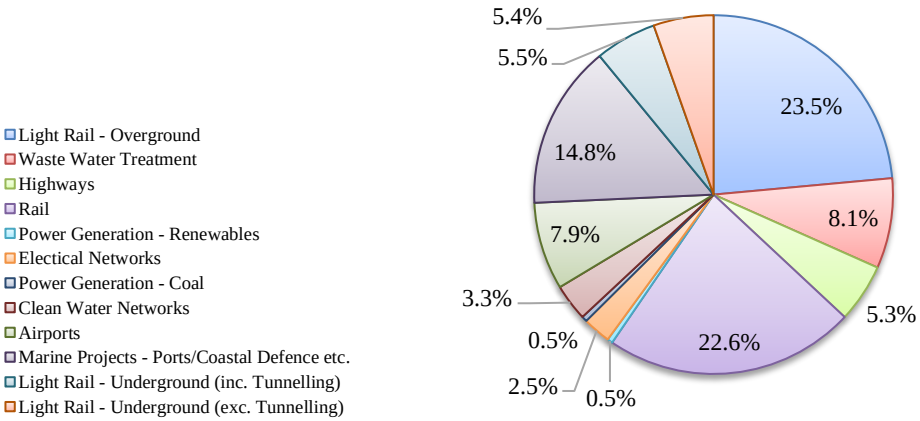


FIG. 2: Database distribution, total number of activities by project type

The weather data including temperature, rainfall, and sunshine, comes from the UK Met Office website (Met Office, 2018). The expected weather is determined by the average weather conditions specific to the region of the project and the month of the year the task is planned for. The averages are calculated from data recorded since the year 1900.

3.2 Data Preparation

The data preparation process aims to identity appropriate topic number in the text files through topic modelling. The input of this process is the construction schedule text file data. The outputs are features and conditional probabilities that will be used for training a classifier for prediction. Generally, a topic modelling method assumes a distribution of topics across a collection of documents, i.e., the schedule files. It then uses the word co-occurrence to predict the distribution of each file to a topic.

Step 1 – Word embedding aims to train a language model to represent the semantic positioning of each work with numeric vectors. Continuous bag-of-works (CBOW) and Skip-gram are most recent methods, which learn word embedding representation and predict target words from context (Pennington et al., 2014), but work in an inverse way (Mikolov et al., 2013). CBOW predicts target words from sources text words; while, Skip-gram predicts source text words from target words (Mikolov et al., 2013). The proposed hybrid method uses Word2Vec, which is a frequently used word vector representation algorithm, to create a neural network for predicting words given their surrounding words from a text corpus. Skip-gram is then used to transform words into multidimensional vectors, followed by Singular Value Decomposition (SVD), which is used to reduce the sparse vectors to 2D eigenvalue vectors. Step 1 converts texts into vectors, which will be used as input in the following steps.

Step 2 – Topic Modelling aims to select an appropriate number of latent topics using model coherence testing and classify a task file into a correct topic (FIG. 3). Topic modelling develops probabilistic generative models to discover the latent semantics embedded in document collection and has demonstrated vast success in modelling and analysing texts (Xie & Xing, 2013). This study uses Gaussian Mixture Modelling (GMM), which is a probabilistic model for representing normally distributed subpopulations within an overall population (Bishop, 2006). Different from other frequently used topic modelling methods – Latent Dirichlet Allocation (LDA) & Latent Semantic Analysis (LSA), which uses a hierarchical process of topic generation (Blei et al., 2003). LDA & LSA works well for large corpora and documents where key words may appear several times, it does not work well for shorter texts. Note that, task description in this context means that the string of text used to describe a single operation in a construction programme. In a construction schedule file, the average task description length is around 5 – 10 words, e.g., “Install Drainage Run 5 to Manhole CP03”.

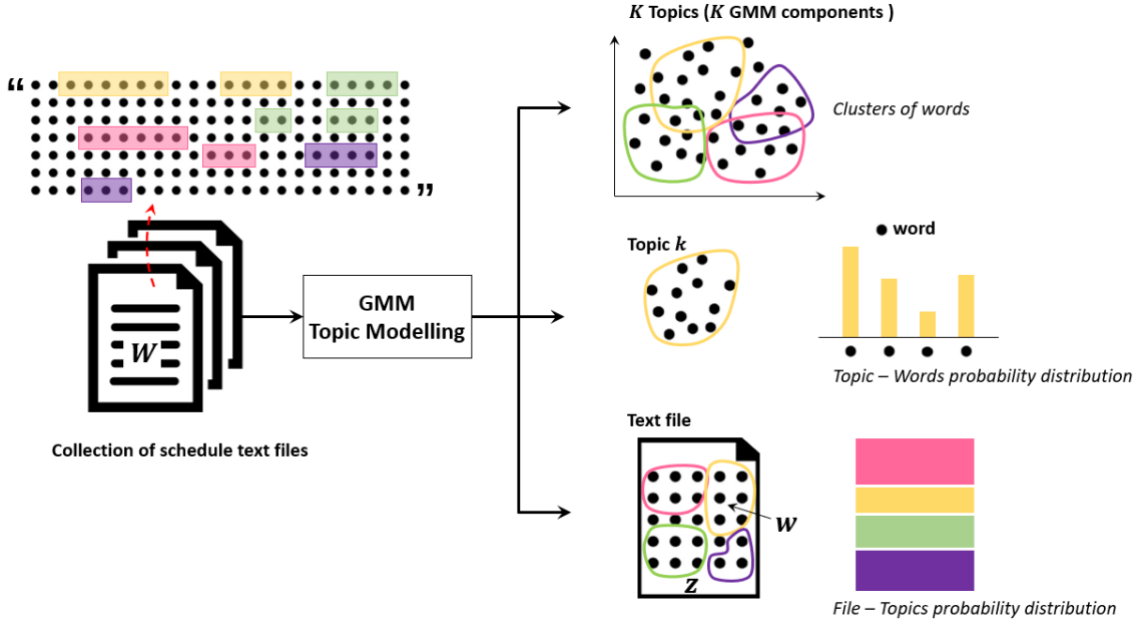


FIG. 3: The tasks of GMM Topic Modelling

To this regard, GMM is leveraged to cluster short texts into topics. The ideal topic count is identified by the point at which the coherence scores move from rapidly changing curve to a plateau or steady increase point (Stevens et al., 2012). Specifically, given a topic z and a set of top N words in z in order of likelihood, and $S^z = \{w_1^z, \dots, w_N^z\}$, the Topic Model Coherence (TMC) score of the GMM (Rangarajan Sridhar, 2015) can be calculated by

$$TMC(z;\, S^z) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} \log \frac{D(w_n^z, w_l^z)+1}{D(w_l^z)}, \qquad \text{Eq. 1}$$

where $D(w)$ is the file frequency of word w and $D(w', w)$ is the co-document frequency of word w and w'. TMC is then averaged across all topics to acquire the mean $\overline{TMC}$, the overall coherence score for the model:

$$\overline{TMC} = \frac{1}{K} \sum_{k=1}^{K} TMC(z_k;\, S^{z_k}), \qquad \text{Eq. 2}$$

where K is the number of topic clusters, namely the latent topics. The notion of latent topics is represented by K components of the GMM. In other words, a K- GMM is fitted to distributed representation of words in 2D space. Naturally, the more topics are assumed, the more coherent the model will become as each topic becomes more specific to certain words. However, selecting the number of topics corresponding with the largest coherence score is not a particularly good solution for clustering. This is because the topic grouping will become sparse and the distinctions between the topics arbitrary, especially with construction domain-specific text datasets. Thus, the point of gradient change to steady increase is normally selected. The GMM provides K different probability score for each word. Summing up each position from 1 – K across the sentence gives a weighted estimation on the topic distribution. The maximum value gives the file topic classification:

$$\sum_{i=1}^{W} p_k^{(w_i)} = \max\left(\sum_{i=1}^{W} p_1^{(w_i)}, \dots, \sum_{i=1}^{W} p_K^{(w_i)}\right), \qquad \text{Eq. 3}$$

where $k \in K$, $w \in W$, and W is a list containing all the works in the task description of a file. p is the probability of work w belonging to topic k.

Step 3 – Feature Generation aims to generate features to be used in the classifier for training. Several features of different types are generated including F1 concurrent tasks in relation to each task, F2 total float values, F3 weather conditions, F4 economic conditions, and F5 schedule quality. Within a project timeline, tasks are either concurrent (can occur simultaneously) or sequential (one task cannot begin until the predecessor). Using the scheduling files, the number of concurrent tasks of each task in a project is counted followed by calculating feature F1, i.e., level of concurrent tasks. To generate F2 total float value, a dummy CPM simulation is run on all projects in the dataset to determine the float values against each task. Specifically, the float value is the subtraction between latest finish and earliest finish of a task in the DAG network. FIG. 4 presents the concept of an example of a 29-day CPM-DAG network. The critical path through the network is identified as the one where the "float" value is zero. It is worth noting that simply taking the "actual start" and "actual finish" criteria from the programme file is not enough for CPM-based simulation. In many cases, where a task is significantly delayed, its succeeding tasks will have commenced before it has completed, even where the task is linked with a "Finish-Start" relationship. This would lead to an overrepresentation of any forecast delay in a subsequent CPM-based simulation. It is therefore necessary to use the Earliest Start date of the succeeding tasks, where this was earlier than the recorded "actual finish" date.

MCS is a probabilistic risk analysis method that models the probability distribution of any risk-prone parameters and simulates the results based on the value randomly selected from the defined distribution (Koulinas et al., 2020). Therefore, MCS focuses on modelling each activity, rather than the logic relationship between activities. Similarly, previous studies that focus on forecasting project delay using MCS did not incorporate the relationship between activities into their model (including Tokdemir et al. (Tokdemir et al., 2019)). Same with MCS, this study looks at the delay risks of each activity on the CPM; therefore, the logic relationship between activities is not included in this study.

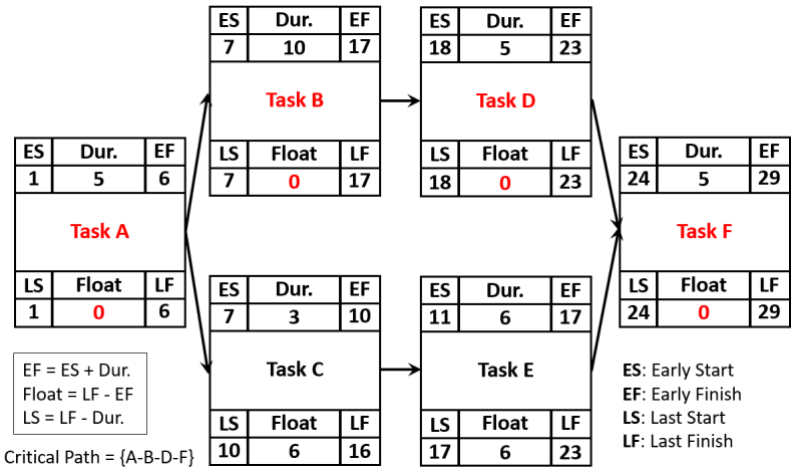


FIG. 4: The CPM-DAG network

Then, F3 weather conditions are determined by the average weather conditions specific to the region of the project and month of the year a task is planned for. Next, F4 economic conditions are an average growth in output in the UK infrastructure sector. F5 schedule quality features are also generated by using Acumen Fuse's default quality criteria. In addition, for model training purpose, the target variables, i.e., the label Y-values, are prepared which are the actual duration growth percentages, calculated using

$$Duration\ Deviaton\ (\%) = \left(\frac{AOD}{ODE}\right) - 1, \qquad \text{Eq. 4}$$

where AOD stands for Actual Observed Duration while ODE for Original Duration Estimate. The process of obtaining AOD is not straightforward. Simply taking the 'actual start' and 'actual finish' criteria from the project file will not enough for CPM-based simulation. In many cases, where a task is significantly delayed, its succeeding tasks will have commenced before it has completed, even where the task is linked with a 'Finish-Start' relationship. This would lead to an overrepresentation of any forecast delay in a subsequent CPM-based simulation. It was therefore necessary to use the earliest start date of the succeeding tasks, where this was earlier than the recorded 'actual finish' date.

Step 4 – Empirical Bayesian Network (EBN), aims to calculate the probability of the duration deviation of each task. The concurrent task (F1), task criticality, i.e., float (F2), weather (F3), and market (F4), are factors to decide whether to refer the estimated duration of a task in the CPM network to the classifier. Specifically, a "Bayesian Trigger (BT)" (FIG. 5) is used as a duration decision decider in the later simulation process.

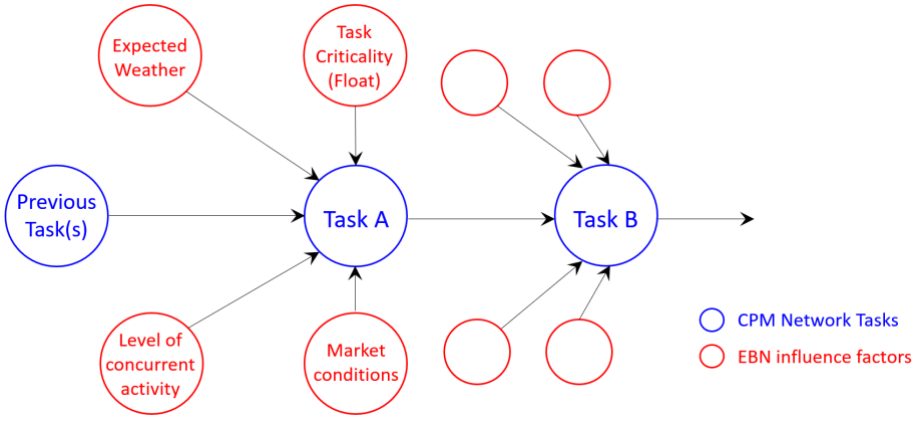


FIG. 5: Bayesian Trigger

The EBN is simplified to a Naïve Bayes (NB) problem, giving a probability that Task A (FIG. 5) will be delayed based on all its known parents. The probability that Task A deviates from its estimated duration can be represented

in Eq. 5. In other words, each of the EBN inputs is converted into binary values. For example, predecessor task delayed ← 1, not delayed ← 0, sector economic output growing ← 1, in decline ← 0 and so on. Then, conditional probability tables are constructed for each variable such that a NB probability calculation can be undertaken to determine the likelihood of task duration variance. Specifically,

$$\text{Probability of deviation given variables: } P(D|W,M,F,C,T) = P(W|D).P(M|D).P(F|D).P(C|D).P(T|D).P(D), \quad \text{Eq. 5}$$

$$\text{Probability of no deviation given variable: } P(\neg D|W,M,F,C,T) = P(W|\neg D).P(M|\neg D).P(F|\neg D).P(C|\neg D).P(T|\neg D).P(\neg D),$$

$$\text{Posterior probability of deviation given observed variable: } P(D|x) = \frac{P(D|W,M,F,C,T)}{P(D|W,M,F,C,T)+P(\neg D|W,M,F,C,T)},$$

where W is the binary representation (BR) of each weather metric expected for that time of year (better than yearly average ← 1, worse than yearly average ← 0), M is the BR of the market conditions at the time of estimate (sector output growing ← 1, sector output in decline ← 0), F is the BR task float as defined by the CPM (higher than average ← 1, lower than average ← 0), C is the BR of the number of concurrent tasks compared against the average for the project (more than average ← 1, fewer than average ← 0), and T is the BR of the known outcome of the previous task in the network (previous task deviated from estimated duration ← 1, previous task did not deviate ← 0). $P(D)$ is the prior probability of deviation from the estimated duration and $P(\neg D)$ is equal to 1 – $P(D)$. The joint probabilities are calculated from the dataset by counting all co-concurrence and dividing them by the total number of occurrences. The joint probabilities associated with each variable change depending on the task family, which helps to factor in the key concept of co-variance for similar tasks. When the probability of deviation has been calculated, a random number generator is used to select a number between 0 and 1. This number is then used as the "deviation decider": if $0 \leq$ random number $\leq$ probability of deviation, then, the Boolean value of "True" is assigned by the BN calculator, the duration is then passed onto the classifier for deviation prediction.

3.3 Training and Simulation

Training and simulation aim to first train an SVM classifier to make project duration deviation prediction using the historical schedule data and leverage MCS to calculate an averaged finishing time of each task over a nominal number of iterations. The inputs of this process are the features generated in Step 3 and the task duration deviation probabilities generated by the EBN in Step 4. The output is the simulated prediction. We demonstrate in FIG. 6 the simulation procedures.

Step 5 – Feature Selection. Irrelevant or partially relevant features can negatively impact the model performance. Therefore, before training a classifier, feature selection is performed in order to identify the most related features and removing the irrelevant or less important features with do not contribute to the target variable in order to achieve better accuracy and reduce the overfitting problem due to noisy data. Kohavi and John (1997) categorized feature selection techniques into two groups: filter method and wrapper method. Compared with the wrapper method, the filter method is less computationally expensive (Weston et al., 2001). Therefore, the filter method is used to select feature. Applicable approaches within the frame of filter method include ANOVA (analysis of variance), chi-square test, and mutual information. ANOVA and Chi-square are the most effective measures (Yang & Pedersen, 1997), but apply on continuous and categorical feature, respectively; whereas mutual information could lead to an NP-hard (note: NP means the nondeterministic polynomial time) optimisation problem (Venkateswara et al., 2016). Since data used in this study is continuous data, the ANOVA F-test classification scoring algorithm is therefore used to evaluate each feature the extent to which that feature can be explained by the data. The higher the score the greater the effect of the feature on the output of the prediction model. Only the highest and most statistically significant features are selected.

Step 6 – SVM Predictive Model aims to predict the expected duration deviation of a task. Radial Basis Function (RBF) kernel is used to improve the predictive performance as it allows SVM to fit the maximum-margin hyperplane in a transformed feature space for nonlinear regression instead of classification. This is favourable because the desired output is unlikely to be an exact repeat of a previous occurrence, rather a numerical prediction which gives a direction of expected deviation and an order of magnitude that stops short of the extremes observed

in the extremely random datasets expected in construction schedules. The trained SVM will be used as input in the next step to simulate the delay of a project.

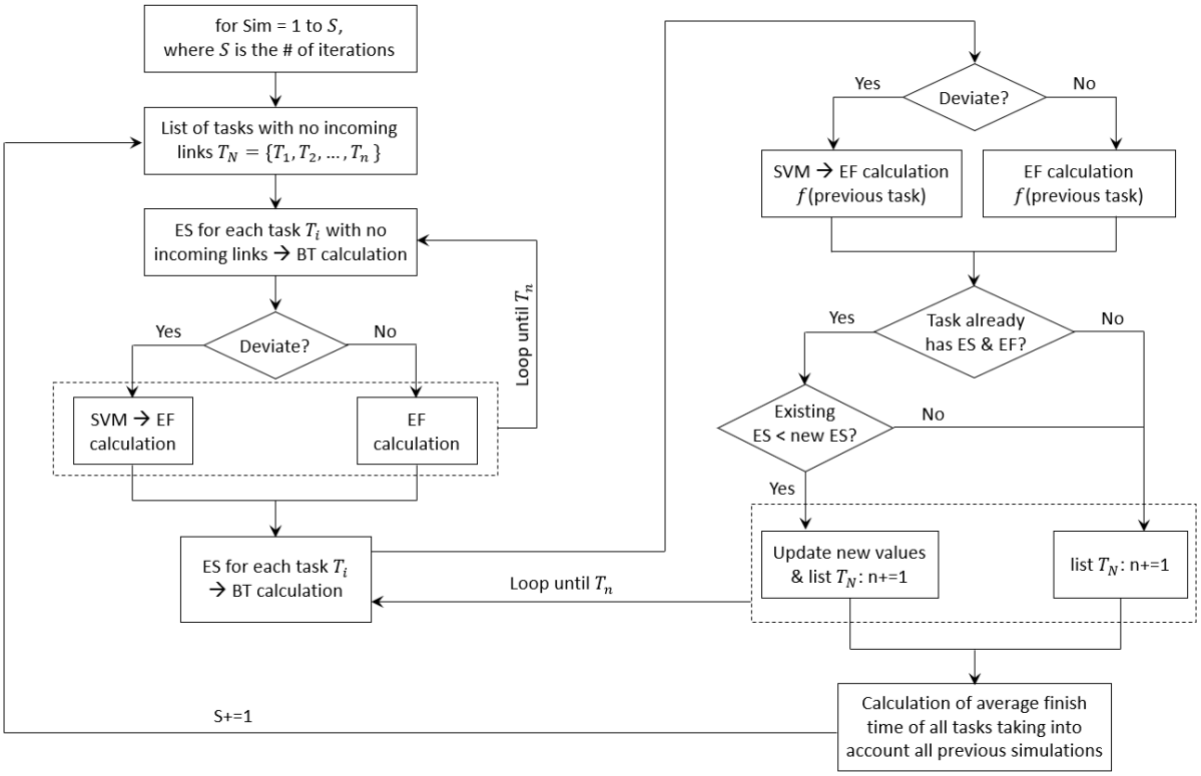


FIG. 6: Proposed Risk Simulation

Step 7 – Risk Simulation aims to calculate the average finishing date of each task over a nominal number of simulation iterations. This gives a time-distributed profile of tasks that can be used as the output of the risk analysis simulation. Specifically, the BT is used to factor in uncertainty based on a traditional MCS, during which, the outcome of the previous task and all other variables become known such that the EBN can then be converted to NB problem. Then, the SVM is used to predict the deviation parameters followed by using a beta distribution to determine the final duration deviation factor to ensure that the duration uncertainty parameters change from one simulation iteration to the next. The distribution is parameterised by the SVM model prediction (Eq. 6). With each iteration, both the SVM input values and the duration deviation estimation change dynamically, giving a broader range of possible values for the duration of each task.

$$Final\ Simulated\ Duration\ Deviation = SVM_i + \left(R_i \times \frac{SVM_i}{3}\right), \quad \text{Eq. 6}$$

where SVM_i is the SVM prediction for a given task and R_i is a random number from a normal distribution expressed as a number of standard deviations from the mean.

4. EXPERIMENTS AND RESULTS

4.1 Experiments for data preparation

Data cleaning. Activity names were first pre-processed before conducting Step 1 Word embedding of the proposed method to ensure the accuracy. Pre-processing steps include tokenisation, lemmatisation, stemming, and removing stop words and meaningless words. Tokenisation is a process that transform text into tokens which are readable in computer language (Manning et al., 2015). Lemmatisation and stemming are used to reduce the effects of inflectional form and words' morphology (Habash et al., 2009). Stop words (e.g., "and" "the") and punctuation were removed, to get eliminate the unmeaningful words. Then, identity names are very common in construction

schedule, with roads, buildings, clients, subcontractors, and locations appearing in nearly every task description. These names were removed because they add very little value in terms of semantic understanding as they are not normally specific to the task, or the type of operation being planned and may not span different projects.

Step 1 – Word embedding. Now, a Skip-gram neural network was used to train the vector space through Word2Vec using words' surrounding words. The hyperparameters of the Word2Vec model are the number of training iterations $N_{\text{iter\_Word2Vec}}$, the context window length $L_{\text{win\_Word2Vec}}$, and the vector size $S_{\text{vect\_Word2Vec}}$. Specifically, empirical experiments were conducted for $N_{\text{iter\_Word2Vec}}$ in a range of 5 to 50. No significant improvement in performance was observed, thus, $N_{\text{iter\_Word2Vec}}$=5 was set as a default value. $L_{\text{win\_Word2Vec}}$ was set to be 8 based on the average concatenated task description and the section heading length. Then, the default vector size $S_{\text{vect\_Word2Vec}}$=100 was used (Das et al., 2015). In total, 5,496 unique words were used in the word embedding process, training across a dataset of 3.1 million co-occurring words. The dimensions of word vector were reduced to 2D using the SVD in Principal Component Analysis. FIG. 7 illustrates the Skip-gram vectorisation result.

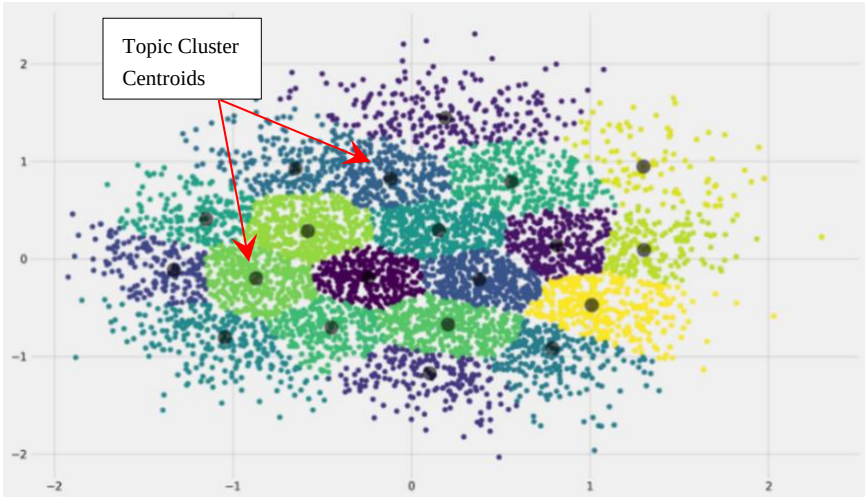


FIG. 7: Proposed Risk Simulation

Step 2 – Topic Modelling. The coherence of the GMM model was evaluated for a range of different topic counts from 1 to 50. FIG. 8 shows the results of the coherence testing for this range. The line settled between 10 and 25 topics, with higher scores towards 20. Note that the design manual for roads and bridges (DMRB, 2019) has 16 volumes. If more topics are allowed for procurement, rail, and mechanical & electrical trades, then 20 is a reasonable number. FIG. 9 demonstrates the topic probability distributions for 2 task descriptions (i.e., Task Description 1 – "install drainage from manhole 3 to manhole 4" and Task Description 2 – "construct filter drainage to catchpit 8"). As shown, despite few marching words and some very generic terms such as "construct" and "install", the two descriptions were both identified as belonging to topic 15. FIG. 10 shows several examples of the word clouds constructed from each of the 20 selected topics.

Step 3 – Feature Generation. Five types of feature were generated, including concurrent tasks, total float, weather conditions, economic factors, and schedule quality features. Given a task, the feature F1 level of concurrent tasks was calculated by calculating the fraction between the number of concurrent tasks of this specific task and the total number of occurrences. Next, F2 total float value associated with each task was generated by running a CPM simulation. Then, we used the UK Met Office website (Met Office, 2018) to generate the F3 weather conditions. Then, F4 economic conditions are defined using the quarterly average growth in output in the UK infrastructure sector as recorded by the Office for National Statistics (ONS, 2019). Acumen Fuse's quality criteria were used to generate the F5 schedule quality features.

Step 4 – EBN construction. We calculated the averaged results of the conditional probability weighting for the binary variables in the EBN analysis. The values varied by GMM topic, with a minimum possible probability of deviation of 0.5% and a maximum of 99.9%. FIG. 11 shows that the two biggest factors influencing the probability of delay were the performance of previous linked tasks and the level of concurrent task on site. The weather

variables made almost no different to the probability of delay and the market conditions exerted only a minor effect in favour of deviation.

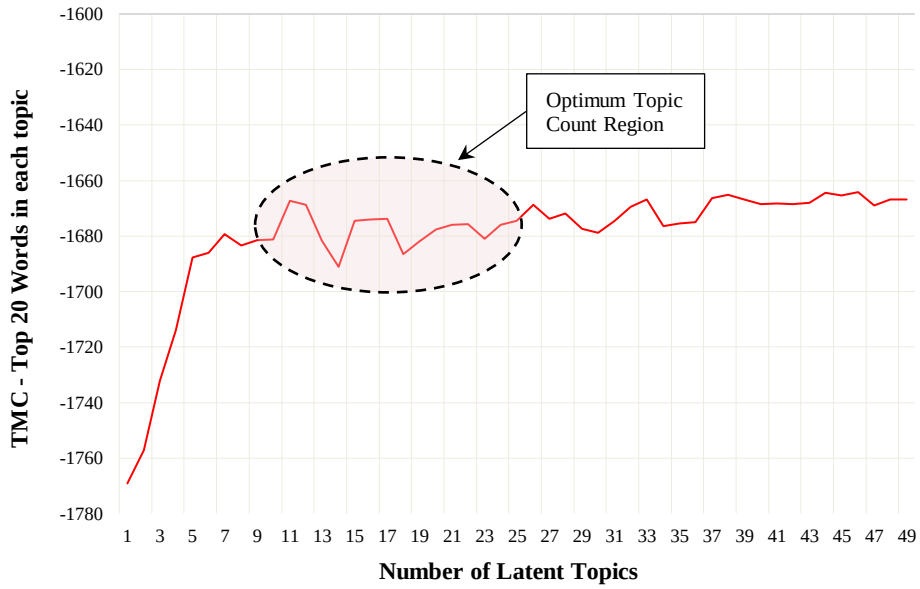


FIG. 8: GMM Topic Coherence Testing

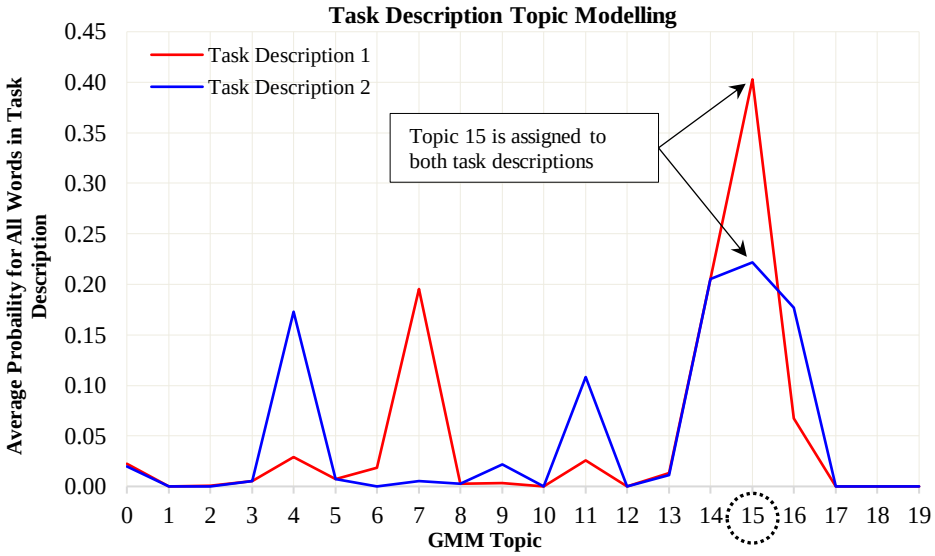
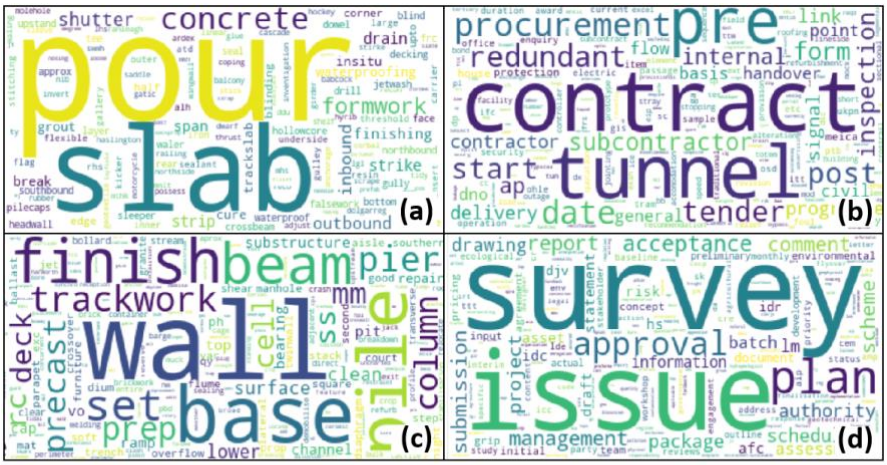


FIG. 9: Topic probability distributions for 2 task descriptions

From the topics identified by the GMM, the likelihood of duration deviation was measured and is presented in FIG. 12 which shows that the least predictable topics were those associated with early-stage works/groundworks and fit-out trades which involve interfaces between many different parties. The prior probability of deviation from the estimated duration $P(D)$ was around 46%. The swing between the most (57%) and least (29%) predictable topics was large at 28%, which is highly significant at the level of hundreds of thousands of tasks.



(a) Topic 4: Insitu reinforced concrete structures

(b) Topic 10: Procurement of materials and subcontracts

(c) Topic 15: Substructures – foundations, beams, columns, piers, etc.

(d) Topic 19: Client/stakeholder interface, review and approval activities

FIG. 10: Examples of Word Clouds drawn from GMM

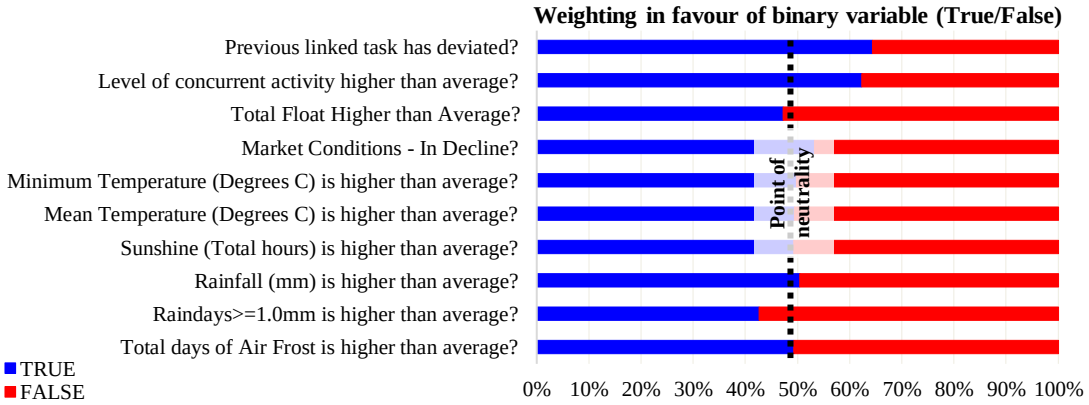


FIG. 11: Task Duration Deviation Probability by Measured Feature

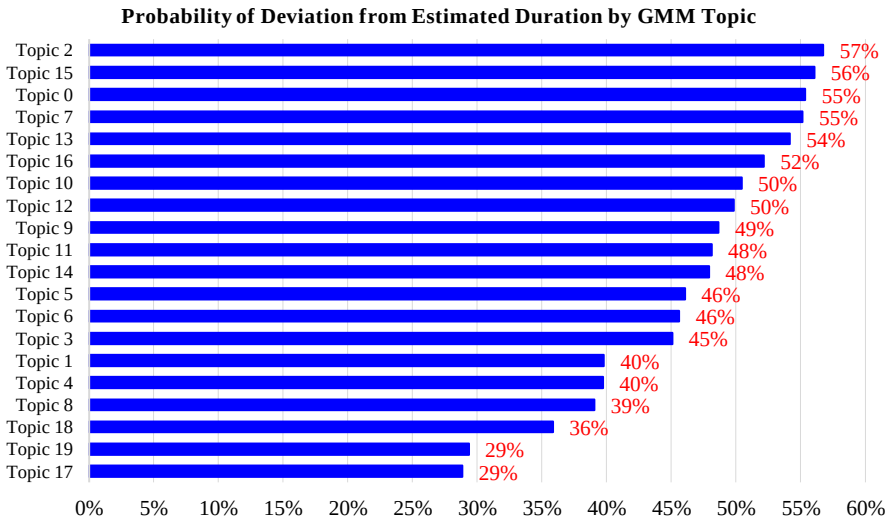


FIG. 12: Probability of deviation from estimation by GMM topic

Task Covariance. FIG. 13 demonstrates the linked task covariance, i.e., the predecessor task variance against observed task variance. A statistically significant correlation ($r=0.22$, $p<0.01$) was obtained using Pearson's correlation coefficient, suggesting that at the scale of megaprojects, with many thousands of activities, the covariance is likely to have a significant effect.

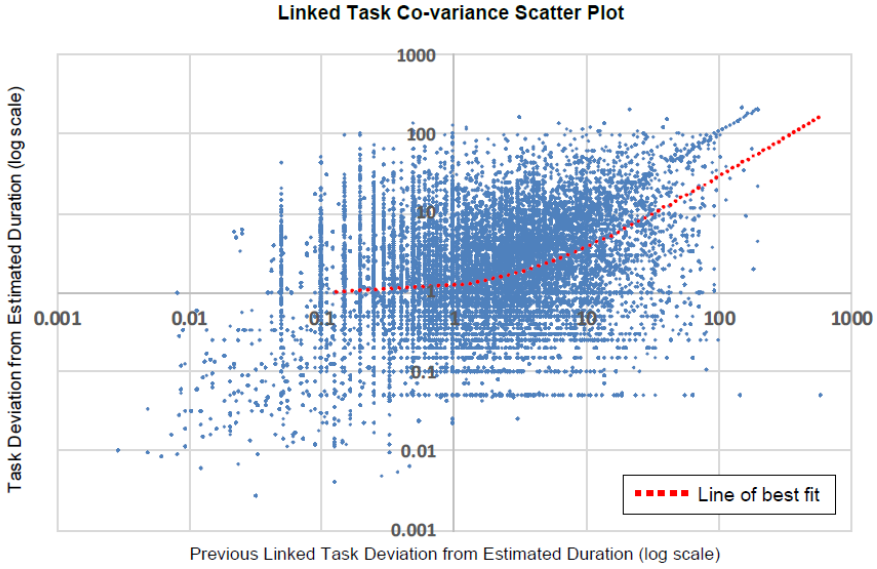


FIG. 13: Covariance between linked tasks

4.2 Experiments for model training and Case Study

Step 5 – Feature Selection. Following the feature generation, each task had around 72 task-specific features available for training in the prediction model. The ANOVA F-test was applied to reduce the number of features down to 13. FIG. 14 shows the 13 features in order of highest to lowest ANOVA F-test score.

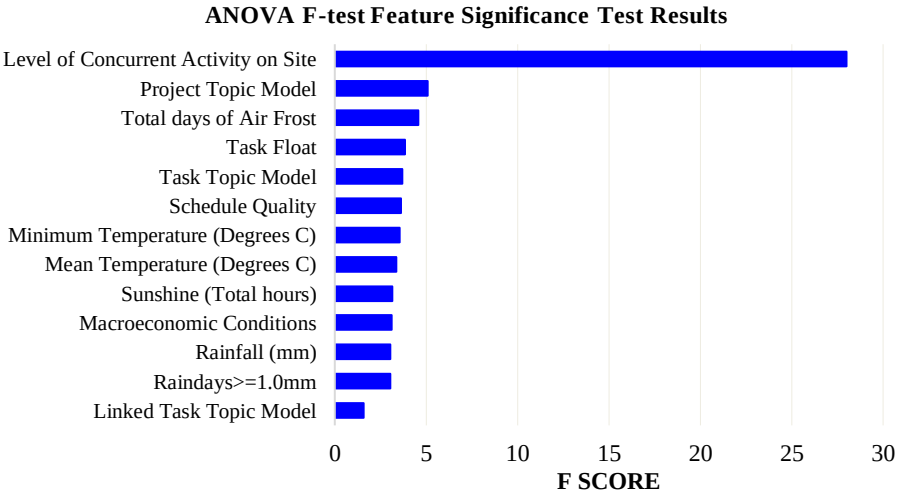


FIG. 14: ANOVA F-test scores of features

Using a confidence interval of 95%, the Level of Concurrent Task on Site was the only significant feature with an ANOVA F-test p-value < 0.05. Other factors i.e., total float, weather, economic conditions, and schedule quality were deleted as they were not statistically significant. This comes as no surprise because, during the construction period i.e., 2007 – 2019 of those selected projects, there were no catastrophic disasters in the UK and no significant fluctuations in local companies' cashflow. Consequently, the level of concurrent tasks was included in the SVM along with the estimated duration, linked task duration deviation, and the full 20-topic probability distribution

from the GMM. The final combination of feature sets was found by analysing the coefficient of the determination R2 value for each set of SVM features: all features R^2_{all} 0.08, the top eight features (FIG. 14) R^2_{top8} 0.09, and the selected four R^2_{set} 0.24.

Step 6 – SVM Predictive Model. The SVM training dataset was split into a training set (80%) and a validation set (20%) (Crowther & Cox, 2005). A sample set of 10 different types of large-scale construction project worth of a total of £1,25bn was randomly selected to be held back from model training to test the proposed hypothesis. A planner's survey was also conducted to enable validation of the project risk simulation procedure. The design of the planning survey was a structured questionnaire based on the following key project success-determining features: (1) The design maturity and scope definition at the time of the schedule estimate (Tulke et al. 2008), (2) The form of contract and payment mechanism (Assaf and Al-Hejji 2006; Suprapto et al. 2016), (3) Experience of the contractors in the relevant field (Chan et al., 2004), and (4) The tender method (including any early contractor involvement) (Suprapto et al., 2016). This information was gathered to help determine to what extent the risk simulations conducted are representative of real construction project complexity and the degree to which the stochastic nature of the simulation process affects simulated outcomes.

Step 7 – Risk Simulation for Case Study. Step 7 randomly selected 10 projects as case studies projects and ran a standard MCS and the proposed hybrid methods on these case study projects using inarticulate risk parameters derived from existing work, i.e., based case = 90%, most likely = estimated duration, worst case = 130%. A binary search function and multiprocessing tools were used to boost the simulation efficiency (i.e., a simulation of 20,000 tasks through 100 iterations take around 4 minutes with 12 processors). The results of 10 case study projects using MCS and the proposed hybrid method are provided in TABLE 2. The results suggest that 8 (indicated as bold and italic) of the 10 projects were better predicted than MCS because the predicted weeks were closer to that of the actual finishing weeks. In general, the proposed hybrid method is significantly more accurate than the MCS. The standard prediction error was 36.6 weeks for MCS while 16.7 weeks for the proposed hybrid method. This suggests that the proposed hybrid method was 2.2 times more accurate than a standard MCS model. In other words, a significant improvement in the prediction accuracy was achieved by 54.4%.

TABLE 2. Simulation Results of 10 case study projects using MCS vs Our method (in weeks)

Project	Nature	Value (million)	Iterations	Actual Finish	MCS (Mean)	Ours (Mean)
1	Road construction	£40	500	+0	+59	***+12***
2	Coastal Protection	£30	500	+57	+13	***+85***
3	Coastal Protection	£50	100	+31	+30	+26
4	Stadium Construction	£70	500	+21	+3	***+26***
5	Airport Renovation	£10	500	+27	+1	+1
6	Port Renovation	£40	500	+26	+1	***+38***
7	Link Road	£15	500	+28	+0	***+36***
8	Link Road	£45	100	+49	+23	***+31***
9	Light Rail	£600	100	+43	–15	***+39***
10	Light Rail	£350	100	–28	+25	***–1***

+ delay, – improvement

With the assistance of the planner's survey, authors analysed these projects with task count histogram by plotting the baseline schedule position against the proposed hybrid method, with a list of the top 5 most delayed tasks. Two examples (Project 1 and Project 3) are demonstrated in FIG. 15. Specifically, the actual finish date of **Project 1** was aligned with contractual complete date. The prediction result indicated that a delay of 12 weeks, which is much better than that of the MCS. The survey of the planner, however, highlighted a number of key issues that required mitigation a cost. One of these was a major weather event, some of the time for which was recoverable through "force majeure" clauses in the contract, much of it, on the other hand, was mitigated at cost. Additionally, there was significant duration-variant around the dry-stone wall, which lined the route. There were significantly underestimations, and the very specialist trade resource required to complete these walls was in short supply. The proposed risk analysis method successfully predicted this, with some of the most significantly delays forecast on the stone walling activities. Despite Project 1 being delivered within the budget and initial schedule, the results highlighted some significant risks which might have enabled the contractor to increase their contingency for these items, potentially enhancing their project margin for the project.

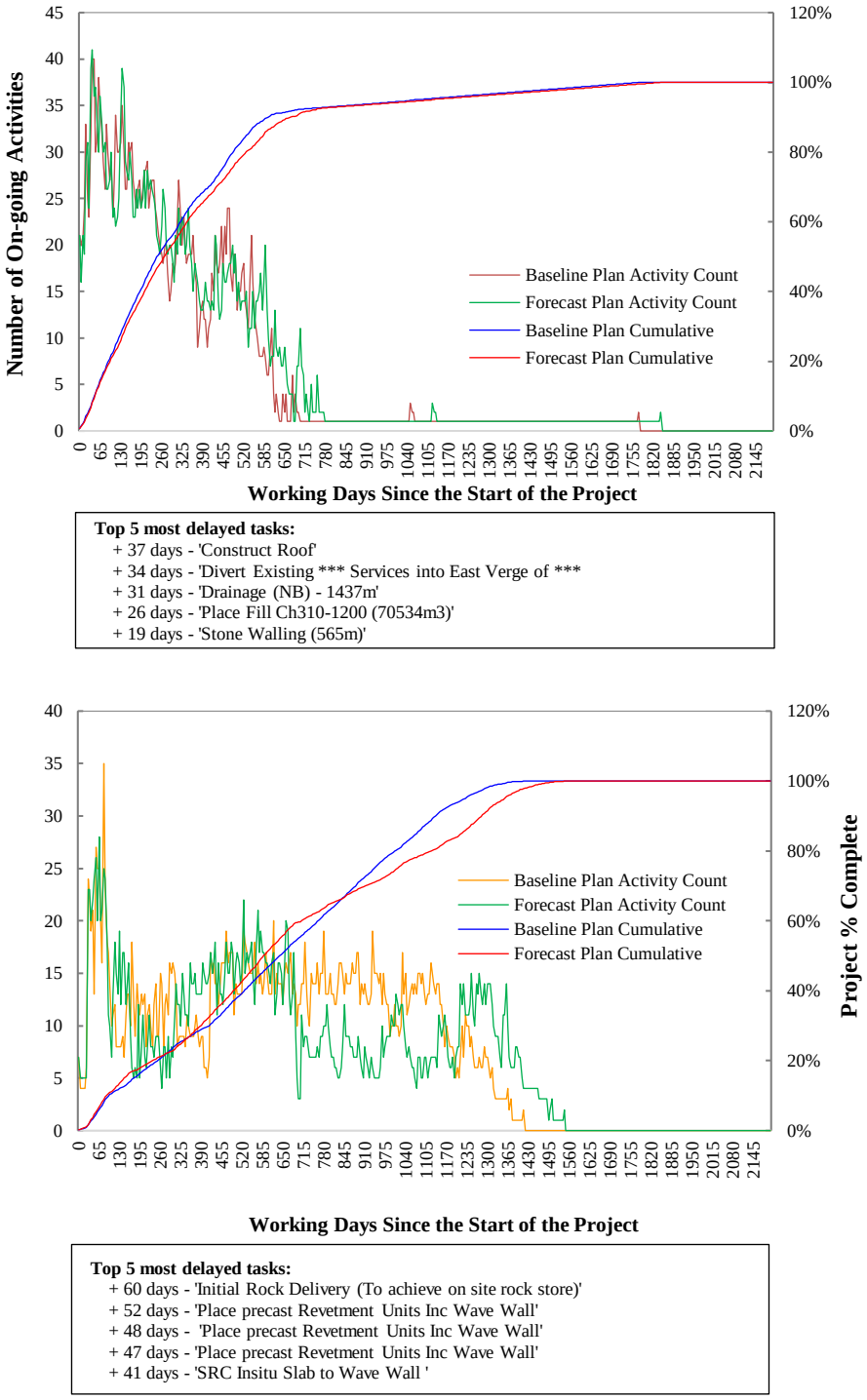


FIG. 15: Project activity account histogram Project 1 (Upper) and Project 3 (Bottom)

Project 2 simulations predicted a significant delay of 85 weeks to the project. Even with knowledge of the large delays suffered on this project, this was a surprising result. This project was undertaken in a coastal marine environment and the key risks to the project were: high winds, high tides, storms, interfaces with third parties, and interfaces with existing infrastructure. This project was significantly delayed as a result of several storm events – causing major washouts, cranage delays as a result of high winds, and the poor condition of the existing flood defences, some of which housed third party-owned equipment. Many of the delays correctly predicted by the

simulation were to the sea-side elements of the project, such as concrete works to the face of the coastal defence. The extreme nature of the forecast delay would almost certainly have raised some concern at project inception and might have helped inform a strategy to avoid the financial losses suffered on this project. Similarities existed between Project 2 and 3, in that they were both in a coastal marine environment and both on the west coast of the UK.

The form of construction for **Project 3**, however, meant that it was significantly less susceptible to deleterious effects of inclement weather and existing sea-side infrastructure. Key risks on this project were concerned with quarry quality and capacity, material delivery productivity, interfaces with environmentally sensitive areas, local residents, and existing land-side infrastructure. The area of risk identified by simulation was concrete works in the marine environment – which would have been just as risky as Project 2 – and rock material deliveries, which presented a genuine concern for the project team. The proposed hybrid method failed to pick up any significant risk to the landside works, which could be understandable, given that the problems experienced (asbestos landfill and 'provisional sum' third party works) could not have been apparent from the initial schedule.

The major issues in **Project 4** were identified as the mechanical and electrical (M&E) interface with internal fit out trades, design change, and land access constraints. The proposed hybrid method correctly assigned significant risk to the M&E installation and fitted out trades as well as the design development activities. There were no risks associated with land access highlighted through simulation, which is perhaps to be expected, given that these client responsibilities are represented by milestones (0-day duration) and usually have no incoming links.

Project 5 singularly failed to predict anything close to the experienced level of delay. This project was delayed due to early-stage design issues, the scope of which was not adequately reflected in the schedule file itself. This was particularly challenging to simulate as the project file only contained 130 tasks. The average project in the training dataset contained 1,400 tasks. It is likely that 130 activities were insufficient to model the risk on a large infrastructure project. This could be an anomalous result due to the small number of tasks for the risk simulation methodology.

Project 6 involved the renovation of a port with significant interfaces with existing structures. The key risks were the piling on the existing dock, the unknown ground conditions and existing substructure, and the condition of the existing aging structures. These were additionally subject to tidal working constraints, which exacerbated the delays experienced as a result of the realization of the above risks. The output succeeded in assigning large time delays to the piling works and tidal interface works.

Project 7 and **Project 8** were similar road construction. Both were delayed by late design, land access constraints, and utility diversion delays. The proposed hybrid method successfully predicted delays to the land constraints, the utility diversions, and the detailed design, all of which made up around 90% of the 50 most delayed tasks. Project 7 was particularly successful compared with MCS, which forecasted no delays to the project.

Project 9 and **Project 10** represented the largest testing projects. The simulations were reasonably successful, in that one correctly predicted a saving (Project 10), while the other was close to the actual position in terms of delay (Project 9). On closer inspection, Project 10 contained many constraints throughout the project, which have perhaps artificially held the completion date at only 1 week of saving. There was no particular trend to the task duration deviation in Project 10; it might because the SVM anticipated lower durations throughout the schedule.

4.3 Comparing with MCS

The null hypothesis H_0 of this work is that the conventional MCS outperforms the proposed method. It is tested by running a trial on a sample of completed projects. A one-tailed paired t-test statistic is used to accept or reject the hypothesis. A significance of p-value of <0.05 (95% confidence interval) is required to reject the null hypothesis.

The hypothesis was tested by using a one-tailed paired t-test:

$$Null\ hypothesis\ Test\ Statistic: t\ Stat = \frac{\sum D}{\sqrt{\frac{n.\sum D^2 - (\sum D)^2}{n-1}}}, \quad \text{Eq. 7}$$

where D is the difference between the error in each of the paired prediction samples, D^2 is the squared error and n is the number of samples.

$$t\ Stat = \frac{203}{\sqrt{\frac{10 \times 7295 - 203^2}{9}}} = 3.41$$

Critical statistic: $t\ Crit$ = 1.833 < 3.418

The significance p-value was < 0.007. Thus, H_0 can be rejected, indicating that the proposed method is significantly more accurate than the MCS. This study trained the hybrid method on 302 projects and tested the hypothesis on 10 projects. The statistical power would be enhanced with an increased confidence level and a reduced margin of error if more trials are investigated. The superiority of the proposed hybrid method than MCS might affect the confidence in hypothesis testing, whereas the robustness of the proposed hybrid method remains strong.

5. DISCUSSIONS

This research presents a hybrid method to generate a time-risk prediction model using significant quantities of historic construction schedule data (302 construction projects with 293,263 tasks). Currently, no single existing technique can conduct risk analysis in complex and large construction projects based on available schedule data. Using collected historic schedule files, the proposed hybrid method starts by using GMM to classify each task description of a schedule file. Then, it constructs an EBN by calculating the conditional probabilities associated with the independent and duration deviation variable followed by transforming these conditional probabilities into NB problems for estimating the deviation duration. Finally, an SVM is used to predict the duration deviation followed by running an MCS to give an averaged prediction. The experimental results on 10 large-scale construction projects suggest that despite the non-trivial standard error of 16.7 weeks, the overall prediction outcomes are statistically proved to be significantly more accurate (54.4%) than a conventional MCS.

The results of ANOVA F-test used to select feature indicate that temperature, sunshine, rainfall and macroeconomic conditions are not significant to test results. Therefore, the proposed method is expected to produce robust results for construction projects undertaken in different regions and different time periods where the industry culture is similar to the modern UK construction industry culture (except extreme weathers and disasters). This study tested the proposed hybrid method mainly on infrastructure construction and heavy construction as summarised in Figure 2, and it holds the potential to be implemented in other industries, such as commercial, manufacturing, oil and gas, and so on. Authors found that weather and macroeconomics condition are not statistically significant features in this proposed method in Section 4.2. Therefore, authors argue that weather and macroeconomics condition during 2007 and 2019 in the UK are not affecting the final results of the proposed hybrid method. However, other project time periods and locations may affect the final results.

The proposed hybrid method estimates the likelihood and consequence of uncertainty based on Bayes' Theorem and historic data. Hence, the proposed hybrid method is more appealing to the industry compared to MCS and its variants (e.g., case-based MC and agent-based MC) that are limited in their scope and by their subjectivity. PERT, another frequently used method, approximates the longest and shortest possible time that each task will task based on scheduler's experience. Hence, the success of PERT replies on scheduler's experience; whereas the proposed hybrid method provides solid statistical foundation for schedulers to forecast uncertainty without such prerequisite.

The proposed method in this study differs from other studies in terms of prediction accuracy and features used. Machine learning based risk prediction studies focus on classifying the level of delays based on certain features. For example, (Gondia et al., 2020) used a number of risk factors to predict the level of project time overrun (i.e., low (< 30%), medium (30% - 60%), and high (> 60%)). However, these studies primarily use as-built schedules to predict the level of project delay of new schedules rather than simulating risk factors to estimate delay durations. Some researchers developed MCS based framework to estimate delay duration (e.g., (Tokdemir et al., 2019) and (Nasir et al., 2003)). The framework developed by (Tokdemir et al., 2019) indicated the project duration would be equal to or shorter than the base case scenario was only 30.6%. The result of our proposed hybrid method is 54.4% more accurate than the MCS. The reason might relate to features used in model training. For example, (Tokdemir et al., 2019) used number of activities and number of workers involved in the project. (Nasir et al., 2003) simulated

project delay in days using a number of risk variables. Whereas the hybrid method in this study included risk variables and also captured the textual data in construction schedules.

Implications to Practice

The proposed method could be used to assist planners in modelling uncertainty of project scheduling while enhancing the level of confidence for quality planning. The topic modelling techniques proposed in this research could be used for a range of other purposes including automatic schedule alignment of costs and 3D objects in 5D modelling. Similar natural language techniques could also be used to monitor sentiment across a company's portfolio of projects, delivering a contemporaneous health check. Construction schedules are possibly the most up-to-date and comprehensive records on large construction projects; harnessing the value of the data they contain is a significant area of opportunity.

In the concurrent business climate, there is an understanding that data have value (Akred & Samani, 2018). Understandably, construction companies may be unwilling to part with their data in the belief that they are giving away value with no financial return. Given the average profit margins (0.5%) and the money wasted on delay (£19bn p.a.) discussed in section 1, however, it is unlikely that any return generated from the relatively small amount of data held by any one construction company will outweigh the financial benefits of improved construction schedule confidence produced by evidence-based, data-rich innovation.

6. CONCLUSIONS

This research presents a new framework for construction schedule risk analysis. It represents the first approach to use machine learning to pre-process noisy construction schedules for a novel hybrid application of Monte Carlo Simulation (MCS), Empirical Bayesian Networks (EBN) and Support Vector Machines (SVM). The proposed solution was trained and tested on large infrastructure projects. It built on the work of several studies discussed in section 3 to describe how machine learning can be used to approach the problem of understanding construction schedules. It has been shown that using large datasets to train prediction models can lead to superior simulation results on a limited sample, doing so with a technique eminently scalable to larger datasets and higher number of test projects. In tackling the problems associated with construction schedule data, some valuable contributions have been made to research around applied artificial intelligence.

There is no single technique that can be applied to the complex task of measuring risk in a noisy sequence of construction operations, using only the data contained with a Gantt chart schedule. Understanding how multiple techniques can be employed to add meaning to a construction schedule dataset, however, is a useful contribution towards leveraging the vast quantities of data available to improve the estimation of time risk on infrastructure and construction projects. Using these techniques, machines can learn about highly complex construction projects without being explicitly programmed, which presents some exciting opportunities for construction practice.

Despite starting out with a larger dataset (560 project files), issues with schedule quality led to a significant diminution of the dataset (302 project files). The accuracy rate of the proposed machine learning-based model could be enhanced with access to a larger quantity of construction schedule data. The final testing set could also be increased to better understand how scalable the method can be. Stricter filtering processes could be employed to ensure that, at the very least, the as-built data are of high quality. For the effects of inaccurate as-built records to become insignificant, the number of tasks collected may need to extend into the tens of millions. It should also be noted that each of the trial analyses benefits significantly from hindsight whereas early planning is still not performed well in many cases – it would be not always have been possible to predict these key risks up-front. Incomplete scope definitions would make project schedules to experience considerable changes. One major limitation of this work is that these front planning aspects were not fully taken into account in the proposed method. This will be incorporated in the future research. Another main limitation is the localising task risks. In the future practice, the project team may take corrective actions to avoid project delay. For example, starting successors before the finishing of its delay predecessors and altering the task sequence. However, it is challenging to track the schedule variation during data collection since the project team alters the schedule without storing the old schedules frequently. Therefore, this study modelled the actual scenario of project execution that includes corrective action, rather than the worst scenario of project execution that excludes corrective action.

Construction schedules are possibly the most up-to-date and comprehensive records on large construction projects. However, getting access to this data remains a challenge. This is also another major limitation of this study since data used in this study are collected from two construction companies only; and thus, results of the proposed method may not be generalised to the whole UK construction industry. Some private concerns (INEIGHT, 2019; nPlan, 2019) are beginning to build datasets on the scale required and are likely to be able to deliver meaningful artificial intelligence-based risk analysis and planning within the next five years. It is worth noting that data access is by agreement. However, schedule data are generally proprietary to the companies who produced them and are shared on the basis of non-disclosure agreements or strict conditions of privacy. To ensure that the industrywide and societal benefits of the continued advances in computer science and information technology are realised for the construction industry, public disclosure of construction schedule data is encouraged – with commercially sensitive details redacted where necessary.

Future work will focus on creating a data integrated schedule platform to integrate Gantt chart schedules with risk variables (e.g., local weather data and supply chain related variables) in order to facilitate more accurate machine learning models. Another stream of future work is creating a construction engineering knowledge dictionary to create a public test set for future researchers in training language models, since existing test sets are designed by computer scientist for general training purposes only which lack the engineering keywords.

ACKNOWLEDGEMENTS

This research work is supported by Laing O'Rourke and Cambridge Construction Engineering Master (CEM) programme. The schedule data and planner's survey data were collected within Laing O'Rourke and Plan Ahead Project Management Group. The authors would like to thank them for their huge supports. Any opinions, findings, and conclusions or recommendations expressed in this research are those if the authors' and do not necessarily reflect the views of Laing O'Rourke, CEM programme, or Plan Ahead Project Management Group.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Apollo at https://doi.org/10.17863/CAM.53890 in accordance with funder data retention policies.

REFERENCES

Akintoye, A. (2000). Analysis of factors influencing project cost estimating practice. *Construction Management and Economics*, *18*(1), 77–89. https://doi.org/10.1080/014461900370979

Akred, J., & Samani, A. (2018). Your Data is Worth More Than You Think. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/your-data-is-worth-more-than-you-think/

Assaf, S. A., & Al-Hejji, S. (2006). Causes of delay in large construction projects. *International Journal of Project Management*, *24*(4), 349–357. https://doi.org/10.1016/j.ijproman.2005.11.010

Attal, A. (2010). Development of Neural Network Models for Prediction of Highway Construction Cost and Project Duration [Ohio University]. In *MS thesis* (Issue August). https://doi.org/10.1016/j.advwatres.2005.10.001

Baker, B. N., Murphy, D. C., & Fisher, D. (2008). Factors Affecting Project Success. In *Project Management Handbook* (pp. 902–919). https://doi.org/10.1002/9780470172353.ch35

Bartlett, J. (2004). *Project Risk Analysis and Management Guide*. High Wycombe: Association for Project Management.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*(1), 58–80. https://doi.org/10.1214/088342304000000116

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Blei, D., Ng, A. Y., & Jordan, M. I. (2000). 10.1162/jmlr.2003.3.4-5.993. *CrossRef Listing of Deleted DOIs*, *1*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Bragadin, M. A., & Kähkönen, K. (2015). Safety, Space and Structure Quality Requirements in Construction

Scheduling. *Procedia Economics and Finance*, *21*(15), 407–414. https://doi.org/10.1016/s2212-5671(15)00193-8

Chan, A. P. C., Scott, D., & Chan, A. P. L. (2004). Factors affecting the success of a construction project. *Journal of Construction Engineering and Management*, *130*(1), 153–155.

Cho, C. S., & Gibson, G. E. (2001). Building project scope definition using project definition rating index. *Journal of Architectural Engineering*. https://doi.org/10.1061/(ASCE)1076-0431(2001)7:4(115)

CIOB. (2009). *Managing the risk of delayed completion in the 21st Century*. *January*.

Crowther, P. S., & Cox, R. J. (2005). A Method for Optimal Division of Data Sets for Use in Neural Networks. In R. Khosla, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 1–7). Springer Berlin Heidelberg.

Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian lda for topic models with word embeddings. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 795–804.

Deltek. (2019). *Acumen Fuse: Project Diagnostics*.

Dey, P., Tabucanon, M. T., & Ogunlana, S. O. (1994). Planning for project control through risk analysis: a petroleum pipeline-laying project. *International Journal of Project Management*, *12*(1), 23–33. https://doi.org/10.1016/0263-7863(94)90006-X

Dikmen, I., & Birgonul, M. T. (2006). An analytic hierarchy process based model for risk and opportunity assessment of international construction projects. *Canadian Journal of Civil Engineering*, *33*(1), 58–68. https://doi.org/10.1139/l05-087

DMRB. (2019). *Design Manual for Roads and Bridges (DMRB)*. Standardsforhighways.Co.Uk.

Drury, T., Hunt, T., Peace, S., Peters, R., Pickavance, K., & Tyerman, D. (2018). Managing the Risk of Delayed Completion in the 21st Century. *Iarjset*. https://doi.org/10.17148/IARJSET.2016.3420

Elmaghraby, S. E. (2000). On criticality and sensitivity in activity networks. *European Journal of Operational Research*, *127*(2), 220–238. https://doi.org/10.1016/S0377-2217(99)00483-X

ElZomor, M., Burke, R., Parrish, K., & Gibson, G. E. (2018). Front-End Planning for Large and Small Infrastructure Projects: Comparison of Project Definition Rating Index Tools. *Journal of Management in Engineering*, *34*(4), 04018022. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000611

Fineman, M. (2010). *Improved risk analysis for large projects: Bayesian networks approach*.

Flyvbjerg, B. (2008). Curbing Optimism Bias and Strategic Misrepresentation in Planning: Reference Class Forecasting in Practice. *European Planning Studies*, *16*(1), 3–21. https://doi.org/10.1080/09654310701747936

Fortin, J., Zieliński, P., Dubois, D., & Fargier, H. (2010). Criticality analysis of activity networks under interval uncertainty. *Journal of Scheduling*, *13*(6), 609–627. https://doi.org/10.1007/s10951-010-0163-3

Gibson, G. E., & Gebken, R. J. (2003). Design quality in pre-project planning: Applications of the project definition rating index. *Building Research and Information*. https://doi.org/10.1080/0961321032000087990

Gładysz, B., Skorupka, D., Kuchta, D., & Duchaczek, A. (2015). Project Risk time Management - A Proposed Model and a Case Study in the Construction Industry. *Procedia Computer Science*, *64*, 24–31. https://doi.org/10.1016/j.procs.2015.08.459

Glenigan. (2015). *UK Industry Performance Report*. 32. https://doi.org/10.3987/Contents-13-8802

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2020). Machine Learning Algorithms for Construction Projects Delay Risk Prediction. *Journal of Construction Engineering and Management*, *146*(1), 04019085. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001736

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Griffith, A. F. (2006). Scheduling practices and project success. *Cost Engineering (Morgantown, West Virginia)*, *48*(9), 24–30.

Guo, H., Wang, L., & Liu, H. (2006). Analysis and Application of Steel Harden Ability Forecasting Model Based on Support Vector Machine. *2006 6th World Congress on Intelligent Control and Automation*, *2*, 7738–7741. https://doi.org/10.1109/WCICA.2006.1713474

Habash, N., Rambow, O., & Roth, R. (2009). MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*, *41*, 62.

Han, S. H., Kim, D. Y., Kim, H., & Jang, W.-S. (2008). A web-based integrated system for international project risk management. *Automation in Construction*, *17*(3), 342–356. https://doi.org/https://doi.org/10.1016/j.autcon.2007.05.012

HM Treasury. (2014). National Infrastructure Plan 2014. In *October* (Issue December). https://doi.org/08-04-2016

Hola, B., & Schabowicz, K. (2010). Estimation of earthworks execution time cost by means of artificial neural networks. *Automation in Construction*, *19*(5), 570–579. https://doi.org/10.1016/j.autcon.2010.02.004

Honek, K., Azar, E., & Menassa, C. C. (2012). Recession Effects in United States Public Sector Construction Contracting: Focus on the American Recovery and Reinvestment Act of 2009. *Journal of Management in Engineering*, *28*(4), 354–361. https://doi.org/10.1061/(asce)me.1943-5479.0000075

Huang, H.-T., & Tserng, H.-P. (2018). A Study of Integrating Support-Vector-Machine (SVM) Model and Market-based Model in Predicting Taiwan Construction Contractor Default. *KSCE Journal of Civil Engineering*, *22*(12), 4750–4759. https://doi.org/10.1007/s12205-017-2129-x

INEIGHT. (2019). *Ineight - Basis Project Management and Con struction Scheduling*. Ineight.Com.

Infrastructure and Projects Authority. (2016). *National Infrastructure Delivery Plan 2016–2021*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statiscal Learning with Applications in R. In *Springer*. https://doi.org/10.1016/j.peva.2007.06.006

Joachims, T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*.

Jun-yan, L. (2012). Schedule Uncertainty Control: A Literature Review. *Physics Procedia*, *33*, 1842–1848. https://doi.org/10.1016/j.phpro.2012.05.293

Kaming, P. F., Olomolaiye, P. O., Holt, G. D., & Harris, F. C. (1997). Factors influencing construction time and cost overruns on high-rise projects in Indonesia. *Construction Management and Economics*, *15*(1), 83–94. https://doi.org/10.1080/014461997373132

Kelley, J. E. (1961). Critical-Path Planning and Scheduling: Mathematical Basis. *Operations Research*, *9*(3), 296–320. https://doi.org/10.1287/opre.9.3.296

Khodakarami, V., & Abdi, A. (2014). Project cost risk analysis: A Bayesian networks approach for modeling dependencies between cost items. *International Journal of Project Management*, *32*(7), 1233–1245. https://doi.org/10.1016/j.ijproman.2014.01.001

Kirytopoulos, K. A., Leopoulos, V. N., & Diamantas, V. K. (2008). PERT vs. Monte Carlo Simulation along with the suitable distribution effect. *International Journal of Project Organisation and Management*, *1*(1), 24–46.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1), 273–324. https://doi.org/https://doi.org/10.1016/S0004-3702(97)00043-X

Koppenjan, J., Veeneman, W., van der Voort, H., ten Heuvelhof, E., & Leijten, M. (2011). Competing management approaches in large engineering projects: The Dutch RandstadRail project. *International Journal of Project Management*, *29*(6), 740–750. https://doi.org/10.1016/j.ijproman.2010.07.003

Koulinas, G. K., Xanthopoulos, A. S., Tsilipiras, T. T., & Koulouriotis, D. E. (2020). Schedule Delay Risk Analysis

in Construction Projects with a Simulation-Based Expert System. *Buildings*, *10*(8), 134.

Lam, K.-C., Lam, M. C.-K., & Wang, D. (2010). Efficacy of Using Support Vector Machine in a Contractor Prequalification Decision Model. *Journal of Computing in Civil Engineering*, *24*(3), 273–280. https://doi.org/10.1061/(asce)cp.1943-5487.0000030

Luu, V. T., Kim, S. Y., Tuan, N. Van, & Ogunlana, S. O. (2009). Quantifying schedule risk in construction projects using Bayesian belief networks. *International Journal of Project Management*, *27*(1), 39–50. https://doi.org/10.1016/j.ijproman.2008.03.003

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2015). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. https://doi.org/10.3115/v1/p14-5010

Met Office. (2018). *Met Office - UK and regional series*. Metoffice.Gov.Uk.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

Moselhi, O., Gong, D., & El-Rayes, K. (2011). Estimating weather impact on the duration of construction activities. *Canadian Journal of Civil Engineering*, *24*(3), 359–366. https://doi.org/10.1139/l96-122

Mubarak, S. A. (2015). *Construction project scheduling and control*. John Wiley & Sons.

Mulholland, B., & Christian, J. (1999). Risk assessment in construction schedules. *Journal of Construction Engineering and Management*, *125*(1), 8–15. https://doi.org/10.1061/(ASCE)0733-9364(1999)125:1(8)

Nasir, D., McCabe, B., & Hartono, L. (2003). Evaluating Risk in Construction-Schedule Model (ERIC-S): Construction schedule risk model. *Journal of Construction Engineering and Management*, *129*(5), 518–527. https://doi.org/10.1061/(ASCE)0733-9364(2003)129:5(518)

Neapolitan, R. E. (2004). *Learning bayesian networks* (Vol. 38). Pearson Prentice Hall Upper Saddle River, NJ.

nPlan. (2019). *nplan - Predicting the outcomes of construction projects to help you understand complexity and risk*. Nplan.Io.

Ökmen, O., & Öztaş, A. (2008). Construction Project Network Evaluation with Correlated Schedule Risk Analysis Model. *Journal of Construction Engineering and Management*, *134*(1), 49. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:1(49)

ONS. (2019). *Construction output in Great Britain: August 2019*. Office for National Statistics.

Ortiz-González, J. I., Pellicer, E., & Howell, G. (2014). Contingency management in construction projects: A survey of spanish contractors. *22nd Annual Conference of the International Group for Lean Construction: Understanding and Improving Project Based Production, IGLC 2014*, 195–206.

Oxford Economics. (2017). *Global Infrastructure Outlook*.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/v1/d14-1162

Peško, I., Mučenski, V., Šešlija, M., Radović, N., Vujkov, A., Bibić, D., & Krklješ, M. (2017). Estimation of Costs and Durations of Construction of Urban Roads Using ANN and SVM. *Complexity*, *2017*, 1–13. https://doi.org/10.1155/2017/2450370

Rangarajan Sridhar, V. K. (2015). *Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words*. 192–200. https://doi.org/10.3115/v1/w15-1526

Salling, K. B., & Leleur, S. (2015). Accounting for the inaccuracies in demand forecasts and construction cost estimations in transport project evaluation. *Transport Policy*, *38*, 8–18. https://doi.org/10.1016/j.tranpol.2014.11.006

Shen, L. Y. (1997). Project risk management in Hong Kong. *International Journal of Project Management*, *15*(2), 101–105. http://www.sciencedirect.com/science/article/pii/S0263786396000452

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference*, 952–961.

Steyn, J. (2018). *Quantitative risk analysis for projects, Available at https://www.ownerteamconsult.com/quantitative-risk-analysis-for-projects/*.

Su, Y., Lucko, G., & Thompson, R. C. (2016). Evaluating performance of critical chain project management to mitigate delays based on different schedule network complexities. *2016 Winter Simulation Conference (WSC)*, 3314–3324.

Suprapto, M., Bakker, H. L. M., Mooi, H. G., & Hertogh, M. J. C. M. (2016). *How do contract types and incentives matter to project performance?* https://doi.org/10.1016/j.ijproman.2015.08.003

Tokdemir, O. B., Erol, H., & Dikmen, I. (2019). Delay Risk Assessment of Repetitive Construction Projects Using Line-of-Balance Scheduling and Monte Carlo Simulation. *Journal of Construction Engineering and Management*, *145*(2), 04018132. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001595

Tulke, J., Nour, M., & Beucke, K. (2008). A Dynamic Framework for Construction Scheduling based on BIM using IFC. *Proceedings of the 17th IABSE Congress, Chicago, USA*.

Vapnik, V. N., & Vapnik, V. N. (2000). The Vicinal Risk Minimization Principle and the SVMs. In V. N. Vapnik (Ed.), *The Nature of Statistical Learning Theory* (pp. 267–290). Springer New York. https://doi.org/10.1007/978-1-4757-3264-1_9

Venkateswara, H., Lade, P., Lin, B., Ye, J., & Panchanathan, S. (2016). Efficient approximate solutions to mutual information based global feature selection. *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/ICDM.2015.140

Vick, S., & Brilakis, I. (2018). Road Design Layer Detection in Point Cloud Data for Construction Progress Monitoring. *Journal of Computing in Civil Engineering*, *32*(5). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000772

Wang, W.-C. C., & Demsetz, L. A. (2000). Application example for evaluating networks considering correlation. *Journal of Construction Engineering and Management*, *126*(6), 467–474. https://doi.org/10.1061/(ASCE)0733-9364(2000)126:6(467)

Ward, S., & Chapman, C. (2003). Transforming project risk management into project uncertainty management. *International Journal of Project Management*, *21*(2), 97–105. https://doi.org/10.1016/S0263-7863(01)00080-1

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2001). Feature selection for SVMs. *Advances in Neural Information Processing Systems*.

Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. *Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013*, 694–703.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In J. D. H. Fisher (Ed.), *The Fourteenth International Conference on Machine Learning (ICML'97)* (Vol. 97, pp. 412–420).

Zhi, H. (1995). Risk management for overseas construction projects. *International Journal of Project Management*, *13*(4), 231–237. https://doi.org/10.1016/0263-7863(95)00015-I

Zio, E. (2013). The Monte Carlo Simulation Method for System Reliability and Risk Analysis. In *Springer Series in Reliability Engineering*. https://doi.org/10.1007/978-1-4471-4588-2