Front. Eng. Manag. 2019, 6(2): 262–274
https://doi.org/10.1007/s42524-019-0015-6

RESEARCH ARTICLE

Houchen CAO, Yang Miang GOH

Analyzing construction safety through time series methods

Abstract The construction industry produces a large amount of data on a daily basis. However, existing data sets have not been fully exploited in analyzing the safety factors of construction projects. Thus, this work describes how temporal analysis techniques can be applied to improve the safety management of construction data. Various time series (TS) methods were adopted for identifying the leading indicators or predictors of construction accidents. The data set used herein was obtained from a large construction company that is based in Singapore and contains safety inspection scores, accident cases, and project-related data collected from 2008 to 2015. Five projects with complete and sufficient data for temporal analysis were selected from the data set. The filtered data set contained 23 potential leading indicators (predictors or input variables) of accidents (output or dependent variable). TS analyses were used to identify suitable accident predictors for each of the five projects. Subsequently, the selected input variables were used to develop three different TS models for predicting accident occurrences, and the vector error correction model was found to be the best model. It had the lowest root mean squared error value for three of the five projects analyzed. This study provides insights into how construction companies can utilize TS data analysis to identify projects with high risk of accidents.

Keywords time series, temporal, construction safety, leading indicators, accident prevention, forecasting

Received April 4, 2018; accepted October 10, 2018

Houchen CAO (✉)
Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Dr., 117566, Singapore
E-mail: caohouchen@hotmail.com

Yang Miang GOH
Safety and Resilience Research Unit (SaRRU), Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Dr., 117566, Singapore

1 Introduction

Time series (TS) data, commonly known as a sequence of data points, are “typically measured at successive times and spaced at uniform time intervals.” (Hwang et al., 2012). These data points and the corresponding errors can then be fitted into a TS model for forecasting and prediction (Ashuri and Lu, 2010). In the construction industry, data are created and recorded periodically in various aspects. Engineering and building data, such as concrete strength and temperatures, are recorded through sensors. By contrast, management data, such as construction cost, materials, schedule, and most importantly, safety data, are tracked by various personnel, such as project managers (PMs) and quantity surveyors. These TS safety data, which are recorded weekly or monthly, can be utilized for TS analysis.

TS analysis has shown considerable potential in forecasting in economics, in predicting weather in meteorology, and in different applications in the medical sector (Hwang, 2011). Similarly, TS methods have been successfully used in the construction industry. However, most of such applications are related to construction costs and demand. Attention on construction safety TS data are lacking. Construction accidents can lead to human suffering, financial losses, and costs. Construction companies should identify projects with high risk of accident so that proactive interventions can be implemented. Hence, the present study aims to apply various TS methods on construction data to identify projects with high risk of accident. Insights into different forecasting techniques are also discussed.

2 Literature review

2.1 Construction safety

The construction industry is prone to various forms of accidents with high frequency rates (Zhou et al., 2015). This field contributes substantially to the overall global fatality and workplace injury rates (International Labour

Organization, 2018). In 2016, Singapore's Occupational Safety and Health Division reported that the local construction sector alone accounted for 24 out of 66 fatal workplace injuries. In addition, the construction industry is also the highest contributing sector of nonfatal, major injuries. It recorded a staggering 153 out of 594 injuries (26%). However, the fatal injury rate of the construction sector decreased to 4.9 per 100000 employed persons, which is the lowest since 2007. Nonetheless, key findings from Singapore's Ministry of Manpower in 2016 stated ominously that "the construction sector remained the top contributor of workplace fatalities." As such, considerable construction data must be utilized in analyzing and possibly improving the sector's workplace safety and health. Therefore, this study aims to prioritize the use of TS methods in analyzing collected construction data and identifying key risks in construction projects.

2.2 Time series (TS) analysis

TS is a set of continuous data points recorded at equal time intervals. Through the historical values of such data points and their corresponding errors, a TS model can recognize trends, uncover underlying relationships, and generate predictions through extrapolation (Hwang et al., 2012). Statistical TS approaches are divided into two broad classes, namely, univariate and multivariate TS modeling.

TS data must be tested for stationarity before a modeling approach is selected. In a stationary TS, the mean, variance, and autocorrelation of all the observations remain constant over time (Nau, 2017). Brockwell and Davis (2002) summarized the process used in most TS analyses (see Fig. 1).

In Step 1, the available data are evaluated through various statistical tests to identify the presence of trend, seasonality, cyclicity, outliers, and drastic changes. In Step 2, differencing is usually adopted for data transformation to achieve a stationary series. By differencing the series, the differences between consecutive observations are obtained to stabilize the overall mean (Hyndman and Athanasopoulos, 2013). Subsequently, the most suitable model is selected in Step 3, whereby different statistics are employed to determine the accuracy of each model. In Step 4, values are forecasted by the inversion of forecasted residuals. The model should always be updated with the latest data available to improve accuracy of prediction (Hwang, 2011).

2.3 Univariate TS

The first category of TS modeling is known as univariate TS analysis. This modeling approach aims to forecast output values (y_t) using only the historical information contained in the past output values ($y_{t-1},...,y_{t-n}$). Univariate TS models can be expressed as follows:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_n y_{t-n}, \tag{1}$$

where y_t is the single variable being modeled as a TS; $\phi_1, \phi_2,...,\phi_n$ are unknown coefficients to be determined by the chosen TS model; and $y_{t-1}, y_{t-2},...,y_{t-n}$ are the lagged values of y_t (Lam and Oshodi, 2016). The common types of univariate TS models are summarized in Table 1.

2.4 Multivariate TS

Multivariate TS analysis can be applied to investigate the existing interactions and co-movements of a group of univariate TS variables (vector TS). To further conduct forecasting, the term "causal" modeling is often applied to this scenario. Causal modeling is used to identify related variables affecting the forecasting variable; this process allows for the selection of an appropriate stochastic model to describe the interrelationships (Fan et al., 2010) among the variables. Brockwell and Davis (2002) denoted the main difference between multivariate TS and univariate models as follows: "The analysis of multivariate time series $\{X_t\}$ is concerned with not only serial dependence within each component series $\{X_{ti}\}$, but also interdependence between the different component series $\{X_{ti}\}$ and $\{X_{tj}\}$, $i \neq j$."

2.4.1 Stationarity

TS data must be differenced if TS is not stationary. The minimum number of differentiation of a set of TS data needed to achieve stationarity is known as the order of integration of TS (Shahandashti and Ashuri, 2013). A unit root test, such as the augmented Dickey–Fuller (ADF) test, is adopted to check the order of integration. These

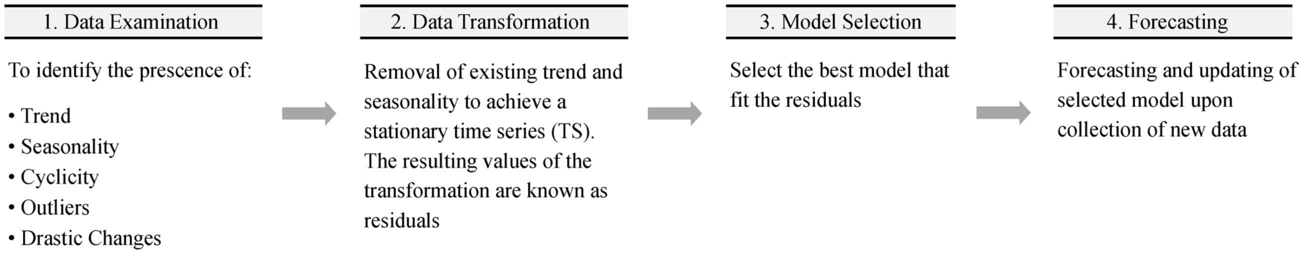

Fig. 1 Time series analysis flowchart

Table 1 Common univariate TS models

Models	Description	Requirements	Type
Moving average (MA)	A method used for smoothing TS; usually a series of arithmetic means (Wohlrabe and Mittnik, 2016).	TS needs to be stationary.	Linear model
Exponential smoothing (ES)	A form of weighted MA whereby the weights decrease exponentially over time (Wohlrabe and Mittnik, 2016). Famous models include Holt–Winters (HW).	TS can be nonstationary.	Can be linear or nonlinear
Autoregression (AR)	Uses correlations of successive lagged observations of itself to forecast values (Wohlrabe and Mittnik, 2016).	TS can be nonstationary, thus resulting in a random walk model.	Linear model
ARMA	Combines AR and MA to generate an accurate predictive stochastic model.	TS needs to be stationary.	Linear model
Box–Jenkins AR Integrated MA (ARIMA)	ARIMA allows for the inclusion of nonstationary TS data in automatic modeling. Here, the "I" of the model stands for integration: the reverse process of differencing data to achieve stationarity (Brockwell and Davis, 2002).	TS can be nonstationary.	Linear model

procedures must be completed before modeling is performed.

2.4.2 Granger causality

The usefulness of a series in forecasting another series must often be checked due to the presence of multiple TS data. The Granger causality test (Granger, 1969) checks whether TS X "causes" TS Y, i.e., whether X is useful in forecasting Y.

2.4.3 Cointegration

Cointegration of variables in multivariate TS is central to the selection of appropriate models for forecasting. Cointegration means that variables do not increasingly drift apart with time and that they may be related in the long run (Shahandashti and Ashuri, 2013). If a stationary linear combination of nonstationary random variables exists, then the set of variables is cointegrated with one another (Buck, 1999). The Johansen test is commonly used to check cointegration; it identifies the number of cointegrating relationships between the groups of TS variables (Shahandashti and Ashuri, 2016).

2.4.4 Selection of multivariate TS models

According to the literature, multivariate TS data can be modeled through three common methods. Figure 2 provides a flowchart of the selection of appropriate multivariate TS models.

Multivariate TS modeling first tests the data for stationarity. The MLR model is used when all data are stationary (i.e., constant mean, variance, and autocorrelation). Otherwise, differencing is performed, followed by the Johansen cointegration test. If no cointegration is

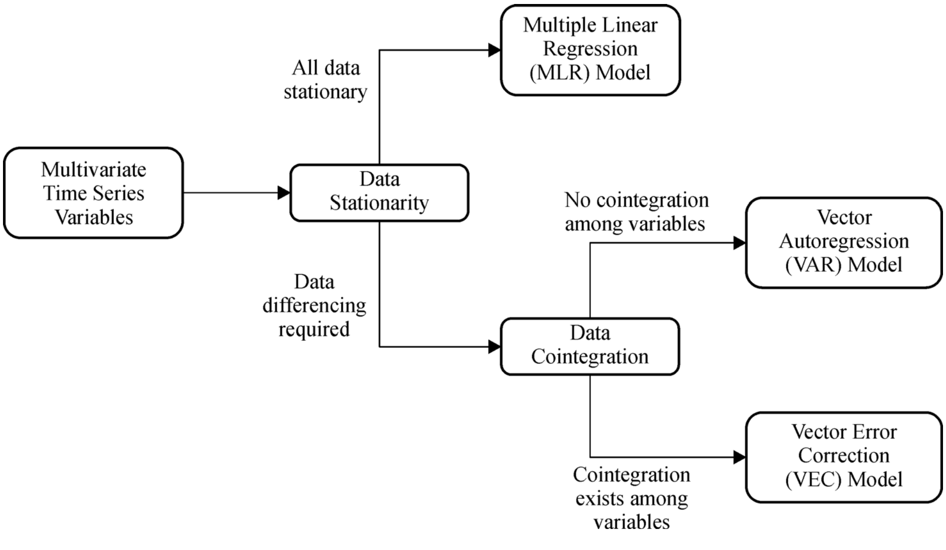

Fig. 2 Multivariate TS model selection flowchart

present, then the VAR model is adopted for TS analysis. However, the VEC model accounts for cointegration among the variables for an accurate analysis.

2.5 Neural networks

Machine learning models have gained popularity in recent years (Ahmed et al., 2010). Artificial neural networks (ANN) have been utilized for analyzing TS data in the construction industry (Schabowicz, 2008; Umit Dikmen and Sonmez, 2010; Chao and Skibniewski, 1995). With sufficient data points, ANN can self-adapt and capture underlying relationships that may not be identified initially. In addition, it can learn from previous experience and update itself accordingly, thereby creating a practical solution to real-world problems (Zhang et al., 1997).

Instead of being a linear statistical model, such as ARIMA, VAR, or VEC, ANNs are nonlinear by nature. In a simple single-hidden-layer feedforward network, the model is deemed a network of three processing units connected by acyclic links. The first is the input layer, where external information, such as multivariate TS data, is received. The last layer is the output layer, where the solution or forecasted data point is obtained. The input and output layers are separated by a pre-specified number of intermediate layers known as hidden layers. ANN in TS forecasting is equivalent to a nonlinear autoregressive (AR) model.

2.6 TS in construction

TS has been applied in different disciplines, such as economics, engineering, medicine, and natural sciences, due to its forecasting capabilities (Hwang, 2011). A database search using specific keywords was conducted on Scopus, ScienceDirect, and Web of Science to gain understanding of TS usage in the construction industry. The keywords included “time series” or “temporal,” “forecasting,” “construction,” and “safety” or “accidents.” The search outcome garnered 24 papers across 10 journals. The search results are summarized in Table 2.

Regarding multivariate TS analysis, VEC modeling was used in nearly 20% of the studies reviewed. By contrast, the Box–Jenkins ARIMA model was the most commonly used technique for univariate analysis. ANN was utilized only once. However, most of the TS analyses were applied to construction topics that are not related to safety. Out of the 24 papers reviewed, only two were focused on construction safety.

The first study was by Lingard et al. (2017) and focused on an analysis of the temporal relationships between the leading and lagging indicators of safety in a large infrastructure project. Through a structured approach, the various indicators of safety were processed and transformed with first differencing and analyzed via VAR modeling. Through VAR and Granger tests, the study identified relationships between the indicators. The second study was by Zhou et al. (2017), and it utilized near-miss accident data for temporal analysis and adopted complex network theory, where the TS of near-misses was analyzed for assistance in risk assessment and safety improvement.

The potential of TS analysis has not been fully utilized for construction safety research. Hence, this study aims to address this research gap.

3 Approach and methods

3.1 Summary of research approach

Figure 3 shows an overview of the research approach for

Table 2 Various TS methods used in current literature

Topic	Univariate Models	Multivariate Models	Others
Safety		• VAR Model (Lingard et al., 2017)	• Complex Network Theory (Zhou et al., 2017)
Construction Productivity	• ARIMA Model (Kim et al., 2015)		
Project Progress	• ES (Batselier and Vanhoucke, 2017)		• Grey Forecasting Model (Lin et al., 2012)
Manpower	• Linear Regression (Bell and Brandenburg, 2003)	• VEC Model (Wong et al., 2011) • VEC Model (Wong et al., 2007) • Multiple Linear Regression (Wong et al., 2008)	• Grey Forecasting Model (Ho, 2010)
Construction Demand	• Box–Jenkins (ARIMA) Model (Fan et al., 2010)	• VAR Model (Sing et al., 2015)	• Grey Forecasting Model (Tan et al., 2015) • Combination of SVM, NNM, and ARIMA models (Lam and Oshodi, 2016)
Construction Costs	• Regression integrated with ARIMA (Ng et al., 2004) • ARIMA Model (Hwang et al., 2012) • Comparative Study between Simple MA (SMA), ARIMA, and HW ES (Ashuri and Lu, 2010)	• VEC Model (Wong and Ng, 2010) • VEC Model (Shahandashti and Ashuri, 2013) • VEC Model (Shahandashti and Ashuri, 2016)	• Data Clustering and Pooling (Yeung and Skitmore, 2012) • Neural Networks (Cao et al., 2015) • Comparative Study using Multiple TS Models (ARMA, VAR, Neural Networks, MA, HW ES) (Hwang, 2011)

this study. EViews 9.5 (IHS Global Inc., 2017) and R (The R Development Core Team, 2018) were used for data processing and modeling. The following subsections discuss each stage in Fig. 3.

3.2 Stage 1: Data cleaning

The construction data set in this study was obtained from the safety division of a private construction company in Singapore (Company X). It contains construction projects ranging from civil works, private redevelopments, and industrial projects to additions and alterations (A&A). The selection of the data set was based on a discussion with Company X's safety manager, consideration of past research, and availability at the time of study. With Company X undertaking projects from the public and private sectors, the current paper analyzed data across various types of works throughout the construction industry. Table 3 summarizes the data collected from Company X.

Data cleaning was conducted via transformation of values and interpolation to ensure high data quality. The cleaning process resulted in five remaining projects with 23 data series each. Table 4 provides information for each remaining project.

With regard to the "project type" and "ownership," these projects presented a good representation of the different types of construction works within the industry. The "contract sum" and "sets of observations available" denoted that these projects were all large-scale works, which consumed multiple years to complete.

Subsequently, Table 5 depicts the pool of 23 potential indicators that may affect construction safety. They are listed in Table 5 as "potential leading indicators" (predictors). Table 5 represents the key output of Stage 1, which was afterward used in Stage 2.

3.3 Stage 2: Data processing

In Stage 2, the five sets of project data underwent statistical tests for identifying suitable predictors. For the following statistical tests and modeling, the final three months of data were removed for an out-of-sample validation test, i.e., out-of-sample data were not used for modeling and were reserved for the evaluation process. This method enabled the acquisition of a fair result for three-month forecasting, which was discussed in the latter stages. A summary of the processing and testing of these sets of data are provided in Fig. 4. To reduce repetition, only the CVL01 data set is described in full detail hereafter.

Stage 1 Data Understanding	**Stage 2** Data Processing	**Stage 3** Modeling	**Stage 4** Evaluation
Extraction of useful data from base dataset	Conduct of statistical tests for processing and selection of final leading indicators of accidents	Modeling of remaining data	Evaluation of models and comparisons
Data Cleaning, Extrapolation and Grouping	General Statistical Tests and TS Statistical Tests	VAR, VEC, ANN and ARIMA	Comparing different models using various selection criterions

Fig. 3 Research design

Table 3 Description of data files

Records	Description	Potential Indicators to be Selected	Attribute Type
Construction Incident Summary	This set of records consists of the number of accidents that occurred each month, with the exact date and nature of injuries.	Number of accidents per month	Ratio
Monthly Project Manager (PM) Inspection Checklist	The monthly inspection checklist is crucial for this research. Each PM inspection checklist contains a set of standard prescribed elements, which contribute directly to the safety of the project (e.g., PPE, overhead protection, falling hazards, and hazardous materials). During each monthly inspection, the PM inspection team gives a score to each element, and a final weighted score that totals the scoring of all elements is generated.	Every element found in the inspection checklist	Interval
Monthly Project Progress	This data set is a record of each project's progress and delay per month for Company X.	Project monthly progress and delay	Ratio
Monthly Project Manpower Statistics	This data set comprises the exact number of man hours on site per month for each project under Company X.	Project man hours per month	Ratio

Table 4 Description of projects selected for modeling

Project Data set Acronym	Project Type	Project Ownership	Contract Sum (S$)	Sets of Observations Available (in Months)
CVL01	Civil works (Train rails)	Public	177.58 M	50
CVL02	Civil works (Train rails)	Public	268.68 M	48
BDG01	Building (Redevelopment)	Private	101.3 M	28
BDG02	Building (Redevelopment)	Private	204.8 M	48
A&A01	Building (Additions & Alterations)	Private	110.5 M	31

Table 5 Finalized pool of potential leading indicators in each set of project data

Number	Potential Leading Indicators	Description
1	Number of Accidents (*Accidents*)	Number of accidents in each project per month
2	Project Progress (*Prog*)	Current project progress, in percentage
3	Project Delay (*Delay*)	Difference between current and planned progress
4	Project Man Hours (*ManHours*)	Amount of man hours input for a project
5	Overall PM Inspection Score (*PMScore*)	Weighted sum of all inspection elements
6	Inspection Element: Personal Protective Equipment (*PPE*)	Wearing of protective equipment (safety harness, belt, lifeline, ear plug, safety boots, face mask, etc.)
7	Inspection Element: Overhead Protection (*OHPro*)	Provision of periphery shelter and exit/entry points; area exposed to falling objects
8	Inspection Element: Excavation Work (*ExWork*)	Provision of shoring systems, warning signage, and barricades along perimeter of excavation
9	Inspection Element: Machine Safety Guarding (*MSG*)	Provision of safety guards at moving parts, blades, and similar parts of machinery
10	Inspection Element: Safe Means of Access (*SafeAccess*)	Access that is free from obstructions and tripping hazards; provision of lighting and directional signage
11	Inspection Element: Operating Crane/Lifting (*Crane*)	Presence of supervisor, signalman, and rigger assist; identification of lifting crew, etc.
12	Inspection Element: Scaffold (*Scaffold*)	Provision of standard erection members, scaffolding access, and load signboard
13	Inspection Element: Tower/Mobile Scaffold (*MobileScaf*)	Similar to above: Scaffold
14	Inspection Element: Mech Elevated Work Platform (*MEWP*)	Provision of working permits; presence of warning signs
15	Inspection Element: Falling Hazard/Opening (*FallHaz*)	Provision of barricades to potential openings on site
16	Inspection Element: Electrical Hazard (*ElecHaz*)	Proper installation of cables; provision of suitable insulation
17	Inspection Element: First-aid Facilities (*FirstAid*)	Provision of first-aid box and room on site
18	Inspection Element: Emergency Preparedness (*EP*)	Provision of firefighting equipment; conduct of emergency drills with maintained records
19	Inspection Element: Handling & Storage of Hazardous Materials (*HazMat*)	Proper disposal of hazardous waste; provision of instructions and manuals
20	Inspection Element: Safe Work Procedure (*SWP*)	Proper procedures for working at height, confined spaces, hot works, etc.
21	Inspection Element: Power Tool Safety (*PTS*)	Proper and safe use of electric tools
22	Inspection Element: Earth Control Measures (*ECM*)	Proper procedures for preventing silt pollution of surrounding environment
23	Inspection Element: Noise, Vector, and Others (*NVO*)	Proper procedures that minimize or prevent excessive noise, vector, etc.

3.3.1 Pearson correlation analysis

Pearson correlation analysis was used to understand the basic linear relationship between the number of accidents and each of the 22 other potential leading indicators. This method measures the direction and strength of association between two continuous variables by creating a line of best fit through the available data. The result of the analysis is a coefficient "*r*," a value ranging from -1 to 1. A value of -1 means a perfect inverse relationship between the two variables, whereas a value of 1 means a perfectly positive linear relationship. A value of 0 shows that no relationship exists between the two variables (Lund A and Lund M, 2013).

In this analysis, the null hypothesis stated that "Accidents" and the 22 other potential leading indicators (PPE, OHPro, ExWork, etc.) were not associated. Rejection of the null hypothesis was at the 5% significance level.

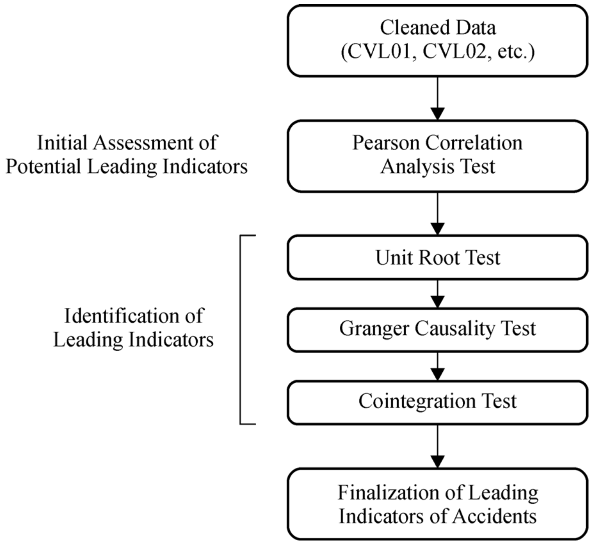

Fig. 4 Summary of Stage 2 data processing

Thus, correlation between "Accidents" and the other indicators was significant at the 10% level. The results of the correlation test for the CVL01 data set are listed in Table 6.

Table 6 CVL01 correlation tests between "accidents" and other indicators

Indicators	Coefficient "*r*"	Test Statistic	p-value
Accidents, Prog	0.439	3.3864	0.001421*
Accidents, Delay	0.444	3.4302	0.001249*
Accidents, ManHours	0.356	2.6418	0.0111*
Accidents, PMScore	−0.196	−1.3816	0.1735
Accidents, PPE	−0.00985	−0.06827	0.9458
Accidents, OHPro	0.0910	0.63028	0.5315
Accidents, ExWork	0.147	1.0285	0.3089
Accidents, MSG	0.221	1.573	0.1223
Accidents, SafeAccess	−0.095	−0.664	0.5101
Accidents, Crane	0.0440	0.3055	0.7613
Accidents, Scaffold	−0.352	−2.6095	0.01206*
Accidents, MobileScaf	−0.332	−2.4387	0.01849*
Accidents, MEWP	−0.0715	−0.49669	0.6217
Accidents, FallHaz	−0.06929	−0.48121	0.6326
Accidents, ElecHaz	−0.2308	−1.6439	0.1067
Accidents, FirstAid	0.2016	1.4264	0.1602
Accidents, EP	−0.06564	−0.45578	0.6506
Accidents, HazMat	−0.01506	−0.1044	0.9173
Accidents, SWP	0.1242	0.86748	0.39
Accidents, PTS	0.1724	1.2127	0.2312
Accidents, ECM	0.098	0.68226	0.4984
Accidents, NVO	0.10199	0.71034	0.4809

Note: * Indicators are correlated.

As observed in Table 6, many of the potential leading indicators were not stationary (e.g., SafeAccess had a p-value of 0.5101; thus, the null hypothesis could not be rejected.). Therefore, all the data points were differenced once, and the ADF test was repeated. The differencing of data calculated the differences between consecutive observations, thereby eliminating trend and seasonality.

After the differencing of all the indicators, all indicators except Prog, Delay, and ManHours became stationary. Hence, the next statistical test, Granger causality, could be applied.

3.3.2 Granger causality test

Proposed by Clive Granger in 1969, the Granger causality test is commonly used to check whether one TS is useful for predicting the next. The null hypothesis of the test states that "*X*" does not Granger-cause "*Y*." Hence, rejecting this null hypothesis denotes that "*X*" is indeed useful in forecasting "*Y*." In addition, "lag" in TS denotes the past period observations of the data set. A lag length of 1 indicates the usage of only the past one month of observations. As this test is lag-specific, lag lengths of two, four, and six months were chosen for demonstrating the relationships between potential leading indicators "*X*" and "Accidents" over different time periods. The results of the test on the stationary potential leading indicators of CVL01 are presented in Table 7.

To interpret the results in Table 7, an example would be Δ *Accidents* & Δ *SafeAccess*. As the p-value at the four lags was less than 0.05, the null hypothesis was rejected. A comparison of the four previous months of SafeAccess and Accidents data statistically indicated that SafeAccess Granger-caused Accidents.

In summary, for the project CVL01, only OHPro, SafeAccess, and HazMat were statistically significant in forecasting accidents, according to the Granger causality test. These indicators are representations of how well safety and health management are being implemented on site. While they do not definitively cause accidents directly, they act as proxies in that failure in managing them signify poor management that contributes to accident occurrences. The Granger causality test supports that the shortlisted variables are predictive of accidents in CVL01.

3.3.3 Cointegration test

The final statistical test for this research was Johansen's cointegration test. Apart from double-checking whether the leading indicators had a certain relationship, this test was critical for final model selection. If cointegration existed among the indicators, then the vector error

Table 7 CVL01 Granger causality test against accidents

Indicators Tested with Δ Accidents	**Lag Length**		
	2	**4**	**6**
Δ *Accidents* & Δ *PMScore*	0.2515	0.6246	0.9027
Δ *Accidents* & Δ *PPE*	0.4134	0.8031	0.6931
Δ *Accidents* & Δ *OHPro*	0.003189*	0.01228*	0.02892*
Δ *Accidents* & Δ *ExWork*	0.3697	0.5058	0.8327
Δ *Accidents* & Δ *MSG*	0.6334	0.7974	0.7699
Δ *Accidents* & Δ *SafeAccess*	0.8179	0.004176*	0.02366*
Δ *Accidents* & Δ *Crane*	0.7892	0.1742	0.2549
Δ *Accidents* & Δ *Scaffold*	0.09519	0.1863	0.1918
Δ *Accidents* & Δ *MobileScaf*	0.129	0.2786	0.3819
Δ *Accidents* & Δ *MEWP*	0.9037	0.8585	0.6584
Δ *Accidents* & Δ *FallHaz*	0.1371	0.3161	0.6377
Δ *Accidents* & Δ *ElecHaz*	0.2972	0.1994	0.634
Δ *Accidents* & Δ *FirstAid*	0.3255	0.4529	0.2639
Δ *Accidents* & Δ *EP*	0.1301	0.1069	0.2292
Δ *Accidents* & Δ *HazMat*	0.14	0.0748	0.005423*
Δ *Accidents* & Δ *SWP*	0.1307	0.1619	0.1267
Δ *Accidents* & Δ *PTS*	0.04684	0.2966	0.3721
Δ *Accidents* & Δ *ECM*	0.9679	0.8669	0.6728
Δ *Accidents* & Δ *NVO*	0.2237	0.4701	0.7294

Note: * Rejection of null hypothesis at 5% confidence level (i.e., Granger-causing Accidents).

correction (VEC) model would be utilized for forecasting (Pfaff, 2008).

The null hypothesis of the Johansen's cointegration test states that the number of cointegrating relationships r present among the indicators are less than or equal to the number of indicators present. For example, if $r \leqslant 0$ is rejected for the vector of leading indicators, then at least one cointegrating relationship exists among the indicators. Similarly, if $r \leqslant 1$ is rejected, then at least two cointegrating relationships exists among the indicators. The CVL01 test results for Accidents with all the selected predictors, Accidents with OHPro, Accidents with SafeAccess, and Accidents with HazMat are presented in Table 8.

According to Table 8 Test A, the null hypothesis (number of cointegrating relationships "r") was rejected at the 5% confidence level up to at most three "r." Thus, four cointegrating relationships existed; i.e., *Accidents*, *OHPro*, *SafeAccess*, and *HazMat* were cointegrated. Similarly, for Tests B to D, the number of cointegrating relationships denoted that all the potential leading indicators were cointegrated in their own respective tests. In consideration of the cointegration present among all the indicators, such results once again supported the fact that *OHPro*, *SafeAccess*, and *HazMat* were the finalized leading indicators of accidents in CVL01.

3.4 Stage 3: Modeling

In Stage 3, the finalized set of predictors was used in the development of the following types of statistical models:

1) VEC model;
2) ARIMA Model;
3) ANN Model.

VEC or VAR is a type of multivariate TS model, ARIMA is a univariate TS technique, and ANN is a commonly used machine learning forecasting method. Root mean squared error (RMSE) was used as the basis for comparison across the models.

3.4.1 VEC modeling

The presence of cointegration (shown in Stage 2) determined the use of the VEC model. A VEC model is a regression-based statistical modeling technique for multivariate TS data. In a slightly modified version of the univariate equation seen in Section 2.2, the multivariate vector equation is deemed as follows:

$$\Delta y_t = \sum_{i=1}^{p-1} \Phi_i \Delta y_{t-i} + \prod y_{t-p} + C + \varepsilon_t.$$

The structure of the equation is as follows. y_t is a $(N*1)$ TS vector data at period t, whereby N = number of variables. Δ stands for first difference operator (i.e., y_t variables are differenced once.); ϕ_i $(i = 1,...,p-1)$ are $(N*N)$ coefficient matrices of endogenous variables; Ð is $(N*N)$ coefficient matrix; C is $(N*1)$ constant vector; and ε_t is $(N*1)$ error term vector (Shahandashti and Ashuri, 2016).

The VEC model included Accidents, OHPro, SafeAccess, and HazMat data from the period of November 2011–September 2015, with the remaining data (October 2015–December 2015) used for out-of-sample predictability testing. Initial vector autoregression tests suggested that the optimal lag length of three months must be used for VEC modeling, the results of which are shown in Table 9.

According to the VEC model's R-squared value, 70% of the variations in accidents per month could be explained by the three previous months' accident rate, overhead protection score, safe access score, and hazardous materials score. RMSE measures the error between two data sets when comparing a predicted value and an observed known value. Thus, a small RMSE value denotes that the predicted and observed values are close.

3.4.2 ARIMA modeling

ARIMA is a popular univariate TS model that combines the AR and moving average (MA) techniques. Created by Box and Jenkins (1976), the ARIMA model is also known as the Box–Jenkins approach. The three-step process starts

Table 8 CVL01 Cointegration test results of various indicators

Test A: Results for *Accidents* and *OHPro*, *SafeAccess*, *HazMat*

Number of Cointegrating Relationships *r*	Trace Statistic	0.05 Critical Value	Probability
None	65.77	47.86	0.0005*
At most 1	37.01	29.80	0.0062*
At most 2	21.73	15.49	0.0050*
At most 3	9.52	3.84	0.0020*

Test B: Results for *Accidents* and *OHPro*

Number of Cointegrating Relationships *r*	Trace Statistic	0.05 Critical Value	Probability
None	22.30	15.50	0.0040*
At most 1	8.06	3.84	0.0045*

Test C: Results for *Accidents* and *SafeAccess*

Number of Cointegrating Relationships *r*	Trace Statistic	0.05 Critical Value	Probability
None	17.54	15.49	0.0243*
At most 1	4.80	3.84	0.0284*

Test D: Results for *Accidents* and *HazMat*

Number of Cointegrating Relationships *r*	Trace Statistic	0.05 Critical Value	Probability
None	16.74	15.49	0.0323*
At most 1	4.96	3.84	0.0260*

Note: *Rejection of null hypothesis at 5% confidence level.

Table 9 Regression and forecast values for VEC model in CVL01

Statistics	Values
R-squared Value	0.70
Forecast RMSE	1.583

with the checking of the autocorrelation function (ACF) and partial ACF (PACF) of the given univariate TS data. TS data are subsequently differenced until they become stationary. Finally, the ACF and PACF plots with Akaike's information criterion are used to identify the most suitable model to be used for forecasting (Lam and Oshodi, 2016).

As ARIMA is a univariate model, only Accidents data themselves are utilized for forecasting. The results of the ARIMA test are depicted in Table 10.

Table 10 ARIMA forecast statistics of CVL01

Statistics	Values
Final Model	ARIMA (0, 1, 1)
ARIMA Forecast RMSEs	1.587

3.4.3 ANN modeling

As described in the Literature review section, a basic feedforward ANN model is used to predict accidents. The statistics in the ANN model are shown Table 11, whereas a pictorial representation of the neural net is presented in Fig. 5.

Table 11 ANN forecast statistics of CVL01

Statistics	Values
Hidden Layers	2 layers, 10 nodes – 2 nodes
Forecast RMSE	0.6858

4 Results and discussion

4.1 Overall leading indicator findings

As observed in the previous section, *Accidents, OHPro, SafeAccess,* and *HazMat* were chosen via the TS process as the finalized leading indicators for project CVL01. The chosen leading indicators for the all five project data sets are listed in Table 12.

In order of frequency of selection, the predictor *Accidents* was selected for all five projects. As the data used were all tagged with time, previous months' number of accident cases contributed largely to future occurrences.

The variable *MobileScaf* (Tower/Mobile Scaffold) was chosen by the TS process for three out of five projects. With a 60% appearance rate, *MobileScaf* should not be classified as an isolated leading indicator for any specific project. However, it may demonstrate an underlying

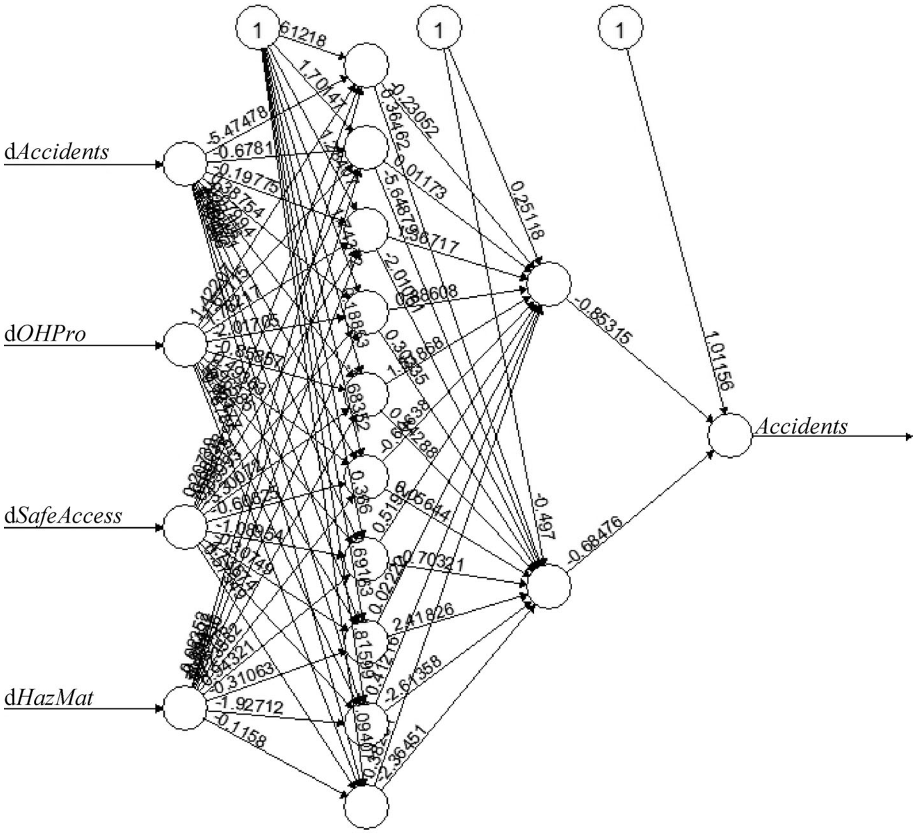

Fig. 5 ANN representation

Table 12 Leading indicators of each project

Project	Selected Input Variables
CVL01 (Train rails – Civil)	*Accidents*, *OHPro*, *SafeAccess*, *HazMat*
CVL02 (Train rails – Civil)	*Accidents*, *PMScore*, *MobileScaf*, *EP*
BDG01 (Redevelopment – Building)	*Accidents*, *Crane*, *MobileScaf*, *SWP*
BDG02 (Redevelopment – Building)	*Accidents*, *PMScore*, *MobileScaf*, *EP*
A&A01 (A&A – Building)	*Accidents*, *SafeAccess*

company-wide issue regarding safety and inspection of tower/mobile scaffolding.

Predictors *PMScore*, *SafeAccess,* and *EP* were selected twice each. While these indicators were not present in all projects, they are important indicators of how well the site is being managed.

Finally, *SWP* and *Crane* were selected only once as predictors for BDG01, whereas *OHPro* and *HazMat* were selected only for CVL01. Thus, these were isolated cases for the respective projects. Nevertheless, non-compliance with any of the above leading indicators (*SWP, Crane, OHPro, HazMat*) will result in serious injuries because these activities are commonplace across the construction sites.

4.2 Forecast model selection and comparison

The comparison of the forecasting models for a number of construction accidents for each project is presented in Table 13. Similarly, RMSE was used as the determining factor for model forecast accuracy.

Using RMSE as the main determinant of predictive accuracy, the VEC model was deemed to be the most accurate in forecasting the construction accidents. By contrast, the ARIMA and ANN models were selected only once each for projects BDG01 and CVL01, respectively. This finding was in conjunction with the results of "Highway construction cost forecasting using vector error correction models" by Shahandashti and Ashuri (2016), whereby VEC models were also found to be the most precise among univariate TS approaches.

Despite being the most robust among the three TS approaches, the VEC models still possessed a relatively high RMSE value in general. Thus, the predictions made should be regarded only as a reference for a project's overall construction safety. However, all the VEC models established a high R-squared value of at least 0.66. Therefore, at least 66% of the variations were explained by the VEC model.

As the only univariate model out of the three, the

Table 13 Forecasting model statistics for each project

Project	VEC Model RMSE/R-squared Value	ARIMA RMSE	ANN RMSE
CVL01 (Train rails –Civil)	1.583/0.70	1.587	0.6858*
CVL02 (Train rails –Civil)	1.355/0.66*	1.717	1.637
BDG01 (Redevelopment –Building)	1.428/0.75	0.9110*	1.483
BDG02 (Redevelopment – Building)	0.9785/0.66*	1.717	1.261
A&A01 (A&A – Building)	0.4710/0.74*	0.9724	1.862

Note: *Model with highest predictive accuracy (lowest RMSE).

ARIMA model considers only past accidents to predict future accident occurrences. As such, it is the least accurate among the models. It does not consider any other variable or indicator in complex construction projects. Furthermore, the PM cannot benefit from any information produced through the ARIMA process, unlike its VEC counterpart. Interestingly, the ARIMA model presented the most accurate prediction for project BDG01 notwithstanding its limitations.

The ANN model was the only machine learning approach adopted in this research. While the model is generally used as a classification technique, this study attempted to feed multiple TS data into its input nodes to forecast a single output, i.e., future number of accidents. Although ANN was selected as the model with best accuracy only once, CVL01 was the project with the most number of observations out of the entire data given by Company X. Thus, CVL01 ANN's RMSE value was lower than its VEC and ARIMA counterparts by nearly 0.9.

4.3 Comparison with similar research

A similar study using the same data set was conducted by Poh et al. (2018) using machine learning classification techniques. The following compares the TS methods with the machine learning classification approach by Poh et al. (2018).

TS techniques take full advantage of the temporal relationship in data sets to study underlying factors or generate forecasts. As such, the data used need to be running in sequential order according to time. In the current work, the data used for analysis were grouped by projects according to time, thereby resulting in approximately 30 to 40 observations points per project. Per contra, machine learning classification techniques are not bound by such rules. No time or date sequence is required given that the training and testing of data do not consider the temporal relationships between the variables. Hence, a number of observation points can be used for forecasting, but such an approach may miss any underlying relationship related to time. In their study, Poh et al. (2018) combined all projects of Company X and employed five different classification techniques to predict all possible future accident occurrences. With more than 300 observations, the machine learning methods of Poh et al. (2018) underwent greater training and consequently produced much better predictions than the method in the current work. Furthermore, their techniques can classify accidents according to severity (major or minor), a factor which cannot be determined by the TS regression models in the present research.

The predictors identified may not be applicable to every project at different points in time. Different projects have varied characteristics that change with time as construction work progresses. In this study, project-specific TS data were used. Thus, the TS approach should be evaluated further with a larger data set than that in this work.

5 Conclusions

This work aimed to analyze construction safety through the use of TS approaches. Through the proposed research methodology, the four-stage process can assist construction companies in forecasting the number of accidents for each project. The input variable *Mobile/ Tower Scaffold* was selected for three of the five projects, whereas *PM Safety Inspection Score*, *Safe Means of Access,* and *Emergency Procedures* were selected for two projects. In addition, the VEC model was identified as the most promising TS model for forecasting construction accidents in terms of RMSE value. However, ARIMA and ANN also produced good results, albeit for only one project each.

Using the TS approach developed herein, a PM can forecast the number of accidents for a project so that interventions can be developed to reduce the likelihood of accidents. In addition, variables such as *MobileScaf* were found to be important predictors for the TS models. Hence, they can be used as a leading indicator of poor safety management. Senior management can also use TS forecasting to allocate additional attention and resources to projects forecasted to have a high number of accidents.

Although the machine learning methods used by Poh et al. (2018) can classify accident severity and occurrence, the TS techniques proposed in this paper provide project-specific TS models that PMs can use to manage their projects proactively. A combination of the TS methods and machine learning methods can be evaluated in future research.

References

Ahmed N K, Atiya A F, Gayar N E, El-Shishiny H (2010). An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29(5–6): 594–621

Ashuri B, Lu J (2010). Time series analysis of ENR construction cost index. Journal of Construction Engineering and Management, 136 (11): 1227–1237

Batselier J, Vanhoucke M (2017). Improving project forecast accuracy by integrating earned value management with exponential smoothing and reference class forecasting. International Journal of Project Management, 35(1): 28–43

Bell L C, Brandenburg S G (2003). Forecasting construction staffing for transportation agencies. Journal of Management Engineering, 19(3): 116–120

Box G E P, Jenkins G M (1976). Time Series Analysis Forecasting and Control. California: Prentice-Hall International, Inc.

Brockwell P J, Davis R A (2002). Introduction to Time Series Forecasting. New York: Springer

Buck A J (1999). Cointegration and error correction. http://www.eco.uc3m.es/~jgonzalo/teaching/EconometriaII/cointegration.htm

Cao M, Cheng M, Wu Y (2015). Hybrid computational model for forecasting taiwan construction cost index. Journal of Construction Engineering and Management, 141(4): 04014089

Chao L C, Skibniewski M J (1995). Neural network method of estimating construction technology acceptability. Journal of Construction Engineering and Management, 121(1): 130–142

Fan R Y C, Ng S T, Wong J M W (2010). Reliability of the Box–Jenkins model for forecasting construction demand covering times of economic austerity. Construction Management and Economics, 28 (3): 241–254

Granger C W J (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3): 424–438

Ho P H K (2010). Forecasting construction manpower demand by gray model. Journal of Construction Engineering and Management, 136 (12): 1299–1305

Hwang S (2011). Time series models for forecasting construction costs using time series indexes. Journal of Construction Engineering and Management, 137(9): 656–662

Hwang S, Park M, Lee H, Kim H (2012). Automated time-series cost forecasting system for construction materials. Journal of Construction Engineering and Management, 138(11): 1259–1269

Hyndman R J, Athanasopoulos G (2013). Forecasting: Principles and Practice. https://otexts.com/ppp2/index.html

HS Global Inc. (2017). EViews 9.5 Student Version Lite. California, USA: IHS Inc.

International Labour Organization (2018). The enormous burden of poor working conditions. http://www.ilo.org/moscow/areas-of-work/occupational-safety-and-health/WCMS_249278/lang–en/index.htm

Kim H, Lee H S, Park M, Ahn C R, Hwang S (2015). Productivity forecasting of newly added workers based on time-series analysis and site learning. Journal of Construction Engineering and Management, 141(9): 05015008

Lam K C, Oshodi O S (2016). Using univariate models for construction output forecasting: Comparing artificial intelligence and econometric techniques. Journal of Management Engineering, 32(6): 04016021

Lin C M, Tserng P H, Ho P S, Young D L (2012). A novel dynamic progress forecasting approach for construction projects. Expert Systems with Applications, 39(3): 2247–2255

Lingard H, Hallowell M, Salas R, Pirzadeh P (2017). Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project. Safety Science, 91: 206–220

Lund A, Lund M (2013). Pearson's correlation using stata. https://statistics.laerd.com/stata-tutorials/pearsons-correlation-using-stata.php

Nau R (2017). Statistical forecasting: Notes on regression and time series analysis. https://people.duke.edu/~rnau/411diff.htm

Ng S T, Cheung S O, Skitmore M, Wong T C Y (2004). An integrated regression analysis and time series model for construction tender price index forecasting. Construction Management and Economics, 22(5): 483–493

Pfaff B (2008). Analysis of Integrated and Cointegrated Time Series with R. 2nd ed. New York: Springer

Poh C Q X, Ubeynarayana C U, Goh Y M (2018). Developing safety leading indicators for construction sites: A machine learning approach. Automation in Construction

Schabowicz B H K, Hoła B (2008). Application of artificial neural networks in predicting earthmoving machinery effectiveness ratios. Archives of Civil and Mechanical Engineering, 8(4): 73–84

Shahandashti S M, Ashuri B (2013). Forecasting engineering news-record construction cost index using multivariate time series models. Journal of Construction Engineering and Management, 139(9): 1237–1243

Shahandashti S M, Ashuri B (2016). Highway construction cost forecasting using vector error correction models. Journal of Management Engineering, 32(2): 04015040

Sing M C P, Edwards D J, Liu H J X, Love P E D (2015). Forecasting private-sector construction works: VAR model using economic indicators. Journal of Construction Engineering and Management, 141(11): 04015037

Tan Y, Langston C, Wu M, Ochoa J J (2015). Grey forecasting of construction demand in Hong Kong over the next ten years. International Journal of Construction Management, 15(3): 219–228

The R Development Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing

Umit Dikmen S, Sonmez M (2010). An artificial neural networks model for the estimation of formwork labour. Journal of Civil Engineering and Management, 17(3): 340–347

Wohlrabe K, Mittnik S (2016). Univariate time series analysis. http://www.finmetrics.statistik.uni-muenchen.de/studium_lehre/sommersemester-2016/tsa_16/univariate_ts_script_ss2016.pdf

Wong J M W, Chan A P C, Chiang Y H (2007). Forecasting construction manpower demand: A vector error correction model. Building and Environment, 42(8): 3030–3041

Wong J M W, Chan A P C, Chiang Y H (2008). Modeling and forecasting construction labor demand multivariate analysis. Journal of Construction Engineering and Management, 134(9): 664–672

Wong J M W, Chan A P C, Chiang Y H (2011). Construction manpower demand forecasting. Engineering, Construction, and Architectural

Management, 18(1): 7–29

Wong J M W, Ng S T (2010). Forecasting construction tender price index in Hong Kong using vector error correction model. Construction Management and Economics, 28(12): 1255–1268

Yeung D, Skitmore M (2012). A method for systematically pooling data in very early stage construction price forecasting. Construction Management and Economics, 30(11): 929–939

Zhang G Q, Patuwo B E, Hu M Y (1997). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14: 35–62

Zhou C, Ding L, Skibniewski M J, Luo H, Jiang S (2017). Characterizing time series of near-miss accidents in metro construction via complex network theory. Safety Science, 98: 145–158

Zhou Z, Goh Y M, Li Q (2015). Overview and analysis of safety management studies in the construction industry. Safety Science, 72: 337–350