Pertanika J. Sci. & Technol. 32 (3): 1103 - 1122 (2024)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

An Accident Prediction Model Based on ARIMA in Kuala Lumpur, Malaysia, Using Time Series of Actual Accidents and Related Data

Boon Chong Choo[1], Musab Abdul Razak[1]*, Mohd Zahirasri Mohd Tohir[1], Dayang Radiah Awang Biak[1] and Syafiie Syam[2]

[1]Safety Engineering Interest Group, Department of Chemical & Environmental Engineering, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia
[2]Department of Chemical and Materials Engineering, Faculty of Engineering Rabigh Brang, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

ABSTRACT

Recently, there has been an emerging trend to analyse time series data and utilise sophisticated tools for optimally fitting time series models. To date, Malaysian industrial accident data is underutilised and lacks informative records. Thus, this paper aims to investigate the Malaysian accident database and further evaluate the optimal forecasting models in accident prediction. The model's input was based on available data from the Department of Occupational Safety and Health, Malaysia (DOSH), from 2018 until 2021, with 80% of the dataset to train the models and the remaining 20% for validation. The negative binomial and Poisson distribution prediction showed a mean absolute percentage error (MAPE) of 33% and 51%, respectively. It indicated that the negative binomial performed better than the Poisson distribution in accident frequency prediction. The available time series accident data were gathered for four years, and stationarity was checked in R Studio software for the Augmented Dickey-Fuller test. The lowest Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and other error values were used to justify the best model, which was the ARIMA(2,0,2)(2,0,0)(12) model. The ARIMA models were considered after the data showed autocorrelation. The MAPE for both ARIMA in R and manual time series were 40% and 49%, respectively. Therefore, the accident prediction by using R Studio would outperform the manually negative binomial and Poisson distribution.

ARTICLE INFO

Article history:
Received: 18 July 2023
Accepted: 06 December 2023
Published: 01 April 2024

DOI: https://doi.org/10.47836/pjst.32.3.07

E-mail addresses:
boonchong93@gmail.com (Boon Chong Choo)
musab@upm.edu.my (Musab Abdul Razak)
zahirasri@upm.edu.my (Mohd Zahirasri Mohd Tohir)
dradiah@upm.edu.my (Dayang Radiah Awang Biak)
ssmahmoud@kau.edu.sa (Syafiie Syam)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

Based on the findings, industrial safety practitioners should report accidents to DOSH truthfully in the era of digitalisation. It could enable future data-driven accident predictions to be carried out.

Keywords: Accident models, accident prediction, digitalisation, Malaysia's accidents, R studio

INTRODUCTION

The rapid growth of industrialisation and the global economy in developing countries like Malaysia has constantly led to industrial accidents, which have emerged as a social problem (Kim et al., 2021). It had been reported by Kim et al. (2021) that the Asian occupational fatality per 100,000 workers was higher than in EU countries, and Malaysia required more effective safety regulations and programmes. There is a legal requirement in Malaysia under the Occupational Safety and Health Act (OSHA) 1994 for accident reporting to the Department of Occupational Safety and Health (DOSH) Malaysia via JKKP 6 or myKKP website. In view of the statistical field and machine learning, such accident reporting can provide continuous analysis and learning processes to prevent unwanted occurrences (Freivalds & Johnson, 1990). Click or tap here to enter text.On accident cost estimation, Rohani et al. (2015) found that the ratio of accident prevention to accident cost is 1:19.6 in Malaysia. However, Kidam et al. (2015) and Choo et al. (2022) highlighted that weakness in Malaysian accident reporting led to poor learning.

The statistical data and analysis of the accident database would be more reliable in the research on accident prevention (Chong & Low, 2014). For example, Abdullah and Wern (2011) studied the accident frequency and revealed the factors of high levels of injuries and fatalities encountered in the construction industry based on accident statistics. Ayob et al. (2018) conducted a descriptive study through a survey to identify the cause (poor risk management) and accident agent (fall from height) in the construction industry. Chong and Low (2014), through statistical data and court cases in the period of 2000-2009, identified and tabulated the causes that contributed to health issues, and the reported main cause of construction accidents were striking objects and falls. Hadi et al. (2017) conducted a survey that found that 94.7% of the workers did not report any accidents to their management and revealed a prevalence of non-reporting accidents in construction sites. As a result, the safety officer in the company may be unaware of the near-miss that happened and not record it in the safety system. Apart from the construction industry, Ali et al. (2017) studied the trend of accidents in the manufacturing industry using descriptive data and found that the number of fatalities and permanent and non-permanent disability increased by 26%, 71% and 64%, respectively. Zein et al. (2015) completed a survey on working postures, revealing the most prominent work involving bending forward and lifting heavy loads, which showed the most significant physical body injury.

In accident prediction, Rohayu et al. (2012) predicted the road accident fatalities for 2020 using the ARIMA model, and the data showed autocorrelation. Manan et al. (2013) reported the first motorcycle accident prediction model in Malaysia using the negative binomial regression model. Malaysia has actively conducted road transportation safety research, but to our knowledge, no industrial accident prediction has been reported in Malaysia. Choo et al. (2022) conducted a literature review on supervised machine learning, and the concept of accident prediction is applied in this paper. Thus, this paper aims to utilise the Malaysian accident database and fit the data for modelling, namely Poisson and negative binomial distribution, for frequency modelling. The time series prediction by using R Studio was also evaluated. The findings of this study can set a foundation for industrial accident prediction in Malaysia.

METHODOLOGY

This study utilised accident data obtained from DOSH in Kuala Lumpur, Malaysia. A total of 1131 industrial accidents reported to DOSH from January 2018 to December 2021 were used in this study. As Zermane et al. (2022) highlighted, the accident data were incomplete, with fewer details, and repetitive with unclear descriptions. Thus, the data were screened pre-processed by removing invalid data (Hajakbari & Minaei-Bidgoli, 2014), resulting in 1047 accident data for this study. The incomplete, redundant, and invalid data, such as non-word text-type data, was excluded from the table. The number of days lost was created from the injuries suffered. For modelling purposes, 80% of the dataset was used in training, whereas the remaining 20% was used as validation.

Frequency Modelling

The number of accidents can be described as the statistical safety indicators (Jian, 2021) and applied in prediction (Attwood et al., 2006). In this research, the primary variable for frequency modelling was the time elapsed between the date of the latest accident and the previous one (Hajakbari & Minaei-Bidgoli, 2014; Esmaili et al., 2021). The frequency distribution was selected based on the relationship between the mean (Attwood et al., 2006) and the variance of annual incidents. Several researchers demonstrated a constant failure rate and assumed no safety-related changes were made (Attwood et al., 2006); therefore, two distributions, Poisson and negative binomial distributions, were used in this study as below.

Poisson distribution as expressed in Equation 1 (If mean and variance of the data are in closed proximity) (Attwood et al., 2006; Ismail & Zamani, 2013; Manan et al., 2013)

$$y \sim p\,(y = y_i) = \left\{\frac{\lambda^{y_i} e^{-\lambda}}{y_i!}\right\},\ y_i \in \{i^1\},\ y_i \geq 0,\ \lambda > 0, \quad [1]$$

where y_i is the number of incidents in the year i, and λ is the annual average number of incidents, with the expected value, E(y), and variance, V(y), equal to λ.

The prior distribution for λ is assumed to follow Gamma-distribution, $\lambda \sim (\alpha, \beta)$ due to uncertainty (Meel & Seider, 2006; Meel et al., 2007) as expressed in Equation 2:

$$p(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}, \alpha > 0, \beta > 0 \quad [2]$$

From Baye's theorem, the posterior distribution, which permits a projection of accident frequency in the future (Meel & Seider, 2006), p (λ | Data), is expressed in Equation 3:

$$p(\lambda \mid \text{Data}) \propto \ell(\text{Data} \mid \lambda)\, p(\lambda)$$

$$\propto (\lambda^S e^{-N_t\lambda})(\lambda^{\alpha-1} e^{-\beta\lambda}) \propto \lambda^{(\alpha+S)-1} e^{-(\beta+N_t)\lambda}, \quad [3]$$

where Data = $(y_0, y_1, \ldots, y_{N_t})$, s = $\sum_{i=0}^{N_t} y_i$, N_t is the number of years, and ℓ (Data | λ) is the Poisson likelihood distribution. Note that p (λ | Data) is also a Gamma distribution, Gamma $(\alpha + S,\ \beta + N_t)$, because λ is distributed according to Gamma (α, β), which is conjugate prior to the Poisson distribution. The mean of the posterior distribution is the weighted average of the means of the prior and likelihood distributions, as expressed in Equation 4:

$$\frac{\alpha+s}{\beta+N_t} = \frac{\beta}{\beta+N_t}\left(\frac{\alpha}{\beta}\right) + \frac{N_t}{\beta+N_t}\frac{s}{N_t}, \quad [4]$$

and the variance of the posterior distribution is $\frac{\alpha+s}{(\beta+N_t)^2}$.

Negative binomial distribution (Poisson-gamma) (Thakali et al., 2016) is used when the Poisson distribution is poor. Denham (2020) and Warner (2015) reported that when the variance exceeds its mean, the data are considered over-dispersed and need a different model instead of Poisson distribution. The number of incidents that occurred in a year is a non-negative and integer-valued result that can be estimated using a negative binomial distribution in Equation 5 for y:

$$y \sim (q)^{\mu}(1-q)^{y_i} \qquad y_i \in \{I^1\}, y_i \geq 0, \mu > 0, q \geq 0, \quad [5]$$

where y_i is the number of incidents in year ith, μ (1-q)/q is the expected annual (mean) of incidents, E(y), and μ (1-q)/q^2 is the expected variance, V(y). Due to uncertainty, the prior distribution for μ is assumed to follow a Gamma distribution as expressed in Equation 6, ~ *Gamma* (α, β):

$$p(\mu) \propto \mu^{\alpha-1} e^{-\beta\mu}, \alpha > 0, \beta > 0, \quad [6]$$

and that for q is assumed to follow a Beta distribution in Equation 7, q~Beta (a, b):

$$p(q) \propto q^{a-1}(1-q)^{b-1}, a > 0, b > 0. \quad [7]$$

From Baye's theorem, the posterior distribution in Equation 8, which permits a projection of accident frequency in the future, p (μ, q | Data), is

$$p(\mu, q \mid Data) \propto l\,(Data \mid \mu, q)p(\mu)p(q)$$

$$\propto q^{n\mu}\,(1-q)^{s}(\mu^{\alpha-1}e^{-\beta\lambda})q^{a-1}(1-q)^{b-1}$$

$$\propto q^{n\mu+a-1}(1-q)^{s+b-1}(\mu^{\alpha-1}e^{-\beta\mu}), \quad [8]$$

where Data = $(y_0, y_1, \ldots, y_{N_t})$, s = $\sum_{i=0}^{N_t} y_i$, N_t is the number of years, and| $l(Data \mid \mu, q)$ is the Negative Binomial likelihood distribution.

Time Series Accident Prediction

Freivalds and Johnson (1990) described that the time series of accident data from the previous week or month would influence the next week or month's data. They reported that accident data varies about a mean value, which applied the concept of time series in accident prediction. The manual time series prediction was analysed in Excel, where the mean value of previous months was the predicted value.

The R package applied in this research was the library "forecast" and "tseries" with "Box-test". The frequency of accidents was expressed in terms of a time series model due to its capability to forecast, interpret, and test hypotheses concerning the data (Sari et al., 2009). The behaviour and pattern of past observations will be assumed to continue in the future. The Augmented Dickey-Fuller test (ADF) confirmed the stationarity of data in R. With the nature of accident data received, the number of accidents did not show a significant relationship among the variables such as date of accident and nationality. Thus, linear regression was not considered in accident prediction. Auto-Regression Integrated Moving Average (ARIMA) was used to predict future values using auto-arima in the forecast package in R. The ARIMA (p, d, q) model consists of expressions identified as the order (p) of the auto-regressive part (AR), with an order of differentiation model (d) and an order (q) for moving average (MA). The seasonal ARIMA (p, d, q) $(P, D, Q)_s$ is a time series model with recurring peaks that represent the order or period of seasonality (Melchior et al., 2021).

Several analyses in R were used for model selection, such as the autocorrelation function (ACF), partial autocorrelation function (PACF), the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC and BIC are the criteria that balance

the model's goodness of fit and complexity and achieve a trade-off between fitting the available data and preventing over-fitting (Esmaili et al., 2021). Therefore, it is reported that it is preferable to have a lower value of AIC and BIC (Abdulqader et al., 2020). The Box-Ljung statistical test in R was used to confirm the correlation of the data.

Model Performance Measure

Two goodness-of-fit measures (Kuşkapan et al., 2021; Thakali et al., 2016) were used to check the model's performance. The first one is the mean absolute error (MAE) (Equation 9), and the other one is the RMSE (Equation 10).

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \quad [9]$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad [10]$$

where y_i is the ith observed accident frequency, $\hat{y}_i$ is the estimated accident frequency for the ith observation, and n is the total observations.

Mean absolute percentage error (MAPE) was used to determine the model accuracy by using the formula as depicted in Equation 11 and evaluating its accuracy in accordance with Table 1 (Weng et al., 2015).

$$\text{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n} \times 100\% \quad [11]$$

Table 1
Scale of evaluation of prediction accuracy

Condition	Assessment
MAPE $\leq$ 10%	Highly accurate prediction
10% < MAPE $\leq$ 20%	Good prediction
20% < MAPE $\leq$ 50%	Reasonable prediction
MAPE > 50%	Inaccurate prediction

RESULTS AND DISCUSSION

In Malaysia, any accident with four lost days must be reported to DOSH through the myKKP website. The database received from DOSH was a relational database stored in a table with accidents recorded in a row and various attributes in columns. An accident dataset of four years (2018-2021) in Kuala Lumpur consisting of the accidents' date, details of injured persons, classification of accidents, incident descriptions, injured parts, type of industries, and related coding for each variable was acquired from DOSH for this study. However, the data were incomplete because containing fewer details and were repetitive with very short descriptions, as highlighted by Zermane et al. (2022). In addition, Mohamad et al. (2019) and Muhamad et al. (2021) highlighted that Malaysian manufacturing companies still lack awareness and are not accustomed to implementing big data analytics due to the high cost of cloud computing services and worry the companies' database being stolen or compromised.

The data were normalised to reduce redundancy and eliminate undesirable data. It was done by removing the duplicated rows in the Excel spreadsheet. In addition, the dataset received had many missing values in various attributes, which is considered the extent of the error. It may be due to several individual and organisational shortcomings in data reporting, collection, management, and processing (Ahmed et al., 2020).

Table 2
Number of accidents in the year 2018 to 2021

Year	Number of Attributes (Column in database)	Number of Accidents	
		Raw Data	Normalised
2018	16	190	188
2019	6	375	338
2020	23	322	305
2021	22	244	216
Total		1131	1047

As shown in Table 2, out of 1131 accident cases reported to DOSH, about 84 cases (7.43%) were missing vital information in the available reported cases. In 2019, the least number of attributes were reported, and most information was missing. It means that a researcher would face difficulty analysing the root causes and the contributing factors to accidents. The number of attributes is the table column found in the database from DOSH. The attributes include age, victim status, date of the accident, accident classification by employers and DOSH officers, a short description of accidents, gender, nationality, type of injuries, accident agent, type of industries, work sector, body part injured and each code of above mentioned which had been defined by DOSH. Hadi et al. (2017) revealed that non-reporting accidents are prevalent in industries where 94.7% of construction workers did not report accidents.

On the other hand, limited research utilised publicly available data from the DOSH website. For example, Rafindadi et al. (2022) and Zermane et al. (2022) analysed data from the DOSH website for fatal fall-related accidents. Their finding showed the limitation in missing vital information in DOSH data. It was urged to adequately document and make the record up to date and international standard. In general, it was observed that 2021 accident reporting was more detailed compared to previous years. It has been discovered that the upper limbs are the most registered injury, followed by the lower limbs. However, the accident records showed a general injury sustained without providing detailed analysis, also reported by Rafindadi et al. (2022).

Frequency Modelling

Different accident prediction models are developed using econometric models such as ARIMA, negative binomial and Poisson models, as Quddus (2008) reported. Due to

the availability of accident data, the distribution was based on the time elapsed between accident dates, as shown in Figure 1. It is the time interval between the date of the latest accident and the previous occurrence (Hajakbari & Minaei-Bidgoli, 2014). Based on Figure 1, accidents were reported daily (shortest time interval between two accidents) in Kuala Lumpur, Malaysia. It was recorded that no accident was reported in two consecutive weeks, and the longest interval was 15 days, which can be considered "zero accident" and is extremely difficult to achieve (Attwood et al., 2006).

The frequency of accidents had been modelled statistically by fitting two distributions, Poisson and negative binomial distribution, as shown in Figures 2(a) and 2(b), respectively. From the data, everyday accidents are reported to DOSH. Both models are mathematical equations through statistical modelling of accident data and show the same right-skewed pattern. The distributions are to be used as the base condition for model development, where the frequency of an accident is significant in risk analysis. By having such a quantitative approach based on past reports, safety practitioners may be able to present unique safety practices to reduce accident frequency since Attwood et al. (2006) highlighted that the prediction of annual accidents was expected to be around the mean value. It indicates that given the same accident data, the impact of the variables on the different models is similar, but the expected frequency obtained from each model is slightly different. In other words, the model could estimate how many accidents would occur under average situations because, to date, no organisation has established a major change in the safety culture. Weng et al. (2015) highlighted the difficulty in predicting accidents with perfect confidence due to their uncertainty. Khattak et al. (2021) reported that the Poisson regression model is the first choice of researchers to adopt for the count data model in the beginning, and the negative binomial model is still in great popularity in the modelling process.

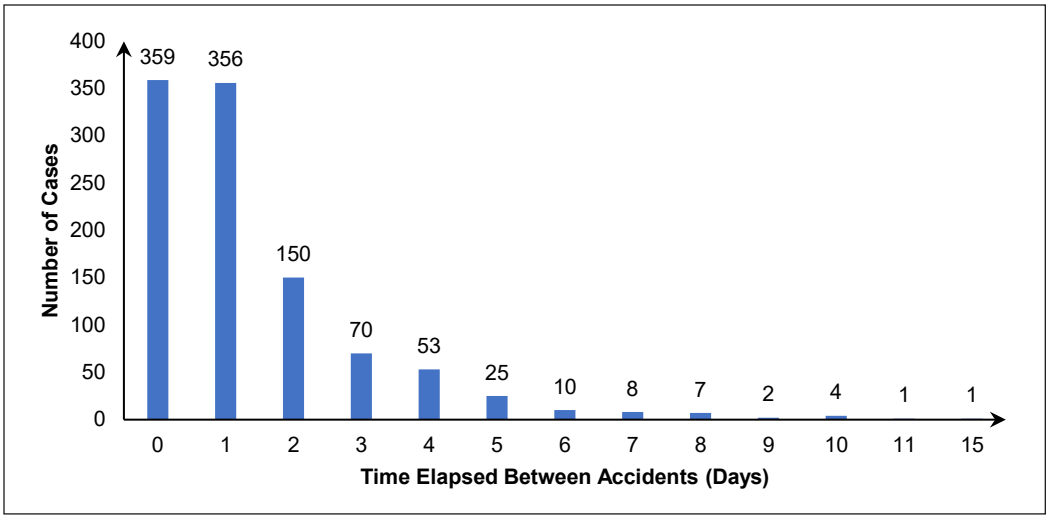
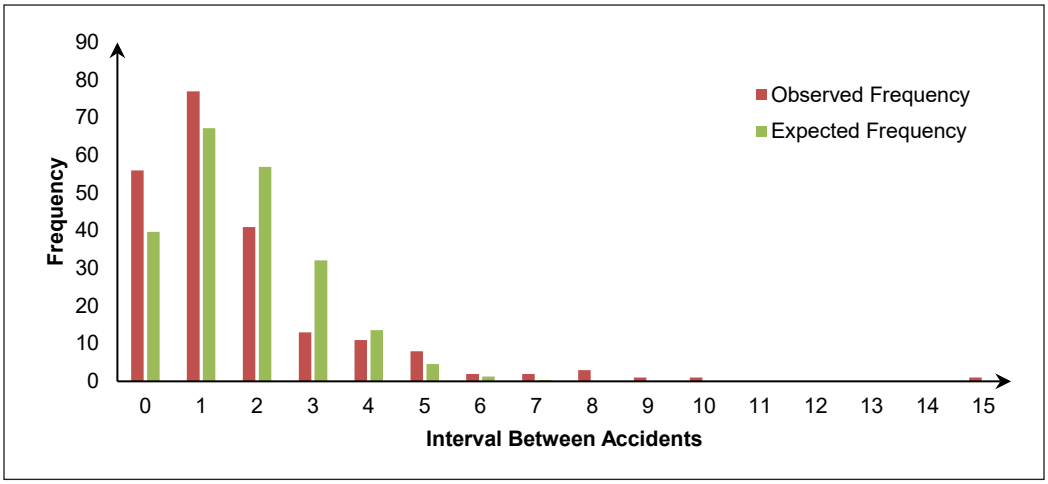
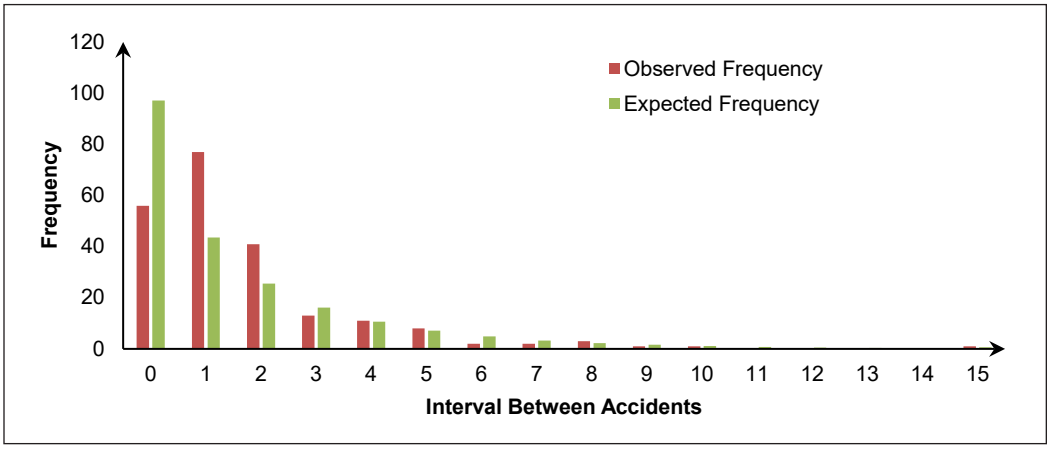


Figure 1. Distribution of the number of accidents for various intervals



(a)



(b)

Figure 2. Frequency modelling: (a) Poisson distribution; (b) Negative binomial distribution

Table 3 tabulated the performance of prediction based on Poisson and negative binomial. The predicted number of accidents using both distributions resulted in 216 cases. The MAPE based on monthly accident cases for the Poisson distribution is 51%, whereas the negative binomial distribution is 33%. The result shows that the negative binomial distribution makes a reasonable prediction compared to the Poisson distribution.

Table 3
Results of distribution accuracy

Distribution	Predicted Accident	MAE	RMSE	MAPE (%)
Poisson	216	4.721	7.974	51
Negative binomial	216	6.393	1.253	33

This finding is consistent with the research by Khattak et al. (2021), where the negative binomial model performed better than the Poisson model. Furthermore, Quddus (2008) found that the negative binomial model application is not statistically significant in serial correlation and non-stationarity in time series of accident data. On the other hand, Meel et al. (2007) utilised the National Response Centre (NRC) database for incident prediction through frequency modelling. Their findings found significantly different predictions using Poisson and negative binomial distribution in different companies. Therefore, the same distribution may not agree better across various companies.

Hajakbari and Minaei-Bidgoli (2014) reported that analysing occupational accident databases using data mining could reveal meaningful patterns that are unable to be provided by traditional statistical methods. Radzuan et al. (2020) believed that the accuracy of the prediction model for road traffic accidents in Malaysia could be increased with more features included, such as vehicle types, gender of driver and others. It is also supported by Alawad et al. (2019), where an increase in the dataset and more attributes would contribute to significant analysis and results. However, Choo et al. (2022) reported that the accident database in DOSH Malaysia is still relatively underutilised, and the data received for this study found many missing values, which may affect the accuracy of the modelling. In addition, Koc et al. (2022) highlighted that their finding for the best occupational accident prediction model for short-term and mid-term was the W-ANN model, and the long-term was the W-MARS model. They also reported that the developed predictive model might show different accuracy for different countries due to the country-specific dataset. In addition, Zhu et al. (2023) revealed that the recent text-based AI tool, ChapGPT, frequently obtained mistakes and errors and needs more effective research. Thus, Malaysians must develop an occupational predictive model to manage safety issues more efficiently and understand what national conditions cause more or fewer accidents.

Time Series Prediction

The ARIMA model is a stochastic time series prediction for short-term forecasting with high accuracy and applies to stationary time series (Li et al., 2021). Figure 3 shows the observed monthly accident data reported to DOSH based on the time between accident dates from January 2018 to December 2021. This period was chosen due to the availability, accuracy and quality of the data received from DOSH. The data fluctuates around the mean value, with no noticeable data sequence trend. Therefore, it is preliminarily determined that the data remains stable and does not change over time.

The ACF and PACF plots from R use a 95% confidence level, as shown in Figures 4(a) and 4(b), which are dashed blue lines indicating the significant threshold level. There are many spikes above the threshold level, and both plots observed tail-off patterns. It is observed that the values of AFC coefficients are gradually declining, and the AFC analysis

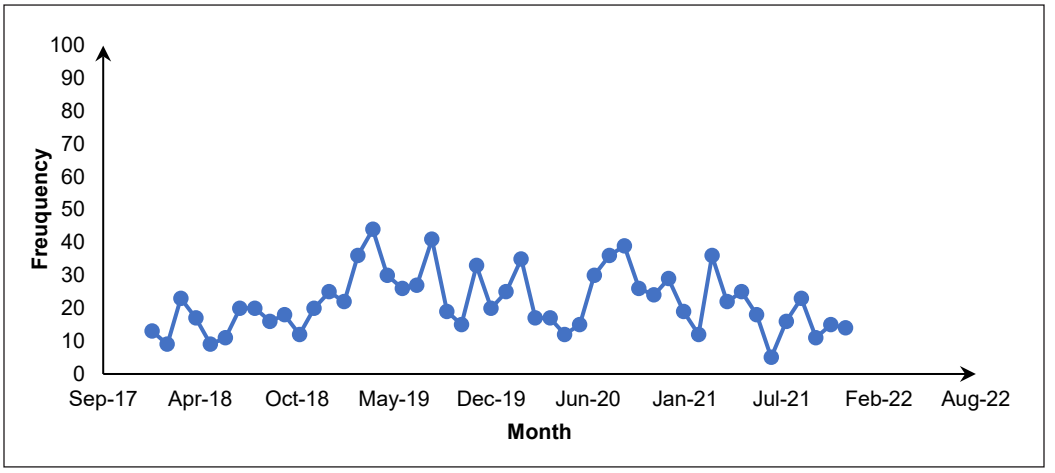
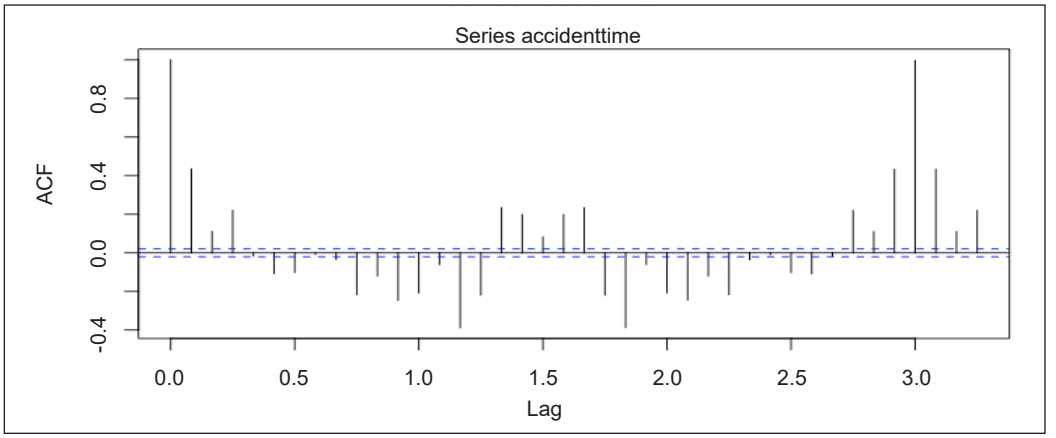
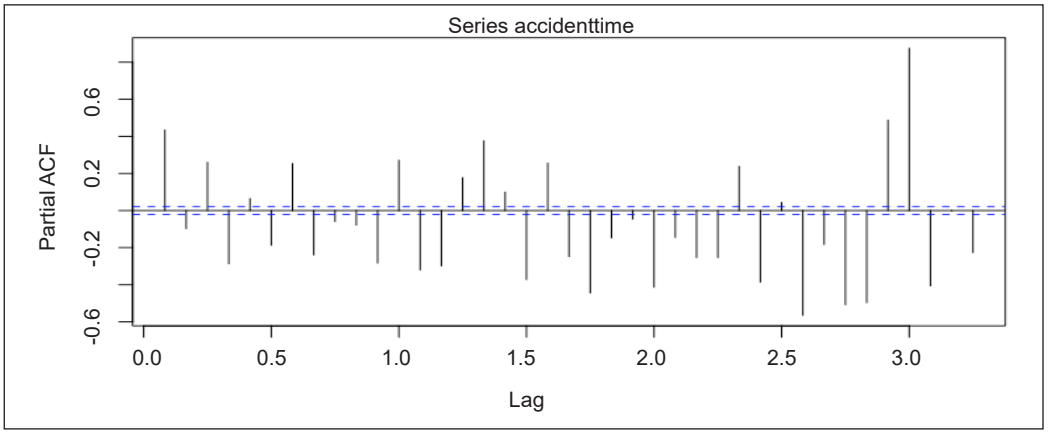


Figure 3. Original time series accident data in KL from January 2018 until December 2021



(a)



(b)

Figure 4. (a) ACF plot; (b) PACF plot

indicates stationary data since the auto-correlation function's fast decay. As highlighted by Alabdulrazzaq et al. (2021), the inspection to visual observation of the data in ACF and PACF helps determine the values for parameters p and q in ARIMA.

The ADF was tested with stationary alternative hypothesis using adf.test in the tseries package. The ADF test in R showed a printed value of -42.701, which is smaller than the p-value of 0.01 and indicates a significant autocorrelation pattern (Abdulqader et al., 2020). Therefore, the null hypothesis was rejected, and the stationary trend was verified. Quddus (2008) also studied the ARIMA model with stationarity time series data based on traffic accidents in Great Britain, UK. On the other hand, Al-Hasani et al. (2019) and Abdulqader et al. (2020) investigated a non-stationary time series of monthly road traffic accidents in Oman and Erbin City of Iraq, respectively. It suggests that different time series accident data must confirm their stationarity before applying the modelling process.

Since the series does not have a trend (it is stationary), it is not applied differencing for ARIMA models. The auto.arima function of the forecast library in R Studio was used to identify the best-suited series data. This function returned the ARIMA model based on the model's generality, which characterises the sample data and the entire population over a given period. Table 4 presents the ARIMA model that best fits the time series data identified, where ARIMA(2,0,2)(2,0,0)(12) was the best-fit model. The selection of the best-fitting model is based on the lowest rate of AIC, corrected AIC (AICc) or BIC, which denotes better generality and more significant potential for maximising the likelihood function (de Souza et al., 2022). Based on the R Studio models fitting, the lowest AIC reported was 58295.35 for ARIMA(2,0,2)(2,0,0)(12). In addition, the forecast package in R had automatically conducted re-fitting without approximation to confirm the best model, with an AIC value of 58313.39, which is also lower compared to other models.

The first part of the model has a second-order self-regression ($p = 2$), no differentiation ($d = 0$) and a second-order moving average ($q = 2$). The $p=2$ indicates that two previous periods are used in the auto-regression of the time series, $q=2$ indicates two lags of the error component, and $d=0$ indicates no differencing transformation required to turn the time series into stationary (Alabdulrazzaq et al., 2021). The other part of the model indicates the developed model for the seasonal component, whose elements only have second-order seasonal self-regression ($P = 2$, $D = 0$ and $Q = 0$). The index of 12 refers to the number of periods per season and the corresponding months for different years. Since the input was monthly time series data, the length of seasonality is 12. Based on Table 4, the ARIMA(0,0,0) with zero mean shows the highest AIC value compared to a non-zero mean of the same model. Alabdulrazzaq et al. (2021) reported that the manual model tends to overfit the data.

Interestingly, Abdulqader et al. (2020) reviewed the studies by other researchers in several countries. For example, in Saudi Arabia, the best fatality forecasting model was

Table 4
Results of ARIMA in R

Models Fitting	Mean	AIC
ARIMA(2,0,2)(1,0,1)(12)	Non-zero	60971.19
ARIMA(0,0,0)	Non-zero	64635.14
ARIMA(0,0,0)	Zero	82371.27
ARIMA(1,0,0)(1,0,0)(12)	Non-zero	62606.65
ARIMA(0,0,1)(0,0,1)(12)	Non-zero	61116.68
ARIMA(2,0,2)(0,0,1)(12)	Non-zero	60989.85
ARIMA(2,0,2)(1,0,0)(12)	Non-zero	61018.40
ARIMA(2,0,2)	Non-zero	61281.70
ARIMA(2,0,2)(2,0,0)(12)	Non-zero	58295.35
ARIMA(2,0,1)(2,0,0)(12)	Non-zero	58749.84
ARIMA(1,0,1)(2,0,0)(12)	Non-zero	60348.56
ARIMA(3,0,1)(2,0,0)(12)	Non-zero	58742.65
ARIMA(2,0,2)(2,0,0)(12)	Non-zero (re-fitting without approximation)	58313.39

ARIMA(1,1,3)(0,1,0) by using historical traffic accident data from 2013 to 2017; the AR of order one also showed the best model to analyse traffic accidents in Al-Qadisiya. On the other hand, ARIMA(1,0,0)(2,1,0)12 showed a good performance model for monthly traffic accidents in India, and ARIMA(1,0,2)(1,0,0)12 for motorcycle injuries study. Several works using different statistical methods have been done with traffic accidents worldwide, where each researcher reported their best model for forecasting. However, it is reported that the best model varies from application to application (Domingos, 2012), although many researchers are trying various models and believe in their efforts' superiority. Li et al. (2021) studied highway transportation accidents in China from 2013 to 2019 and applied the ARIMA modelling process.

Freivalds and Johnson (1990) presented ARIMA's Box-Jenkins modelling procedures, where model selection is based on the sum of errors with less than infinity. Thus, ARIMA(2,0,2)(2,0,0) was selected by R. The model established may not be perfect but best suits the available data set and returned the smallest standard error. After fitting the best prediction model, a residual analysis indicated a serial correlation in the data (Quddus, 2008). The statistical tests of Box-Ljung were performed in R, and the p-value is less than the 5% significance level; the residuals are dependent on each other where there is serial correlation and without white noise.

Based on Figure 5 and Table 5, the number of predicted accidents in R (214 cases) and manual time series (229 cases) are low compared to the number of actual accidents (216 cases). Each month, the prediction in R and manual time series was not equal to the actual accident. The number of accidents in actual and prediction shows less difference in April 2021 compared to other months. For the nature of time-series accident data, the

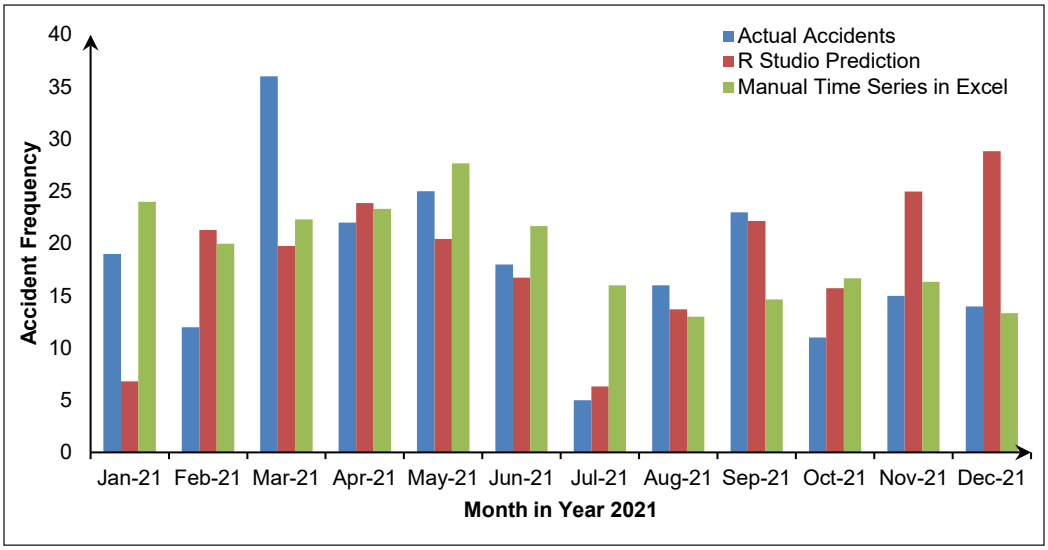


Figure 5. Comparison of actual and predicted accidents

Table 5
Comparison of manual time series and R Studio prediction accuracy

Model	Predicted Accident	MAE	RMSE	MAPE (%)
Manual time series	229	6.181	7.784	49
R Studio	214	6.621	8.537	40

predicted number will vary around the mean of previous months. The predicted number of accidents in R was observed to be a difference of 2 cases from the actual reported cases at 95% confidence, which shows that ARIMA(2,0,2)(2,0,0) produced a reasonably accurate prediction. Compared to the manual time series method, there was a 13-case difference from the actual reported cases. It shows that R Studio is able to forecast more accurately than manual time series. Alabdulrazzaq et al. (2021) presented their forecast for COVID-19 cases using the ARIMA model, which was accurate despite the dynamic conditions of the daily disease data.

They highlighted that utilisation of software packages to facilitate the automated selection of ARIMA's model in R Studio returned the best-fit model. Table 5 tabulates the prediction accuracy based on MAE, RSME, and MAPE. The MAPE of the prediction in R is 40%, and the manual time series is 49%, which shows that the ARIMA model can make a reasonable prediction. The finding was supported by Attwood et al. (2006), where the number of accidents prediction indicated that the number of future accidents happened was around the mean value of the past number of accidents. Alabdulrazzaq et al. (2021) highlighted that predicted values will not necessarily equal actual observed values but use scale-dependent accuracy measurement, as shown in Table 1. Rohayu et al. (2012) also

found that the ARIMA model performed better than Poisson and Negative Binomial for road accident prediction. Quddus (2008) highlighted that the performance of the model could be measured based on MAE, MAPE and RMSE, where the smaller the value, the better the fit of the model.

On the other hand, Abdulqader et al. (2020) reported that their best model was ARIMA(0,1,1)(1,0,1)12 with an MAE of 23.11, which fits predicted accident injuries. In this study, the number of accidents from April 2021 until September 2021 was close to the forecast values, and there were also decreased and increased forecasted values reported. There are the same mean values of actual accident cases and forecasted accidents in 2021, which is 18 cases. This research used time series analysis to contribute to accident modelling and forecasting, which agrees with Marhavilas et al. (2013). It was also supported by de Souza et al. (2022), where ARIMA successfully applied modelling for time series forecasting.

Table 6 shows the performance of accident data for actual and predicted data using different approaches. From the data analysis, the mean value for R Studio and manual time series prediction is comparable to actual accident data, whereas both negative binomial and Poisson distributions recorded lower values. The negative binomial and Poisson distribution variance significantly differed from the actual accident data. Besides that, the negative binomial and Poisson standard deviation also show a large difference compared to the actual accident data. However, the standard deviation of both R Studio and manual time series prediction is lower than the actual accident data, showing that the R Studio prediction is more accurate than other approaches. On the other hand, in a modelling study conducted by Bora et al. (2020), they observed that lower standard error was considered precise, and the model developed from R was reasonably accurate.

Table 6
Statistical performance of the actual accident and predicted data

Statistical Analysis	Actual Data	R Studio	Manual Time Series	Negative Binomial	Poisson
Mean	18.0	18.4	19.0	14.0	13.0
Variance	58.2	43.6	20.5	596.5	479.2
Standard Deviation	8.0	6.9	4.7	25.2	22.6

CONCLUSION AND FUTURE RESEARCH

This paper investigated the DOSH accident data across various industries in KL and pointed out that the database was incomplete with missing values. This study analyses the accident data and time series models generated by the auto.arima function in R. The time series data considered in the study represented monthly industrial accidents in KL from January 2018 to December 2021, totalling 1047 cases. Upon investigation, the stakeholders in accident reporting shall report more detailed information, which could be useful for future

research. The data fit the stationary time series curve, as the Augmented Dickey-Fuller test confirmed. The ARIMA(2,0,2)(2,0,0)(12) model fits predicted numbers and results in the best model in this time series data. The model has been validated against 20% of the actual accident data. The model generated in R outperformed the Poisson and negative binomial model with the lowest MAPE (40%). The MAPE for the manually calculated time series model was 49%.

The results of this study support the idea that auto.arima function from the forecast R package would be a significant improvement in forecasting accident frequency from a safety perspective. Based on the findings, industrial safety practitioners should report accidents truthfully in the era of digitalisation. It could enable future data-driven accident predictions to be carried out. The main bottleneck of the study was the lack of informative records reported and the access to data in DOSH, which resulted in the underutilisation of DOSH data. In addition, the number of datasets used in training can influence the study's results. It can be verified in future by including more data and adjustments to the model. The same data could be tested using Phyton or Matlab to compare their accuracy.

ACKNOWLEDGEMENT

We thank the Master of Process Safety and Loss Prevention, Department of Chemical and Environmental Engineering, Universiti Putra Malaysia, for providing beneficial insight into improving safety culture in Malaysian industries. This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Abdulqader, Q. M., Hassan, M. T., & Ahmad, K. H. (2020). Building a mathematical SARIMA model for forecasting the number of monthly injured people by traffic accidents in Erbil City. *Technology Reports of Kansai University*, *62*(3), Article 909916.

Abdullah, D. N. M. A., & Wern, G. C. M. (2011). An analysis of accidents statistics in Malaysian construction sector. *International Conference on E-Business, Management and Economics*, *3*(1), 1–4.

Ahmed, A., Sadullah, A. F. M., Yahya, A. S., Akhtar, M. N., & Azam, Q. (2020). How accurate are locations in Malaysian accident data? Development of a rectification procedure based on nested filtered search technique. *Transportation Research Procedia*, *48*, 1125–1141. https://doi.org/10.1016/j.trpro.2020.08.138

Alabdulrazzaq, H., Alenezi, M. N., Rawajfih, Y., Alghannam, B. A., Al-Hassan, A. A., & Al-Anzi, F. S. (2021). On the accuracy of ARIMA based prediction of COVID-19 spread. *Results in Physics*, *27*, Article 104509. https://doi.org/10.1016/j.rinp.2021.104509

Alawad, H., Kaewunruen, S., & An, M. (2019). Learning from accidents: Machine learning for safety at railway stations. *IEEE Access*, *8*, 633–648. https://doi.org/10.1109/ACCESS.2019.2962072

Al-Hasani, G., Khan, A. M., & Al Reesi, H. (2019). Diagnostic time series models for road traffic accidents data modelling, analysis and forecasting road traffic accidents view project design and verification of

safety critical embedded systems view project. *International Journal of Applied Statistics and Economics*, *2*, 19–26.

Ali, D., Yusof, Y., & Adam, A. (2017). Safety culture and issue in the Malaysian manufacturing sector. *MATEC Web of Conferences*, *135*, Article 00031. https://doi.org/10.1051/matecconf/201713500031

Attwood, D., Khan, F., & Veitch, B. (2006). Validation of an offshore occupational accident frequency prediction model - A practical demonstration using case studies. *Process Safety Progress*, *25*(2), 160–171. https://doi.org/10.1002/prs.10128

Ayob, A., Shaari, A. A., Zaki, M. F. M., & Munaaim, M. A. C. (2018). Fatal occupational injuries in the Malaysian construction sector-causes and accidental agents. *IOP Conference Series: Earth and Environmental Science*, *140*(1), Article 012095. https://doi.org/10.1088/1755-1315/140/1/012095

Bora, B., Chattopadhyaya, S., & Kumar, R. (2020). Development of mathematical model for friction stir welded joint using "R" programming. *Materials Today: Proceedings*, *27*(3), 2142–2146. https://doi.org/10.1016/j.matpr.2019.09.083

Chong, H. Y., & Low, T. S. (2014). Accidents in Malaysian construction industry: Statistical data and court cases. *International Journal of Occupational Safety and Ergonomics*, *20*(3), 503–513. https://doi.org/10.1080/10803548.2014.11077064

Choo, B. C., Razak, M. A., Radiah, A. B. D., Tohir, M. Z. M., & Syafiie, S. (2022). A review on supervised machine learning for accident risk analysis: Challenges in Malaysia. *Process Safety Progress, 41*(S1), S147-S158. https://doi.org/10.1002/prs.12346

Denham, B. E. (2020). Poisson and negative binomial regression. In *Categorical Statistics for Communication Research* (1st ed.: pp. 74–94). John Wiley & Sons.

de Souza, J. A. F., Silva, M. M., Rodrigues, S. G., & Santos, S. M. (2022). A forecasting model based on ARIMA and artificial neural networks for end–OF–life vehicles. *Journal of Environmental Management*, *318*, Article 115616. https://doi.org/10.1016/j.jenvman.2022.115616

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Esmaili, N., Buchlak, Q. D., Piccardi, M., Kruger, B., & Girosi, F. (2021). Multichannel mixture models for time-series analysis and classification of engagement with multiple health services: An application to psychology and physiotherapy utilization patterns after traffic accidents. *Artificial Intelligence in Medicine*, *111*, Article 101997. https://doi.org/10.1016/j.artmed.2020.101997

Freivalds, A., & Johnson, A. B. (1990). Time-series analysis of industrial accident data. *Journal of Occupational Accidents*, *13*(3), 179–193. https://doi.org/10.1016/0376-6349(90)90020-V

Hadi, N. A. A., Tamrin, S. B. M., Guan, N. Y., How, V., & Rahman, R. A. (2017). Association between non-reporting of accident and contributing factors in Malaysia's construction industry. *The Japanese Journal of Ergonomics, 53*(Supplement2), S648-S651. https://doi.org/10.5100/jje.53.S648

Hajakbari, M. S., & Minaei-Bidgoli, B. (2014). A new scoring system for assessing the risk of occupational accidents: A case study using data mining techniques with Iran's Ministry of labor data. *Journal of Loss Prevention in the Process Industries*, *32*, 443–453. https://doi.org/10.1016/j.jlp.2014.10.013

Ismail, N., & Zamani, H. (2013). Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. *Casualty Actuarial Society E-Forum, 41*(20), 1–28.

Jian, T. (2021, October 22-24). *Statistical analysis and countermeasures study on major accidents of highway and waterway in China*. [Paper presentation]. 6th International Conference on Transportation Information and Safety (ICTIS), Wuhan, China. https://doi.org/10.1109/ICTIS54573.2021.9798527

Khattak, M. W., Pirdavani, A., De Winne, P., Brijs, T., & De Backer, H. (2021). Estimation of safety performance functions for urban intersections using various functional forms of the negative binomial regression model and a generalized Poisson regression model. *Accident Analysis and Prevention*, *151*, Article 105964. https://doi.org/10.1016/j.aap.2020.105964

Kidam, K., Abidin, Z. Z., Sulaiman, Z., Hashim, M. H., Ripin, A., Ali, M. W., Safuan, H. M., Haron, S., Othman, N., Zakaria, Z. Y., Fandi, F. M., Masri, M. F., Hassan, S. A. H. S., Ali, N. M., Ahmad, A., & Asri, H. (2015). Current status of industrial accident learning in Malaysia. *Journal of Occupational Safety and Health*, *12*(1), 1-4.

Kim, S., Lee, J., & Kang, C. (2021). Analysis of industrial accidents causing through jamming or crushing accidental deaths in the manufacturing industry in South Korea: Focus on non-routine work on machinery. *Safety Science*, *133*, Article 104998. https://doi.org/10.1016/j.ssci.2020.104998

Koc, K., Ekmekcioğlu, Ö., & Gurgun, A. P. (2022). Accident prediction in construction using hybrid wavelet-machine learning. *Automation in Construction*, *133*, Article 103987. https://doi.org/10.1016/j.autcon.2021.103987

Kuşkapan, E., Çodur, M. Y., & Atalay, A. (2021). Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms. *Accident Analysis and Prevention*, *155*, Article 106098. https://doi.org/10.1016/j.aap.2021.106098

Li, X., Liu, Y., Fan, L., Shi, S., Zhang, T., & Qi, M. (2021). Research on the prediction of dangerous goods accidents during highway transportation based on the ARMA model. *Journal of Loss Prevention in the Process Industries*, *72*, Article 104583. https://doi.org/10.1016/j.jlp.2021.104583

Manan, M. M. A., Jonsson, T., & Várhelyi, A. (2013). Development of a safety performance function for motorcycle accident fatalities on Malaysian primary roads. *Safety Science*, *60*, 13–20. https://doi.org/10.1016/j.ssci.2013.06.005

Marhavilas, P. K., Koulouriotis, D. E., & Spartalis, S. H. (2013). Harmonic analysis of occupational-accident time-series as a part of the quantified risk evaluation in worksites: Application on electric power industry and construction sector. *Reliability Engineering and System Safety*, *112*, 8–25. https://doi.org/10.1016/j.ress.2012.11.014

Meel, A., O'Neill, L. M., Levin, J. H., Seider, W. D., Oktem, U., & Keren, N. (2007). Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, *20*(2), 113–127. https://doi.org/10.1016/j.jlp.2006.10.003

Meel, A., & Seider, W. D. (2006). Plant-specific dynamic failure assessment using Bayesian theory. *Chemical Engineering Science*, *61*(21), 7036–7056. https://doi.org/10.1016/j.ces.2006.07.007

Melchior, C., Zanini, R. R., Guerra, R. R., & Rockenbach, D. A. (2021). Forecasting Brazilian mortality rates due to occupational accidents using autoregressive moving average approaches. *International Journal of Forecasting*, *37*(2), 825–837. https://doi.org/10.1016/j.ijforecast.2020.09.010

Mohamad, E., Shern, T. Y., Jamli, M. R., Mohamad, N. A., Rahman, M. A. A., Salleh, M. R., Oktavianty, O., & Ito, T. (2019, September 25-27). *Readiness of Malaysian manufacturing firms in implementing industry 4.0.* [Paper presentation]. Proceedings of the 29th Design Engineering and Systems Division Lecture Meeting of the Japan Society of Mechanical Engineers, Sendai, Japan. https://doi.org/10.1299/jsmedsd.2019.29.1201

Muhamad, M. Q. B., Syed Mohamad, S. J. A. N., & Mat Nor, N. (2021). When digital intelligent taking over: Addressing SMEs readiness on Industry 4.0 in Malaysia. *International Journal of Academic Research in Business and Social Sciences*, *11*(1), 543-551. https://doi.org/10.6007/ijarbss/v11-i1/8335

Quddus, M. A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident Analysis and Prevention*, *40*(5), 1732–1741. https://doi.org/10.1016/j.aap.2008.06.011

Radzuan, N. Q., Hassan, M. H. A., Majeed, A. P. P. A., Musa, R. M., Razman, M. A. M., & Kassim, K. A. A. (2020). Predicting serious injuries due to road traffic accidents in Malaysia by means of artificial neural network. In Z. Jamaludin & M. N. A. Mokhtar (Eds.), *Intelligent Manufacturing and Mechatronics* (pp. 75-80) Springer. https://doi.org/10.1007/978-981-13-9539-0_8

Rafindadi, A. D. U., Napiah, M., Othman, I., Mikić, M., Haruna, A., Alarifi, H., & Al-Ashmori, Y. Y. (2022). Analysis of the causes and preventive measures of fatal fall-related accidents in the construction industry. *Ain Shams Engineering Journal*, *13*(4), Article 101712. https://doi.org/10.1016/j.asej.2022.101712

Rohani, J. M., Atan, H., Hamid, W. H. W. Johari, M. F., & Ramly, E. (2015). Occupational accident cost estimation: A case study in wood based related industries. *Journal of Occupational Safety and Health*, *12*(1), 109–116.

Rohayu, S., Rahim, S. A. S. M., Marjan, J. M., & Voon, W. S. (2012). *Predicting Malaysian road fatalities for year 2020*. Malaysian Institute of Road Safety Research (MIROS)

Sari, M., Selcuk, A. S., Karpuz, C., & Duzgun, H. S. B. (2009). Stochastic modeling of accident risks associated with an underground coal mine in Turkey. *Safety Science*, *47*(1), 78–87. https://doi.org/10.1016/j.ssci.2007.12.004

Thakali, L., Fu, L., & Chen, T. (2016). Model-based versus data-driven approach for road safety analysis: Do more data help? *Transportation Research Record*, *2601*(1), 33–41. https://doi.org/10.3141/2601-05

Warner, P. (2015). Poisson regression. *Basic Biostatistics for Medical and Biomedical Practitioners, 41*(3), 591-595.

Weng, J., Qiao, W., Qu, X., & Yan, X. (2015). Cluster-based lognormal distribution model for accident duration. *Transportmetrica A: Transport Science*, *11*(4), 345–363. https://doi.org/10.1080/23249935.2014.994687

Zein, R. M., Halim, I., Azis, N. A., Saptari, A., & Kamat, S. R. (2015). A survey on working postures among Malaysian industrial workers. *Procedia Manufacturing*, *2*, 450–459. https://doi.org/10.1016/j.promfg.2015.07.078

Zermane, A., Tohir, M. Z. M., Baharudin, M. R., & Yusoff, H. M. (2022). Risk assessment of fatal accidents due to work at heights activities using fault tree analysis: Case study in Malaysia. *Safety Science*, *151*, Article 105724. https://doi.org/10.1016/j.ssci.2022.105724

Zhu, J. J., Jiang, J., Yang, M., & Ren, Z. J. (2023). ChatGPT and environmental research. *Environmental Science & Technology, 57*(46), 17667–17670. https://doi.org/10.1021/acs.est.3c01818