scientific reports

Check for updates

OPEN

Forecasting occupational accidents in Turkey using multivariate ARMAX and NLARX models

Selcan Kaplanvural[1]✉, Eren Tosyalı[2] & İsmail Ekmekçi[3]

Occupational accidents remain a critical issue in Turkey, with significant social and economic consequences, and understanding accident trends is essential for developing effective prevention strategies. This study employed both linear AutoRegressive Moving Average with Exogenous Input (ARMAX) and Nonlinear AutoRegressive with Exogenous Input (NLARX) models to forecast future occupational accidents using four accident-related populations (y_1, y_2, y_3, y_4) derived from official insurance records. Due to the lack of consistently reported monthly data from the Social Security Institution (SSI), exogenous variables such as sectoral, economic, or demographic indicators were not incorporated, and the models were therefore identified based solely on the endogenous accident dynamics. The ARMAX identification process yielded a relatively large set of candidate parameters, which were subsequently evaluated using statistical significance criteria; only coefficients with *p* values below 0.05 and confidence intervals excluding zero were retained for interpretation. Model performance was evaluated using the normalized mean squared error (NMSE), which was computed separately for the training period, the test (out-of-sample forecasting) period, and the full dataset for each model. This multi-level evaluation enabled a consistent comparison of in-sample fitting accuracy, out-of-sample generalization capability, and overall predictive performance across ARMAX and NLARX models. The significance-based analysis revealed distinct linear dynamic structures across the output groups, with the (y_1, y_2) populations characterized by a larger number of moderate-magnitude significant coefficients, whereas (y_3, y_4) exhibited fewer but more dominant linear effects. The ARMAX model produced the lowest NMSE values across the training, test, and full datasets for most populations, demonstrating particularly strong and stable predictive accuracy for y_1 and y_2. The NLARX model yielded the best performance for y_1 and showed comparable NMSE values to ARMAX for y_2 and y_4, although it exhibited higher forecasting errors for y_3, especially in the test period. Overall, the results indicated that while NLARX was capable of capturing nonlinear patterns in specific cases, the ARMAX framework provided a more robust, interpretable, and consistently generalizable representation of the dominant temporal dynamics governing occupational accident trends. These findings highlighted the potential of multivariate time series models to support evidence-based decision-making in occupational safety planning and policy development.

Keywords Occupational accidents, ARMAX, NLARX, Time series forecasting, OHS modelling

Occupational Health and Safety (OHS) is an interdisciplinary field dedicated to protecting the well-being of workers by preventing workplace injuries, illnesses and accidents[1,2]. Occupational accidents pose a major threat to public health and workforce productivity[3], and understanding their causes and predicting their occurrence remain central concerns in OHS research[4]. Statistical analysis of accident records enables the identification of risk-prone areas and critical factors, supporting the development of targeted prevention strategies[5]. Early efforts, such as those by Andersson and Lagerlöf, emphasized the importance of comprehensive accident databases and laid the foundation for modern OHS data system[6].

As accident databases have expanded, a wide range of data-driven analytical techniques has emerged in OHS research. Beyond Natural Language Processing[7], recent studies have employed machine-learning classifiers to predict accident severity[8], clustering algorithms to identify hidden risk patterns in high-volume datasets[9], and

[1]Vocational School of Health Services, Occupational Health and Safety, Istanbul Bilgi University, 34387 Istanbul, Turkey. [2]Department of Mechatronics Engineering, Faculty of Engineering and Natural Sciences, Istanbul Bilgi University, 34060 Istanbul, Turkey. [3]Department of Industrial Engineering, Faculty of Engineering, Istanbul Commerce University, 34840 Istanbul, Turkey. ✉email: selcan.cicek@bilgi.edu.tr

Bayesian approaches for probabilistic assessment of accident likelihood[10]. These advanced analytical methods are also increasingly applied in high-risk sectors. In the construction industry, deep-learning frameworks such as CNN and LSTM architectures have been used to forecast accident frequency and model leading safety indicators[11]. Similarly, in the mining sector, deep-learning and text-mining approaches have been utilized to automatically classify hazardous events and uncover latent risk themes from unstructured safety records, providing valuable support for proactive hazard investigation and accident prevention[12].

Parallel to these text-based approaches, time series models have been widely used to forecast accident trends, particularly through univariate methods such as AutoRegressive Integrated Moving Average (ARIMA)[13]. ARIMA-based studies have been applied across various industrial settings, including Indian factories[14], mining operations in Pakistan[15], and steel manufacturing[16]. While these approaches offer reliable short-term forecasts[17–19], their univariate nature limits their ability to capture interactions among different accident types[20]. To address this limitation, researchers have explored multivariate and structure-based modeling approaches including unobserved component models[21], complex networks[22], and Bayesian frameworks[23] highlighting the need to consider inter dependencies among accident populations. Furthermore, recent comparative time-series studies such as those evaluating ARIMA, beta-ARMA, and KARMA models for occupational accident mortality in Brazil underscore both the widespread reliance on classical autoregressive approaches[24]. Long-term analyses from high-risk industrial sectors likewise indicate that accident dynamics often follow hierarchical and interdependent patterns, where shifts in one accident category can propagate through the broader safety system[25]. These findings reinforce the necessity of adopting multivariate modeling techniques capable of representing coupled temporal behaviors. Despite these advances, there remains a notable gap in the application of multivariate system identification techniques particularly AutoRegressive Moving Average with Exogenous Input (ARMAX) and Nonlinear ARX (NLARX) to occupational accident forecasting. These methods can capture nonlinear, dynamic relationships among multiple correlated variables, such as minor injuries, major injuries, and fatalities, while providing interpretable model structures[26].

Building on our earlier work, where we developed an ODE system inspired by the epidemiological SIR model to describe the propagation of occupational accidents[27], the present study adopts a data-driven, multivariate system identification perspective. In the previously proposed conceptual model, a hierarchical accident progression was assumed, in which the insured worker population primarily contributed to minor accidents, while major accidents represented the dominant transition state leading to fatal outcomes. Motivated by this structural assumption, the accident dynamics are modeled through two coupled but separately identified subsystems, namely (y_1, y_2) capturing the insured–minor accident interaction and (y_3, y_4) representing the major accident-fatality relationship. Using real-world insurance data from Turkey, we model the dynamic interplay among these accident severity groups and forecast future trends using ARMAX and NLARX frameworks. Within this framework, ARMAX models are employed to capture dominant linear temporal dependencies, while NLARX models are used to investigate potential nonlinear dynamics inherent in the accident propagation process. Model performance is evaluated through the normalized mean squared error (NMSE), providing a transparent and robust validation metric for time-series forecasting in occupational health and safety (OHS) research.

This study makes three key contributions to the occupational safety forecasting literature: (i) it provides the first multivariate system-identification framework for modeling the coupled dynamics of insured, minor, major, and fatal accident populations in Turkey; (ii) it offers a systematic comparison of linear ARMAX and nonlinear NLARX approaches within a unified data-driven architecture; and (iii) it demonstrates how cross-lag dependencies among accident categories can be leveraged to generate interpretable and policy-relevant forecasts for national OHS planning. The structure of the paper is as follows: Section "Material and methods" introduces the dataset and methodology. Section "Results" presents the model results and performance evaluation. Section "Conclusion" summarizes key findings and outlines recommendations for future research.

Material and methods

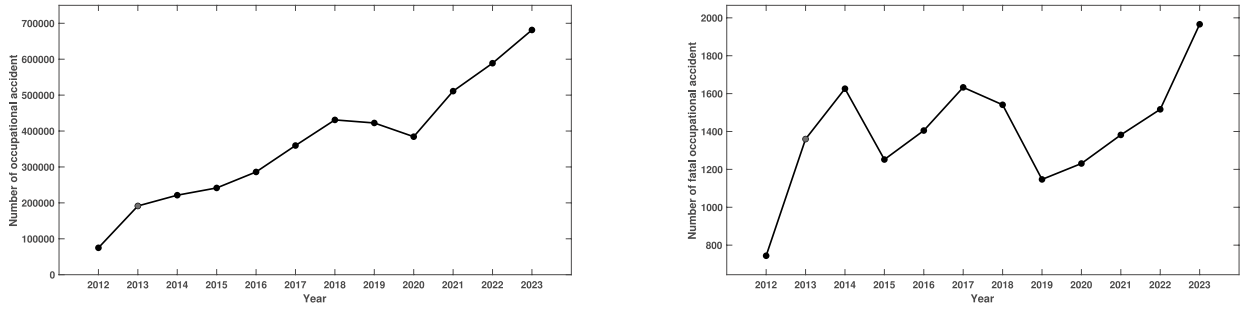
Data description

In Turkey, as in many other countries, occupational accidents present serious challenges to public health and workforce productivity[28]. According to official data published by the Social Security Institution (SSI), at least 1500 workers have lost their lives in occupational accidents each year[29]. Fatal accidents are particularly prevalent in high-risk sectors such as construction, mining, and transportation[30–32]. The persistently high rate of occupational accidents constitutes a major safety concern. The enactment of Occupational Health and Safety Law No. 6331 in 2013 marked a turning point in improving data reliability and strengthening supervision and preventive efforts[33]. According to 2023 statistics, a total of 681,401 occupational accidents were reported, including 1966 fatalities, underscoring the urgency of adopting multidimensional approaches to workplace safety[29].

The SSI provides detailed and regularly updated statistics on insured workers, workplace conditions, occupational accidents, and occupational diseases[34]. Employees are classified into three main categories based on their insurance status: 4a (private sector employees), 4b (civil servants), and 4c (self-employed individuals). The available data are further segmented by sector (industry-specific rates), time (annual and monthly trends), geography (provincial distribution), demographics (age and gender), work status (whether the employee was working at the time of the accident), and duration of incapacitation (ranging from one day to five or more days).

Figure 1 shows the number of insured workers exposed to occupational and fatal accidents in Turkey between 2012 and 2023.

As illustrated in Fig. 1, the implementation of Occupational Health and Safety Law No. 6331, marked in red, significantly impacted the reporting mechanisms for occupational accidents. Although a sharp decline occurred during the COVID-19 pandemic, recent years have seen a renewed increase in both the number of insured workers involved in occupational accidents and those involved in fatal accidents.



(a) (b)

Fig. 1. (**a**) Number of insured individuals exposed to occupational accidents. (**b**) Number of insured individuals exposed to fatal occupational accidents in Turkey (2012–2023).

Population	Description
A	The average of the total insured population under 4a
y_1	The number of accident-free insured individuals
y_2	The number of accident-minor insured individuals
y_3	The number of accident-major insured individuals
y_4	The number of accident-fatal insured individuals

Table 1. The classification of occupational accident populations.

In this study, the analysis begins in 2013, aligning with the enforcement of the occupational health and safety law to ensure data consistency and reliability. The dataset extends through 2023, providing a comprehensive view of the period. The time series consists of 132 monthly data points, covering the period from January 2013 to December 2023. The year 2023 was selected as the endpoint of the dataset, as it corse that result in death.

The classification of populations used in this study is summarized in Table 1. All population values are presented cumulatively and scaled per one million individuals (1/1,000,000).

Times series analysis

Time series analysis is a statistical technique that focuses on analyzing data collected or recorded at specific time intervals, enabling researchers and analysts to identify trends, patterns, and seasonal variations over time[35]. These analyses are very important in various fields, including economics, finance, environmental science and engineering, as they enable the prediction of future values based on past data[36–38]. In this context, Yule made great contribution to time series analysis, especially with his work on auto-regressive (AR) models, and has laid the foundation for modern statistical studies in this field[39]. The AR model is denoted as $AR(p)$ of order p. The formulation of the model is expressed as follows Eq. (1)[40].

$$Y_t = \alpha + \sum_{i=1}^{p} \phi_i Y_{t-i} + e_t \tag{1}$$

where Y_t is the value of the time series at time t. α is the constant, ϕ is the AR coefficients for each lag i. Y_{i-t} is the lagged values of the series (i.e. $Y_{t-1}, Y_{t-2}, \quad , Y_{t-p}$), e_t is the error term at time t.

The auto-regressive moving average (ARMA) model combines the components of (AR) and the moving average (MA)[41]. The formula of the ARMA model of order (p, q) is given by Eq. (2)[13].

$$Y_t = \alpha + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j} + e_t \tag{2}$$

In the ARMA model, Y_t, Y_{t-i}, e_t and e_{t-j} represent variables, where Y_t is the current observation, Y_{t-i} represents the past observation, e_t is the current error term and e_{t-j} represents the past error terms. This model is estimated using both the past observation Y_{t-i} and the error terms e_{t-j}[42].

In another study, Box and Jenkins present the Auto-Regressive Integrated Moving Average (ARIMA) methodology as a statistical approach to analyze time series data.[13]. The main difference between the ARIMA and ARMA models is that ARIMA can handle non-stationary time series by applying differencing, while ARMA is designed for stationary time series[43]. The mathematical expression of the ARIMA model (p,d,q) is shown in Eq. (3)[44].

$$(1-\sum_{i=1}^{p}\phi_i B^i)(1-B)^d Y_t = \alpha + \sum_{j=1}^{q}\theta_j B^j e_t \tag{3}$$

here B represents the back-shift operator, defined as $B^k Y_t = Y_{t-k}$ and d denotes the number of times the series must be differenced to achieve stationarity.

The Autoregressive Moving Average with Extra Input (ARMAX) model is a version of the ARMA model that incorporates exogenous factors X into the modeling process[45]. The general form of the ARMAX model (p, q, r) is given as follows Eq. (4)[46]:

$$Y_t = \alpha + \sum_{i=1}^{p}\phi_i Y_{t-i} + \sum_{j=0}^{q}\theta_j e_{t-j} + \sum_{k=1}^{r}\delta_k X_{t-k} + e_t \tag{4}$$

here X_{tik} represents the exogenous variables,and p, q and r are the orders of the AR, MA and X terms, respectively[47].

The polynomial AR model (PAR or ARX) is derived from the AR model for nonlinear time series[48]. This approach is particularly useful when the relationship between the current value and past values is non-linear. It enables a more accurate representation of the time series dynamics by capturing complex patterns and dependencies in the data[49]. The PAR model of order d is expressed as Eq. (5)[50]

$$Y_t = \alpha + \sum_{i=1}^{p}\phi_i Y_{t-i} + \sum_{j=1}^{d}\theta_j Y_{t-j}^2 + \cdots \sum_{k=1}^{d}\delta_k Y_{t-k}^d + e_t \tag{5}$$

where $\sum_{j=1}^{d}\theta_j Y_{t-j}^2 + \cdots \sum_{k=1}^{d}\delta_k Y_{t-k}^d$ represents polynomial relationships involving past values of Y_t, including squared, cubed, or higher-degree terms[50,51].

Results

Model order

The workflow for the research presented in this study is summarized in Fig. 2. The ARMAX and NLARX models representing linear and nonlinear autoregressive approaches were employed to forecast future trends in occupational accidents in Turkey, where the SSI's monthly accident records (2013–2023) were organized into four population groups (y_1, y_2, y_3, y_4) as defined in Table 1. The workflow begins by structuring the SSI accident records into a multivariate time series and applying basic preprocessing. The data are then split into estimation and validation sets, after which two modelling paths are pursued: a linear ARMAX model with orders selected using information-based criteria, and a nonlinear NLARX model employing polynomial regressors. Both models are assessed through short- and long-term forecasts using NMSE metrics. The workflow concludes with a comparative evaluation of model performance and interpretation of the dynamic relationships among accident groups.

In line with the system identification workflow adopted in this study, the ARMAX model was estimated, validated, and forecasted using MATLAB's System Identification Toolbox. While Fig. 2 provides a conceptual overview of the modeling pipeline, the detailed implementation steps of data preparation, ARMAX and NLARX model estimation, validation, and performance assessment are formally described in Algorithm 1.

Input: Monthly SSI accident data (2013–2023): y_1, y_2, y_3, y_4
Output: Forecasted accident trends and performance metrics
Data preparation
    Import data, handle missing values, and detrend all series.
    Split data into training and testing subsets.
Linear modelling (ARMAX/ARMA)
    Select AR orders using arxstruc and selstruc .
    Set $n_b = 0$ (no exogenous inputs) and choose MA orders.
    Estimate the multivariate ARMAX model.
    Perform short-term prediction and long-term forecasting.
Nonlinear modelling (NLARX)
    Define linear and polynomial regressors.
    Estimate the NLARX model using nlarx .
    Generate predictions and forecasts.
Model evaluation and interpretation
    Compute NMSE and compare ARMAX and NLARX results.
    Interpret interdependencies among accident categories.

Algorithm 1. ARMAX–NLARX Modelling Workflow

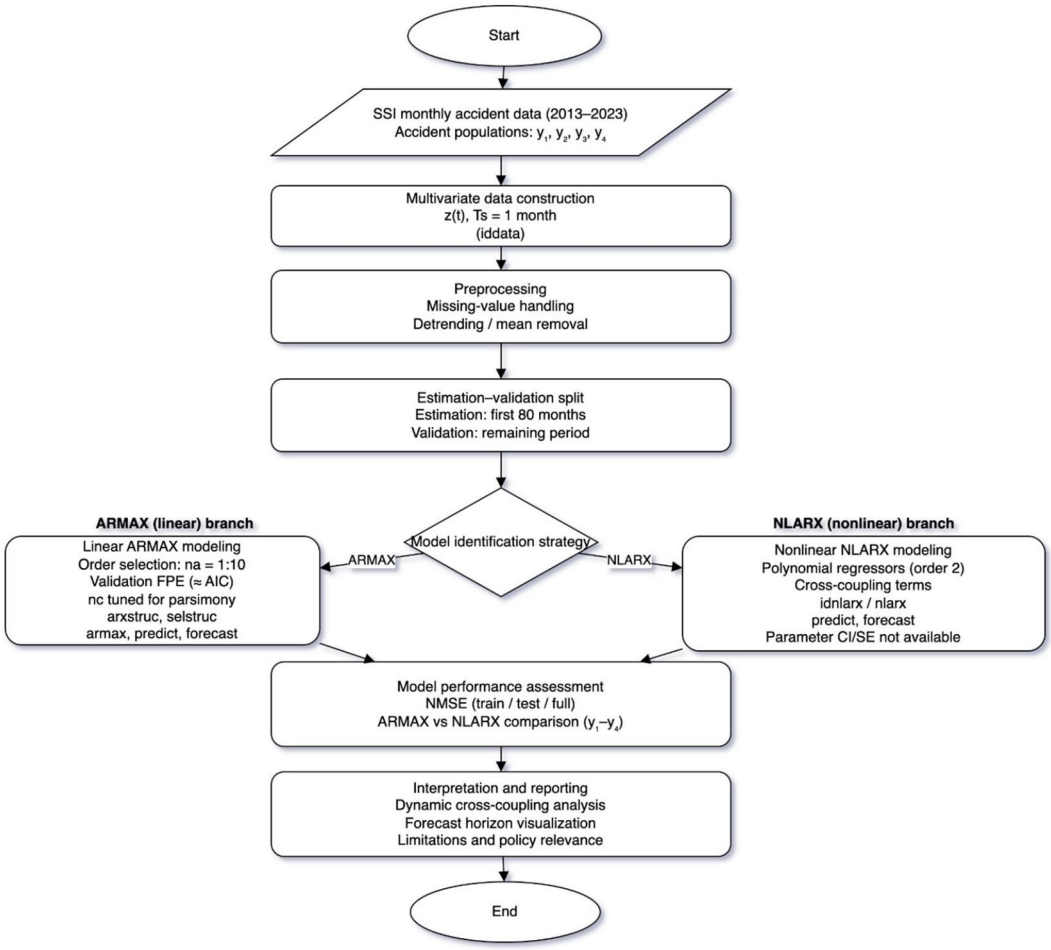


Fig. 2. Minimal workflow of the proposed ARMAX–NLARX modeling framework for multivariate occupational accident time series. Monthly SSI accident data (2013–2023) are organized into a multivariate `iddata` structure, preprocessed, and divided into estimation and validation sets. Linear ARMAX and nonlinear NLARX models are identified in parallel, followed by in-sample prediction and out-of-sample forecasting. Model performance is evaluated using NMSE across all accident populations $(y_1–y_4)$, enabling a systematic comparison of linear and nonlinear dynamics.

As illustrated in Fig. 3, the test NMSE curves for all output variables exhibit a clear stabilization behavior around $N_{\text{train}} = 80$. For smaller training lengths, the NMSE values show pronounced variability, indicating insufficient data to robustly capture the underlying system dynamics. Conversely, increasing the training length beyond $N_{\text{train}} = 80$ does not yield consistent or significant improvements in prediction accuracy across the outputs, and in some cases leads to marginal degradation, suggesting the onset of overfitting or numerical sensitivity. The emergence of this plateau in the test NMSE curves indicates diminishing returns with respect to additional training data. Therefore, $N_{\text{train}} = 80$ is selected as a reliable and well-balanced compromise that ensures stable generalization performance while avoiding unnecessary model complexity.

ARMAX model: numerical evaluation and results

Before interpreting these polynomials, it is important to clarify the structural form of the ARMAX model used in this study. Although the ARMAX framework allows the inclusion of external (exogenous) variables, the present study does not incorporate any exogenous inputs. The dataset obtained from the Social Security Institution (SSI) provides only accident-related populations (y_1, y_2, y_3, y_4) without consistent monthly economic, sectoral, demographic, or regulatory indicators that could serve as external predictors for the 2013–2023 period. Therefore, the ARMAX models used in this study operate solely on internal dynamics autoregressive terms and cross-lag interactions among the accident populations resulting in a multivariate ARMA-type structure without an $X(t)$ component.

The first eighty data points of each population were used as the training set (*ze*), while the remaining data formed the test set (*zv*). To appropriately capture short-term fluctuations and cyclical components in the data, a detrending procedure was applied prior to model estimation.

To determine suitable autoregressive orders n_a, we employed MATLAB's `arxstruc selstruc` routine. This procedure evaluates candidate models with autoregressive orders from 1 to 10 using a cross validation

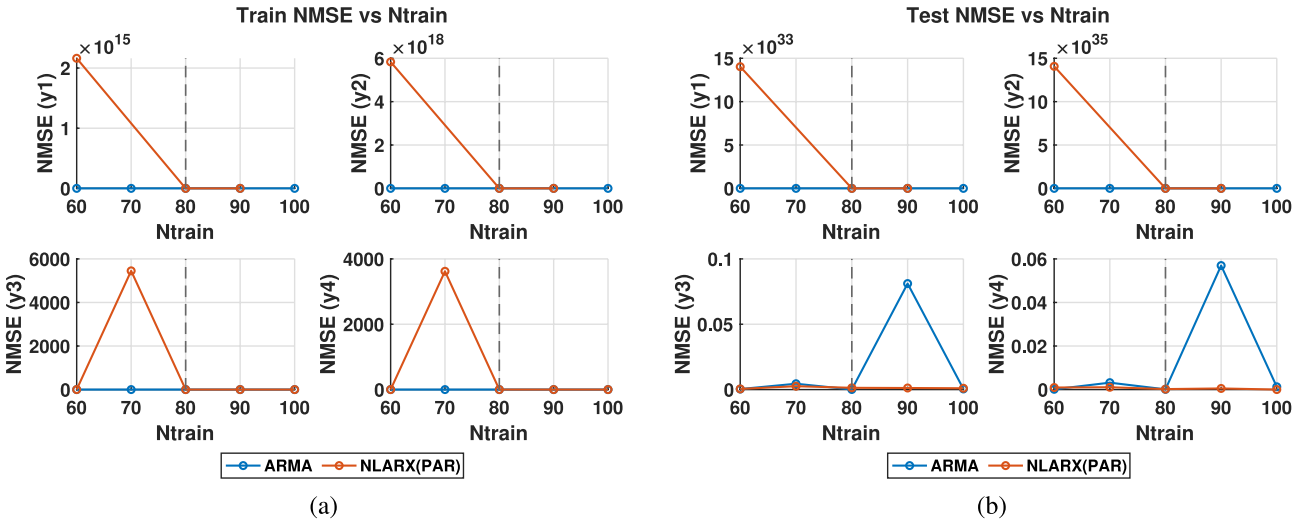


Fig. 3. Variation of the normalized mean squared error (NMSE) with respect to the training length N_{train} for the ARMAX and NLARX models. Panel (a) shows the training NMSE results, while panel (b) presents the test NMSE results for all four outputs (y_1–y_4). The vertical dashed line indicates the selected training length ($N_{\text{train}} = 80$), corresponding to a stable NMSE region across outputs. Test NMSE values are emphasized for model selection, as they better reflect generalization performance.

Population	n_a	n_c
y_1	[5, 4; 4, 5]	[5, 5]
y_2	[5, 4; 4, 5]	[5, 5]
y_3	[10, 9; 6, 7]	[10, 7]
y_4	[10, 9; 6, 7]	[10, 7]

Table 2. Selected ARMAX model orders n_a and n_c.

framework: each candidate structure is estimated on the training subset and evaluated on the validation subset. For every candidate, the Final Prediction Error (FPE) criterion is computed, which for a model with *k* parameters and residual variance $\hat{\sigma}^2$ is defined as:

$$\text{FPE} = \hat{\sigma}^2 \frac{N+k}{N-k}, \tag{6}$$

where *N* denotes the number of estimation samples.

Since FPE is mathematically equivalent to Akaike's Information Criterion (AIC) up to a multiplicative constant, minimizing FPE corresponds to an AIC-based model selection strategy that naturally penalizes overly complex models[52]. Thus, the final n_a values were selected as those minimizing the validation FPE for each population. Following standard practice in multivariate ARMA modeling, the n_c values were chosen to match the complexity of the AR component, ensuring a balanced and identifiable polynomial structure. Since no exogenous input signals are included in the dataset, the n_b term is not applicable and was omitted from the model specification.

The resulting n_a and n_c values used for the four accident populations are reported in Table 2.

The general form of the ARMAX model (multiple-input multiple-output) is shown in Eq. (7)

$$A(z)y(t) = -A_i(z)y_i(z) + C(z)e(t) \tag{7}$$

where $A(z)$ represents the AR polynomial for the populations, and $-A_i(z)y_i(z)$ corresponds to the cross-interactions in the model. Population y_1 is correlated with population y_2. In addition, y_3 and y_4 are correlated. $A_i(z)$ captures these cross-dependencies and reflects the influence of each group populations. The term $C(z)e(t)$ is the error term or noise polynomial, which is part of the MA model. The AR and MA components for the populations y_1, y_2, y_3, and y_4 are provided in Table 3.

In the Table 3, the coefficients in each polynomial are estimated based on the data, and the behavior of the system is modeled by these relationships. To clarify the practical meaning of these coefficients, it is helpful to illustrate how they describe the interaction dynamics among accident populations. For example, in the case of $y_1(t)$, the cross-term $A_2(z) = 0.9995z^{-1}$ indicates that an increase in minor injuries (y_2) in the previous month has an almost one-to-one positive effect on the current value of y_1. This suggests that fluctuations in

Population	Polynomial	Coefficients	Explanation
$y_1(t)$	$A(z)$	$1 - 1.017z^{-1} + 1.016z^{-2}$	Dependency of $y_1(t)$ on its own past values
	$A_2(z)$	$0.9995z^{-1}$	Effect of previous time steps of $y_2(t)$ on $y_1(t)$ (cross-dependency)
	$C(z)$	$1 - 0.7804z^{-1} - 0.1907z^{-2}$	Effect of past error terms
$y_2(t)$	$A(z)$	$1 - 1.017z^{-1} + 1.016z^{-2}$	Dependency of $y_2(t)$ on its own past values
	$A_1(z)$	$0.9995z^{-1}$	Effect of previous time steps of $y_1(t)$ on $y_2(t)$ (cross-dependency)
	$C(z)$	$1 - 0.7804z^{-1} - 0.1907z^{-2}$	Effect of past error terms
$y_3(t)$	$A(z)$	$1 - 2.004z^{-1} + 1.009z^{-2}$	Dependency of $y_3(t)$ on its own past values
	$A_4(z)$	$-0.312z^{-1}$	Effect of previous time steps of $y_4(t)$ on $y_3(t)$ (cross-dependency)
	$C(z)$	$1 - 0.6833z^{-1} - 0.1148z^{-2}$	Effect of past error terms
$y_4(t)$	$A(z)$	$1 - 0.9604z^{-1} - 0.6297z^{-2}$ $+0.1162z^{-3} + 0.5052z^{-4}$	Dependency of $y_4(t)$ on its own past values
	$A_3(z)$	$0.0009556z^{-1} - 0.001454z^{-2}$ $+1.658 \times 10^{-5}z^{-3}$	Effect of previous time steps of $y_3(t)$ on $y_4(t)$ (cross-dependency)
	$C(z)$	$1 + 0.1664z^{-1} - 0.4016z^{-2}$ $-0.5183z^{-3} - 0.2465z^{-4}$	Effect of past error terms

Table 3. Polynomials and coefficients for populations $y_1(t)$, $y_2(t)$, $y_3(t)$, and $y_4(t)$.

y_2 propagate directly into y_1, reflecting strong coupling between these two accident categories. Similarly, the negative coefficient in $A_4(z)$ for the y_3 equation ($A_4(z) = -0.312z^{-1}$) implies that an increase in the number of major injuries (y_4) in the previous period tends to reduce the current value of y_3, capturing an inverse short-term relationship. These examples demonstrate how the AR and cross-interaction terms collectively characterize the dynamic dependencies among accident populations, providing insights into how changes in one category influence the others over time.

In this context, the numerical results of ARMAX model are illustrated in Fig. 4.

In Fig. 4, the ARMAX model results are presented for the outputs y_1, y_2, y_3, and y_4. Each time series is divided into two segments by a dashed vertical line, representing the training and forecasting phases. The blue curve denotes the measured SSI data, while the orange and pink curves correspond to the model prediction during the training phase and the forecast during the testing phase, respectively. The first 80 months of data were used for model training, and the remaining 52 months were reserved for forecasting. The NMSE values for the training, testing, and complete datasets were computed using the `calNMSE` function in MATLAB[53], which evaluates the normalized squared error between the measured and estimated data in Eq. (8).

$$NMSE = \frac{\Sigma(g(x,y)) - f(x,y))^2}{\Sigma(f(x,y))^2} \tag{8}$$

here $g(x, y)$ and $f(x, y)$ denote the actual data and the synthetic data, respectively. The error performance of the ARMAX model for each population is summarized in Table 4. The NMSE values indicate that the model achieves a high level of accuracy during the training (prediction) phase for all outputs. Specifically, the training NMSE values are on the order of 10^{-7} for y_1 and 10^{-5} for y_2, y_3, and y_4, demonstrating an excellent in-sample fit. During the testing phase, which corresponds to multi-step open-loop forecasting, the NMSE values increase as expected, reflecting the increased difficulty of long-horizon prediction. Nevertheless, the test NMSE values remain below 10^{-4} for all populations, indicating that the ARMAX model preserves satisfactory generalization capability across the forecasting interval.

The statistical significance of the estimated ARMAX parameters for the two output groups is summarized in Tables 5 and 6. The ARMAX identification procedure produced a total of 49 parameters for each output-group model. Rather than reporting the full set, the tables list only the coefficients that are statistically significant at the 5% level. Specifically, we retained parameters satisfying $p < 0.05$ and exhibiting 95% confidence intervals that do not include zero. Parameters failing these criteria were omitted because their estimated effects were statistically indistinguishable from zero and therefore provide limited interpretability with respect to the dominant linear temporal structure. For the (y_1, y_2) model, multiple coefficients satisfy the significance criteria and display small p values together with confidence intervals excluding zero, indicating that a subset of the ARMAX terms consistently contributes to the linear temporal dependencies of the outputs. Notably, p_4 and p_{22} have comparatively large negative estimates (approximately -1.65) with tight confidence intervals (e.g., for p_4, $[-2.80, -0.51]$), suggesting persistent negative temporal effects. In contrast, parameters such as p_5, p_{12}, and p_{26} are positive (approximately 0.74 to 1.06), pointing to compensatory linear contributions. Several of these terms exhibit extremely small p values (often on the order of 10^{-5} to 10^{-6}), strengthening the evidence that they represent robust linear structure. At the same time, the presence of additional parameters with comparatively large standard errors (not listed due to the significance filter) suggests a degree of redundancy in the parametrization, which is not uncommon in multivariate ARMAX models fitted to datasets of limited length. In contrast, the (y_3, y_4) model is characterized by fewer but more pronounced significant parameters with substantially larger magnitudes. In particular, p_{18} and p_{19} display dominant coefficients (approximately

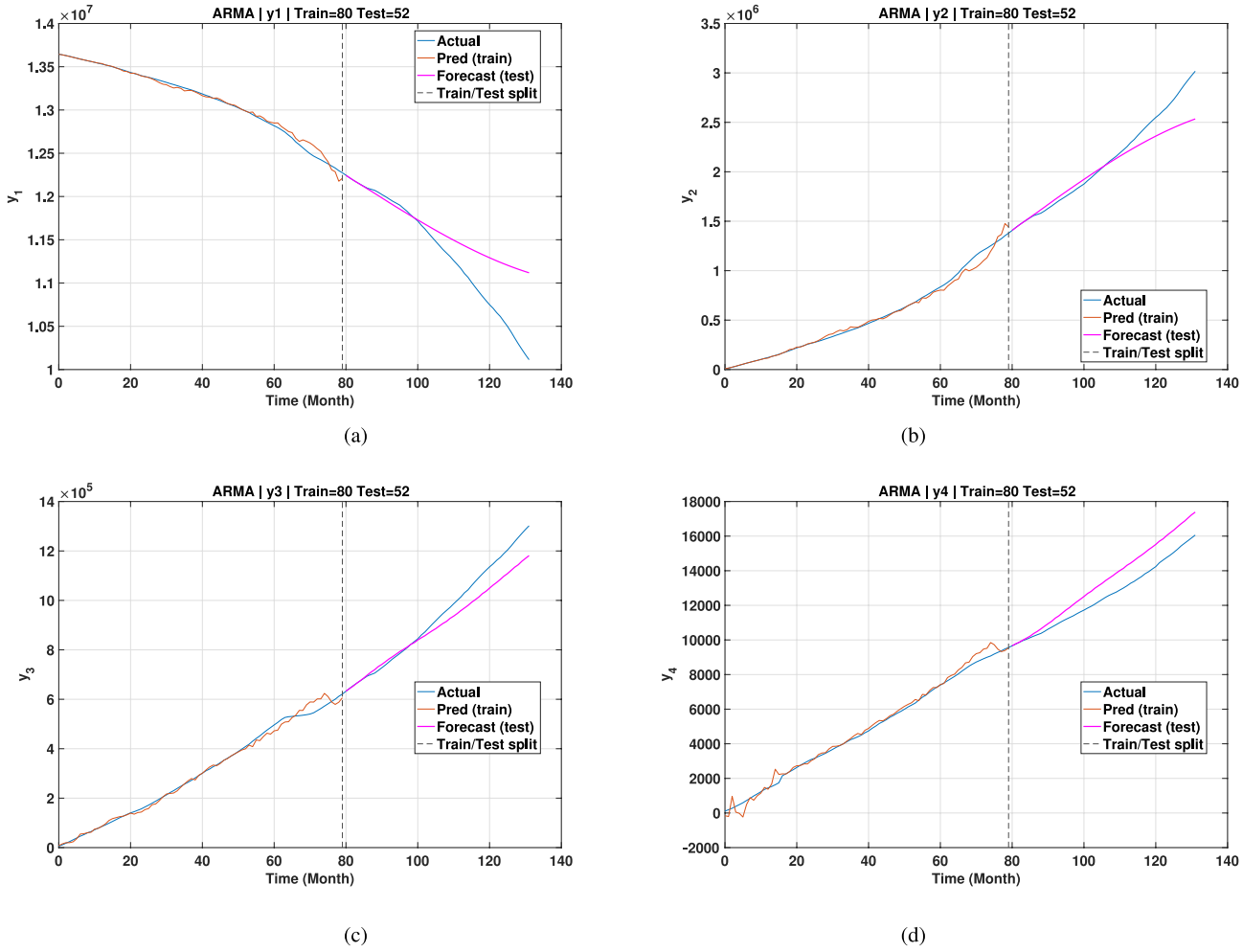


Fig. 4. ARMAX model results for monthly occupational accident data. Panels show the temporal evolution of (**a**) accident-free insured individuals (y_1), (**b**) insured individuals with minor accidents (y_2), (**c**) insured individuals with major accidents (y_3), and (**d**) insured individuals with fatal accidents (y_4). The solid blue lines denote the observed data, the dashed red lines represent in-sample model predictions over the training period (first 80 months), and the solid magenta lines indicate out-of-sample forecasts over the test period (remaining 52 months). The vertical dashed line marks the boundary between the training and test intervals.

Output	Train (Prediction) NMSE	Test (Forecast) NMSE	Full NMSE
y_1	1.15×10^{-7}	2.47×10^{-5}	3.25×10^{-6}
y_2	4.35×10^{-5}	1.16×10^{-4}	4.33×10^{-5}
y_3	2.68×10^{-5}	7.17×10^{-5}	2.68×10^{-5}
y_4	2.99×10^{-5}	1.07×10^{-4}	3.53×10^{-5}

Table 4. ARMAX model performance in terms of NMSE for training prediction and testing multi-step forecasting.

-11.33 and 13.16, respectively) with extremely small p values ($< 10^{-6}$), indicating more clearly identifiable linear dynamics governed by a smaller set of influential temporal dependencies. Additional significant terms (e.g., p_{33}, p_{34}, and p_{35}) exhibit smaller magnitudes but tight confidence intervals, suggesting that the model also captures finer-scale linear effects reliably. Overall, the significance-based parameter analysis supports the suitability of the ARMAX framework for capturing the principal linear characteristics of both output groups, while the differing distributions (in both number and magnitude) of significant coefficients are consistent with the variations observed in their respective forecasting performances.

Parameter	Estimate	Std. Error	$CI_{95\%}$ Low	$CI_{95\%}$ High	*p* value
p_4	− 1.654531	0.582605	− 2.796415	− 0.512646	0.00451
p_5	0.741829	0.182512	0.384112	1.099546	4.81×10^{-5}
p_8	− 0.411854	0.008652	− 0.428811	− 0.394897	$< 10^{-6}$
p_9	0.501119	0.004596	0.492111	0.510127	$< 10^{-6}$
p_{10}	− 1.115868	0.235594	− 1.577624	− 0.654111	2.18×10^{-6}
p_{12}	1.063190	0.397061	0.284965	1.841415	0.00741
p_{17}	− 0.411854	0.009462	− 0.430400	− 0.393308	$< 10^{-6}$
p_{18}	0.501119	0.004907	0.491501	0.510737	$< 10^{-6}$
p_{22}	− 1.654531	0.581963	− 2.795157	− 0.513905	0.00447
p_{23}	0.741829	0.182887	0.383377	1.100282	5.02×10^{-5}
p_{24}	− 1.115868	0.233994	− 1.574488	− 0.657248	2.07×10^{-6}
p_{26}	1.063190	0.395627	0.287776	1.838605	0.00720

Table 5. Statistical significance of ARMA parameters identified for the (y_1, y_2) output pair (only $p < 0.05$).

Parameter	Estimate	Std. Error	$CI_{95\%}$ Low	$CI_{95\%}$ High	*p* value
p_1	− 1.090313	0.444396	− 1.961313	− 0.219313	0.0141
p_6	− 1.018297	0.465850	− 1.931345	− 0.105248	0.0288
p_{18}	− 11.326142	0.391167	− 12.092815	− 10.559469	$< 10^{-6}$
p_{19}	13.155061	0.409644	12.352173	13.957949	$< 10^{-6}$
p_{33}	0.019229	0.005256	0.008928	0.029531	2.54×10^{-4}
p_{34}	− 0.013273	2.669×10^{-4}	− 0.013797	− 0.012750	$< 10^{-6}$
p_{35}	0.002953	1.683×10^{-4}	0.002623	0.003283	$< 10^{-6}$
p_{36}	− 1.324360	0.550075	− 2.402487	− 0.246234	0.0161

Table 6. Statistical significance of ARMA parameters identified for the (y_3, y_4) output pair (only $p < 0.05$).

NLARX model: numerical evaluation and results

NLARX models possess the capability to capture nonlinear relationships between variables. In scenarios where the relationship between occupational accidents and contributing factors such as workplace conditions, safety measures, or employee behavior exhibits nonlinearity, NLARX models are better suited to represent these complexities effectively. The basic formulation used in this model is given in Eq. (10).

$$y(t) = f(y(t-1), y(t-2), \ldots, y(t-n), u(t), u(t-1), \ldots u(t-m)) + e(t) \tag{9}$$

here $y(t)$ represents the number of occupational accident populations at the t time. $f(.)$ is a nonlinear function that maps the past outputs and inputs to the current output. $y(t-1), y(t-2), \ldots, y(t-n)$ are the past values of the output variable, $u(t), u(t-1), \ldots, u(t-m)$ are the exogenous input variables and $e(t)$ is the error term.

NLARX models possess the capability to capture nonlinear relationships between variables. In scenarios where the relationship between occupational accidents and contributing factors such as workplace conditions, safety measures, or employee behavior exhibits nonlinearity, NLARX models are better suited to represent these complexities effectively. The basic formulation used in this model is given in Eq. (10).

$$y(t) = f(y(t-1), y(t-2), \ldots, y(t-n), u(t), u(t-1), \ldots u(t-m)) + e(t) \tag{10}$$

here $y(t)$ represents the number of occupational accident populations at the t time. $f(.)$ is a nonlinear function that maps the past outputs and inputs to the current output. $y(t-1), y(t-2), \ldots, y(t-n)$ are the past values of the output variable, $u(t), u(t-1), \ldots, u(t-m)$ are the exogenous input variables and $e(t)$ is the error term. The exogenous input was set to zero, as no consistent monthly exogenous indicators were reported by SSI.

In this study, the nonlinear function $f(\cdot)$ is implemented using MATLAB's polynomial regression based NLARX structure. No neural networks, wavelet networks, or advanced nonlinear architectures are employed. Instead, the nonlinear mapping is constructed from linear regressors, squared terms (such as $y_i(t-1)^2$), and interaction terms (such as $y_1(t-1)y_2(t-1)y_3(t-1)y_4(t-1)$), enabling the model to capture nonlinear dependencies among the accident populations while maintaining interpretability. During parameter estimation, MATLAB iteratively updates the model coefficients to minimize prediction error, ensuring that the final NLARX model provides an accurate nonlinear representation of the data. Following the definition of the nonlinear

structure and the learning procedure, the NLARX model is applied to the occupational accident dataset as described below.

The NLARX model utilizes the populations y_1, y_2, y_3, and y_4 as described in Table 1. The numerical results of the NLARX model are computed using MATLAB. The first eighty data points of each population are designated as the training set (*ze*), while the remaining data points constitute the test set (*zv*). Additionally, the dataset is de-trended to remove any long-term trends or biases.

In this model, the inputs include the populations y_1, y_2, y_3 and y_4, their squared terms $(y_1(t-1)^2$, $y_2(t-1)^2, y_3(t-1)^2, y_4(t-1)^2)$ and their interaction terms $(y_1(t-1) * y_2(t-1) * y_3(t-1) * y_4(t-1))$. These terms are incorporated into the system to capture potential nonlinear relationships and interactions among the variables. The initial coefficients for the populations y_1 and y_2 were determined as $[-0.394, 0.00565]$, while the coefficients for y_3 and y_4 were initialized as $[0.0028, -2.3175 \times 10^{-4}, -1.1382 \times 10^{-5}, 3.9849 \times 10^{-6}]$. These coefficients were subsequently optimized during the learning process to improve the model's performance. Unlike ARMAX models, NLARX models are identified through nonlinear optimization, and the estimation procedure does not yield an analytical parameter covariance matrix. As a result, standard errors, confidence intervals, and hypothesis tests for individual parameters cannot be computed in a straightforward manner. Therefore, the evaluation of the NLARX models in this study focuses on predictive performance metrics rather than parameter-level statistical inference.

Figure 5 presents the results obtained using the NLARX model. The blue curve corresponds to the measured data, while the orange curve shows the one-step-ahead predictions generated during the training phase using data up to the 80th month. The pink curve represents the multi-step open-loop forecasts produced for the period between the 80th and 132nd months. The prediction performance of the NLARX model is evaluated using

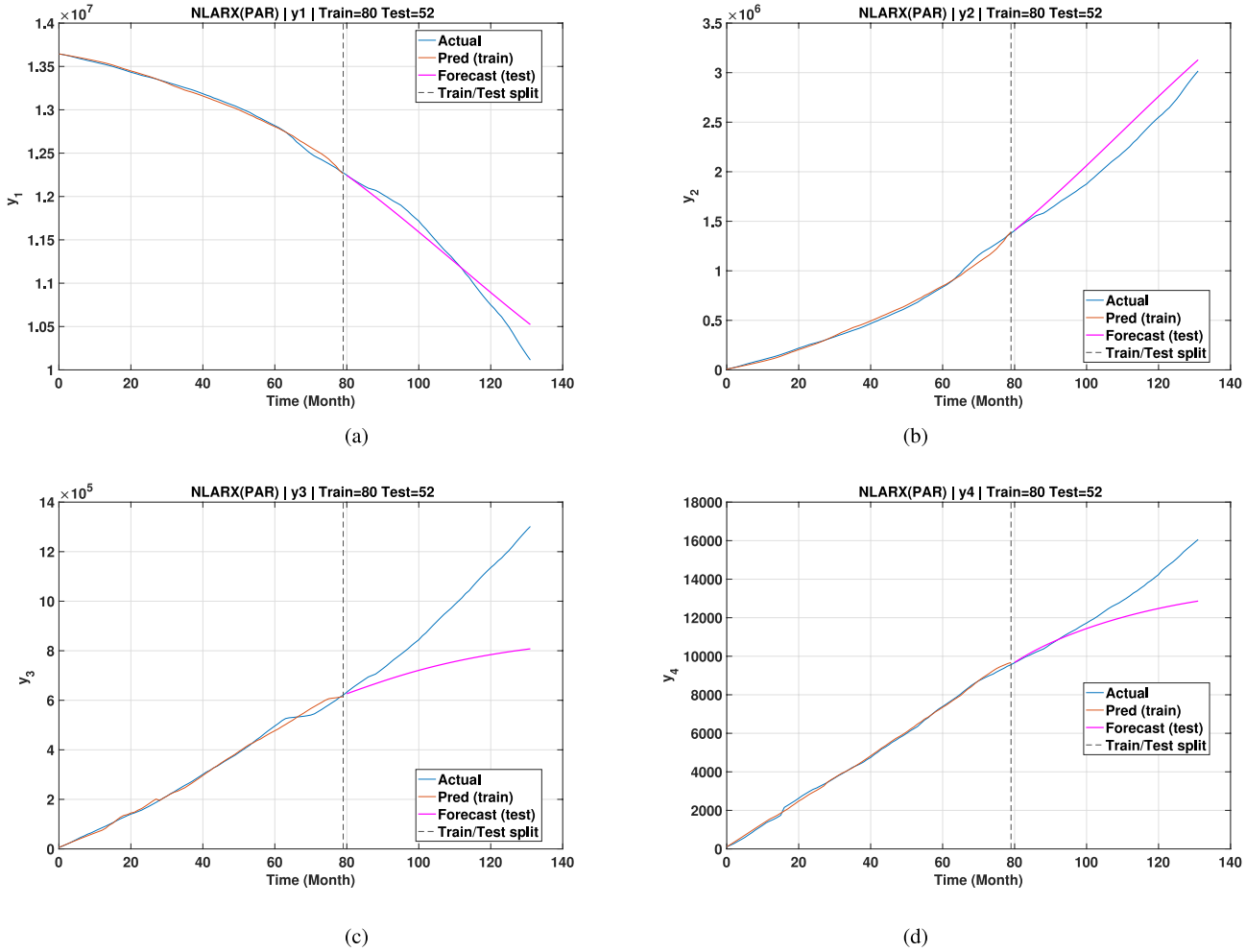


Fig. 5. NLARX (parallel structure) model results for monthly occupational accident data. Panels depict the temporal evolution of (**a**) accident-free insured individuals (y_1), (**b**) insured individuals with minor accidents (y_2), (**c**) insured individuals with major accidents (y_3), and (**d**) insured individuals with fatal accidents (y_4). Solid blue lines represent the observed data, dashed red lines indicate in-sample predictions obtained from the NLARX model over the training period (first 80 months), and solid magenta lines show out-of-sample forecasts over the test period (remaining 52 months). The vertical dashed line denotes the boundary between the training and test intervals.

Output	Train (Prediction) NMSE	Test (Forecast) NMSE	Full NMSE
y_1	5.65×10^{-8}	3.34×10^{-6}	4.56×10^{-7}
y_2	2.15×10^{-5}	1.17×10^{-4}	4.19×10^{-5}
y_3	1.28×10^{-5}	1.34×10^{-3}	4.36×10^{-4}
y_4	3.66×10^{-6}	2.21×10^{-4}	6.74×10^{-5}

Table 7. NLARX model performance in terms of NMSE for training prediction and testing multi-step forecasting.

the NMSE, computed between the measured outputs and the corresponding model responses. As reported in Table 7, the training NMSE values lie in the range of 10^{-8} to 10^{-5}, indicating an excellent in-sample prediction capability of the NLARX model. During the forecasting phase, the NMSE values increase and range from 10^{-6} to 10^{-3}, which is expected due to the accumulation of errors in open-loop multi-step prediction. Despite this increase, the obtained NMSE levels demonstrate that the NLARX model maintains a satisfactory forecasting performance while capturing the underlying nonlinear dynamics of the system.

Conclusion

Building on our previous work utilizing a system of ordinary differential equations based on the SIR model, this study introduces a data-driven, multivariate time-series framework to analyze and forecast occupational accident dynamics in Turkey. In contrast to univariate forecasting approaches, the proposed framework explicitly models cross-lag interactions among accident populations and therefore enables the assessment of interdependencies between accident categories rather than treating each series in isolation. The primary aim was to predict future occupational accident trends by examining the coupled evolution of four accident-related populations (y_1, y_2, y_3, y_4) derived from SSI monthly records (2013–2023).

Consistent with the hierarchical propagation mechanism assumed in our earlier conceptual model, the multivariate identification was implemented through two coupled but separately estimated subsystems: (y_1, y_2) representing the insured–minor accident interaction and (y_3, y_4) representing the major accident–fatality relationship. Within this setting, linear ARMAX models (estimated without exogenous inputs) and polynomial-regression-based NLARX models were constructed and evaluated using the NMSE, computed separately for the training (prediction), test (multi-step forecasting), and full datasets.

The numerical results demonstrate that the ARMAX model provides strong and consistent predictive performance across all populations. In terms of full-data NMSE, ARMAX achieved 3.25×10^{-6} for y_1, 4.33×10^{-5} for y_2, 2.68×10^{-5} for y_3, and 3.53×10^{-5} for y_4, indicating accurate in-sample fitting and stable generalization to the out-of-sample forecasting interval.

In comparison, NLARX achieved its best performance for y_1 with a full-data NMSE of 4.56×10^{-7}, outperforming ARMAX for this population, and produced comparable accuracy for y_2 (4.19×10^{-5}). However, its forecasting errors were substantially higher for y_3 (4.36×10^{-4}) and y_4 (6.74×10^{-5}), revealing reduced robustness for the major and fatal accident dynamics. This discrepancy is consistent with the observation that the (y_3, y_4) subsystem is characterized by fewer but more dominant linear effects, whereas NLARX estimated via nonlinear optimization without an analytical parameter covariance matrix is best interpreted through predictive performance rather than parameter-level statistical inference.

Beyond predictive accuracy, the ARMAX parameter significance analysis provided interpretable evidence of distinct linear structures across the two output groups: the (y_1, y_2) subsystem exhibited a larger number of moderate-magnitude significant coefficients, while the (y_3, y_4) subsystem was governed by fewer but markedly larger-magnitude significant parameters. Taken together, these findings suggest that the dominant temporal dynamics of occupational accident trends in Turkey are largely captured by linear multivariate dependencies, which explains the strong and stable performance of the ARMAX framework across training and forecasting horizons. At the same time, the improved NLARX performance for y_1 indicates that nonlinear components may contribute in more stable accident-related populations, and that nonlinear models can be beneficial in specific cases.

Overall, the results confirm that multivariate ARMAX/NLARX modeling provides an effective data-driven tool for forecasting occupational accident trends and for examining coupled dynamics among accident categories. The ARMAX approach stands out as a robust and interpretable baseline suitable for forecasting and policy planning, while NLARX can complement it by capturing subtle nonlinear patterns where they exist.

These modeling insights may support occupational safety management by enabling more informed planning and targeted interventions. By identifying coupled temporal patterns among accident populations, decision-makers may better prioritize preventive actions such as enhanced safety training, proactive inspections, and optimized resource allocation to mitigate future occupational accidents.

From a policy perspective, the results offer concrete guidance for occupational safety planning. The strong and stable performance of the ARMAX model across all accident populations indicates that linear multivariate forecasts can be used by public authorities as a reliable early-warning tool for monitoring accident trends. In particular, the (y_3, y_4) subsystem, representing major accidents and fatalities, is governed by a small number of high-impact dynamics, suggesting that policy efforts should prioritize targeted inspections, stricter enforcement in high-risk activities, and focused preventive measures rather than broad, uniform interventions. Operationally, upward shifts in forecasted major accidents (y_3) may justify pre-scheduled targeted inspections in high-risk

sectors, while short-term increases in predicted fatality trends (y_4) may warrant intensified supervision, mandatory toolbox trainings, or sector-specific audits during elevated-risk periods.

Despite the strengths of the proposed framework, some limitations should be acknowledged when interpreting the results.Because consistently reported monthly exogenous indicators were not available, the models focus solely on endogenous interactions among accident populations. The moderate length of the SSI time series also places natural constraints on the identification of more complex multivariate especially nonlinear structures. Furthermore, NLARX models do not provide analytical parameter uncertainty measures, which shifts the emphasis toward forecasting performance rather than coefficient-level interpretation.Finally, the analysis is based on national aggregated data; sector- or region-specific datasets may reveal additional patterns not captured here.

Future research could therefore benefit from integrating reliable exogenous indicators (e.g., sector, occupation, experience level, accident cause), employing longer or higher-resolution datasets, and applying the proposed methodology to disaggregated high-risk sectors such as construction and mining, which would enhance both explanatory depth and policy relevance.

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and can be accessed via the official database of the Turkish Social Security Institution (SSI) at the following link: https://www.sgk.gov.tr/Istatistik/Yillik/fcd5e59b-6af9-4d90-a451-ee7500eb1cb4

Received: 29 September 2025; Accepted: 9 January 2026
Published online: 19 January 2026

References

1. Zanko, M. & Dawson, P. Occupational health and safety management in organizations: A review. *Int. J. Manag. Rev.* **14**, 328–344. https://doi.org/10.1111/j.1468-2370.2011.00319.x (2012).
2. World Health Organization. Occupational health. https://www.who.int/health-topics/occupational-health (n.d). Accessed 20 February 2025.
3. Suharni, S. Prioritizing occupational safety, health, and wellbeing; An analysis of imperative public health trends. *Health Educ. Health Promot.* **12**, 243–254. https://doi.org/10.58209/hehp.12.2.243 (2024).
4. Ciarapica, F. E. & Giacchetta, G. Classification and prediction of occupational injury risk using soft computing techniques: An Italian study. *Saf. Sci.* **47**, 36–49. https://doi.org/10.1016/j.ssci.2008.01.006 (2009).
5. Attwood, D., Khan, F. & Veitch, B. Occupational accident models-where have we been and where are we going?. *J. Loss Prev. Process Ind.* **19**, 664–682. https://doi.org/10.1016/j.jlp.2006.02.001 (2006).
6. Andersson, R. & Lagerlöf, E. Accident data in the new Swedish information system on occupational injuries. *Ergonomics* **26**, 33–42. https://doi.org/10.1080/00140138308963310 (1983).
7. Eker, H. Natural language processing risk assessment application developed for marble quarries. *Appl. Sci.* **14**, 9045 (2024).
8. George, M. R., Nalluri, M. R. & Anand, K. Application of ensemble machine learning for construction safety risk assessment. *J. Inst. Eng. (India) Ser. A* **103**, 989–1003 (2022).
9. Deng, F., Gu, W., Zeng, W., Zhang, Z. & Wang, F. Hazardous chemical accident prevention based on k-means clustering analysis of incident information. *IEEE Access* **8**, 180171–180183 (2020).
10. Zhang, L., Wu, X., Skibniewski, M. J., Zhong, J. & Lu, Y. Bayesian-network-based safety risk analysis in construction projects. *Reliab. Eng. Syst. Saf.* **131**, 29–39 (2014).
11. Zhong, B., Pan, X., Love, P. E., Ding, L. & Fang, W. Deep learning and network analysis: Classifying and visualizing accident narratives in construction. *Autom. Constr.* **113**, 103089 (2020).
12. Tingjiang, T., Enyuan, W., Ke, Z. & Changfang, G. Research on assisting coal mine hazard investigation for accident prevention through text mining and deep learning. *Resour. Policy* **85**, 103802 (2023).
13. Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. *Time Series Analysis: Forecasting and Control* 5th edn. (Wiley, 2015).
14. Rajaprasad, S. Prediction of fatal accidents in Indian factories based on Arima. *Prod. Eng. Arch.* **18**, 24–30. https://doi.org/10.30657/pea.2018.18.04 (2018).
15. Shah, K. S., Khan, S., Abbass, N., Rahman, A. & Khan, N. M. Analysis and forecast of mining fatalities in Cherat coal field, Pakistan: Analysis and forecast of mining fatalities in Cherat coal field, Pakistan. *Int. J. Econ. Environ. Geol.* **11**, 22–26. https://doi.org/10.46660/ijeeg.v11i4.293 (2020).
16. Verma, A., Dhalmahapatra, K. & Maiti, J. Forecasting occupational safety performance and mining text-based association rules for incident occurrences. *Saf. Sci.* **159**, 106014. https://doi.org/10.1016/j.ssci.2022.106014 (2023).
17. Grigonytė, E. & Butkevičiūtė, E. Short-term wind speed forecasting using arima model. *Energetika* **62** (2016).
18. Cho, M., Hwang, J. & Chen, C. Customer short term load forecasting by using Arima transfer function model. In *Proceedings 1995 International Conference on Energy Management and Power Delivery EMPD'95*, vol. 1, 317–322 (IEEE, 1995).
19. Stergiou, K. Short-term fisheries forecasting: comparison of smoothing, arima and regression techniques. *J. Appl. Ichthyol.* **7**, 193–204 (1991).
20. Mannering, F. L. & Bhat, C. R. Analytic methods in accident research: Methodological frontier and future directions. *Anal. Methods Accident Res.* **1**, 1–22 (2014).
21. Carnero, M. C. & Pedregal, D. J. Modelling and forecasting occupational accidents of different severity levels in Spain. *Reliab. Eng. Syst. Saf.* **95**, 1134–1141 (2010).
22. Zhou, C., Ding, L., Skibniewski, M. J., Luo, H. & Jiang, S. Characterizing time series of near-miss accidents in metro construction via complex network theory. *Saf. Sci.* **98**, 145–158. https://doi.org/10.1016/j.ssci.2017.06.012 (2017).
23. Bairami-Khankandi, S., Bolbot, V., BahooToroody, A. & Goerlandt, F. A systems-theoretic approach using association rule mining and predictive bayesian trend analysis to identify patterns in maritime accident causes. *Reliab. Eng. Syst. Saf.* 110911. https://doi.org/10.1016/j.ress.2025.110911 (2025).
24. Melchior, C., Zanini, R. R., Guerra, R. R. & Rockenbach, D. A. Forecasting Brazilian mortality rates due to occupational accidents using autoregressive moving average approaches. *Int. J. Forecast.* **37**, 825–837 (2021).
25. Cao, Z., Zhou, T., Miao, S., Wang, L. & Wang, Z. Exploring the economic occupational health, safety, and fatal accidents in high-risk industries. *BMC Public Health* **25**, 433 (2025).
26. Ljung, L. *System Identification* (Linkoping University, 1995).
27. Kaplanvural, S., Tosyalı, E. & Ekmekçi, İ. A new approach to determine occupational accident dynamics by using ordinary differential equations based on sir model. *Sci. Rep.* **14**, 23985. https://doi.org/10.1038/s41598-024-74938-9 (2024).

28. Unsar, S. & Sut, N. General assessment of the occupational accidents that occurred in turkey between the years 2000 and 2005. *Saf. Sci.* **47**, 614–619. https://doi.org/10.1016/j.ssci.2008.08.001 (2009).
29. Social Security Provincial/Central Directorates, D. o. A., General Directorate of Service Delivery & Management, F. Work accident and occupational diseases statistics.
30. Ceylan, H., Kaplan, A. & Bekar, M. High-risky sectors in terms of work accidents in turkey. *Int. J. Eng. Res. Dev.* **14**, 45–57. https://doi.org/10.29137/umagd.937910 (2022).
31. Öney, Ö., Samanlı, S. & Özmen, S. Madencilik sektöründeki ölümlü iş kazalarının analizi. *Karaelmas J. Occup. Health Saf.* **2**, 53–61. https://doi.org/10.33720/kisgd.479803 (2018).
32. Ceylan, H. Türkiye'de inşaat sektöründe meydana gelen iş kazalarının analizi. *Int. J. Eng. Res. Dev.* **6**, 1–6. https://doi.org/10.29137/umagd.346068 (2014).
33. Korkmaz, A. & Avsallı, H. Çalışma hayatında yeni bir dönem: 6331 sayılı iş sağlığı ve güvenliği yasası. *Süleyman Demirel Üniversitesi Fen-Edebiyat Fakültesi Sosyal Bilimler Dergisi* **2012**, 153–167 (2012).
34. Social Security Provincial/Central Directorates, General Directorate of Service Delivery, Department of Actuarial and Fund Management. https://www.sgk.gov.tr/Istatistik/Yillik/fcd5e59b-6af9-4d90-a451-ee7500eb1cb4 (2023). Accessed 06.02.2023.
35. Akgül, I. Time series analysis and forecasting models. *Recommend. Marmara Univ. Soc. Sci. Inst. J.* **1**, 52–69 (1994).
36. Nerlove, M., Grether, D. M. & Carvalho, J. L. *Analysis of Economic Time Series: A Synthesis* 1st edn. (Academic Press, 2014).
37. Kim, Y. S., Rachev, S. T., Bianchi, M. L., Mitov, I. & Fabozzi, F. J. Time series analysis for financial market meltdowns. *J. Bank. Finance* **35**, 1879–1891. https://doi.org/10.1016/j.jbankfin.2010.12.007 (2011).
38. Baek, J., Cho, Y. & Koo, W. W. The environmental consequences of globalization: A country-specific time-series analysis. *Ecol. Econ.* **68**, 2255–2264. https://doi.org/10.1016/j.ecolecon.2009.02.021 (2009).
39. Yule, G. U. VII. On a method of investigating periodicities disturbed series, with special reference to wolfer's sunspot numbers. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Papers Math. Phys. Char.* **226**, 267–298. https://doi.org/10.1098/rsta.1927.0007 (1927).
40. Kaur, J., Parmar, K. S. & Singh, S. Autoregressive models in environmental forecasting time series: a theoretical and application review. *Environ. Sci. Pollut. Res.* **30**, 19617–19641. https://doi.org/10.1007/s11356-023-25148-9 (2023).
41. Patterson, K. An introduction to arma models. In *Unit Root Tests in Time Series: Key Concepts and Problems*, 68–122 (Springer, 2011).
42. Makridakis, S. & Hibon, M. Arma models and the box-Jenkins methodology. *Journal of forecasting***16**, 147–163. https://doi.org/10.1002/(SICI)1099-131X(199705)16:3<147::AID-FOR652>3.0.CO;2-X (1997).
43. Jadhav, V., Chinnappa, R. B. & Gaddi, G. Application of Arima model for forecasting agricultural prices. *J. Agricult. Sci. Technol. (JAST)* **19**, 981–992 (2017).
44. Nihan, N. L. & Holmesland, K. O. Use of the box and Jenkins time series technique in traffic forecasting. *Transportation* **9**, 125–143. https://doi.org/10.1007/BF00167127 (1980).
45. Prasad, G. L. V., Teja, B. R., Govathoti, S. & Dhanikonda, S. Leveraging arma and armax time-series forecasting models for rainfall prediction. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 353–357 (IEEE, 2023).
46. Likothanassis, S. & Demiris, E. Armax model identification with unknown process order and time-varying parameters. In *Signal Analysis and Prediction*, 175–184 (Springer, 1998).
47. Akal, M. Forecasting turkey's tourism revenues by armax model. *Tour. Manag.* **25**, 565–580. https://doi.org/10.1016/j.tourman.2003.08.001 (2004).
48. Christini, D. J. et al. Application of linear and nonlinear time series modeling to heart rate dynamics analysis. *IEEE Trans. Biomed. Eng.* **42**, 411–415. https://doi.org/10.1109/10.376135 (1995).
49. Karakuş, O., Kuruoğlu, E. E. & Altınkaya, M. A. One-day ahead wind speed/power prediction based on polynomial autoregressive model. *IET Renew. Power Gener.* **11**, 1430–1439. https://doi.org/10.1049/iet-rpg.2016.0972 (2017).
50. Kuruoğlu, E. E. Nonlinear least LP-norm filters for nonlinear autoregressive α-stable processes. *Digital Signal Process.* **12**, 119–142. https://doi.org/10.1006/dspr.2001.0416 (2002).
51. Koh, T. & Powers, E. Second-order volterra filtering and its application to nonlinear system identification. *IEEE Trans. Acoust. Speech Signal Process.* **33**, 1445–1455 (1985).
52. Akaike, H. Fitting autoregreesive models for prediction. In *Selected Papers of Hirotugu Akaike*, 131–135 (Springer, 1969).
53. Ionescu, C. M., Birs, I. R., Copot, D., Muresan, C. I. & Caponetto, R. Mathematical modelling with experimental validation of viscoelastic properties in non-Newtonian fluids. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **378**, 20190284. https://doi.org/10.1098/rsta.2019.0284 (2020) https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2019.0284/1318028/rsta.2019.0284.pdf.

Acknowledgements

We would like to thank Turkish Social Security Institution to share and release accidental data regularly on their website.

Author contributions

S.K.: Writing—original draft, Data Curation, Methodology, Software, Visualization. E.T.: Methodology, Software,Supervision, Conceptualization. I.E: Writing—review and editing, Supervision, Formal analysis.

Funding

No funding was obtained for this study.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

This study did not require ethical approval as it utilized publicly available data from the Turkish Social Security Institution (SSI), which does not include any personal or identifiable information.

Additional information

Correspondence and requests for materials should be addressed to S.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026